# Theme Transformer

## Symbolic Music Generation with Theme-Conditioned Transformer

*IEEE Transactions on Multimedia*

**Yi-Jen Shih**[1,2], Shih-Lun Wu[1,2], Frank Zalkow[3], Meinard Müller[3], Yi-Hsuan Yang[2]

[1]National Taiwan University, Taiwan
[2]Academia Sinica, Taiwan
[3]International Audio Laboratories Erlangen, Germany

# About me

- Ian Shih
- B.S. in Electrical Engineering at National Taiwan University
- Part-time Research Assistant at Music and AI Lab
- Love playing some improvisation on piano (SoundCloud)
- Research Interest:
  - Music Generation (Prof.  Yi-Hsuan Yang)
  - Visual Grounded Speech Models (Prof. Hung-Yi Lee)
- Website: atosystem.github.io

# Outline

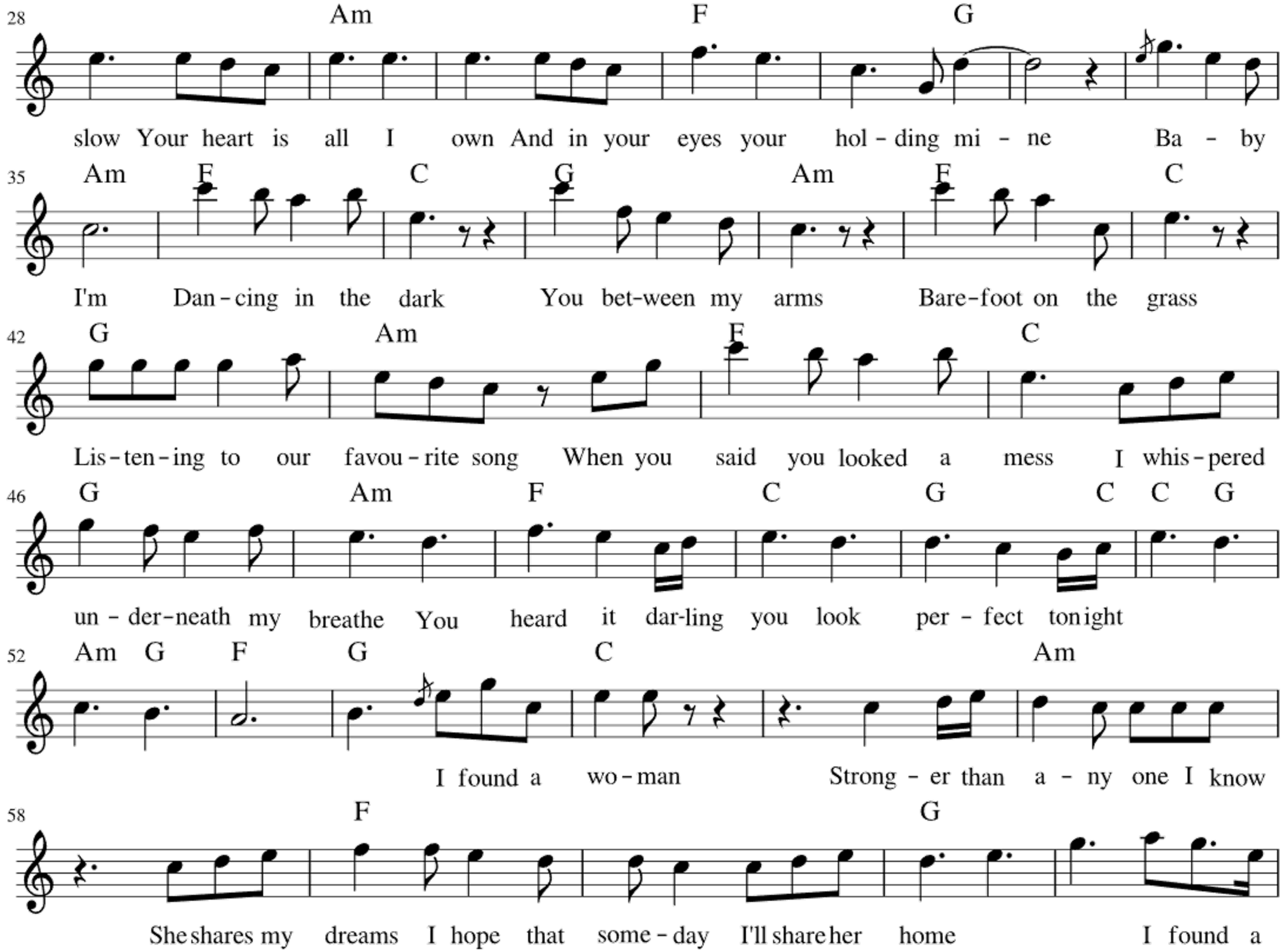
- Overview
- Technical Background
- Results
- Conclusion & Contribution

# Outline

- Overview
- Technical Background
- Results
- Conclusion & Contribution

# Overview

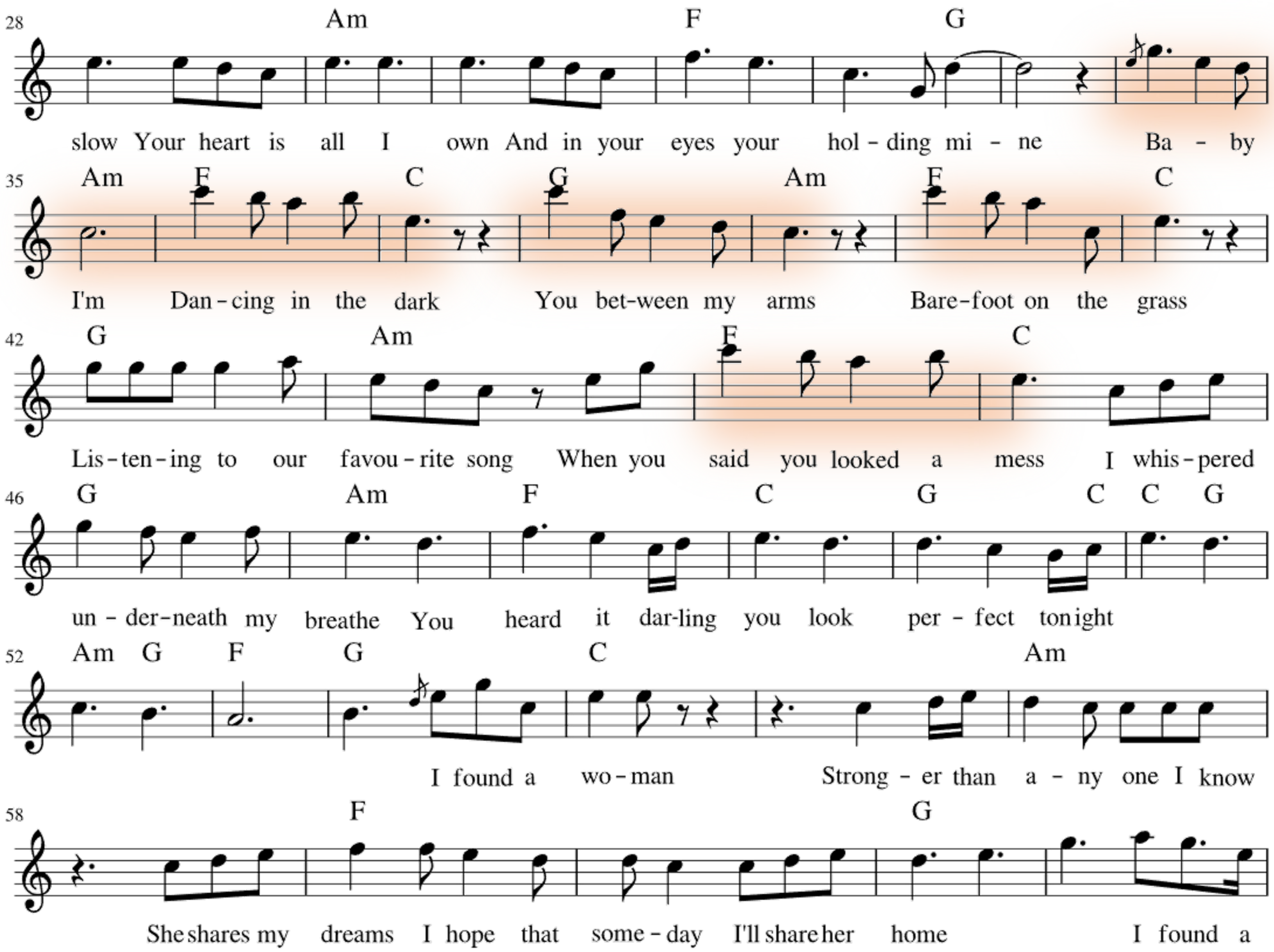
Excerpt from Perfect – Ed Sheeran



28
Am F G
slow Your heart is all I own And in your eyes your hol-ding mi - ne Ba - by
35
Am F C G Am F C
I'm Dan-cing in the dark You bet-ween my arms Bare-foot on the grass
42
G Am F C
Lis-ten-ing to our favou-rite song When you said you looked a mess I whis-pered
46
G Am F C G C C G
un-der-neath my breathe You heard it dar-ling you look per-fect tonight
52
Am G F G C Am
I found a wo-man Strong - er than a - ny one I know
58
F G
She shares my dreams I hope that some-day I'll share her home I found a

