MelodyGLM: Pre-Training with Musical N-Gram for Melody Generation and Editing

Bachelor Thesis

HUANG Zhijie

Chu Kochen Honor College, Zhejiang University
Department of Computer Science and Technology

2023.06

Melody Pretraining for Generation & Editing

Background: Deep Learning & Intelligent Music

Paradigm Shift of DL

Pretraining +

Language → ChatGPT

Image/Video → stability.ai Stable Diffusion

Music → ?

Scale & Generality

Reference:
[1] ChatGPT, OpenAI. [2] Stable Diffusion, Stability AI.
[3] Briot, J.-P., & Pachet, F. (2020). Deep learning for music generation: Challenges and directions. Neural Computing and Applications.
[4] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). *Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.*

Melody Pretraining for Generation & Editing

Background: Deep Learning & Intelligent Music

Pretraining + Music → ?

Weakness

- Few datasets. Small/poor quality. *(scale)*
- Good at only one particular task. *(generality)*

Goal

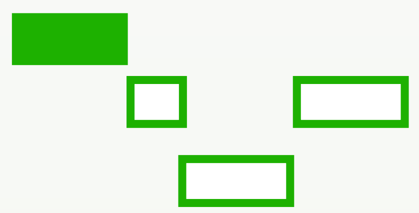
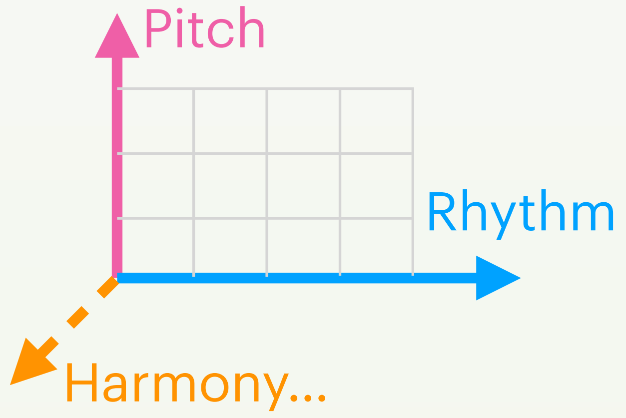
- Knowledge transfer from large-scale datasets. *(scale)*
- Unify music generative tasks. *(generality)*

Reference:
[1] ChatGPT, OpenAI. [2] Stable Diffusion, Stability AI.
[3] Briot, J.-P., & Pachet, F. (2020). Deep learning for music generation: Challenges and directions. Neural Computing and Applications.
[4] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing.

Pretraining based on **Musical N-Gram**

Challenge: Tailor Pretraining to Music

Music vs. Language

- A single note expresses nothing. ♪ *vs* 音
- Vague entity boundary.  *vs* *Words have* ***boundaries.***
- Multiple Dimensions. 

vs *Text is linear.*

Challenge: Difficult to adopt NLP techniques directly.
Boundary and dimension of music are under exploration.

