# Encoder-Only Transformers for Melodic Harmonization: Representation Emergence and Inference Strategies

**Maximos Kaliakatsos-Papakostas** MAXIMOSKP@HMU.GR
**Dimos Makris** DIMOS@HMU.GR
**Konstantinos Soiledis** KSOILEDIS@HMU.GR
**Konstantinos-Theodoros Tsamis** KTSAMIS@HMU.GR
*Department of Music Technology and Acoustics Hellenic Mediterranean University, Greece Archimedes, Athena RC, Greece*

**Editors:** D. Herremans, K. Bhandari, A. Roy, S. Colton, M. Barthet

## Abstract

This paper addresses the problem of melodic harmonization –the automatic generation of harmonic accompaniments that complement a given melody– using non-autoregressive, encoder-only transformer models operating on a synchronized melody–harmony time grid. The proposed framework allows flexible conditioning, such as fixing chords at specific positions, while maintaining high generative quality. Comparative experiments show that single-encoder models outperform dual-encoder architectures despite using fewer parameters. Interestingly, harmony-related attention patterns emerge even when harmony tokens remain fully masked during training, and models using only cross-attention achieve comparable results, suggesting implicit modeling of harmony–harmony relations. Different inference unmasking strategies further reveal notable effects on harmonic structure and coherence.

**Keywords:** Melodic harmonization; Non-autoregressive transformer; Encoder-only architecture; Attention dynamics

## 1. Introduction

Transformer architectures have emerged as powerful sequence modeling frameworks across domains such as language, vision, and music Vaswani et al. (2017); Huang et al. (2018). Within symbolic music generation, melodic harmonization –the task of generating a harmonic accompaniment given a melody–poses unique challenges: it requires local melodic compatibility and long-range harmonic coherence. Harmonization therefore serves as a strong testbed for studying how sequence models integrate and structure multiple musical dimensions across time.

Early neural approaches to automatic harmonization relied on recurrent architectures Lim et al. (2017); Yeh et al. (2021); Chen et al. (2021); Yi et al. (2022), while more recent transformer-based methods Rhyu et al. (2022); Huang and Yang (2024); Wu et al. (2024a); Bhandari et al. (2025) typically frame harmonization as a sequence-to-sequence translation task, generating harmony autoregressively. However, autoregressive decoding enforces a left-to-right generation order, limiting flexibility when imposing harmonic constraints (e.g., fixed cadences or key modulations) prior to generation. Some non-autoregressive approaches perform diffusion in a continuous approximation of the discrete token space Mittal et al.

(2021); Lv et al. (2023), while others apply diffusion in the latent space of a VAE Zhang et al. (2023).

Non-autoregressive, encoder-only transformers offer a compelling alternative. Such models have been explored in other domains through the lens of discrete diffusion or masked token modeling, as in MaskGIT Chang et al. (2022); Austin et al. (2021), which iteratively refines partially masked token sequences. These approaches enable flexible conditioning and substantially faster generation compared to autoregressive models.

Applied to melodic harmonization, this framework represents melody and harmony on a synchronized temporal grid and progressively unmask harmony tokens until a complete harmonization is produced Kaliakatsos-Papakostas et al. (2025). This formulation naturally supports partial conditioning and parallel generation, aligning more closely with the inherently bidirectional nature of harmonic reasoning.

Recent work has shown that training such models effectively requires careful curriculum design. When harmony tokens are gradually unmasked during training, the model learns to rely on melodic context for harmonic inference rather than trivial self-copying patterns. Interestingly, even when all harmony tokens remain masked for long portions of training, harmony-related self-attention patterns still emerge in the generative encoder. This suggests that the model develops an implicit sense of harmonic organization purely through its exposure to melodic structures.

Experiments presented in this paper with cross-attention-only architectures, where neither melody nor harmony use self-attention, reveal that performance remains comparable to full-attention models. This finding hints that cross-attention can, to some degree, internalize self-like relational patterns among harmony tokens. Similar phenomena have been observed in multimodal transformers Tsai et al. (2019); Alayrac et al. (2022), where cross-modal layers spontaneously capture intra-modal dependencies despite the absence of explicit self-attention mechanisms.

