TASTE: TAXONOMY-AWARE SYMBOLIC PIANO GENERATION FROM TEXTUAL DESCRIPTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-conditioned symbolic music generation has attracted growing interest with the development of generative models. Yet two questions remain underexplored: what aspects of music the conditioning text should describe, and how the corresponding data should be collected and trained to improve controllability and text–music alignment. We address these challenges with TASTE, a taxonomy-aware framework for symbolic music generation with a main focus on piano pieces. First, we introduce a taxonomy of musical elements that specifies how a piece can be captured through textual description in a musically grounded way. The taxonomy comprises a set of metadata tags (genre, emotion, section plan, etc.) together with a free-form description of music imagery. This structure decomposes textual input to match the generation goal more coherently and precisely. Second, we build a data pipeline for collecting, annotating, and retrieving public symbolic piano resources, yielding a corpus of 432K MIDI files labeled at multiple levels of granularity. Finally, we design a staged training procedure that leverages this multi-granularity data across two model families: an autoregressive model and a non-autoregressive diffusion model. Both objective and subjective evaluations show that our models achieve superior musical-element control, text-description alignment, and overall music quality. We will release the full pipeline, all annotated datasets, and the model checkpoints.

1 INTRODUCTION

Symbolic music generation offers a structured and editable complement to audio generation, well suited to composition, arrangement, and human-in-the-loop creation. Driven by recent advances in generative models, a growing number of systems now generate symbolic music directly from textual description, letting users steer composition with language. Yet before asking how to generate music from text, we should ask a more basic question: **what makes a good description of music?**

Existing text-conditioned systems seldom confront this question directly and tend toward one of two extremes. Some (Bhandari et al., 2025; Wu et al., 2026) condition on free-form captions, which are expressive but whose musical content is uncontrolled and hard to quantify. Others (Lu et al., 2023; Wang et al., 2025) restrict control to a fixed set of attributes such as key, tempo, instrument, and genre, which are precise but cover only a narrow portion of what a user wish to express.

These two extremes are symptoms of a deeper issue: a textual description of music conveys two kinds of information, and each design captures only one. The first is objective and well-defined musical attributes such as key, genre, emotion, and section plans of structure, which rest on a shared vocabulary and admit discrete labels. The second is subjective and open-ended music imagery that evokes a scene, a mood, or the symbolic impression conjured by music pieces. A real-world music description from users usually interleaves both. Consider "*an anxious, arpeggiated passage with flexible timing, evoking a gathering storm that breaks into release*": it combines discrete attributes with an imagined scene and a temporal progression, reducing neither to a few metadata fields nor to an undifferentiated caption. We call the resulting mismatch between the descriptions current systems can handle and the intentions users actually hold **the description-intention gap**.

To close this gap, we should call for an explicit taxonomy of musical description that unifies both kinds of information while keeping them distinct. Such a taxonomy serves two purposes at once. For users, it offers a principled guide for expressing intent with sufficient scope and controllability.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

For models, it provides a structured training signal, so that generation will more faithfully reflect what a description specifies. Concretely, it separates a set of structured attribute tags from a free-form imagery description, allowing each to be annotated, learned, and evaluated on its own terms. Besides, we should also admit that sometimes imagery inevitably carries traces of some attributes, such as emotion, which is harmless as long as it also contains the correct attribute tags,

Answering the first question reshapes the second: **how can we collect data annotated in our designed taxonomy?** Existing datasets rarely provide such labels. MidiCaps (Melechovský et al., 2024) builds captions from metadata and extracted features, which favors properties easy to read off a MIDI file, leaving music imagery uncovered; large-scale MIDI collections like (Bradshaw & Colton, 2025) offer abundant data but none of the taxonomy labels we need; and piano-specific labeled datasets (Bang et al., 2024) remain small. Closing this gap calls not only for adding a few fields to an existing dataset, but also for a data pipeline organized around the taxonomy to collect public symbolic piano data, annotate each piece by matching each condition to suitable evidence, and propagate labels from reliably paired data to weakly labeled data at scale. This yields multiple granularities so that training will not leave uncertain supervision dominate.

To address these challenges by proposing **TASTE**, a taxonomy-aware framework that connects musical intent, data construction, and model training for symbolic piano generation. The taxonomy includes **tags** for objective, finite-vocabulary attributes, a **section plan** for temporal form, and an **imagery description** for the subjective scene or impression a piece evokes. The pipeline and training recipe answer how such data can be collected and used: we annotate each taxonomy channel from the evidence best suited to it, propagate labels from reliably paired data, and yield a corpus of 432K MIDI files labeled at multiple granularities. Our contributions are threefold:

1. We propose an intent-oriented taxonomy of music description that separates objective tags, section plans, and music imagery. We tie each channel to its annotation evidence and evaluation to narrow the description-intention gap.
2. We build a taxonomy-driven data pipeline to match each condition to a suitable source of evidence and propagate labels across tiers, yielding a three-tier piano corpus of 10K, 51K, 371K pieces at multi-granular labels.
3. We design a staged training recipe, instantiate it in both autoregressive and latent diffusion models. On both objective and subjective evaluations, TASTE achieves the superior performance on attribute agreement, full-prompt alignment, temporal structure, and musicality to other baselines. Music samples and other details can be found at https://taste-piano.github.io/.

2 Related Work

Text-conditioned Symbolic Generation. Unconditional piano models generate expressive performances without textual control (Wu & Yang, 2023; Bradshaw et al., 2025). Attribute-conditioned generators support emotion (Hung et al., 2021; Huang et al., 2024), composer style (Yao & Chen, 2025; Jiang et al., 2026), genre (Lehmkuhl et al., 2025), and composition identity (Wang et al., 2025). FIGARO combines learned and expert features, while MuseCoco maps text to fixed attributes (von Rütte et al., 2023; Lu et al., 2023). MidiCaps and MetaScore construct captions from extracted features and community metadata, respectively (Melechovský et al., 2024; Xu et al., 2025). Using MidiCaps, text2midi introduces caption cross-attention, while MIDI-LLM adapts a pretrained language model for text-to-MIDI generation (Bhandari et al., 2025; Wu et al., 2026).

Descriptions of music and piano performance. MusicCaps and Song Describer provide expert and listener-written captions (Agostinelli et al., 2023; Manco et al., 2023). Piano-focused work pairs performances with text, aligns audio, MIDI, and language for retrieval, and supports text-guided rendering (Bang et al., 2024; 2025; Zhang et al., 2025a). TASTE instead organizes descriptions into separately supervised tags, section plans, and imagery for symbolic generation.

Weak supervision and multimodal distillation. Confident learning detects likely label errors (Northcutt et al., 2021), language-aligned audio and symbolic representations support probes and retrieval (Wu et al., 2023; 2025), and section fusion identifies recurring sections (Salamon et al., 2021). TASTE coordinates these methods through its taxonomy, matching fields to suitable evidence and selectively revising earlier labels under restricted rules.

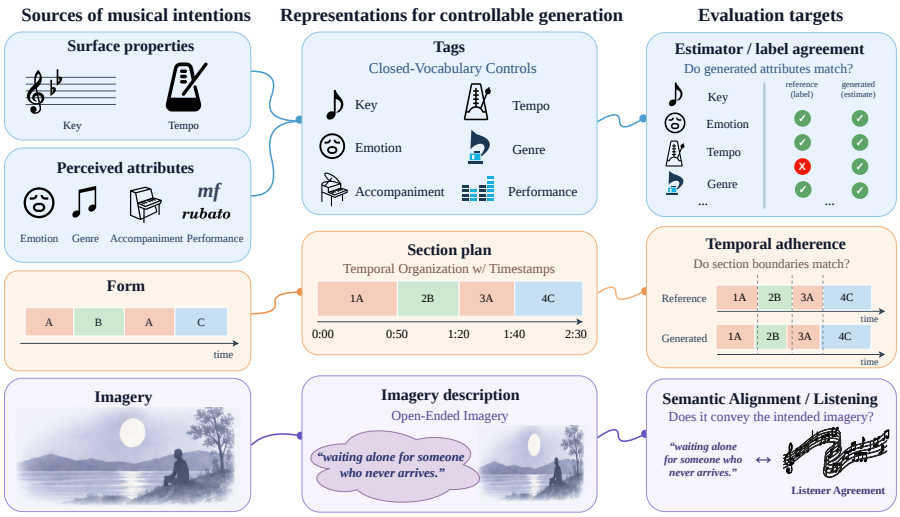


Figure 1: The TASTE taxonomy. Sources of musical intention (left) map to three condition channels (middle), each with its own evaluation target (right).

3 Taxonomy of Music Description

How a description is formatted has consequences on two sides: it bounds what a user can express, and determines what a model can be trained to control. A useful taxonomy should reflect how the aspects of a music description genuinely differ: in the vocabulary they draw on, in whether they hold globally or over time, and in the evidence their labels can rely on. These yield three channels from musical intentions in the left and middle of Figure 1: tags, section plan, and imagery description.

3.1 From musical intentions to taxonomy channels

Tags. Tags cover properties that fit a closed vocabulary and hold for the piano piece globally. They include surface properties namely key and tempo, and perceived attributes namely emotion, genre, accompaniment, and performance style. Among these properties, key and tempo are well handled by algorithmic estimation, while emotion or performance style requires perceptual interpretation. Performance style crosses two established dimensions of expressive performance (Gabrielsson, 2003; Widmer & Goebl, 2004) into four classes: steady, flexible, connected, or detached. Because each field has a predefined label space, extracted attributes can be compared field by field.

Section plan. Musical form is objective as one can verify from the notes whether a section returns. In addition to global tags, section plan represents as an ordered sequence of section identities with durations. Letters A–H identify sections within a piece, and a repeated letter marks recurrence. This ties form control to boundary and self-similarity analysis (Foote, 2000; Salamon et al., 2021), so that boundary placement and recurrence can be evaluated separately.