Solution →

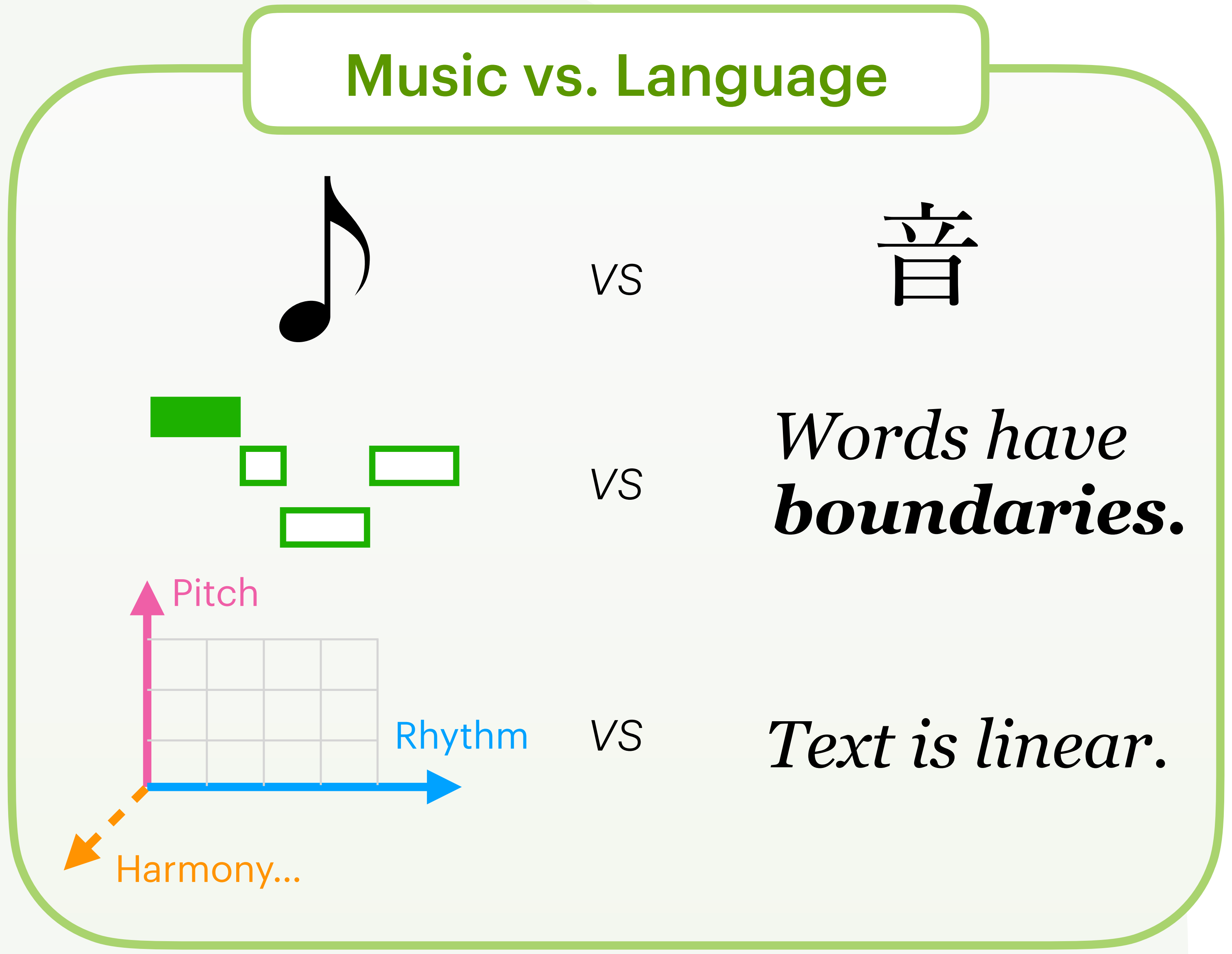
Musical N-Gram

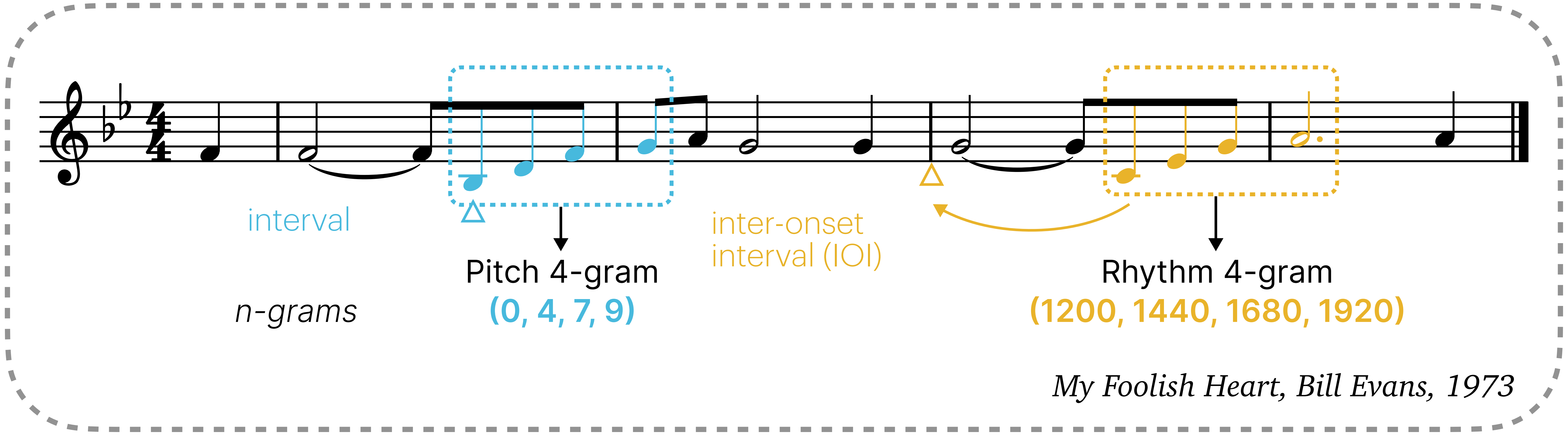
- Statistically capture musical semantic boundaries.
- Explore musical vocabulary among different dimensions.
- Design masking strategy tailored to characteristics of music.

Reference:
[1] Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., & Liu, T.-Y. (2021). MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training.

Pretraining based on **Musical N-Gram**

Method: Musical N-gram Masking Strategy





Common Patterns
vocabulary, phrases and idioms

- ***Context:*** relationship among notes
- ***Boundary:*** high-level semantics
- ***Dimension:*** both pitch and rhythm

Reference:
[1] Xiao, D., Li, Y.-K., Zhang, H., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding.
[2] Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., & Shoham, Y. (2020). PMI-Masking: Principled masking of correlated spans.

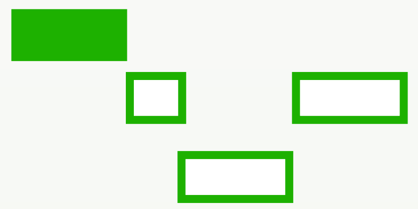
Pretraining based on **Musical N-Gram**

Method: Musical N-gram Masking Strategy

Music vs. Language

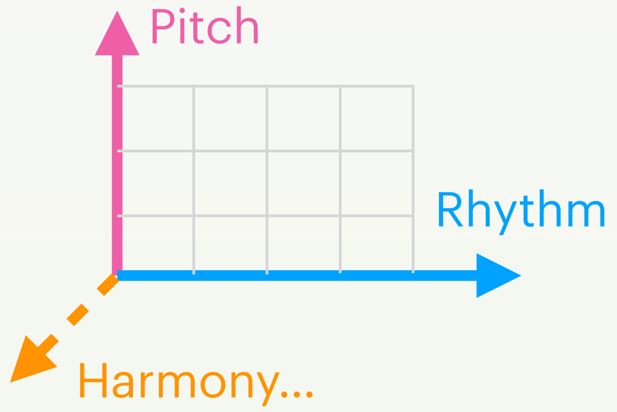
VS

音



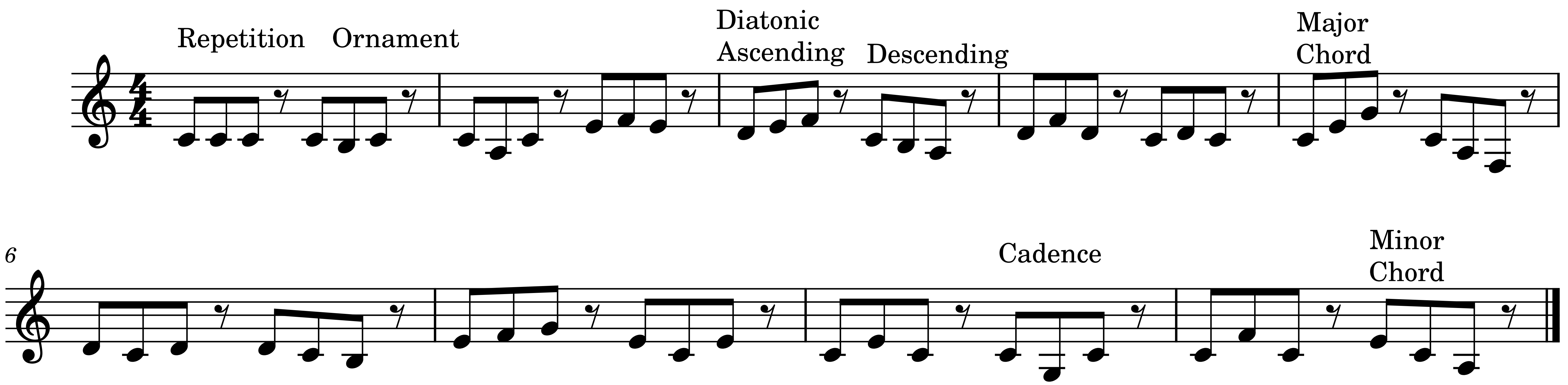
VS

Words have ***boundaries.***



VS

Text is linear.



Example: Most significant pitch 3-grams extracted from the Wikifonia dataset.