Finally, inference in non-autoregressive harmonization models introduces new challenges: since generation need not proceed sequentially, the order of unmasking can be chosen flexibly. We therefore investigate several unmasking strategies—starting from cadential positions, high-confidence predictions, or random tokens—and analyze their musical and structural implications through both quantitative metrics and attention visualizations.

In summary, this work explores how encoder-only, non-autoregressive transformers learn harmonic structure under different attention and unmasking regimes. Beyond their practical advantages for constraint-based harmonization, the observed emergent behaviors offer new insights into how structured musical relations can arise from weak or indirect supervision. Future work will further examine the underlying mechanisms and draw connections to broader studies on emergent attention dynamics in multimodal and self-supervised learning systems. The code of the work presented in this paper is available online[1].

## 2. Method

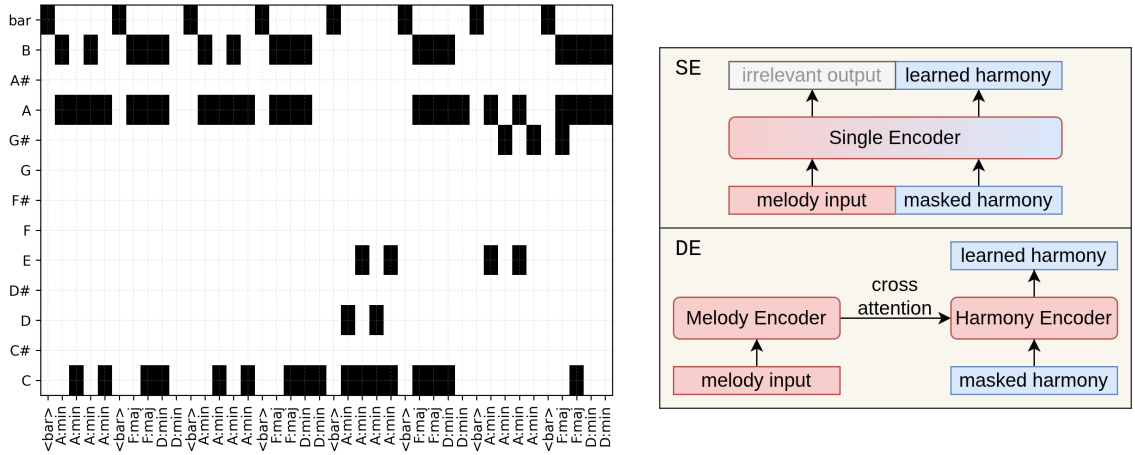
This section describes the melody and harmony representations, the single- and dual-encoder architectures, the training procedure, and the inference strategies.

1. https://github.com/NeuraLLMuse/EncoderOnlyMelHarmSelfCross.git

### 2.1. Melody and harmony representation

A quarter-note resolution is sufficient to capture all harmonic details in the datasets used for training and testing, with no overlapping chords within the same segment. Melody events occurring within each quarter note are grouped and represented as a binary *pitch-class* piano roll with an additional binary column marking bar boundaries. Formally, the pitch-class matrix is defined as $\mathbf{PC} \in \{0,1\}^{L\times 13}$, where $L$ is the number of quarter-note steps. The first 12 columns correspond to the 12 pitch classes, following Rhyu et al. (2022), where active pitch classes are indicated by 1. The 13th column is zero everywhere except at bar onsets, where it is 1 and all other pitch-class columns are zero.

Harmony is represented as a sequence of chord tokens from a fixed vocabulary $\mathcal{V}$, denoted $\mathbf{y} \in \mathcal{V}^L$. Chord symbols are normalized following the `mir_eval` convention Raffel et al. (2014) (e.g., `Cmaj7` instead of `C`△). The vocabulary includes $12 \times 29 = 348$ chord types (12 pitch classes $\times$ 29 chord qualities). Harmony is aligned to the same quarter-note grid: if a chord spans multiple steps, it is repeated for its duration. For example, a `C:maj7` spanning two beats occupies two grid positions. Special tokens handle missing or padding cases: `<nc>` denotes "no chord," `<pad>` fills trailing positions beyond the harmonization length, and `<bar>` marks bar boundaries. Both melody and harmony representations thus encode bar-level structure explicitly. Figure 1 (a) shows an example segment from the test dataset.




