YuE2: Unifying Symbolic and Audio Music Generation at Frontier Quality

Anonymous authors
Paper under double-blind review

Abstract

Symbolic models make musical composition explicit, while audio models generate complete recordings with composition largely implicit. We introduce **YuE2**, which unifies symbolic and audio music generation in one model at **frontier song quality**. Through **symbolic planning**, the model writes an editable melody-and-chord score, expands it into semantic music tokens, and renders full-song audio. The score specifies musical structure while leaving performance details to audio generation. Automatically constructed score and semantic targets let the model learn this generation process from recordings without pre-existing aligned scores. On WildSongBench, YuE2 scores 6.73 on SongBench (Wu et al., 2026) Global Avg with a two-candidate budget, exceeding all evaluated public baselines. YuE2 (best-of-8) achieves 6.96, the highest observed mean among all evaluated systems. In expert listening, YuE2 (best-of-8) receives 57.3% of overall preferences versus 30.5% for Suno v4.5 (Team Suno, 2025), and 40.4% versus 39.9% for Suno v5 (Suno, 2025). Within the same checkpoint, symbolic planning improves overall quality and musicality: 49.3% of overall judgments favor planning versus 34.6% without it. The generated audio follows the particular planned melody and harmony; score edits change the requested content while largely preserving unedited musical material. Without cover-specific training, the same checkpoint generates zero-shot covers from full scores of 948 unseen works, surpassing both evaluated cover systems on all eight work-identity retrieval metrics. This score-to-audio mapping also enables **agentic music editing**. An external language-model agent turns user feedback into score, style, and lyric revisions, which YuE2 renders as complete songs. Audio examples: `https://yue2-0001.github.io`.

1 Introduction

A finished song contains both a composition and its performance. Symbolic models expose notes and musical structure (Huang et al., 2018; Chen et al., 2024), while audio models generate complete recordings with composition largely implicit (Agostinelli et al., 2023). A score makes melody, harmony, rhythm, and form explicit while leaving aspects of timing, timbre, vocal expression, and production open to interpretation (Wu et al., 2021). Planning the full score could coordinate melody and harmony across phrases and sections before audio generation. Can one model compose an editable score and realize it as a complete recording at frontier song quality?

We introduce YuE2, which **unifies symbolic composition and audio generation in one model at frontier full-song quality** (Figure 1). Given a style request and lyrics, it writes an editable melody-and-chord score, expands it into semantic music tokens, and generates acoustic latents that a decoder converts into a complete recording. One autoregressive–non-autoregressive (AR–NAR) Mixture-of-Transformers (Deng et al., 2025) jointly models these three representations. The score makes compositional decisions explicit and editable; the later representations supply the detail needed to turn those decisions into sound. We construct score and semantic targets offline from recordings, enabling training without pre-existing aligned scores.

Under the standard two-candidate protocol, YuE2 leads the evaluated public systems on SongBench Global Avg; best-of-8 has the highest observed mean among all evaluated systems. Expert listening favors best-of-8 over Suno v4.5 and yields nearly balanced preferences against Suno v5, while

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.

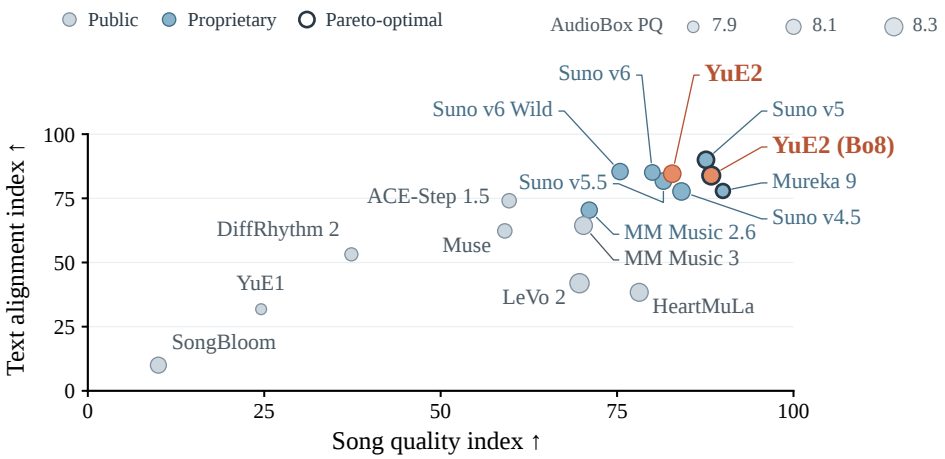

Figure 1: **Explicit symbolic planning at frontier song quality.** Both YuE2 settings exceed every evaluated public baseline on both displayed indices; best-of-8 (Bo8) lies on the observed Pareto front. Both settings generate melody-and-chord plans, selecting from two or eight candidates on 192 WildSongBench prompts. Higher and further right is better. The axes are relative indices, not percentages: quality combines standardized SongBench (Wu et al., 2026) and SongEval (Yao et al., 2025) averages; alignment combines standardized MuQ-MuLan (Zhu et al., 2025), AllMusicCaps (Alonso-Jiménez et al., 2026), and Qwen3-Omni (Xu et al., 2025) prompt adherence. Larger bubbles indicate higher AudioBox production quality (Tjandra et al., 2025). Black outlines mark settings for which no other setting improves one index without lowering the other. Table 1 gives individual metrics; Appendix E.4 gives index definitions, weights, and sensitivity analyses.

Suno v6 receives more preferences (Suno, 2026). These results establish competitiveness with the evaluated proprietary systems.

We also compare the unified model with a separate language model and diffusion Transformer (Xu et al., 2026). Both systems use the same representations and melody-and-chord planning; experts favor the unified model for overall quality (53.4% versus 35.6%).

Symbolic planning improves perceived song quality within the same checkpoint. With prompts, lyrics, candidate budget, and decoder fixed, experts favor planning for overall quality (49.3% versus 34.6%), musicality, melody, and chord progression. The score also provides selective control: recordings agree much more closely with their own plans than with other plans for the same prompt. Local score edits change targeted content while preserving most unedited melody and harmony, despite regenerating the whole song.

The learned score-to-audio mapping also enables zero-shot covers of unseen compositions without cover-specific training. Across 948 held-out works, YuE2 with full scores surpasses both evaluated cover systems on all eight work-identity retrieval measures and on automatic measures of production quality and musicality (Section 3.6). The readable score also supports agentic music editing. An external language-model agent can read the score, revise its melody, harmony, and form from user feedback, and invoke YuE2 to perform the revised composition.

Contributions.

1. We introduce YuE2, which unifies symbolic and audio music generation through symbolic planning and reaches frontier full-song quality.
2. Within-checkpoint expert comparisons show that symbolic planning improves perceived song quality. With melody-and-chord planning in both systems, experts also prefer the unified model to a baseline with separate generators.
3. We show that the score governs musical content and supports selective editing and zero-shot covers through the same learned score-to-audio mapping. The readable score also serves as a natural interface for language-model agents to revise the composition from user feedback, enabling agentic music editing.

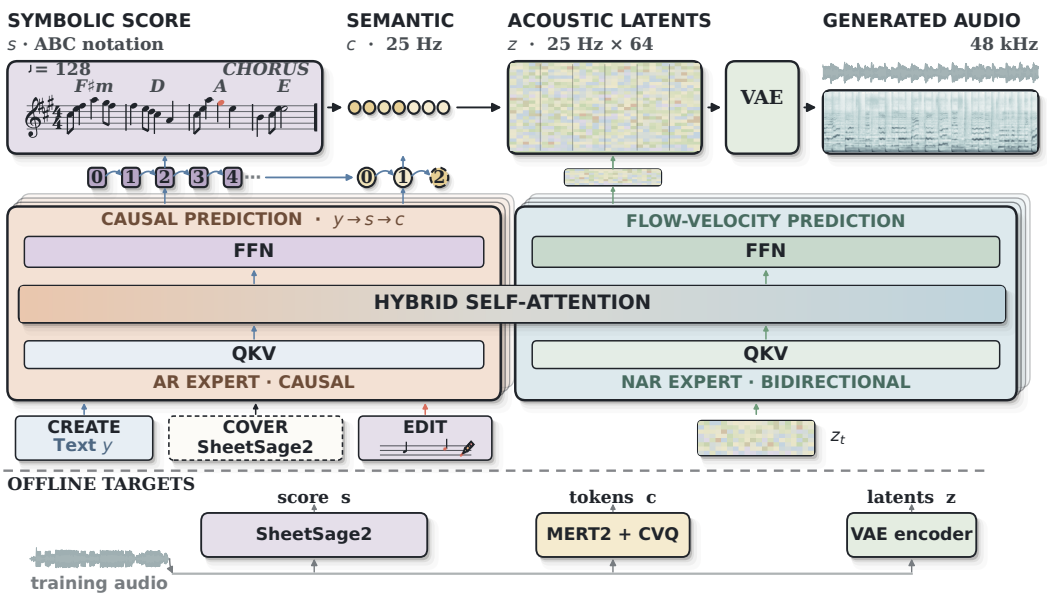

Figure 2: **Composing in symbols, performing in audio.** Given style and lyrics y, YuE2's autoregressive (AR) expert predicts score s, then semantic tokens c. Its non-autoregressive (NAR) expert generates acoustic latents z from noisy latents z_t through flow matching. A variational autoencoder (VAE) decodes z into stereo audio. The experts share attention with separate query/key/value projections (QKV) and feed-forward networks (FFN). **Creation** generates the score; **covering** transcribes one with SheetSage2; **editing** supplies a revised score. Offline targets from recordings are SheetSage2 scores, MERT2 tokens with clustered vector quantization (CVQ), and VAE latents.

2 Symbolic Planning for Full-Song Generation

2.1 From a Symbolic Composition to Audio

Let y denote the style description and lyrics, s a score, c semantic music tokens, and z continuous acoustic latents. YuE2 jointly models these three representations in one backbone and factorizes generation as

$$p_\theta(s, c, z \mid y) = p_{\theta,\mathrm{AR}}(s \mid y)\, p_{\theta,\mathrm{AR}}(c \mid y, s)\, p_{\theta,\mathrm{NAR}}(z \mid y, s, c). \tag{1}$$

The score s specifies melody, harmony, rhythm, and form in a readable, editable representation, while leaving performance details open. The semantic tokens c form a compact, discrete representation of recorded music. Their training targets are obtained by quantizing features from our MERT2 audio encoder. Conditioned on the score and text, YuE2 predicts these tokens autoregressively at 25 Hz. The acoustic generator then uses y, s, and c to produce continuous latents $z \in \mathbb{R}^{T\times 64}$ at the same rate. A separately trained variational autoencoder (VAE) (Kingma & Welling, 2013) decoder converts these latents into 48-kHz stereo audio.

The score is serialized in ABC, a text-based music notation, and tokenized with byte-pair encoding (Sennrich et al., 2016). Headers specify tempo, meter, and key; barlines and section labels organize time and form; chord symbols specify harmony; vocal and instrumental voices specify melody. The autoregressive stream generates the complete score before the semantic sequence, with boundary markers separating the two spans. ABC shares the text tokenizer, whereas semantic codes occupy a disjoint 32,768-token range. Type-specific output masks separate the two vocabularies; ABC is sampled without grammar constraints.

2.2 Learning and Realizing the Plan

One backbone, two information flows. Score and semantic tokens are predicted sequentially; acoustic generation can use the completed plan and bidirectional context. The 28-layer, 3.58B-

parameter backbone supports both computations through separate AR and NAR normalization, attention projections, and multilayer-perceptron experts in each layer (Figure 2). Discrete positions attend causally to preceding text, score, and semantic tokens and cannot read acoustic targets. Acoustic positions attend to all conditioning tokens and one another. Their inputs combine projected noisy latents with flow-time and frame-position embeddings. A linear head predicts the flow velocity. At inference, this velocity guides successive updates from Gaussian noise to acoustic latents, with all frames updated in parallel at each step.

Supervision from recordings. We introduce MERT2, an audio encoder, and SheetSage2, a full-song transcriber, to construct supervision from recordings without pre-existing aligned scores. SheetSage2-AR (hereafter SheetSage2) uses MERT2's full-context representations to recover melody, chords, rhythm, and form as a readable score. A separate MERT2 tokenizer branch uses causal self-attention to produce 25-Hz semantic tokens. These scores and tokens, together with continuous latents from a separately trained VAE, provide YuE2's offline training targets. The corpus consists primarily of CC0 music and licensed synthetic data. SheetSage2 achieves the highest scores on 12 of 15 benchmark–metric pairs in our full-song transcription comparison. Appendix C details the annotation and tokenizer models and their benchmark performance.

Objectives and training. Next-token cross-entropy $\mathcal{L}_{\mathrm{AR}}$ supervises the score and semantic spans. For conditional flow matching (Lipman et al., 2022), let z_0 be the clean acoustic latent, $\epsilon \sim \mathcal{N}(0, I)$, and $z_t = (1-t)z_0 + t\epsilon$ for flow time $t \in [0, 1]$. We minimize

$$\mathcal{L} = 0.25\mathcal{L}_{\mathrm{AR}} + \underbrace{\frac{1}{64|\mathcal{Z}|}\sum_{i\in\mathcal{Z}} \|v_\theta(z_t, t, y, s, c)_i - (\epsilon_i - z_{0,i})\|_2^2}_{\mathcal{L}_{\mathrm{FM}}}, \tag{2}$$

where $\mathcal{Z}$ contains non-padding acoustic positions. We train on approximately 346,000 hours of music. We pretrain the AR model from random initialization for 60,000 updates, then initialize the NAR experts from their AR counterparts. Joint training uses 40,000 updates, followed by 6,000 annealing updates. Joint training and annealing use whole-song packing in a 24,576-position context. In both phases, we mix four tasks in equal proportions: the full sequence, omission of the score, omission of semantic tokens, or omission of both. All four retain acoustic targets, so the same checkpoint learns generation both with and without symbolic planning. Text and lyrics may be dropped separately or together for classifier-free guidance. Appendix B gives the optimization settings.

2.3 Creation, Editing, and Cover Generation

All three operations use the same learned score-to-audio mapping; only the source of the score changes. **Creation** generates s from style and lyrics, then c and z. **Editing** fixes a revised score $\tilde{s}$ as the prefix and regenerates c and z for the entire song. **Cover generation** supplies a score transcribed from a reference recording and a new style description. Learning this mapping from individual recordings therefore enables covers without training on original–cover pairs.

3 Experiments

We evaluate song quality and the effects of unified modeling and symbolic planning, then test score–audio agreement, selective editing, and zero-shot covers.

3.1 Evaluation Setup

WildSongBench (WSB) contains 192 in-the-wild song-generation requests across 15 genre categories. Automatic metrics assess song quality, prompt adherence, and lyric accuracy. The standard comparison selects one of two candidates per prompt by lyric accuracy, using the lower phoneme error rate (PER). Best-of-8 selects from eight candidates, prioritizing SongBench Musicality, then Qwen3-Omni (Xu et al., 2025) prompt adherence (Q3O), then PER. Both reported YuE2 settings use melody-and-chord planning.

We use expert listening as the primary evidence for perceived quality. The Arena panel comprises 62 paid annotators with conservatory training or audio AI research backgrounds and contributes 3,602

retained pairwise evaluations across the system, planning, and architecture comparisons. Audio pairs have hidden identities and randomized A/B order; experts listen to at least 30 seconds per song and pass consistency checks. Overall quality judges the whole song; musicality judges composition and artistic appeal; audio quality judges fidelity and mixing. Preference shares retain ties and exclude unable-to-judge responses. Arena tests and confidence intervals use cluster-robust (CR1) standard errors clustered by evaluator and song; Appendix E.7 gives the complete protocol and criterion-specific counts.

3.2 Frontier Song Quality with Symbolic Planning

Automatic evaluation. Table 1 compares publicly released and proprietary systems in separate subtables on 192 WSB prompts. SongBench Global Avg and SongEval (Yao et al., 2025) Avg average seven and five quality dimensions, respectively. AudioBox Aesthetics (Tjandra et al., 2025) scores production quality (PQ); MuQ-MuLan (Zhu et al., 2025) and AllMusicCaps (Alonso-Jiménez et al., 2026) score alignment with lyric-free style descriptions. Q3O measures prompt adherence; lower PER indicates better lyric accuracy. SongBloom (Yang et al., 2025) uses a fixed held-out 10-second audio prompt because it cannot accept style text.

Table 1: **YuE2 leads public models on SongBench and competes with proprietary systems.** WSB, 192 prompts. Both YuE2 settings use symbolic planning. Standard: lower PER of two candidates; best-of-8 (Bo8): Musicality, Q3O, then PER. Mus.: Musicality; AMCaps: AllMusicCaps; MM: MiniMax. Bold/underline: best/second-best means within each subtable.

(a) Publicly released models

	SongBench		SongEval		AudioBox	Text alignment			Lyrics
Model	Mus.↑	Avg↑	Mus.↑	Avg↑	PQ↑	MuLan↑	AMCaps↑	Q3O↑	PER↓
YuE1 (Yuan et al., 2025)	4.0847	4.9165	3.1524	3.2150	7.8683	0.2623	0.2882	3.7301	0.3638
SongBloom (Yang et al., 2025)	3.4493	4.2350	3.2048	3.2051	8.1539	0.2697	0.1926	3.0287	0.1919
LeVo 2 (Lei et al., 2026)	5.4590	6.3247	3.9819	4.0234	**8.3966**	0.3542	0.2680	3.9458	0.2612
ACE-Step 1.5 (Gong et al., 2026)	5.1588	6.0118	3.8051	3.8465	8.0518	0.4372	0.3869	4.5809	0.0746
HeartMuLa† (Yang et al., 2026)	5.4963	6.2483	**4.5329**	**4.5519**	8.2933	0.3823	0.2786	3.4907	0.1071
DiffRhythm 2 (Jiang et al., 2025b)	4.4775	5.2428	3.4711	3.5245	7.9782	0.3782	0.3255	4.0870	0.1841
Muse (Jiang et al., 2026)	5.1692	6.0349	3.7362	3.7887	8.0517	0.3937	0.3466	4.4038	0.3342
MM Music 3 (MiniMax, 2026b)	5.3482	6.2830	4.0431	4.0994	8.2825	0.3928	0.3609	4.4362	**0.0627**
YuE2	**5.9075**	**6.7316**	4.2151	4.2625	8.2598	**0.5068**	**0.4054**	**4.6819**	0.0844

(b) Proprietary systems

	SongBench		SongEval		AudioBox	Text alignment			Lyrics
Model	Mus.↑	Avg↑	Mus.↑	Avg↑	PQ↑	MuLan↑	AMCaps↑	Q3O↑	PER↓
Suno v5 (Suno, 2025)	5.9918	6.8721	4.3051	4.3579	8.1698	**0.5428**	**0.4353**	4.5907	0.0810
Suno v4.5 (Team Suno, 2025)	5.8317	6.6995	4.3198	4.3666	8.2541	0.5022	0.3873	4.4149	**0.0580**
Suno v5.5 (Shulman, 2026)	5.8087	6.7150	4.1497	4.2152	8.1955	0.5089	0.3917	4.5914	0.0596
Suno v6 (Suno, 2026)	5.6558	6.5562	4.2635	4.3086	8.1296	0.4916	0.4305	4.6258	0.0758
Suno v6 Wild (Suno, 2026)	5.5644	6.4195	4.1716	4.2199	8.1785	0.4999	0.4316	4.5898	0.0745
MM Music 2.6 (MiniMax, 2026a)	5.4437	6.3222	4.0560	4.0985	8.1711	0.4251	0.3670	4.5688	0.2455
Mureka 9 (Mureka, 2026)	6.0488	6.9377	**4.3619**	**4.4111**	8.0226	0.4394	0.4102	4.6368	0.1169
YuE2	5.9075	6.7316	4.2151	4.2625	8.2598	0.5068	0.4054	4.6819	0.0844
YuE2 (Bo8)	**6.2666**	**6.9632**	4.2506	4.2960	**8.2714**	0.5051	0.3980	**4.7009**	0.0979

†HeartMuLa uses SongEval and AudioBox in post-training. Appendix E.3 details evaluator reuse and selection.

Leading public models. With two candidates, YuE2 leads evaluated public systems on SongBench Musicality and Global Avg (6.7316; LeVo 2: 6.3247), MuLan, AllMusicCaps, and Q3O (Table 1a).

Competitive with proprietary systems. Best-of-8 achieves the highest observed SongBench Global Avg (6.9632 versus 6.9377 for Mureka 9; Table 1b). Both YuE2 settings exceed every evaluated proprietary system on AudioBox PQ, a metric unused for selection.

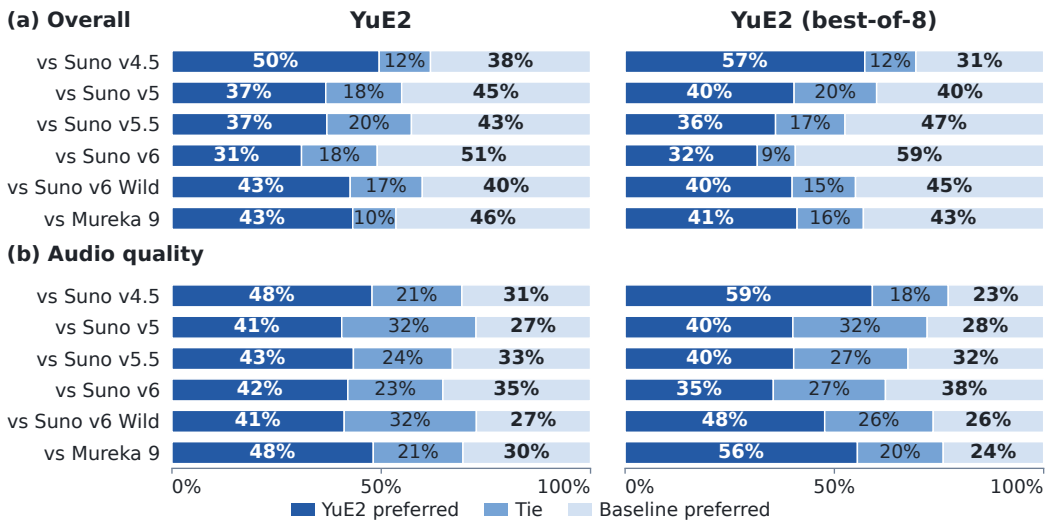

Figure 3: **Competitive song quality in expert comparisons with proprietary generators.** Experts assess overall song quality (top) and audio fidelity and mixing (bottom). Both YuE2 settings use melody-and-chord planning with two candidates (left) or eight (right). Bars show win/tie/loss percentages among judged responses; Appendix E.7.2 gives counts and CR1 intervals.