# Overview - Theme

- Themes
- *Sequentia*
- Motivic Development
- Music Expectancy



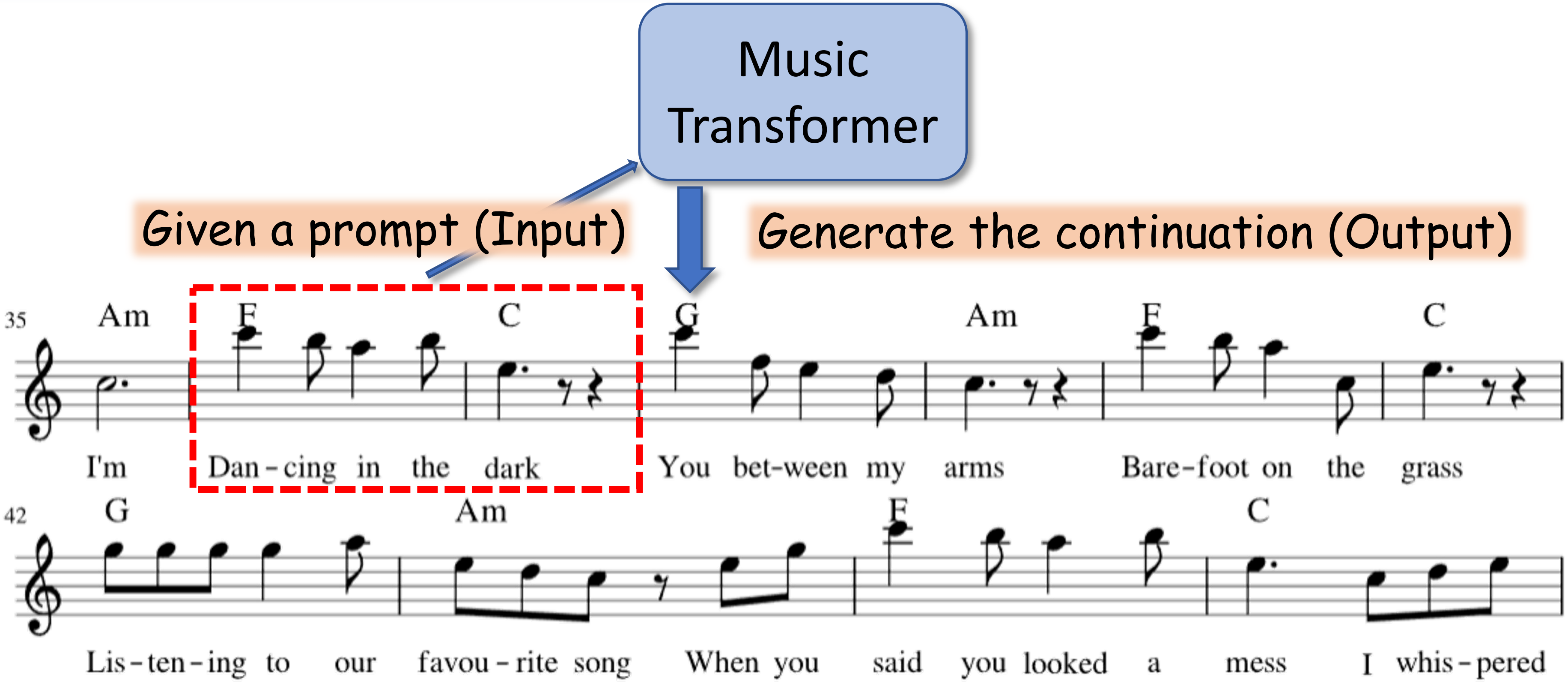
28
Am F G
slow Your heart is all I own And in your eyes your hol - ding mi - ne Ba - by
35
Am F C G Am F C
I'm Dan - cing in the dark You bet-ween my arms Bare-foot on the grass
42
G Am F C
Lis - ten - ing to our favou - rite song When you said you looked a mess I whis - pered
46
G Am F C G C C G
un - der-neath my breathe You heard it dar-ling you look per - fect ton ight
52
Am G F G C Am
I found a wo - man Strong - er than a - ny one I know
58
F G
She shares my dreams I hope that some - day I'll share her home I found a

# Theme is crucial in music composition

But how do recent model generate music?

# Overview

## Prompt Conditioned Music Generation



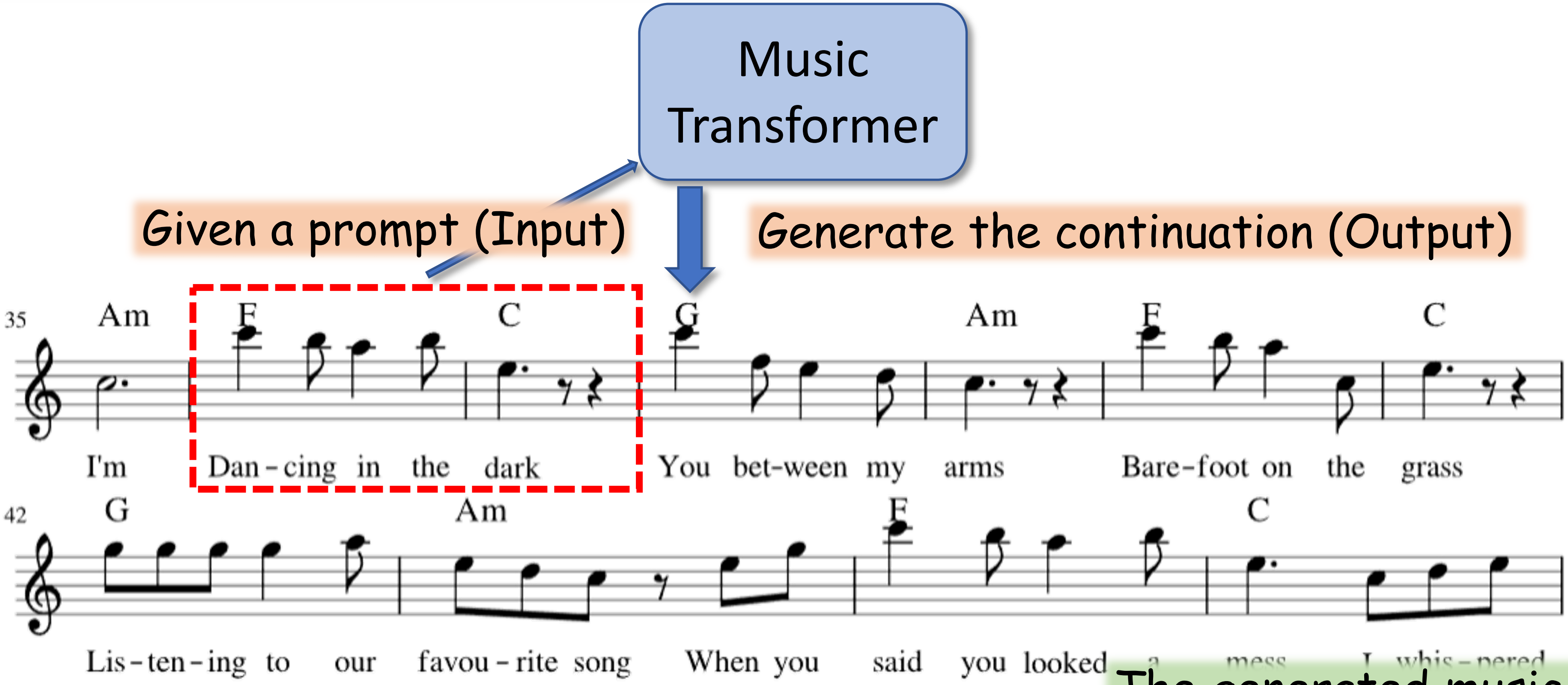
Music Transformer

Given a prompt (Input)

Generate the continuation (Output)

# Overview

## Prompt Conditioned Music Generation



Given a prompt (Input)

Generate the continuation (Output)

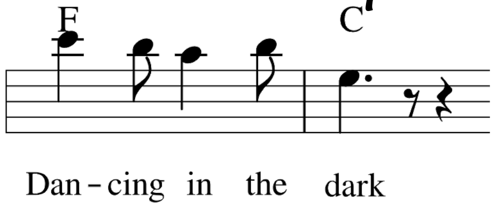
The generated music has no knowledge of theme *variation* and *repetitions*

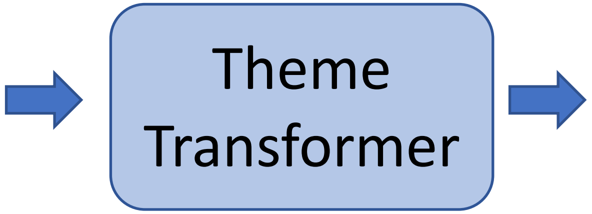
# How to teach models to compose music base on a given Theme?