(a) Music representation

(b) `SE` and `DE` architectures

Figure 1: (a) Example of a pitch-class piano roll ($13\times T$ matrix) and the respective harmony tokens as x-axis labels. (b) Overview of `SE` and `DE` architectures.

### 2.2. Model architectures

The proposed transformer architectures, abstractly illustrated in Figure 1 (b), are based on BERT Devlin et al. (2019) and adapted for generation through masked language modeling (MLM). Two variants are explored: (a) a **single-encoder** model (`SE`), where the input sequence jointly encodes melody and harmony information, and (b) a **dual-encoder** model (`DE`), with a dedicated melody encoder and a harmony-generative encoder connected via

cross attention. Both models predict chord tokens conditioned on a melodic context and on a varying proportion of visible (unmasked) harmony tokens.

During inference, the harmony sequence is initially fully masked using `<mask>` tokens. The model then iteratively unmasks tokens in $t$ steps, providing at each step a partially masked harmony input $\mathbf{y}_{\text{in}}^{(t)}$. Although accelerated multi-token unmasking strategies exist Kaliakatsos-Papakostas et al. (2025), we focus here on single-token unmasking for clarity.

During training, the models learn to estimate the conditional distribution:

$$p_\theta\left(\mathbf{y}_{\text{target}}^{(k)} \mid \mathbf{y}_{\text{in}}^{(k)}, \mathbf{m}\right), \tag{1}$$

where $\mathbf{y}_{\text{target}}^{(k)}$ denotes the subset of harmony tokens to be predicted at training step $k$, and $\mathbf{m}$ is the melody matrix $\mathbf{PC} \in \{0,1\}^{L\times 13}$.

The melody matrix is first projected through a linear layer before entering the melody encoder of either architecture. The harmony input (masked and unmasked tokens) is passed through an embedding layer. In the `SE` model, the transformer output corresponding to the harmony portion is used to compute a cross-entropy loss for predicting masked harmony tokens, while the melody portion of the output is ignored. In the `DE` model, the melody encoder provides contextual information to the harmony decoder via cross attention, enabling the latter to learn to reconstruct harmony tokens at its output.

### 2.3. Training and inference

At the beginning of training, all harmony tokens are masked, and only the melody is visible. This setup compels the model to establish cross-attention pathways between melody and harmony. As training progresses, harmony tokens are gradually revealed, transitioning from full masking to partial visibility. This progression enables the model to learn both extreme regimes: full reliance on melody and partial self-reliance on visible harmony context. Interestingly, models trained entirely in the fully masked regime still produce high-quality harmonizations, as discussed in Section 3.

The number of visible harmony tokens at training step $k$ is defined as

$$\#\text{unmasked} = \min(\lfloor v \cdot L \rfloor,\, L-1)\,, \tag{2}$$

where the visible fraction $v$ follows

$$v = \left(\frac{k}{k_{\text{total}}}\right)^5, \tag{3}$$

with $k$ the current training step and $k_{\text{total}}$ the total number of steps. The exponent of 5 allocates roughly half of the training duration to the fully masked regime; similar values produce comparable performance.

Let $\mathcal{H}$ denote the set of all harmony tokens, $\mathcal{M}^{(k)} \subseteq \mathcal{H}$ the set of masked positions, and $\mathcal{U}^{(k)} = \mathcal{H} \setminus \mathcal{M}^{(k)}$ the visible tokens at step $k$. The model input is defined as

$$y_i^{(k)} = \begin{cases} y_i, & i \in \mathcal{U}^{(k)}, \\ \texttt{<mask>}, & i \in \mathcal{M}^{(k)}. \end{cases} \tag{4}$$

The prediction targets are the masked positions, $\mathbf{y}_{\text{target}}^{(k)} = \{y_i \mid i \in \mathcal{M}^{(k)}\}$, and the MLM loss is computed as

$$\mathcal{L}^{(k)} = -\sum_{i \in \mathcal{M}^{(k)}} \log p_\theta\left(y_i \mid \mathbf{m}, \mathbf{y}_{\text{in}}^{(k)}\right). \tag{5}$$

At inference time, generation begins from a fully masked harmony sequence and proceeds for $L$ unmasking steps. At each step, one masked token is selected and predicted according to one of five unmasking strategies:

`start` Sequentially from the first to the last token, mimicking autoregressive decoding.

