MIDI-VALLE: Improving Expressive Piano Performance Synthesis Through Neural Codec Language Modelling

Jingjing Tang[1], Xin Wang[2], Zhe Zhang[2], Junichi Yamagishi[2], Geraint Wiggins[13], György Fazekas[1]

[1]Centre for Digital Music, Queen Mary University of London, UK,

[2]National Institute of Informatics, Japan

[3]Vrije Universiteit Brussel, Belgium

ISMIR 2025

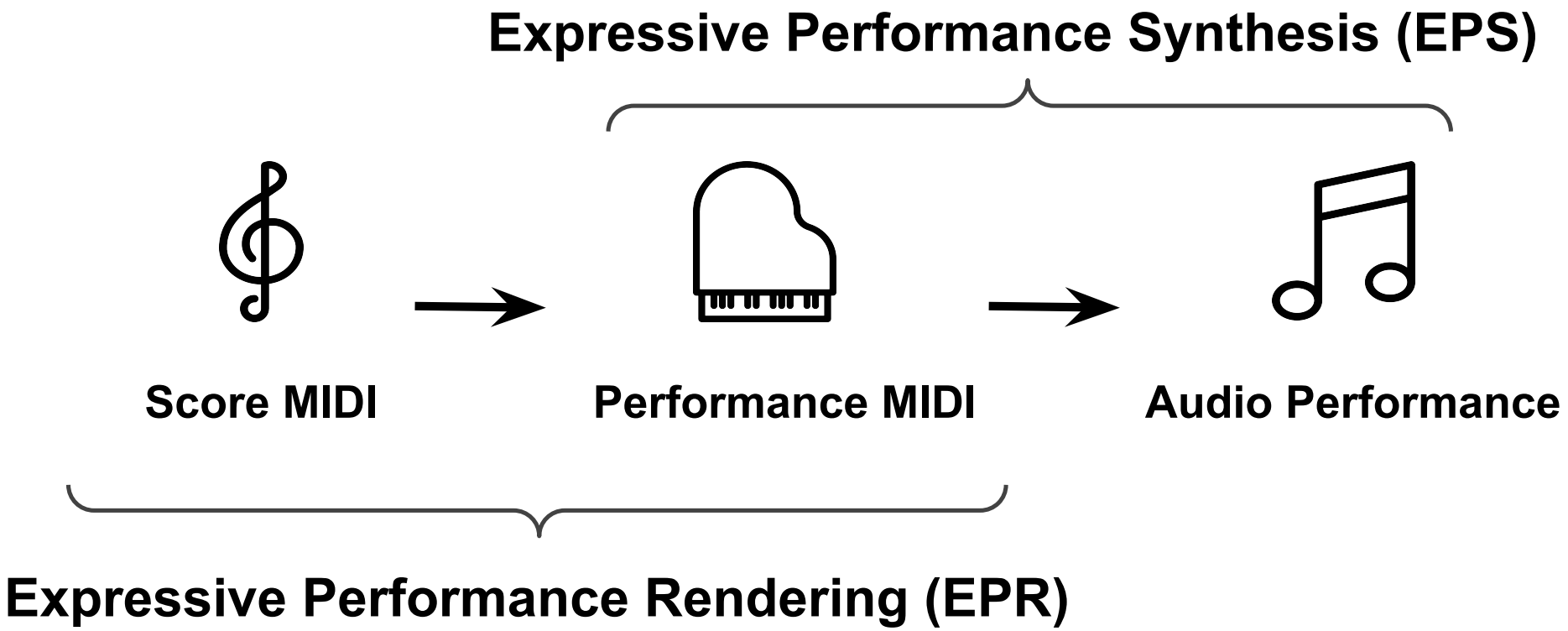
NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

AIM AI + MUSIC

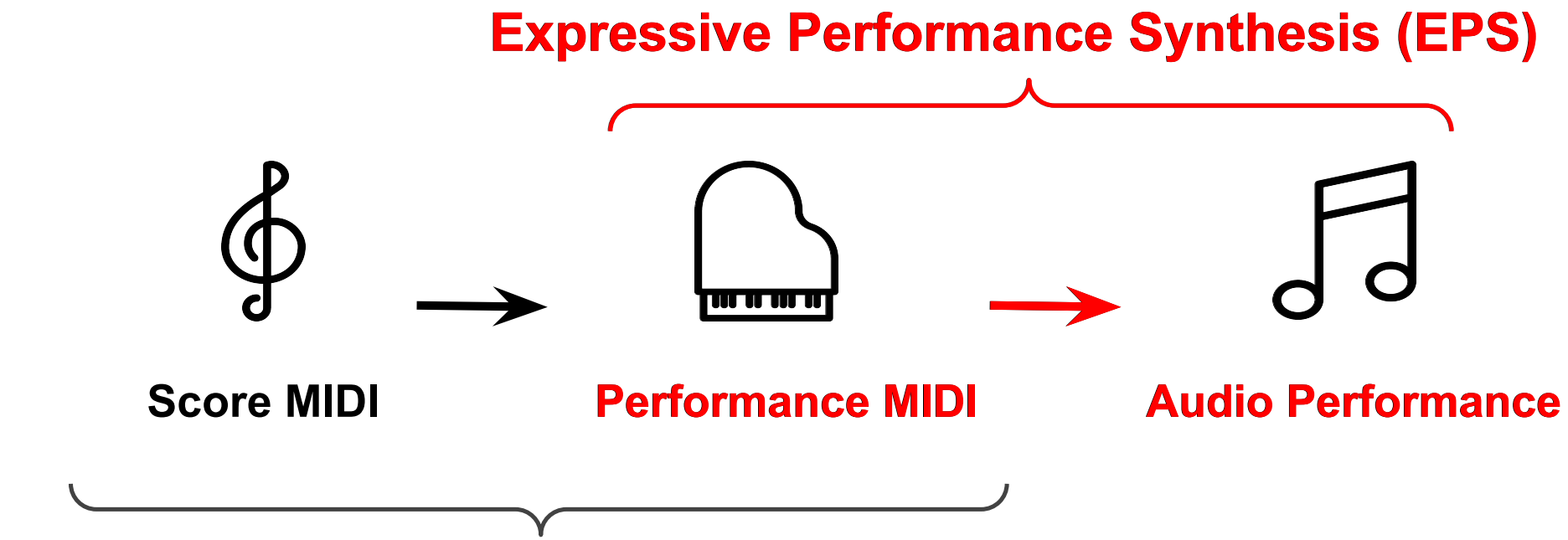
Queen Mary University of London

UKRI UK Research and Innovation

Background


Expressive Performance Synthesis (EPS)
Score MIDI
Performance MIDI
Audio Performance
Expressive Performance Rendering (EPR)

Background


Expressive Performance Synthesis (EPS)
Score MIDI
Performance MIDI
Audio Performance
Expressive Performance Rendering (EPR)

