VI. Supplemental Materials: Extra Tables and Figures

TABLE IV: Main attributes shared by the implemented Transformer decoders in Section III.

Attr.	Description	Value
T_{tgt}	target sequence length	1,024
T_{mem}	Transformer-XL memory length	1,024
L	# self-attention layers	12
n_{head}	# self-attention heads	10
d_{e}	token embedding dimension	320
d	hidden state dimension	640
d_{ff}	feed-forward dimension	2,048
d_{c}	condition embedding dimension	512
# params		58.7∼62.6 mil.

TABLE V: Main attributes of our *MuseMorphose* model.

Attr.	Description	Value
T	target sequence length	1,280
L	# self-attention layers	24
$L_{\textbf{enc}}$	# encoder self-attention layers	12
$L_{\textbf{dec}}$	# decoder self-attention layers	12
n_{head}	# self-attention heads	8
d_{e}	token embedding dimension	512
d	hidden state dimension	512
d_{ff}	feed-forward dimension	2,048
$d_{\boldsymbol{z}}$	latent condition dimension	128
$d_{\boldsymbol{a}}$	attribute embedding dimension (each)	64
# params	—	79.4 mil.

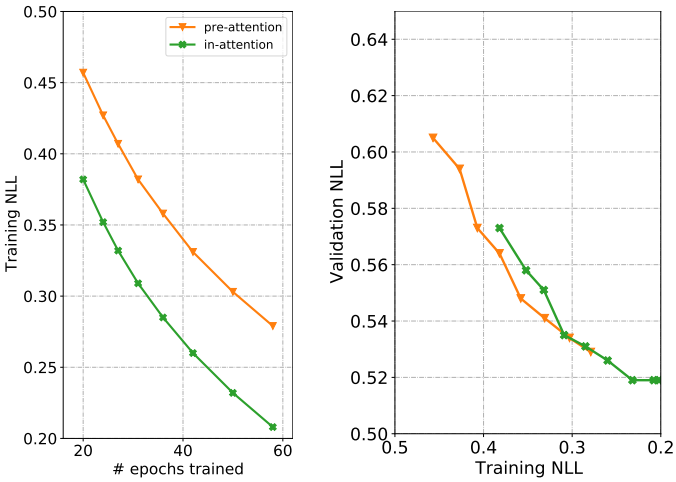
TABLE VI: The vocabulary used to represent songs in *LPD-17-cleansed* dataset, which is adopted in Section III.

Event type	Description	# tokens
Bar	beginning of a new bar	1
Sub-beat	position in a bar, in 32nd note steps (𝅘𝅥𝅯)	32
Tempo	32∼224 bpm, in steps of 3 bpm	64
Pitch*	MIDI note numbers (pitch) 0∼127	1,757
Velocity*	MIDI velocities 3∼127	544
Duration*	multiples (1∼64 times) of 𝅘𝅥𝅯	1,042
All events	—	**3,440**

*: unique for each of the 17 tracks (instruments)

TABLE VII: The vocabulary used to represent piano songs in *AILabs.tw-Pop1K7* dataset, on which *MuseMorphose* (see Section IV) is trained.

Event type	Description	# tokens
Bar	beginning of a new bar	1
Sub-beat	position in a bar, in 16th note steps (𝅘𝅥𝅯)	16
Tempo	32∼224 bpm, in steps of 3 or 6 bpm	54
Pitch	MIDI note numbers (pitch) 22∼107	86
Velocity	MIDI velocities 40∼86	24
Duration	multiples (1∼16 times) of 𝅘𝅥𝅯	16
Chord	chord markings (root & quality)	133
All events	—	**330**



(a) # epochs vs. training loss (b) Training loss vs. val. loss

Fig. 6: Training dynamics of *pre-attention* and *in-attention* Transformers on *LPD-17-cleansed* dataset from the 20th epoch onwards (best viewed in color).