`end` From the last to the first token, prioritizing cadential regions Allan and Williams (2004).

`random` Selecting masked positions uniformly at random.

`certain` Selecting the position with the lowest logit entropy (highest model confidence).

`uncertain` Selecting the position with the highest logit entropy (lowest model confidence).

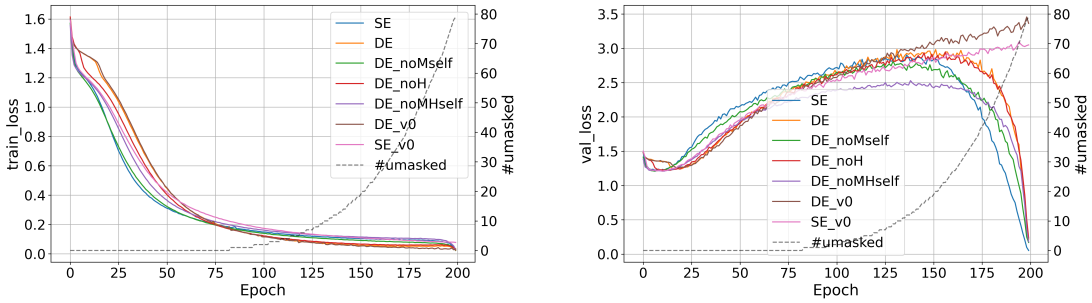
Once a position is selected, the model samples a prediction from $\hat{\mathbf{y}}^{(t)} \sim p_\theta(\cdot \mid \mathbf{m}, \mathbf{y}_{\text{in}}^{(t)})$, and updates the input sequence: $\mathbf{y}_{\text{in}}^{(t+1)} = \mathbf{y}_{\text{in}}^{(t)} \cup \hat{\mathbf{y}}^{(t)}$. All experiments used nucleus sampling ($p = 0.9$) with temperature 0.2.

All models had 8 layers and 8 heads per layer for each encoder – one encoder for the `SE` and two for the `DE` architectures. Models were trained using AdamW with a learning rate of $1 \times 10^{-4}$, batch size 8, for 200 epochs. For models trained with the gradual unmasking curriculum, the final-epoch version was retained, as it encompasses all curriculum stages. For models trained entirely with masked harmony, the checkpoint with the lowest validation loss was used. Training was performed on three NVIDIA RTX 3080 GPUs. The loss was averaged over tokens and batches.

## 3. Results

Experiments are conducted on a curated version of the HookTheory dataset Yeh et al. (2021) (15,440 MIDI lead sheets), following previous harmonization studies Rhyu et al. (2022); Huang and Yang (2024). To reflect harmonic rhythm, redundant chord repetitions within bars are removed, and all pieces are transposed to C major or A minor using the Krumhansl key-finding algorithm Krumhansl (2001). The split comprises 14,679 training and 761 validation/test pieces (95/5%). Training and validation losses are illustrated in Figure 2. Training accuracy (i.e., percentage of correctly unmasked tokens) for all architectures reached between 90-95% during the all-masked harmony epochs and increased to over 99% as harmony tokens were gradually unmasked. The `v0` (no unmasking) versions reached over 98%. Test-set accuracy reached over 65% for all architectures during the all-masked harmony epochs (remained so for the `v0` versions) and reached over 98% as the unmasked input tokens gradually increased.

Generated melodic harmonizations are evaluated both *in-domain* (HookTheory test split) and *out-of-domain* (650 curated jazz standards). Each model generates harmonizations for the melodies in these sets, which are evaluated against ground-truth harmonies using established chord- and rhythm-based metrics Sun et al. (2021); Wu et al. (2024b): **CHE**




(a) Training loss `DE` (b) Validation loss

Figure 2: Training and validation loss for all examined models.

(Chord Histogram Entropy), **CC** (Chord Coverage), **CTD** (Chord Tonal Distance), **CTnCTR** (Chord Tone ratio), **PCS** (Pitch Consonance Score), **MCTD** (Melody–Chord Tonal Distance), **HRHE**, **HRC**, and **CBS**. Average ground-truth statistics for both datasets are shown in Table 1. In future extensions, we also plan to supplement quantitative metrics with qualitative listening studies and curated harmonization examples, enabling a more perceptual assessment.