Expert preferences. Experts judge the selected outputs in Figure 3. They favor best-of-8 over Suno v4.5 for overall quality (57.3% versus 30.5%) and give nearly balanced preferences against Suno v5 (40.4% versus 39.9%). Suno v6 receives more preferences (59.3% versus 31.6%), despite its lower SongBench Avg. Audio quality is a particular strength: best-of-8 receives an average preference of 58.9% across six proprietary baselines (CR1 95% CI 55.6–62.2%), with ties split equally between systems and each baseline weighted equally. These results establish competitiveness with the evaluated proprietary systems.

Unified versus separate generation. To test whether learning discrete prediction and acoustic generation together improves song quality, we compare YuE2's unified backbone with a separate language model and diffusion Transformer (LM+DiT) following Qwen-Music (Xu et al., 2026). Both use melody-and-chord planning and the same representations. Experts prefer the unified model for overall quality: 53.4% versus 35.6% ($p = 0.0084$). The systems share training-data volume, prompts, and selection; parameter counts are 3.58B and 3.4B, respectively. AR token budgets, batch sizes, and update counts match, and the DiT receives 900,000 updates. Appendix D.1 gives the training comparison and results for all six listening criteria.

3.3 Symbolic Planning Improves Perceived Song Quality

The same checkpoint supports generation with and without symbolic planning. Without planning, the model generates semantic tokens from style and lyrics, then acoustic latents. With planning, it first generates a melody-and-chord score that conditions both subsequent stages. We hold prompts, lyrics, the two-candidate budget, selection rule, and decoder fixed.

Figure 4 shows a preference for planning: 49.3% versus 34.6% for overall quality ($p = 0.0070$), and 45.0% versus 29.4% for musicality ($p = 0.0141$), using two-sided CR1 tests. Preferences also favor planning for melody and chord progression.

Musical development in a paired example. Figure 5 illustrates how recognizable melodic variation connects verse and chorus in one pair. The full scores also show instrumental fills between vocal phrases, varied rhythmic phrasing, and a more extended ending with planning (Appendix G). These qualitative observations complement the aggregate listening results above.

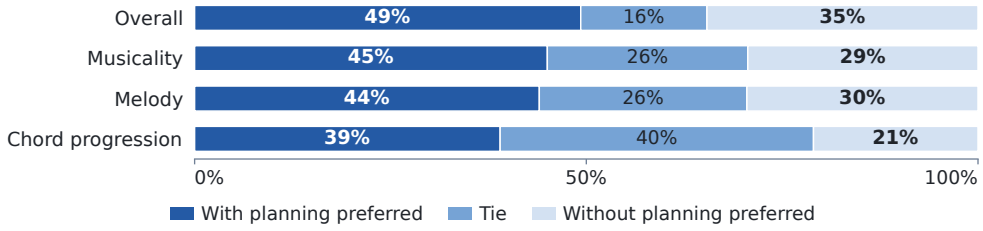

Figure 4: **Symbolic planning improves perceived song quality.** Expert win/tie/loss (%): 211 judgments each for overall quality and musicality; 200 each for melody and chord progression. Listeners come from the same expert pool (Appendix E.7).

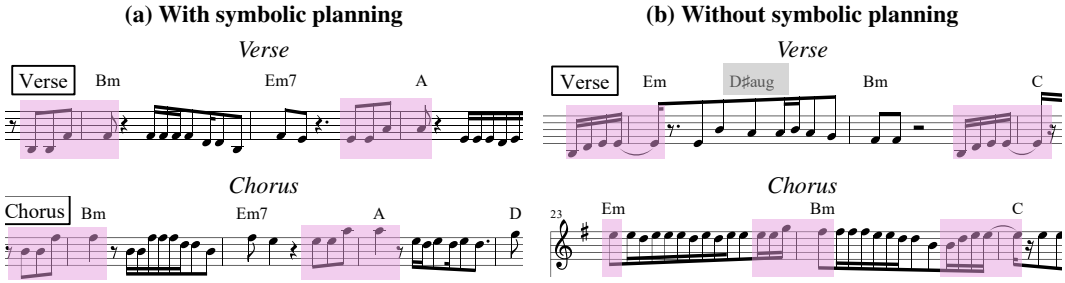

Figure 5: **Planning develops a recurring melodic idea in this paired example.** Both songs use the same prompt and lyrics. Read each staff from left to right; higher notes indicate higher pitch, and letters above the staff name the accompanying chords. Purple marks phrases with repeated lyric openings. **With planning (left),** the verse varies a low–low–high–high note pattern; the chorus recalls it an octave higher with longer phrase endings, linking the sections through recognizable variation. **Without planning (right),** the verse repeats the same B–D–E–E pattern, with less variation and a less direct melodic connection to the chorus. Gray marks a chord transition where experts hear the bass descent break, weakening harmonic continuity. The left score starts from the generated plan and the right from a SheetSage2 transcription; experts corrected both against the completed recordings for display, without regenerating audio. Appendix G gives the full scores and further analysis.

3.4 The Planned Composition Is Realized in Audio

For the score to serve as an editing interface, its musical decisions must survive audio generation. On WSB, we compare two scores generated from the same text prompt and the recordings produced from them. For both melody and chords, the correct score–audio pairing has higher total similarity than the swapped pairing in every evaluable case. Appendix F.1 gives the setup, exclusions, and metric definitions.

For each WSB prompt, we retain two generated scores, one recording from each, and two recordings generated without planning, without quality-based selection. A frozen SheetSage2 analyzer transcribes every recording; music experts check and correct the transcriptions. Each score is compared with its own recording, the recording from the other score for the same prompt, and a recording generated without a score. The two controls test whether agreement reflects the particular score rather than the shared text prompt.

Melody and chord similarity are one minus normalized edit distance, using sounding pitches and root-and-triad labels, respectively; these scores compare event content independently of timing. Table 2 reports prompt-averaged agreement for the three conditions. Corresponding recordings have much higher agreement than either control.

The model also follows temporal and structural decisions. Rhythm F1 compares consecutive vocal onset intervals in beats; tempo reports absolute log BPM error and accuracy within 8% of the target. Key scores measure exact tonic-and-mode agreement and give partial credit for related keys. Form compares section roles, order, and content; boundary F1 matches section starts within two beats after one global time shift. Corresponding audio achieves rhythm F1 of 0.9396, tempo accuracy

Table 2: **Audio realizes the particular planned composition on WSB.** Mismatched audio realizes the other score under the same prompt and lyrics. Rhythm reports vocal onset-interval F1. Values are prompt means. Tempo accuracy allows 8% error; lower log error is better.

	Melody	Chords	Rhythm	Key		Tempo		Form	
Audio	Seq. ↑	Seq. ↑	F1 ↑	Exact ↑	Wtd. ↑	Log err. ↓	Acc. ↑	Content ↑	Bound. ↑
Corresponding	**0.9464**	**0.9246**	**0.9396**	**0.9307**	**0.9515**	**0.0142**	**0.9869**	**0.7799**	**0.8724**
Mismatched	0.2160	0.2473	0.6119	0.3004	0.3629	0.1131	0.6806	0.1547	0.1454
Without score	0.2055	0.1897	0.5875	0.2234	0.2831	0.1468	0.5366	0.1236	0.1245

of 0.9869, and section-boundary F1 of 0.8724, exceeding both controls throughout the table. The melody and harmony advantage persists when allowing a song-wide pitch shift, extending beyond key and register differences (Appendix F.1).

3.5 Score Edits Change the Requested Musical Content

Score–audio agreement shows that generated music follows its plan. We now intervene on the score to test whether the model can change the requested musical content while preserving unedited melody and harmony. Despite regenerating the whole song, local edits preserve 90.16–94.34% of unedited melody and harmony. They attain 84.17% target-pitch accuracy, 79.54% target-chord agreement, and 73.43% relative onset accuracy (Table 3).

Table 3: **Score editing on WSB.** Scores range from 0 to 100; higher is better.

Edit	Adherence	Preservation	
		Melody	Harmony
Melody	84.17	93.10	94.05
Harmony	79.54	93.43	90.16
Rhythm	73.43	94.34	93.08
Key	90.58	—	—
Tempo	95.68	—	—

We edit each WSB prompt's first generated score. Melody, harmony, and rhythm edits target up to four bars of the first chorus; harmony substitutions also extend to matching phrases. Key and tempo edits are global and have no preservation score. We hold lyrics, style, model, decoder, and sampling settings fixed. Each applicable condition uses two seeds; all recordings are retained, including unedited controls. Appendix F.2 gives edit eligibility, sample counts, and the full protocol.

Table 3 separates adherence to the requested edit from preservation of unedited melody and harmony in the source score. We assess musical content using transcriptions from the SheetSage2 family, reviewed and corrected by music experts; madmom (Böck et al., 2016) estimates tempo. Weighted key accuracy gives partial credit for related keys (Yuan et al., 2023), and tempo Acc2 (Schreiber et al., 2020) also accepts estimates at related beat rates, such as half or double the target tempo. Missing or ambiguous edited events count as adherence misses; preservation uses evaluable comparisons. We average seeds and edit strengths within songs, then weight songs equally.

SongBench Global Avg remains close to the unedited controls: 6.64–6.73 across melody, harmony, rhythm, and key edits. Controls score 6.70 for melody, harmony, and key; the rhythm comparison uses its own matched source scores and seeds, with a control mean of 6.73. Appendix F.2 reports coverage, quality, and lyric retention.

3.6 Zero-Shot Covers Preserve Musical Identity Across Styles

Beyond self-generated scores, the same mapping also supports covers: we supply scores transcribed from existing recordings to test whether musical identity is preserved under a new style. We evaluate the same checkpoint on 948 eligible works from the SHS100K test partition (NovaFrost, 2018), without original–cover training pairs or cover-specific fine-tuning. Work-level checks exclude these works and their alternate performances from training. We render each source score with its lyrics in two contrasting styles using two seeds per style, retaining all 3,792 outputs per method. Using each generated cover as the query, CLEWS (Serrà et al., 2025) and Discogs-VINet (Araz et al.,

2024) search a gallery of 10,546 recordings, with the exact source recording excluded. Retrieval of other performances of the source work tests whether the generated cover preserves its musical identity across the change in style. We evaluate four retrieval metrics per encoder. MuLan measures target-style alignment; PQ and Mus. denote production quality and musicality.

We compare SongEcho (Li et al., 2026) and ACE-Step 1.5 through their native cover interfaces, and remove chords or the entire score within YuE2 (Table 4). Without cover-specific training, full-score YuE2 leads both evaluated cover systems on all eight work-identity retrieval measures, production quality, and Musicality. Removing chords lowers retrieval mean average precision (mAP). Removing the entire score reduces CLEWS mAP from 0.647 to 0.006 and Discogs-VINet mAP from 0.288 to 0.004, even though the source lyrics and target style remain unchanged. Relaxing score conditioning raises target-style alignment and automatic quality. Appendix E.6 gives the full comparison.

Our qualitative assessment with music experts links this pattern to harmony's role in musical style: retaining source harmony favors fidelity to the original work, while relaxing it allows greater adaptation to the target style. The cover task uses a fixed source score; in the planning preference study, YuE2 composes one for the requested style. These results motivate two ways to adapt a source composition: relaxing harmonic conditioning during cover generation, or explicitly revising the score through agentic music editing.

Table 4: **Zero-shot cover results.** CLEWS and VINet report mAP. Higher is better; bold and underlining mark the best and second-best means.

Method	Retrieval		Alignment	Quality	
	CLEWS	VINet	MuLan	PQ	Mus.
SongEcho	0.419	0.122	0.366	6.862	3.286
ACE-Step 1.5	0.024	0.006	0.166	6.918	3.689
YuE2 (full score)	**0.647**	**0.288**	0.382	8.044	5.104
Without chords	<u>0.598</u>	<u>0.179</u>	<u>0.417</u>	<u>8.117</u>	<u>5.490</u>
Without score	0.006	0.004	**0.474**	**8.186**	**5.691**

3.7 Agentic Music Editing

ABC represents the composition as text that an external language-model agent can read and revise. The agent translates user feedback into changes to the score, style, and lyrics, then invokes the same YuE2 checkpoint to render a new complete song. In *The Last Train*[1], a researcher and agent rework a Mandarin pop song as English jazz. When the researcher finds a transition abrupt, the agent revises the bridge lyrics, slows the notated tempo, and writes a saxophone interlude built around *Twinkle, Twinkle, Little Star*. The researcher directs the revisions and selects the final recording by listening. The score is both the plan that guides YuE2 and the composition that the user and agent continue to write together (Appendix F.3).

Related work. Music ControlNet (Wu et al., 2023) and Joint Audio and Symbolic Conditioning (JASCO) (Tal et al., 2024) condition audio on supplied musical controls; Qwen-Music (Xu et al., 2026) conditions covers on relative vocal pitch contours from a reference recording. MusiCoT (Lam et al., 2025) plans quantized audio–text embeddings; ACE-Step 1.5 (Gong et al., 2026) plans tempo, key, duration, and structure. Music Transformer (Huang et al., 2018) and SymPAC (Chen et al., 2024) generate symbolic music, while MIDI-DDSP (Wu et al., 2021) renders expressive instrumental performances. BACH (Wang et al., 2025) generates ABC scores for external vocal and accompaniment synthesis. Seed-Music (Bai et al., 2024) outlines a proprietary lead-sheet-to-song pipeline but does not isolate the quality benefit of symbolic planning. YuE2 jointly models readable full-song melody-and-chord scores, semantic tokens, and acoustic latents in one backbone. Within-checkpoint listening comparisons establish the quality benefit of planning.

Conclusion. YuE2 unifies symbolic composition and audio generation at frontier song quality. Within-checkpoint planning improves perceived quality; the same score-to-audio mapping supports selective editing, zero-shot covers, and continued revision by users and language-model agents. Limitations appear in Appendix A.

[1]Editing conversation, scores, and audio: https://yue2-0001.github.io/#agentic-music-editing.

AI USE STATEMENT

We used large language models to assist with code implementation, literature review, manuscript organization, and editing for clarity. The research also uses synthetic music and model-generated annotations: Qwen3-Omni supplies musical and vocal attributes, and Qwen3-ASR (Shi et al., 2026) transcribes lyrics; the learned transcription and tokenization models construct the symbolic and semantic training targets described in the paper. Model-based evaluation and candidate selection are identified in the experimental protocols. An external language-model agent assists with composition revisions in the qualitative editing case. We check AI-assisted implementations and results against source code, configurations, and evaluation records, and review the manuscript and citations. The authors are responsible for the final work's accuracy and integrity, including AI-assisted content.

ETHICS STATEMENT

Training primarily uses CC0 music and synthetic data, with most synthetic training data supplied under license. The listening campaign includes a paid expert panel and anonymous online participation sessions; reported Arena figures use expert responses. The protocol hides system identities, randomizes presentation order, and applies consistency checks. We report aggregate judgments rather than identifying participants. Song generation, cover generation, and editing can facilitate reuse of compositions, imitation of performances, and misleading attribution. Retrieval and quality metrics do not determine permission to reuse a work or a voice. Users must consider applicable rights, consent, and attribution when supplying recordings or distributing generated music.

REPRODUCIBILITY STATEMENT

We will publicly release WildSongBench and model weights for YuE2, MERT2, and SheetSage2, and open-source the evaluation code for reproducing the reported WildSongBench scores.

Section 2 specifies the representation sequence, attention structure, objectives, and training mixture. Appendix B gives architecture and optimization settings; Appendix C documents target construction and its supporting models. Section 3 states the main evaluation and selection protocols, with metric definitions, counts, and CR1 statistics in the evaluation appendices. The anonymous demo linked in the abstract provides audio examples.

REFERENCES

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Muhammad Waseem Akram, Stefano Dettori, Valentina Colla, and Giorgio Carlo Buttazzo. ChordFormer: A conformer-based architecture for large-vocabulary audio chord recognition. *arXiv preprint arXiv:2502.11840*, 2025.

Pablo Alonso-Jiménez, Xavier Lizarraga-Seijas, Xavier Serra, and Dmitry Bogdanov. AllMusicCaps: Album reviews as complementary supervision for music CLAP. *arXiv preprint arXiv:2608.25244*, 2026.

R. Oguz Araz, Xavier Serra, and Dmitry Bogdanov. Discogs-VI: A musical version identification dataset based on public editorial metadata. *arXiv preprint arXiv:2410.17400*, 2024.

Ye Bai, Haonan Chen, Jitong Chen, Zhuo Chen, Yi Deng, Xiaohong Dong, Lamtharn Hantrakul, Weituo Hao, Qingqing Huang, Zhongyi Huang, Dongya Jia, Feihu La, Duc Le, Bochen Li, Chumin Li, Hui Li, Xingxing Li, Shouda Liu, Wei-Tsung Lu, Yiqing Lu, Andrew Shaw, Janne Spijkervet, Yakun Sun, Bo Wang, Ju-Chiang Wang, Yuping Wang, Yuxuan Wang, Ling Xu, Yifeng Yang, Chao Yao, Shuo Zhang, Yang Zhang, Yilin Zhang, Hang Zhao, Ziyi Zhao, Dejian Zhong, Shicen Zhou, and Pei Zou. Seed-Music: A unified framework for high quality and controlled music generation. *arXiv preprint arXiv:2409.09214*, 2024.

Sebastian Böck, Filip Korzeniowski, Jan Schlüter, Florian Krebs, and Gerhard Widmer. madmom: A new python audio and music signal processing library. In *Proceedings of the ACM Conference on Multimedia*, pp. 1174–1178, 2016. doi: 10.1145/2964284.2973795.

Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra. The MTG-Jamendo dataset for automatic music tagging. In *Machine Learning for Music Discovery Workshop, International Conference on Machine Learning*, Long Beach, California, United States, 2019. URL https://hdl.handle.net/10230/42015.

Haonan Chen, Jordan B. L. Smith, Janne Spijkervet, Ju-Chiang Wang, Pei Zou, Bochen Li, Qiuqiang Kong, and Xingjian Du. SymPAC: Scalable symbolic music generation with prompts and constraints. *arXiv preprint arXiv:2409.03055*, 2024.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-Audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Junyu Dai, Xinyue Fan, Weiqin Li, Xiangang Li, Yunjia Li, Bin Ma, Yukun Ma, Chongjia Ni, Yufei Shi, Biao Tian, Haoxu Wang, Menglin Wu, Jianwei Yu, Huaicheng Zhang, Han Zhao, Shengkui Zhao, and Haina Zhu. Pushing the frontier of full-song generation: Hierarchical autoregressive planning meets flow-matching rendering. *arXiv preprint arXiv:2607.20253*, 2026. URL https://arxiv.org/abs/2607.20253v3.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, Guang Shi, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Heinrich Dinkel, Zhiyong Yan, Yongqing Wang, Junbo Zhang, Yujun Wang, and Bin Wang. Scaling up masked audio encoder learning for general audio classification. In *Proceedings of Interspeech*, pp. 547–551, 2024. doi: 10.21437/Interspeech.2024-246.

Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *Proceedings of the International Society for Music Information Retrieval Conference*, 2022.

Omar Eldeeb and Martin E. Malandro. Barwise section boundary detection in symbolic music using convolutional neural networks. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 847–854, 2025. doi: 10.5281/zenodo.17811497.

Vsevolod Eremenko, Emir Demirel, Baris Bozkurt, and Xavier Serra. Audio-aligned jazz harmony dataset for automatic chord transcription and corpus-based research. In *Proceedings of the 19th International Society for Music Information Retrieval Conference*, pp. 483–490, 2018. doi: 10.5281/zenodo.1291834. URL https://zenodo.org/records/1291834.

FluidSynth Developers. FluidSynth: Software synthesizer based on the SoundFont 2 specifications. GitHub software repository, 2026. URL https://github.com/FluidSynth/fluidsynth/tree/9567aeea2df24db3cd69c9f7d8efc00ce22173fa.

Francesco Foscarin, Jan Schlüter, and Gerhard Widmer. Beat this! accurate beat tracking without DBN postprocessing. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 962–969, 2024. doi: 10.5281/zenodo.14877491.

Junmin Gong, Yulin Song, Wenxiao Zhao, Sen Wang, Shengyuan Xu, Jing Guo, and Xuerui Yang. ACE-Step 1.5: Pushing the boundaries of open-source music generation. *arXiv preprint arXiv:2602.00744*, 2026.

Masataka Goto, Hiroki Hashiguchi, Takuichi Nishimura, and Ryuichi Oka. RWC music database: Popular, classical and jazz music databases. In *Proceedings of the International Conference on Music Information Retrieval*, pp. 287–288, 2002. doi: 10.5281/zenodo.1416474.

Yicheng Gu, Junan Zhang, Jerry Li, Zhizheng Wu, and Lauri Juvela. Frequency-aware self-supervised music representation learning. *arXiv preprint arXiv:2606.25713*, 2026.

Chunbo Hao, Ruibin Yuan, Jixun Yao, Qixin Deng, Xinyi Bai, Yanbo Wang, Wei Xue, and Lei Xie. SongFormer: Scaling music structure analysis with heterogeneous supervision. *arXiv preprint arXiv:2510.02797*, 2025.

Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M. Dai, Matthew D. Hoffman, Monica Dinculescu, and Douglas Eck. Music transformer. *arXiv preprint arXiv:1809.04281*, 2018.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Eric J. Humphrey and Juan Pablo Bello. Four timely insights on automatic chord estimation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 673–679, 2015. doi: 10.5281/zenodo.1417549.

Changhao Jiang, Jiahao Chen, Zhenghao Xiang, Zhixiong Yang, Hanchen Wang, Jiabao Zhuang, Xinmeng Che, Jiajun Sun, Hui Li, Yifei Cao, Shihan Dou, Ming Zhang, Junjie Ye, Tao Ji, Tao Gui, Qi Zhang, and Xuanjing Huang. Muse: Towards reproducible long-form song generation with fine-grained style control. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 22492–22512. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1129.

Junyan Jiang. osu2017: A rhythm game–based dataset for beat tracking, chord recognition, and key estimation. Dataset, version osu2017, 2026. URL https://github.com/instr3/osu-dataset/tree/osu2017.

Junyan Jiang, Ke Chen, Wei Li, and Gus Xia. Large-vocabulary chord transcription via chord structure decomposition. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 644–651, 2019. URL https://archives.ismir.net/ismir2019/paper/000078.pdf.

Junyan Jiang, Bilei Zhu, and Gus Xia. The music labels itself: Content-based MIDI chord and key estimation without human labels. In *Late-Breaking/Demo Extended Abstracts, International Society for Music Information Retrieval Conference*, 2025a. URL https://ismir2025program.ismir.net/lbd_445.html.