Background

Existing Expressive Performance Synthesis (EPS) Models for Piano

- Differentiable Digital Signal Processing (DDSP) models
- Adaptation from conventional **Text-to-Speech** (TTS) models (Spectrograms + Vocoder)

Background

Existing Expressive Performance Synthesis (EPS) Models for Piano

- Differentiable Digital Signal Processing (DDSP) models
- Adaptation from conventional **Text-to-Speech** (TTS) models (Spectrograms + Vocoder)

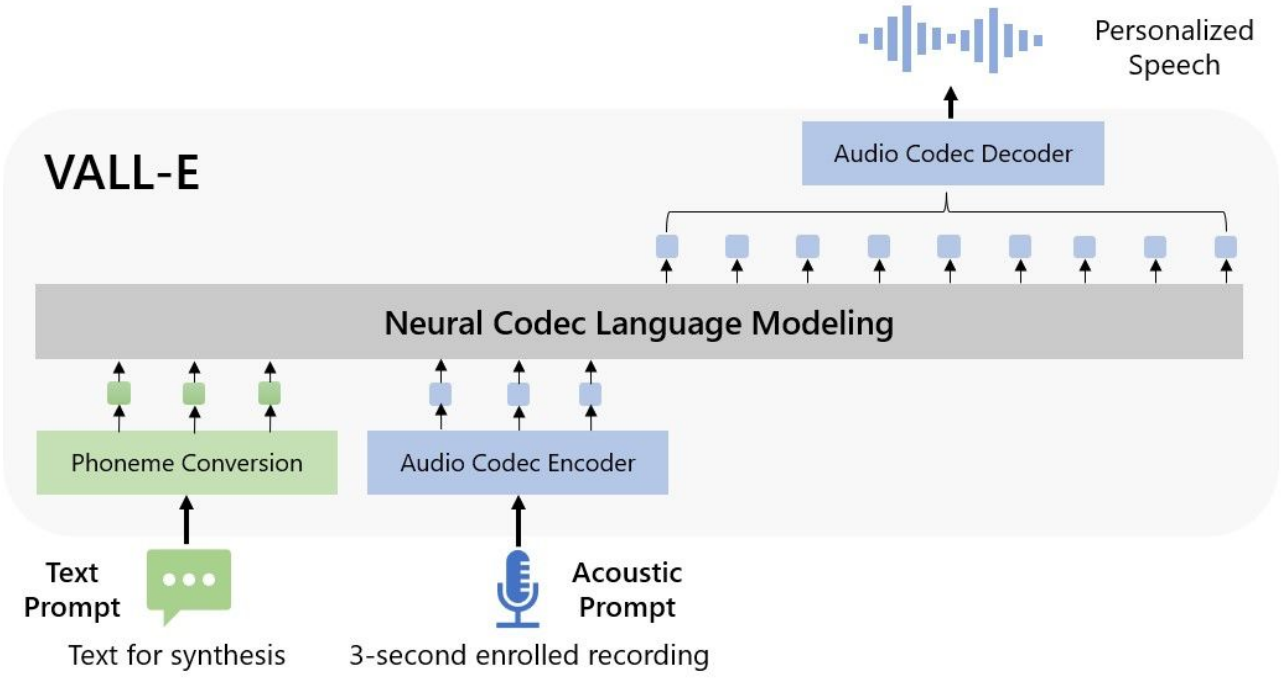
Remaining Challenges

Limited Generalisation: Difficulty in handling unseen timbres, styles, and acoustic environments.

Restricted Control: Current model designs limit fine control over output acoustic characteristics.

Integration Issue in Two-Stage Pipeline: Differences in how EPR (e.g. tokenisation; feature encoding) and EPS models (piano-rolls) represent MIDI cause inconsistencies, leading to loss of precise timing and expressiveness and reducing synthesis quality.

Adapting VALL-E[1] for EPS



Generalisation: Fine-tune EnCodec[2] to compress and tokenise piano audio

Integration: Align MIDI and audio representations more consistently

Control: Extend zero-shot adaptation to piano synthesis, allowing control over acoustical conditions

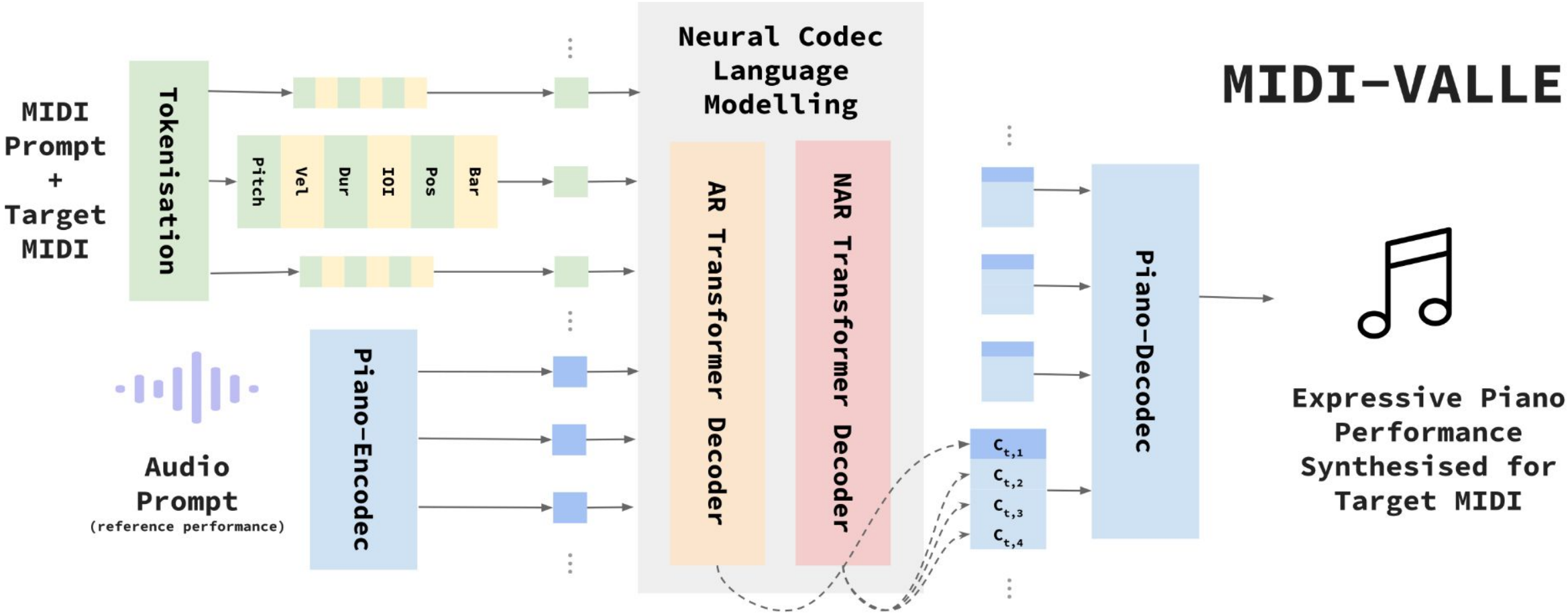
Neural codec language modelling can produce expressive piano audio from MIDI!