# Overview

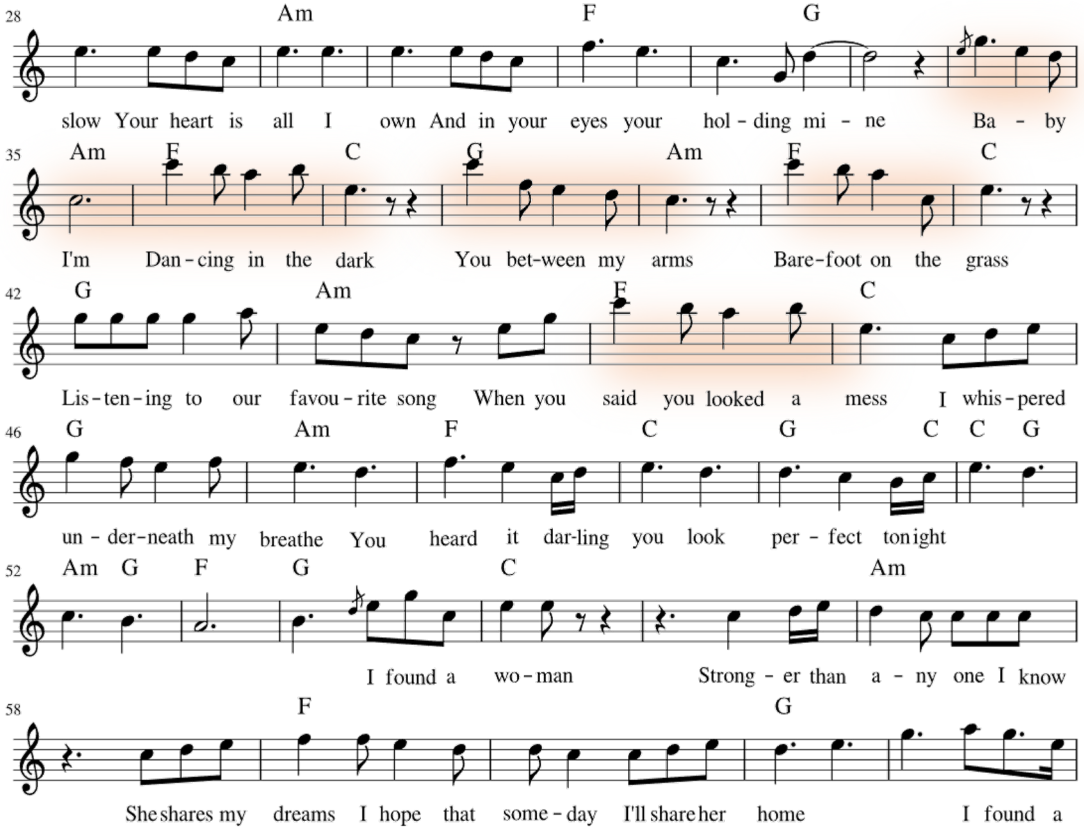
**Theme Conditioned Music Generation**

*Theme (Input)*




Theme Transformer

*Entire Song (Output)*

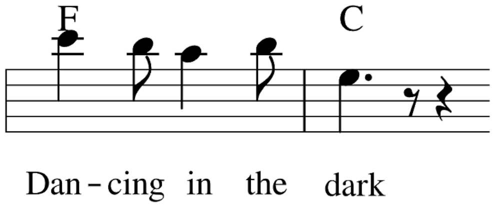

28
Am F G
slow Your heart is all I own And in your eyes your hol-ding mi-ne Ba-by
35
Am F C G Am F C
I'm Dan-cing in the dark You bet-ween my arms Bare-foot on the grass
42
G Am F C
Lis-ten-ing to our favou-rite song When you said you looked a mess I whis-pered
46
G Am F C G C C G
un-der-neath my breathe You heard it dar-ling you look per-fect tonight
52
Am G F G C Am
I found a wo-man Strong-er than a-ny one I know
58
F G
She shares my dreams I hope that some-day I'll share her home I found a

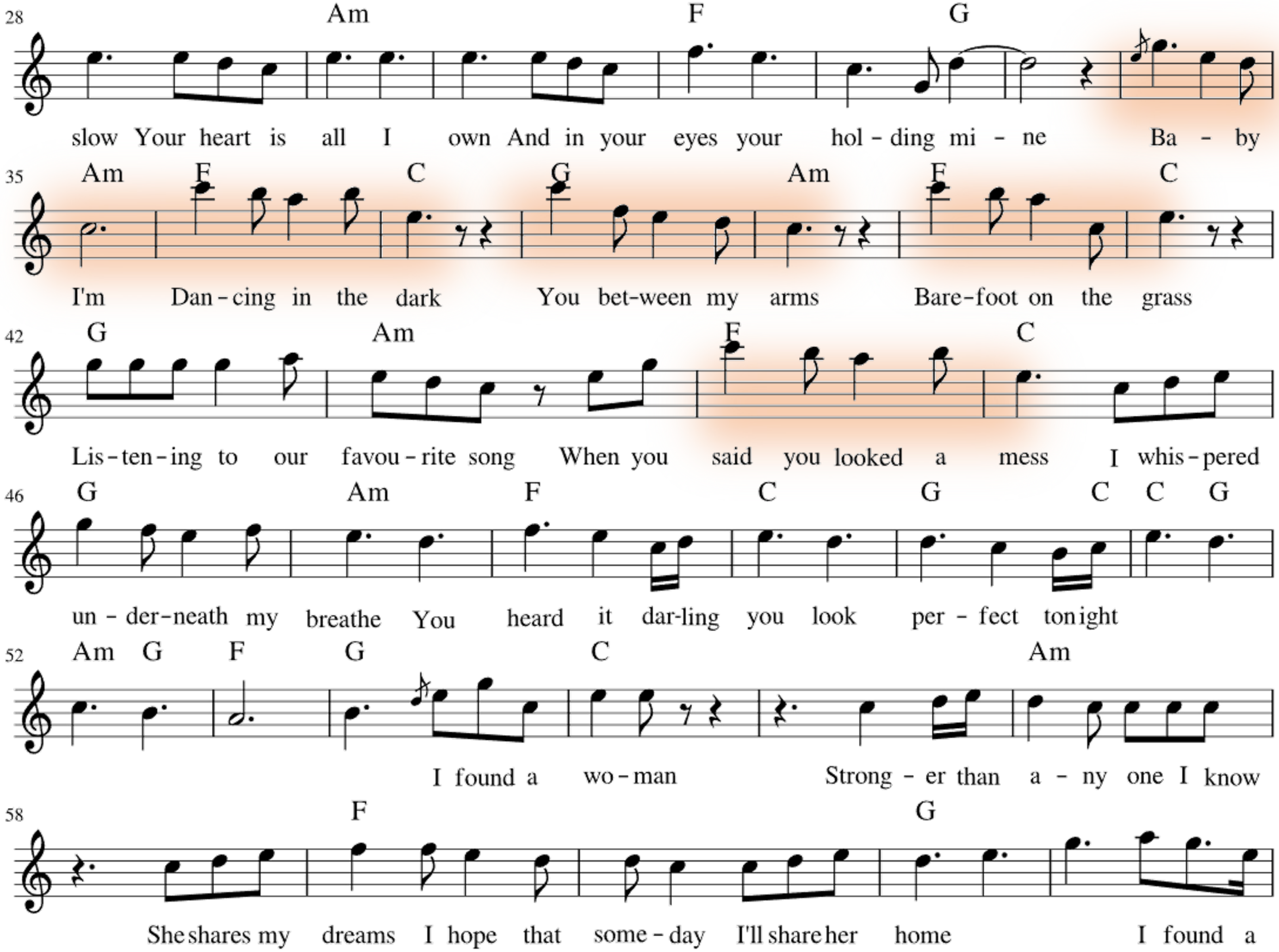
# Overview - Difficulties

- Definition of Musical Theme is quite ***ambiguous***
- ***Lack of Dataset*** for Musical Theme Annotations
- Recent Music Generation Models have problems recognizing “**Theme**”, not to mention ***variations*** and ***repetitions***

# Overview

## Theme Retrieval

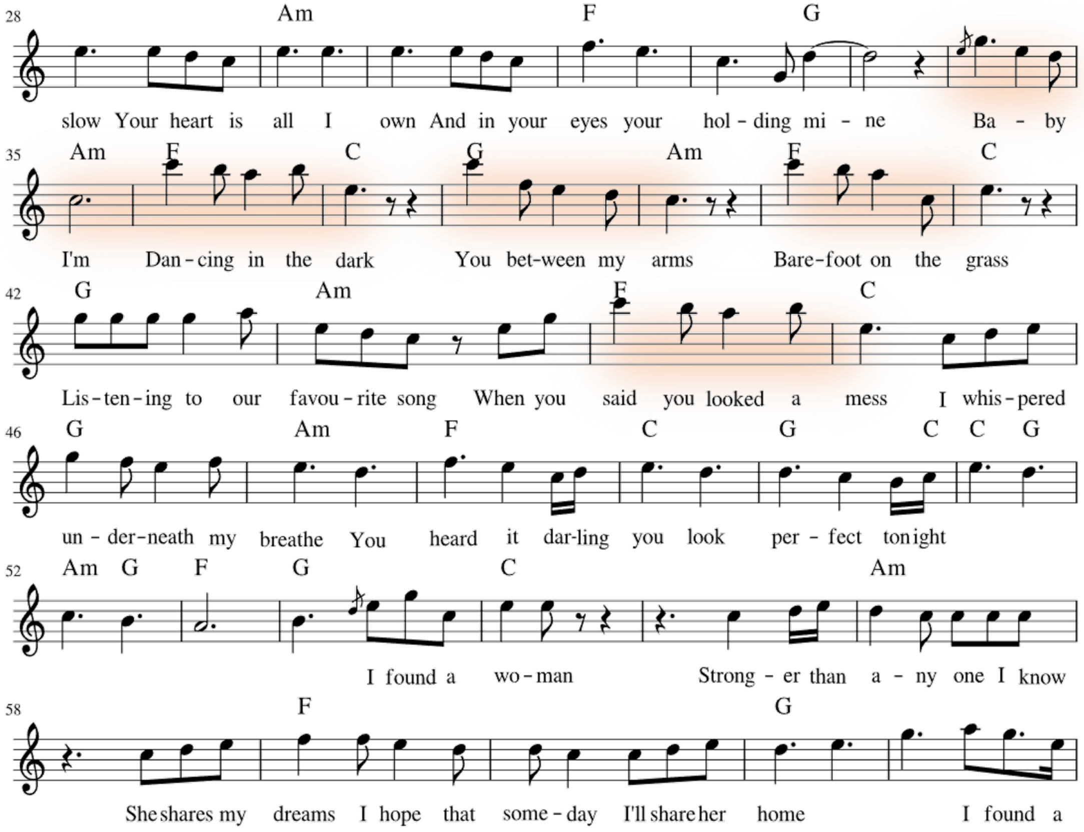



28 Am F G
slow Your heart is all I own And in your eyes your hol-ding mi-ne Ba-by
35 Am F C G Am F C
I'm Dan-cing in the dark You bet-ween my arms Bare-foot on the grass
42 G Am F C
Lis-ten-ing to our favou-rite song When you said you looked a mess I whis-pered
46 G Am F C G C C G
un-der-neath my breathe You heard it dar-ling you look per-fect tonight
52 Am G F G C Am
I found a wo-man Strong-er than a-ny one I know
58 F G
She shares my dreams I hope that some-day I'll share her home I found a

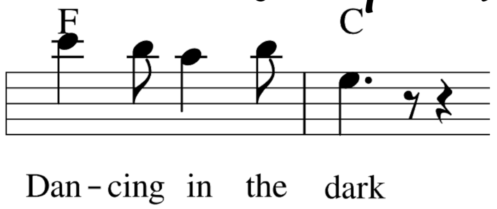
# Overview

**Theme Conditioned Music Generation**

Entire Song (Output)