Common Patterns

vocabulary, phrases and idioms

- ***Context:*** relationship among notes
- ***Boundary:*** high-level semantics
- ***Dimension:*** both pitch and rhythm

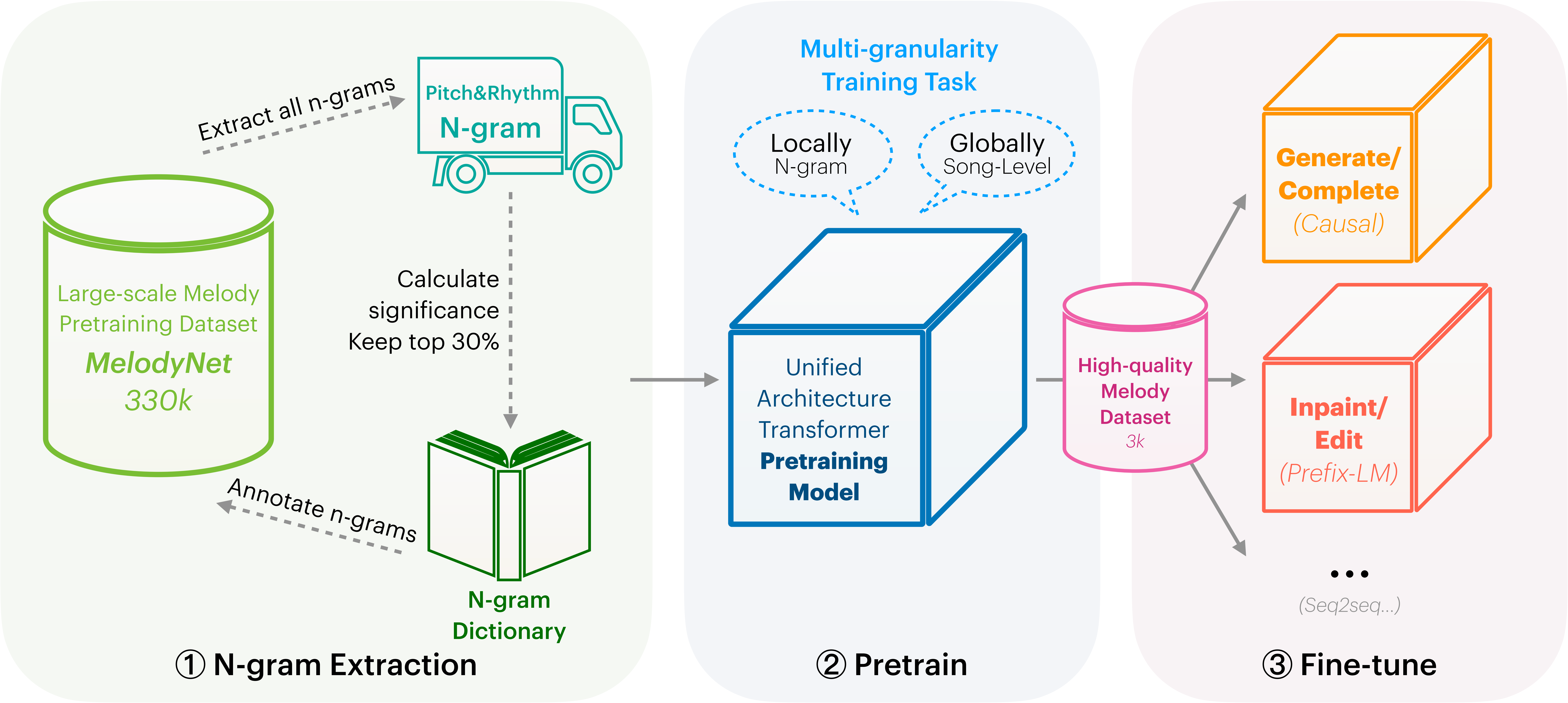
Reference:

[1] Xiao, D., Li, Y.-K., Zhang, H., Sun, Y., Tian, H., Wu, H., & Wang, H. (2021). ERNIE-Gram: Pre-Training with Explicitly N-Gram Masked Language Modeling for Natural Language Understanding.

[2] Levine, Y., Lenz, B., Lieber, O., Abend, O., Leyton-Brown, K., Tennenholtz, M., & Shoham, Y. (2020). PMI-Masking: Principled masking of correlated spans.