[1] Chen, Sanyuan, et al. "Neural codec language models are zero-shot text to speech synthesizers." IEEE Transactions on Audio, Speech and Language Processing (2025).
[2] Défossez, Alexandre, et al. "High Fidelity Neural Audio Compression." Transactions on Machine Learning Research.(2023)

Dataset

	ATEPP[3] Dataset (>700h, >10k recordings)	Maestro[4] Dataset (~200h, ~1k recordings)
Performance Diversity	Multiple pianists, composers, and live settings	Limited to competition recordings
MIDI source	Deep learning-based transcription	High-quality, precise recordings
Acoustic Environment	Varied real-world acoustic environments	Controlled, consistent acoustic settings

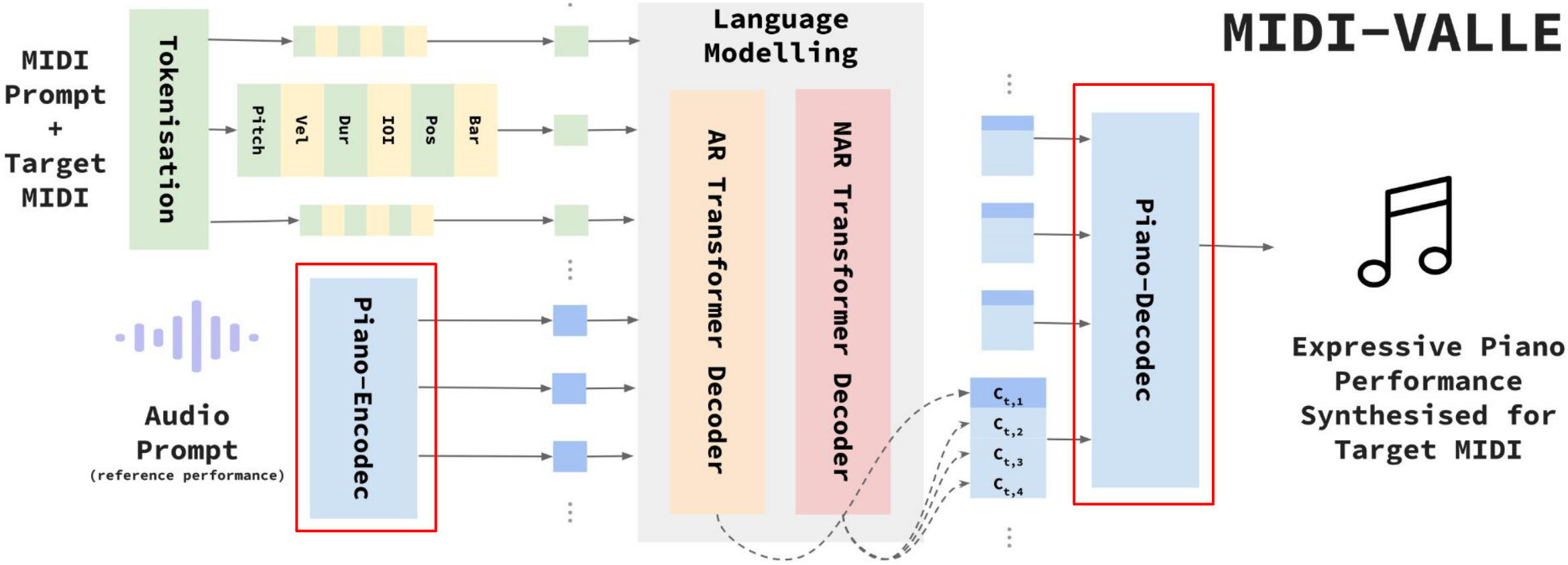
[3] H. Zhang, J. Tang, S. R. Rafee, S. Dixon, G. A. Wiggins, and G. Fazekas, “ATEPP: A Dataset of Automatically Transcribed Expressive Piano Performance,” in International Society for Music Information Retrieval Conference, Dec. 2022, pp. 446–453.
[4] C. Hawthorne, A. Stasyuk, A. Roberts, et al., “Enabling factorized piano music modeling and generation with the MAESTRO dataset,” in International Conference on Learning Representations, 2019.



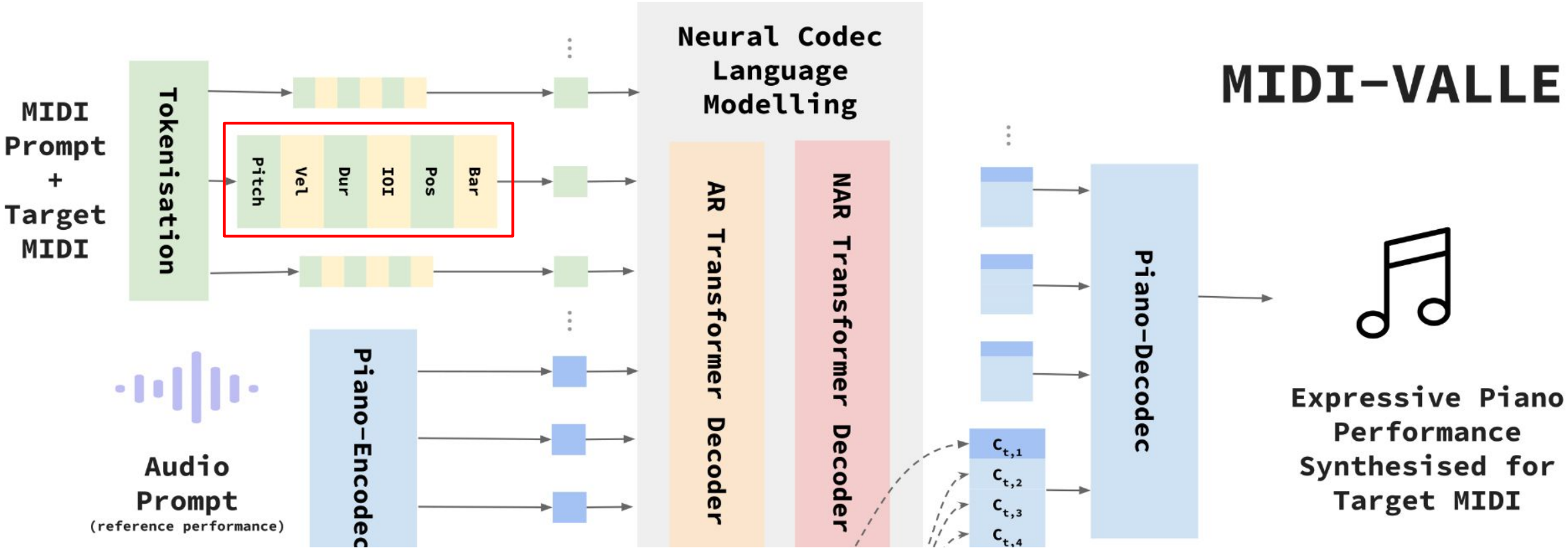
[5] Copet, Jade, et al. "Simple and controllable music generation." Advances in Neural Information Processing Systems 36 (2023): 47704-47720.
[6] Zhu, Hongyuan, et al. "MusicBERT: A self-supervised learning of music representation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

🎵 Audio Tokenisation

- Fine-tune EnCodec from MusicGen[5] on ATEPP dataset to create **Piano-Encodec**.
- Use **Residual Vector Quantisation** to produce discrete audio tokens as 4 codebooks.