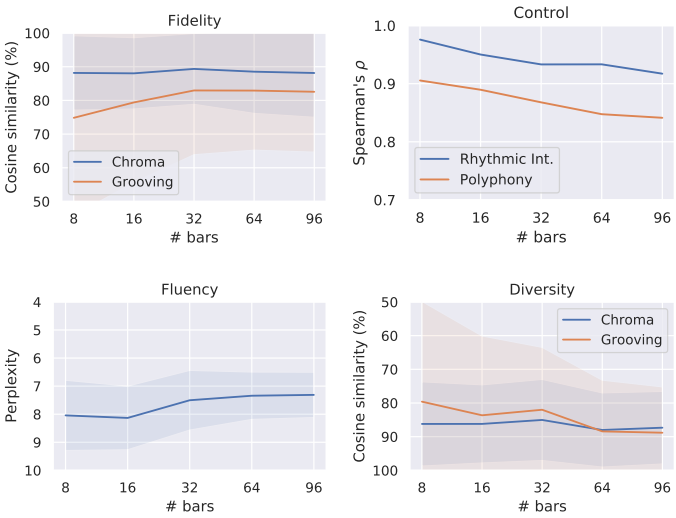
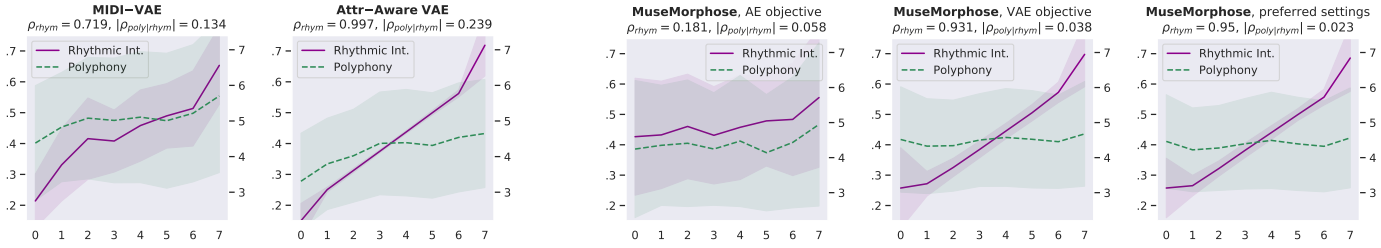
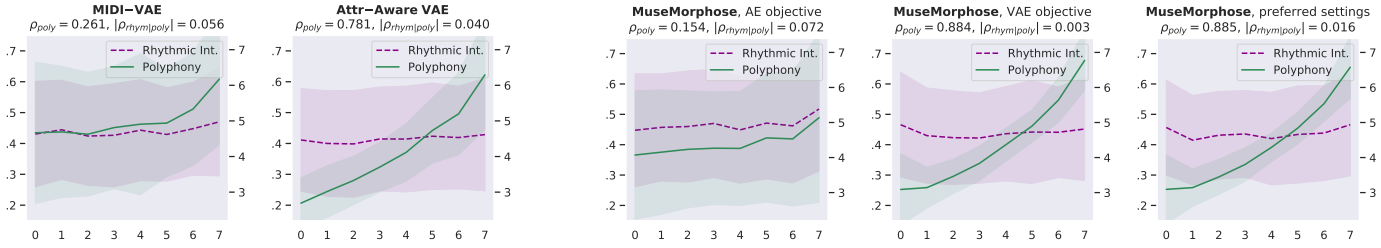


Fig. 7: Evaluation on MuseMorphose's generations of different lengths. (Y-axis for "Fluency" and "Diversity" plots are inverted since these two metrics are the lower the better; shaded regions indicate ±1 std from the mean. Best viewed in color.)



(a) On controlling **rhythmic intensity** (*x-axis*: user-specified $\tilde{a}^{\text{rhym}}$; *y-axis*: s^{rhym} (left), s^{poly} (right) computed from the resulting generations; error bands indicate ± 1 std from the mean).



(b) On controlling **polyphony** (*x-axis*: user-specified $\tilde{a}^{\text{poly}}$; *y-axis*: s^{rhym} (left), s^{poly} (right) computed from the resulting generations; shaded regions indicate ± 1 std from the mean).

Fig. 8: Comparison of the models on attribute controllability. We desire both a high correlation ρ_a and a low $|\rho_{a'|a}|$, where a is the attribute in question, while a' is not.