Imagery description. The remaining intentions resist any finite label set (except partially from emotion). Human beings may ask for music that conjures a scene, feeling, or symbolic impression from visual imagery (Juslin & Västfjäll, 2008). An imagery description expresses such that the music may evoke, e.g., "waiting alone for someone who never arrives." It differs from tags in kind, not merely in vocabulary but also in its one-to-many mapping, as many descriptions may suit one piece and none is uniquely correct. Yet it is not unconstrained, for instance, a storm scene normally is open to anxious or angry emotions, but clearly not calm. The imagery channel is therefore a language condition whose target is perceptual compatibility rather than a single truth.

3.2 Connecting representation, annotation, and evaluation

The separation matters beyond the model interface, because each channel calls for different supervision and evaluation. Tags can be annotated and audited within finite vocabularies, section

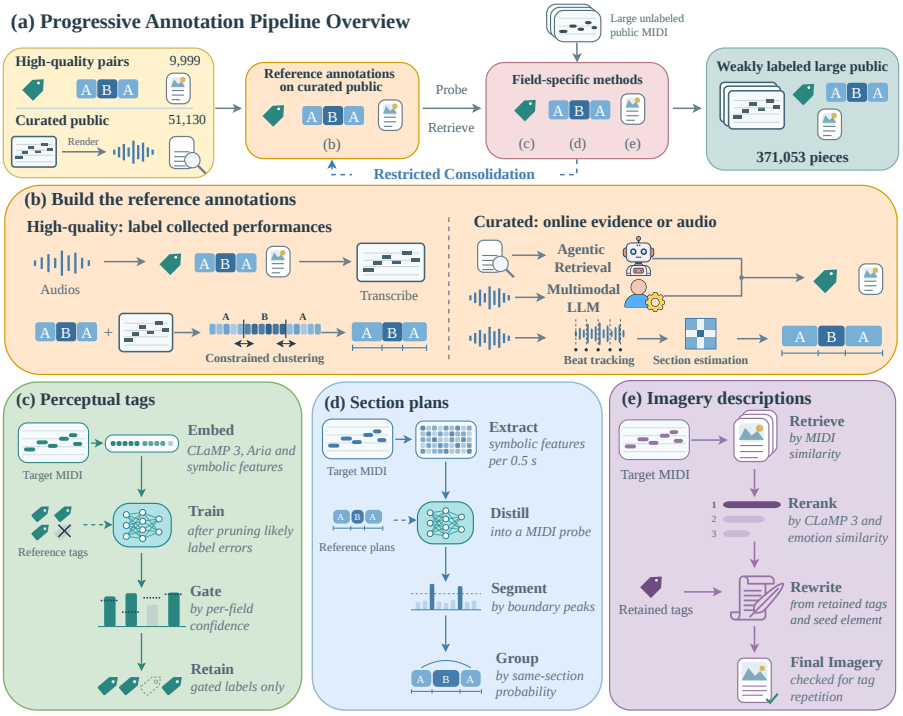


Figure 2: Progressive annotation across three tiers. Forward arrows gain scale at the cost of label reliability, while the dotted backward arrow is the restricted consolidation pass of Section 4.3.

plans require temporal alignment, and imagery descriptions need contextual or perceptual grounding. Evaluation follows the same division instead of a single caption-similarity score (Figure 1, right): estimators and critics read back generated attributes, realized structure is compared with the plan, and listeners judge imagery, with text–music similarity as a joint proxy for tags and imagery.

4 Multi-Tier Annotation Pipeline

The annotation pipeline addresses how to collect data labeled in our taxonomy. We organizes 432K MIDI files into three tiers that trade label reliability for scale (Figure 2 and Table 1). We first build reference annotations in the high-quality and curated tiers by matching each taxonomy channel to suitable evidence, then propagate them to the large public tier through field-specific weak supervision, and finally revise selected earlier labels in a restricted backward pass.

4.1 Reference annotations in the high-quality and curated tiers

We collect high-quality tier from public AI generated piano performances, and hire people to annotate four perceptual tags, imagery description and section sequence. We then transcribe the audio to MIDI with Transkun (Yan & Duan, 2024). Of 9,999 collected pairs, 9,949 have an imagery description, and we hold out 100 of them as evaluation prompts.

In the curated tier, we collect multiple public symbolic music dataset in total of 51,130 pieces and infer their labels through one of two annotation routes. For 40,652 pieces, a language-model agent searches the web for evidence about the work using identity metadata such as composer, title, and movement, annotating each field the evidence supports and abstaining otherwise. The remaining 10,478 pieces lack identity metadata or receive at most one field from the agent. For them, a multimodal LLM describes the rendered audio, and a language model normalizes this description into tags and an imagery description. On both routes, imagery describes a plausible association without restating the tags, and a genre given by the source dataset takes priority (Appendix B).

We then derive key, tempo, and timed section plans for both tiers. Key is estimated from MIDI pitch-class profiles (Temperley, 2006) and tempo from beats tracked in audio with Beat This! (Foscarin et al., 2024) (Appendix B.3). A timed plan needs boundaries as well as identities that en-

Table 1: Annotation tiers and the label evidence.

Tier	Pieces	Evidence source
High-quality paired	9,999	Paired tags, imagery, section labels; estimated key, tempo, timing
Curated public	51,130	Web evidence or audio descriptions
Large public	371,053	Gated tag and section probes, imagery retrieval

Table 2: Tag precision and tier coverage.

Field	**Precision**	**Coverage**	**Binding rule**
Emotion	0.600	86.6%	precision
Genre	0.936	100.0%	precision
Accompaniment	0.538	75.0%	coverage floor
Performance	0.588	75.0%	coverage floor

code recurrence. High-quality pairs supply the section sequence but no timestamps, so we place boundaries by constrained clustering over symbolic and Aria features (Bradshaw et al., 2025), alternating dynamic-programming boundary placement with centroid re-estimation as in Lloyd's algorithm (Lloyd, 1982). For curated performances, initial plans come from the tracked beats and deep-embedding section fusion (Salamon et al., 2021).

4.2 Scaling annotations through field-specific weak supervision

Running web search or audio annotation for each of the 371,053 pieces in the large tier would be costly. We instead propagate the reference annotations with a method suited to each channel: classification for perceptual tags, temporal prediction for section plans, and retrieval with rewriting for imagery descriptions. Key and tempo are estimated directly from MIDI (Appendix B.3).

Noise-aware tag probes. We predict perceptual tags with probes on CLaMP 3 (Wu et al., 2025) and Aria (Bradshaw et al., 2025) embeddings and symbolic statistics of timing, dynamics, polyphony, and pedal use. Class-balanced linear classifiers predict emotion, genre, and accompaniment, and two binary classifiers for timing and articulation jointly determine performance style. We develop the probes on the 61,129 pieces of the two smaller tiers: we select features and calibrate per-field confidence thresholds on the validation set, prune likely label errors with a confident-learning rule on out-of-fold predictions (Northcutt et al., 2021), and refit the production probes on the pruned full set (Appendix B.4). Where source metadata provide a genre, we map and keep it instead of the probe prediction.

A per-field confidence gate trades agreement with the reference annotations against coverage. For each field, we take the threshold that reaches 0.60 validation precision and relax it only as far as needed to label at least 75% of the large tier. Predictions below the gate are omitted from the training conditions. Emotion and genre meet the precision target, whereas accompaniment and performance style are limited by the coverage floor and keep less precise labels (Table 2).

Section-plan distillation. We distill the timed section plans of the two smaller tiers into the *structure probe*, a 1.8M-parameter transformer over frame-level symbolic features. A boundary head segments a performance, and a grouping head gives recurring segments the same section identity. This MIDI-native annotator needs neither audio nor metadata and thus covers the whole large tier.

Imagery retrieval and rewriting. For each target performance, we first retrieve candidate descriptions from the smaller tiers by Aria embedding similarity (Bradshaw et al., 2025). A ridge regression projects the target's Aria embedding into the CLaMP 3 text space (Wu et al., 2025), and we rerank the candidates by their similarity to this projection and by agreement between the two pieces' predicted emotion distributions. A language model then adapts the selected description to the target, keeping the mood and energy conveyed by the retrieved description and the target's predicted tags while changing the concrete setting (Appendix B.6).

4.3 Restricted backward consolidation

The trained tag and structure probes can also revise the smaller tiers, filling gaps and correcting labels whose evidence is indirect, and we apply them conservatively. A gated prediction may replace any perceptual tag on audio-annotated pieces, but on evidence-annotated pieces, whose web evidence rarely covers texture or rendition, it may only fill or replace accompaniment and performance style, with a stricter threshold for replacing performance labels (Appendix B.7). The structure probe

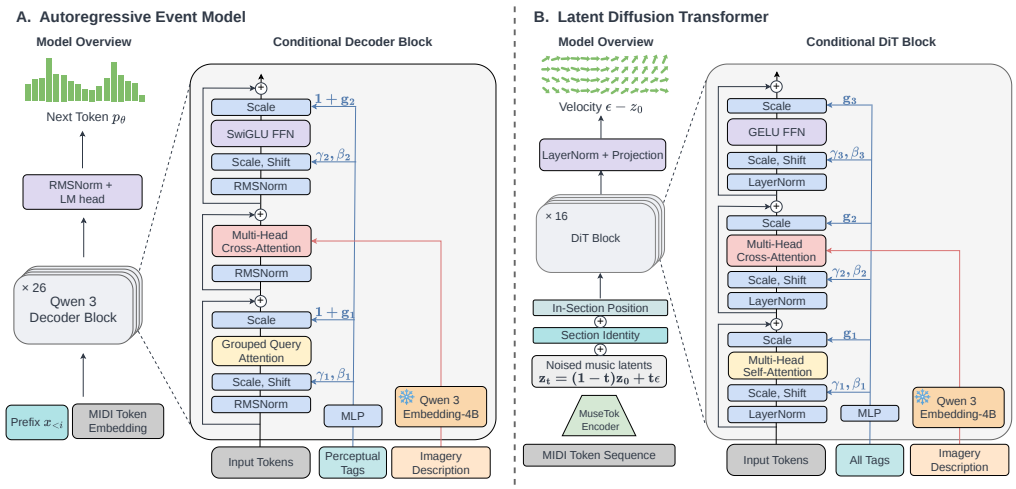


Figure 3: TASTE model families. (A) AR reads key, tempo, and the section plan as a token prefix. (B) DiT receives the section plan as per-position embeddings. In both, summed tag embeddings drive adaptive normalization, and imagery enters via cross-attention to a frozen text encoder.

replaces the curated tier's initial audio-derived plans, but on high-quality pairs it refines only the boundary timestamps and keeps the supplied identities and order. High-quality-tier tags and all imagery descriptions remain unchanged. The backward pass thus aligns labeling conventions across tiers while maintaining the reliability of each field.