Table 1: Average metric values for all pieces in the test set (*in-domain*) and jazz set (*out-of-domain*) datasets.

| Ground truth | CHE | CC | CTD | CTnCTR | PCS | MCTD | HRHE | HRC | CBS |
|---|---|---|---|---|---|---|---|---|---|
| Test set | 1.4078 | 4.9485 | 0.9748 | 0.7769 | 0.4060 | 1.4139 | 0.4542 | 1.9710 | 0.2314 |
| Jazz set | 2.2027 | 11.6471 | 0.8208 | 0.8297 | 0.3145 | 1.4042 | 0.5093 | 2.0607 | 0.2426 |

### 3.1. Effect of Unmasking Order

We first evaluate five unmasking strategies during inference: `start`, `end`, `certain`, `uncertain`, and `random`. For this comparison we use the single-encoder (`SE`) model; all other models produced similar results. Mean absolute error (MAE) is computed between generated and reference harmonizations across all metrics.

Results (Table 2) show that the `certain` strategy—unmasking tokens for which the model exhibits the highest confidence—consistently outperforms others in both in-domain and out-of-domain settings. This suggests that harmonization generation benefits from data-driven uncertainty guidance rather than fixed-order decoding. Notably, the same ranking of strategies holds across all metrics, implying robust inference behavior independent of musical style.

### 3.2. Ablation Study: Architectural Insights

We next compare single-encoder (`SE`) and dual-encoder (`DE`) architectures and their ablations under the `certain` unmasking regime (Table 3).

Table 2: Comparison of unmasking order strategies during inference in the *in-domain* test set and *out-of-domain* jazz set using the `SE` model architecture. Mean absolute errors (MAEs) are calculated, and the smallest differences per metric are shown in bold. Results are presented in ascending order of average MAE, which is show in the last column.

| Instance | CHE | CC | CTD | CTnCTR | PCS | MCTD | HRHE | HRC | CBS | avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| In-domain / Test set | | | | | | | | | | |
| certain | **1.3235** | **4.8536** | **0.9126** | 0.7933 | 0.3940 | 1.4158 | **0.4520** | **2.0383** | **0.1225** | 1.3673 |
| start | 1.4521 | 5.4406 | 0.9467 | 0.7949 | 0.3910 | 1.4161 | 0.5312 | 2.2243 | 0.1489 | 1.4829 |
| end | 1.5202 | 5.8311 | 0.9525 | 0.7961 | 0.3879 | 1.4166 | 0.5865 | 2.3509 | 0.1640 | 1.5562 |
| random | 1.6076 | 6.3113 | 0.9862 | 0.7962 | 0.3968 | **1.4125** | 0.7235 | 2.7269 | 0.2094 | 1.6856 |
| uncertain | 1.7405 | 7.1583 | 0.9971 | **0.7877** | **0.3848** | 1.4178 | 0.8159 | 2.8958 | 0.2624 | 1.8289 |
| Out-of-domain / Jazz set | | | | | | | | | | |
| certain | **1.8768** | **8.6660** | **0.8723** | 0.8002 | **0.3890** | 1.3851 | **0.5401** | **2.5351** | **0.1098** | 1.9083 |
| start | 2.0047 | 9.5180 | 0.9115 | **0.7872** | 0.3924 | 1.3860 | 0.6048 | 2.6907 | 0.1252 | 2.0467 |
| end | 2.0551 | 9.8634 | 0.8978 | 0.8051 | 0.3863 | 1.3807 | 0.6424 | 2.7514 | 0.1381 | 2.1023 |
| random | 2.1457 | 10.7059 | 0.9466 | 0.8047 | 0.3948 | **1.3782** | 0.8397 | 3.2182 | 0.2006 | 2.2927 |
| uncertain | 2.2651 | 11.8899 | 0.9767 | 0.8021 | 0.3935 | 1.3810 | 0.9283 | 3.3795 | 0.2399 | 2.4729 |

The `SE` model achieves the best overall results, particularly in rhythm-related metrics for the out-of-domain jazz set, despite having less than half the parameters of `DE`. In-domain, `DE_noM` (dual encoder without melody self-attention) performs slightly better, indicating that cross-attention can compensate for missing melody self-context. Surprisingly, models trained with fully masked harmony throughout training (`v0`) do not collapse, supporting the hypothesis that harmonic structure can be indirectly inferred from melodic patterns alone. Even more strikingly, the `DE_noMH` variant (no self-attention in either encoder) remains functional, suggesting that cross-attention alone can partially encode both melody–harmony and harmony–harmony dependencies—a key insight for future investigation.