Yuepeng Jiang, Huakang Chen, Ziqian Ning, Jixun Yao, Zerui Han, Di Wu, Meng Meng, Jian Luan, Zhonghua Fu, and Lei Xie. DiffRhythm 2: Efficient and high fidelity song generation via block flow matching. *arXiv preprint arXiv:2510.22950*, 2025b.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Peter Knees, Ángel Faraldo, Perfecto Herrera, Richard Vogl, Sebastian Böck, Florian Hörschläger, and Mickael Le Goff. Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 364–370, 2015. doi: 10.5281/zenodo.1414996. URL https://archives.ismir.net/ismir2015/paper/000246.pdf.

Yuexuan Kong, Gabriel Meseguer-Brocal, Vincent Lostanlen, Mathieu Lagrange, and Romain Hennequin. S-KEY: Self-supervised learning of major and minor keys from audio. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890222.

Max W. Y. Lam, Yijin Xing, Weiya You, Jingcheng Wu, Zongyu Yin, Fuqiang Jiang, Hangyu Liu, Feng Liu, Xingda Li, Wei-Tsung Lu, Hanyu Chen, Tong Feng, Tianwei Zhao, Chien-Hung Liu, Xuchen Song, Yang Li, and Yahui Zhou. Analyzable chain-of-musical-thought prompting for high-fidelity music generation. *arXiv preprint arXiv:2503.19611*, 2025.

Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 387–392, 2009. doi: 10.5281/zenodo.1417647.

Shun Lei, Yaoxun Xu, Zhiwei Lin, Huaicheng Zhang, Wei Tan, Hangting Chen, Jianwei Yu, Yixuan Zhang, Chenyu Yang, Haina Zhu, Shuai Wang, Zhiyong Wu, and Dong Yu. LeVo: High-quality song generation with multi-preference alignment. *arXiv preprint arXiv:2506.07520*, 2025. URL https://arxiv.org/abs/2506.07520v3.

Shun Lei, Huaicheng Zhang, Dapeng Wu, Yaoxun Xu, Lishi Zuo, Wei Tan, Hangting Chen, Guangzheng Li, Jianwei Yu, Zhiyong Wu, and Dong Yu. LeVo 2: Stable and melodious song generation via hierarchical representation modeling and progressive post-training. *arXiv preprint arXiv:2606.30642*, 2026.

Sifei Li, Yang Li, Zizhou Wang, Yuxin Zhang, Fuzhang Wu, Oliver Deussen, Tong-Yee Lee, and Weiming Dong. SongEcho: Towards cover song generation via instance-adaptive element-wise linear modulation. *arXiv preprint arXiv:2602.19976*, 2026.

Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. MERT: Acoustic music understanding model with large-scale self-supervised training. In *International Conference on Learning Representations*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

MiniMax. MiniMax Music 2.6: Four stories we want to tell, 2026a. URL https://www.minimax.io/news/music-26.

MiniMax. MiniMax Music 3, 2026b. URL https://github.com/MiniMax-AI/MiniMax-Music3.

Mureka. Mureka API Platform Changelog, 2026. URL https://platform.mureka.ai/docs/en/changelog.html. Official changelog entry for the enhanced Mureka 9 model, 9 April 2026.

Oriol Nieto, Matthew McCallum, Matthew E. P. Davies, Andrew Robertson, Adam Stark, and Eran Egozy. The Harmonix set: Beats, downbeats, and functional segment annotations of western popular music. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 565–572, 2019. doi: 10.5281/zenodo.3527870. URL https://archives.ismir.net/ismir2019/paper/000068.pdf.

NovaFrost. Metadata for SHS100K and codes. Dataset repository, 2018. URL https://github.com/NovaFrost/SHS100K2/tree/a81043a6c90f378480b3c519e7f4def3c528ff33.

Aurian Quelennec, Pierre Chouteau, Geoffroy Peeters, and Slim Essid. MATPAC++: Enhanced masked latent prediction for self-supervised audio representation learning. *arXiv preprint arXiv:2508.12709*, 2025.

Sunita Sarawagi and William W. Cohen. Semi-markov conditional random fields for information extraction. In *Advances in Neural Information Processing Systems*, volume 17, 2004. URL https://papers.nips.cc/paper_files/paper/2004/hash/eb06b9db06012a7a4179b8f3cb5384d3-Abstract.html.

Hendrik Schreiber, Julián Urbano, and Meinard Müller. Music tempo estimation: Are we done yet? *Transactions of the International Society for Music Information Retrieval*, 3(1):111–125, 2020. doi: 10.5334/tismir.43. URL https://transactions.ismir.net/articles/10.5334/tismir.43.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-1162.

Joan Serrà, R. Oguz Araz, Dmitry Bogdanov, and Yuki Mitsufuji. Supervised contrastive learning from weakly-labeled audio segments for musical version matching. *arXiv preprint arXiv:2502.16936*, 2025.

Noam Shazeer. GLU variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR technical report. *arXiv preprint arXiv:2601.21337*, 2026.

Mikey Shulman. Suno v5.5: More expressive. more you., 2026. URL https://suno.com/blog/v5-5.

Mohammad Soleymani, Michael N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pp. 1–6. Association for Computing Machinery, 2013. doi: 10.1145/2506364.2506365.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Suno. Introducing v5 – the world's best music model, 2025. URL https://suno.com/release-notes/introducing-v5-the-world-s-best-music-model.

Suno. What's new in v6?, 2026. URL https://help.suno.com/en/articles/13924801.

Or Tal, Alon Ziv, Itai Gat, Felix Kreuk, and Yossi Adi. Joint audio and symbolic conditioning for temporally controlled text-to-music generation. *arXiv preprint arXiv:2406.10970*, 2024. URL https://arxiv.org/abs/2406.10970.

Team Suno. Introducing v4.5, 2025. URL https://blog.suno.com/blog/introducing-v4-5.

David Temperley and Trevor de Clercq. Statistical analysis of harmony and melody in rock music. *Journal of New Music Research*, 42(3):187–204, 2013. doi: 10.1080/09298215.2013.788039. URL https://doi.org/10.1080/09298215.2013.788039.

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. *arXiv preprint arXiv:2502.05139*, 2025.

George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302, 2002. doi: 10.1109/TSA.2002.800560.

Tongxi Wang, Yang Yu, Qing Wang, and Junlang Qian. Via score to performance: Efficient human-controllable long song generation with bar-level symbolic notation. *arXiv preprint arXiv:2508.01394*, 2025.

Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Xianbin Gu, and Gus Xia. POP909: A pop-song dataset for music arrangement generation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pp. 38–45, 2020. URL https://archives.ismir.net/ismir2020/paper/000089.pdf.

Minz Won, Yun-Ning Hung, and Duc Le. A foundation model for music informatics. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2024. doi: 10.1109/ICASSP48485.2024.10448314.

Dapeng Wu, Shun Lei, Wei Tan, Guangzheng Li, Yunzhe Wang, Huaicheng Zhang, Lishi Zuo, and Zhiyong Wu. SongBench: A fine-grained multi-aspect benchmark for song quality assessment. *arXiv preprint arXiv:2604.25937*, 2026.

Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J. Bryan. Music ControlNet: Multiple time-varying controls for music generation. *arXiv preprint arXiv:2311.07069*, 2023. URL https://arxiv.org/abs/2311.07069.

Yusong Wu, Ethan Manilow, Yi Deng, Rigel Swavely, Kyle Kastner, Tim Cooijmans, Aaron Courville, Cheng-Zhi Anna Huang, and Jesse Engel. MIDI-DDSP: Detailed control of musical performance via hierarchical modeling. *arXiv preprint arXiv:2112.09312*, 2021.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, Baosong Yang, Bin Zhang, Ziyang Ma, Xipin Wei, Shuai Bai, Keqin Chen, Xuejing Liu, Peng Wang, Mingkun Yang, Dayiheng Liu, Xingzhang Ren, Bo Zheng, Rui Men, Fan Zhou, Bowen Yu, Jianxin Yang, Le Yu, Jingren Zhou, and Junyang Lin. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

Jin Xu, Kangdi Wang, Ruibin Yuan, Shun Lei, Xiong Wang, Xize Cheng, Xueyao Zhang, Yang Zhang, Yiheng Chen, Yongqi Wang, Yue Wang, Zhifang Guo, Zihan Liu, Zijian Lin, Dake Guo, Hangrui Hu, Lei Xie, Linhan Ma, Wei Xue, Wenxiang Guo, Xinfa Zhu, Xipin Wei, Yangze Li, Yuanjun Lv, Yuxuan Wang, Yunfei Chu, and Zhiyong Wu. Qwen-Music technical report. *arXiv preprint arXiv:2607.11699*, 2026.

Sarthak Yadav, Sergios Theodoridis, and Zheng-Hua Tan. AudioMAE++: Learning better masked audio representations with SwiGLU FFNs. In *2025 IEEE 35th International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2025. doi: 10.1109/MLSP62443.2025.11204339.

Chenyu Yang, Shuai Wang, Hangting Chen, Wei Tan, Jianwei Yu, and Haizhou Li. SongBloom: Coherent song generation via interleaved autoregressive sketching and diffusion refinement. *arXiv preprint arXiv:2506.07634*, 2025.

Dongchao Yang, Yuxin Xie, Yuguo Yin, Zheyu Wang, Xiaoyu Yi, Gongxi Zhu, Xiaolong Weng, Zihan Xiong, Yingzhe Ma, Dading Cong, Jingliang Liu, Zihang Huang, Jinghan Ru, Rongjie Huang, Haoran Wan, Peixu Wang, Kuoxi Yu, Helin Wang, Liming Liang, Xianwei Zhuang, Yuanyuan Wang, Dingdong Wang, Haohan Guo, Junjie Cao, Zeqian Ju, Songxiang Liu, Yuewen Cao, Heming Weng, and Yuexian Zou. HeartMuLa: A family of open sourced music foundation models. *arXiv preprint arXiv:2601.10547*, 2026.

Jixun Yao, Guobin Ma, Huixin Xue, Huakang Chen, Chunbo Hao, Yuepeng Jiang, Haohe Liu, Ruibin Yuan, Jin Xu, Wei Xue, Hao Liu, and Lei Xie. SongEval: A benchmark dataset for song aesthetics evaluation. *arXiv preprint arXiv:2505.10793*, 2025.

Ruibin Yuan, Yinghao Ma, Yizhi Li, Ge Zhang, Xingran Chen, Hanzhi Yin, Le Zhuo, Yiqi Liu, Jiawen Huang, Zeyue Tian, Binyue Deng, Ningzhi Wang, Chenghua Lin, Emmanouil Benetos, Anton Ragni, Norbert Gyenge, Roger Dannenberg, Wenhu Chen, Gus Xia, Wei Xue, Si Liu, Shi Wang, Ruibo Liu, Yike Guo, and Jie Fu. MARBLE: Music audio representation benchmark for universal evaluation. In *Advances in Neural Information Processing Systems*, 2023.

Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, Xinrun Du, Zhen Ye, Tianyu Zheng, Zhengxuan Jiang, Yinghao Ma, Minghao Liu, Zeyue Tian, Ziya Zhou, Liumeng Xue, Xingwei Qu, Yizhi Li, Shangda Wu, Tianhao Shen, Ziyang Ma, Jun Zhan, Chunhui Wang, Yatian Wang, Xiaowei Chi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Shansong Liu, Lingrui Mei, Peng Li, Junjie Wang, Jianwei Yu, Guojian Pang, Xu Li, Zihao Wang, Xiaohuan Zhou, Lijun Yu, Emmanouil Benetos, Yong Chen, Chenghua Lin, Xie Chen, Gus Xia, Zhaoxiang Zhang, Chao Zhang, Wenhu Chen, Xinyu Zhou, Xipeng Qiu, Roger Dannenberg, Jiaheng Liu, Jian Yang, Wenhao Huang, Wei Xue, Xu Tan, and Yike Guo. YuE: Scaling open foundation models for long-form music generation. *arXiv preprint arXiv:2503.08638*, 2025.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *arXiv preprint arXiv:1910.07467*, 2019.

Chuanxia Zheng and Andrea Vedaldi. Online clustered codebook. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22741–22750, 2023. doi: 10.1109/ICCV51070.2023.02084.

Haina Zhu, Yizhi Zhou, Hangting Chen, Jianwei Yu, Ziyang Ma, Rongzhi Gu, Yi Luo, Wei Tan, and Xie Chen. MuQ: Self-supervised music representation learning with mel residual vector quantization. *arXiv preprint arXiv:2501.01108*, 2025.

A LIMITATIONS

Automatic metrics can disagree with expert preferences: Suno v6 receives more overall preferences than YuE2 (best-of-8) despite its lower SongBench Global Avg. We do not systematically test how quality and planning gains vary with model size, training data, or training compute. Lyric accuracy and perceived quality across a broader range of languages remain untested.

B ARCHITECTURE AND TRAINING DETAILS

Training data and annotation. The generator uses approximately 346,000 hours of music (6.36 million full-song examples), primarily CC0 and licensed synthetic recordings. MERT2 foundation pretraining uses approximately 700,000 hours, and SheetSage2-AR transcription training uses approximately 28,400 hours. We adapt the automatic annotation pipeline of YuE (Yuan et al., 2025): Qwen3-Omni annotates musical and vocal attributes, SongFormer (Hao et al., 2025) supplies section boundaries and functional labels, and Qwen3-ASR transcribes lyrics. The resulting annotations provide text and lyric conditions; symbolic scores and semantic tokens are constructed by the models described below.

Table 5: **Main YuE2 architecture.**

Property	Value
Parameters	approximately 3.58B
Layers / hidden width / feed-forward width	28 / 2,048 / 6,144
Query heads / key–value groups / head width	16 / 8 / 128
Context length	24,576 positions
Text vocabulary	151,643 entries
Semantic vocabulary	32,768 entries
Acoustic representation	64 dimensions at 25 Hz
Normalization / MLP	RMSNorm (Zhang & Sennrich, 2019) / SwiGLU (Shazeer, 2020)
Training precision	bfloat16

Training tasks. During joint training and annealing, we mix four tasks within the same model (Table 6). The first two retain semantic tokens while varying whether a symbolic score is included, enabling comparison of generation with and without symbolic planning using the same checkpoint. Every task trains acoustic generation with flow matching; next-token loss supervises the included score and semantic tokens.

Table 6: **Four training tasks mixed within one YuE2 model.** The four tasks are sampled in a balanced ratio during joint training and annealing.

Training task	AR prediction targets	Acoustic conditioning
Full generation	Score, then semantic tokens	Text, lyrics, score, and semantic tokens
No symbolic planning	Semantic tokens	Text, lyrics, and semantic tokens
No semantic tokens	Score	Text, lyrics, and score
Direct audio generation	None	Text and lyrics

During training, score and semantic-token sequences derived from the target recording are supplied as conditioning prefixes. Text and lyrics may be dropped separately or together for classifier-free guidance.

Training schedule. We first pretrain an autoregressive model from random initialization and use its 60,000-update checkpoint to initialize the joint AR–NAR model. The NAR Transformer experts are initialized from their AR counterparts. We then perform 40,000 joint-training updates, followed by a 6,000-update annealing stage. We reset the optimizer when starting joint training and retain its state for annealing. Table 7 specifies the learning-rate schedules and context lengths for these three phases.

Table 7: **Training phases and optimization schedules for YuE2.**

Phase	Updates used	Max. context (positions)	Learning-rate schedule
AR pretraining	60,000	16,384	5,000-step linear warmup to 10^{-4}; cosine decay with a 10^{-5} floor at step 95,000.
Joint AR–NAR training	40,000	24,576	5,000-step linear warmup to 3×10^{-4}; cosine decay with a 10^{-4} floor at step 150,000.
Annealing	6,000	24,576	No warmup; cosine decay from approximately 2.726×10^{-4} to 3×10^{-5} over 6,000 updates.

Steps are counted from the start of each phase. The AR and joint-training checkpoints are taken before their scheduled cosine-decay endpoints.

Optimization and packing. All three phases use bfloat16, a global batch size of 256, gradient clipping at 1.0, and 64 H800 GPUs. Adam (Kingma & Ba, 2014) uses $\beta = (0.9, 0.98)$ and weight decay 0.05 during AR pretraining, and $\beta = (0.9, 0.95)$ and weight decay 0.1 during joint training and annealing. Songs are packed without temporal splitting, and attention and positional indices reset at document boundaries.

Cover supervision. Training uses aligned conditioning inputs and targets from individual recordings, without original–cover pair supervision. Cover evaluation uses the general song-generation checkpoint without cover-specific fine-tuning, on SHS100K works absent from its training corpus (Appendix E.6).

Inference interfaces. Creation generates a score; editing and cover generation supply a score as a fixed prefix.

Table 8: **Score conditioning for creation, editing, and cover generation.**

Operation	Symbolic score	Generated tokens and latents
Create	$s \sim p(s \mid y)$	$c, z \sim p(c, z \mid y, s)$
Edit	user-modified $\tilde{s}$	$\tilde{c}, \tilde{z} \sim p(c, z \mid y, \tilde{s})$
Cover	$\hat{s}$ transcribed from reference	$c', z' \sim p(c, z \mid y', \hat{s})$

y' combines the source lyrics with the target style description.

C Representation and Transcription Details

C.1 Constructing Symbolic and Semantic Supervision

Learning this generation hierarchy requires scores and semantic sequences aligned with the training audio. Ordinary song corpora lack this paired supervision. We construct it from recordings with complementary MERT2-based models: SheetSage2-AR transcribes the training scores, and a separate semantic tokenizer produces the dense music-token sequence.

Learning the audio representation. Different audio encoders emphasize different aspects of the same recording. Motivated by the hypothesis that independently learned representations can reveal complementary projections of shared underlying structure (Huh et al., 2024), we construct MERT2's prediction targets through offline *Multi-View Target Synthesis*. Time-aligned features from frozen MuQ (Zhu et al., 2025) and Qwen2-Audio-Instruct (Chu et al., 2024) encoders are compressed through one shared four-level residual vector-quantized bottleneck. Separate decoders reconstruct both source feature spaces from the same quantized state. We cache its four code streams as prediction targets. Stage 1, *Foundation Pretraining*, trains a bidirectional MERT2 encoder to predict these targets at masked frames.

Symbolic targets. For full-song transcription, Stage 2-FS, *Full-Song Adaptation*, continues pretraining with bidirectional context on complete recordings lasting 30–360 seconds, producing MERT2-FS. Its features support *SheetSage2-AR* (hereafter SheetSage2), an autoregressive transcriber that converts a recording into chronological events with beat timestamps quantized to 10 ms, for conversion to ABC lead-sheet notation. Its training targets are generated entirely by the non-autoregressive *SheetSage2-Prober*, which uses structured decoding to establish the beat grid before decoding the remaining musical attributes.

Semantic targets. A separate branch adapts the Stage 1 encoder to produce a compact semantic sequence for autoregressive prediction. Following Qwen-Music (Xu et al., 2026), *Causal Adaptation* aligns self-attention with this prediction order. Tokenization operates offline, combining causal self-attention with noncausal convolution and temporal normalization. *Supervised Fine-Tuning* adds lyric, mel, and chroma objectives, and *Semantic Quantization* learns the single 32,768-entry clustered vector quantizer (CVQ) used by YuE2. We adapt the online updates of CVQ-VAE (Zheng & Vedaldi, 2023) to move rarely selected entries toward current encoder features. Codebook utilization exceeds 99% on the validation set. Appendix C.3 gives the update rule and defines utilization; Appendix C reports the representation and transcription benchmarks and their evaluation scope.

C.2 MERT2 MARBLE Evaluation

We evaluate MERT2-30s and its full-song continuation, MERT2-FS, on MARBLE (Yuan et al., 2023). MERT2-30s is pretrained on 30-second excerpts; MERT2-FS continues pretraining on complete recordings lasting 30–360 seconds. Downstream tagging uses MARBLE's standard 30-second task excerpts. Table 9 reports the MARBLE results, and Table 10 gives the corresponding layer and learning rate. The baseline table contains MERT (Li et al., 2024), Dasheng (Dinkel et al., 2024), MuQ (Zhu et al., 2025), MusicFM (Won et al., 2024), AudioMAE++ (Yadav et al., 2025), MAT-PAC++ (Quelennec et al., 2025), and the M2D-Large and PupuJEPA models reported by Gu et al. (2026).

Table 9: MERT2-30s leads on 14 of 15 MARBLE (Yuan et al., 2023) metrics against the listed external baselines. Baseline scores are from Tables III and V of Gu et al. (2026) (arXiv v3). AudioMAE++, MATPAC++, and M2D-Large are that study's retrained models; the remaining external baselines use its evaluations of released checkpoints or its own PupuJEPA models. Results cover MagnaTagATune (MTT) (Law et al., 2009), GiantSteps (Knees et al., 2015), GTZAN (Tzanetakis & Cook, 2002), EmoMusic (Soleymani et al., 2013), and MTG-Jamendo (Bogdanov et al., 2019). MERT2-FS (full-song) continues pretraining on complete recordings lasting 30–360 seconds.