5 Models and Staged Training

The taxonomy specifies what information conditions generation, and the data tiers determine how supervision for these conditions enters training. We instantiate this design in two model families (Figure 3), an autoregressive decoder (AR) and a non-autoregressive latent diffusion transformer (DiT), which both keep the three channels separate while adapting each channel to the architecture.

5.1 Piano Representation

Both models build on an Aria-style tokenizer (Bradshaw et al., 2025) with a 25 ms grid (Figure 4 in Appendix C.1). Each note is a combined pitch–velocity token with 12 velocity levels, followed by onset and duration tokens, and pedal events have their own tokens in both models. Window markers every five seconds anchor onsets in absolute time, replacing the bar tokens of metrical tokenizers (Huang & Yang, 2020; Hsiao et al., 2021). In AR, key, tempo, and the section plan form an in-sequence prefix, in which each plan entry gives a section identity and its duration in windows. The active section identity is repeated after every window marker, so the model receives both a prospective plan and a local structural cue. The DiT's codec encodes only the note and pedal events, and the DiT receives key, tempo, and the plan as separate conditions (Section 5.3).

5.2 Autoregressive Model

The AR backbone is a 0.89B-parameter Qwen3-style decoder (Team, 2025) with a 4,096-token context. We pretrain it for five passes over about 817K performances of the pruned Aria-MIDI subset (Bradshaw & Colton, 2025), with section-aligned crops and pitch and tempo augmentation, for 4.1M sequences and 11.44B tokens in total. Because key, tempo, and section plans can be estimated for any MIDI file (Section 4), their tokens are present from pretraining, so the model learns how they relate to musical events before the remaining conditions are introduced.

The four perceptual tags inject to every decoder layer through AdaLN-Zero-style modulation (Peebles & Xie, 2023), with their embeddings summed and mapped to a 512-dimensional code. The

imagery description enters via cross-attention to the token states of a frozen Qwen3-Embedding-4B (Zhang et al., 2025b). Adding these modules gives a 1.235B conditioanl generator (Appendix C).

5.3 Non-autoregressive latent diffusion

The DiT generates in a continuous latent space. A MuseTok-based window VAE (Huang et al., 2026) encodes each one-second window of a performance, with a 0.5 s hop, into a 128-dimensional latent. A 16-layer, 397.6M-parameter diffusion transformer (Peebles & Xie, 2023) is trained on these latents with a rectified-flow objective (Liu et al., 2023): given a music latent z_0, Gaussian noise ϵ, and a timestep t, it receives $z_t = (1-t)z_0 + t\epsilon$ and predicts the velocity $\epsilon - z_0$. Like AR, the DiT is first pretrained on the same corpus with key, tempo, and section plans.

The DiT differs from AR in how key, tempo, and the section plan enter. All six tags, key and tempo included, are embedded, summed, and combined with the timestep embedding into the AdaLN-Zero code that modulates every block (Appendix C.4). The section plan is aligned with the latent sequence: each position receives embeddings of its section identity, position within the current section, and its absolute position (Vaswani et al., 2017). The imagery description enters through cross-attention to the same states as in AR. At inference, we apply classifier-free guidance (Ho & Salimans, 2022) and decode latents with a grammar-constrained MuseTok-based decoder (Appendix C).

6 Evaluation

Following the division of Section 3, we first test whether each taxonomy channel is realized in the generated music, through agreement with the requested tags and descriptions and adherence to section plans. We then examine music quality and listener judgments. Appendix D.5 analyzes the effect of staged training and ablation expereiment with no large weak tier training (pre-training–curated tier–high-quality tier).

6.1 Evaluation setup

Generation protocol and baselines. We use the 100 held-out prompts of the high-quality tier and draw eight samples per prompt from each model. Every prompt contains six tags and an imagery description. AR generates its own section plan, while DiT additionally receives the plan and duration of the reference performance.

The conditional baselines are MuseCoco (Lu et al., 2023), text2midi (Bhandari et al., 2025), MIDI-LLM (Wu et al., 2026), EMO-Disentangler (Huang et al., 2024), and Piano Transformer (Lehmkuhl et al., 2025), each receiving only the controls it supports. The caption-conditioned text2midi and MIDI-LLM receive all tags and the imagery in one caption, with non-piano tracks masked at decoding. Aria (Bradshaw et al., 2025) and Compose & Embellish (Wu & Yang, 2023) serve as unconditional baselines for music quality. Appendix D details the evaluation pipeline and data processing.

Condition agreement. Key and tempo are read back with the annotation estimators. For the perceptual tags, we train logistic-regression critics on CLaMP 3 and Aria embeddings of the 9,849 high-quality training pieces, and scoring the 100 reference performances with the same critics shows how reliably each field can be read back. Full-prompt alignment is the cosine similarity, under CLaMP 3 (Wu et al., 2025) and LAION-CLAP (Wu et al., 2023), between the generated music and a caption that combines the perceptual tags with the imagery description.

Music quality. We compute Fréchet Audio Distance in VGGish space (Kilgour et al., 2019) and Fréchet Music Distance in CLaMP 3 space (Retkowski et al., 2024; Wu et al., 2025) against MAESTRO (Hawthorne et al., 2019) for classical and POP909 (Wang et al., 2020) for pop. Audiobox Aesthetics scores for production quality (PQ) and content enjoyment (CE) (Tjandra et al., 2025) serve as proxies for listener ratings. All audio-based metrics, including those of the reference performances, use audio rendered from MIDI with shared synthesizer settings (Appendix D).

Table 3: Tag agreement and full-prompt alignment (↑). Reference: the 100 held-out high-quality MIDIs, excluded from ranking. m: mode-only key agreement. q: emotion agreement over four quadrants. †: agreement with the key declared by the model. Bold and underline mark the best and second best.

Model	Key	Tempo	Emotion	Genre	Accomp.	Perf.	CLaMP 3	CLAP
Reference	1.000	0.820	0.840	0.830	0.550	0.530	0.132	0.338
MuseCoco	0.582^{m}	0.636	0.575^{q}	0.386	–	–	–	–
text2midi	0.181	0.456	0.276	0.380	0.205	0.270	0.054	0.273
MIDI-LLM	0.340	0.713	0.285	0.370	0.201	0.226	0.059	0.292
EMO-Disentangler	$0.752^{\dagger}$	–	$\mathbf{0.611^{q}}$	–	–	–	–	–
Piano Transformer	–	–	–	0.458	–	–	–	–
TASTE (AR)	<u>0.767</u>	<u>0.730</u>	<u>0.585</u>	<u>0.509</u>	<u>0.296</u>	<u>0.273</u>	**0.141**	<u>0.335</u>
TASTE (DiT)	**0.993**	**0.809**	0.584	**0.603**	**0.326**	**0.350**	<u>0.123</u>	**0.347**

6.2 Controllability and text–music alignment

Table 3 shows that both TASTE models match or exceed the caption-conditioned baselines on all six tag fields and both full-prompt alignment scores. Emotion agreement roughly doubles, from at most 0.285 to 0.584–0.585, and CLaMP 3 alignment more than doubles. TASTE also exceeds the attribute-conditioned baselines on their own fields, except for EMO-Disentangler's quadrant-level emotion. Gains on both the discrete tags and the language-based alignment indicate that the two kinds of condition can be learned jointly within one model. The two families differ in emphasis: DiT, sampled with classifier-free guidance, leads on most tag fields, whereas AR, sampled without guidance, still attains the highest CLaMP 3 alignment and matches DiT on emotion.

Accompaniment and performance style remain the hardest fields. Even on the reference performances, the critics agree with their labels only slightly more than half of the time, and these are the two fields whose large-tier labels fall below the precision target (Table 2). Both observations point to label reliability as the limiting factor for these fields, in training as well as in evaluation.

6.3 Temporal structure

We evaluate section boundaries and recurrence with the structure probe of Section 4.2, which predicts both boundaries and section identities, and a training-free Foote novelty curve (Foote, 2000). *Boundary F1* matches planned and detected boundaries one-to-one within 5 s. *Novelty contrast* divides the novelty near planned boundaries by the piece's median novelty. *Pairwise F1* checks whether the plan and the detector assign the same section identity to each pair of frames, testing whether recurring segments are grouped together.

Table 4: Temporal structure (↑)

System	Boundary F1	Contrast (random)	Pairwise F1
Reference	0.896	2.52 (1.72)	0.769
AR	0.517	1.99 (1.69)	0.608
DiT	0.868	3.18 (1.78)	0.775

Table 4 shows that DiT closely realizes its supplied section plan. Its boundary F1 of 0.868 approaches the 0.896 of the reference performances, and its pairwise F1 matches the reference. Its novelty contrast of 3.18, against 1.78 at random positions and 2.52 for the references, shows clear changes in musical features at planned transitions. AR also places changes near its self-generated boundaries, but its smaller contrast gap (1.99 versus 1.69) and boundary F1 of 0.517 indicate a weaker link between the plan prefix and the events that follow. These results suggest that aligning the plan with the generated sequence supports reasonable realization of form, even for AR sampled without guidance (Appendix D.4).

