

FUMENLAB: PLAN-CONDITIONED JOINT GENERATION AND REFINEMENT OF OSU!TAIKO CHARTS

Ryo Mitoma

Yokohama National University

mitoma-ryo-ph@ynu.jp

ABSTRACT

We describe FumenLab for the MIREX 2026 Rhythm Game Chart Generation task. An osu!taiko chart maps music to don (drum-centre) and kat (rim) hits, with long drumroll and denden objects. The system has two core learned stages: a structural planner predicts bar-level controls, then a plan-conditioned joint model generates note timing, don/kat class, and long-note starts. Downstream stages decode local colour patterns, refine large-note labels, classify long-note type and duration, predict Go-Go Time and performance controls, and apply conservative repair. On a fixed 50-chart development subset, the selected note-generation pipeline obtains 0.4939 four-class ordinary-note F1, 0.3756 large-note F1, and 0.7462 onset F1 after long notes are expanded to sixteenth-note hits; all 50 outputs pass native parsing. These are configuration-selection results, not hidden-test or player-evaluation results.

1. TASK AND SYSTEM

In osu!taiko, players strike the drum centre for don notes and the rim for kat notes, while drumrolls and denden occupy intervals. Chart generation must choose musically appropriate timing, note class, size, and long-note boundaries for each difficulty. The MIREX task supplies audio and uninherited timing points for Kantan, Futsuu, Muzukashii, Oni and Inner Oni [1]. Figure 1 shows our plan-conditioned generator and downstream note, Go-Go and validation stages. Supplied tempo and beat positions are preserved while Go-Go flags may change.

1.1 Audio and Plan Conditioning

The audio encoder is frozen MERT v2 FullSong [2, 5, 6]. Audio is decoded to 24 kHz mono; one-based block outputs 12, 23 and 24 provide three 1024-dimensional feature vectors per frame. The implementation records frame centres at $15 + 40n$ ms in source-audio coordinates. For each beat, 32 centres at $(i + 0.5)/32$ select nearest encoder frames; the detail stream is interpolated to output positions $q/48$. Online features are extracted once and reused downstream; no prepared training-feature arrays are needed at inference.

The structural planner (configuration E2, 13.14M trainable parameters) uses a learned layer mixture, a 1024-to-512 projection, difficulty and sinusoidal position embeddings, and four 8-head Transformer layers with feed-forward width 2048 and dropout 0.1. Duration-weighted frame overlap produces bar features. Its sigmoid head predicts five controls in $[0, 1]$, supervised from reference charts: circle density per beat divided by four (clipped at one); the fraction of adjacent circle gaps at most 0.26 beats; adjacent colour-switch fraction; longest unoccupied span divided by bar length; and merged long-note occupancy divided by bar duration. The epoch-16 planner is frozen before exporting predicted plans for Joint training; reference plans and planner gradients are absent from that stage. The predicted five-vector is projected to width 256 and added to each corresponding beat summary in Joint.

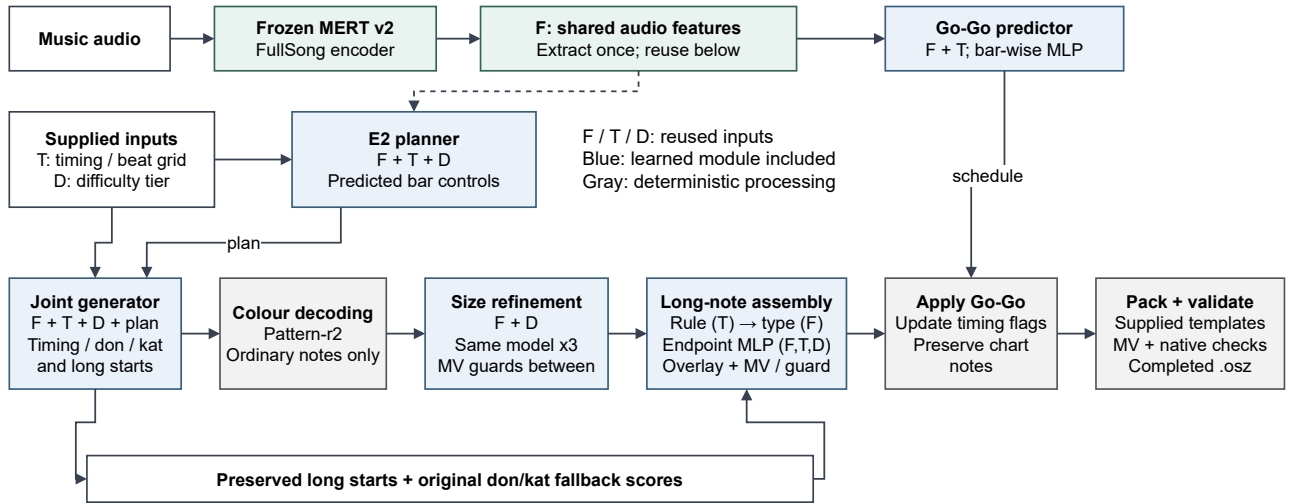
1.2 Joint Timing, Colour, and Long Starts