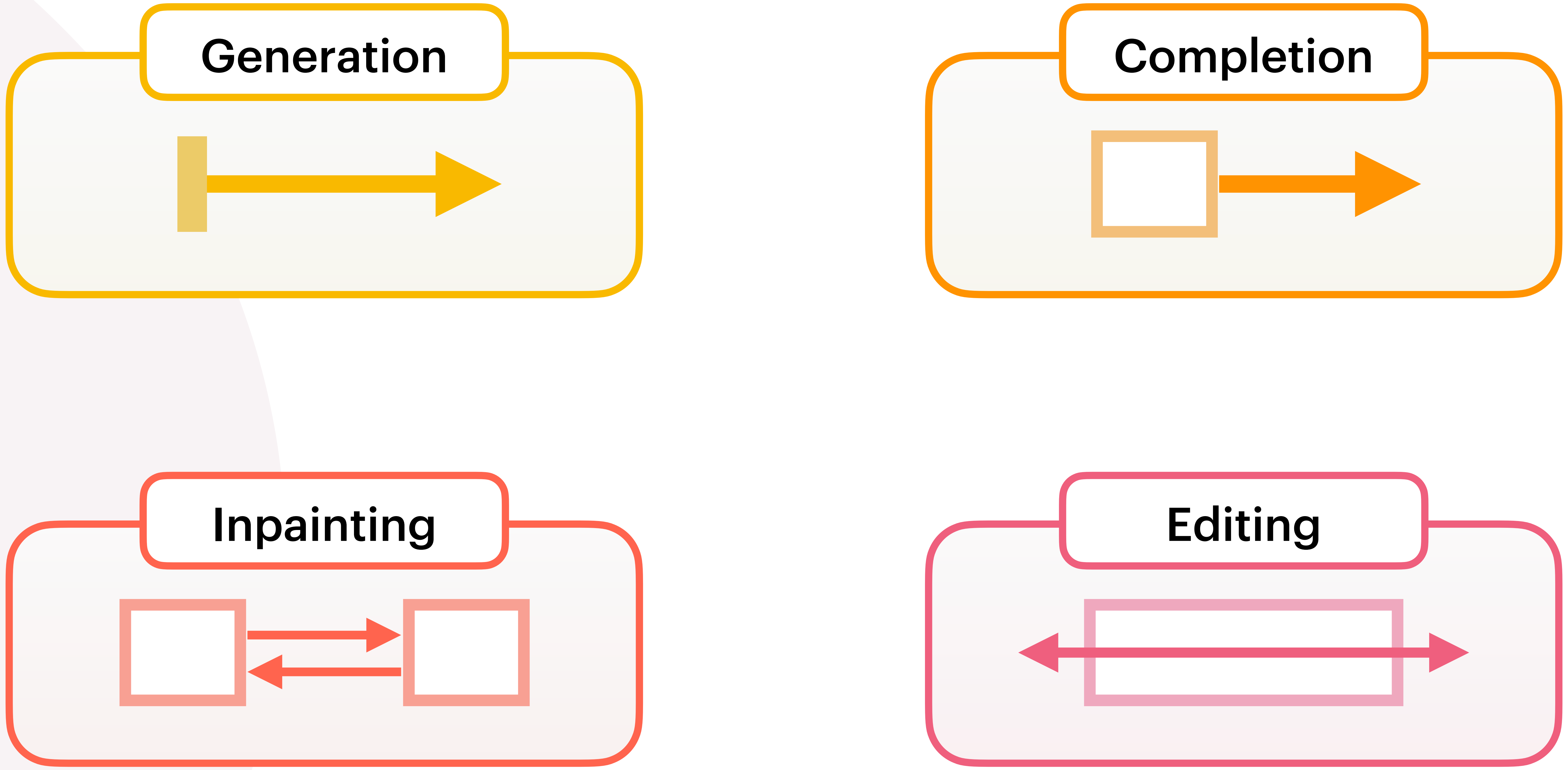
Pretraining based on Musical N-Gram

Method Overview



Pretraining based on Musical N-Gram

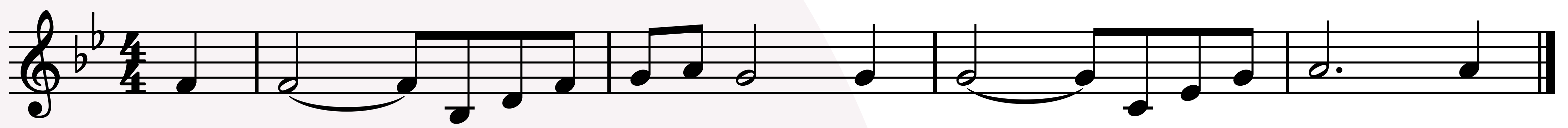
Downstream Tasks

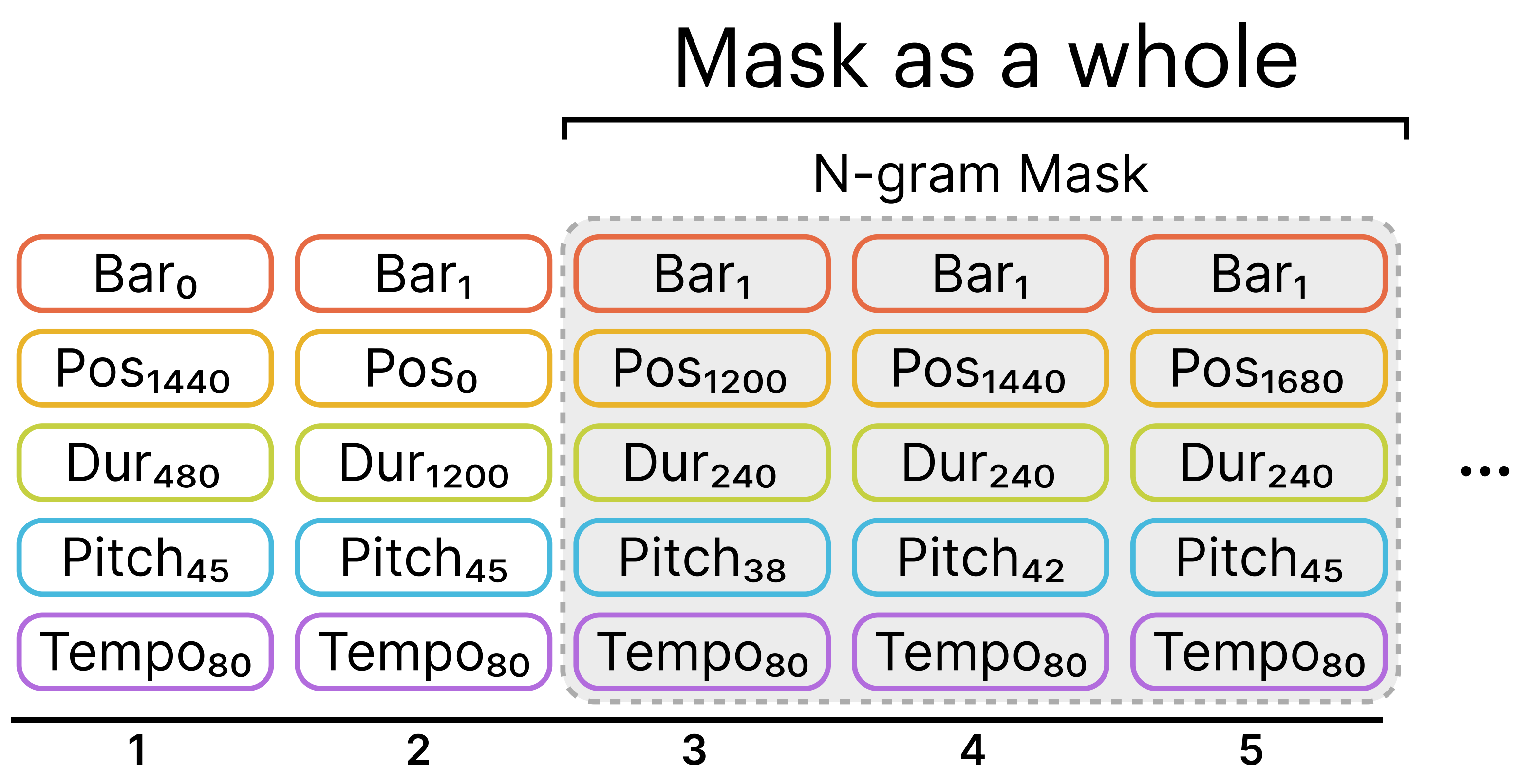


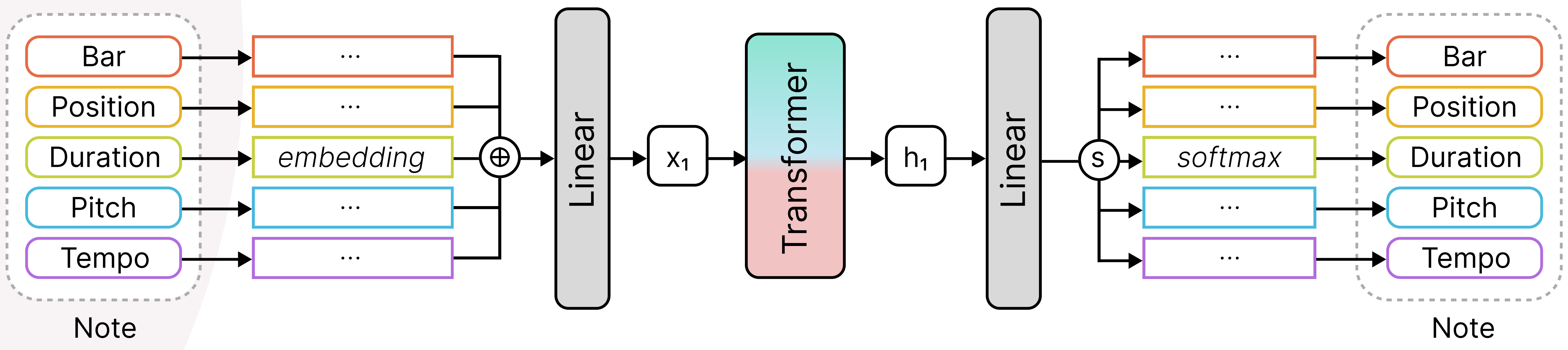
Pretraining based on Musical N-Gram

Method: Melody Encoding

Compound word music encoding.
A unit of one single note.





