Extended Results

Beyond File-Level Splits: Rethinking Symbolic Music Evaluation via Multi-View Alignment

Companion to the manuscript · 21 November 2026

Key results appear in manuscript Tables V–X. This companion provides expanded experimental results in Tables S1–S9; experimental protocols are in the manuscript appendix.

Calibration and alignment

Table S1 adds operating thresholds and fixed-threshold F1 intervals to manuscript Table V. Calibration is independent and frozen before evaluation. Intervals use 2,000 work-group bootstrap replicates (seed 20260921), pooling query-level TP/FP/FN and taking the 2.5th/97.5th percentiles. Recall includes unretrieved positives.

Table S1: Independent dataset calibration and MVA-SWI F1 intervals (%). MVA/Union are calibrated F1; Union is CAug-BERT ∪ CLaMP, with thresholds listed in that order. Fixed/calibrated intervals use transferred/independent thresholds.

Dataset	MVA threshold	MVA F1	Union thresholds	Union F1	Fixed 95% CI	Calibrated 95% CI
SHS	0.5933	72.99	0.91/0.78	39.68	[47.68, 85.71]	[51.28, 85.71]
ASAP	0.5899	99.73	0.95/0.84	60.40	[99.40, 99.94]	[99.40, 99.93]
ATEPP	0.5864	98.24	0.98/0.88	68.20	[96.79, 98.78]	[97.15, 99.09]
LMD-clean	0.5686	82.35	0.87/0.85	51.12	[75.59, 83.55]	[78.77, 85.81]
Vienna 4×22	0.6033	100.00	0.89/0.81	90.92	[99.35, 100.00]	[100.00, 100.00]
PianoVAM	0.4924	89.17	0.94/0.88	57.33	[64.41, 93.55]	[71.50, 95.09]

Table S2 holds candidates and all 57 features fixed. Prefix lengths are in segments of two MIDI quarter-note units. The deployed setting is zero shift with a 512-segment cap. P, R, and F1 denote pairwise precision, recall, and their harmonic mean; TP/FP/FN are pair counts.

Table S2: Pitch-shift and prefix-length sensitivity. Changes in pairwise precision, recall, and F1 are in percentage points relative to zero shift at 512 segments in the same cohort.

Dataset	Shift	Prefix	ΔP	ΔR	ΔF1	TP	FP	FN
LMD-clean	Zero	512	0.00	0.00	0.00	1385	55	652
	Top-3	512	−2.17	+3.63	+1.64	1459	93	578
	Zero	768	−0.07	0.00	−0.02	1385	56	652
	Top-3	768	−2.23	+3.63	+1.61	1459	94	578
ATEPP	Zero	512	0.00	0.00	0.00	3263	64	76
	Top-3	512	−0.32	−0.03	−0.18	3262	75	77
	Zero	768	0.00	0.00	0.00	3263	64	76
	Top-3	768	−0.35	−0.03	−0.19	3262	76	77
ASAP	Zero	512	0.00	0.00	0.00	4969	8	19
	Top-3	512	−0.28	−0.02	−0.15	4968	22	20
	Zero	768	0.00	0.00	0.00	4969	8	19
	Top-3	768	−0.28	0.00	−0.14	4969	22	19

On LMD-clean, Top-3 at 512 recovers 74 additional positives and introduces 38 additional false positives. Extending the prefix alone recovers no additional positives. These aggregate tests do not separately evaluate late-only shared passages or isolate ordered versus unordered matching with identical features.

Timing and training-duration sensitivity

For ASAP, timing is the median annotated interbeat interval in seconds; a pair's ratio is the larger median divided by the smaller. Beat annotations are used only for this diagnostic, while the detector operates in MIDI quarter-note units. ASAP-equal retains pairs with equal annotated beat counts. ATEPP uses note-span duration as a proxy, which may also reflect content differences. The deployed detector threshold is fixed; recall includes retrieval misses. A work can contribute pairs to both strata. These strata do not directly measure local rubato.

Table S3: End-to-end recall stratified by the ratio of larger to smaller timing measurements. ASAP-equal restricts to equal annotated beat counts. ATEPP uses note-span duration. Retrieved/accepted counts refer to reference-positive pairs.