The beat-grid backbone is inspired by ITGPT’s hierarchical beat representation [4]. It combines within-beat audio detail with one bidirectional attention block at each group size of 1, 4, 16, 32 and 64 beats, followed by one global block (feed-forward width 512). Three segment-masked temporal detail blocks use dilations 1, 4 and 16. Our adaptation uses frozen MERT features, predicted plan conditioning, and joint taiko labels instead of ITGPT’s separate autoregressive step-selection model.

The Joint model has 4.56M trainable parameters, width 256, 8 attention heads, within-beat detail width 128 and dropout 0.1. Audio layer mixing/projection, difficulty and BPM conditioning are learned with the task backbone. It outputs logits z_0, z_D, z_K, z_L at each of 48 positions per beat; $p = \text{softmax}(z)$ represents empty, don, kat and long start. The onset score is $q = p_D + p_K + p_L$. Training applies weighted binary cross-entropy to $(z_0, \log \sum_{c \neq 0} e^{z_c})$ and conditional type cross-entropy to (z_D, z_K, z_L) at reference onsets, with equal loss coefficients.

Decoding first accepts $q \geq \tau_d$, with frozen thresholds (0.7, 0.8, 0.8, 0.8, 0.8) in increasing difficulty order. It then chooses $\arg \max(p_D, p_K, 6p_L)$; the development-selected multiplier changes only type selection, not q . Times are rounded to integer milliseconds; coincident predictions retain the largest onset score. The binary-onset model was trained from fresh whole-model initialization for 1,500 updates; Joint continues those weights, and its selected checkpoint is at 500 additional updates.

For ordinary don/kat sequences, decoder r2 enumerates



Repeat for all five difficulty tiers. Go-Go prediction uses audio + bar timing, not generated notes or difficulty.

Figure 1. Inference repeated for five templates. Long starts bypass ordinary-note processing; provisional intervals are typed before learned endpoint prediction. Shared MERT features condition learned stages.

complete tilings by rhythm units of 2, 3, 4, 6, or 8 notes. It combines joint-model colour scores with training-corpus transition and motif probabilities. The fixed transition strength is 0.25, the motif share inside that term is 0.25, change cost is zero, and the rarity threshold is 0.99. Decoding is split at every long start: it may change ordinary-note colours but cannot move, delete, or recolour a long start.

1.3 Large and Long Notes

The selected size stage applies the same 15.81M-parameter MERT-v2-conditioned, ModernBERT-based refiner [3] three times, using its epoch-20 checkpoint and threshold 0.5. Its learnable audio pooling and 14 chart features feed eight width-384 ModernBERT layers (6 heads, local attention 128, global attention every third layer) and a KEEP/make-big/make-small head. Each pass changes only ordinary-note size labels. Between passes, the existing MV Problem-level finisher guard converts unusable large notes to small notes before the next pass consumes the chart.

A key design challenge was choosing a long-note representation that would preserve ordinary onset timing and don/kat classification. We generate long-note starts jointly with ordinary notes, then estimate their type and duration in downstream stages.

