METEOR: Melody-aware Texture-controllable Symbolic Music Re-Orchestration via Transformer VAE

Dinh-Viet-Toan Le[1] and **Yi-Hsuan Yang**[2]
[1]Univ. Lille, CNRS, Inria, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France
[2]National Taiwan University, Taiwan
dinhviettoan.le@univ-lille.fr, yhyangtw@ntu.edu.tw

Abstract

Re-orchestration is the process of adapting a music piece for a different set of instruments. By altering the original instrumentation, the orchestrator often modifies the musical texture while preserving a recognizable melodic line and ensures that each part is playable within the technical and expressive capabilities of the chosen instruments. In this work, we propose METEOR, a model for generating Melody-aware Texture-controllable re-Orchestration with a Transformer-based variational auto-encoder (VAE). This model performs symbolic instrumental and textural music style transfers with a focus on melodic fidelity and controllability. We allow bar- and track-level controllability of the accompaniment with various textural attributes while keeping a homophonic texture. With both subjective and objective evaluations, we show that our model outperforms style transfer models on a re-orchestration task in terms of generation quality and controllability. Moreover, it can be adapted for a lead sheet orchestration task as a zero-shot learning model, achieving performance comparable to a model specifically trained for this task.

1 Introduction

Re-orchestration refers to the musical arrangement of an existing music piece for a different set of instruments [Cacavas, 1975]. In the context of popular music, this notion is often associated with “song covers”. A key similarity between the original piece and its re-orchestration often lies in maintaining melodic fidelity. In Western music, which is predominantly homophonic, a primary melody is typically supported by an accompanying background [Young and Roens, 2022]. Moreover, in the composition process, effective orchestration requires knowledge of writing for various instruments by combining their timbres, while being restricted by their physical limitations [Adler and Hesterman, 1989].

Going further, re-orchestration extends beyond simply reassigning parts of the original piece to instruments in a new ensemble. It often involves altering the overall *musical texture* of the piece to suit artistic goals or ensemble constraints. Musical texture refers to how different musical streams are written, organized, and combined [Huron, 1989]. An orchestral score can be described by global characteristics, such as instrument groupings or part diversity, and part-specific attributes such as rhythmicity or repetitiveness [Le *et al.*, 2022].

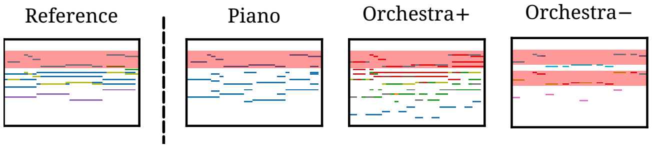


Figure 1: METEOR’s re-orchestration task. The model can reorchestrate a reference for multiple instrumentations (*e.g.* solo piano, or orchestra) with texture controls, with more (orchestra+) or less (orchestra-) “polyphonicity” and “rhythmic intensity” (cf. Section 3.1). The models ensures melodic fidelity (red highlight) with fine-grained controls (melodic instrument choice and pitch range).

In the field of symbolic music generation, re-orchestration can be considered as a *style transfer* task, for which a model is designed to replicate a reference piece while altering high-level musical attributes. However, existing style transfer systems may be inadequate for specifically a re-orchestration task. They often focus on band arrangements [Zhao *et al.*, 2024; Luo *et al.*, 2024] which restricts the instrument choices to a fixed and small ensemble and does not allow fine-grained selection of instrumentation. Moreover, these systems often overlook or even disregard the melodic fidelity of the generated content. For example, according to FIGARO [von Rütte *et al.*, 2023]: “some salient features such as melodies are often not preserved”.

Beyond the instrumentation choice, re-orchestration implies textural controls, for which style transfer systems have also been implemented. This control is often performed at a piece-level [Lu *et al.*, 2023] or bar-level [Wu and Yang, 2023b]. For orchestral music – more generally, multi-track music – such control can also occur at the track level.

In this study, we present METEOR, a model for **Me**lody-aware **Te**xture-controllable re-**Or**chestration (Section 3). The model is designed to achieve the following (Figure 1):

- Multi-track music re-orchestration: the model automatically orchestrates a reference multi-track piece, with the instrumentation possibly specified by the user.
- Texture-controllability: textural attributes can be controlled at both bar and track levels.

Model	Multi-track	Texture controllability		Melodic fidelity	Instrument choice		Open-source[1]
		Track-level	Bar-level		Full ensemble	Melody	
MuseMorphose [Wu and Yang, 2023b]	✗	✗	✓	✗	✗	✗	✓
MuseBarControl [Shu *et al.*, 2024]	✗	✗	✓	✗	✗	✗	✗
FIGARO [von Rütte *et al.*, 2023]	✓	✓	✓	✗	✗	✗	✓
PopMAG [Ren *et al.*, 2020]	✓	✗	✗	✓ *(ind. track)**	*fixed (6)*†	✗	✗
GetMUSIC [Lv *et al.*, 2023]	✓	✗	✗	✓ *(ind. track)**	*fixed (5)*†	✗	✗
BandControlNet [Luo *et al.*, 2024]	✓	✓	✓	✓ *(ind. track)**	*fixed (6)*†	✗	✗
AccoMontage-band [Zhao *et al.*, 2024]	✓	✗	*implicit***	✓ *(ind. track)**	*implicit***	✗	✓
METEOR (ours)	✓	✓	✓	✓	✓	✓	✓

Table 1: Models related to the style transfer sub-tasks performed by METEOR. (^{1}) We consider models to be open-source when both the code and trained models are publicly available. (*) The melody is added *a posteriori* as an independent track, in contrast with METEOR where the melodic instrument is chosen *among* the chosen instrumentation. (**) This model mimics the texture and the instrumentation of an *already existing* source: the choices are not explicit. (†) These models only handle a *fixed* number of instrument types (*e.g.* 5 or 6).