Dataset		**MTT**		**GiantSteps**	**GTZAN**		**EmoMusic**	
Task		Tagging		Key	Genre	Beat	Emotion	
Model	**# Params**	ROC↑	AP↑	$\text{Acc.}^{\text{Refined}}$↑	Acc.↑	F1^{beat}↑	R^2_{V}↑	R^2_{A}↑
MERT-Large (Li et al., 2024)	330M	90.6	37.9	64.1	77.6	86.8	56.7	76.1
Dasheng-1.2B (Dinkel et al., 2024)	1.2B	91.5	40.4	58.0	81.4	87.7	57.4	75.0
MuQ (Zhu et al., 2025)	310M	90.5	38.5	63.2	83.8	90.1	58.3	76.4
MusicFM (Won et al., 2024)	330M	90.9	38.3	63.0	84.1	90.2	57.2	74.4
AudioMAE++ (Yadav et al., 2025)	307M	91.2	39.5	61.7	80.3	90.0	59.0	75.7
MATPAC++ (Quelennec et al., 2025)	307M	90.6	38.2	63.7	81.4	90.1	57.8	74.7
M2D-Large (Gu et al., 2026)	307M	91.0	39.2	65.0	83.8	90.0	57.4	74.8
PupuJEPA-Large (Gu et al., 2026)	307M	91.7	40.8	66.1	86.9	**91.0**	62.5	76.8
PupuJEPA-Huge (Gu et al., 2026)	632M	91.3	39.7	64.8	85.9	90.5	62.0	<u>78.5</u>
MERT2-30s	632M	**91.91**	**41.29**	<u>66.97</u>	**91.72**	<u>90.59</u>	<u>63.23</u>	**80.01**
MERT2-FS	632M	<u>91.74</u>	<u>41.20</u>	**67.05**	<u>90.69</u>	90.57	**63.52**	78.14

Dataset		**MTG-Jamendo**							
Task		Instrument		Mood / theme		Genre		Top 50	
Model	**# Params**	ROC↑	AP↑	ROC↑	AP↑	ROC↑	AP↑	ROC↑	AP↑
MERT-Large (Li et al., 2024)	330M	75.5	18.8	75.3	13.5	86.1	18.0	82.6	29.1
Dasheng-1.2B (Dinkel et al., 2024)	1.2B	75.0	19.0	76.1	15.5	85.5	18.8	82.4	29.6
MuQ (Zhu et al., 2025)	310M	74.8	19.1	73.7	13.2	85.4	19.1	83.0	30.2
MusicFM (Won et al., 2024)	330M	74.6	18.5	74.9	14.1	85.3	19.4	81.9	29.7
AudioMAE++ (Yadav et al., 2025)	307M	77.1	19.9	75.6	14.0	86.3	18.9	83.1	31.1
MATPAC++ (Quelennec et al., 2025)	307M	77.2	19.7	75.1	14.1	85.7	19.6	82.5	30.2
M2D-Large (Gu et al., 2026)	307M	76.6	19.3	74.6	14.3	85.5	19.2	82.5	29.6
PupuJEPA-Large (Gu et al., 2026)	307M	<u>78.4</u>	21.2	76.2	15.3	86.1	20.1	82.8	30.5
PupuJEPA-Huge (Gu et al., 2026)	632M	77.6	20.5	75.9	14.7	85.9	20.1	83.1	30.7
MERT2-30s	632M	**80.27**	<u>22.89</u>	**79.44**	**16.68**	**88.01**	**21.22**	**84.18**	**32.17**
MERT2-FS	632M	**80.27**	**23.51**	<u>78.74</u>	<u>15.74</u>	<u>87.98</u>	<u>20.66</u>	<u>84.13</u>	<u>31.62</u>

Naming note. The authors renamed PupuJEPA to PupuM2D after the paper's initial release. We retain PupuJEPA.

MERT2-30s exceeds every listed external baseline on 14 of 15 MARBLE metrics across nine tasks. Both checkpoints exceed every listed external baseline on all ten MTT/MTG ROC/AP measures. MERT2-FS exceeds every listed external baseline on 13 of 15 metrics and remains within 1.87 points of MERT2-30s across all 15.

Table 10: **MARBLE probe hyperparameters for the reported MERT2 results.** L_k denotes the output of Conformer layer k (numbered 1–24). "All" concatenates the 24 layer representations and applies a learned linear projection to 1,024 dimensions.

	MERT2-30s		MERT2-FS	
Task	**Representation**	**Learning rate**	**Representation**	**Learning rate**
GTZAN genre	L23	5×10^{-3}	L24	5×10^{-4}
GTZAN beat	L21	1×10^{-3}	L23	1×10^{-3}
GiantSteps key	L4	1×10^{-3}	L23	1×10^{-3}
EmoMusic	All	5×10^{-5}	L24	5×10^{-4}
MTT	L22	1×10^{-3}	L23	1×10^{-3}
MTG Instrument	L14	1×10^{-3}	L12	1×10^{-3}
MTG MoodTheme	L16	1×10^{-3}	L13	1×10^{-3}
MTG Genre	L19	1×10^{-3}	L16	1×10^{-3}
MTG Top50	L13	1×10^{-3}	L22	1×10^{-3}

C.3 Semantic Tokenizer Retention

Pre-VQ and post-VQ denote the outputs of Conformer layers 12 and 14, respectively, from the same frozen tokenizer (zero-based indexing). Layer 14 is the first Conformer layer after quantization; both representations have 1,024 dimensions at 25 Hz. The 32,768-entry cosine quantizer uses 32-dimensional code vectors and operates at 25 Hz (375 bits/s). The quantized branch is fully active during evaluation.

Online codebook updates. Our cosine quantizer adapts CVQ-VAE (Zheng & Vedaldi, 2023) with $K = 32{,}768$ entries $e_j \in \mathbb{R}^{32}$. Let p_j be the exponential moving average of code j's assignment frequency, with decay $\rho = 0.99$. Each training step samples an anchor a_j from up to 2,048 encoder features gathered across workers, with probability determined by cosine similarity to e_j, then updates

$$e_j \leftarrow (1-\gamma_j)e_j + \gamma_j a_j, \qquad \gamma_j = \exp\left(-\frac{10Kp_j}{1-\rho} - 10^{-3}\right). \tag{3}$$

Usage is aggregated across workers and the updated codebook is synchronized. The loss combines codebook loss, commitment loss weighted by 0.5, and a contrastive term with temperature 0.07. Codebook utilization is the fraction of entries selected at least once over the validation pass at the selected checkpoint.

We compare the MERT2 tokenizer with LeVo 2's generation-time 16,384-entry residual vector quantization (RVQ) output (1,024 dimensions at 25 Hz; 350 bits/s) (Lei et al., 2026) and EnCodec's quantizer-decoded latent at 1.5 kbit/s (128 dimensions at 75 Hz) (Défossez et al., 2022). The five tasks are GTZAN genre, EmoMusic emotion, GiantSteps key, MagnaTagATune tagging, and GTZAN beat tracking. Within each task, every representation uses the same split, probe, optimization recipe, and seed (1234). For MagnaTagATune, we average the selected frame representations over each 30-second excerpt before classification and report macro AUROC and average precision. Emotion reports the mean R^2 across arousal and valence. Beat features are linearly interpolated from their native rates to MARBLE's 100-Hz label grid and evaluated with a 70-ms beat-F1 tolerance. The downstream input width follows each representation; all other probe settings remain fixed. Model selection uses validation AUROC for MTT and validation beat F1 for beat tracking.

Post/pre is the ratio of post-VQ to pre-VQ task-probe scores. The post-VQ state is strongest among the quantized representations on five of six reported metrics.

Table 11: **At 375 bits/s, the MERT2 tokenizer leads five of six metrics among the evaluated quantized representations.** Scores are multiplied by 100 (higher is better), with the best post-VQ result per column in bold. Emotion reports the mean R^2 across arousal and valence. Rate, feature dimension, and nominal bitrate characterize each representation; external controls are LeVo 2 (Lei et al., 2026) and EnCodec (Défossez et al., 2022).

	Token stream			Genre	Emotion	Key	MTT		Beat
Representation	Rate (Hz)	Dim.	kb/s	Acc. ↑	R^2 ↑	Score ↑	AUROC ↑	AP ↑	F1 ↑
MERT2 tokenizer pre-VQ	25	1024	–	84.14	66.19	56.94	91.60	40.02	89.81
MERT2 tokenizer post-VQ	25	1024	0.375	**65.86**	**54.49**	52.20	**90.06**	**35.65**	**86.38**
LeVo 2 post-VQ	25	1024	0.350	44.83	30.17	**61.23**	86.72	30.79	79.72
EnCodec post-VQ	75	128	1.500	31.72	26.51	15.76	80.84	22.32	76.50

C.4 SheetSage2-AR Benchmark Evaluation

Table 12 compares the validation-selected SheetSage2-AR checkpoint with SheetSage1 (Donahue et al., 2022), madmom (Böck et al., 2016), and task-specific systems. The comparison covers GTZAN (Tzanetakis & Cook, 2002), osu2017 (Jiang, 2026), GiantSteps (Knees et al., 2015), Chords1217 (Humphrey & Bello, 2015), Jazz Audio-Aligned Harmony (JAAH) (Eremenko et al., 2018), HarmonixSet (Nieto et al., 2019), RWC-Pop (Goto et al., 2002), and Rock Corpus (RS 200) (Temperley & de Clercq, 2013).

A single SheetSage2-AR model achieves the highest scores on 12 of the 15 benchmark–metric pairs. SheetSage2 substantially improves over SheetSage1 in melody, chord, key, and downbeat estimation, and surpasses task-specific models on several benchmarks. These transcription results support SheetSage2-AR as the source of full-song score annotations for YuE2. Cross-validation and training-overlap reference values are marked separately.

Table 12: **SheetSage2-AR: one model for full-song lead-sheet analysis.** The comparison includes SheetSage1 (Donahue et al., 2022), madmom (Böck et al., 2016), and task-specific systems cited beside their scores. SheetSage2-AR achieves the highest scores on 12 of 15 benchmark–metric pairs. Scores are percentages; higher is better.

Task	Benchmark	Metric	SheetSage1	Madmom	Prior specialist	SheetSage2-AR
Beat	GTZAN	F1↑	86.07	86.07	**89.01** (Foscarin et al., 2024)	86.27
	osu2017		91.80	91.80	89.19 (Foscarin et al., 2024)	**93.01**
Downbeat	GTZAN	F1↑	64.65	64.65	78.28 (Foscarin et al., 2024)	**80.45**
	osu2017		83.47	83.47	85.90 (Foscarin et al., 2024)	**92.90**
Key	GiantSteps	Score↑	43.89	74.62	72.09 (Kong et al., 2025)	**77.73**
	GTZAN		54.56	72.05	74.43 (Kong et al., 2025)	**75.77**
Chord	osu2017	Maj/min↑	79.82	77.42	86.55 (Akram et al., 2025)	**90.08**
	Chords1217		72.98	83.52*	**83.94**† (Akram et al., 2025)	83.81
	JAAH		39.77	51.24	59.45 (Akram et al., 2025)	**64.50**
Structure	HarmonixSet	Accuracy↑	—	—	80.03 (Hao et al., 2025)	**80.51**
		F1 (0.5 s)↑	—	—	**70.63** (Hao et al., 2025)	67.96
		F1 (3 s)↑	—	—	79.50 (Hao et al., 2025)	**82.86**
Melody	RWC-Pop	Vocal F1↑	62.71	—	62.71 (Donahue et al., 2022)	**82.51**
		Full F1↑	64.02	—	64.02 (Donahue et al., 2022)	**75.29**
	Rock Corpus	Vocal F1↑	49.19	—	49.19 (Donahue et al., 2022)	**67.08**

Bold marks the best value for each benchmark–metric pair; an em dash denotes unavailable results. Structure F1 measures section boundaries at the stated tolerance. Melody F1 uses a 50-ms onset tolerance and pitch classes. SheetSage1 rhythm entries reuse Madmom outputs; the melody specialist repeats SheetSage1. Beat This! scores average three released training seeds. *The madmom chord model was trained on data that include Chords1217. †ChordFormer uses five-fold evaluation, with one held-out model per track. Its original scorer pools metric-eligible durations; we recompute per-track recall and weight it by each reference span to match the other models' scoring protocol. SheetSage1 and SheetSage2-AR each evaluate one fixed checkpoint on all 1,217 tracks. Appendix C.5 gives the evaluation protocols and score sources.

C.5 SheetSage2-AR and SheetSage2-Prober

SheetSage2-AR builds on audio-to-lead-sheet transcription (Donahue et al., 2022) to jointly predict beat, downbeat, meter, key, chord, musical structure, and lead melody. Melody notes are assigned to vocal (Track 0) or instrumental (Track 1) voices. The benchmarked model is *SheetSage2-AR*; the separate *SheetSage2-Prober* produces its training annotations.

Autoregressive transcriber. SheetSage2-AR pairs a full-context MERT2-FS encoder with a six-layer RoFormer decoder (Su et al., 2024). The encoder backbone is frozen, with trainable low-rank adapters and a learned mixture of its layer outputs. It processes audio in windows of up to 300 seconds. Given encoder output $\mathbf{h}$ and event sequence $\mathbf{e} = (e_1, \ldots, e_N)$, the decoder factorizes the conditional distribution as

$$p(\mathbf{e} \mid \mathbf{h}) = \prod_{n=1}^{N} p(e_n \mid e_{<n}, \mathbf{h}). \tag{4}$$

The benchmark table reports the validation-selected checkpoint. All attributes share one discrete event vocabulary. Events are chronological, with a fixed order when several events have the same onset. The vocabulary is defined as follows.

- **Absolute beat time.** Each beat onset is represented by a 100-Hz-quantized timestamp such as ⟨time_8.48s⟩. Events between beats are located relative to the most recent timestamp.

- **Beat position and meter.** The token ⟨beat_1_of_4⟩ marks a downbeat. Other tokens, such as ⟨beat_4_of_4⟩, give the remaining metrical positions. Supported meters are $2/4$, $3/4$, $4/4$, $3/8$, and $6/8$.
- **Relative subbeat position.** A token such as ⟨subbeat_shift_0.5⟩ advances the event cursor by half a beat. Shifts are quantized to quarter-beat intervals, and no shift is emitted between simultaneous events.
- **Melody pitch and track.** A melody onset such as ⟨pitch_78_track_0⟩ contains a MIDI pitch and source track. The full-melody task includes both tracks; the vocal-melody task includes only Track 0.
- **Note duration.** A pitch may be followed by a quantized token such as ⟨duration_0.5⟩. Durations come from a fixed template; omitting the token denotes the shortest default duration.
- **Chord.** Chords encode root and quality, for example ⟨chord_F#:maj⟩. The vocabulary also includes a no-chord symbol. A chord is emitted when it changes and repeated at downbeats to anchor each measure.
- **Key.** Key changes use 24 root–mode combinations, such as ⟨key_F#:major⟩. The active key is emitted at the first beat and subsequently only when it changes.
- **Musical structure.** Categorical tokens such as ⟨structure_chorus⟩ mark the active section at the beginning of a window and at later segment changes.

Table 13 shows the ordering of timestamps, metrical positions, score attributes, melody events, durations, and relative shifts in a short decoded sequence.

Table 13: **Example event sequence decoded by SheetSage2-AR.**

```
<time_8.48s> <beat_4_of_4>                                          // beat
  <pitch_80_track_0> <duration_1.0> <subbeat_shift_1.0>
<time_9.00s> <beat_1_of_4>                                          // downbeat
  <structure_chorus> <chord_F#:maj>
  <pitch_82_track_0> <duration_0.5> <subbeat_shift_0.5>
  <pitch_73_track_0> <duration_0.5> <subbeat_shift_0.5>
<time_9.52s> <beat_2_of_4>                                          // beat
  <pitch_78_track_0> <duration_0.5>...
```

Generating aligned annotations. SheetSage2-AR requires multitask labels with a shared vocabulary and timing convention. Existing music information retrieval (MIR) datasets provide complementary annotations with differing coverage. We first train the non-autoregressive SheetSage2-Prober, fine-tuning MERT2 through low-rank adapters and task-specific prediction heads. Masking losses for unannotated attributes allows training on partially annotated recordings.

Prober supervision includes an expanded lead-sheet corpus based on the data used by SheetSage1 (Donahue et al., 2022), an expanded version of POP909 (Wang et al., 2020), the training split of HarmonixSet (Nieto et al., 2019), and MIDI-rendered audio from the MIDI AutoLabel Dataset (MALD) (Jiang et al., 2025a) and the Segmented Lakh MIDI Subset (SLMS) (Eldeeb & Malandro, 2025). MALD supplies automatically inferred chord and key annotations; SLMS supplies section boundaries. We render MIDI with FluidSynth (FluidSynth Developers, 2026) and the FluidR3 GM SoundFont.[2]

The prober predicts frame-level activations, with melody predicted on a subdivided beat grid. Task-specific conditional random field (CRF)-based structured decoding (Sarawagi & Cohen, 2004) combines these neural scores with temporal constraints to recover discrete labels: rhythm decoding first establishes beats, downbeats, and meter, then key, chord, structure, and melody are decoded on the resulting grid. We apply the prober to a separate corpus of real audio recordings to obtain labels

[2]Official distribution: https://member.keymusician.com/Member/FluidR3_GM/index.html.

with a common vocabulary and timing convention. SheetSage2-AR is trained exclusively on these generated labels using next-token cross-entropy: 441,094 training recordings (28,434 hours) and 884 validation recordings. A deterministic builder converts the decoded events to ABC, and a parser checks the exported notation.

Comparison with SheetSage2-Prober. Additional experiments compare SheetSage2-AR with SheetSage2-Prober under the same evaluation protocol (Table 14). Although SheetSage2-AR is trained exclusively on labels generated by SheetSage2-Prober, it achieves higher scores on 10 of 15 benchmark–metric pairs. SheetSage2-AR generates the attributes in one chronological event sequence using grammar-constrained autoregressive decoding, replacing the Prober's task-specific CRF decoders.

Table 14: **SheetSage2-AR compared with SheetSage2-Prober.** The models use the same benchmark cohorts and scoring protocol as Table 12. All scores are percentages; **bold** marks the higher value in each row.

Task	Benchmark	Metric	SheetSage2-Prober	SheetSage2-AR
Beat	GTZAN	F1 ↑	82.93	**86.27**
	osu2017	F1 ↑	92.28	**93.01**
Downbeat	GTZAN	F1 ↑	78.74	**80.45**
	osu2017	F1 ↑	92.79	**92.90**
Key	GiantSteps	Score ↑	**78.29**	77.73
	GTZAN	Score ↑	72.62	**75.77**
Chord	osu2017	Maj/min ↑	**90.43**	90.08
	Chords1217	Maj/min ↑	**84.29**	83.81
	JAAH	Maj/min ↑	62.94	**64.50**
Structure	HarmonixSet	Accuracy ↑	**80.78**	80.51
		F1 (0.5 s) ↑	66.95	**67.96**
		F1 (3 s) ↑	81.65	**82.86**
Melody	RWC-Pop	Vocal F1 ↑	**83.08**	82.51
		Full F1 ↑	75.00	**75.29**
	Rock Corpus	Vocal F1 ↑	65.98	**67.08**

Melody F1 uses a 50-ms onset tolerance and pitch-class matches, with pitches folded across octaves. Prober uses a fixed zero global rhythm offset before downstream processing. Structure F1 measures section boundaries at the indicated tolerance.

Evaluation protocol. Table 12 uses the validation-selected SheetSage2-AR checkpoint. Inference converts audio to mono at 24 kHz and applies grammar-constrained greedy decoding in bfloat16 (BF16) with batch size one and a maximum output length of 5,120 tokens. Full songs are decoded in 300-second windows with 100-second overlap. Events in the overlap are supplied as a prefix to the next window, and only events in its newly covered tail are retained. We use neither pitch-shift test-time augmentation nor a global timing shift.

For SheetSage2-Prober, we rerun the fixed multitask and beat-conditioned melody models from audio, using 24-kHz mono input, BF16, batch size one, and 300-second multitask windows (30 seconds for GTZAN). Rhythm decoding uses 10-ms resolution and a fixed zero global offset for every dataset. Melody uses 30-second windows with 15-second overlap and four subdivisions per predicted beat, followed by Markov structured decoding, with no subsequent timestamp shift. No reference beats, per-song timing alignment, or pitch-shift augmentation are used. We score the vocal and full-melody outputs separately, before their merger into a lead sheet. Both variants use the same reference tracks, evaluation spans, and metric implementations.

Rhythm results are macro averages of per-track F1 with a 70-ms matching tolerance: 999 GTZAN tracks for beat, 993 for downbeat, and 142 osu2017 (Jiang, 2026) tracks for both. Following Beat This! (Foscarin et al., 2024), we discard reference and predicted beat/downbeat events before 5 s for

every model and use the same reference tracks. We rerun Beat This! under its original inference settings, without DBN postprocessing, and average per-track F1 across its three released training seeds. On the original 993-track GTZAN subset, beat and downbeat F1 match the published means and standard deviations at the reported precision. Our beat evaluation retains the six additional tracks lacking downbeat annotations for every model. Madmom is rescored from saved predictions; the SheetSage1 rhythm entries reuse this backend.

Key evaluation selects the longest predicted key segment and averages the corrected Music Information Retrieval Evaluation eXchange (MIREX) weighted score over 604 GiantSteps and 837 GTZAN tracks.[3] The scoring assigns 0.5 to same-mode dominant or subdominant errors, 0.3 to relative major/minor errors, and 0.2 to parallel major/minor errors.

Chord results average per-track maj/min chord-symbol recall, weighted by each reference annotation span, over all 142 osu2017 tracks and 1,217 Chords1217 tracks. The osu2017 release fixes the evaluated track collection; its beat annotations derive from cleaned game beatmaps, while chord and key annotations were added separately. All available chord predictions use the same references and evaluation time spans. SheetSage1 is independently run on both complete collections; we score its structured harmony output directly. Chords1217 merges the Isophonics, McGill Billboard, and MARL annotations into 1,235 tracks (Humphrey & Bello, 2015); removing 18 duplicate recordings leaves 1,217. The MARL portion contains 195 USPop and 100 RWC-Pop tracks. The original benchmark defines five cross-validation folds, whereas our SheetSage1, SheetSage2-AR, and SheetSage2-Prober evaluations use fixed checkpoints on the full collection. Madmom has documented training overlap with Chords1217. We rerun the five released ChordFormer (Akram et al., 2025) models with the official default chord vocabulary and HMM decoder. On osu2017, we average the five models' component probabilities before decoding; on Chords1217, each track uses only its held-out fold model. Both results are rescored with the same references, time spans, and aggregation as SheetSage2. The implementation extends the chord-structure decomposition code of Jiang et al. (2019). No model is trained or tuned on osu2017.

We additionally evaluate chord transcription on all 113 recordings in JAAH (Eremenko et al., 2018). SheetSage1 uses its released Jukebox-based model and native structured harmony outputs; madmom uses its released CNN feature extractor and CRF chord decoder. Both retain the settings used for the other chord benchmarks. ChordFormer uses the same five-model probability ensemble, default vocabulary, and HMM as on osu2017. All checkpoints and inference settings are fixed for this comparison, without JAAH tuning. The published chord annotations are intersected with each recording's audio span, without extending shorter annotations or excluding songs. This removes 9.15 seconds of trailing no-chord labels beyond the audio, leaving 24,991.97 seconds of reference annotations. We report `mir_eval` maj/min recall, computed per track and weighted by its complete evaluated reference span. Out-of-vocabulary reference segments are excluded from the per-track denominator; they are not relabeled as no chord. Eligible reference duration, including no-chord labels, is 93.87% for maj/min. Using the untrimmed reference spans changes each reported score by less than 0.03 percentage points.

On JAAH, SheetSage2-AR exceeds ChordFormer by 5.06 percentage points in maj/min recall. The maj/min difference between AR and Prober is 1.56 percentage points.

Structure evaluation uses the 200-song HarmonixSet test split, seven section classes, and a common 0.2-second frame grid. Corpus-level accuracy pools frame counts across all songs; boundary F1 pools matched, reference, and predicted boundary counts at 0.5- and 3-second tolerances. All rescored models use the full reference duration; a non-final `end` label denotes silence.