A provisional long interval ends one beat before the next generated onset, or at the earlier of audio end and start plus 1,000 ms if none follows. Nonpositive intervals revert to ordinary small notes using the original joint don/kat argmax. Four duration-weighted temporal bins concatenate the three MERT layers into 12,288 features. A 12,288–32–3 GELU MLP and a 12,288–32–2 roll-size hard gate classify valid provisional intervals as normal drumroll, large drumroll, or denden.

A learned endpoint then replaces each valid interval’s

provisional end, preserving its start and type. The 0.984M-parameter MLP (15,374–64–1, GELU, dropout 0.1) pools MERT features within ± 250 ms of the start, next onset, and their 25/50/75% intermediate positions, plus 14 gap, tempo, beat-phase, tier and type features. Audio end substitutes for a missing next onset. Standardized inputs predict $\log(1 + \text{duration}/\text{gap})$; decoded durations are bounded to 20–30,000 ms and audio end, then rounded to milliseconds. A prediction beyond the next onset reverts to the provisional end. Type classification is not rerun on the revised interval.

1.4 Go-Go Time, Controls, and Final Processing

Go-Go Time is predicted once per bar. The selected 0.393M-parameter model takes the 3072-dimensional, duration-weighted bar means of MERT v2 FullSong layers 12, 23, and 24, followed by a linear 128-unit layer, GELU, dropout 0.1, and one output logit. A bar is active at the calibration-set threshold 0.3779998. No singleton-joining or other interval-merging rule is used. The model consumes no notes, tier, scroll velocity, or reference Go-Go labels at inference.

The scroll-velocity stage uses a section-level predictor: section feature means and standard deviations feed a two-layer Transformer of width 128. Tiers 3–5 use its predictions; tiers 1–2 retain constant scroll velocity one. Hit-sound volume uses an epoch-6 MLP with a separate volume head. Dense inference runs on the public time grid, quantizes outputs in 5% increments, and uses no teacher event times. Mapset Verifier (MV) v2.0.0 [7] then runs through our pinned serverless CLI adapter with conservative deterministic repairs. Raw bytes are preserved; final native parsing and an outer MV pass check validity and stability.

Component	AdamW LR / decay	Batch unit	Selected	Training and selection
E2 planner	$2 \times 10^{-4} / 0.01$	8 charts	Epoch 16	Up to 20 epochs; development macro-MAE; patience 4.
Joint	$10^{-4} / 0.01$	64 chunks	Update 500	Binary pretraining 1,500 updates; then up to 1,000 Joint updates; evaluate every 100.
Size refiner	$1.566 \times 10^{-4} / 0.01$	512 windows	Epoch 20	60 epochs; checkpoints at 20/40/60; guarded downstream comparison.
Long type / size	$2 \times 10^{-3} / 10^{-4}$	64 intervals	Epochs 16 / 1	Up to 120 epochs; calibration macro-F1, then NLL.
Long endpoint	$10^{-3} / 10^{-3}$	128 intervals	Epoch 32	Train-derived calibration MAE; retrain on all 2,152 intervals.
Go-Go	$10^{-3} / 10^{-3}$	4,096 bars	Epoch 1	9 epochs; calibration BCE; threshold by calibration F1.

Table 1. Training for the generator, refiners, and Go-Go predictor. Constant learning rates; bf16 autocast except CPU float32 endpoint training. Joint batch 64 accumulates 16 microbatches of 4; size batch 512 accumulates microbatches of 32.

2. DATA AND TRAINING

The corpus uses the general ranked osu!taiko collection in the permitted 5 May 2019–31 December 2025 period, not only Featured Artist material. We use the recommended normalized difficulty-name mapping, retain the five requested tiers, and exclude unrecognized names and Hell Oni. The pre-existing calibration partition is not used; calibration is derived from training groups. Materialization also excludes one four-chart development group with duplicate beatmap IDs. The effective split contains 1,341 training charts from 363 source groups and 248 development charts from 64 groups. Recorded source-group, audio-byte, chart-byte and beatmapset identities do not overlap between splits, although this does not guarantee acoustic separation across re-encodings. Independently pretrained audio encoders remain frozen; we do not use pretrained Mapperatorinator audio-chart weights.