### 3.3. Attention Dynamics

Figure 3 visualizes averaged attention maps across layers and heads for representative models. Even when harmony tokens remain masked during all training epochs (`DE_v0`), coherent self-attention structures emerge in the harmony encoder. When melody self-attention is removed (`DE_noM`), harmony self-attention reorganizes, seemingly compensating for missing melodic structure. Cross-attention in `DE_noM` remains similar to the full model, while in `DE_noMH` it becomes diffuse, implying an adaptive redistribution of representational load. These emergent behaviors highlight the model's ability to develop internal harmonic organization even under heavily constrained or degenerate training regimes. A complete analysis should compare these attention patterns with those of randomly initialized encoders. We leave this comparison to future work, but note that such baselines would clarify which structures truly reflect learned harmonic representations.maximoskalpap@gmail.com

Table 3: Comparison of ablations in the *in-domain* test set and *out-of-domain* jazz set using the `certain` unmasking order. Mean absolute errors (MAEs) are calculated, and the smallest differences per metric are shown in bold. Results are presented in ascending order of average MAE, which is show in the last column.

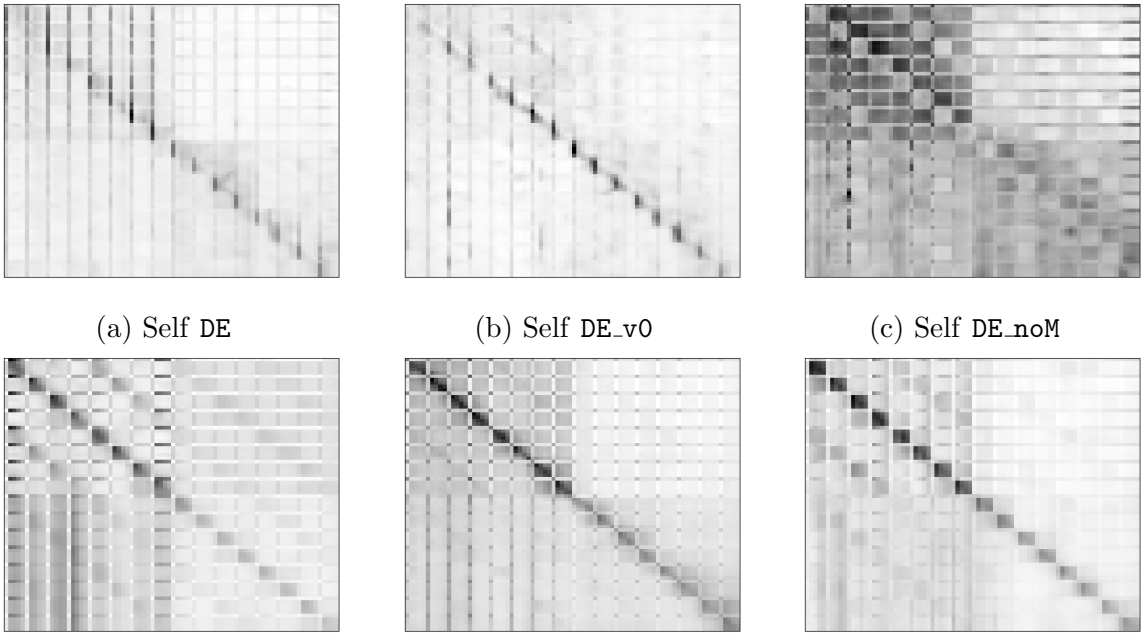
| Instance | CHE | CC | CTD | CTnCTR | PCS | MCTD | HRHE | HRC | CBS | avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| In-domain / Test set | | | | | | | | | | |
| DE_noM | **1.2017** | **4.0778** | 0.9547 | **0.7920** | 0.4217 | 1.4117 | **0.4492** | 2.0831 | 0.1263 | 1.2798 |
| SE | 1.3235 | 4.8536 | **0.9126** | 0.7933 | **0.3940** | 1.4158 | 0.4520 | **2.0383** | **0.1225** | 1.3673 |
| DE | 1.3181 | 4.6293 | 0.9235 | 0.7895 | 0.4235 | 1.4105 | 0.7327 | 2.7639 | 0.2220 | 1.4681 |
| DE_v0 | 1.4143 | 5.2916 | 0.9595 | 0.7981 | 0.4288 | 1.4049 | 0.8867 | 3.1161 | 0.2928 | 1.6214 |
| DE_noH | 1.4295 | 5.3351 | 0.9547 | 0.8006 | 0.4264 | 1.4065 | 0.8821 | 3.1069 | 0.2871 | 1.6254 |
| DE_noMH | 1.3928 | 5.3259 | 0.9484 | 0.8056 | 0.4349 | 1.4046 | 0.9980 | 3.4024 | 0.3093 | 1.6691 |
| SE_v0 | 1.5219 | 5.9354 | 0.9741 | 0.7999 | 0.4345 | **1.4045** | 1.0972 | 3.6240 | 0.3869 | 1.7976 |
| Out-of-domain / Jazz set | | | | | | | | | | |
| SE | 1.8768 | 8.6660 | 0.8723 | 0.8002 | 0.3890 | **1.3851** | **0.5401** | **2.5351** | **0.1098** | 1.9083 |
| DE_noH | 1.7924 | 8.1879 | 0.8040 | 0.7630 | 0.3801 | 1.4018 | 0.9228 | 3.3510 | 0.2539 | 1.9841 |
| DE_noM | 1.7932 | 8.1935 | **0.8030** | **0.7626** | 0.3795 | 1.4016 | 0.9253 | 3.3548 | 0.2543 | 1.9853 |
| DE_v0 | **1.7919** | 8.1879 | **0.8030** | 0.7631 | 0.3800 | 1.4015 | 0.9275 | 3.3662 | 0.2535 | 1.9861 |
| SE_v0 | **1.7919** | **8.1860** | **0.8030** | 0.7630 | **0.3794** | 1.4022 | 0.9276 | 3.3700 | 0.2537 | 1.9863 |
| DE_noMH | 1.7925 | 8.1954 | 0.8040 | 0.7631 | 0.3800 | 1.4016 | 0.9263 | 3.3662 | 0.2543 | 1.9871 |
| DE | 1.7935 | 8.2030 | 0.8040 | 0.7633 | 0.3800 | 1.4016 | 0.9299 | 3.3700 | 0.2549 | 1.9889 |