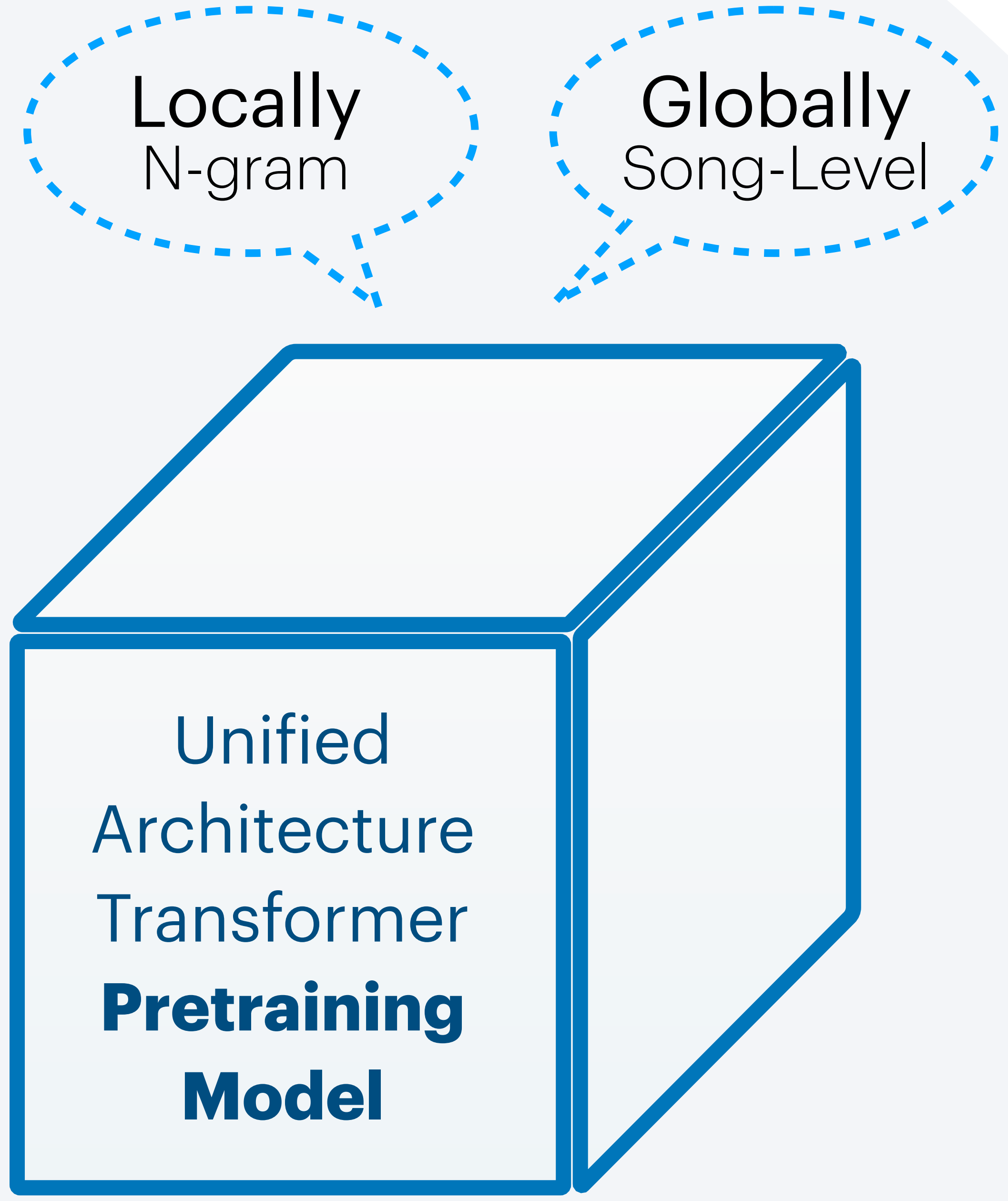


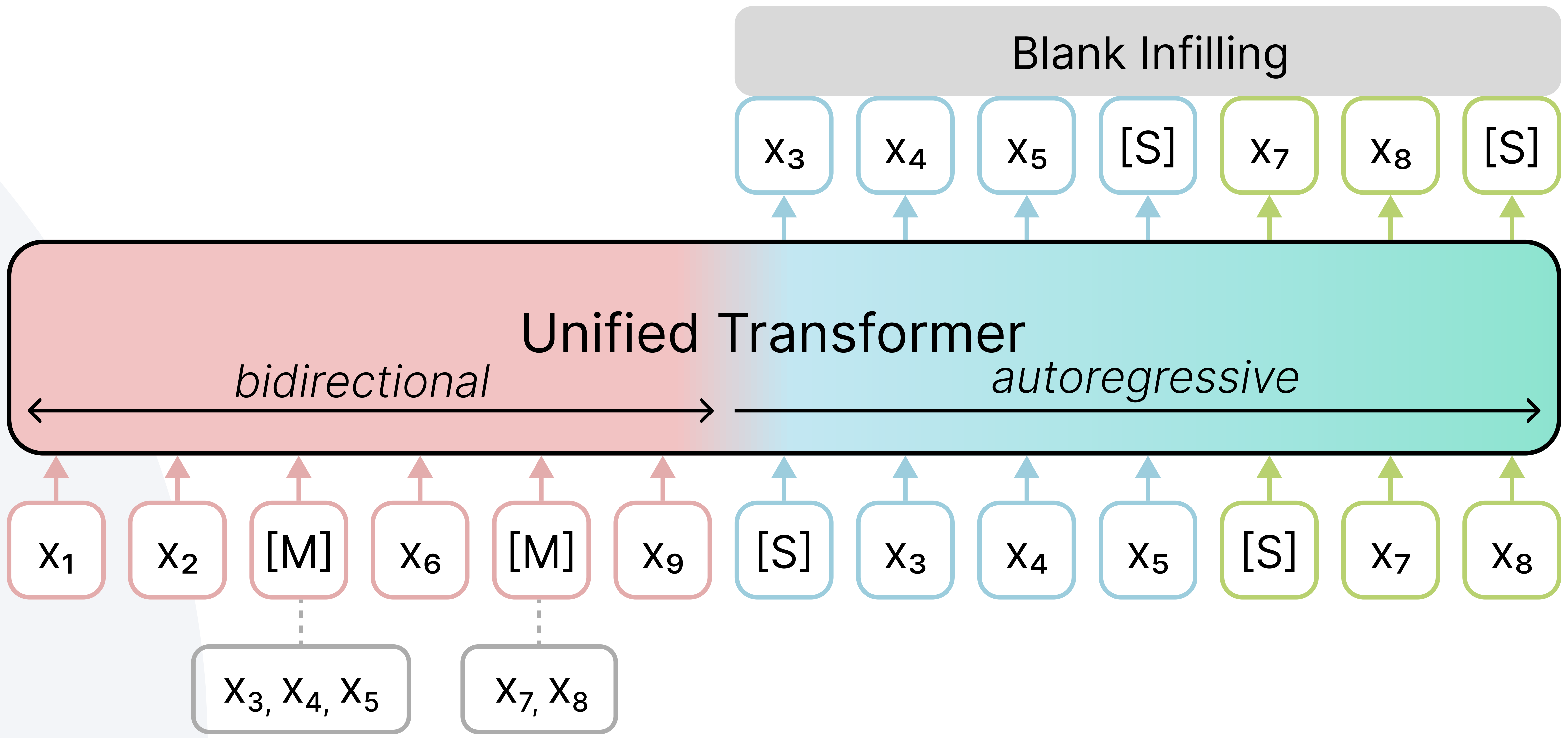
Reference:
[1] Zeng, M., Tan, X., Wang, R., Ju, Z., Qin, T., & Liu, T.-Y. (2021). MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training.