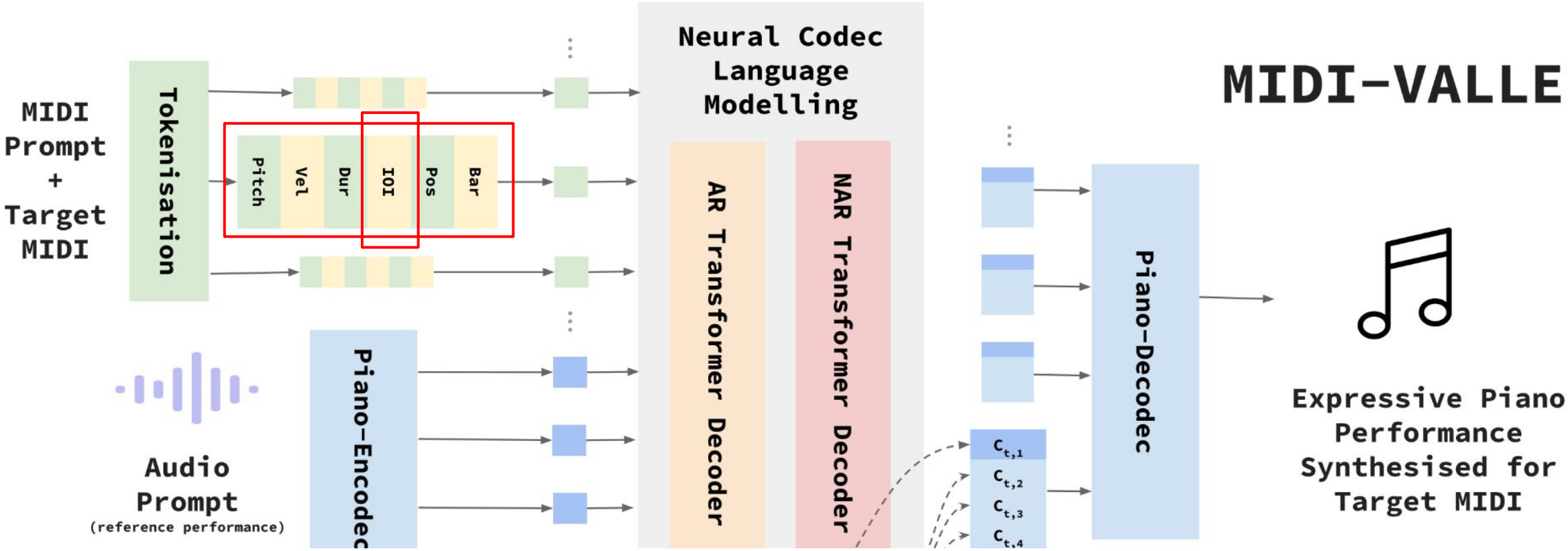


[5] Copet, Jade, et al. "Simple and controllable music generation." Advances in Neural Information Processing Systems 36 (2023): 47704-47720.
[6] Zhu, Hongyuan, et al. "MusicBERT: A self-supervised learning of music representation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.



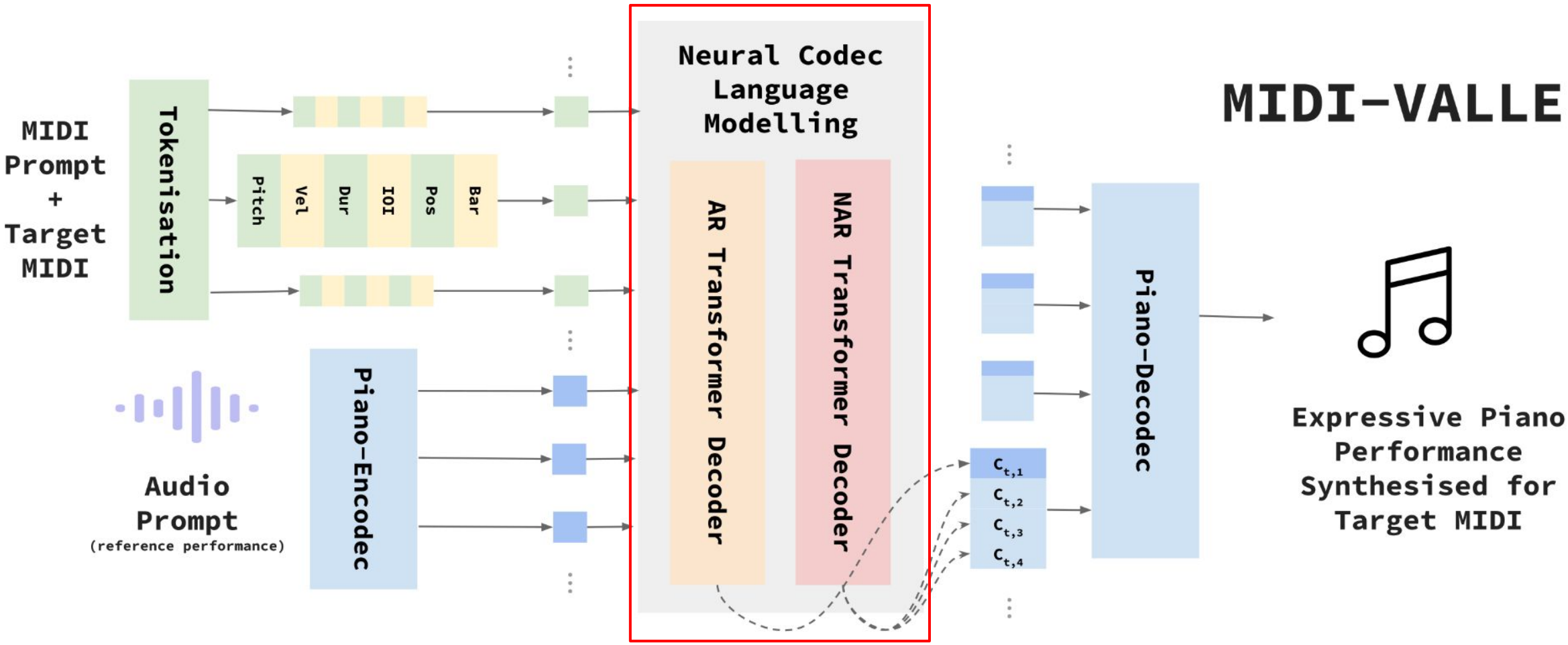
MIDI Tokenisation

- Use the **Octuple**[6] MIDI representation, encoding each note with features
- Include **Inter-Onset Interval** (IOI) tokens to explicitly model expressive onset timing.