28 Am F G
slow Your heart is all I own And in your eyes your hol - ding mi - ne Ba - by
35 Am F C G Am F C
I'm Dan - cing in the dark You bet-ween my arms Bare-foot on the grass
42 G Am F C
Lis - ten - ing to our favou - rite song When you said you looked a mess I whis - pered
46 G Am F C G C C G
un - der-neath my breathe You heard it dar-ling you look per - fect tonight
52 Am G F G C Am
I found a wo - man Strong - er than a - ny one I know
58 F G
She shares my dreams I hope that some - day I'll share her home I found a

Theme (Input)



Theme Transformer

# Outline

- Overview
- Technical Background
    - Theme Retrieval
    - Theme based Music Generation
- Results
- Conclusion & Contribution

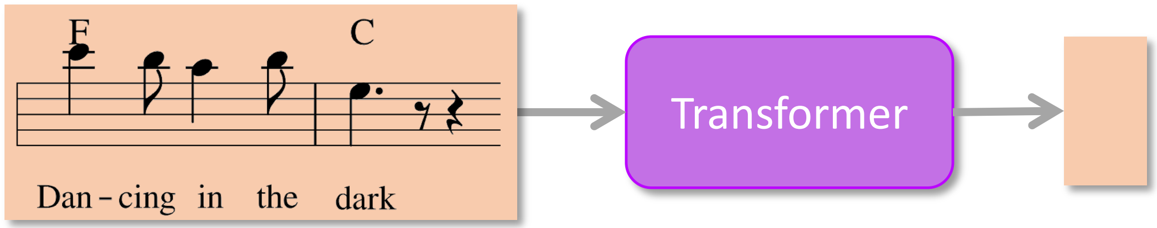
# Theme Retrieval

- Previous Works
  - String-based
    - Correlative matrix (Hsu et al., 2001)
  - Geometric-based
    - COSIATEC (Meredith et al., 2010)
    - RECURSIA-RRT (Meredith, 2019)
- Requires hyperparameters tuning and prone to noise in data

# Theme Retrieval

## Encode Melody into Vector Space

***Music Segment***



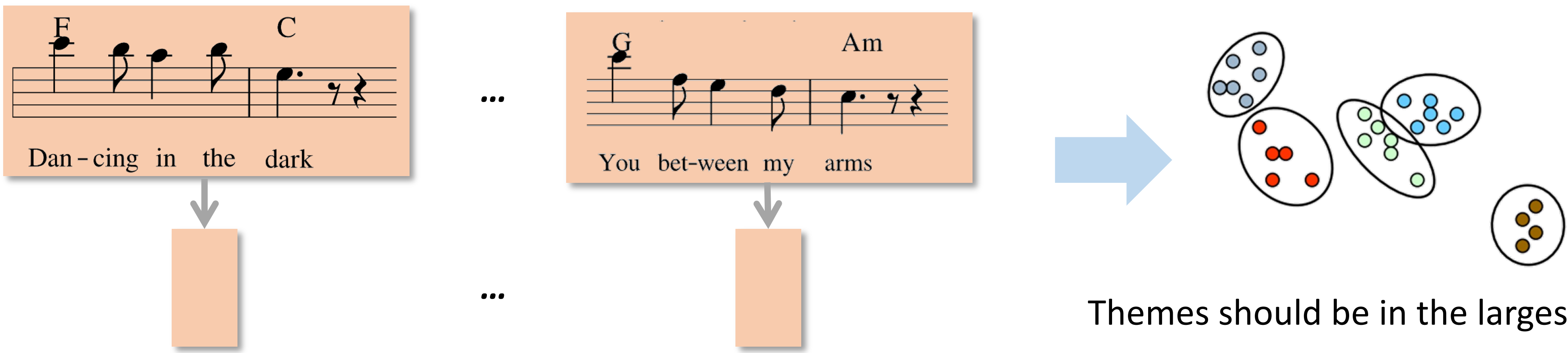
Transformer

***Vector***

## Extract Theme

***Music Segments***

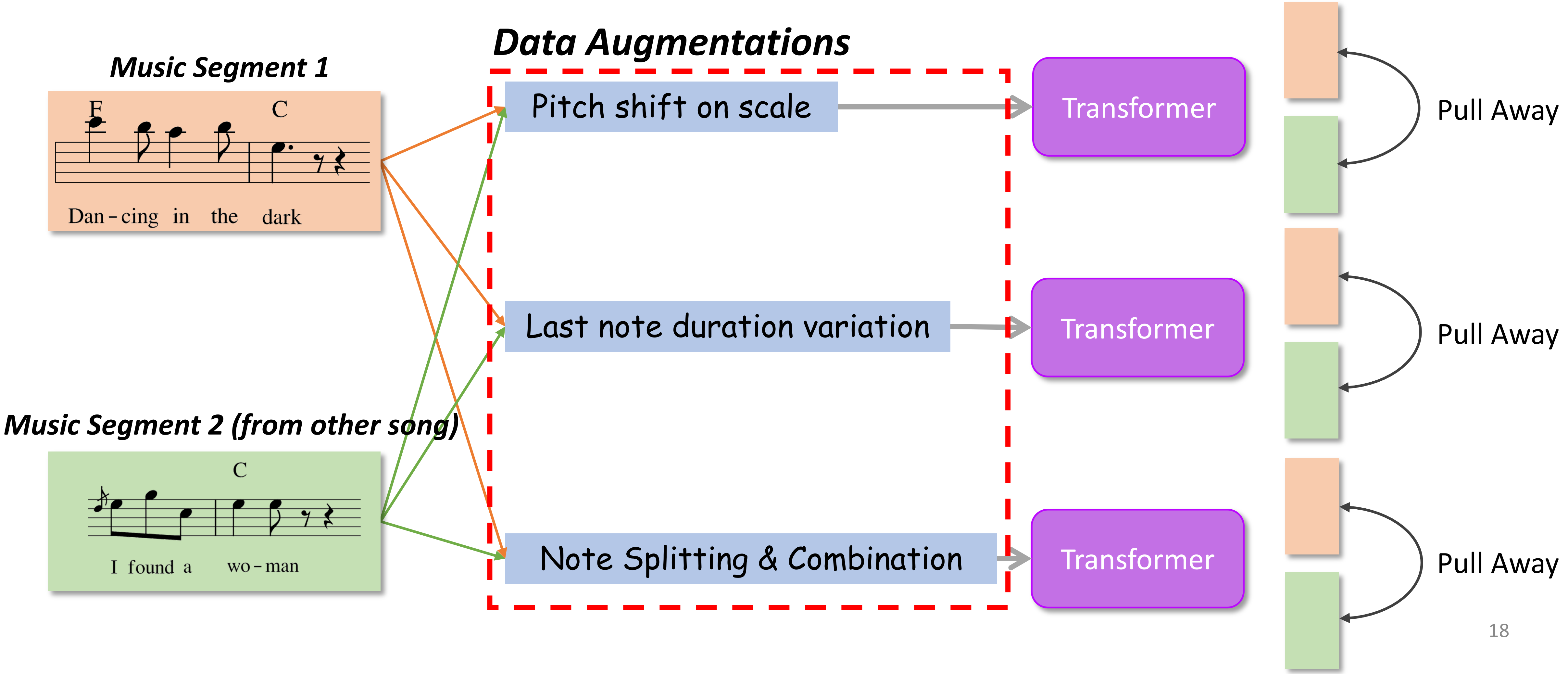
***Density based clustering***



Themes should be in the largest cluster

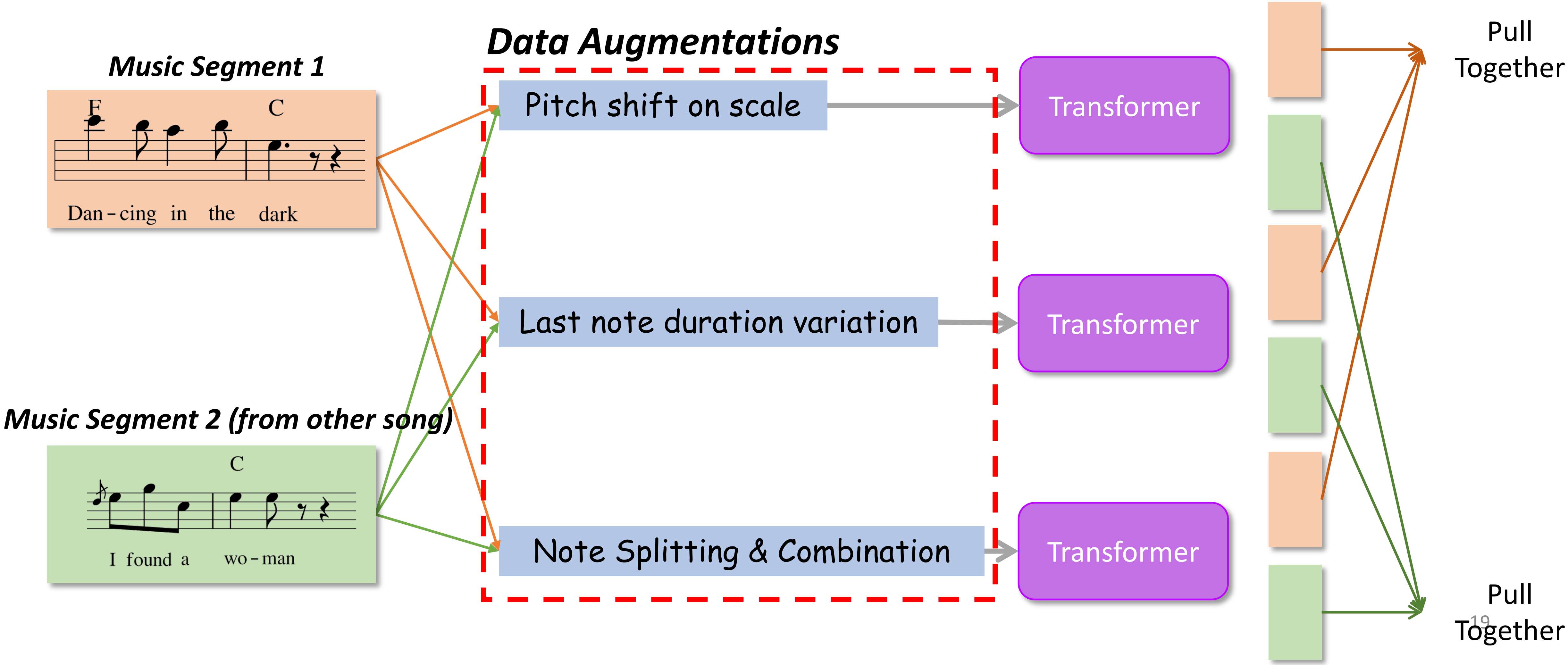
# Theme Retrieval

- Adopt idea from SimCLR (Chen et al., 2020)