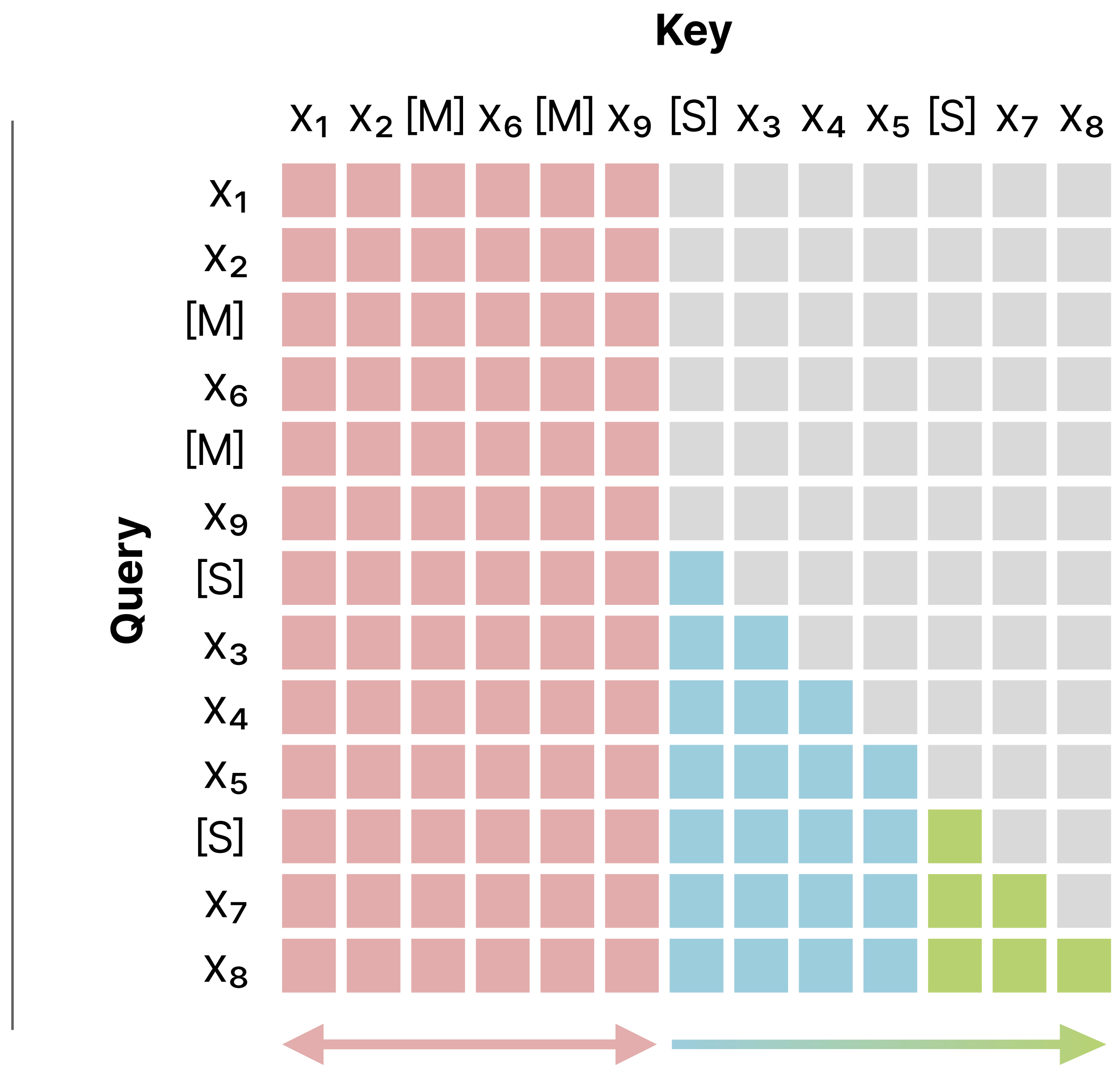
Pretraining based on Musical N-Gram

Method: Model Architecture & Pretraining Task

$$\mathcal{L} = \mathcal{L}_{pitch} + \mathcal{L}_{rhythm} + \mathcal{L}_{song}$$







Multi-granularity & Multi-dimension Task

Unified Architecture & Span-infilling Objective

Various downstream tasks with different formats

generation, editing, inpainting, refinement...

Reference:
[1] Du, Z., Qian, Y., Liu, X., Ding, M., Qiu, J., Yang, Z., & Tang, J. (2022). GLM: General Language Model Pretraining with Autoregressive Blank Infilling.
[2] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer.

Experiments

Experimental Settings & Evaluation Metrics

Ablation Studies

Effectiveness of the main components

1. **Masking strategy**
   - Random Span *(SpanBERT)*
   - Random Bar *(MusicBERT)*
   - Single Span *(MASS)*
   - **N-gram *(Ours)***
2. **Pretraining framework**
   - GPT-style
   - No pretraining
3. **Multi-granularity objective**
   - N-gram
   - **N-gram + Single Span**

Hyperparameters of n-gram extraction

1. Masking ratio
   - 20%, 40%, **60%**, 80%
2. N-gram extraction strategy
   - jointly for pitch & rhythm
   - **separately for pitch & rhythm**
3. N-gram length *(N = ?)*
   - 3–5, 3–8, **3–12**

Baseline Comparison

Compare with baselines on downstream tasks

1. **Melody generation**
   - Compound Word Transformer (2021)
   - Music Transformer (2018)
2. **Melody inpainting**
   - VLI (2021)

Metrics

1. **Objective evaluation**
   - Perplexity
   - Consistency *pitch histogram*
   - Rhythmicity *average IOI*
   - Structure *structure error*
   - Diversity *distinct pitch n-gram percentage*
2. **Subjective evaluation**
   - Consistency
   - Rhythmicity
   - Structure
   - Overall Quality

Experiments

Comparison: Melody Generation & Inpainting

Baseline Comparison

Compare with baselines on downstream tasks

1. **Melody generation**
   - Compound Word Transformer (2021)
   - Music Transformer (2018)
2. **Melody inpainting**
   - VLI (2021)

MelodyGLM

- Good consistency & rhythm
- Excellent structure & long-term coherence
- Nice musical creativity & diversity