## 4. Conclusions

This paper presented a non-autoregressive approach to melodic harmonization based on an encoder-only transformer trained with a synchronized melody–harmony grid. The study compared multiple architectural and inference variants, showing that single-encoder models can achieve superior harmonization quality with fewer parameters than dual-encoder counterparts. Furthermore, the results highlighted two intriguing phenomena: (i) harmony-related self-attention patterns emerge even when harmony tokens remain fully masked during training, and (ii) models relying solely on cross-attention can perform comparably to those using both self- and cross-attention, suggesting that cross-attention may implicitly capture harmony–harmony relations.

Experiments on inference unmasking strategies revealed that different decoding orders –such as starting from cadential regions or from high-confidence tokens– can meaningfully affect the musical structure and coherence of generated harmonies. Future work will focus on understanding the mechanisms underlying these emergent attention behaviors, refining inference scheduling strategies, and exploring the cognitive and music-theoretical interpretations of non-autoregressive harmonization dynamics.

This study focuses on objective chord and rhythm based metrics, which provide reliable and widely used indicators of harmonization quality. Nevertheless, perceptual aspects such as naturalness and stylistic preference are not captured by such metrics. Incorporating listening-based evaluations is an important direction for future work and could further validate the musical qualities observed in our generated harmonizations. Another promising direction involves human-in-the-loop evaluation of controllable harmonization, assessing



(a) Self `DE` (b) Self `DE_v0` (c) Self `DE_noM`

(d) Cross `DE` (e) Cross `DE_noMH` (f) Cross `DE_noM`

Figure 3: Average attention maps in the harmony decoding encoder of some ablations across all layers and heads, averaged across melodic harmonizations of all test data with the `certain` unmasking method.

how flexible unmasking strategies and partial conditioning support real-world compositional workflows.