# Theme Retrieval

- Adopt idea from SimCLR (Chen et al., 2020)



# Theme Retrieval

- Contrastive loss
  - $$-\log \frac{\exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_j)/\alpha)}{\sum_k \ \mathbf{1}_{[k \neq i]} \exp(\mathrm{sim}(\mathbf{z}_i, \mathbf{z}_k)/\alpha)}$$
- Apply DBSCAN to cluster music segments
  - $D(S_i, S_j) = \|\mathrm{Emb}(S_i) - \mathrm{Emb}(S_j)\|_2$
- Regard the largest segment as “Theme”
- Results: (F1 retrieval with human annotators)

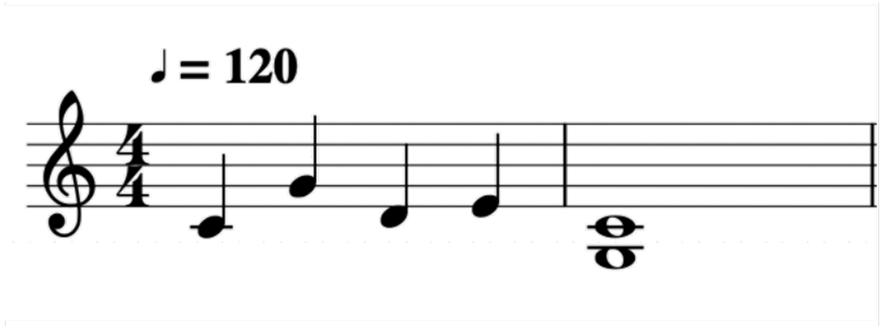
| | CL (proposed) | CL w/o Note Duration Augmentation | CL w/o Pitch Shift Augmentation | CM | COSIATEC |
|---|---|---|---|---|---|
| **Average F1** | **.378** | .220 | .336 | .345 | .297 |

# Outline

- Overview
- Technical Background
    - Theme Retrieval
    - Theme based Music Generation
- Results
- Conclusion & Contribution

# Theme-based Music Generation

- Background – Representation **REMI** (Hung et al., 2018)



We added additional ***Theme-Start***, ***Theme-End*** tokens to represent Theme Regions

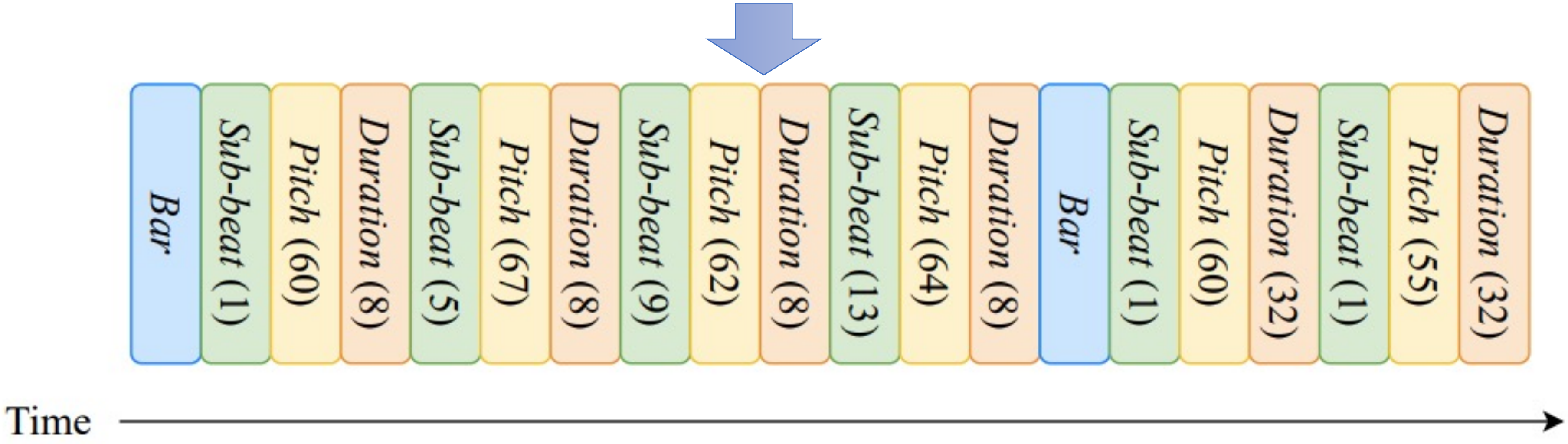


Figure from Chou et al., 2021

# Theme-based Music Generation

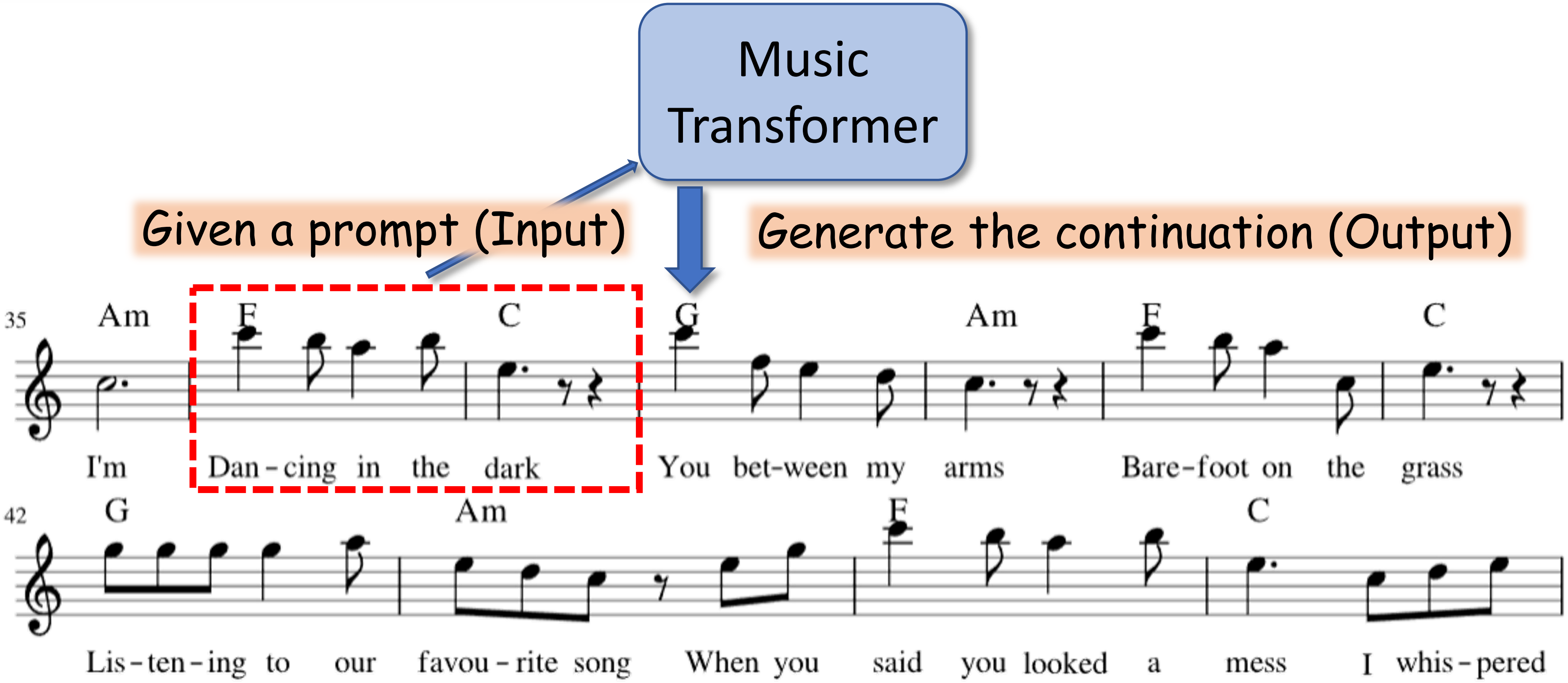
- Background – Autoregressive Model

$$p(x_t|x_{<t})$$

- Recently works employ Transformer as main backbone
    - Music Transformer (Huang et al., 2018)
    - Pop Music Transformer (Hung and Yang, 2020)
    - Compound Words Transformer (Hsiao et al., 2021)
- Train by minimizing Negative log-likelihood
    - $-\sum_{t=1}^{T} \log p(x_t \mid x_{<t})$