Table 1: Results on 32-bar melody generation compared with SOTA baselines

Model	$OA(PCH)\uparrow$	$OA(IOI)\uparrow$	$SE\downarrow$	DN_{short}	DN_{medium}	DN_{long}
Ground Truth	-	-	-	1.4737	3.5765	7.6436
MT[3]	0.8399	0.9203	2.63%	0.8580	2.2132	5.5436
CWT[5]	0.9168	0.9573	4.84%	1.2454	3.2014	**7.2011**
MelodyGLM	**0.9783**	**0.9636**	**1.75%**	**1.3118**	**3.2642**	7.1856

Table 2: Results on 4-bar melody inpainting compared with SOTA baseline

Model	$OA(PCH)\uparrow$	$OA(IOI)\uparrow$	$SE\downarrow$	DN_{short}	DN_{medium}	DN_{long}
Ground Truth	-	-	-	1.9466	3.9951	7.2187
VLI[18]	0.9760	0.9615	0.58%	2.0867	4.1792	7.3696
MelodyGLM	**0.9907**	**0.9671**	**0.25%**	**1.926**	**3.9651**	**7.1957**

Note: *Greater overlapped area (OA) and lesser structure error (SE) is better.*
For other metrics, closer to ground truth is better.

Experiments

Ablation: Masking Strategy, Ratio & Pretraining Framework

Figure 1: PPL under different masking strategies and ratio on two downstream tasks

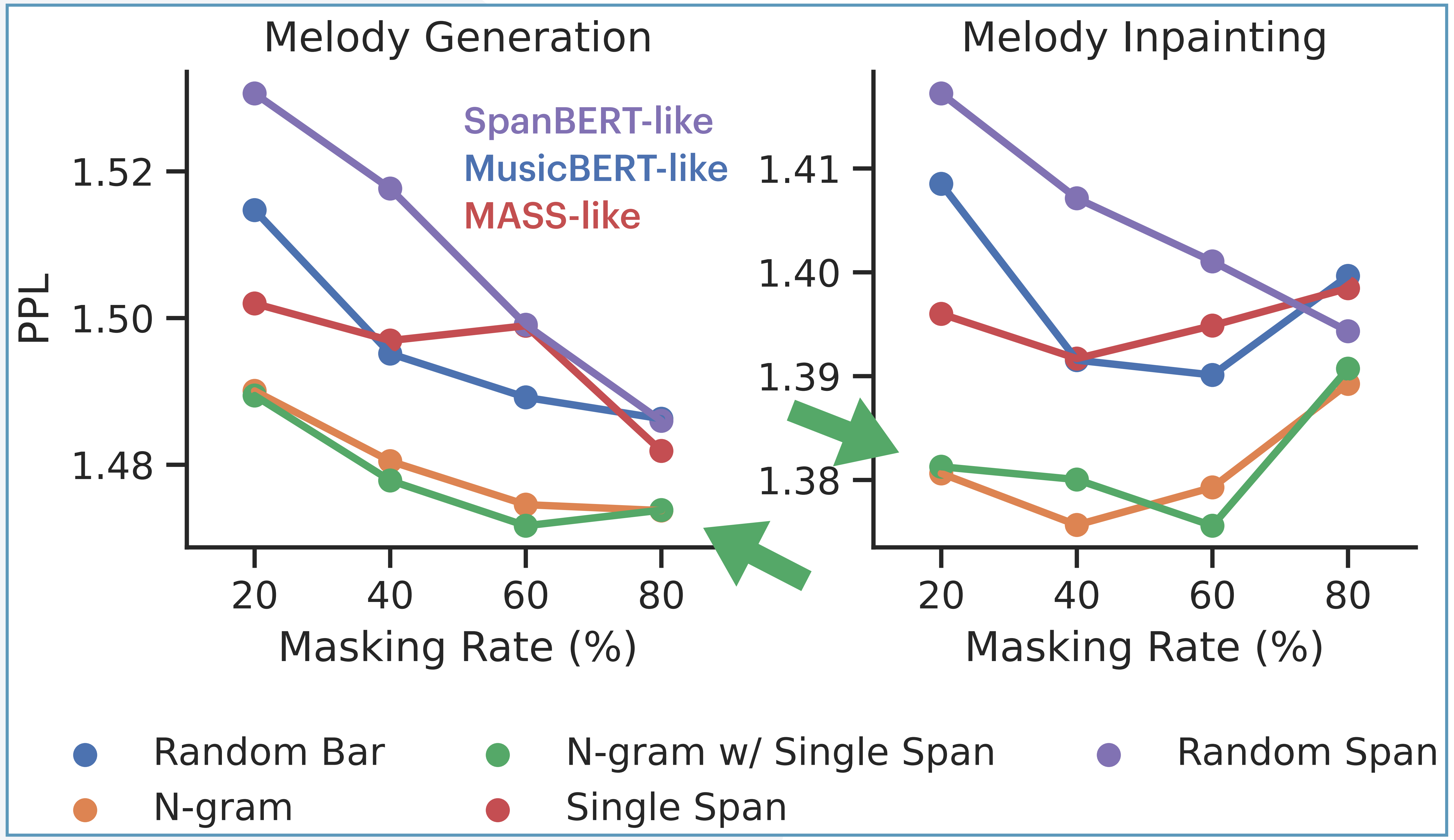
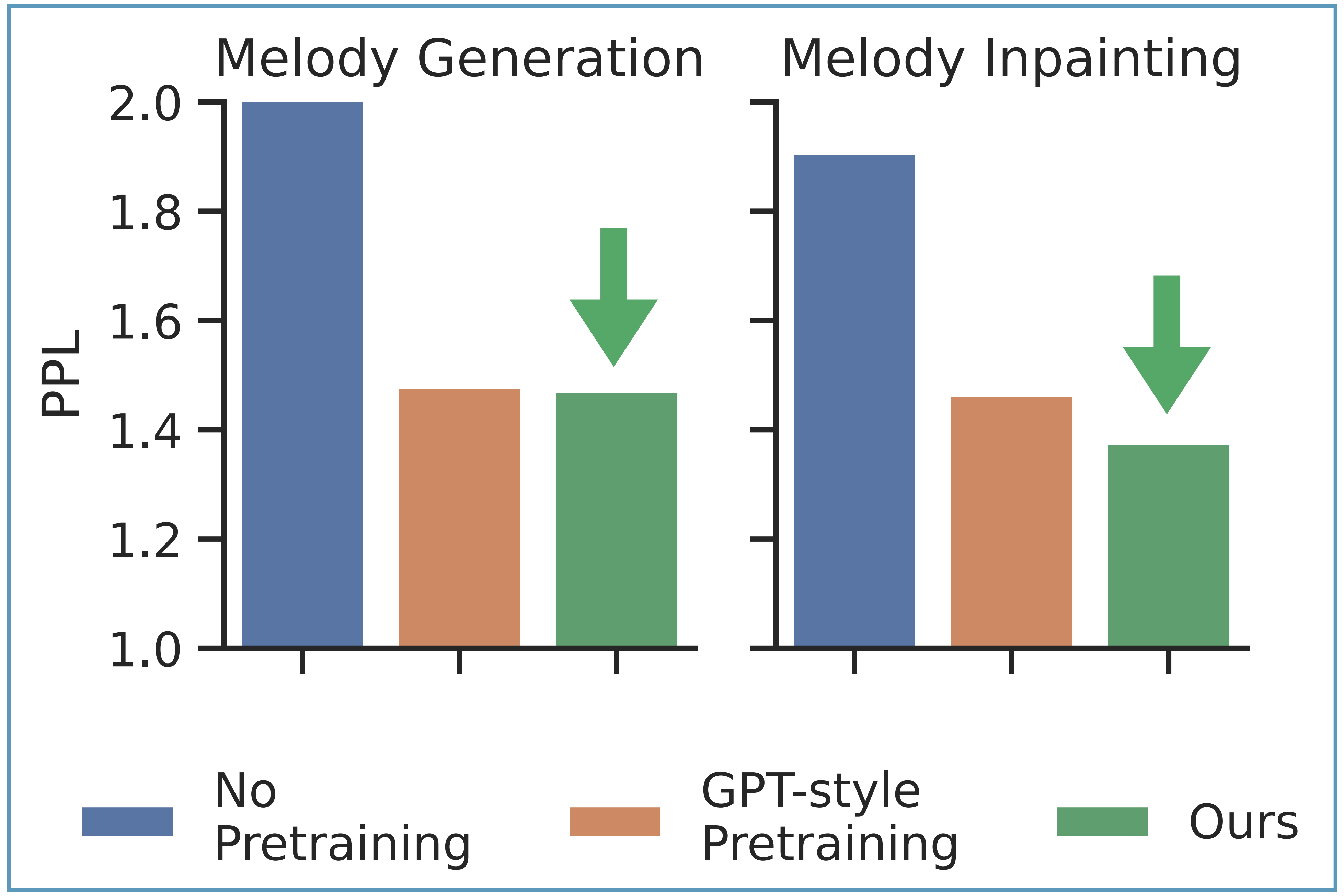