- Melodic fidelity: the melody is preserved in the re-orchestrated piece, with the option for the user to select the melodic instrument.

To our best knowledge, METEOR is the first deep generative model that offers both *instrumental* and *texture-based* style transfers with melodic fidelity. Moreover, we show that our model can perform a lead sheet orchestration task without further training in a zero-shot manner. The main approach relies on an extension of MuseMorphose, a Transformer-based VAE, with bar- and track-level token constraints and inference guidance for melodic fidelity. Section 4 provides an objective evaluation demonstrating METEOR's effectiveness in bar- and track-level controllability, melodic fidelity, and melodic instrument playability. A subjective evaluation further supports that it generates higher-quality re-orchestrations than baseline models. We share audio extracts of generations on a demo page and open source code and model weights[1].

2 Related Works

Re-orchestration is a task which can be associated with multi-track music *style transfer*, which aims at generating a multi-track piece by taking a multi-track reference and altering musical characteristics to reflect a specific style [Dai *et al.*, 2018]. Style transfer can refer to composer style transfer, where a music style is applied to a reference content [Cífka *et al.*, 2020]. Though, our study focuses on two types of music style transfers: *instrumental* style transfer, where the instrumentation of the reference piece is altered and *texture-based* style transfer, where high-level musical features from the reference are adjusted to generate a new piece. We specifically explore these tasks within the context of *homophonic* music, which consists of a melody supported by an accompaniment. Multiple models have been developed to address sub-tasks, with their strengths and limitations summarized in Table 1.

2.1 Multi-track Homophonic Music Generation

Several models have been developed for multi-track music free generation [Ens and Pasquier, 2020; Liu *et al.*, 2022; Dong *et al.*, 2023] which can generate music without an initial musical reference. Comparatively, few studies have explicitly addressed the task of re-orchestration [von Rütte *et al.*, 2023]. Closest style transfer models for this task focus on band arrangements [Ren *et al.*, 2020; Lv *et al.*, 2023; Luo *et al.*, 2024; Zhao *et al.*, 2024]. Such models usually only consider a fixed-number instrumental ensemble composed of generic instruments, such as drums, piano, or strings. Moreover, while band music is usually written using a homophonic texture, defined as a primary melody supported by an accompaniment [Benward, 2018], the melodic part is often overlooked. These models either discard the melodic content [von Rütte *et al.*, 2023] or only generate the accompaniment and insert the melodic content *a posteriori* into a track played by a *fixed* instrument such as a synthesizer [Luo *et al.*, 2024] or a "lead" track [Zhao *et al.*, 2024], without strict physical restrictions like its ambitus or register. However, in styles such as Western classical orchestral music, the melody is assigned to a specific instrument or a group of instruments which can change throughout the piece to achieve particular timbre effects [Adler and Hesterman, 1989]. In such cases, the melody must respect the instrument's limitations, such as its range.

Adapting multi-track style transfer models for re-orchestration is not direct. In particular, AccoMontage-band [Zhao *et al.*, 2024], designed for lead sheet band arrangement, suffers from several limitations for re-orchestration. Beyond the *a posteriori* melody insertion, it must rely on transcribing the multi-track input into a lead sheet used as input by the model. This dependency leads to challenges: the simplification of the textural content and a risk of transcription errors.

2.2 Texture-Based Style Transfer

Musical texture characterizes how musical streams are organized and describes their content [Huron, 1989]. Texture-based style transfer systems often offer control on a reference piece over attributes such as rhythmic density [Wu and Yang, 2023b] or pitch distributions [von Rütte *et al.*, 2023]. Multiple levels of controllability can be defined *i.e.* global, time-varying (often bar-level), and track-level (for multi-track music) controllabilities. Global features characterize the whole generated sequence; time-varying attributes only impact a

[1] https://github.com/dinhviettoanle/meteor

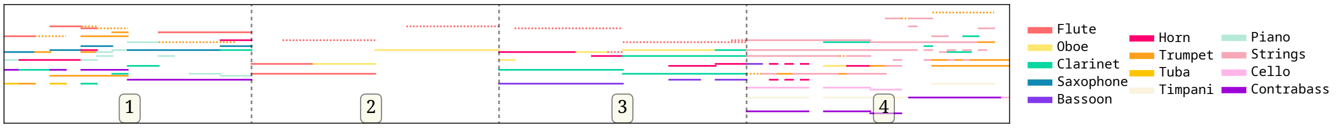


Figure 2: Pianoroll of a 8-bar re-orchestration generation by METEOR with various textural and instrumentation constraints changing each 2 bars. Melody is dashed. (1) Automatic instrumentation, no textural changes. (2) Flute + oboe duet, with a melodic flute, low polyphonicity. (3) Wind quintet, with a melodic flute, low rhythmicity. (4) Classical orchestra, with a melodic trumpet, high rhythmicity and polyphonicity.

single bar; and track-level controllability affects a single track either globally or locally at the bar level.