Targets use the nearest valid cell on the supplied 48-position beat grid. The native-boundary encoder rejects cell collisions and unrepresentable boundaries rather than dropping events. Circles map to don/kat regardless of size; drumroll and denden starts share the long-start label. Long interiors and ends are empty after endpoint labels are removed.

The E2 planner uses masked smooth-L1 supervision of normalized bar controls. Joint uses empty/onset weights 0.25/1 and inverse-square-root training-frequency type weights, normalized to mean one (don/kat/long: 0.2039/0.2026/2.5935). Its long-start multiplier and checkpoint are selected on the development data, so their reported behaviour is not an independent test estimate. Size training draws clean, hidden-big and inflated-small corruptions in proportions 0.4/0.3/0.3, changing up to three eligible notes per window without changing timing or colour. Loss averages the global means of operation classes present in the batch, preventing KEEP from dominating. Long-type and roll-size MLPs use inverse-frequency class-weighted negative log-likelihood. Three-class train/calibration/development interval counts are 1,912/236/460; the binary head uses only true rolls (355/74/83). Calibration groups are taken from training; development groups are unchanged. Calibration se-

lects macro-F1, breaking ties by NLL; a maximum of 120 epochs and 15 stale epochs limits training. Decoder r2 estimates transition and motif counts only from the training split.

Table 1 summarizes optimization. Models are trained separately with frozen audio features; Joint continues the 1,500-update binary-onset model while training its backbone, feature adapter and expanded type head with fixed predicted E2 plans. Joint training uses shuffled, non-overlapping chunks of at most 256 beats, with gradient accumulation. Additional update 500 was selected using the recorded ± 20 ms development metrics. Size training completed 60 epochs; epoch 20 was retained from downstream development comparisons.

Endpoint training uses smooth-L1 loss and gradient norm clipping at one, with seed 20260912. An initial 1,916/236 train/calibration split selects epoch 32 by mean endpoint MAE on clean and synthetically shifted calibration inputs (maximum 80 epochs, patience 10). The selected model is retrained from scratch on all 2,152 training intervals for 32 epochs; its training inputs use no synthetic shifts. The four candidate configurations were subsequently compared on development-generated starts.

Go-Go uses 1,207 training charts/326 audio groups and 134 calibration charts/37 separate train-derived groups. It minimizes BCE against soft bar occupancy (seed 20260912). Development charts do not select its epoch or threshold.

Events and timing headers remain in the osu-chart frame, origin zero, rate one. Endpoint pooling converts chart query positions to audio-aligned coordinates through the shared legacy +20 ms policy; features and metrics receive no generic offset.

3. DEVELOPMENT EVIDENCE

The fixed-50 sample uses seeded-hash ranks within development event-count quartiles (quotas 13/12/13/12; 38 source groups), independently of outputs. With a common ± 20 ms pooled micro evaluation, the selected Joint and learned endpoint pipeline obtains four-class ordinary-note F1 0.493896, large-note F1 0.375602, and long-expanded onset F1 0.746166; all 50 charts pass native parsing. The

Model	Prec.	Recall	Bar F1	Time IoU
MERT v2 MLP128	75.6%	81.0%	78.2%	63.8%
B2 epoch 16	58.4%	60.4%	59.4%	41.9%
All bars active	27.6%	100.0%	43.3%	27.6%

Table 2. Common-population Go-Go development comparison.

control scores are 0.486922, 0.353134, and 0.734296. Both Joint models use MERT v2 and fixed predicted E2 plans, so this is final-configuration selection rather than an isolated encoder or plan-input ablation.

For the 11 generated long starts matched within 20 ms under both selected-Joint endpoint conditions, learned endpoints reduce endpoint MAE from 569.545 to 452.909 ms. Requiring a matched start within 20 ms and its paired end within 250 ms raises F1 from 0.059172 to 0.094675. The small matched set limits this evidence, and it does not establish that the representation itself caused the timing or classification changes.