## Acknowledgments

This work has been partially supported by project MIS 5154714 of the National Recovery and Resilience Plan Greece 2.0 funded by the European Union under the NextGenerationEU Program.

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

Moray Allan and Christopher Williams. Harmonising chorales by probabilistic inference. *Advances in neural information processing systems*, 17, 2004.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Keshav Bhandari, Sungkyun Chang, Tongyu Lu, Fareza R Enus, Louis B Bradshaw, Dorien Herremans, and Simon Colton. Improvnet–generating controllable musical improvisations with iterative corruption refinement. *arXiv preprint arXiv:2502.04522*, 2025.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022.

Yi-Wei Chen, Hung-Shin Lee, Yen-Hsing Chen, and Hsin-Min Wang. Surprisenet: Melody harmonization conditioning on user-controlled surprise contours. *arXiv preprint arXiv:2108.00378*, 2021.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer, 2018. URL https://arxiv.org/abs/1809.04281.

Jingyue Huang and Yi-Hsuan Yang. Emotion-driven melody harmonization via melodic variation and functional representation. *arXiv preprint arXiv:2407.20176*, 2024.

Maximos Kaliakatsos-Papakostas, Dimos Makris, Konstantinos Soiledis, Konstantinos-Theodoros Tsamis, Vassilis Katsouros, and Emilios Cambouropoulos. Diffusion-inspired masked language modeling for symbolic harmony generation on a fixed time grid. *Applied Sciences*, 15(17):9513, 2025.

Carol L Krumhansl. *Cognitive foundations of musical pitch*. Oxford University Press, 2001.

Hyungui Lim, Seungyeon Rhyu, and Kyogu Lee. Chord generation from symbolic melody using blstm networks. *arXiv preprint arXiv:1712.01011*, 2017.

Ang Lv, Xu Tan, Peiling Lu, Wei Ye, Shikun Zhang, Jiang Bian, and Rui Yan. Getmusic: Generating any music tracks with a unified representation and diffusion framework. *arXiv preprint arXiv:2305.10841*, 2023.

Gautam Mittal, Jesse Engel, Curtis Hawthorne, and Ian Simon. Symbolic music generation with diffusion models. *arXiv preprint arXiv:2103.16091*, 2021.

Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. Mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, volume 10, page 2014, 2014.

Seungyeon Rhyu, Hyeonseok Choi, Sarah Kim, and Kyogu Lee. Translating melody to chord: Structured and flexible harmonization of melody with transformer. *IEEE Access*, 10:28261–28273, 2022.

Chung-En Sun, Yi-Wei Chen, Hung-Shin Lee, Yen-Hsing Chen, and Hsin-Min Wang. Melody harmonization using orderless nade, chord balancing, and blocked gibbs sampling. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4145–4149. IEEE, 2021.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Shangda Wu, Yashan Wang, Xiaobing Li, Feng Yu, and Maosong Sun. Melodyt5: A unified score-to-score transformer for symbolic music processing. *arXiv preprint arXiv:2407.02277*, 2024a.

Shangda Wu, Yue Yang, Zhaowen Wang, Xiaobing Li, and Maosong Sun. Generating chord progression from melody with flexible harmonic rhythm and controllable harmonic density. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):4, 2024b.

Yin-Cheng Yeh, Wen-Yi Hsiao, Satoru Fukayama, Tetsuro Kitahara, Benjamin Genchel, Hao-Min Liu, Hao-Wen Dong, Yian Chen, Terence Leong, and Yi-Hsuan Yang. Automatic melody harmonization with triad chords: A comparative study. *Journal of New Music Research*, 50(1):37–51, 2021.

Li Yi, Haochen Hu, Jingwei Zhao, and Gus Xia. Accomontage2: A complete harmonization and accompaniment arrangement system. *arXiv preprint arXiv:2209.00353*, 2022.

Jincheng Zhang, György Fazekas, and Charalampos Saitis. Fast diffusion gan model for symbolic music generation controlled by emotions. *arXiv preprint arXiv:2310.14040*, 2023.