MIDI Tokenisation

- Use the **Octuple**[6] MIDI representation, encoding each note with features
- Include **Inter-Onset Interval** (IOI) tokens to explicitly model expressive onset timing.



[5] Copet, Jade, et al. "Simple and controllable music generation." Advances in Neural Information Processing Systems 36 (2023): 47704-47720.
[6] Zhu, Hongyuan, et al. "MusicBERT: A self-supervised learning of music representation." Proceedings of the 29th ACM International Conference on Multimedia. 2021.

Evaluation

Dataset	Genre	MIDI Type	RE*
ATEPP [8]	classical	Transcribed	Live & Studio
Maestro [6]	classical	Recorded	Competition
Pijama [31]	jazz	Transcribed	Live & Studio

RE → Recording Environments

Evaluation: Objective Metrics

Fréchet Audio Distance (**FAD**) with Piano-Encodec

Spectrogram distortion (MAE)

Chroma distortion (MSE)

Evaluation: Objective Metrics

Fréchet Audio Distance (**FAD**) with Piano-Encodec
Spectrogram distortion (MAE)
Chroma distortion (MSE)

Model	Dataset	FAD ↓	Spec. ↓	Chroma ↓
Encodec [21]	ATEPP	–	0.304 ± .005	0.478 ± .011
Piano-Enc.	ATEPP	0.685	0.123 ± .002	0.140 ± .002
	Maestro	0.984	0.135 ± .002	0.139 ± .001
	Pijama	1.133	0.143 ± .003	0.137 ± .001

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
ATEPP				
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
Maestro				
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
Pijama				
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
		ATEPP		
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
		Maestro		
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
		Pijama		
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
ATEPP				
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
Maestro				
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
Pijama				
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
		ATEPP		
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
		Maestro		
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
		Pijama		
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
ATEPP				
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
Maestro				
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
Pijama				
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

Model	Ref.	FAD ↓	Spec. ↓	Chroma ↓
ATEPP				
M2A [3]	GT[1]	11.014	0.218 ± .005	0.421 ± .017
	RC[2]	11.463	0.214 ± .004	0.464 ± .017
MV	GT	3.329	0.219 ± .005	0.436 ± .012
	RC	2.659	0.199 ± .005	0.442 ± .012
Maestro				
M2A [3]	GT	34.479	0.230 ± .003	0.387 ± .007
	RC	33.753	0.224 ± .003	0.427 ± .007
MV	GT	11.281	0.231 ± .004	0.428 ± .009
	RC	9.168	0.206 ± .003	0.420 ± .009
Pijama				
M2A [3]	GT	274.153	0.312 ± .010	0.471 ± .009
	RC	267.969	0.293 ± .008	0.509 ± .010
MV	GT	102.022	0.322 ± .010	0.558 ± .014
	RC	97.634	0.298 ± .009	0.584 ± .015

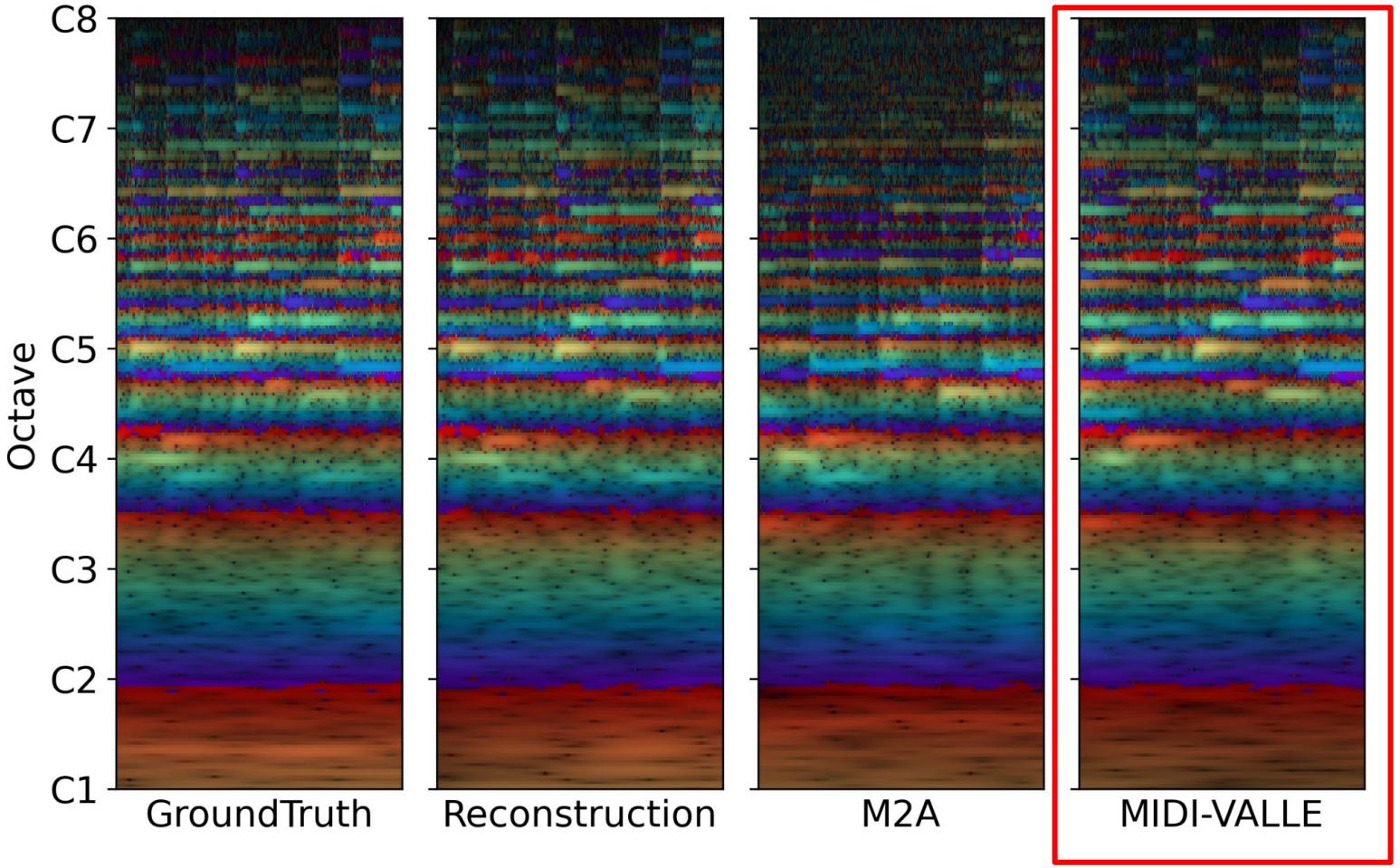
Fréchet Audio Distance (**FAD**) **Spec**trogram distortion
Chroma distortion