# Prompt-based Music Generation

## Prompt Conditioned Music Generation



Music Transformer

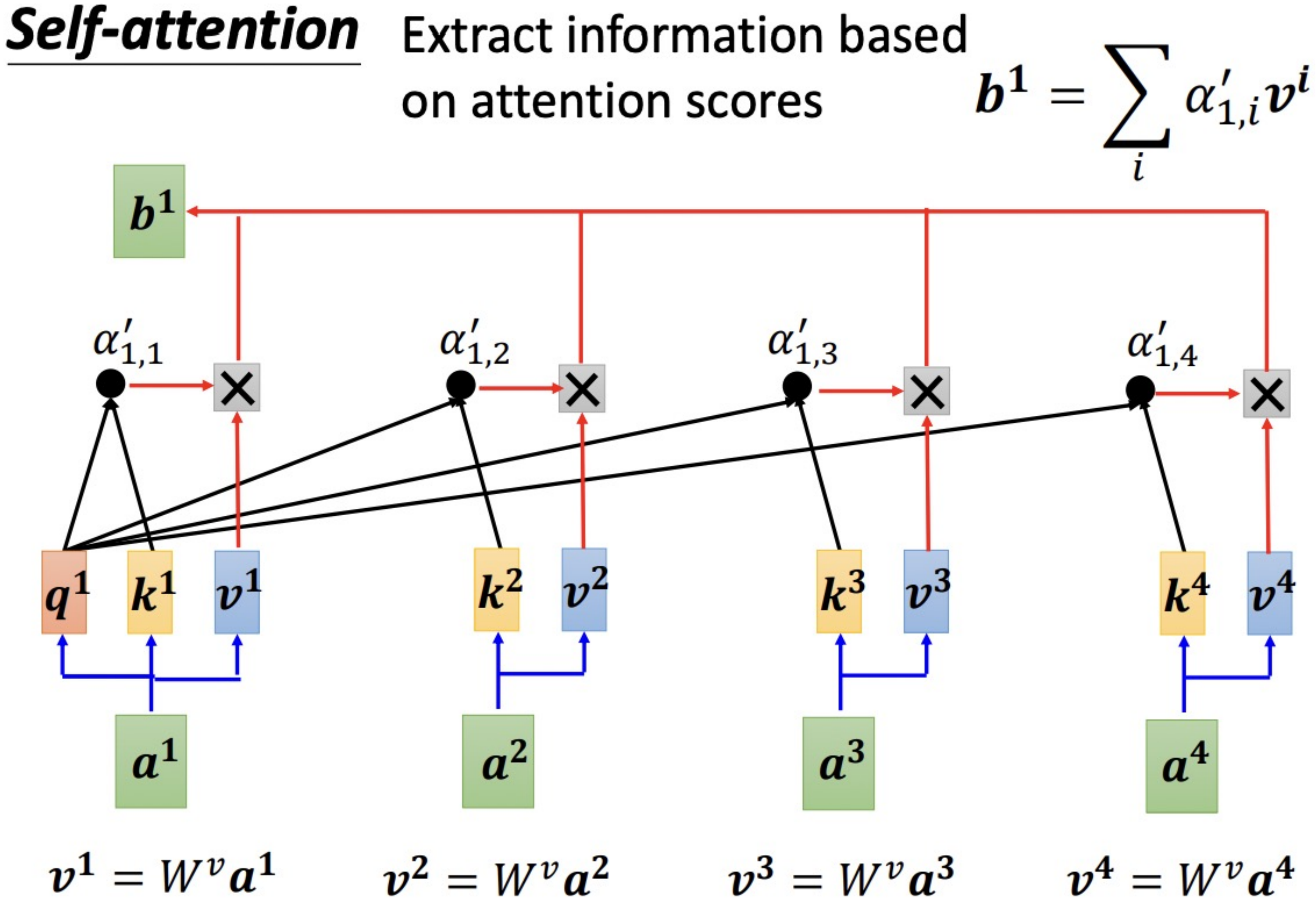
Given a prompt (Input)

Generate the continuation (Output)

35
Am
F
C
G
Am
F
C
I'm Dan-cing in the dark You bet-ween my arms Bare-foot on the grass
42
G
Am
F
C
Lis-ten-ing to our favou-rite song When you said you looked a mess I whis-pered

# Theme-based Music Generation

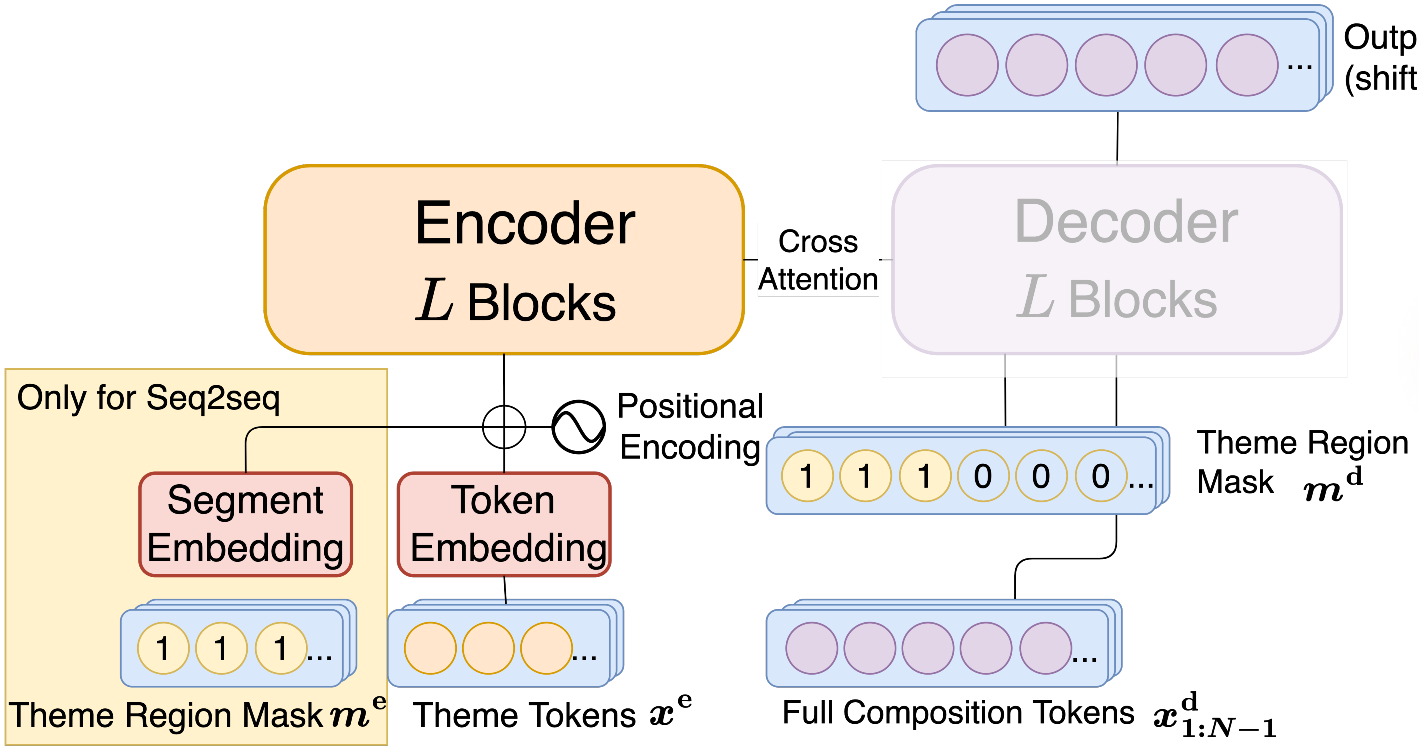
- Problem for prompt-based method



The NLL loss can be minimized without considering the "themes"

Figure from Prof. Hung-yi Lee

# Theme-based Music Generation



Output (shifted left)

Encoder
$L$ Blocks

Cross Attention

Decoder
$L$ Blocks

Only for Seq2seq

Positional Encoding

Segment Embedding

Token Embedding

1 1 1 ...

Theme Region Mask $\boldsymbol{m}^{\mathbf{e}}$

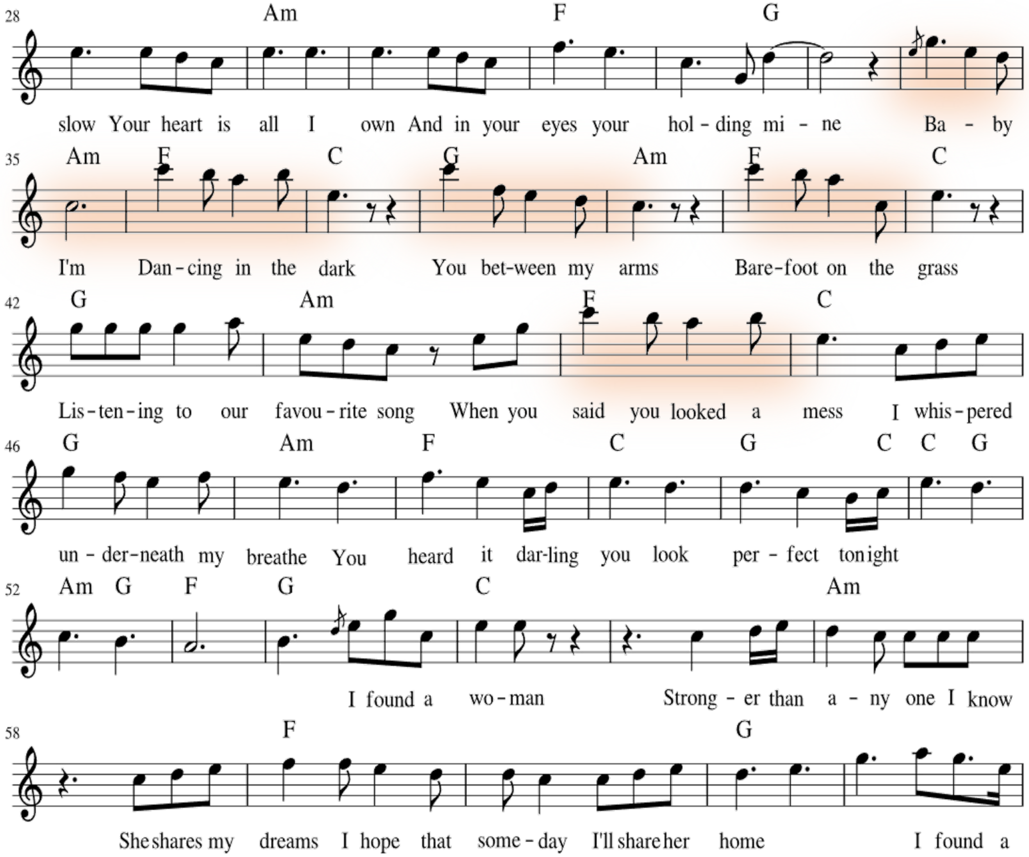
Theme Tokens $\boldsymbol{x}^{\mathbf{e}}$

1 1 1 0 0 0 ...