Melody results average per-song onset/pitch-class F1 over all 100 RWC-Pop songs. Matching uses a 50-ms onset tolerance and 50-cent pitch tolerance after folding pitches across octaves; note offsets are excluded. All systems use the same references and onset criterion, with the fixed Prober timing convention described above.

[3] Official MIREX 2025 Audio Key Detection task: `https://music-ir.org/mirex/wiki/2025:Audio_Key_Detection`.

We additionally evaluate vocal melody on all 200 recordings in Rock Corpus (RS 200) (Temperley & de Clercq, 2013), using the released timed vocal-note lists.[4] We retain the supplied note onsets and pitches and use the same 50-ms onset/pitch-class metric, without time alignment or annotation correction. The six recordings whose reference lists contain no pitched notes receive zero F1 and remain in the 200-song macro average. Restricting the average to the 194 nonempty references gives 50.71, 68.02, and 69.15 for SheetSage1, Prober, and AR, respectively. Reference onsets within each measure are derived from metrical positions and bar timestamps. We evaluate onset and pitch-class agreement; note offsets are excluded. SheetSage1 uses its released Jukebox-based model and automatic beats, with six-measure chunks and the existing four-/two-measure fallback. Its single melody output is compared with the vocal reference; AR uses its vocal-melody prompt. Prober uses its vocal output under the fixed inference protocol above. No checkpoints or inference settings are tuned on these annotations.

D Additional System Comparisons

Table 15: **Comparison with proprietary systems on WildSongBench (WSB).** All metrics cover 192 prompts. Mus.: Musicality; Avg: mean of all SongBench (Wu et al., 2026) or SongEval (Yao et al., 2025) dimensions; PQ: AudioBox production quality (Tjandra et al., 2025); MuLan: MuQ-MuLan (Zhu et al., 2025); AMCaps: AllMusicCaps (Alonso-Jiménez et al., 2026); Q3O: Qwen3-Omni prompt adherence (Xu et al., 2025); PER: phoneme error rate. Both YuE2 settings use symbolic planning. The standard comparison selects the lower-PER output from two candidates per prompt for every system; YuE2 best-of-8 (Bo8) selects from eight, prioritizing SongBench Musicality, then Q3O, then PER. Appendix E.3 details selection. MM abbreviates MiniMax. Bold and underlining mark the best and second-best observed means; higher is better except PER.

	SongBench		**SongEval**		**AudioBox**	**Text alignment**			**Lyrics**
Model	Mus.↑	Avg↑	Mus.↑	Avg↑	PQ↑	MuLan↑	AMCaps↑	Q3O↑	PER↓
Suno v5 (Suno, 2025)	5.9918	6.8721	4.3051	4.3579	8.1698	**0.5428**	**0.4353**	4.5907	0.0810
Suno v4.5 (Team Suno, 2025)	5.8317	6.6995	<u>4.3198</u>	<u>4.3666</u>	8.2541	0.5022	0.3873	4.4149	**0.0580**
Suno v5.5 (Shulman, 2026)	5.8087	6.7150	4.1497	4.2152	8.1955	<u>0.5089</u>	0.3917	4.5914	<u>0.0596</u>
Suno v6 (Suno, 2026)	5.6558	6.5562	4.2635	4.3086	8.1296	0.4916	0.4305	4.6258	0.0758
Suno v6 Wild (Suno, 2026)	5.5644	6.4195	4.1716	4.2199	8.1785	0.4999	<u>0.4316</u>	4.5898	0.0745
MM Music 2.6 (MiniMax, 2026a)	5.4437	6.3222	4.0560	4.0985	8.1711	0.4251	0.3670	4.5688	0.2455
Mureka 9 (Mureka, 2026)	<u>6.0488</u>	<u>6.9377</u>	**4.3619**	**4.4111**	8.0226	0.4394	0.4102	4.6368	0.1169
YuE2	5.9075	6.7316	4.2151	4.2625	<u>8.2598</u>	0.5068	0.4054	<u>4.6819</u>	0.0844
YuE2 (Bo8)	**6.2666**	**6.9632**	4.2506	4.2960	**8.2714**	0.5051	0.3980	**4.7009**	0.0979

[4]Official melodic transcriptions and timed note lists: `https://rockcorpus.midside.com/melodic_transcriptions.html`.

Table 16: **Comparison with publicly released systems on WSB.** Prompts, metrics, abbreviations, and YuE2 selection budgets follow Table 15. Both YuE2 rows use symbolic planning. HeartMuLa† reports using SongEval and AudioBox during post-training to filter supervised fine-tuning data and construct DPO preference pairs (Yang et al., 2026). Bold and underlined values indicate the best and second-best scores across all rows, respectively. Appendix E.3 gives the baseline protocols.

	SongBench		SongEval		AudioBox	Text alignment			Lyrics
Model	Mus.↑	Avg↑	Mus.↑	Avg↑	PQ↑	MuLan↑	AMCaps↑	Q3O↑	PER↓
YuE1 (Yuan et al., 2025)	4.0847	4.9165	3.1524	3.2150	7.8683	0.2623	0.2882	3.7301	0.3638
SongBloom (Yang et al., 2025)	3.4493	4.2350	3.2048	3.2051	8.1539	0.2697	0.1926	3.0287	0.1919
LeVo 2 (Lei et al., 2026)	5.4590	6.3247	3.9819	4.0234	**8.3966**	0.3542	0.2680	3.9458	0.2612
ACE-Step 1.5 (Gong et al., 2026)	5.1588	6.0118	3.8051	3.8465	8.0518	0.4372	0.3869	4.5809	<u>0.0746</u>
HeartMuLa† (Yang et al., 2026)	5.4963	6.2483	**4.5329**	**4.5519**	<u>8.2933</u>	0.3823	0.2786	3.4907	0.1071
DiffRhythm 2 (Jiang et al., 2025b)	4.4775	5.2428	3.4711	3.5245	7.9782	0.3782	0.3255	4.0870	0.1841
Muse (Jiang et al., 2026)	5.1692	6.0349	3.7362	3.7887	8.0517	0.3937	0.3466	4.4038	0.3342
MM Music 3 (MiniMax, 2026b)	5.3482	6.2830	4.0431	4.0994	8.2825	0.3928	0.3609	4.4362	**0.0627**
YuE2	<u>5.9075</u>	<u>6.7316</u>	4.2151	4.2625	8.2598	**0.5068**	**0.4054**	<u>4.6819</u>	0.0844
YuE2 (Bo8)	**6.2666**	**6.9632**	<u>4.2506</u>	<u>4.2960</u>	8.2714	<u>0.5051</u>	<u>0.3980</u>	**4.7009**	0.0979

D.1 Unified versus Separate Generation

Having established the benefit of symbolic planning, we ask whether semantic-token prediction and acoustic generation should be learned separately or jointly. Many recent song generators pair an autoregressive language model with a separate diffusion Transformer (LM+DiT) (Yang et al., 2026; Xu et al., 2026; Lei et al., 2025; 2026; Dai et al., 2026). We test whether bringing these stages into YuE2's unified model improves song quality over this widely used design, with the representations and planning method held fixed.

Protocol. We compare YuE2's unified AR–NAR Mixture-of-Transformers (MoT) (Deng et al., 2025) with a separate LM+DiT baseline following Qwen-Music (Xu et al., 2026). We match the training-data volume between the two systems. The LM and DiT each have 1.7B parameters (3.4B in total), comparable to the 3.58B MoT. The LM and MoT's AR component use the same token budget, batch size, and number of training updates. The DiT is trained for 900,000 updates. Both systems use the same semantic tokenizer, acoustic VAE, and melody-and-chord symbolic planning, so the comparison tests unified versus separate generation with the same representations and planning method.

We evaluate both systems on the same 192 WSB prompts and lyrics, generating two candidates per prompt. Each candidate is transcribed four times with ASR, its lowest PER is retained, and the lower-PER candidate is selected. The Arena expert panel compares the resulting anonymized audio pairs under the same randomized presentation, minimum listening duration, and consistency checks as the planning experiment. Experts judge overall quality, musicality, text alignment, audio quality, vocals, and accompaniment separately, with ties and unable-to-judge responses allowed. The direct comparison with planning contains 209 retained expert evaluations, yielding 206–208 judged responses per criterion. Figure percentages include ties and exclude unable-to-judge responses; Appendix E.7.2 reports the criterion-specific counts, CR1 intervals clustered by evaluator and song, and two-sided p-values.

Results. With melody-and-chord symbolic planning in both systems, experts prefer YuE2's unified MoT to separate LM+DiT. For overall quality, 53.4% of judged responses favor YuE2 and 35.6% favor separate LM+DiT ($p = 0.0084$); the rest are ties. The corresponding audio-quality shares are 48.5% and 29.6% ($p = 3.06 \times 10^{-4}$; Figure 6). Preferences also favor YuE2 for musicality, vocals, and accompaniment; text alignment is nearly balanced at 34.0% versus 33.5%.

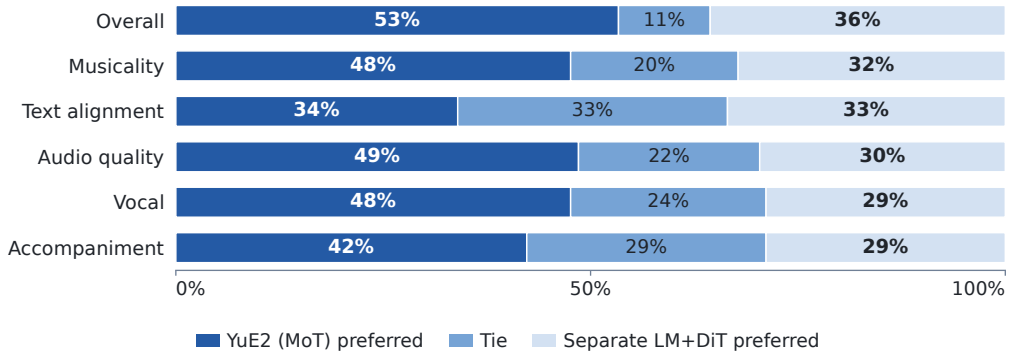

Figure 6: **Experts prefer unified generation on five of six criteria; text alignment is nearly balanced.** YuE2's Mixture-of-Transformers (MoT) (Deng et al., 2025) learns semantic-token prediction and acoustic generation in one backbone; the baseline uses a separate language model and diffusion Transformer (LM+DiT) (Xu et al., 2026). Both use melody-and-chord planning, the same training-data volume and representations, and the same prompts and two-candidate selection protocol. Dark blue favors YuE2, medium blue denotes ties, and light blue favors LM+DiT. Bars show percentages of judged expert responses (206–208 per criterion), excluding unable-to-judge answers.

E Data and Evaluation Protocols

WildSongBench (WSB) contains 192 in-the-wild requests across 15 genre buckets, with 94 Chinese and 98 English source-language labels and multilingual requests retained.

E.1 Qwen3-Omni Prompt Alignment (Q3O)

We use Q3O to denote our Qwen3-Omni-based prompt-alignment score. It measures how well a generated song follows the musical requirements in its WildSongBench request. We evaluate seven dimensions: vocal timbre, vocal setup (solo, duet, group, gender, or instrumental), genre, mood, instrumentation and production, tempo, and key. We use Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) in two passes. A text-only pass assigns each dimension an importance weight $w_{i,d} \in \{0, \ldots, 5\}$; zero means that the request does not specify that dimension. For each prompt, the same dimension weights are used across all evaluated systems and sampling settings. An audio pass then assigns each active dimension a match score $s_{i,d} \in \{0, \ldots, 5\}$, where five is a strong match and zero is a clear miss.

For request i, we compute

$$C_i = \frac{\sum_d w_{i,d} s_{i,d}}{\sum_d w_{i,d}}, \qquad \text{Q3O} = \frac{1}{N} \sum_{i=1}^{N} C_i. \tag{5}$$

An active key term uses the standard Music Information Retrieval Evaluation eXchange (MIREX) weighted agreement between the requested key and a `madmom` CNN estimate (Böck et al., 2016), rescaled from $[0, 1]$ to $[0, 5]$. The reported system score averages all $N = 192$ requests.

Audio-scoring prompt. For reproducibility, the exact instruction and user-message template are shown below. Angle-bracketed fields are filled directly from each WSB record; `DIMENSIONS` and `OUTPUT_SCHEMA` contain only dimensions with positive prompt weight.

Q3O: exact audio-scoring prompt

SYSTEM MESSAGE

```
You are a strict music prompt-control evaluator.

Evaluate whether the provided audio satisfies the user's open-format music
request. Use only the user's raw prompt, style tags, and raw genre as the target.
Do not use any standardized or annotated prompt fields as ground truth.

Return JSON only. Do not wrap it in markdown.

Score each dimension independently:
- score: integer 0-5. 5 means the audio strongly matches the user's request for
  this dimension. 0 means it clearly contradicts or misses the request.
- reason: short reason. Mention the requested target and the audible evidence.

Evaluate only the dimensions listed below. Do not add extra dimensions.

Dimensions:
{{DIMENSIONS}}

Output schema:
{{OUTPUT_SCHEMA}}
```

USER MESSAGE TEMPLATE · AUDIO ATTACHED

```
Evaluate this generated song against the following open-format user request.

Title: <TITLE>
Language: <LANGUAGE>
Raw genre: <RAW_GENRE>
Style tags: <STYLE_TAGS>

Raw user prompt:
<RAW_USER_PROMPT>
```

E.2 Production Quality and Audio–Text Alignment

Each additional metric is evaluated on the same selected audio as the other WSB metrics, with complete coverage of 192 songs per system. System scores are arithmetic means over songs. AudioBox Aesthetics (Tjandra et al., 2025) estimates production quality (PQ), including clarity, fidelity and spatial presentation. Its full-song score uses length-weighted 10-s windows.

For MuQ-MuLan (Zhu et al., 2025) and AllMusicCaps (Alonso-Jiménez et al., 2026), the text target is the same standardized style description for each prompt, containing genre, instrumentation, mood and vocal attributes but no lyrics. Thus these scores measure alignment with the style annotation. They complement Q3O, which evaluates alignment with the original user request. Audio is decoded to 24-kHz mono. MuQ-MuLan uses its serial full-song inference with 10-s clips. AllMusicCaps uses the released model with a trained text encoder and SigReg regularization: it averages embeddings over complete, non-overlapping 10-s audio clips, normalizes the mean and text embeddings, then computes cosine similarity. Audio shorter than 10 s is zero-padded; an incomplete trailing clip in longer audio is omitted.

E.3 Selection and Evaluator Reuse

Both YuE2 settings use melody-and-chord symbolic planning. For every system, the standard comparison selects the lower-PER output from two candidates per prompt. Candidate budgets count outputs available for selection. For YuE2, the publicly released baselines, and Suno v6 and v6 Wild (Suno, 2026), each candidate's PER is the lowest across four ASR transcriptions. MiniMax Music 3 (MiniMax, 2026b) uses its official caption rewriter; MM Music abbreviates MiniMax Music in the tables.

For best-of-8, we remove candidates dominated jointly in SongBench Musicality, Q3O and PER, retain candidates within 0.10 of the highest Musicality, then within 0.10 of the highest remaining Q3O, and select the lowest PER. Reported best-of-8 scores evaluate the selected recordings.

HeartMuLa uses SongEval to filter supervised fine-tuning data and, with AudioBox, to construct a DPO preference set; its final model merge includes the resulting DPO model (Yang et al., 2026). The tables flag this evaluator reuse and include all systems in the per-column rankings.

E.4 Calculation of the Quality and Alignment Indices

Figure 1 maps each system to a song quality index (horizontal axis) and a text alignment index (vertical axis), with bubble area encoding production quality. Its 17 points use the same selected recordings and candidate-selection protocols as Tables 15–16.

Prompt aggregation and standardization. For each metric m, let $v_m(s, i)$ denote the score for system s on WSB prompt i. All inputs have complete coverage of the same 192 prompts. Each prompt contributes equally to the system mean:

$$x_m(s) = \frac{1}{192} \sum_{i=1}^{192} v_m(s, i). \tag{6}$$

SongBench Global Avg averages all seven emitted dimensions within a song; SongEval Global Avg averages all five. MuLan and AllMusicCaps use their audio–text similarities, and Q3O uses the prompt-weighted 0–5 score defined in Appendix E.1. We standardize each system-level metric against the same reference set $\mathcal{B}$: all 15 external systems in Tables 15–16, including HeartMuLa and both Suno v6 variants. The two YuE2 settings are excluded from the reference statistics. We compute

$$\mu_m = \frac{1}{15} \sum_{b \in \mathcal{B}} x_m(b), \qquad \sigma_m^2 = \frac{1}{15} \sum_{b \in \mathcal{B}} (x_m(b) - \mu_m)^2, \qquad z_m(s) = \frac{x_m(s) - \mu_m}{\sigma_m}. \tag{7}$$

This gives each input a common reference mean and variance before weighting. The population standard deviation uses the 15 external system means with divisor 15.

Weighted composites. The quality and alignment composites are

$$Q(s) = \tfrac{2}{3} z_{\text{SongBench Avg}}(s) + \tfrac{1}{3} z_{\text{SongEval Avg}}(s), \tag{8}$$

$$A(s) = \tfrac{1}{3} \big(z_{\text{MuLan}}(s) + z_{\text{AMCaps}}(s) + z_{\text{Q3O}}(s) \big). \tag{9}$$

We give SongBench twice the weight of SongEval to emphasize its fine-grained assessment of contemporary song generators. This more recent evaluator covers seven aspects of song quality and uses expert calibration, blind listening, and annotation quality control (Wu et al., 2026). Its authors also report stronger agreement with expert Musicality ratings than SongEval (Yao et al., 2025) on their out-of-distribution test set. SongEval contributes a complementary assessment of song aesthetics. Alignment weights all three standardized inputs equally. The composites summarize standardized metric means under these weights for visualization; the original metrics are the primary reported measurements.

Display coordinates. For display, each composite $C \in \{Q, A\}$ is rescaled to an index on 0–100 axes:

$$I_C(s) = 10 + 80 \, \frac{C(s) - C_{\min}}{C_{\max} - C_{\min}}. \tag{10}$$

The endpoints are the observed extrema across all 17 settings. The lowest and highest composites map to relative display coordinates of 10 and 90. This positive affine mapping preserves weights, rankings, and pairwise dominance without changing the 15-system standardization reference.

YuE2 and YuE2 (Bo8) obtain quality indices of 82.82 and 88.34, respectively, and alignment indices of 84.59 and 83.89. Both settings exceed each public baseline on both indices. Mureka 9 has the highest quality index (90), while Suno v5 has the highest alignment index (90).

Pareto outlines and bubble areas. A setting t dominates s when $I_Q(t) \geq I_Q(s)$ and $I_A(t) \geq I_A(s)$, with at least one strict inequality. Black outlines identify settings dominated by none of the other 16 settings, using the observed means without rounding. YuE2 (Bo8), Suno v5, and Mureka 9 form the overall empirical Pareto front.

Bubble area separately encodes AudioBox production quality (PQ), with larger areas indicating higher scores; area ratios are not PQ ratios. PQ is excluded from the two-axis Pareto calculation.

Weight sensitivity. We also inspect quality weights from 0.5 to 0.8 for SongBench and replace both global averages with the respective Musicality scores. For either choice of dimensions, YuE2 exceeds every public baseline throughout the 0.6–0.8 weight interval. At the selected 2:1 weighting, the Musicality composite is highest for YuE2 (Bo8), whereas the global-average composite is highest for Mureka 9. The teaser uses global averages to summarize all quality dimensions under the specified 2:1 weighting.

E.5 Musicality and Model Size

Figure 7 compares musicality and parameter count for the eight public baselines and both YuE2 settings. Among the two-candidate settings, YuE2 has the highest observed Musicality index (81.63) with 3.58B parameters. It uses 34.0% fewer parameters than HeartMuLa (5.43B, index 78.74), the strongest other public system on this index. Six of the eight baselines have both more parameters and a lower index; DiffRhythm 2 and SongBloom are smaller and have lower indices.

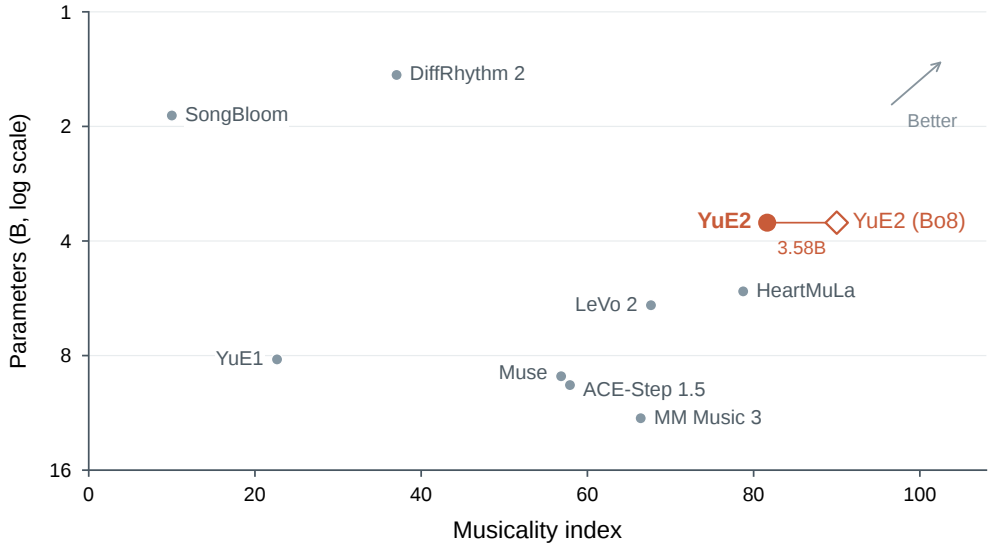

Figure 7: **YuE2 achieves the highest Musicality index among the evaluated public generators, with fewer parameters than six of eight baselines.** Results use 192 WildSongBench prompts. The index combines standardized SongBench and SongEval Musicality scores with 2:1 weights; it is not a percentage. Moving right indicates higher Musicality index; moving up indicates fewer parameters on a reversed logarithmic axis. Circles denote the two-candidate settings; the hollow diamond denotes YuE2 (best-of-8). The connecting segment identifies the same 3.58B-parameter model under these two selection protocols. Counts include generation networks and conditioning encoders, excluding audio tokenizers, VAEs, and vocoders. Appendix E.5 gives the index and counting details.