6.4 Music quality

Table 5 shows that TASTE combines broad condition control with competitive music quality. Both models rank first or second in MAESTRO audio distance and POP909 symbolic distance, and both

Table 5: Objective music quality. V/C: Fréchet distance in VGGish/CLaMP 3 space. Bold and underline mark the best and second best, excluding the reference row.

Model	MAESTRO V ↓	MAESTRO C ↓	POP909 V ↓	POP909 C ↓	PQ ↑	CE ↑
Reference	0.497	193.427	0.662	239.368	8.007	7.333
MuseCoco	1.045	372.282	0.964	280.778	7.709	6.879
text2midi	1.067	296.012	1.113	277.089	7.789	6.953
MIDI-LLM	1.425	282.214	2.025	250.544	7.586	6.840
EMO-Disentangler	0.979	356.460	**0.560**	250.884	7.944	<u>7.448</u>
Piano Transformer	0.745	<u>196.619</u>	1.542	247.119	7.761	7.180
Aria (unconditional)	0.831	320.512	1.114	241.098	7.864	6.868
Compose & Embellish	2.049	383.140	0.898	219.129	**7.984**	**7.502**
TASTE (AR)	<u>0.426</u>	**170.697**	<u>0.605</u>	<u>201.652</u>	<u>7.946</u>	7.410
TASTE (DiT)	**0.309**	239.704	0.808	**196.491**	7.945	7.319

improve on the unconditional Aria model in all six metrics. AR has the lowest MAESTRO symbolic distance and DiT the lowest MAESTRO audio distance, suggesting that the two representations preserve different aspects of the reference distributions. Their production-quality scores are nearly identical (7.946 and 7.945) and close to the best, Compose & Embellish (7.984). Content enjoyment favors Compose & Embellish and EMO-Disentangler, but our models remain competitive.

6.5 Listening study

A listening study measures perceived quality and imagery alignment for both TASTE models, text2midi, MIDI-LLM, and the reference performances, using 24 held-out prompts and the first sample of each system. Twenty-two participants each judged 15 random blind 30-second excerpts, rating overall quality from 1 to 5 and then choosing which of two imagery descriptions fits better.

Table 6: Listening study.

System	Quality (1–5) ↑	Imagery accuracy (%) ↑
Reference	3.85 ± 0.85	69.7
text2midi	2.71 ± 0.91	54.5
MIDI-LLM	<u>2.82 ± 1.08</u>	51.5
TASTE (AR)	$\mathbf{3.26 \pm 0.79}$	**80.3**
TASTE (DiT)	2.73 ± 1.00	<u>75.8</u>

Table 6 shows that listeners match TASTE outputs to their imagery descriptions even more often than the reference. By contrast, text2midi and MIDI-LLM stay near chance, in line with their low emotion agreement (Table 3). Even for the reference performances, listeners choose the unaligned description in 30% of trials, which further supports that imagery admits multiple valid descriptions (Section 3). Quality ratings separate the two TASTE models more than the objective metrics do (Table 5). AR is rated highest among all models, whereas DiT is rated on par with the worst baseline although it outperforms that baseline on all automatic metrics. DiT thus aligns with the imagery about as reliably as AR, but with lower perceived quality.

7 Conclusion

We presented TASTE, which addresses the description-intention gap by asking what makes a good description of music and how data for such descriptions can be collected. Its taxonomy separates tags, section plans, and imagery descriptions, a multi-tier pipeline annotates each channel from suitable evidence, and staged training carries these labels into autoregressive and diffusion models. With the channels kept separate from annotation to generation, both models follow tags and imagery more closely than caption-conditioned baselines, listeners match their outputs to the intended imagery well above chance, and AR's music quality remains competitive.

Yet the results suggest that accompaniment and performance style, the fields with the least reliable labels, remain the hardest to control, making label reliability the next bottleneck. These findings motivate treating a description as a structured specification whose channels each receive suitable evidence and model pathway. TASTE provides a basis for extending this specification toward richer musical intentions while preserving the freedom to realize the same request in different ways.

AI USE STATEMENT

We used large language models (LLMs) to edit and polish the writing of this paper. All research decisions, including the method design, experiments, and conclusions, were made by the authors, who verified every result and all text. The LLMs that annotate data in our pipeline (Section 4) are part of the proposed method and are described in Appendix B.

ETHICS STATEMENT

This work generates symbolic piano music from descriptive conditions. Our corpora combine publicly released piano datasets with high-quality paired data, and the released annotations will follow the licenses of their source datasets. No condition field encodes composer or performer identity, so the models are not designed to imitate specific artists. The listening study collected anonymous responses and recorded no personal information beyond self-reported years of musical training and the playback device. We are not aware of other ethical concerns specific to this work.

REPRODUCIBILITY STATEMENT

We will release the codebase for annotation, corpus construction, training, inference, and evaluation, together with the annotated datasets and model checkpoints. Section 4 and Appendices A–B specify the label vocabularies, annotation routes, estimators, probe training, confidence gates, and retrieval. Section 5 and Appendix C give the architectures, per-stage training configurations, and sampling settings. Section 6 and Appendix D describe the evaluation protocol, critics, baseline interfaces, rendering, and listening study. Data splits, listening prompts, and random-position baselines follow deterministic rules or fixed seeds. Because web-search and LLM-based annotation can vary with time, model version and random seed, the released annotations allow the training and evaluation stages to be reproduced without repeating these steps.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse H. Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. Musiclm: Generating music from text. *CoRR*, abs/2301.11325, 2023.

Hayeon Bang, Eunjin Choi, Megan Finch, Seungheon Doh, Seolhee Lee, Gyeong-Hoon Lee, and Juhan Nam. PIAST: A multimodal piano dataset with audio, symbolic and text. *CoRR*, abs/2411.02551, 2024.

Hayeon Bang, Eunjin Choi, Seungheon Doh, and Juhan Nam. Pianobind: A multimodal joint embedding model for pop-piano music. In *Proc. ISMIR*, pp. 391–398, 2025.

Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorien Herremans. Text2midi: Generating symbolic music from captions. In *Proc. AAAI*, pp. 23478–23486. AAAI Press, 2025.

Louis Bradshaw and Simon Colton. Aria-midi: A dataset of piano MIDI files for symbolic music modeling. In *Proc. ICLR*, 2025.

Louis Bradshaw, Alexander Spangher, Honglu Fan, Stella Biderman, and Simon Colton. Scaling self-supervised representation learning for symbolic piano performance. In *Proc. ISMIR*, pp. 451–459, 2025.

Jongho Choi and Kyogu Lee. Pop2piano : Pop audio-based piano cover generation. In *Proc. ICASSP*, pp. 1–5. IEEE, 2023.

Simon Dixon. Automatic extraction of tempo and beat from expressive performances. *Journal of New Music Research*, 30:39–58, 2001. doi: 10.1076/jnmr.30.1.39.7119.

Drew Edwards, Simon Dixon, and Emmanouil Benetos. Pijama: Piano jazz with automatic MIDI annotations. *Trans. Int. Soc. Music. Inf. Retr.*, 6(1):89–102, 2023.

Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *IEEE International Conference on Multimedia and Expo (I)*, pp. 452. IEEE Computer Society, 2000.

Francesco Foscarin, Jan Schlüter, and Gerhard Widmer. Beat this! accurate beat tracking without DBN postprocessing. In *Proc. ISMIR*, pp. 962–969, 2024.

Alf Gabrielsson. Music performance research at the millennium. *Psychology of Music*, 2003.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse H. Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *Proc. ICLR*, 2019.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *CoRR*, abs/2207.12598, 2022.

Wen-Yi Hsiao, Jen-Yu Liu, Yin-Cheng Yeh, and Yi-Hsuan Yang. Compound word transformer: Learning to compose full-song music over dynamic directed hypergraphs. In *Proc. AAAI*, pp. 178–186. AAAI Press, 2021.

Jingyue Huang, Ke Chen, and Yi-Hsuan Yang. Emotion-driven piano music generation via two-stage disentanglement and functional representation. In *Proc. ISMIR*, pp. 335–342, 2024.

Jingyue Huang, Zachary Novack, Phillip Long, Yupeng Hou, Ke Chen, Taylor Berg-Kirkpatrick, and Julian McAuley. Musetok: Symbolic music tokenization for generation and semantic understanding. In *Proc. ICASSP*, 2026.

Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *Proc. ACM Multimed.*, pp. 1180–1188, 2020.

Hsiao-Tzu Hung, Joann Ching, Seungheon Doh, Nabin Kim, Juhan Nam, and Yi-Hsuan Yang. EMOPIA: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation. In *Proc. ISMIR*, pp. 318–325, 2021.

Xunyi Jiang, Mingyang Yao, Jingyue Huang, and Julian J. McAuley. Composer vector: Style-steering symbolic music generation in a latent space. *CoRR*, abs/2604.03333, 2026.

Patrik N. Juslin and Daniel Västfjäll. Emotional responses to music: The need to consider underlying mechanisms. *Behav. Brain Sci.*, 2008.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet audio distance: A reference-free metric for evaluating music enhancement algorithms. In *Proc. Interspeech*, pp. 2350–2354, 2019.

Jonathan Lehmkuhl, Ábel Ilyés-Kun, Nico Bremes, Cemhan Kaan Özaltan, Frederik Muthers, and Jiayi Yuan. Generating piano music with transformers: A comparative study of scale, data, and metrics. *CoRR*, abs/2511.07268, 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *Proc. ICLR*. OpenReview.net, 2023.

Stuart P. Lloyd. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2):129–136, 1982.