Bar-level textural controls are implemented by MuseMorphose [Wu and Yang, 2023b] and MuseBarControl [Shu *et al.*, 2024] for single-track piano music and FIGARO [von Rütte *et al.*, 2023] for multi-track music. BandControlNet [Luo *et al.*, 2024] adds track-level controls. AccoMontage-band [Zhao *et al.*, 2024] addresses a multi-track lead sheet arrangement task through texture transfer, but its controllability is limited. The model applies the texture of a "texture donor" to a musical content, restricting texture controllability to the set of *pre-existing* texture donors.

MuseMorphose [Wu and Yang, 2023b] appears to be a promising model for our task, offering fine-grained textural controls at a bar level. Though, a straightforward multi-track extension of the model may be insufficient for addressing the re-orchestration task, for example, due to the lack of melodic control. In contrast, the ideas introduced in Compose & Embellish [Wu and Yang, 2023a], a lead sheet piano arrangement model, provide insights that could address the issue of melodic control, in particularly through its approach of interleaving one-bar segments of melody and accompaniment, serving as an inspiration for the design of our model.

3 Methods

In this section, we introduce METEOR, a Transformer-based VAE model for multi-track re-orchestration with instrumentation controllability, bar- and track-wise texture controllability and melodic fidelity. We first present the musical attributes considered for textural controllability, and the technical contributions, particularly the tokenization strategies developed for this task.

3.1 Textural Attributes

METEOR is a model designed for both instrumental and textural style transfer (Figure 2). Specifically, its textural style transfer function enables the control of various textural attributes. We consider two levels of controllability: "bar-wise" (*i.e.* all tracks may be influenced by the control attribute) and "bar- and track-wise" (*i.e.* each track can be individually controlled at a bar level). We first consider bar-wise control attributes following [Wu and Yang, 2023b].

- *Rhythmic intensity* (or *rhythmicity*): number of sub-beats having at least one note played within a bar containing B sub-beats, regardless of the track. With $\mathbf{1}(\cdot)$ the indicator function, $s^{\text{rhym}} = \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}(n_{\text{onset},b} \geq 1)$.
- *Polyphonicity*: average number of notes played (hit or held) during a sub-beat in a bar containing B sub-beats, including all tracks. We consider $s^{\text{poly}} = \frac{1}{B}\sum_{b=1}^{B}(n_{\text{onset},b} + n_{\text{hold},b})$.

Each bar is characterized by a raw value of polyphonicity and rhythmicity. These raw values are then split into 8 bins with a similar number of bars in each bin. These bins are set according to distribution of polyphonicity and rhythmicity values in the dataset.

For finer-grained control, we propose "bar-wise and track-wise" control attributes aiming at controlling each instrument individually among those initially selected at a bar level.

- *Average pitch*: average pitch of the set of pitches $\{p_1, \ldots, p_M\}$ played in a track t in a bar, expressed in MIDI value and rounded to the nearest ten.

$$p_{\text{t}}^{\text{avg}} = \text{round}\left(\frac{1}{M}\sum_{i=1}^{M} p_i, 10\right)$$

Levels of average pitches are thus divided into 13 classes, spanning from 10 to 130. For instance, this attribute can be used to assign high register to melodic instruments, and low register for bass parts.

- *Pitch diversity*: number of different pitch classes played in a track in a bar.

$$p_{\text{t}}^{\text{diversity}} = \left|\left\{p_i \mod 12 \,\middle|\, i = 1, 2, \ldots, M\right\}\right|$$

Levels of pitch diversity are divided into 13 classes, spanning from 0 to 12. Low pitch diversity can relate to bass parts, repeated notes or arpeggios, while high pitch diversity can encourage passing notes, embellishments or extended chords.

3.2 Tokenization, Model & Control Strategies

METEOR's architecture is based on MuseMorphose [Wu and Yang, 2023b], originally developed for piano style transfer. The model implements a VAE based on Transformers encoders and decoders (Figure 3). We first extend its initial REMI tokenization [Huang and Yang, 2020] using the REMI+ tokenization [von Rütte *et al.*, 2023] which handles multi-track music. Based on early experiments, we implement REMI+ using a vertical parsing[2], where notes are grouped and ordered based on time rather than track. We also rely on a "pitch class + octave encoding" of the pitches [Li *et al.*, 2023] instead of absolute MIDI values, in particular, to handle melodies independently of the original octave register.

For instrumentation controllabillity, the user can select the playing instruments from a subset of 64 instruments defined

[2] https://musiclang.github.io/tokenizer

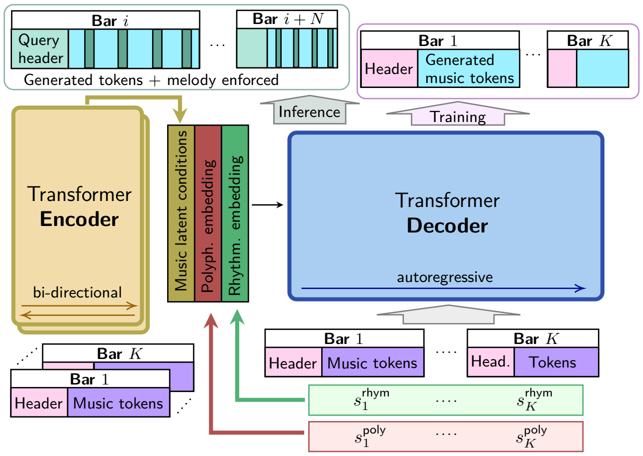


Figure 3: Architecture of METEOR, based on MuseMorphose. The musical content in each bar is preceded by a *header* describing the playing instruments in this bar and track-wise controls. During training, the model is trained to reconstruct K bars. At inference time, the user can specify different headers for each bar and starts the generation of N bars starting from bar $i < K$ (*i.e.* the user can ask to generate only from a sub-part of the full piece). The inference is guided with melody constraints at a beat level.

in [Dong *et al.*, 2023] or the ensemble can be automatically defined by the model. Instrument selection is handled through `DescriptionTrack-[track]` tokens that indicate instruments playing in a bar which are added in a *header* at the start of each bar in the token sequence.