Musicality index. We combine the Musicality dimensions of SongBench (Wu et al., 2026) and SongEval (Yao et al., 2025), using the prompt aggregation and 15-system standardization reference in Appendix E.4:

$$M(s) = \tfrac{2}{3} z_{\text{SongBench Musicality}}(s) + \tfrac{1}{3} z_{\text{SongEval Musicality}}(s). \tag{11}$$

We apply the same affine display mapping, with extrema across all 17 WSB settings mapped to 10 and 90; the public subset is not renormalized. The index summarizes observed metric means under these weights, rather than measuring a percentage. Appendix E.4 reports weight sensitivity. HeartMuLa's raw SongEval score remains included, with its reported evaluator reuse disclosed in Appendix E.3.

Parameter counts. We count the evaluated checkpoints' language models, generative diffusion networks, and conditioning encoders, including frozen parameters, embeddings, output heads, and retained auxiliary heads. Shared weights count once. We exclude audio tokenizers, VAEs, vocoders, deterministic inverse modules, and fixed buffers; generative diffusion networks within codec pipelines remain included. Counts come from checkpoint tensor shapes and model configurations. MiniMax Music 3's count covers its released music model and excludes the external caption rewriter, whose size is undisclosed. Best-of-8 uses the same model with a larger candidate budget and a different selection rule (Appendix E.3).

E.6 Cover Generation and Version Retrieval

Zero-shot scope and supervision. The evaluated SHS100K (NovaFrost, 2018) works, including their alternate performances, are absent from YuE2's training corpus. We verified this using CLEWS (Serrà et al., 2025) and Discogs-VINet (Araz et al., 2024) to check for work-level overlap with the training corpus. YuE2 learns from individual recordings and their aligned annotations, without pairing an original performance with a cover as the target. We use the general song-generation checkpoint without cover-specific fine-tuning. Zero-shot therefore refers to both unseen works and transfer without cover-specific training. The source score is supplied only at inference through the existing conditioning interface; SheetSage2 remains an external analyzer.

The baseline training procedures differ. SongEcho (Li et al., 2026) trains a melody encoder and modulation modules on Suno70k while freezing its generative backbone. ACE-Step 1.5 (Gong et al., 2026) includes multi-task training for audio manipulation, with cover generation supported through quantized source latents. We compare the released systems under their respective training and inference procedures.

Data and selection. SHS100K (NovaFrost, 2018) groups recordings by musical work. We use its public test partition, which lists 10,547 recordings of 1,692 works. The evaluated subset contains 948 source recordings, each from a different work: 947 are labeled English and one Chinese. The retrieval gallery contains 10,546 available recordings from all 1,692 works. Source and style-target candidates are 60–300 s long and come from works with at least three usable versions. Sources must have at least 20 automatically transcribed lyric units, vocal/style annotations, complete symbolic transcriptions, and two annotated alternative versions. Selection also requires a chorus of at least four bars and successful conversion of its score variants. All 948 eligible sources among the 1,349 candidates are included.

Source composition and target style. Qwen3-ASR-1.7B (Shi et al., 2026) transcribes source lyrics and identifies their language; SheetSage2 transcribes melody, chords, beats, key, and form. Qwen3-Omni-30B-A3B-Instruct (Xu et al., 2025) describes each recording's genre, instrumentation, voice, tempo, and mood, with song titles, artist names, and lyrics excluded from its instructions. We choose two alternative recordings of the same work to maximize the sum of pairwise Jaccard distances between the source and two target tag sets. Each alternative provides a style description, rather than a waveform to reconstruct. The scores, lyrics, and style descriptions are automatic annotations.

Generation conditions. The full-score condition supplies the source melody-and-chord transcription as a fixed prefix. The condition without chords rebuilds that transcription without the chord timeline, retaining vocal and instrumental melody voices, beats, key, and form; this also changes chord-dependent serialization boundaries. The condition without a score supplies an empty score prefix but retains lyrics and target style. YuE2 consumes no source waveform during rendering. All three conditions use the same generator checkpoint and 48-kHz stereo decoder used for metric evaluation. Autoregressive sampling uses temperature 1.0, top-p 0.95, top-k 100, and a repetition

penalty of 1.2 over 50 tokens. Acoustic generation uses 32 midpoint ODE steps. The acoustic token budget is 25 times the source duration in seconds, capped at 9,000; generation may terminate earlier.

SongEcho (Li et al., 2026) uses the source recording to extract its melody condition, with 60 sampling steps, guidance scale 15, and $\omega = 10$. ACE-Step 1.5 (Gong et al., 2026) uses the released XL supervised-fine-tuned model in native cover mode, with 50 steps, guidance scale 7, shift 3, cover strength 0.8, cover noise strength 0, and language-model thinking disabled. Both receive the source lyrics and target style text. Each method generates two samples for each of the two target styles, for 3,792 outputs per method and 18,960 in total. Retrieval and embedding alignment average all outputs without candidate selection; Q3O coverage is specified below.

Retrieval and metric definitions. Generated audio is the query. The same gallery is used for CLEWS (Serrà et al., 2025) and Discogs-VINet (Araz et al., 2024), with 10,545 candidates remaining after exclusion of the exact source recording. All remaining recordings of the source work are relevant, including those used to derive target style descriptions. CLEWS uses its released SHS model, segment embeddings extracted with a 5-s hop, and the mean distance of up to ten greedily matched segment pairs, with no segment reused. Discogs-VINet uses its released Discogs-VI model and cosine similarity of normalized track embeddings. Their scores are reported separately.

Average precision averages the precision at every rank occupied by a relevant recording; mAP averages this value across queries. MRR averages the reciprocal rank of the first relevant recording. Hit@k is the percentage of queries with at least one relevant recording among the first k results, rather than the fraction of all relevant recordings retrieved. Higher values are better for all four measures. The table reports descriptive means over generated queries, with four queries grouped under each source work.

Target-style control and coverage. MuQ-MuLan (Zhu et al., 2025), AllMusicCaps (Alonso-Jiménez et al., 2026), and Q3O (Xu et al., 2025) compare each generated recording with the target style description used to condition it, excluding lyrics. MuQ-MuLan uses serial full-song inference at 24 kHz; AllMusicCaps averages embeddings from non-overlapping 10-s clips before computing audio–text cosine similarity, as in Appendix E.2. Both metrics cover all 3,792 outputs per method. The supplementary AllMusicCaps comparison gives full-score YuE2 a mean of 0.302, compared with 0.349 for SongEcho, the highest value among these five conditions.

Q3O uses the weighted 0–5 score in Appendix E.1, with the target style text as the request. It covers the common available-case subset of 3,286 outputs per method (86.7%), matched by source, target style and seed. These cover 933 works: 710 contribute four outputs per method and 223 contribute two. All active dimensions are valid, and the reported means use the stored weights for these records. The 506 unscored outputs per method are outside this availability-based subset. Qwen3-Omni supplies both target style annotations and Q3O judgments.

Audio quality. AudioBox Aesthetics (Tjandra et al., 2025) evaluates production quality (PQ) on the same generated recordings. SongBench (Wu et al., 2026) evaluates Musicality on each complete waveform after conversion to 24-kHz mono audio. It covers all 3,792 outputs per method, matched by source, target style, and seed, without candidate selection. Both quality columns report mean scores at three decimal places.

Interpreting the score conditions. Cover generation pairs a fixed source score with a contrasting target style. Our qualitative assessment with music experts links the observed metric ordering to the role of harmony in musical style: retaining the source harmony favors fidelity to the original work, while relaxing it allows greater adaptation to the target style. Retrieval measures work identity; alignment compares the audio with the target description; PQ and Musicality evaluate the audio itself.

After averaging four outputs per work, Musicality is higher than with the full source score for 798 of 948 works (84.2%) when source chords are removed, and for 864 works (91.1%) when no score is supplied or generated. These are descriptive comparisons of work means, with no output selection.

In this fixed-source cover task, the full score yields the strongest work-identity retrieval, while the relaxed conditions achieve higher alignment and quality scores.

Table 17: **YuE2 with full scores leads both evaluated cover systems on all eight retrieval measures without cover-specific training.** SHS100K (NovaFrost, 2018) evaluation across 948 works, with 3,792 outputs per method and no candidate selection. These works are absent from YuE2's training corpus. Each query ranks 10,545 recordings after excluding its source. All metrics are higher-is-better.

	CLEWS				Discogs-VINet				Alignment		Quality	
Method	mAP↑	MRR↑	Hit@1↑	Hit@5↑	mAP↑	MRR↑	Hit@1↑	Hit@5↑	MuLan↑	Q3O↑	PQ↑	Mus.↑
SongEcho	0.419	0.536	48.4	58.8	0.122	0.227	16.6	28.4	0.366	4.474	6.862	3.286
ACE-Step 1.5	0.024	0.036	2.4	4.1	0.006	0.014	0.6	1.4	0.166	4.190	6.918	3.689
YuE2 (full score)	**0.647**	**0.748**	**71.3**	**78.6**	**0.288**	**0.438**	**37.5**	**50.3**	0.382	4.273	8.044	5.104
Without chords	0.598	0.715	67.3	76.1	0.179	0.304	23.6	36.6	0.417	4.482	8.117	5.490
Without score	0.006	0.008	0.3	0.9	0.004	0.008	0.2	0.8	**0.474**	**4.837**	**8.186**	**5.691**

Hit@k is in percent; Q3O ranges from 0 to 5. PQ: AudioBox production quality; Mus.: SongBench Musicality. **Bold**: column best; underline: second best.

E.7 Expert Listening Evaluation

Recruitment and coverage. The campaign comprised an online study recruited through X, GitHub, and Hugging Face, and a paid expert panel. The online study contributed 837 retained evaluations from 289 participation sessions, assessing overall quality only. The expert panel comprised 62 annotators with conservatory training or audio AI research backgrounds, including Central Conservatory of Music, Shanghai Conservatory of Music, and Xinghai Conservatory of Music. The Arena export records 351 participations and 4,439 retained pairwise evaluations as of 21 September 2026. The reported Arena preferences use the 3,602 expert evaluations: 2,545 frontier, 636 planning, and 421 architecture evaluations, covering 192 prompts. Each expert evaluation contains separate answers for overall quality, musicality, text alignment, audio quality, vocals, and accompaniment. Within the same listening study, four listeners drawn from the same expert pool also assessed melody and chord progression, as described below.

Evaluation criteria. Table 18 defines the six reported criteria, drawing on the musical, acoustic, and control dimensions in YuE (Yuan et al., 2025) and the song-quality dimensions in SongBench (Wu et al., 2026). Overall quality is a direct judgment of the complete song; it is not computed by averaging the other five responses.

Table 18: **Definitions of the six expert-listening criteria.**

Criterion	**Definition**
Overall quality	Overall preference for the complete song, considering its composition, performance, sound quality, and fit to the request.
Musicality	Musical appeal and expressive coherence, including memorable melodic phrases, convincing harmony, rhythmic flow, effective arrangement, and development across sections.
Text alignment	Adherence to the supplied style description and lyrics, including the requested genre, mood, instruments, vocal configuration, tempo, and accurate delivery of the words.
Audio quality	Fidelity of the complete recording: audible detail, balanced levels, separation of sound sources, and freedom from unintended noise, clipping, distortion, or synthesis artifacts.
Vocals	Quality of the singing: natural vocal tone, stable pitch and timing, intelligible pronunciation, and expressive phrasing.
Accompaniment	Quality of the instrumental performance: believable timbres, clear articulation, expressive dynamics, and coordination among instruments and with the voice.

Listening and consistency checks. System identities were hidden and A/B presentation order was randomized. The expert protocol required at least 30 seconds of listening to each song. Every batch of 32 questions contained five hidden consistency checks, with at least four correct responses required to pass the batch. The checks used either identical audio on both sides, which required a tie, or a previously presented pair with the audio order reversed, which required the same preference after restoring the model identities. The analysis retains non-check expert responses that were neither rejected nor part of an invalidated group at the export cutoff. Online responses are retained only from accepted rounds, excluding check items and responses marked for exclusion from voting.

Generation and selection. The planning comparison holds the checkpoint, prompt, candidate budget, and decoder fixed. The LM+DiT architecture baseline follows Qwen-Music (Xu et al., 2026) and uses the same training-data volume as the unified MoT. Its LM and DiT each have 1.7B parameters, for 3.4B in total compared with MoT's 3.58B. The LM and MoT's AR component use the same token budget, batch size, and number of training updates; the DiT is trained for 900,000 updates. Both systems use the same semantic tokenizer, acoustic VAE, and symbolic planning method. The main architecture comparison uses melody-and-chord symbolic planning, the same 192 prompts, and two candidates per prompt. Each candidate is transcribed four times with ASR, and its lowest PER is retained. The lower-PER candidate is then selected for each prompt. YuE2 (best-of-8) selects from eight candidates, prioritizing SongBench Musicality, then Q3O, then PER.

E.7.1 Melody and Chord Evaluation

Four listeners drawn from the same expert pool evaluated 100 song pairs generated with and without symbolic planning, holding the prompts, checkpoint, candidate budget, and decoder fixed. Each listener assessed 50 pairs, and each pair received two evaluations from different listeners. Presentation order was randomized and system identities were hidden. Listeners assessed melody, chord progression, and overall quality separately. Melody concerns phrase shape, development, and memorability; chord progression concerns harmonic coherence, movement, and fit with the melody. Overall quality has the meaning given in Table 18. Figure 4 reports the melody and chord-progression judgments. Each criterion has 200 judgments, and preference shares use the corresponding criterion's denominator.

E.7.2 Statistical Protocol

We recover model identities from the recorded A/B presentation order. For each criterion, W and L count preferences for the first- and second-named systems, and T counts ties. Figure percentages use $n = W + T + L$ as the denominator; the tables separately report unable-to-judge answers as U. The tie-adjusted preference is

$$S = \frac{W + 0.5T}{W + T + L}. \tag{12}$$

Average comparisons weight model pairs equally and responses equally within each pair. The frontier averages cover the six proprietary baselines. The planning average combines melody-and-chord versus no planning with melody-only versus no planning. The architecture average combines the planning and no-planning comparisons.

Expert standard errors use two-way CR1 clustering by evaluator and song; online standard errors use one-way CR1 clustering by anonymous participation session. Pointwise 95% t intervals use the smaller evaluator/song cluster count minus one as the degrees of freedom for experts, and the session count minus one for online responses. Two-sided tests compare S with 0.5. We report each test's unadjusted p-value. Best-of-8 average audio-quality preference is 58.9% (CR1 95% CI 55.6–62.2%; $p = 1.70\times10^{-6}$). Tables 19 and 20 provide the counts, CR1 intervals, and p-values for every expert direct and average comparison.

E.7.3 Complete Frontier Preferences

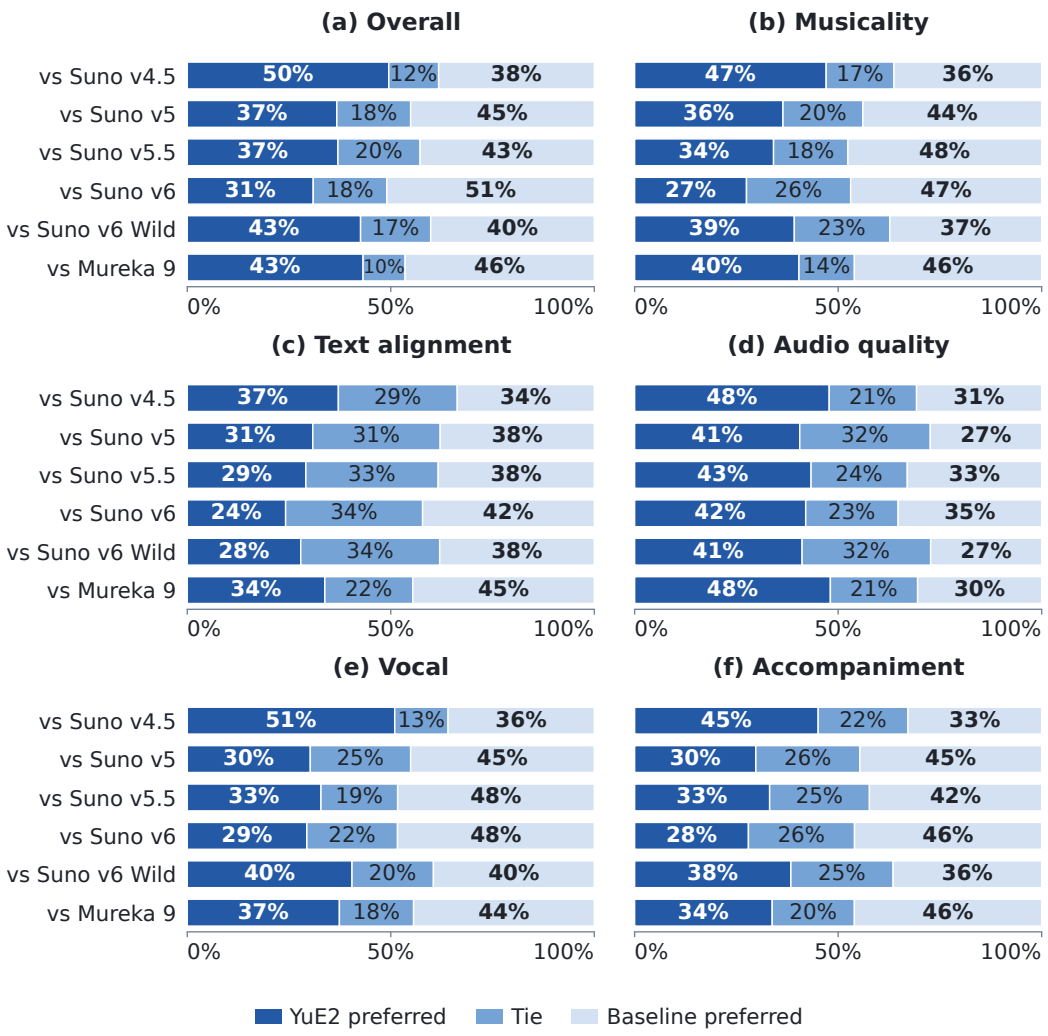

Figure 8: **Expert preferences for YuE2 versus six proprietary song generators across all six criteria.** Audio-quality preferences favor YuE2 against all six baselines; overall preferences favor it in two comparisons. YuE2 uses melody-and-chord planning and selects the candidate with fewer lyric errors from two generated songs. Baselines are Suno v4.5 (Team Suno, 2025), v5 (Suno, 2025), v5.5 (Shulman, 2026), v6 and v6 Wild (Suno, 2026), and Mureka 9 (Mureka, 2026). Dark blue favors YuE2, light blue favors the named baseline, and medium blue denotes ties. Baseline order is shared by the left and right panels. Each bar sums to 100% of judged responses, excluding unable-to-judge answers. Counts and CR1 intervals are in Appendix E.7.2.

For audio quality, experts prefer YuE2 to each of the six proprietary baselines more often than they prefer the reverse (Figure 8). Overall preferences favor YuE2 in two of the six comparisons. Preferences for musicality, text alignment, vocals, and accompaniment vary by baseline.

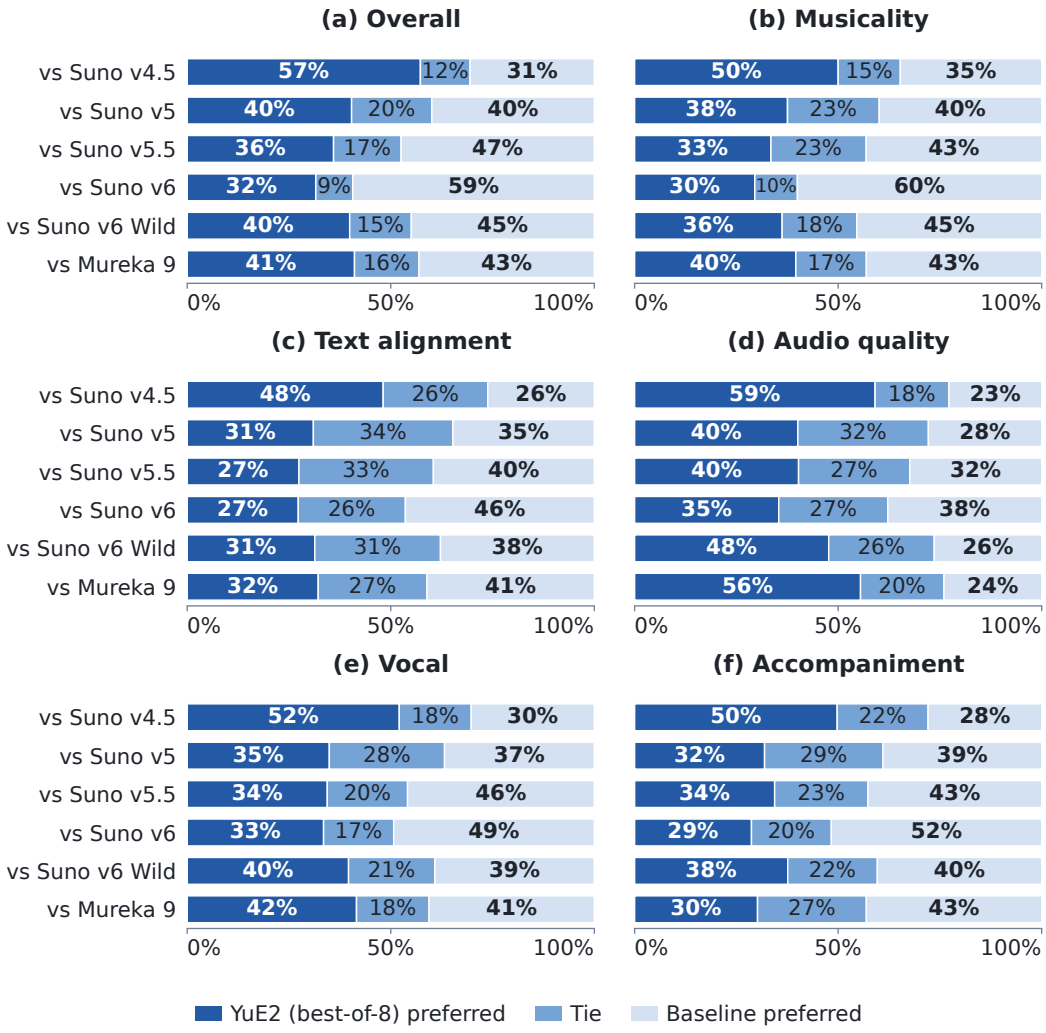

Figure 9: **Expert preferences for YuE2 (best-of-8) versus six proprietary song generators across all six criteria.** Overall preferences favor best-of-8 over Suno v4.5, are nearly balanced against v5, and favor v6 over best-of-8. Audio-quality preferences favor best-of-8 in five of six comparisons. YuE2 uses melody-and-chord planning and selects from eight candidates by musicality, prompt adherence, then lyric accuracy. Baselines are Suno v4.5 (Team Suno, 2025), v5 (Suno, 2025), v5.5 (Shulman, 2026), v6 and v6 Wild (Suno, 2026), and Mureka 9 (Mureka, 2026). Dark blue favors YuE2 (best-of-8), light blue favors the named baseline, and medium blue denotes ties. Baseline order is shared by the left and right panels. Each bar sums to 100% of judged responses, excluding unable-to-judge answers. Counts and CR1 intervals are in Appendix E.7.2.