MV refers to **MIDI-VALL**
GT refers to the **groudtruth** performance recording

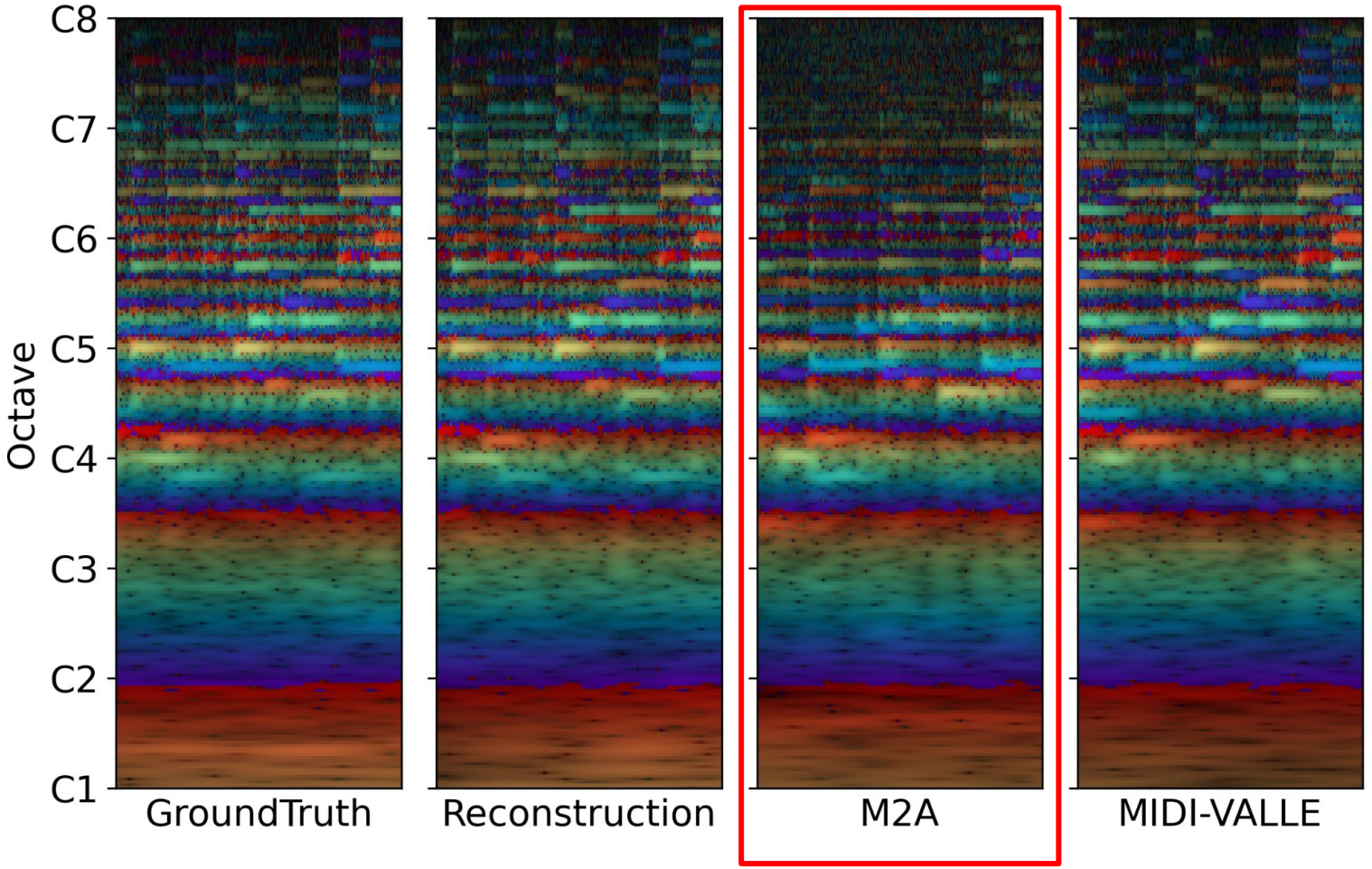
M2A[7] is the baseline
RC indicates audio **reconstructed** via Piano-Encodec.

[7] Tang, Jingjing, et al. "Towards an integrated approach for expressive piano performance synthesis from music scores." ICASSP. IEEE, 2025.

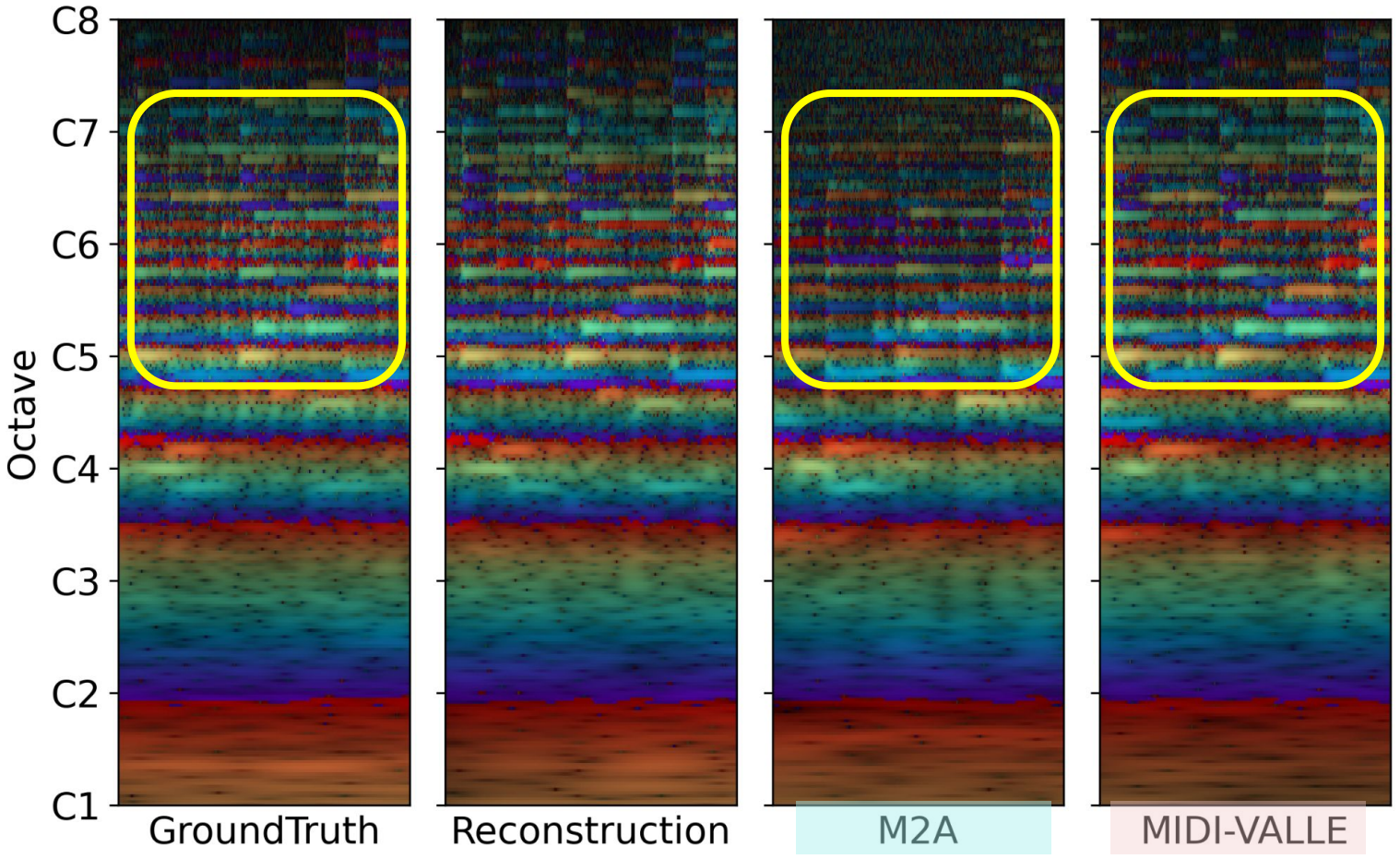
Colour → Frequency and **Lightness → Amplitude**