For texture controllability, the model implements multiple controls over various textural attributes (Section 3.1). For bar- and track-wise controls, `PitchAvg-[track]-[level]` tokens and `PitchDiversity-[track]-[level]` tokens are added jointly in this header to describe the average pitch and pitch diversity level of each track. A token sequence is shown in Figure 4. Following MuseMorphose, the bar-wise polyphonicity and rhythmicity classes are encoded in a separate sequence of bar-level conditions, which is embedded and concatenated with the latent vector and used as condition in the decoder through an "in-attention" mechanism.

Regarding the overall training process, the model is trained as an end-to-end model on the SymphonyNet dataset composed of 46k multi-track pieces [Liu *et al.*, 2022]. Similar to MuseMorphose, the loss function used is a β-VAE objective with free bits. The resulting model is 67M parameter-large and is trained for one week on a single RTX 6000 24GB GPU.

3.3 Inference Guidance for Melodic Fidelity

Melodies are crucial elements in music, as they often make a piece easily recognizable [Stefani, 1987]. Thus, a key focus of our model is melodic fidelity, ensuring that the original melody is preserved in the generated extract, with possibly different textures in the accompaniment parts. Models preserving the melody often insert *a posteriori* a track containing the melody played by a generic instrument (*e.g.* synthesizer), which prevents any melodic ornamentation [Le *et al.*, 2022] and restricts its integration into the queried ensemble. Thus, we propose an *inference guidance* process designed to ensure the melodic fidelity in a more flexible way in the generation.

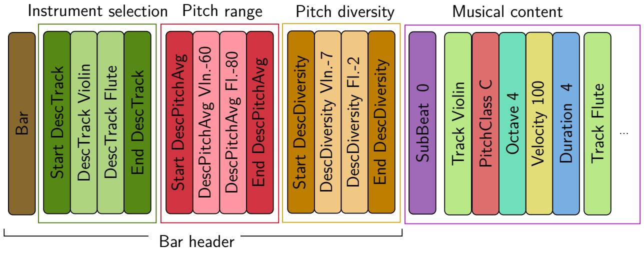


Figure 4: Example of token sequence for a bar with a violin and a flute. As indicated in the *bar header*, the violin plays in a medium register and has a high pitch diversity, while the flute is in the upper range, with a low pitch diversity.

First, the melody is identified in the original piece during a pre-processing step using a bar-wise and track-wise skyline algorithm. The melody in each bar is estimated as being the track with the highest average pitch within that bar[3].

The instrument playing the melody is first chosen by the model or can be specified by the user. In particular, the model or the user may choose to use different instruments to play the melody in different bars of the generated piece. The melody notes are then generated alongside with the re-orchestration using *inference guidance*: tokens identified as melody in the original piece are treated as beat-level conditions at inference time. Following Figure 3 (top left), after a `Bar` token and the enforced *header* describing this bar, each `Sub-beat` token generated by the model is followed by an enforced `Track` token corresponding to the chosen melodic instrument, along with the tokens corresponding to the melody note played at this time position (*i.e.* pitch class, octave, duration, and velocity). The next tokens (*i.e.* all the accompaniment tokens until the next melodic tokens) are then generated auto-regressively. In particular, we do not restrict the model to generate additional notes played by the melodic instrument. In further experiments (see Table 3), we allow the model to infer `Octave` tokens to evaluate its relation with the instruments' register.

3.4 Zero-Shot Lead Sheet Orchestration

While METEOR has been specifically trained for a re-orchestration task, if can be adapted into a lead sheet orchestration model without requiring further training, effectively performing as a zero-shot learning model. The model takes as input a lead sheet provided as a multi-track MIDI file, composed of a melodic track and a second track with block chords. By interpreting the lead sheet as a low-rhythmicity multi-track piece, METEOR is able to orchestrate this lead sheet with specific instruments by increasing the rhythmicity.

4 Evaluation

In this section, we first present an objective evaluation to assess our model's performance in terms of fidelity and controllability. This objective evaluation is then supported by a user study conducted as a subjective evaluation.

[3] This assumption is a compromise as the melody can possibly be misdetected (*e.g.* melodic bassoon or cello). Further improvements may be implemented with track role identification [Guo *et al.*, 2019].

4.1 Baseline Models

We compare METEOR with two open-source and state-of-the-art style transfer models and adapt them as multi-track re-orchestration models:

- FIGARO [von Rütte *et al.*, 2023]: This multi-track style transfer model can directly perform the re-orchestration task. For the evaluation, we focus on the proposed "note density" controls, which corresponds to the rhythmic intensity in our work.
- AccoMontage-band [Zhao *et al.*, 2024]: This model is originally designed to take a lead sheet as input and generate a multi-track pop band arrangement. We adapt this model to evaluate its performance as a re-orchestration system. To this end, we first pre-process a multi-track input its lead sheet representation *i.e.* we extract the melody as the skyline stream and the chords using the Chorder package[4]. This extracted lead sheet is then used as the input for the model which generates the re-orchestration of the initial input.

We also consider a multi-track extension of MuseMorphose [Wu and Yang, 2023b], initially developed for piano textural style transfer, in which the original REMI tokenization is replaced with a REMI+ tokenization.