Best-of-8 receives more overall preferences than Suno v4.5 and nearly balanced preferences against Suno v5. Suno v6 receives more overall preferences (Figure 9). For audio quality, experts prefer best-of-8 in five of the six comparisons; Suno v6 receives more preferences in the remaining comparison. The six-baseline average audio-quality preference is 58.9% (Appendix E.7.2).

E.7.4 Complete Planning and Architecture Comparisons

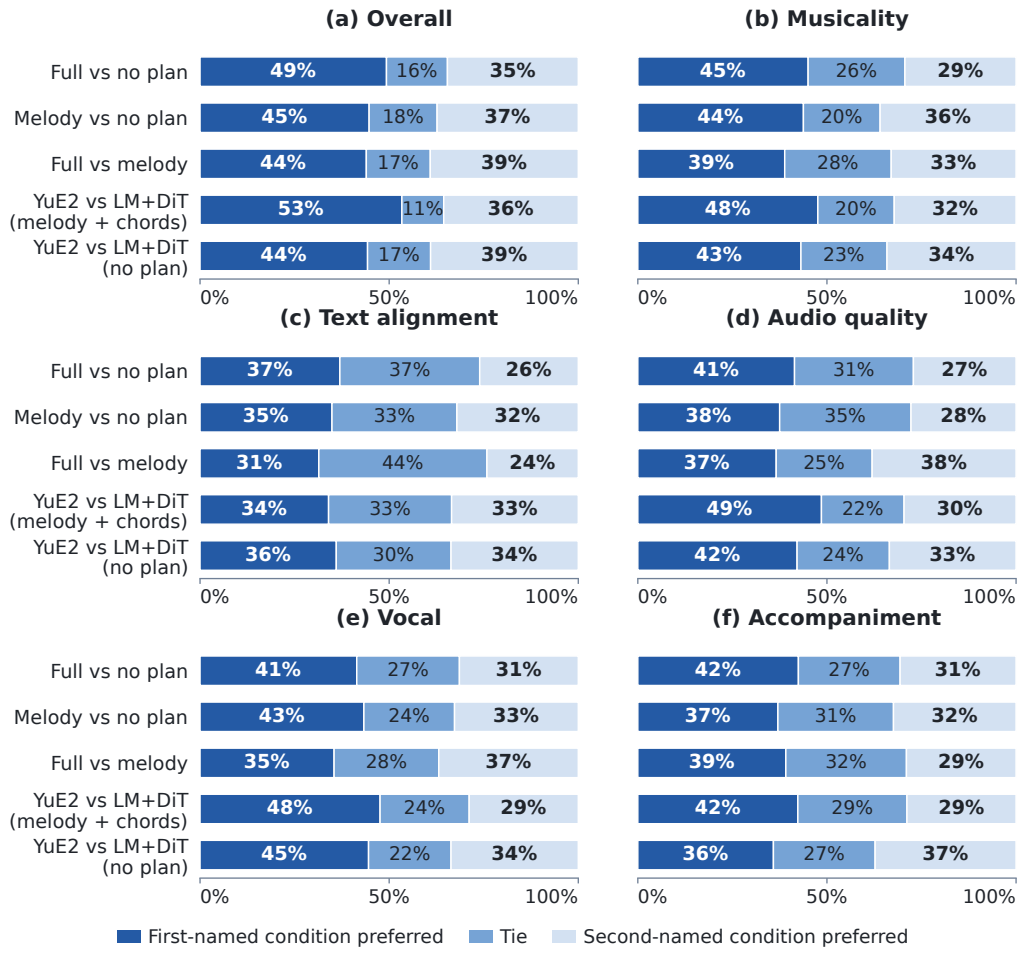

Figure 10: **Expert preferences across six criteria for symbolic-planning ablations and YuE2 versus separate LM+DiT.** Full planning receives more preferences than no planning on every criterion; unified generation is preferred overall both with and without planning. Each panel contains five direct comparisons. Full planning includes melody and chords; melody-only planning omits chords. The first three rows compare planning conditions within YuE2. The last two compare YuE2's AR–NAR Mixture-of-Transformers (Deng et al., 2025) with a separate language model and diffusion Transformer (LM+DiT), with and without planning. Dark blue favors the first-named condition in each row; light blue favors the second. Medium blue denotes ties. Each bar sums to 100% of judged responses, excluding unable-to-judge answers. Counts and CR1 intervals are in Appendix E.7.2.

Melody-and-chord planning receives more preferences than generation without planning in all six dimensions, with the largest preference margins in overall quality and musicality (Figure 10). Melody-only planning also receives more preferences than generation without planning across the six dimensions. Adding chords to the melody plan increases overall, musicality, text-alignment, and accompaniment preferences; audio quality and vocals are nearly balanced.

Preference shares favor the unified MoT over separate LM+DiT for overall quality, musicality, audio quality, and vocals both with and without symbolic planning. With melody-and-chord planning, accompaniment preferences also favor the MoT, while text alignment is nearly balanced. Without planning, text-alignment preferences also favor the MoT, while accompaniment is nearly balanced. The statistical tables distinguish these direct comparisons from averages over the two planning conditions or the two architecture conditions.

Table 19: **Expert pairwise preferences.** W, T, and L count preferences for the first-named system, ties, and preferences for the second-named system. $n = W + T + L$; U counts unable-to-judge answers. Preference and intervals are percentages. Intervals and tests use two-way CR1 clustering by evaluator and song. p tests a tie-adjusted preference of 50% (two-sided).

Criterion	Counts					Preference (%)		p
	W	T	L	n	U	Estimate	CR1 95% CI	
YuE2 (best-of-8) vs Mureka 9								
Overall	83	32	87	202	0	**49.0**	39.7–58.3	0.8316
Musicality	80	35	87	202	0	**48.3**	40.7–55.8	0.6477
Text alignment	65	54	83	202	0	**45.5**	37.8–53.3	0.2523
Audio quality	111	41	48	200	2	**65.8**	57.8–73.7	2.26×10^{-4}
Vocal	84	36	82	202	0	**50.5**	43.1–57.9	0.8930
Accompaniment	61	54	87	202	0	**43.6**	35.0–52.1	0.1369
YuE2 (best-of-8) vs Suno v4.5								
Overall	122	26	65	213	3	**63.4**	57.7–69.1	2.26×10^{-5}
Musicality	108	33	75	216	0	**57.6**	50.9–64.4	0.0269
Text alignment	103	55	56	214	2	**61.0**	55.4–66.6	2.76×10^{-4}
Audio quality	127	39	49	215	1	**68.1**	62.8–73.5	1.65×10^{-8}
Vocal	112	38	65	215	1	**60.9**	54.1–67.8	0.0024
Accompaniment	107	48	60	215	1	**60.9**	55.0–66.8	5.02×10^{-4}
YuE2 (best-of-8) vs Suno v5								
Overall	86	42	85	213	0	**50.2**	44.0–56.5	0.9402
Musicality	80	48	85	213	0	**48.8**	43.2–54.4	0.6765
Text alignment	66	73	74	213	0	**48.1**	42.3–54.0	0.5212
Audio quality	85	68	59	212	1	**56.1**	50.3–62.0	0.0400
Vocal	74	60	78	212	1	**49.1**	41.8–56.3	0.7956
Accompaniment	68	62	83	213	0	**46.5**	40.8–52.1	0.2159
YuE2 (best-of-8) vs Suno v5.5								
Overall	78	36	103	217	1	**44.2**	38.4–50.0	0.0518
Musicality	73	51	94	218	0	**45.2**	39.0–51.3	0.1220
Text alignment	59	71	85	215	3	**44.0**	39.0–48.9	0.0184
Audio quality	87	59	70	216	2	**53.9**	48.9–59.0	0.1221
Vocal	75	43	100	218	0	**44.3**	38.9–49.6	0.0359
Accompaniment	75	50	93	218	0	**45.9**	40.8–51.0	0.1095
YuE2 (best-of-8) vs Suno v6								
Overall	66	19	124	209	1	**36.1**	28.0–44.2	0.0012
Musicality	62	22	126	210	0	**34.8**	27.5–42.1	1.06×10^{-4}
Text alignment	57	55	97	209	1	**40.4**	33.9–46.9	0.0047
Audio quality	74	56	79	209	1	**48.8**	40.9–56.7	0.7631
Vocal	70	36	103	209	1	**42.1**	34.4–49.8	0.0444
Accompaniment	60	41	108	209	1	**38.5**	30.9–46.1	0.0038
YuE2 (best-of-8) vs Suno v6 Wild								
Overall	87	33	98	218	1	**47.5**	40.2–54.7	0.4883
Musicality	79	40	99	218	1	**45.4**	39.3–51.6	0.1406
Text alignment	68	67	82	217	2	**46.8**	40.9–52.7	0.2759
Audio quality	103	56	57	216	3	**60.6**	55.7–65.6	7.51×10^{-5}
Vocal	86	46	85	217	2	**50.2**	44.1–56.4	0.9404
Accompaniment	82	48	88	218	1	**48.6**	42.7–54.6	0.6455

Continued on next page

Table 19 continued

Criterion	Counts					Preference (%)		p
	W	T	L	n	U	Estimate	CR1 95% CI	
YuE2 vs Mureka 9								
Overall	92	22	99	213	1	**48.4**	40.8–55.9	0.6655
Musicality	86	29	98	213	1	**47.2**	38.4–55.9	0.5219
Text alignment	72	46	95	213	1	**44.6**	38.2–51.0	0.0977
Audio quality	101	45	64	210	4	**58.8**	50.3–67.3	0.0422
Vocal	80	39	95	214	0	**46.5**	37.6–55.3	0.4306
Accompaniment	72	43	98	213	1	**43.9**	37.0–50.8	0.0836
YuE2 vs Suno v4.5								
Overall	101	25	78	204	3	**55.6**	49.2–62.0	0.0831
Musicality	96	34	74	204	3	**55.4**	49.4–61.4	0.0782
Text alignment	75	59	68	202	5	**51.7**	46.4–57.0	0.5137
Audio quality	98	44	63	205	2	**58.5**	52.9–64.2	0.0039
Vocal	105	27	74	206	1	**57.5**	50.6–64.4	0.0327
Accompaniment	92	45	67	204	3	**56.1**	49.7–62.5	0.0607
YuE2 vs Suno v5								
Overall	79	39	97	215	1	**45.8**	38.1–53.6	0.2825
Musicality	78	42	94	214	2	**46.3**	40.2–52.3	0.2196
Text alignment	66	67	81	214	2	**46.5**	40.5–52.5	0.2496
Audio quality	86	68	58	212	4	**56.6**	49.8–63.4	0.0583
Vocal	65	53	97	215	1	**42.6**	36.2–48.9	0.0226
Accompaniment	64	55	96	215	1	**42.6**	35.8–49.3	0.0323
YuE2 vs Suno v5.5								
Overall	77	42	89	208	0	**47.1**	39.9–54.4	0.4295
Musicality	71	38	99	208	0	**43.3**	35.9–50.6	0.0729
Text alignment	60	67	79	206	2	**45.4**	39.0–51.8	0.1529
Audio quality	88	48	67	203	5	**55.2**	47.8–62.5	0.1647
Vocal	68	39	100	207	1	**42.3**	35.6–49.0	0.0244
Accompaniment	69	51	88	208	0	**45.4**	37.0–53.9	0.2825
YuE2 vs Suno v6								
Overall	65	38	107	210	1	**40.0**	32.2–47.8	0.0131
Musicality	58	54	99	211	0	**40.3**	33.6–46.9	0.0050
Text alignment	51	71	89	211	0	**41.0**	35.5–46.4	0.0017
Audio quality	87	47	73	207	4	**53.4**	43.2–63.5	0.5060
Vocal	62	47	102	211	0	**40.5**	32.9–48.1	0.0158
Accompaniment	59	55	97	211	0	**41.0**	33.6–48.4	0.0186
YuE2 vs Suno v6 Wild								
Overall	89	36	84	209	2	**51.2**	44.0–58.4	0.7416
Musicality	82	49	78	209	2	**51.0**	45.7–56.3	0.7185
Text alignment	58	71	79	208	3	**45.0**	39.0–50.9	0.0949
Audio quality	86	66	57	209	2	**56.9**	49.0–64.9	0.0852
Vocal	85	42	83	210	1	**50.5**	43.4–57.5	0.8929
Accompaniment	81	53	77	211	0	**50.9**	43.7–58.2	0.7933

Continued on next page

Table 19 continued

Criterion	Counts					Preference (%)		p
	W	T	L	n	U	Estimate	CR1 95% CI	
YuE2 vs YuE2 with melody-only planning								
Overall	91	35	81	207	0	**52.4**	46.9–57.9	0.3819
Musicality	80	58	68	206	1	**52.9**	47.3–58.5	0.3033
Text alignment	65	92	50	207	0	**53.6**	48.2–59.0	0.1845
Audio quality	75	52	78	205	2	**49.3**	42.8–55.8	0.8219
Vocal	73	57	76	206	1	**49.3**	44.2–54.4	0.7762
Accompaniment	81	66	60	207	0	**55.1**	50.1–60.0	0.0443
YuE2 vs YuE2 without planning								
Overall	104	34	73	211	1	**57.3**	52.1–62.6	0.0070
Musicality	95	54	62	211	1	**57.8**	51.6–64.0	0.0141
Text alignment	78	78	55	211	1	**55.5**	49.5–61.4	0.0732
Audio quality	87	66	57	210	2	**57.1**	51.0–63.3	0.0229
Vocal	87	57	66	210	2	**55.0**	49.1–60.9	0.0927
Accompaniment	90	57	65	212	0	**55.9**	50.4–61.4	0.0357
YuE2 with melody-only planning vs YuE2 without planning								
Overall	97	39	81	217	0	**53.7**	47.9–59.5	0.2064
Musicality	95	44	78	217	0	**53.9**	47.4–60.4	0.2316
Text alignment	75	71	69	215	2	**51.4**	46.1–56.7	0.5973
Audio quality	81	75	60	216	1	**54.9**	49.9–59.8	0.0532
Vocal	94	52	71	217	0	**55.3**	50.6–60.0	0.0295
Accompaniment	80	66	70	216	1	**52.3**	46.7–57.9	0.4110
YuE2 vs Separate LM+DiT								
Overall	111	23	74	208	1	**58.9**	52.4–65.4	0.0084
Musicality	99	42	67	208	1	**57.7**	52.3–63.1	0.0065
Text alignment	70	67	69	206	3	**50.2**	45.6–54.9	0.9162
Audio quality	100	45	61	206	3	**59.5**	54.6–64.4	3.06×10^{-4}
Vocal	99	49	60	208	1	**59.4**	55.3–63.4	2.69×10^{-5}
Accompaniment	88	60	60	208	1	**56.7**	51.7–61.8	0.0097
YuE2 without planning vs Separate LM+DiT without planning								
Overall	93	35	82	210	2	**52.6**	47.1–58.1	0.3456
Musicality	91	48	72	211	1	**54.5**	49.8–59.2	0.0602
Text alignment	76	64	71	211	1	**51.2**	45.2–57.2	0.6950
Audio quality	88	51	70	209	3	**54.3**	49.5–59.2	0.0807
Vocal	94	46	71	211	1	**55.5**	49.2–61.7	0.0882
Accompaniment	76	57	79	212	0	**49.3**	42.2–56.4	0.8416

Table 20: **Average expert preferences.** Model pairs receive equal weight. Preference and intervals are percentages; n counts judgments and U counts unable-to-judge answers. Intervals and tests use two-way CR1 clustering by evaluator and song. p tests a tie-adjusted preference of 50% (two-sided).

Criterion	Counts		Preference (%)		p
	n	U	Estimate	CR1 95% CI	
YuE2: average over six proprietary systems					
Overall	1259	8	**48.0**	43.8–52.2	0.3498
Musicality	1259	8	**47.2**	43.6–50.8	0.1263
Text alignment	1254	13	**45.7**	42.4–49.0	0.0110
Audio quality	1246	21	**56.6**	51.6–61.6	0.0105
Vocal	1263	4	**46.6**	42.7–50.6	0.0971
Accompaniment	1262	5	**46.7**	42.8–50.6	0.0920
YuE2 (best-of-8): average over six proprietary systems					
Overall	1272	6	**48.4**	44.8–52.0	0.3813
Musicality	1277	1	**46.7**	43.4–50.0	0.0484
Text alignment	1270	8	**47.6**	44.7–50.6	0.1111
Audio quality	1268	10	**58.9**	55.6–62.2	1.70×10^{-6}
Vocal	1273	5	**49.5**	45.8–53.2	0.7931
Accompaniment	1275	3	**47.3**	44.4–50.3	0.0730
Architecture: average of planning and no-planning comparisons					
Overall	418	3	**55.8**	51.8–59.7	0.0050
Musicality	419	2	**56.1**	52.2–60.0	0.0027
Text alignment	417	4	**50.7**	47.1–54.3	0.6928
Audio quality	415	6	**56.9**	53.2–60.5	3.98×10^{-4}
Vocal	419	2	**57.4**	53.5–61.3	3.86×10^{-4}
Accompaniment	420	1	**53.0**	49.0–57.0	0.1353
Planning: average of melody-and-chord and melody-only comparisons					
Overall	428	1	**55.5**	51.1–59.9	0.0153
Musicality	428	1	**55.9**	51.5–60.2	0.0086
Text alignment	426	3	**53.4**	49.2–57.7	0.1143
Audio quality	426	3	**56.0**	51.9–60.1	0.0048
Vocal	427	2	**55.1**	51.3–59.0	0.0098
Accompaniment	428	1	**54.1**	49.6–58.6	0.0752

F Additional Composition, Editing, and Cover Results

F.1 Generated-Score Agreement with Audio

Table 21: **Score–audio consistency on WSB.** Mismatched audio uses the other score for the same prompt. Rhythm reports vocal onset-interval F1. Values are prompt means. Tempo accuracy allows 8% error; lower log error is better.

	Melody	Chords	Rhythm	Key		Tempo		Form	
Audio	Seq.↑	Seq.↑	F1↑	Exact↑	Wtd.↑	Log err.↓	Acc.↑	Content↑	Bound.↑
Corresponding	**0.9464**	**0.9246**	**0.9396**	**0.9307**	**0.9515**	**0.0142**	**0.9869**	**0.7799**	**0.8724**
Mismatched	0.2160	0.2473	0.6119	0.3004	0.3629	0.1131	0.6806	0.1547	0.1454
Without score	0.2055	0.1897	0.5875	0.2234	0.2831	0.1468	0.5366	0.1236	0.1245

Melody and harmony. For a nonempty reference event sequence R and recovered sequence P, we use

$$S(R, P) = 1 - \frac{D_{\text{edit}}(R, P)}{\max(|R|, |P|)}, \tag{13}$$

where global edit distance assigns cost zero to exact matches and unit cost to insertions, deletions, and substitutions. Both endpoints and every event enter the comparison. The melody sequence uses sounding pitches in semitones: tied notes are merged, separately articulated repeated notes remain separate, and pitches in different octaves are distinct. Each chord is represented by its root and triad quality (major, minor, diminished, augmented, or suspended second/fourth), omitting extensions and inversions. Consecutive chords with the same representation are merged. Unknown recovered labels cannot match known reference labels. These scores compare the resulting note and chord sequences in order, independently of onset times and durations. Edit-distance alignment can resume matching subsequent notes or chords after an insertion or deletion.

Rhythm. We compare consecutive vocal onset intervals, including intervals across rests, in quarter-note beats. The maximum number of one-to-one matches that preserve interval order defines precision and recall over all recovered and target intervals, respectively. A target interval d matches an estimate e when $|e - d| \leq \min(0.125, 0.25d)$ beats. The relative bound prevents doubling short intervals from counting as a match, while the absolute cap limits deviations across long rests. Their harmonic mean is the rhythmic interval F1. This pitch-independent measure compares notated onset spacing on the transcriber's quantized beat-subdivision grid, without a time warp. Matching rhythmic patterns can receive credit across different melodies. The measure excludes note-off articulation and expressive microtiming.

Global transposition. The primary scores use the specified pitches. For a companion comparison, we shift the reference by one integer number of semitones that maximizes the mean of the available melody and chord similarities. The search includes all shifts that keep reference melody pitches in the MIDI range 0–127; without melody, only a shift modulo 12 is identifiable. Ties prefer the smallest absolute shift, then the smaller signed shift. This single shift applies to melody, chord roots, key, and all sections. Key and form never choose their own shifts. Local octave errors and inconsistent transposition between melody and accompaniment remain errors. The same procedure applies to every score–audio pairing.

Key. We compare the declared score key with the recovered tonic pitch class and major/minor mode, treating enharmonic spellings as equivalent. Alongside exact agreement, we use the key-relation weights used in MARBLE (Yuan et al., 2023): 1 for the same key, 0.5 for a predicted tonic a perfect fifth above the reference in the same mode, 0.3 for relative major/minor, 0.2 for parallel major/minor, and 0 otherwise. The fifth credit is directional. For a single target key, scores average recovered key regions by their original audio duration, from the first to the last key annotation; internal unknown gaps receive zero. For changing keys, we use the melody correspondence, or chord correspondence if melody is absent, and weight agreement at event onsets by reference event duration. Deleted reference events receive zero. These scores measure agreement with the tonic and mode declared in the score.

Tempo. For each recovered beat interval, let Δq be its length in quarter notes and Δt its duration in seconds. The tempo estimate $\hat{b}$ is the Δq-weighted median of $60\Delta q/\Delta t$, computed before any time alignment. Against target BPM b, we report $|\log(\hat{b}/b)|$ and accuracy under $|\hat{b}/b - 1| \leq 0.08$. A separate accuracy permits $\min_{k\in\{0.5,1,2\}} |\hat{b}/(kb) - 1| \leq 0.08$, accounting for half- and double-tempo ambiguity. We report this companion alongside the primary specified-tempo accuracy. Every evaluated score specifies one song-wide tempo.