Peiling Lu, Xin Xu, Chenfei Kang, Botao Yu, Chengyi Xing, Xu Tan, and Jiang Bian. Musecoco: Generating symbolic music from text. *CoRR*, abs/2306.00110, 2023.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. The song describer dataset: a corpus of audio captions for music-and-language evaluation. *CoRR*, abs/2311.10057, 2023.

Jan Melechovský, Abhinaba Roy, and Dorien Herremans. Midicaps: A large-scale MIDI dataset with text captions. In *Proc. ISMIR*, pp. 858–865, 2024.

Curtis G. Northcutt, Lu Jiang, and Isaac L. Chuang. Confident learning: Estimating uncertainty in dataset labels. *J. Artif. Intell. Res.*, 70:1373–1411, 2021.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. ICCV*, pp. 4172–4182. IEEE, 2023.

Jan Retkowski, Jakub Stepniak, and Mateusz Modrzejewski. Frechet music distance: A metric for generative symbolic music evaluation. *CoRR*, abs/2412.07948, 2024.

Justin Salamon, Oriol Nieto, and Nicholas J. Bryan. Deep embeddings and section fusion improve music segmentation. In *Proc. ISMIR*, pp. 594–601, 2021.

Qwen Team. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025.

David Temperley. *Music and Probability*. The MIT Press, 2006. ISBN 9780262257077. doi: 10.7551/mitpress/4807.001.0001. URL https://doi.org/10.7551/mitpress/4807.001.0001.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta audiobox aesthetics: Unified automatic quality assessment for speech, music, and sound. *CoRR*, abs/2502.05139, 2025.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proc. NeurIPS*, pp. 5998–6008, 2017.

Dimitri von Rütte, Luca Biggio, Yannic Kilcher, and Thomas Hofmann. FIGARO: controllable music generation using learned and expert features. In *Proc. ICLR*. OpenReview.net, 2023.

Yashan Wang, Shangda Wu, Jianhuai Hu, Xingjian Du, Yueqi Peng, Yongxin Huang, Shuai Fan, Xiaobing Li, Feng Yu, and Maosong Sun. Notagen: Advancing musicality in symbolic music generation with large language model training paradigms. In *Proc. IJCAI*, pp. 10207–10215. ijcai.org, 2025.

Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, and Gus Xia. POP909: A pop-song dataset for music arrangement generation. In *Proc. ISMIR*, pp. 38–45, 2020.

Gerhard Widmer and Werner Goebl. Computational models of expressive music performance: The state of the art. *Journal of New Music Research*, 33:203–216, 2004. doi: 10.1080/09298210420003178O4.

Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *ACL (Findings)*, volume ACL 2025 of *Findings of ACL*, pp. 2605–2625. Association for Computational Linguistics, 2025.

Shih-Lun Wu and Yi-Hsuan Yang. Compose & embellish: Well-structured piano performance generation via A two-stage approach. In *Proc. ICASSP*, pp. 1–5. IEEE, 2023.

Shih-Lun Wu, Dave Carlton, Ryan Mikayawa, Yoon Kim, Chris Donahue, and Cheng-Zhi Anna Huang. MIDI-LLM: Improving text-to-MIDI music generation via adapting large language models. In *Proc. ISMIR*, 2026.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *Proc. ICASSP*, pp. 1–5. IEEE, 2023.

Weihan Xu, Julian J. McAuley, Taylor Berg-Kirkpatrick, Shlomo Dubnov, and Hao-Wen Dong. Generating symbolic music from natural language prompts using an llm-enhanced dataset. In *Proc. ISMIR*, pp. 215–222, 2025.

Yujia Yan and Zhiyao Duan. Scoring time intervals using non-hierarchical transformer for automatic piano transcription. In *Proc. ISMIR*, pp. 973–980, 2024.

Mingyang Yao and Ke Chen. From generality to mastery: Composer-style symbolic music generation via large-scale pre-training. In *Proc. AIMC*, 2025.

Huan Zhang, Akira Maezawa, and Simon Dixon. Renderbox: Expressive performance rendering with text control. *CoRR*, abs/2502.07711, 2025a.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176, 2025b.

A Condition specification

A.1 Tag vocabularies and annotation scope

Table 7 defines the classes of the six tag fields. Each field takes one class, and an abstention becomes the null. On the audio route, the normalizer also applies three scope rules: emotion follows the affect of greatest cumulative duration, accompaniment and performance style describe the first clearly identifiable section, and explicitly broken chords take precedence over the broader ostinato class.

Table 7: Tag vocabularies and annotation scope.

Field	Classes and operational definition
Key	24 classes, 12 major and 12 minor tonics: one global estimate per piece.
Tempo	3 classes from the estimated bpm: *slow* < 75, *moderate* $\in [75, 115)$, *fast* ≥ 115.
Emotion	*happy* brightness, playfulness, or triumph. *sad* sorrow, longing, or grief. *calm* serenity or reflection without persistent distress. *angry* confrontation or defiance. *anxious* tension, restlessness, or foreboding.
Genre	*classical* Baroque through modern and neoclassical. *jazz* ragtime, stride, swing, blues, jazz ballads. *pop* pop-song covers, pop ballads, New Age and easy-listening styles. Hybrids take the strongest overall identity.
Accompaniment	*chordal* block or voiced chords. *ostinato* a repeated driving cell. *arpeggiated* broken or rolling chords. *sustained* held tones, pedal points, or sparse long harmony. *contrapuntal* independently moving voices.
Performance	Two binary axes crossed into four classes: timing *steady* (strict, even, motoric) versus *flexible* (rubato, elastic, breathing), and articulation *connected* (legato, singing, smooth) versus *detached* (staccato, crisp, percussive).

A.2 Section identities and imagery

Section identities are letters assigned in order of first appearance within each piece, so a repeated letter denotes recurring material. A piece uses at most eight identities (A–H), and each section's duration is counted in five-second windows.

Imagery descriptions express a scene, atmosphere, or experience without restating musical tags. On the evidence route, the scene is grounded in the retrieved context, and on the audio route, it is derived from the audio description. In the large tier, rewriting keeps the mood and energy conveyed by a retrieved description and the target's predicted tags while changing the setting (Appendix B.6).

B Corpus construction and validation

B.1 Corpus accounting and splits

Table 8 lists the sources of the curated tier. PIAST-YT and Pop2Piano recordings are transcribed to piano MIDI, while the other sources supply MIDI directly.

The large tier contains 371,053 pieces from the deduplicated Aria-MIDI release (Bradshaw & Colton, 2025), and pretraining uses its pruned release. A note is that tiers are not disjoint by work.

Table 8: Source composition of the curated tier.

Source	Pieces
Aria-MIDI named subset (Bradshaw & Colton, 2025)	32,522
PIAST-YT (Bang et al., 2024)	8,669
Pop2Piano cover index (Choi & Lee, 2023)	5,866
PiJAMA (Edwards et al., 2023)	2,777
POP909 (Wang et al., 2020)	909
EMOPIA-derived pieces (Hung et al., 2021)	387
Total	51,130

The high-quality tier contributes 9,849 audio–text training pairs after removing 50 pairs with empty imagery and holding out 100 pairs for evaluation.

Probe selection and calibration use a 61,129-piece benchmark that joins the curated tier with all high-quality pairs and groups them by work identity. Clips of one recording share a group, and each high-quality piece forms its own group. These groups define a deterministic 80/20 split as well as the folds of five-fold pruning.

B.2 Curated annotation protocols

Evidence route. A Qwen3.6-35B-A3B agent issues at most two web searches and reads at most one page per piece, using only sources whose identity matches the piece: creator, catalogue number, and movement for classical works, and title and artist for songs. Typical sources are encyclopedia entries, program notes, and lyrics pages. It annotates each perceptual field separately and abstains when the evidence is insufficient. Source-provided genre and single-valued emotion labels take precedence. The imagery is written as an original scene grounded in the retrieved context.

Audio route and normalization. Pieces without identifiable work metadata go directly to audio annotation, as do pieces for which the evidence route accepts at most one field. Gemini 3.5 Flash describes the genre, affect, texture, timing, articulation, and development of the rendered audio. Qwen3.6-35B-A3B then normalizes this description into emotion, accompaniment, and performance style, following the scope rules of Appendix A, and into a short imagery description that keeps the scene or atmosphere but omits musical terminology. Genre takes the label of the source dataset.

B.3 Key and tempo estimation

Key is the major or minor key whose Kostka–Payne profile, as reported by Temperley (2006, Table 6.1), correlates best with the piece's duration-weighted pitch-class profile. Durations are measured in seconds under the MIDI tempo map, percussion is excluded, and both profiles are centered and normalized before comparison.

In the smaller tiers, tempo is $60/\mathrm{median}(\Delta t)$ over the inter-beat intervals from Beat This! (Foscarin et al., 2024) and tracks that fail or fall outside 20–400 bpm use the MIDI estimator instead. This estimator, which also serves the whole large tier, is a simplified inter-onset-interval histogram inspired by Dixon (2001). It collects lag-k onset differences for $k \leq 8$ within 0.12–3.0 s, folds their tempos into 50–180 bpm, and smooths a 1 bpm histogram weighted by $1/\sqrt{k}$. The estimate is the weighted mean within 2 bpm of the peak.