For the objective metrics, we compare these baselines with two versions of our model: "METEOR without inference guidance", which includes bar- and track-level controllability but with ablated melody constraints (Section 3.3), and "METEOR" which includes these melody constraints.

4.2 Objective Metrics

We first consider objective metrics to evaluate the full piece fidelity with respect to the reference piece, both overall and specifically for the melody. We also consider a metric to evaluate the instrument realisticness in terms of pitch distribution.

- *Overall fidelity* – Following [von Rütte *et al.*, 2023], we consider the overall fidelity as the chroma similarity between the original piece and the generation, defined as the average of bar-wise cosine similarities between bar-wise chroma vectors.
- *Melodic fidelity* – For a piece, let $X_{b,\text{mel}}$ the token sequence representing the melody in bar b. For a track t in the generation, let $X_{b,t}$ the token sequence of one track t at this bar b. We consider the Levenstein edit distance between two sequences $d(\cdot,\cdot)$ and normalize it so that $|d(X_1, X_2)| \leq 1$ for X_1 and X_2 two sequences. We define the melodic fidelity of a track t at a bar b as $d(X_{b,\text{mel}}, X_{b,t})$. By taking the minimum of these distances among the tracks, we aim at selecting the track which is playing the melody within a bar. Therefore, the smaller the distance, the greater the melodic fidelity. Namely, we define the melodic fidelity φ_b at a bar b as:
$$\varphi_b = 1 - \min_{t \in \text{tracks}} d(X_{b,\text{mel}}, X_{b,t})$$
Finally, we define the melodic fidelity $\varphi_{\text{mel}} \in [0,1]$ of a full multi-track generation of N bars as the average of these bar-wise fidelities: $\varphi_{\text{mel}} = \frac{1}{N}\sum_{b=1}^{N} \varphi_b$
- *Pitch distribution similarity per instrument* – To evaluate the re-orchestration instrumental realisticness, we compare the distribution of pitches per instruments between a generated content and a reference dataset. Let P_i (resp. Q_i) the distribution of pitches played by the instrument i in a reference dataset[5] (resp. in the generated music). We consider that i is among the I available instruments. For JSD$(\cdot||\cdot)$ the Jensen–Shannon divergence, we define the instrument pitch distribution similarity $\rho \in [0,1]$:
$$\rho = \frac{1}{I}\sum_{i=1}^{T} \left(1 - \text{JSD}(D_i||Q_i)\right)$$

Regarding textural controllability, we then consider bar-level metrics for polyphonicity and rhythmicity and bar- and track-level metrics for average pitch and pitch diversity.

- *Bar-controllability* – Polyphonicity and rhythmicity are evaluated at the bar level by including all tracks. Following [Wu and Yang, 2023b], we consider the Spearman correlation between the user-specified polyphonicity or rhythmicity class and the class computed from the model generations given the user inputs.
- *Track-controllability* – Average pitch and pitch diversity are also evaluated with a Spearman correlation between the user input and the class computed from the generation, for each track and each bar.

For this evaluation, each model generates 20 samples of 8 bars each, with the reference pieces and the control signals randomly selected and the instruments chosen by the models.

4.3 Subjective Metrics

Following these objective metrics, we conduct a user study to compare METEOR with the two baseline models. We evaluate the quality of the generations on the task of re-orchestration (multi-track to multi-track) and lead sheet orchestration (lead sheet to multi-track). For both tasks, participants listen to a 8-bar long reference (multi-track piece or lead sheet) and samples generated by the 3 models (Section 4.1). For the first task, they are asked to rate the generation contents on a 6-point Likert scale from 0 (very low) to 5 (very high) based on the following criteria and guidelines:

- **Overall musicality**: how enjoyable is the music?
- **Naturalness of the generation**: to what degree does the piece meet your expectations for musical plausibility?
- **Textural fidelity with the reference**: how does the extract reflect the reference "mood" (calmness, energy...)?
- **Convincing use of instruments**: how well do the instruments blend together within the overall arrangement?
- **Content coherency with the reference**: how much do you recognize the reference by listening to the sample?

The same aspects are evaluated for the lead sheet orchestration task, without "textural fidelity" and with a "creativity" criterion (how inventive while being still pleasant to hear, is

[4] https://github.com/joshuachang2311/chorder

[5] This reference dataset includes the SymphonyNet dataset and an equal number of pieces from the LakhMIDI dataset.

	Model		Overall fidelity ↑	Melodic fidelity ↑	Instr. pitch similarity ↑	Bar-controllability ↑		Track-controllability ↑	
						Rhyth.	Polyph.	Pitch diver.	Avg. pitch
1	FIGARO		.735 ±.24	.271 ±.08	.617 ±.14	.867	–	–	–
2	AccoMontage-band		.756 ±.10	.338* ±.09	.583 ±.17	–	–	–	–
3	Multi-track MuseMorphose		**.932** ±.10	.527 ±.13	.696 ±.18	.941	.936	–	–
4	METEOR (w/o inference guidance)		.918 ±.11	.491 ±.16	.755 ±.13	**.972**	**.951**	**.929**	**.926**
5	METEOR		.927 ±.10	**.632** ±.18	**.780** ±.12	.950	.932	.897	.821
6	Multi-track MuseMorphose	*Flute-oboe duet*	.888	.479	–	.919	.710	–	–
7		*Woodwind quintet*	.932	.519	–	.956	.875	–	–
8		*Classical orchestra*†	.947	.511	–	.921	.676	–	–
9	METEOR (w/o infer. guidance)	*Flute-oboe duet*	.837	.457	–	.967	.782	.949	.873
10		*Woodwind quintet*	.903	.519	–	.971	.860	.961	.853
11		*Classical orchestra*†	.917	.493	–	.975	.898	.958	.798
12	METEOR	*Flute-oboe duet*	.837	.650	–	.936	.715	.786	.711
13		*Woodwind quintet*	.909	.651	–	.927	.862	.889	.875
14		*Classical orchestra*†	.912	.720	–	.953	.867	.909	.780