Musical form. We first merge adjacent same-role fragments, concatenating all their musical events. This tolerates splitting one verse into two annotations while retaining the content of a repeated chorus. For sections i and j, let c_{ij} be the mean of their available melody and chord sequence similarities. Their pair similarity is $\mathbf{1}[r_i = r_j]c_{ij}$, where r is the section role. Global ordered section alignment assigns cost 1 to an insertion or deletion and cost $1 - \mathbf{1}[r_i = r_j]c_{ij}$ to a substitution. As above, similarity is one minus total cost divided by the longer section count. Unknown roles receive

no match credit. Reference sections without observable melody or harmony are unresolved. The score measures section roles and order using their melody and harmony content; boundary timing is evaluated by the separate boundary F1 below. Repeated sustained material with indistinguishable content can match a prolonged occurrence.

Section boundaries. After the same adjacent-role merging, we compare internal section starts, excluding the trivial recording start and end. One-to-one boundary matches within two quarter-note beats define F1 over all target and estimated boundaries. Only a global translation is allowed, estimated as the median target-minus-recovered onset difference of exact pitch matches in the melody sequence alignment; absent anchors give a zero shift. Boundaries do not fit their own alignment, and no scaling or local warp is used. Each control receives the same procedure. Corresponding-audio F1 is 0.8724 with translation and 0.5358 at the original recovered beat origin. The first measures relative section placement, while the second also includes start-position and analyzer beat-origin discrepancies.

Companion results. Allowing one global transposition leaves the corresponding-audio melody and chord scores at 0.9465 and 0.9245. The other same-prompt recording reaches only 0.3152 and 0.4658, respectively (Table 22). The content advantage therefore persists after removing global key and register differences. Individual key or chord scores can decrease slightly because the shared shift optimizes the joint melody/chord objective. Half/double-tempo tolerance raises corresponding-audio tempo accuracy from 0.9869 to 1.0000; the specified-tempo result remains the primary comparison.

Table 22: **Companion comparisons allowing global transposition or tempo ambiguity.** The first five rows share one melody/chord-derived pitch shift per comparison. The last row permits half or double the target tempo, with the same 8% tolerance. Higher is better. Values are prompt means.

Measure	Corresponding	Mismatched	Without score
Melody sequence ↑	**0.9465**	0.3152	0.3050
Chord sequence ↑	**0.9245**	0.4658	0.4199
Key: exact agreement ↑	**0.9280**	0.7527	0.6429
Key: weighted agreement ↑	**0.9489**	0.8157	0.7320
Musical form ↑	**0.7798**	0.2749	0.2276
Tempo: half/double tolerant ↑	**1.0000**	0.7094	0.5654

Denominators and aggregation. All 768 recordings were analyzed without transcription failures. Four of 384 generated scores are truncated and invalid references. Of the remaining 380, one lacks vocal notes and another lacks explicit chords, leaving 379 scores for each melody/harmony metric; rhythm also covers 379. Key and tempo cover 380 scores; form covers 378 because two references have unresolved section boundaries or content. One of these has a zero-duration section marker and is also excluded from boundary F1, leaving 379 boundary targets. Each dimension covers 191 prompts. Missing reference content is undefined; a nonempty reference with empty recovered content scores zero. Unknown predictions remain in the denominator. We average available candidates within each prompt, then weight prompts equally. Paired comparisons retain the same valid cases in both conditions.

Sensitivity to missing reference content. Melody and harmony similarities remain 0.9343 and 0.9122 when all 192 prompts are retained and undefined bounded scores receive zero credit. Under this convention, corresponding-audio means are 0.9414 for weighted key agreement, 0.9766 for tempo accuracy, 0.7696 for form, 0.9277 for rhythm, and 0.8628 for boundary F1. This lower-bound convention does not apply to the unbounded tempo error.

The two-candidate assignment task compares the sum of corresponding scores with the sum under the swapped assignment. Four truncated scores affect three of the 192 WSB prompts, leaving 189 complete pairs. Melody excludes one further prompt lacking vocal notes; chords exclude a different prompt lacking explicit chords. Each task therefore includes 188 prompts, with 187 in common.

The correct assignment has a higher similarity sum in all 188 cases for each metric, without ties. This assignment accuracy measures identification of each score's corresponding recording from the two candidates.

Measurement checks and scope. The metrics were developed on this recording cohort. Music experts checked and corrected each SheetSage2 transcription. Halving both rhythm tolerances leaves corresponding-audio F1 at 0.9396; doubling them gives 0.9506. Boundary windows of one and four beats give F1 0.8709 and 0.8800, respectively. Agreement remains high across the tested rhythm and section-boundary tolerances.

F.2 Controlled Score Editing on WildSongBench

Setup and aggregation. We use the first generated score for each of the 192 WildSongBench prompts. Key and tempo edits apply to 191 parsable scores; melody, harmony and rhythm edits apply to 190, 181 and 107 scores, respectively. Two seeds per condition yield 3,844 recordings, including unedited controls. All outputs are retained, including 32 at the six-minute cap. Lyrics, style, checkpoint, sampling settings and the main evaluation acoustic decoder remain fixed. We average seeds and edit strengths within each song, then weight songs equally. Rhythm conditions use matched source scores and generation seeds.

Interventions. Key edits transpose all voices, chord roots, bass notes and key declarations by $-5, -2, +2, +5$ semitones. Tempo edits multiply notated BPM by 0.8 or 1.2 and round to the nearest integer; evaluation uses this rounded target. Melody and rhythm edits affect up to four bars of the first chorus. Melody edits raise vocal pitches by two scale steps. Rhythm edits exchange the durations of adjacent quarter and eighth notes with distinct pitches: $(1, 1/2) \leftrightarrow (1/2, 1)$ in quarter-note beats. We select nonoverlapping pairs in score order, preserving pitches, each pair's start and end, other voices, chords and tempo. The 107 eligible scores contain 210 pairs; eligibility depends only on the source score. Both rhythm conditions use the training ABC serialization rules and retain each source's unit note length. Independent ABC-to-MIDI checks verify pitch and timing.

Harmony edits process complete chord spans in score order within up to four bars at the start of the first chorus. We shift each root and any specified bass by the first of $(+5, +7, +2, +3)$ semitones that avoids matching either current neighbor in both root and triad quality. Chord quality, timing, and vocal notes remain fixed, and we repeat these substitutions in matching phrases. Bar-aligned windows within sections qualify by identical chord progressions, or by melody pitch-class agreement and onset F1 of at least 0.80 plus chord agreement of at least 0.75 (at least four notes; $1/8$-beat onset tolerance). Transferred chords must match the original labels and boundaries, and overlapping substitutions must agree. Edits cover 43.86% of score duration on average.

Attainment and preservation. Harmony adherence and all content-preservation metrics use SheetSage2-AR transcriptions, with BF16 inference, 300-second windows, 200-second overlap and 100-second lookahead. Melody and rhythm adherence, key scores and the MIDI controls use SheetSage2-Prober transcriptions. For editing evaluation, music experts review and correct the transcriptions of generated songs. For melody and harmony attainment and content preservation, up to eight unchanged notes on either side locate local edits, with edited pitches masked during alignment. Melody matching maximizes ordered onset matches within one beat, breaking ties by total onset error; attainment is exact target-pitch accuracy on changed notes. Harmony attainment measures duration-weighted root-and-triad agreement in the designated chorus window, aligned through unchanged vocal notes. Missing events and unresolved contexts score zero for attainment.

For each preservation metric, we exclude undefined scores, average available seeds within each song, and weight evaluable songs equally. For melody and rhythm edits, preservation compares unchanged melody and chord sequences with the source score, penalizing insertions, omissions and substitutions. These comparisons resolve 368 melody-edit and 207 rhythm-edit recordings. Harmony preservation compares the full vocal melody sequence with the source score and measures duration-weighted agreement on unchanged chords outside all edited spans. Chord comparisons use time supported by vocal-note anchors in 170 songs.

Rhythm measurement. We measure onset placement relative to unchanged neighboring onsets, allowing audio to differ in absolute offset and local tempo. We first align full pitch sequences by unit-cost edit distance and identify note correspondences shared by all optimal alignments. Consecutive uniquely matched unchanged onsets define local spans of at most eight quarter-note beats. Within each span, we resolve alignment ties using unchanged onsets within $1/4$ beat of their linearly interpolated source positions, retaining the minimum pitch-edit cost. Edited onsets do not enter this constraint. A correspondence is resolved only when every remaining optimal alignment assigns the same equal-pitch partner without a deletion alternative.

For an edited pair (i, j), let k be the next resolved unchanged onset after j within eight beats of i, s denote source onsets in quarter-note beats, and t the corresponding audio onsets. The normalized interval is

$$\widehat{\Delta}_{ij} = \frac{t_j - t_i}{t_k - t_i}(s_k - s_i). \tag{14}$$

An onset succeeds if this interval is within $1/8$ beat of the edited target and closer to that target than to the original interval. Missing or ambiguous correspondences count as misses in Table 3; coverage is 80.36%. On pairs observable in both conditions (74.95% coverage), mean target-onset error falls from 0.501 to 0.037 quarter-note beats; 96.79% of these pairs move closer to the target.

For each source, we synthesize three instrumental MIDI controls: the target exchange, the unchanged score, and a $1/4$-beat shift in the opposite direction. The target controls achieve 94.08% onset accuracy; unchanged and opposite-shift controls score zero. Table 23 reports these controls alongside generated songs, with correspondence coverage and conditional accuracy.

Table 23: **Rhythm exchange accuracy and correspondence coverage.** All scores are percentages.

Condition	Onset accuracy	Coverage	Conditional accuracy
Unedited	0.00	88.05	0.00
Edited	73.43	80.36	91.37
MIDI: target exchange	94.08	94.08	100.00
MIDI: unchanged	0.00	94.08	0.00
MIDI: opposite shift	0.00	93.54	0.00

Onset accuracy includes every edited pair; unresolved correspondences score zero. Coverage is the fraction of pairs with resolved correspondences; conditional accuracy uses those pairs. We average pairs within recordings, seeds within sources, and sources equally. All conditions are scored against the edited target. MIDI rows are instrumental controls. Conditional accuracy is defined for 195/214, 188/214, 102/107, 102/107, 102/107 recordings, respectively.

Key and tempo metrics. Weighted key scores compare recovered tonic and mode with the edited target using the weights in Appendix F.1, averaging over original audio duration between the first and last annotation; unknown intervals score zero.

Tempo uses madmom (Böck et al., 2016) 0.16.1 recurrent beat activations and dynamic Bayesian network tracking at 100 frames/s, with a 30–300 BPM range and transition parameter 100. We take the median instantaneous BPM between consecutive detected beats; neither the score nor the target enters beat tracking. Acc2 (Schreiber et al., 2020) accepts estimates within 4% of the target multiplied by any of $\{1/3, 1/2, 1, 2, 3\}$; Acc1 accepts only the target beat level. Across 764 recordings, Acc2 is 95.68% and Acc1 is 75.92%.

Song quality and lyrics. SongBench (Wu et al., 2026) quality averages seven dimensions. Greedy audio-only Qwen3-ASR (Shi et al., 2026) transcripts supply phoneme error rate (PER), using tone-marked pinyin for Chinese and English phonemes for English. Character error rate (CER) retains all scripts and digits after Unicode normalization, case folding and removal of section tags and punctuation. Table 24 reports quality and lyric retention for melody, harmony, rhythm, and key edits.

Table 24: **Song quality and lyric retention after editing on WildSongBench.**

Condition	Quality ↑	PER ↓	CER ↓
Unedited	6.704	0.184	0.181
Melody	6.717	0.191	0.181
Harmony	6.674	0.179	0.175
Key	6.639	0.190	0.188
Unedited (rhythm)	6.728	0.214	0.197
Rhythm	6.731	0.185	0.175

Quality averages seven SongBench dimensions. PER and CER are phoneme and character error rates. Rhythm rows use matched source scores and generation seeds.

F.3 Agentic Music Editing: A Case Study

Adapting a song to a new style can require revising the composition itself. YuE2's readable ABC score gives an external language-model agent direct access to melody, harmony, and form. In *agentic music editing*, a user shapes the song through musical dialogue: the agent turns requests into revisions of the score, style description, and lyrics, while the same YuE2 model renders each version as a complete song.

The Last Train begins as Mandarin pop and is reworked into English jazz through collaboration between a researcher and a language-model agent. Nine curated steps and fourteen rendered versions capture changes to harmony, melodic phrasing, and lyrics, including a saxophone interlude built around *Twinkle, Twinkle, Little Star*. The researcher steers the musical direction and chooses the final recording by listening; retained alternatives and corrections show how the composition develops through feedback. The score is both the plan that guides YuE2 and the composition that the user and agent continue to write together.

F.4 Illustrative Planning Comparison

Musical development in a paired example. To illustrate how planning affects musical development, Figure 5 compares expert-corrected scores of songs generated with planning (left) and without planning (right), using the same prompt and lyrics. On the left, the **purple** motif varies pitch level and intervals while retaining its low–low–high–high contour, then returns an octave higher with longer endings in the chorus. This recognizable variation strengthens thematic continuity and makes the motif more memorable. The right **purple** verse repeats B–D–E–E four times, with little variation and weaker chorus recall, leaving the melody more static. In the complete scores (Appendix G), the left **orange** fills connect vocal phrases and prepare a smoother chorus entry; comparable responses are sparse in the right transcription after the introduction. In the left **blue** passages, syncopation (accents between the main beats) and varied note lengths animate the vocal phrasing, while the right's continuous sixteenth-note texture sounds more monotonous. The **green** closing line is reprised on the left, allowing the ending to settle; the right's single statement feels abrupt. Harmonic motion also sounds more continuous on the left. In expert listening, the right **gray** $E\mathrm{m}$–$D\sharp\mathrm{aug}$–$B\mathrm{m}$ transition interrupts the descending bass motion implied by its opening. This is a judgment of the recorded performance, not a claim that the chord progression is intrinsically invalid; neither score separately transcribes bass.

G FULL SCORES FOR THE PLANNING COMPARISON

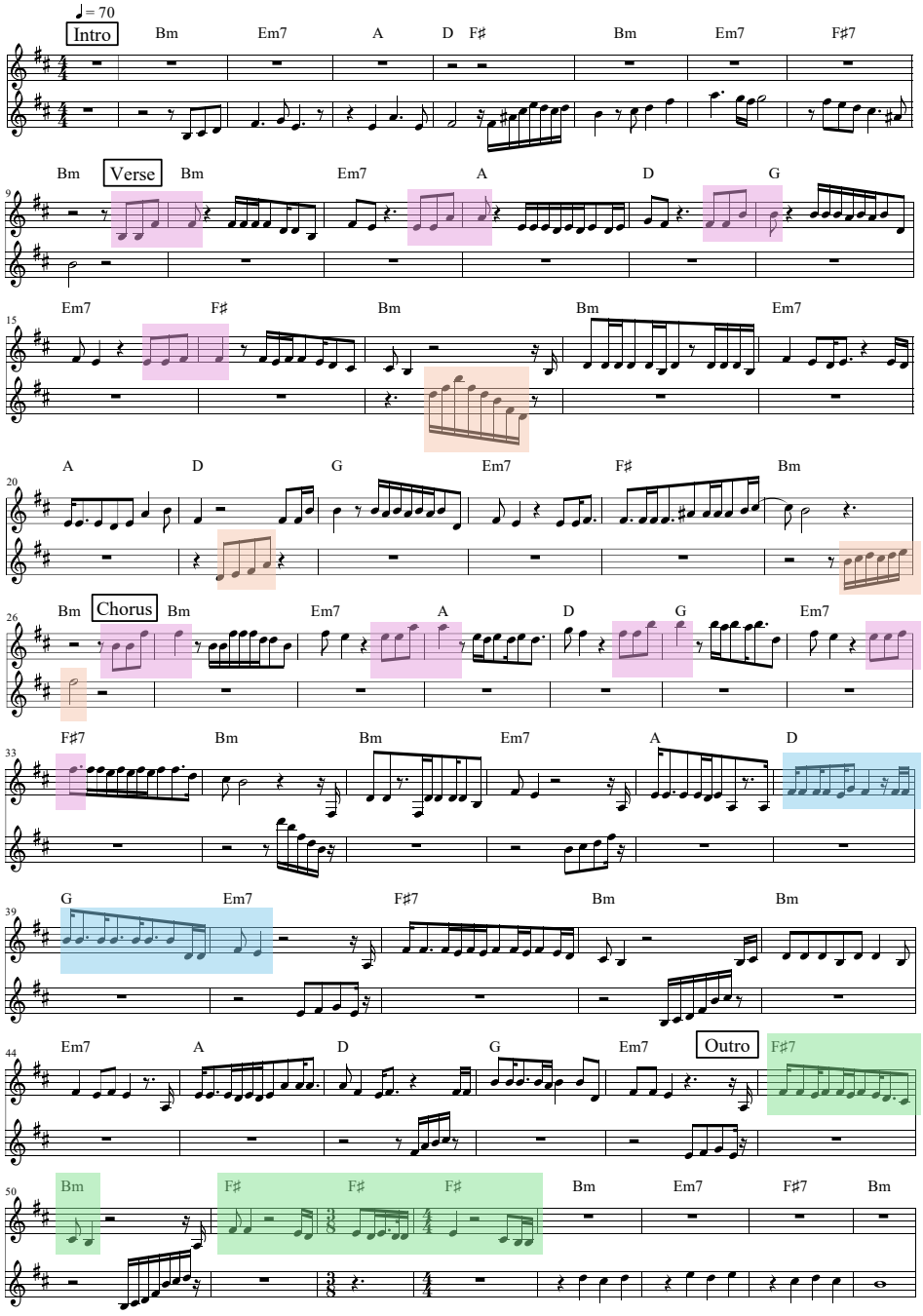

Figure 11: **With planning: melodic variation, instrumental responses, and a more extended ending.** This song uses the same prompt and lyrics as the song without planning in Figure 12. In each staff pair, the upper line shows the vocal melody and the lower line the instrumental melody; letters above name chords. **Purple:** a short melodic idea keeps its low–low–high–high shape while varying its pitches, then returns an octave higher with longer endings in the chorus. **Orange:** instrumental fills answer the voice between phrases and lead into the chorus. **Blue:** varied note lengths and accents between main beats (syncopation) give the phrasing rhythmic variety. **Green:** repeating the closing vocal line gives the ending more time to settle. Experts corrected the generated plan against the completed recording for display; these corrections were not used to regenerate audio.

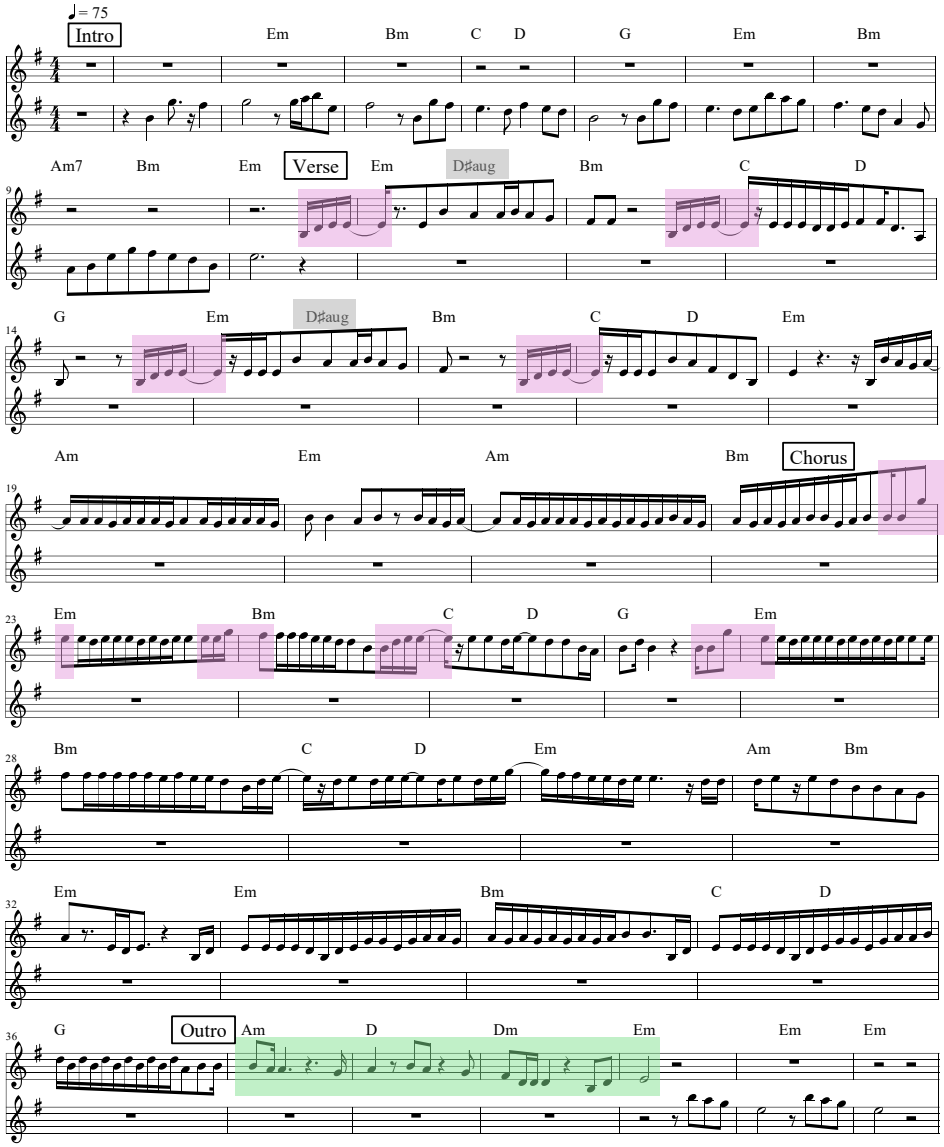

Figure 12: **Without planning: less melodic variation and a shorter close in this paired example.** The song uses the same prompt and lyrics as Figure 11. In each staff pair, the upper line shows vocal melody and the lower line instrumental melody; letters above name chords. **Purple:** the verse repeats the same B–D–E–E figure four times, with a less direct connection to the chorus than in the planned song. **Green:** the closing vocal line is stated once, making the ending feel more abrupt. **Gray:** experts hear the $E\mathrm{m}$–$D\sharp\mathrm{aug}$–$B\mathrm{m}$ chord transition as interrupting a descending bass motion. This is a listening interpretation: the score has no separate bass transcription, and the chord progression is not intrinsically invalid. Experts corrected the SheetSage2 transcription against the completed recording for display, without regenerating audio.

Instrumental and rhythmic contrast. The planned song's orange fills connect vocal phrases; comparable instrumental responses are sparse after the unplanned song's introduction. Rests in the lower staff do not mean that all accompaniment disappears. The planned song's blue passages vary note lengths and accent placement; the unplanned song's dense runs of equal-length sixteenth notes sound more monotonous in the compared passages. These observations describe this pair, while Figure 4 reports the aggregate listening results.