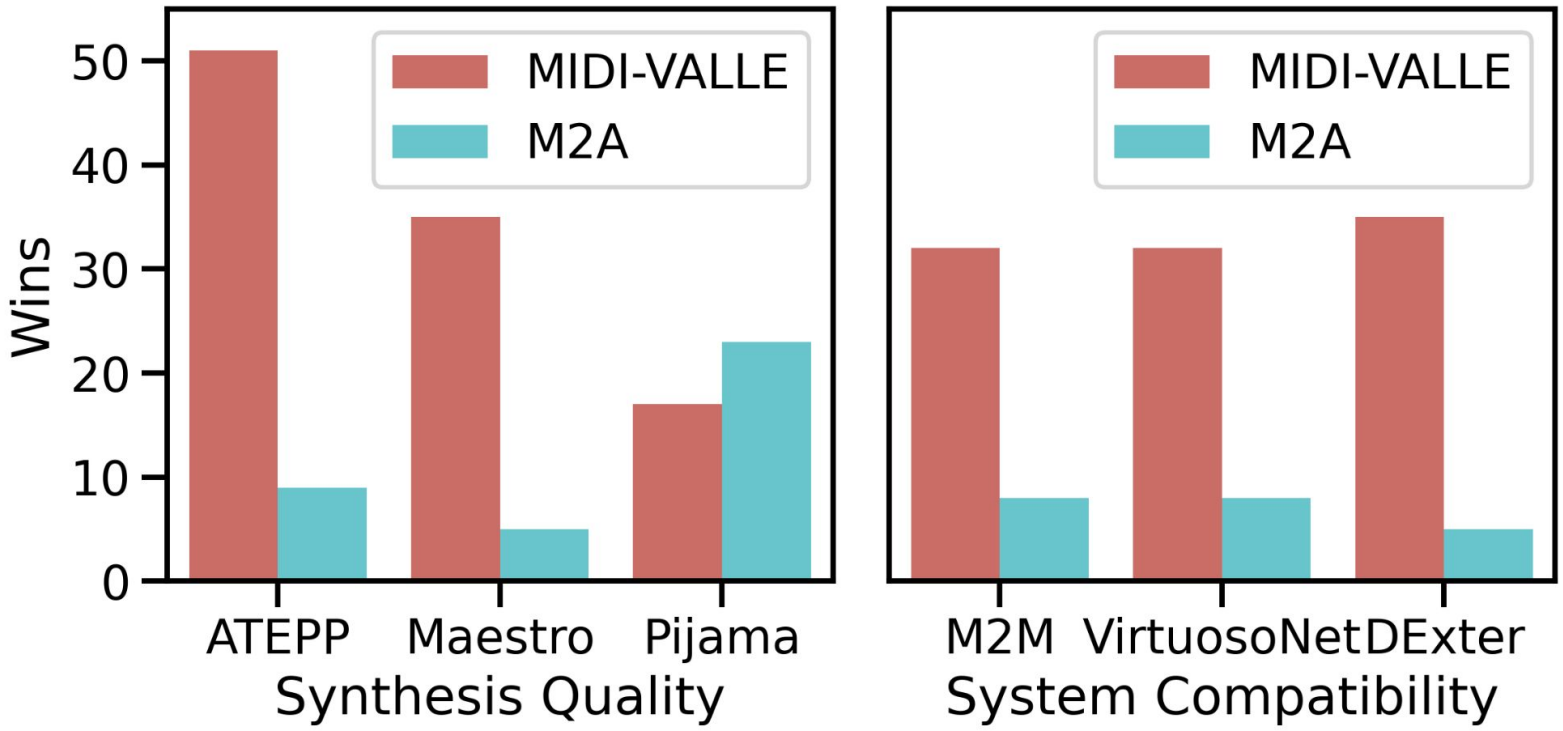
Colour → Frequency and **Lightness → Amplitude**