Figure 2: PPL under different pretraining framework on two downstream tasks

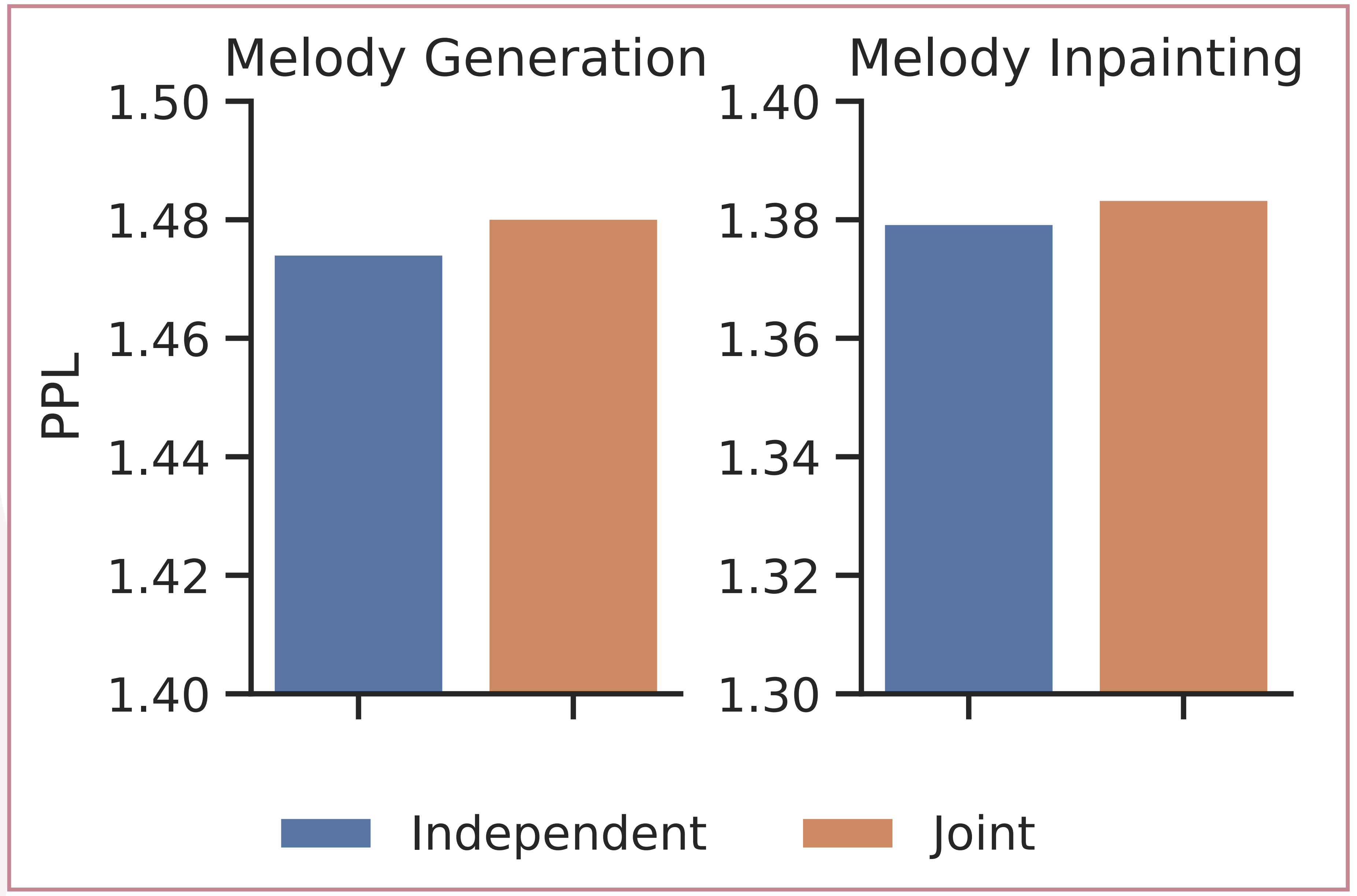


- Musical n-gram masking outperforms others.
- Multi-granularity training brings performance gain.
- GPT-style pretraining still struggles a bit on inpainting.
- Ours pretraining framework generalizes on two tasks.

Experiments

Ablation: Dimension & Scale of N-gram Extraction

Figure 3: PPL under different dimensions of n-gram extraction on two downstream tasks



- **Independent modeling for different musical dimensions** benefits two tasks.

MelodyGLM

Conclusion

MelodyGLM

Pre-Training with Musical N-Gram for Melody Generation and Editing

Contributions:

- Introduce **the paradigm of pretrain–fine-tune** to music generation.
- Design musical N-gram masking strategy and multi-granularity, multi-task training objective **tailored to music.**
- MelodyGLM **outperforms previous SOTA methods** on melody generation and inpainting, enabling various generative tasks.