B.4 Tag probes and confidence gates

Each probe uses CLaMP 3 music embeddings (768 dimensions) (Wu et al., 2025), Aria embeddings (512) (Bradshaw et al., 2025), and 32 symbolic statistics summarizing timing, pitch, dynamics, polyphony, and pedal. We apply median imputation and standardization before class-balanced logistic regression. Performance style combines two binary probes, predicting flexible timing at

$p_{\text{flex}} \geq 0.45$ and detached articulation at $p_{\text{det}} \geq 0.75$, with joint confidence

$$\min\{\max(p_{\text{flex}}, 1 - p_{\text{flex}}), \max(p_{\text{det}}, 1 - p_{\text{det}})\}.$$

Pruning follows a confident-learning rule (Northcutt et al., 2021). For supplied label y_i and work-grouped out-of-fold probabilities $p_i(c)$, define the class self-confidence $\mu_c = |I_c|^{-1} \sum_{i \in I_c} p_i(c)$, where $I_c = \{i : y_i = c\}$. Example i is removed only when $\arg\max_c p_i(c) \neq y_i$ and $p_i(y_i) < \mu_{y_i}$. Table 9 compares the pruned and unpruned fits: pruning helps emotion, accompaniment, and articulation most, leaves genre and timing nearly unchanged, and slightly lowers the combined performance style. The production probes are refit on the retained labels of the full benchmark, and calibration uses train-side fits evaluated on the held-out split.

Table 9: Tag-probe macro-F1 before and after pruning. Kept: fraction of training labels retained.

	Emotion	**Genre**	**Accomp.**	**Timing**	**Articulation**	**Performance**
Unpruned	0.5353	0.9086	0.4445	0.6527	0.7256	0.5119
Pruned	0.5629	0.9130	0.4621	0.6523	0.7405	0.5027
Kept	55.3%	93.1%	50.8%	65.4%	81.0%	–

For field f, θ_f^{prec} is the lowest validation confidence at which precision reaches 0.60 with at least 50 retained examples, and θ_f^{floor} is the quantile of production scores that retains 75% of the large tier. The applied gate is

$$\theta_f = \min(\theta_f^{\text{prec}}, \theta_f^{\text{floor}}),$$

and predictions below it are omitted from the conditions. For accompaniment and performance style, the coverage floor binds and precision falls below the target (Table 10). Genre uses mapped source metadata where available (276,948 pieces) and the probe elsewhere.

Table 10: Confidence thresholds, validation precision, and large-tier coverage. Genre coverage includes source metadata.

Field	θ^{prec}	θ^{floor}	**Applied**	**Val. prec.**	**Coverage**	**Binding rule**
Emotion	0.7319	0.8807	0.7319	0.600	86.6%	precision
Genre	0.3968	1.0000	0.3968	0.936	100.0%	precision
Accompaniment	0.9873	0.8097	0.8097	0.538	75.0%	coverage floor
Performance	0.9319	0.9083	0.9083	0.588	75.0%	coverage floor

B.5 Section supervision

The audio teacher for the curated tier tracks beats with Beat This! (Foscarin et al., 2024) and then applies beat-synchronous deep, few-shot, and CQT features with spectral clustering and section fusion (Salamon et al., 2021). We use hierarchy level 4, a 12-second minimum section, $\mu = \gamma = 0.5$, and seed 0.

For high-quality pieces, the supplied letter sequence constrains segmentation. At a 0.5-second frame rate, segments sharing a letter share a centroid. With segment lengths d_i, letters L_i, interior cuts b_i, and total length T, the objective is

$$\sum_{i=1}^{N} \sum_{t \in \text{seg}_i} \|x_t - \mu_{L_i}\|^2 + \lambda_{\text{dur}} \sum_{i=1}^{N} \frac{(d_i - T/N)^2}{T/N} - \lambda_{\text{nov}} \sum_{i=1}^{N-1} \text{nov}(b_i).$$

Dynamic programming optimizes the cuts for fixed centroids, and up to twelve Lloyd iterations (Lloyd, 1982) alternate between cut placement and centroid updates until convergence. We use $\lambda_{\text{dur}} = 0.10$, a 10-second novelty window, and an 8-second minimum segment. The first pass uses chroma, weighted by two, together with six texture statistics and $\lambda_{\text{nov}} = 0$. A second pass,

initialized from those cuts, uses standardized, smoothed, unit-normalized Aria layer-7 features with $\lambda_{\text{nov}} = 1.0$. On 50 audited pieces with 236 reference boundaries, the mean absolute boundary error is 3.11 s, compared with 6.3 s for equal spacing.

The MIDI-native structure probe has 1,817,847 parameters: 26 symbolic features at 0.5-second resolution feed a four-layer, four-head transformer of width 192, with a boundary head and a 64-dimensional grouping head. Gaussian boundary targets use $\sigma = 1.5$ s. The grouping head reads the encoder states through a stop-gradient and learns a balanced same-section classification over teacher spans. Training uses 10 epochs, batch size 16, learning rate 3×10^{-4}, a 95/5 piece split, and threefold repetition of high-quality training pieces.

With a 5-second tolerance, a 12-second peak spacing, and each subset's validation-optimal threshold, the checkpoint that labeled the curated tier and the high-quality boundaries agrees with its teachers at a boundary F1 of 0.525 on 2,553 curated pieces and 0.878 on 503 high-quality pieces. It also labels the pretraining corpus and evaluation samples. Production labeling of the public tiers uses a 20-second spacing, boundary threshold 0.15, and agglomerative merging at same-section probability 0.65, capped at eight identities. Adjacent equal identities are merged, and a final span shorter than five seconds is absorbed. For high-quality pieces, each supplied boundary moves to the nearest unused probe peak (threshold 0.30, 12-second spacing) within ± 4 s, provided that neighboring boundaries stay at least 12 s apart and the boundary stays at least 5 s from the end.

B.6 Imagery retrieval and rewriting

The candidate pool contains 60,783 descriptions with paired Aria and CLaMP 3 text embeddings. Ridge regression with regularization 1.0 maps normalized Aria vectors into the text space. For a query piece q, we retrieve its 50 nearest candidates in Aria space and select

$$j^* = \arg \max_{j \in \mathcal{N}_{50}(q)} \left[0.6 \cos(W a_q, t_j) + 0.3 \cos(e_q, e_j) \right],$$

where a_q is the query's Aria embedding, t_j the CLaMP 3 text embedding of candidate j, and e the predicted emotion distribution. Qwen3.5-27B-FP8 rewrites the selected scene, keeping the mood and energy conveyed by the scene and the target's predicted tags, including tempo, while changing the setting. Lexical checks reject rewrites that use musical terminology, drift into another script, or contain fewer than eight words, and such rewrites are regenerated. We also use 100 held-out prompt-MIDI pairs for evaluation.

B.7 Restricted backward consolidation

On audio-annotated pieces, a prediction that passes the production gate of Table 10 may replace any of the four perceptual tags. On evidence-annotated pieces, the probes fill missing accompaniment and performance labels under the same gates and leave emotion and genre untouched. Existing accompaniment labels there are replaced at the production gate. Replacing an evidence-derived performance label requires a confidence of at least 0.9982, which limits such overrides to 3,013 of the 40,652 evidence-annotated pieces (7.41%). Predictions below the applicable threshold leave the original label unchanged.

The structure probe supplies the curated plans and snaps high-quality boundaries with the production settings of Appendix B.5, preserving the supplied identities. High-quality tags and all imagery descriptions keep their original annotations.

C Model, training, and inference specification

C.1 Representation and augmentation

Figure 4 shows a tokenized AR sequence. The vocabulary contains 2,120 active identifiers, padded to 2,176 embedding entries. Alongside note and timing events, it includes eight section identities, 24 key classes, three tempo classes, ten plan digits, and an interior-crop start marker. Prefix and event tokens share the 4,096-token context.

Pretraining crops start at a randomly selected eligible section and extend over whole sections. An overlong section is cut at a window boundary, and the plan is recomputed to fit. The end token

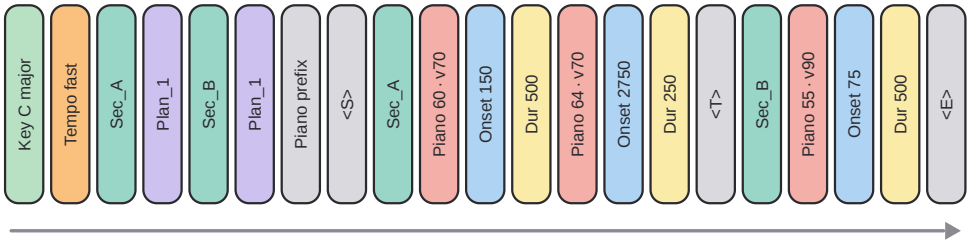


Figure 4: Tokenization example. The prefix holds key, tempo, and the section plan, where Sec_A, Plan_1 means that section A lasts one five-second window. <T> anchors each window and is followed by the active section identity, and <S> and <E> mark the start and end of the performance.

is kept only when the crop reaches the end of the piece. Conditional rows always start at the true opening.

Augmentation draws an integer transposition uniformly from $[-5, +5]$ semitones and a time factor a uniformly from $[0.90, 1.10]$. Notes and key transpose together, event times and section boundaries scale by a, and a tempo of b bpm becomes b/a. Velocity is unchanged. AR pretraining uses one clean pass followed by four augmented passes. Every other stage, including DiT pretraining, places the clean pass last.

C.2 Autoregressive conditioning

Table 11 specifies the Qwen3-style backbone (Team, 2025). For each of the four perceptual fields f, the selected tag a_f indexes an embedding table E_f that includes a null entry for a missing or dropped label. The shared code is

$$c = \mathrm{SiLU}\left(W_c \sum_f E_f[a_f] + b_c \right).$$

Each layer projects c into a shift, scale, and gate (β, γ, g) for its self-attention and feed-forward branches. For branch F and pretrained normalization N, the update is

$$h' = h + (1 + g) \odot F((1 + \gamma) \odot N(h) + \beta).$$

The modulation heads start at zero, so the $1 + g$ gate preserves the pretrained branch at initialization.

Imagery enters through a cross-attention branch placed between self-attention and the feed-forward branch:

$$h' = h + W_o \,\mathrm{Attn}(W_q \,\mathrm{RMSNorm}(h), W_k S, W_v S).$$

Here S contains the final-layer token states of a frozen Qwen3-Embedding-4B (Zhang et al., 2025b), truncated at 64 tokens without pooling. A learned null context token is always prepended, and a missing imagery description attends to it alone. The output projection W_o starts at zero, so adding both conditioning pathways initially preserves the pretrained model.