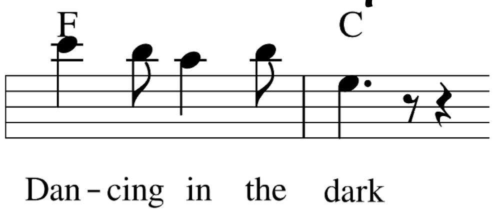
Theme Region Mask $\boldsymbol{m}^{\mathbf{d}}$

Full Composition Tokens $\boldsymbol{x}^{\mathbf{d}}_{\mathbf{1:N-1}}$

## *Entire Song (Output)*


28
Am F G
slow Your heart is all I own And in your eyes your hol - ding mi - ne Ba - by
35
Am F C G Am F C
I'm Dan-cing in the dark You bet-ween my arms Bare-foot on the grass
42
G Am F C
Lis-ten-ing to our favou-rite song When you said you looked a mess I whis-pered
46
G Am F C G C C G
un - der-neath my breathe You heard it dar-ling you look per - fect tonight
52
Am G F G C Am
I found a wo-man Strong - er than a - ny one I know
58
F G
She shares my dreams I hope that some-day I'll share her home I found a

## *Theme (Input)*



# Theme-based Music Generation

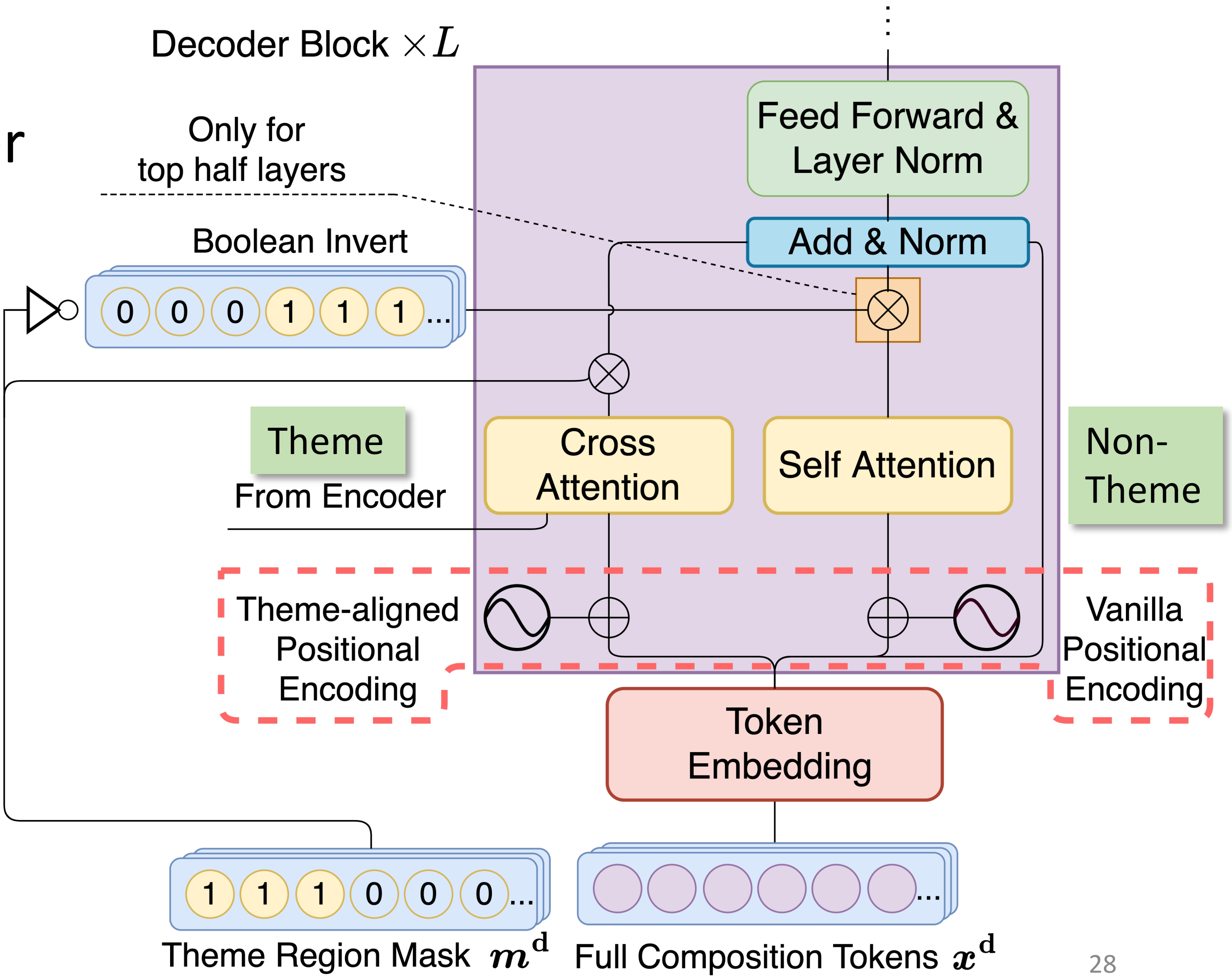
- Propose Theme Transformer
- Gating Mechanism

$$\mathbf{h}_t^l = \begin{cases} m_t\,\mathbf{h}_t^{l,(\text{cross})} + (1-m_t)\,\mathbf{h}_t^{l,(\text{self})}, & l > L/2 \\ m_t\,\mathbf{h}_t^{l,(\text{cross})} + \mathbf{h}_t^{l,(\text{self})}, & l \le L/2 \end{cases}$$

- Theme Positional Encoding

$$p_i^{\text{self}} = i \quad p_i^{\text{cross}} = i - \max_{(m_k^{\text{d}}=0)\wedge(0\le k<i)} k$$

- 2 Memory Networks
  - Theme
  - Non-Theme



# Outline

- Overview
- Technical Background
- Results
- Conclusion & Contribution

# Results

- Evaluation Metrics
  - Pitch Class Consistency
    - Overlapping Area of chroma histograms of two bars
  - Melody Inconsistency
    - The min distance of all the segments compared to the first one

      $D(S_1, S_*)$ $\quad D(S_i, S_j) = \|\mathrm{Emb}(S_i) - \mathrm{Emb}(S_j)\|_2$
  - Grooving consistency : coherence in rhythm

# Results

- Proposed Evaluation Metrics
  - Theme Inconsistency
    - the inconsistency between theme regions

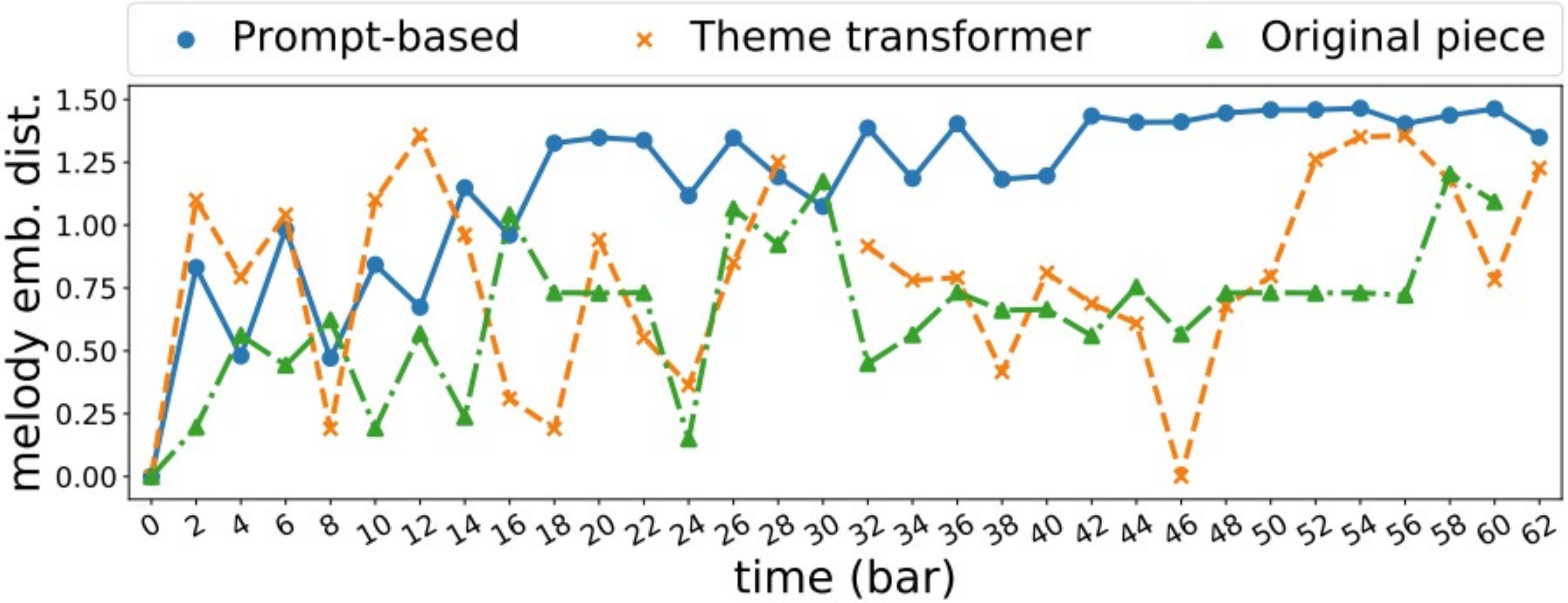
$$\frac{2}{N(N-1)} \sum_{i,j} D(\Gamma_i, \Gamma_j)$$