Cohort	Ratio	Groups	Pairs	Retrieved/ accepted	R (%)
ASAP	≤ 1.20	164	4809	4809/4793	99.67
	> 1.20	41	179	179/176	98.32
ASAP-equal	≤ 1.20	159	4454	4454/4440	99.69
	> 1.20	40	169	169/166	98.22
ATEPP	≤ 1.20	300	2593	2592/2558	98.65
	> 1.20	185	746	740/705	94.50

The controlled exposure study uses 100 manually confirmed LMD test–donor pairs selected from highly similar candidates. Unrelated and Same replace the same 1,264 of 38,374 training windows with unrelated or same-work material, matching tokens. Each duration uses three paired seeds and the same groups. Reduction first averages paired-seed losses within each group, computes its relative Unrelated-to-Same reduction, then averages equally over groups. Thus the reduction need not equal the ratio of the displayed aggregate NLLs. The 95% confidence intervals use 10,000 paired work-group bootstrap replicates with trained seeds fixed and exclude cohort-selection and new-training uncertainty.

Table S4: MIDI-GPT exposure effects at different training durations. Reduction averages within-group relative NLL changes; 95% CIs resample groups with trained seeds fixed.

Updates	Unrelated NLL	Same NLL	Reduction [95% CI] (%)
5,000	0.7664	0.7346	4.39 [3.75, 5.04]
10,000	0.7010	0.6080	14.44 [12.87, 16.10]
20,000	0.7318	0.5256	30.16 [27.35, 33.02]

The 5,000- and 10,000-update models were separately retrained with the original 20,000-update learning-rate schedule; they are not intermediate checkpoints from one trajectory. These runs used RTX 3090/4090 servers, with the third seed's two conditions on different machines. The original 20,000-update results used RTX 5090. The reductions grow across these runs, but the hardware differences preclude assigning the entire change to update count alone.

Independent audit and resource use

The Fast auditor uses 49 features and is distinct from the 57-feature v1.9 core. Candidate and pair recall are the fractions of all eligible positives retrieved and accepted. Query coverage is the fraction of queries with a reference positive that yield an accepted pair. The seven-corpus replay uses a separate checkpoint and grouping configuration; validation on the three labeled cohorts does not establish missed-version rates in unlabeled corpora.

Table S5: Frozen Fast auditor versus the v1.9 core on independently labeled cohorts. Values are percentages.

Cohort	Config.	Candidate recall	Pair precision	Pair recall	F1	Query coverage
ATEPP	Fast	98.53	98.37	97.36	97.86	99.80
	v1.9	99.79	98.08	97.72	97.90	99.67
LMD-clean	Fast	66.67	91.70	62.89	74.61	76.09
	v1.9	88.61	96.18	67.99	79.67	79.51
ASAP	Fast	99.80	99.70	99.66	99.68	99.80
	v1.9	100.00	99.84	99.62	99.73	99.80

For each unlabeled corpus below, two reviewers independently assessed 100 random accepted edges, 20 targeted weak bridges, and five components selected for size, including the largest. Disagreements were reconciled. The targeted samples describe failure modes and do not estimate their population prevalence. Of 680 inspected component members, 38 differ from the modal work and six remain uncertain.

Table S6: Manual audit of four unlabeled corpora. Edge and bridge entries are same-work/different-work/uncertain counts. Components are selected for size; mixed components contain confirmed members from more than one work.

Corpus	Random edges	Targeted bridges	Members	Mixed / audited	Outside modal work	Uncertain members
PDMX	95/3/2	17/2/1	186	2/5	6	1
LMD-full	91/7/2	15/4/1	120	3/5	13	1
Aria-MIDI	93/4/3	16/3/1	194	2/5	11	3
GigaMIDI	94/5/1	16/4/0	180	2/5	8	1

The audit covers 3,685,318 input files across seven frozen manifests. Failed parses remain unresolved singleton groups. End-to-end time includes parsing through grouping on a shared 128-vCPU, 256-GB server; jobs can overlap, so their durations are not a single sequential wall-clock total. Peak and per-corpus memory measurements are unavailable.

Table S7: Input files, successful parses, and elapsed audit time per corpus job. Jobs can overlap.

Corpus	Files	Parsed	Minutes
POP909	2,898	2,897	0.70
ASAP	1,067	1,067	0.63
PDMX	222,820	222,769	132.21
GigaMIDI	2,136,218	2,134,556	530.11
MAESTRO	1,276	1,276	0.88
LMD-full	178,561	175,923	128.10
Aria-MIDI	1,142,478	1,142,478	454.11

The 4,352 parse failures and other missed relations can lower inferred exposure, while false bridges can raise it. Random-split summaries use 200 80/10/10 partitions and report 2.5th/97.5th percentiles across partitions, conditional on inferred groups. Official MAESTRO/GigaMIDI splits use their manifest metadata.

Budget-matched losses and reference construction

Manuscript Table VIII reports the original five downstream training pools, residual exposure, and clean-token retention. Calibrated union appears in Fig. 6 and is excluded from those five pools. The separate budget-matched study below supplies absolute NLLs and training-file counts in addition to the key deviations reported in manuscript Table IX.