C.3 Autoregressive training

Table 12 summarizes the training configuration of each stage. All stages use AdamW with weight decay 0.1, 1% warm-up, cosine decay to 10% of the peak learning rate, gradient clipping at 1.0, bfloat16 precision, seed 0, and four RTX A6000 GPUs. Token-budget batching allows 12,288 tokens per device microbatch. Conditional training drops the four perceptual tags and the imagery together with probability 0.10, replacing them with their null conditions, while key, tempo, and plan tokens remain in the event sequence.

The first large-public pass orders pieces from sparse to complete perceptual labels, and later passes are shuffled. The high-quality training count follows the exclusions and split in Appendix B. Architecture, conditioning, and augmentation remain fixed across the conditional stages.

Table 11: Autoregressive architecture. Parameter counts exclude the frozen text encoder.

Parameter	Value
Layers / hidden width / context	26 / 1,536 / 4,096
Query heads / key–value heads / head width	24 / 6 / 64
Feed-forward width (SwiGLU) / embedding table	6,144 / 2,176
Normalization / rotary base / embeddings	RMSNorm + QK-norm / 100,000 / tied
Backbone parameters	892,881,664
Tag tables (emotion, genre, accomp., perf.)	$(6, 4, 6, 5) \times 1{,}536$
Shared trunk (linear + SiLU)	$1{,}536 \to 512$
Per-layer tag head, zero-initialized	$512 \to 6 \times 1{,}536$
Cross-attention heads × width, per layer	16×64
Text-state width / null context tokens	2,560 / 1
Added parameters / conditional total	341,888,512 / 1,234,770,176

Table 12: Autoregressive training configuration by stage.

	Pretrain	Large public	Curated	High-quality
Distinct pieces per pass	816,682	369,553	50,130	9,849
Passes / training rows	5 / 4,083,410	6 / 2,217,318	3 / 150,390	5 / 49,245
Validation rows	1,518	1,500	1,000	100
Training tokens	11.44B	6.60B	429.72M	155.97M
Optimizer updates	43,708	12,616	1,644	607
Peak learning rate	2×10^{-4}	1.5×10^{-4}	5×10^{-5}	2.5×10^{-5}
Token budget per update / accumulation	294,912 / 6	589,824 / 12	294,912 / 6	294,912 / 6
Condition dropout	–	0.10	0.10	0.10

C.4 Latent representation and diffusion conditioning

The MuseTok-based codec (Huang et al., 2026) maps each one-second window of the token sequence to a 128-dimensional posterior mean. Windows start every 0.5 s, yielding two overlapping latents per second. Source rows are cropped to at most 190 s, at a section boundary where possible, and the model has a 384-position canvas (192 s). Batches are length-bucketed, and padding is excluded from attention and loss.

Table 13 specifies the diffusion transformer (Peebles & Xie, 2023). Its conditioning code sums a transformed timestep embedding with a transformed sum of label embeddings for the six tags and two flags. The completeness flag is set when a row ends at the end of its piece, and the interior-start flag when it begins after the opening. Per-position embeddings of section identity, within-section position, and absolute position carry the plan. Imagery uses the same frozen text states as AR, with a null token. Each block's AdaLN-Zero head supplies a shift, scale, and gate for each of its self-attention, cross-attention, and feed-forward branches. The modulation heads and the final output projection start at zero, so every branch is initially switched off.

Table 13: Latent diffusion architecture with a 384-position canvas.

Parameter	Value
Hidden width / layers / heads	1,024 / 16 / 16
Feed-forward ratio (GELU) / latent width	4× / 128
Canvas positions / conditioning-code width	384 / 512
AdaLN-Zero head output width per block	$9 \times 1{,}024$
Cross-attention context width	2,560
Normalization	LayerNorm, no affine; QK-RMSNorm
Parameters	397,558,400

C.5 DIFFUSION OBJECTIVE AND TRAINING

The rectified-flow objective minimizes the mean squared velocity error over non-padded positions. One timestep per sequence is drawn from a logit-normal distribution, $t = \sigma(u)$ with $u \sim \mathcal{N}(\mu, 1)$, where $\mu = -0.5$ during pretraining and -0.25 during conditional training.

Classifier-free dropout (Ho & Salimans, 2022) independently nulls three condition groups with probability 0.10 each: (1) perceptual tags and imagery, (2) key, tempo, and the completeness and interior-start flags, and (3) the section plan. These drops add to the missing or dropped conditions inherited from the corpus.

All stages use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, warm-up and cosine decay to 10% of the peak learning rate, gradient clipping at 1.0, bfloat16 computation and gradient compression, and an exponential moving average with decay 0.9995. The moving-average weights initialize the next stage and are used for inference. Training uses two RTX 5090 GPUs with per-device batch size 24 and accumulation 8, for a global batch of 384. As in AR, a pass is one clean or augmented copy of the corpus and keep original pass last (Table 14).

Table 14: Diffusion training configuration. Curated (b) continues (a) with a restarted schedule. Pretraining rows per pass are listed as clean / augmented.

	Pretrain	**Large public**	**Curated**		**High-quality**
			(a)	**(b)**	
Steps	120,000	120,000	24,000	36,000	9,000
Peak learning rate ($\times 10^{-5}$)	10	10	5	3	2
Warm-up steps	3,000	2,000	500	500	300
Timestep mean μ	-0.5	-0.25	-0.25	-0.25	-0.25
Passes visited	2	3	3	3	3
Rows per pass	816,633 / 816,641	369,552	50,129	50,129	9,849
Validation rows	1,518	1,500	1,000	1,000	100

C.6 INFERENCE SETTINGS

AR receives key and tempo tokens, the perceptual tags, and the imagery description. It first samples its section-plan prefix and then the performance events, with grammar masking to control the transition from plan to events. Sampling uses temperature 0.98, nucleus probability 0.98, no top-k truncation, no guidance, and at most 4,094 new tokens.

DiT receives the section plan, latent length, and completeness flag of the reference performance. For 19 of the 100 prompts, the reference row is capped at the 4,096-token AR context, so DiT receives the plan and length of its opening, with the completeness flag unset. Fifty uniform Euler steps integrate from $t = 1$ to 0 with the guided velocity

$$v_\varnothing + 3.0(v_c - v_\varnothing),$$

where the null branch drops all three condition groups. Sampled latents are destandardized and decimated to integer-second offsets before grammar-constrained decoding in 20-second chunks, with a cap of 1,024 tokens per chunk and the end token masked because the duration is supplied.

D EVALUATION AND EXTENDED RESULTS

D.1 EVALUATION SAMPLES AND RENDERING

Objective results use all eight samples for each of the 100 held-out prompts, except that 12 failed MuseCoco decodes are excluded, leaving 788. The reference row contains the 100 paired performances. Appendix C.6 gives the sampling settings.

All systems and references set are rendered from MIDI with FluidSynth and the TimGM6mb soundfont at a shared gain. CLaMP 3 inputs are standardized to 960 ticks per beat and 120 bpm while

preserving performance timing. Distances use all 1,276 MAESTRO v3 and all 909 original POP909 performances. Audiobox scores are averaged over 10-second windows within each clip and then over clips.

D.2 Baseline interfaces

Table 15 lists the inputs each baseline receives. MuseCoco accepts only the key mode, so its key agreement is scored on mode alone (m); quadrant-conditioned models are scored on quadrant-folded emotion (q). EMO-Disentangler receives no key but declares one in its first stage, so its key agreement is measured against that declared key (†).

Table 15: Baseline input mappings. Emotion quadrants map happy to Q1, angry/anxious to Q2, sad to Q3, and calm to Q4.

System	Received conditions and adaptation
MuseCoco (Lu et al., 2023)	Attribute-to-music stage: piano, emotion quadrant, genre, key mode, tempo class, 4/4 meter, and a length request over 60 s. Native top-k 15, temperature 1.0, and 512–2,560 tokens. Its tempo cutoffs (76/120 bpm) differ from ours (75/115).
text2midi (Bhandari et al., 2025), MIDI-LLM (Wu et al., 2026)	All six tags and imagery in a caption. Tempo classes become Adagio 65, Andante 95, or Allegro 135 bpm. Decoding masks non-piano programs and drum notes.
EMO-Disentangler (Huang et al., 2024)	Emotion quadrant, decomposed into valence and arousal across its two stages. The first stage samples and declares its own key.
Piano Transformer (Lehmkuhl et al., 2025)	Genre prefix only.
Aria (Bradshaw et al., 2025), Compose & Embellish (Wu & Yang, 2023)	Unconditional piano generation for quality evaluation.

Caption-conditioned baselines receive: *"A {emotion} {genre} solo piano piece featuring only piano. Set in {key} with a 4/4 time signature, it moves at an {tempo word} tempo of {bpm} bpm, with {accompaniment} accompaniment, played {performance}. {imagery description}"*.

D.3 Critics and interpretation

Key and tempo. Key and tempo are read back with the estimators of Appendix B.3. Reference key agreement is 1.000 by construction. Reference tempo labels come from audio beat tracking, and the MIDI estimator agrees with them on 0.820 of the reference pieces.

Tag critics. The tag critics apply class-balanced logistic regression to concatenations of CLaMP 3 (768-d) and Aria (512-d) embeddings. Their five-fold accuracies for emotion, genre, accompaniment, and performance style are 0.869, 0.852, 0.543, and 0.491. The lower accuracy for accompaniment and performance style matches their lower agreement on the reference set in Table 3.

External diagnostics. The external emotion critic (E4) is a ReLU MLP with 512 hidden units fitted on 1,078 EMOPIA clips (Hung et al., 2021), with song-grouped five-fold accuracy 0.745. Zero-shot genre (G0) selects the most similar LAION-CLAP text embedding among "A {class} solo piano piece." for the three genres. Table 16 adds both diagnostics to the main comparison. AR leads E4 and G0 among the evaluated generators, whereas DiT leads on the genre tag critic, indicating that the ranking depends on the domain in which a critic was trained.