Table 2: Objective metrics for the re-orchestration task, with automatic choice and user-defined ensembles. (*) we evaluate the generated content only, without the inserted melodic track. (†) Classical orchestra includes 11 instruments (4 woodwinds, 2 brasses, timpani, 4 strings).

the audio extract). We let the model choose the melodic track automatically (or randomly for FIGARO and AccoMontage-band) in the re-orchestration task. Instead, for lead sheet orchestration, we insert *a posteriori* the melodic track played by a synthesizer, following the method of AccoMontage-band. This ensures a fair comparison of all models in terms of melody perception by the listener, allowing for a focused comparison between the generated accompaniments.

The survey consists of 6 pieces for the re-orchestration task and 4 for lead sheet orchestration, chosen to ensure diversity. For each piece, the instrumentation is fixed for all models, including different cases: where the number of target instruments is smaller or greater than the source instruments. Each model generates four re-orchestrations for each 6 pieces. Participants are randomly assigned to one of the four groups, with each group evaluating a different set of samples. A total of 24 participants for the re-orchestration task, and 13 for lead sheet orchestration have answered the survey. They have various musical backgrounds, from individuals with no musical experience (15%) to professional musicians (8%), with a majority of amateur (46%) to intermediate musicians (31%).

4.4 Results

Objective evaluation Quantitative metrics are summarized in Table 2 (rows 1–5). MuseMorphose and the two versions of METEOR manage to outperform baseline models in all metrics. With FIGARO, they outperform AccoMontage-band in pitch distribution fidelity, as both are trained on orchestral instruments while AccoMontage-band is trained on band instruments. MuseMorphose and the two METEORs achieve comparable overall fidelities and adding melodic constraints naturally leads to an improvement in melodic fidelity. Though, the latter does not reach a perfect score, as *inference guidance* does not prevent the melodic instrument from adding extra notes beyond the exact melody, a phenomenon which can be found in orchestral music, often referred to as "decorative melody" [Le *et al.*, 2022]. This increase in melodic fidelity results in a drop in controllability metrics compared to METEOR w/o melody. This may result from using independent control methods, either latent or token-based, for beat-, bar- and track-level attributes. The compatibility between latent space-based or token-based controls remains unexplored and could be investigated in future research to improve the understanding of controllable models.

Instrumentation impact We further study the impact of the chosen instrumentation on our models' performances (Table 2, rows 6–14). We select three musical ensembles: woodwind duet, quintet, and classical orchestra, assigning the melody to the flute in each case. For METEORs, increasing the number of instruments helps the model maintaining better fidelity to the reference piece and improves bar-wise attributes. With more instruments, the model has a larger instrumental flexibility and a broader range of options to assign each track a part that aligns with the control signals. Moreover, all the models demonstrate better bar-wise polyphonicity controllability when the instrumentation is chosen automatically (rows 3–5) compared to each user-defined ensembles. In other words, they manage to effectively select the most suitable ensemble to match the requested polyphonicity.

Melodic instrument range playability We then study the playability of generations in terms of physical constraints of the melodic instrument. Unlike generic instruments such as synthesizers [Luo *et al.*, 2024; Zhao *et al.*, 2024], orchestral instruments are limited in their range and usually play in a specific register [Rimsky-Korsakov, 1964].

To evaluate such range playability, we generate five extracts from the same original reference without textural control attributes and assign the melody to an instrument. Based on our pitch class-based tokenization, we let the model infer the `Octave` tokens of the melody notes, while the other components (pitch class and duration) are enforced. As presented in Table 3, the model manages to generate instrumental parts which match with their usual register, with still a

Melodic instrument	Average note in instr. range (register bounds)	Average pitch in generations	Out of range generated notes
Flute	F5 (B3-C7)	D5	0.0%
Bassoon	E3 (B♭1-B4)	B♭2	4.8%
Trumpet	A4 (F#3-C6)	G4	4.3%
Violin	A5 (G3-B7)	C5	1.6%
Cello	B♭3 (C2-A5)	F3	3.8%

Table 3: Average pitch of melodic instruments with octave inference in the generated music compared to their real instrumental range.

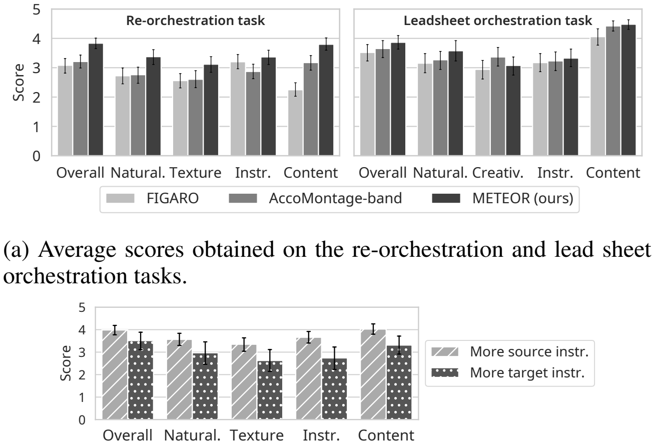
limited amount of out of range notes. However, while the difference between the generated average pitch and its middle-range note is below a fourth for woodwinds and trumpet, the average generated pitch for the cello and the violin are much lower (*e.g.* a sixth lower than the midpoint note of the violin's full register). Violin parts and, more generally, string parts, are indeed typically written below the extreme high register of the instrument [Adler and Hesterman, 1989, p. 52].