In a separate matched size-stage comparison, repeated MERT-v2 refinement improved large-note joint F1 from 0.2985 to 0.3407 and four-class joint F1 from 0.4584 to 0.4629 over the heterogeneous three-refiner sequence. Starting charts and repair policies were identical (± 10 ms pooled micro scores); both produced 50/50 structurally valid charts. This supports applying one refiner three times, not an encoder ablation or whole-system performance claim.

Table 2 compares saved Go-Go predictions on the same 248 development charts and 27,305 bars. A reference bar is positive at 50% active occupancy. Time IoU covers valid-feature bars totalling 37,015 seconds. B2 is a separate 7.82M audio-conditioned next-event model; its epoch 16 was selected on a fixed-50 development subset, whereas the MLP uses train-derived calibration. Saved predictions are integrated over identical bars.

The Go-Go MLP outperforms B2 and all-active on this development population. Source-group macro F1 is 67.2/52.5/40.4%, respectively; this is configuration selection, not an architecture ablation.

Separate all-248 plan-input diagnostics changed joint F1 by +0.0010 and +0.0030 from common checkpoints, while lowering long-start F1; they do not establish an isolated plan-conditioning gain.

4. REPRODUCTION AND LIMITATIONS

Source, checkpoints and this report are supplied directly to the organizers with an immutable Git commit. The accompanying instructions and manifest specify execution, setup and complete hashes. Access remains restricted as required. The runtime requires Windows, CUDA/bf16 and a SHA-verified artifact map relative to the extracted package; a fresh recipient GPU run has not been executed.

Generation fills all five supplied templates, preserving audio, tempo and beat positions. Go-Go updates timing flags and adds inherited points. Structural validity means passing section, timing/order, object-type and decoded-

endpoint checks, separately from MV findings and human playability. Remaining MV warnings are not claimed to be eliminated.

Repeated development-set selection precludes independent-test claims. No hidden-test, expert, community or significance results are available; note scores do not measure enjoyment or creativity. Endpoint accuracy remains low and depends on provisional rules. Bar-resolution Go-Go can miss short regions or entire songs. The scroll-velocity and hitsound-volume controls lack player evaluation. Difficulty ordering and entire-song playability need separate evaluation.

5. ACKNOWLEDGMENT

This work was supported by the Joint Research Promotion Program (Project C) of the Graduate School of Environment and Information Sciences, Yokohama National University.

6. AI USAGE STATEMENT

AI tools assisted drafting, implementation and evidence checking.

7. REFERENCES

- [1] MIREX, “2026: Rhythm Game Chart Generation,” 2026. https://music-ir.org/mirex/wiki/2026:Rhythm_Game_Chart_Generation, accessed 12 September 2026.
- [2] Y. Li et al., “MERT: Acoustic music understanding model with large-scale self-supervised training,” in *ICLR*, 2024. <https://arxiv.org/abs/2306.00107>.
- [3] B. Warner et al., “Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference,” 2024. <https://arxiv.org/abs/2412.13663>.
- [4] M. O’Malley, “ITGPT: A Transformer Based Architecture for the Generation of Dance Dance Revolution and In the Groove Charts,” arXiv:2607.14148, 2026. <https://arxiv.org/abs/2607.14148>.
- [5] R. Yuan et al., “MARBLE: Music Audio Representation Benchmark for Universal Evaluation,” in *Advances in Neural Information Processing Systems*, vol. 36, pp. 39626–39647, 2023. <https://arxiv.org/abs/2306.10548>.
- [6] Multimodal Art Projection, “MERT-v2-FullSong,” model release, 2026. <https://huggingface.co/m-a-p/MERT-v2-FullSong>, accessed 12 September 2026.
- [7] Mapset Verifier contributors, “Mapset Verifier,” version 2.0.0, 2026. <https://github.com/Naxesss/MapsetVerifier/releases/tag/v2.0.0>.