  - Theme Uncontrollability
    - the differences between theme regions and the given condition

$$\frac{1}{N} \sum_{i=1}^{N} D(c_{1:\tau}, \Gamma_i)$$

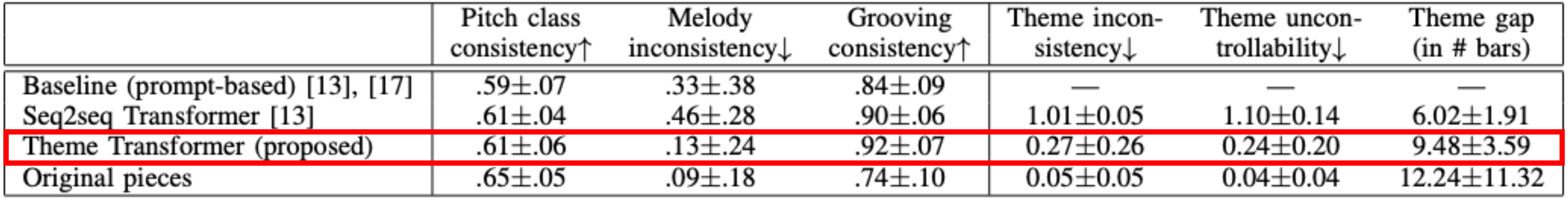
  - Theme Gap
    - Gaps between Theme Regions

# Results


● Prompt-based
× Theme transformer
▲ Original piece
melody emb. dist.
1.50
1.25
1.00
0.75
0.50
0.25
0.00
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34
36
38
40
42
44
46
48
50
52
54
56
58
60
62
time (bar)

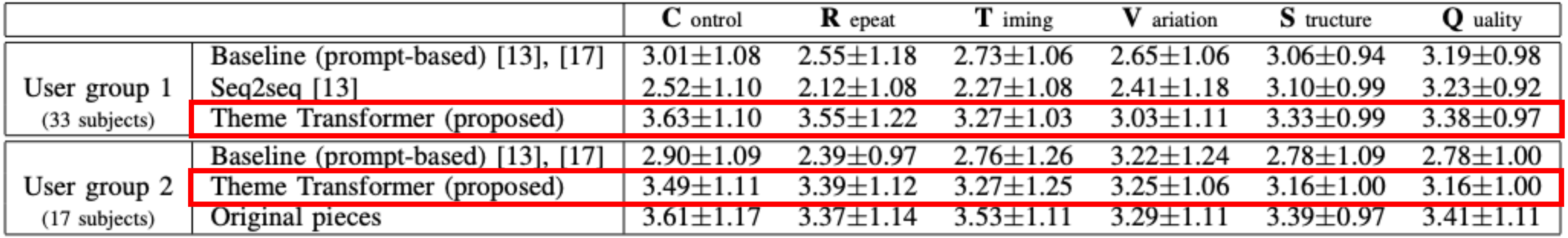
# Results

- Objective Evaluation



| | Pitch class consistency↑ | Melody inconsistency↓ | Grooving consistency↑ | Theme incon-sistency↓ | Theme uncon-trollability↓ | Theme gap (in # bars) |
|---|---|---|---|---|---|---|
| Baseline (prompt-based) [13], [17] | .59±.07 | .33±.38 | .84±.09 | — | — | — |
| Seq2seq Transformer [13] | .61±.04 | .46±.28 | .90±.06 | 1.01±0.05 | 1.10±0.14 | 6.02±1.91 |
| Theme Transformer (proposed) | .61±.06 | .13±.24 | .92±.07 | 0.27±0.26 | 0.24±0.20 | 9.48±3.59 |
| Original pieces | .65±.05 | .09±.18 | .74±.10 | 0.05±0.05 | 0.04±0.04 | 12.24±11.32 |

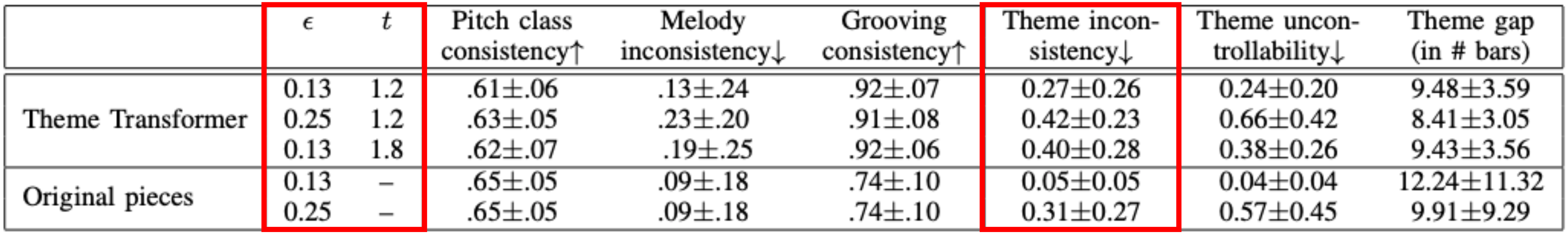
- Subjective Evaluation (Total ***50*** participants)



| | | **C** ontrol | **R** epeat | **T** iming | **V** ariation | **S** tructure | **Q** uality |
|---|---|---|---|---|---|---|---|
| User group 1<br>(33 subjects) | Baseline (prompt-based) [13], [17] | 3.01±1.08 | 2.55±1.18 | 2.73±1.06 | 2.65±1.06 | 3.06±0.94 | 3.19±0.98 |
| | Seq2seq [13] | 2.52±1.10 | 2.12±1.08 | 2.27±1.08 | 2.41±1.18 | 3.10±0.99 | 3.23±0.92 |
| | Theme Transformer (proposed) | 3.63±1.10 | 3.55±1.22 | 3.27±1.03 | 3.03±1.11 | 3.33±0.99 | 3.38±0.97 |
| User group 2<br>(17 subjects) | Baseline (prompt-based) [13], [17] | 2.90±1.09 | 2.39±0.97 | 2.76±1.26 | 3.22±1.24 | 2.78±1.09 | 2.78±1.00 |
| | Theme Transformer (proposed) | 3.49±1.11 | 3.39±1.12 | 3.27±1.25 | 3.25±1.06 | 3.16±1.00 | 3.16±1.00 |
| | Original pieces | 3.61±1.17 | 3.37±1.14 | 3.53±1.11 | 3.29±1.11 | 3.39±0.97 | 3.41±1.11 |

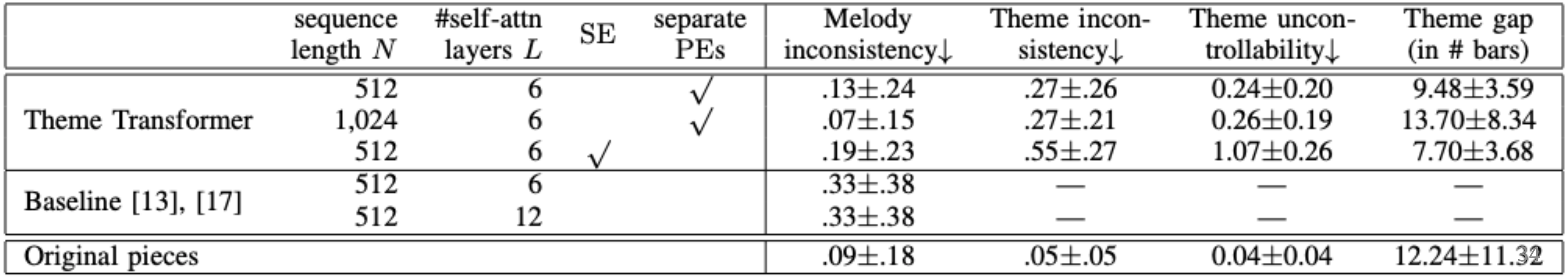
# Results

- Ablation Studies on Temperature and Sampling



| | $\epsilon$ | $t$ | Pitch class consistency↑ | Melody inconsistency↓ | Grooving consistency↑ | Theme inconsistency↓ | Theme uncontrollability↓ | Theme gap (in # bars) |
|---|---|---|---|---|---|---|---|---|
| Theme Transformer | 0.13 | 1.2 | .61±.06 | .13±.24 | .92±.07 | 0.27±0.26 | 0.24±0.20 | 9.48±3.59 |
| | 0.25 | 1.2 | .63±.05 | .23±.20 | .91±.08 | 0.42±0.23 | 0.66±0.42 | 8.41±3.05 |
| | 0.13 | 1.8 | .62±.07 | .19±.25 | .92±.06 | 0.40±0.28 | 0.38±0.26 | 9.43±3.56 |
| Original pieces | 0.13 | – | .65±.05 | .09±.18 | .74±.10 | 0.05±0.05 | 0.04±0.04 | 12.24±11.32 |
| | 0.25 | – | .65±.05 | .09±.18 | .74±.10 | 0.31±0.27 | 0.57±0.45 | 9.91±9.29 |

- Ablation Studies on Model Architecture



| | sequence length $N$ | #self-attn layers $L$ | SE | separate PEs | Melody inconsistency↓ | Theme inconsistency↓ | Theme uncontrollability↓ | Theme gap (in # bars) |
|---|---|---|---|---|---|---|---|---|
| Theme Transformer | 512 | 6 | | √ | .13±.24 | .27±.26 | 0.24±0.20 | 9.48±3.59 |
| | 1,024 | 6 | | √ | .07±.15 | .27±.21 | 0.26±0.19 | 13.70±8.34 |
| | 512 | 6 | √ | | .19±.23 | .55±.27 | 1.07±0.26 | 7.70±3.68 |
| Baseline [13], [17] | 512 | 6 | | | .33±.38 | — | — | — |
| | 512 | 12 | | | .33±.38 | — | — | — |
| Original pieces | | | | | .09±.18 | .05±.05 | 0.04±0.04 | 12.24±11.32 |

# Outline

- Overview
- Technical Background
- Results
- Contribution

# Contributions

- Proposed an Unsupervised Method for Theme Retrieval
- The first work to introduce Theme-based Symbolic Music Generation
- Design Theme-based Evaluation Metrics
- Our method outperform previous music generation works

# Thanks for listening

Ian Shih

Email: yjshih23@gmail.com

Website: atosystem.github.io