Subjective evaluation The results from our user study on each task and criterion are presented in Figure 5.

Re-orchestration task – METEOR outperforms in four of the five criteria on average (Figure 5a, left). In particular, it holds significant advantage over the two other models on the overall musicality and naturalness (t-test: $p < .01$ for both). Further analysis highlights notable insights on other criteria.

- Texture fidelity. METEOR achieves significantly better results than the baseline models, in particular, compared to AccoMontage-band ($p < .01$). This may be attributed to the lead sheet input which simplifies the original piece by reducing it to melody and chords, losing crucial textural characteristics and making it challenging for the model to generate a similar musical texture.
- Instrumental use. METEOR and FIGARO show comparable performances and both outperform AccoMontage-band. Given that the ensembles have been set to standard orchestral instruments, this shows that AccoMontage-band, which was trained with pop band instruments, can weakly adapt to unseen instruments. However, when comparing scenarios with varying numbers of target instruments relative to source instruments (Figure 5b), METEOR performs better on instrumentation reduction and weaker when the target ensemble is larger than the reference on all criteria. This may be attributed to the need for generating longer sequences for these larger ensembles, highlighting a potential limitation in the model's ability to capture long-term dependencies.
- Content coherency. FIGARO has an average score significantly lower than METEOR and AccoMontage-band ($p < 1\text{e}{-6}$). As noted in the original study, FIGARO often fails to preserve the melody, highlighting that content coherency is strongly influenced by the retention of the melodic line. This effect is supported by the observation that its content fidelity is more comparable to other models in the lead sheet orchestration task, where the melody is inserted unchanged *a posteriori*.

Lead sheet orchestration task – Across all metrics, METEOR achieves performance ranging from comparable to better than the other models (Figure 5, right). In particular, while AccoMontage-band has been specifically trained on this task, it only outperforms METEOR on average on the creativity criterion. This zero-shot learning ability highlights METEOR's versatility in performing tasks closely related to orchestration with comparable performances with state-of-the-art models.



(a) Average scores obtained on the re-orchestration and lead sheet orchestration tasks.

(b) Impact of the number of source and target instruments on METEOR's re-orchestrations.

Figure 5: Subjective evaluation results. A 6-point Likert scale ranging from 0 (very low preference) to 5 (very high preference) is used.

5 Conclusion & Future Directions

In this study, we present METEOR, a model for texture-controllable multi-track style transfer with a focus on melodic fidelity specifically trained for a task of re-orchestration. The model performs this task through token constraints at a bar- and track-level, with inference guidance for melodic fidelity. On a re-orchestration task, METEOR outperforms multi-track style transfer models on subjective and objective evaluations. We show that our model can be adapted into a lead sheet orchestrator and is comparable to a model trained for this task.

Limitations & future directions Our study focuses on bar- and track-level controllability, excluding piece-level controllability [Lu *et al.*, 2023], which consequently disregards high-level structures, such as repeated musical phrases, which are fundamental in music composition [Shih *et al.*, 2023]. Future work towards multi-level multi-track style transfer may include a model able to perform style transfer at these three levels. Such a controllable model could be integrated into an orchestrator's workflow as a co-creative tool, allowing both broad orchestration drafts and detailed refinements. From a musical perspective, although METEOR succeeds in ensuring that melodic instruments fit their range constraints, their technical playability (*e.g.* convenient fingerings, breath considerations, logical articulations) have not been thoroughly studied and are systematically overlooked in music generation studies. Ensuring playability in relation to instrumental constraints, timbre effects, and instrument groupings [Goodchild and McAdams, 2018] would be a significant advancement towards automatic humanly playable orchestration.

Ethical Statement

Our work focuses on automatic music generation, raising potential concerns about the ownership of the generated content. Though, our study emphasizes human-machine co-creativity, particularly by enabling fine-grained control over the textural and instrumental properties of the generated content. These controls are still limited regarding the style of music due to the choice of the training dataset which inherently exhibits a bias towards a Western instrumentation and a Western tonal style of music.

Moreover, our study's evaluation relies on a survey presented as a user listening test. In this survey, no personal information was retrieved and the data was not used for other purposes than the current study.

Finally, our study is based on a deep learning approach, which may have an energy consumption impact due to the computational power required for model development, training, and evaluation. Although we did not precisely monitor any hardware power consumption during this study, an approximation[6] [Lacoste *et al.*, 2019] of a training of our model, limited to a duration of one week on our hardware, reaches a consumption of 7 kgCO$_2$eq.

Acknowledgements

The work of Dinh-Viet-Toan Le was supported by grant ANR-20-THIA-0014 program "AI_PhD@Lille" and a bourse MERMOZ from Région Hauts-de-France. The work of Yi-Hsuan Yang was supported by a grant from the National Science and Technology Council of Taiwan (NSTC 112-2222-E002-005-MY2) and a grant from Google Asia Pacific. The authors would like to thank the anonymous reviewers for their valuable feedback which helped improve the article, Ken Déguernel and Chih-Pin Tan for their comments on earlier versions of this manuscript, and Louis Bigo and Mikaela Keller for fruitful discussions.