Full-prompt alignment. CLaMP 3 and CLAP report the mean music–text cosine similarity for "A {emotion} {genre} solo piano piece with {accompaniment} accompaniment, played in a {performance} style. {imagery description}". This caption evaluates tags and imagery jointly, whereas the listening study presents the imagery without tags.

Table 16: Tag agreement with external diagnostics (↑). E/G/A/P: emotion/genre/accompaniment/performance. E4: external emotion. G0: zero-shot genre. Ranking and the marks m, q, and † follow Table 3.

Model / stage	Key	Tempo	E4	E	G0	G	A	P	CLaMP 3	CLAP
Reference	1.000	0.820	0.630	0.840	0.500	0.830	0.550	0.530	0.132	0.338
MuseCoco	0.582^{m}	0.636	0.280	0.575^{q}	0.315	0.386	–	–	–	–
text2midi	0.181	0.456	0.268	0.276	0.330	0.380	0.205	0.270	0.054	0.273
MIDI-LLM	0.340	0.713	0.281	0.285	0.391	0.370	0.201	0.226	0.059	0.292
EMO-Disentangler	$0.752^{\dagger}$	–	0.477	$\mathbf{0.611^{q}}$	–	–	–	–	–	–
Piano Transformer	–	–	–	–	0.375	0.458	–	–	–	–
TASTE (AR)	0.767	0.730	**0.514**	0.585	**0.444**	0.509	0.296	0.273	**0.141**	0.335
TASTE (DiT)	**0.993**	**0.809**	0.406	0.584	0.414	**0.603**	**0.326**	**0.350**	0.123	**0.347**

D.4 Temporal evaluation

The retrained structure probe of Appendix B.5 segments generations without any priors. Because we placed the boundaries based on the probe of the high-quality plans (Appendix B.7), the reference boundary F1 is partly by construction. Evaluation uses boundary threshold 0.30, a 20 s minimum peak spacing, grouping threshold 0.65, at most eight identities, merging of adjacent equal identities, and absorption of a final span shorter than five seconds. The complementary Foote-style novelty curve (Foote, 2000) uses sustain- and velocity-weighted chroma plus texture features at 2 fps, cosine self-similarity, and a Gaussian-tapered checkerboard kernel with a 10-second half-width.

Boundary precision, recall, and F1 use greedy one-to-one matching within 5 s, with counts pooled over the cohort. Planned boundaries within 2.5 s of the last note-off are excluded. Contrast is the mean peak novelty within ± 2.5 s of a piece's planned boundaries divided by its median interior novelty, and Tables 4 and 19 report the cohort median. The random-position comparison uses equally many seeded interior positions with the same local-maximum rule. Pairwise F1 compares same-section frame pairs at 0.5-second resolution and is averaged over pieces. Single-section plans are excluded from boundary and contrast scores but retained for pairwise F1.

D.5 Effect of staged training

Tables 17, 18, and 19 report control, quality, and temporal structure after each conditional stage, together with an AR ablation that skips the large-public stage. The two refinement stages contribute different kinds of control. Curated supervision raises genre agreement in both families, most clearly for AR, but lowers emotion agreement and boundary F1. High-quality refinement then brings the largest gains in emotion agreement, CLaMP 3 alignment, and boundary F1, which all peak at this stage in both families, while AR's genre agreement partly falls back. Curated evidence thus mainly informs genre, whereas high-quality pairs improve affective, semantic, and temporal alignment.

Accompaniment and performance style develop less consistently across stages, in line with their less reliable labels. For DiT, the large-public stage already attains the highest tempo, accompaniment, and CLAP scores, which the later stages do not improve. In temporal structure, AR's recurrence agreement peaks at the curated stage, whereas DiT improves both boundary and recurrence agreement at the high-quality stage. Learning when to change and learning which material should recur thus respond differently to supervision, particularly when AR generates the plan along with the music.

Refinement also shifts the output distribution toward pop: symbolic distance to POP909 falls at each stage, and audio distance to MAESTRO rises from the first to the final stage. Production-quality and content-enjoyment scores are highest after the high-quality stage in both families.

The stage-wise rows cannot show how much the large-public stage itself contributes, since every later stage starts from it. To isolate it, we train AR from the pretrained model on the curated and then high-quality tiers with otherwise identical settings (*w/o large public*). Compared with the full recipe, the final model changes only modestly. Emotion, genre, and accompaniment agreement drop by 0.019 to 0.044, CLaMP 3 alignment from 0.141 to 0.131, and boundary and pairwise F1 from

0.517 to 0.464 and from 0.608 to 0.592. Key, tempo, performance style, and the external critics stay level or slightly improve, and the quality metrics change little in either direction. Most of the control thus comes from the curated and high-quality tiers, while the large public tier adds a few points on emotion, genre, accompaniment, CLaMP 3 alignment, and temporal structure.

The ablation also ends with a higher validation loss on the high-quality tier (1.715 versus 1.695). In an informal check of a few samples, its outputs sounded structurally simpler and of lower quality, a gap that the automatic quality metrics may understate, as they did for DiT in Section 6.5.

Table 17: Control and alignment across training stages (↑), with 800 samples per row. + denotes continued training. *w/o large public*: AR trained on the curated and high-quality tiers only, not ranked. Notation follows Table 16.

Model / stage	Key	Tempo	E4	E	G0	G	A	P	CLaMP 3	CLAP
AR										
Large public	0.642	0.731	0.440	0.395	0.388	0.451	0.295	0.281	0.083	0.339
+ Curated	0.559	0.682	0.406	0.370	0.434	**0.615**	0.236	0.300	0.090	0.332
+ High-quality	0.767	0.730	**0.514**	**0.585**	**0.444**	0.509	0.296	0.273	**0.141**	0.335
w/o large public	0.781	0.754	0.516	0.566	0.458	0.481	0.252	0.299	0.131	0.333
DiT										
Large public	0.991	**0.886**	0.376	0.372	0.409	0.540	**0.367**	0.349	0.087	**0.351**
+ Curated	0.974	0.696	0.255	0.329	0.374	0.561	0.275	0.319	0.082	0.345
+ High-quality	**0.993**	0.809	0.406	0.584	0.414	0.603	0.326	**0.350**	0.123	0.347

Table 18: Quality across training stages. Metrics follow Table 5, and rows follow Table 17.

	MAESTRO ↓		POP909 ↓		PQ ↑	CE ↑
Model / stage	V	C	V	C		
AR						
Large public	0.210	172.422	0.998	286.869	7.775	7.252
+ Curated	0.319	175.372	1.177	226.742	7.781	7.168
+ High-quality	0.426	**170.697**	**0.605**	201.652	**7.946**	**7.410**
w/o large public	0.445	165.549	0.592	206.130	7.947	7.416
DiT						
Large public	0.220	213.041	0.871	239.708	7.878	7.303
+ Curated	**0.209**	233.151	0.944	217.005	7.858	7.211
+ High-quality	0.309	239.704	0.808	**196.491**	7.945	7.319

Table 19: Temporal structure across training stages (↑). P/R: boundary precision/recall. Parentheses give random-position contrast. Rows follow Table 17.

	Boundary @5 s ↑			Contrast ↑	Pairwise
Model / stage	P	R	F1	(random)	F1 ↑
Reference	0.876	0.918	0.896	2.52 (1.72)	0.769
AR					
Large public	0.384	0.301	0.338	1.88 (1.66)	0.600
+ Curated	0.362	0.275	0.313	1.82 (1.65)	0.623
+ High-quality	0.567	0.475	0.517	1.99 (1.69)	0.608
w/o large public	0.518	0.420	0.464	1.93 (1.67)	0.592
DiT					
Large public	0.832	0.770	0.799	3.02 (1.84)	0.721
+ Curated	0.788	0.702	0.743	2.98 (1.84)	0.707
+ High-quality	**0.867**	**0.869**	**0.868**	**3.18** (1.78)	**0.775**

D.6 LISTENING PROTOCOL

The 24 listening prompts are sampled with a fixed seed under emotion quotas (5 angry, 5 anxious, 5 calm, 5 happy, 4 sad) from the prompts for which every system produced at least 10 seconds of audio. Clips use the opening 30 seconds of the first sample, with a one-second fade at the cut and peak normalization.

The 120 clips are distributed over eight questionnaire versions of 15 scored items, with three items per system and no prompt repeated within a version; an unscored practice item comes first. Each blind audio item first asks for overall quality from 1 (Bad) to 5 (Excellent) and then reveals two imagery descriptions for a forced choice.

D.7 LISTENING RESULTS BY MUSICAL TRAINING

Table 20 splits the listening results by the participants' self-reported years of musical training. Listeners with more training rate every system lower, and training level correlates with a participant's mean quality rating (Spearman $\rho = -0.48$) but not with imagery accuracy ($\rho = -0.07$). Both groups still rate the reference highest and AR second, and both match TASTE outputs to their descriptions at the same rate (78.3% and 77.8% over AR and DiT). The imagery results thus hold for listeners with and without musical training.

Table 20: Listening results by years of musical training: 10 participants with less than two years and 12 with two or more. Imagery accuracy is in percent; chance is 50%.

	Quality (1–5)		**Imagery accuracy (%)**	
System	**<2 years**	**≥2 years**	**<2 years**	**≥2 years**
Reference	4.00	3.72	70.0	69.4
text2midi	2.90	2.56	56.7	52.8
MIDI-LLM	2.93	2.72	50.0	52.8
TASTE (AR)	3.37	3.17	76.7	83.3
TASTE (DiT)	2.97	2.53	80.0	72.2
All systems	3.23	2.94	66.7	66.1