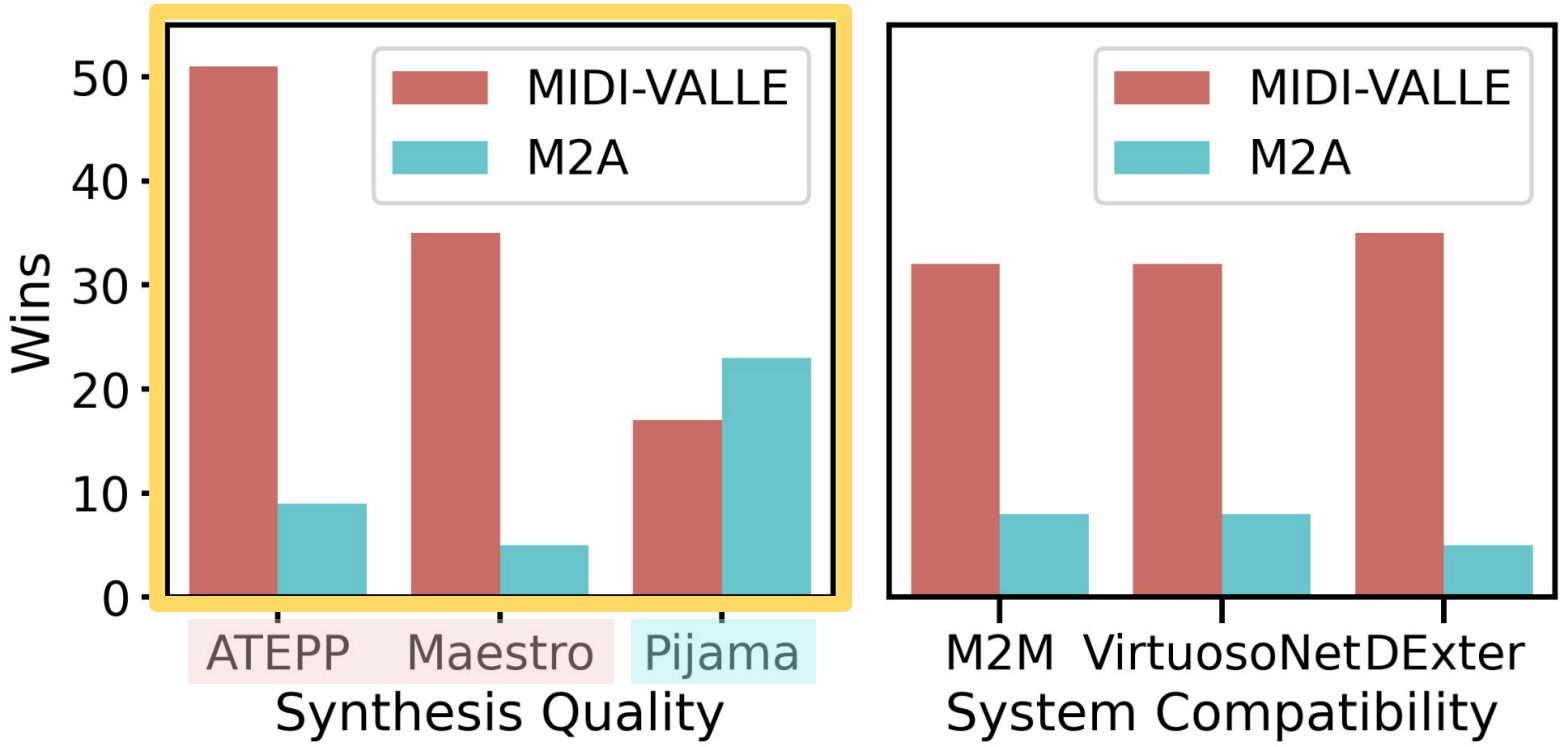
Colour → Frequency and **Lightness → Amplitude**



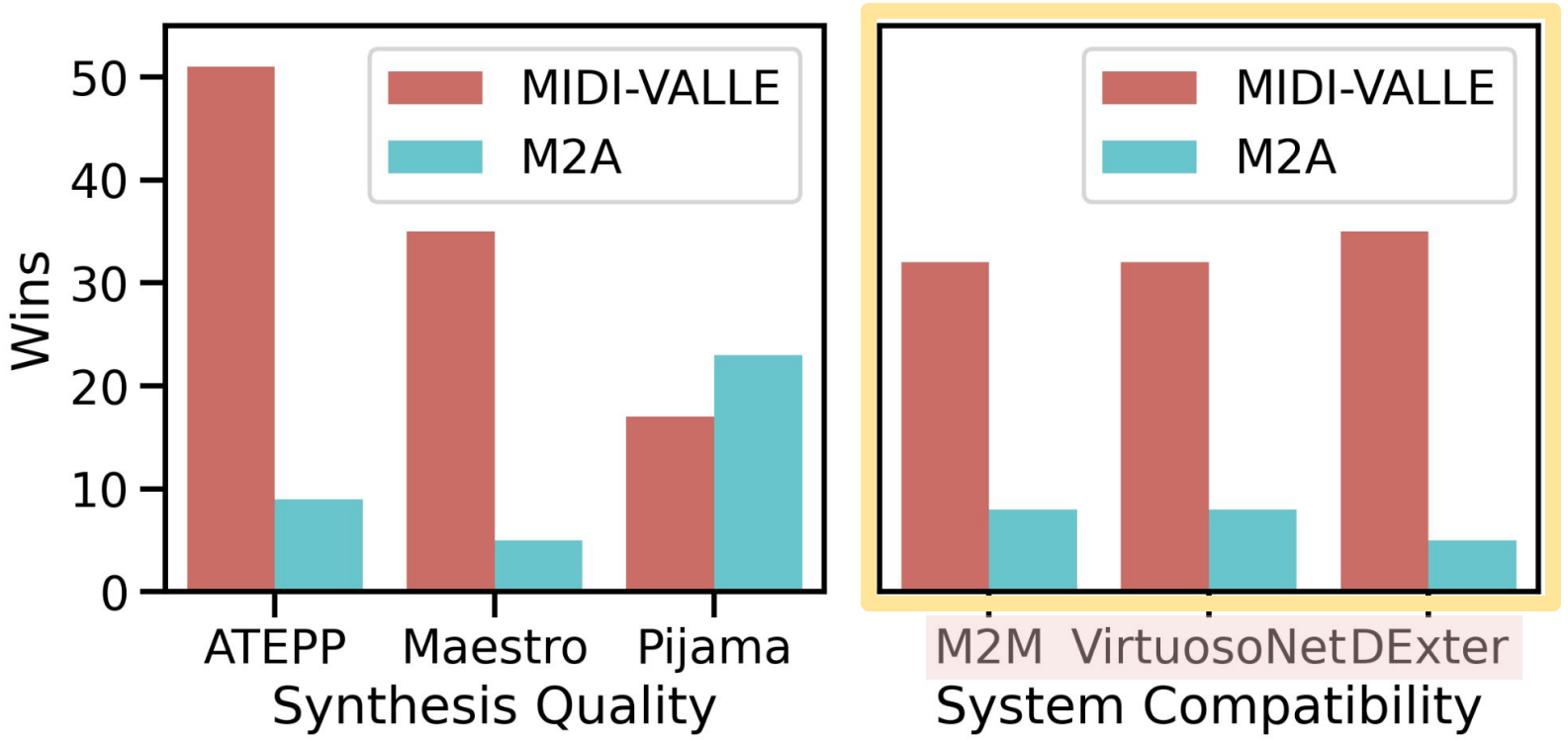
Evaluation: Listening Test



Evaluation: Listening Test



Evaluation: Listening Test



Listening Sample

Excerpt from *Mozart: Piano Sonata No. 13 in B-Flat Major, K. 333: I. Allegro* by András Schiff

More Samples

Thank you for listening!

Demo, codes and checkpoints are all available!

MIDI-VALLE: Improving Expressive Piano Performance Synthesis Through Neural Codec Language Modelling

Jingjing Tang, Xin Wang, Zhe Zhang, Junichi Yamagishi, Geraint Wiggins, György Fazekas

ISMIR 2025

NII 大学共同利用機関法人 情報・システム研究機構 国立情報学研究所 National Institute of Informatics

AI + MUSIC

Queen Mary University of London

UK Research and Innovation