References

[Adler and Hesterman, 1989] Samuel Adler and Peter Hesterman. *The study of orchestration*, volume 2. WW Norton New York, NY, 1989.

[Benward, 2018] Bruce Benward. *Music in theory and practice*. McGraw Hill Higher Education, 2018.

[Cacavas, 1975] John Cacavas. *Music arranging and orchestration*. Alfred Music, 1975.

[Cífka *et al.*, 2020] Ondřej Cífka, Umut Şimşekli, and Gaël Richard. Groove2groove: One-shot music style transfer with supervision from synthetic data. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2638–2650, 2020.

[Dai *et al.*, 2018] Shuqi Dai, Zheng Zhang, and Gus G Xia. Music style transfer: A position paper. *arXiv preprint arXiv:1803.06841*, 2018.

[Dong *et al.*, 2023] Hao-Wen Dong, Ke Chen, Shlomo Dubnov, Julian McAuley, and Taylor Berg-Kirkpatrick. Multitrack music transformer. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Ens and Pasquier, 2020] Jeff Ens and Philippe Pasquier. Mmm: Exploring conditional multi-track music generation with the transformer. *arXiv preprint arXiv:2008.06048*, 2020.

[Goodchild and McAdams, 2018] Meghan Goodchild and Stephen McAdams. Perceptual Processes in Orchestration. In *The Oxford Handbook of Timbre*. Oxford University Press, 2018.

[Guo *et al.*, 2019] Rui Guo, Dorien Herremans, and Thor Magnusson. Midi miner–a python library for tonal tension and track classification. *arXiv preprint arXiv:1910.02049*, 2019.

[Huang and Yang, 2020] Yu-Siang Huang and Yi-Hsuan Yang. Pop Music Transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA, 2020. Association for Computing Machinery.

[Huron, 1989] David Huron. Characterizing musical textures. In *International Computer Music Conference (ICMC 1989)*, pages 131–134, 1989.

[Lacoste *et al.*, 2019] Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*, 2019.

[Le *et al.*, 2022] Dinh-Viet-Toan Le, Mathieu Giraud, Florence Levé, and Francesco Maccarini. A corpus describing orchestral texture in first movements of classical and early-romantic symphonies. In *Proceedings of the 9th International Conference on Digital Libraries for Musicology*, DLfM '22, page 27–35, New York, NY, USA, 2022. Association for Computing Machinery.

[Li *et al.*, 2023] Yuqiang Li, Shengchen Li, and George Fazekas. Pitch Class and Octave-Based Pitch Embedding Training Strategies for Symbolic Music Generation. In *Proceedings of the 16th International Symposium on Computer Music Multidisciplinary Research (CMMR)*, page 86–97, Tokyo, Japan, November 2023. Zenodo.

[Liu *et al.*, 2022] Jiafeng Liu, Yuanliang Dong, Zehua Cheng, Xinran Zhang, Xiaobing Li, Feng Yu, and Maosong Sun. Symphony generation with permutation invariant language model. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference*, pages 551–558. ISMIR, December 2022.

[Lu *et al.*, 2023] Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. MuseCoco: Generating symbolic music from text. *arXiv preprint arXiv:2306.00110*, 2023.

[Luo *et al.*, 2024] Jing Luo, Xinyu Yang, and Dorien Herremans. BandControlNet: Parallel transformers-based

[6] https://mlco2.github.io/impact

steerable popular music generation with fine-grained spatiotemporal features. *arXiv preprint arXiv:2407.10462*, 2024.

[Lv *et al.*, 2023] 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.

[Ren *et al.*, 2020] Yi Ren, Jinzheng He, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Popmag: Pop music accompaniment generation. In *Proceedings of the 28th ACM international conference on multimedia*, pages 1198–1206, 2020.

[Rimsky-Korsakov, 1964] Nikolay Rimsky-Korsakov. *Principles of orchestration*. Courier Corporation, 1964.

[Shih *et al.*, 2023] Yi-Jen Shih, Shih-Lun Wu, Frank Zalkow, Meinard Müller, and Yi-Hsuan Yang. Theme transformer: Symbolic music generation with theme-conditioned transformer. *IEEE Transactions on Multimedia*, 25:3495–3508, 2023.

[Shu *et al.*, 2024] Yangyang Shu, Haiming Xu, Ziqin Zhou, Anton van den Hengel, and Lingqiao Liu. Musebarcontrol: Enhancing fine-grained control in symbolic music generation through pre-training and counterfactual loss. *arXiv preprint arXiv:2407.04331*, 2024.

[Stefani, 1987] Gino Stefani. Melody: A popular perspective. *Popular Music*, 6(1):21–35, 1987.

[von Rütte *et al.*, 2023] Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. FIGARO: Controllable music generation using learned and expert features. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[Wu and Yang, 2023a] Shih-Lun Wu and Yi-Hsuan Yang. Compose & Embellish: Well-structured piano performance generation via a two-stage approach. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Wu and Yang, 2023b] Shih-Lun Wu and Yi-Hsuan Yang. MuseMorphose: Full-song and fine-grained piano music style transfer with one transformer VAE. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1953–1967, 2023.

[Young and Roens, 2022] Gregory Young and Steve Roens. Form, texture, style, and harmonic language. In *Insights into music composition*, pages 46–55. Routledge, 2022.

[Zhao *et al.*, 2024] Jingwei Zhao, Gus Xia, Ziyu Wang, and Ye Wang. Structured multi-track accompaniment arrangement via style prior modelling. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.