The separate ATEPP study in Table S8 matches the six conditions' model, initial weights across three seeds, 20,000 updates, 11,316 windows, and 5,884,369 scored training tokens. Its fixed test set has 121 works, 152 files, 1,192 windows, and 602,471 scored tokens. Exact deduplication uses SHA-256 and normalized MIDI-event fingerprints. Random removal stratifies by usable windows and note-token length. The original reference minimally repairs the MVA-SWI pool with work labels at the same file budget. The independent reference excludes validation/test works from the original pool by ATEPP labels and selects windows in deterministic hash order under fixed window and prompt-length quotas.

Table S8: Budget-matched ATEPP results. Exposure is measured using ATEPP work labels. NLL is averaged over works. Dev. is signed relative NLL compared with the independent reference. Intervals are paired across works and conditional on the trained seeds. The independent reference has zero deviation by definition.

Condition	Train files	Exposed works	NLL	Dev. (%)	Paired 95% interval (%)
Uncorrected	1,140	119/121	3.6009	−21.77	[−22.93, −20.59]
Exact deduplication	1,140	119/121	3.6009	−21.77	[−22.93, −20.59]
Random removal	484	72/121	4.1693	−10.11	[−11.31, −8.93]
MVA-SWI	484	0/121	4.7421	+2.15	[+1.85, +2.48]
Original ref.	484	0/121	4.6860	+0.89	[+0.63, +1.19]
Independent ref.	492	0/121	4.6482	0.00	[0.00, 0.00]

NLL is averaged over valid tokens within files, files within works, and seeds. Deviation is the mean signed per-work relative difference from the independent reference. Confidence intervals pair works and condition on trained seeds. The independent reference has zero deviation by definition.

Manuscript Fig. 6 uses a different ATEPP Uncorrected pool (1,223 files, 121/121 exposed test works) from Fig. 7 and Table S8 (1,140 files, 119/121). The latter also matches windows, scored tokens, and updates and uses an independently selected reference. The 4.41% absolute and −21.77% signed deviations concern different trained models and reference losses, not just different sign conventions. Fig. 6 gives paired-seed SD; Fig. 7/Table S8 give work-bootstrap intervals conditional on trained seeds.

Automatic continuation diagnostics

The evaluation fixes 120 ATEPP works and 119 LMD-clean reference groups across five conditions and two models. Ten training seeds and ten generations per work/group/checkpoint give 239,000 excerpts. Generation replicates are pooled within each seed; Table S9 reports the mean and SD across ten seeds. The calibrated union added to manuscript Fig. 6 is excluded. Decoding generates 512 tokens from a 128-token prompt with temperature 1.0 and top-$p = 0.95$.

Table S9: Automatic continuation diagnostics (mean ± SD across ten training seeds). Melody, rhythm, and long-passage matches quantify agreement with held-out continuations. FMD measures embedding-distribution distance to the shared reference continuations.

	MIDI-GPT				Transformer			
Training condition	Melody sim. (%)	Rhythm sim. (%)	Passage match (%)	FMD ↓	Melody sim. (%)	Rhythm sim. (%)	Passage match (%)	FMD ↓
ATEPP								
Uncorrected	36.09 ± 0.78	59.14 ± 0.86	22.42 ± 0.92	7.49 ± 0.07	39.21 ± 0.21	61.26 ± 1.35	23.89 ± 0.71	6.78 ± 0.03
Matched random removal	29.03 ± 0.96	49.76 ± 0.91	10.92 ± 0.94	8.68 ± 0.08	32.27 ± 0.31	52.12 ± 1.16	11.69 ± 0.42	7.87 ± 0.01
MVA-SWI isolation	22.49 ± 0.65	40.88 ± 1.03	3.64 ± 0.27	10.14 ± 0.07	25.53 ± 0.64	43.05 ± 1.24	4.06 ± 0.46	9.13 ± 0.07
CAugBERT ∪ CLaMP	19.02 ± 1.00	36.60 ± 0.42	2.28 ± 0.10	11.04 ± 0.06	22.21 ± 0.45	39.07 ± 0.86	3.03 ± 0.29	9.96 ± 0.07
Label-isolated reference	22.05 ± 1.04	40.80 ± 0.95	3.33 ± 0.17	10.02 ± 0.08	25.29 ± 0.48	43.35 ± 1.24	3.94 ± 0.43	9.06 ± 0.07
LMD-clean								
Uncorrected	33.80 ± 1.81	55.66 ± 0.39	21.90 ± 1.86	9.00 ± 0.08	37.58 ± 1.11	60.40 ± 1.31	22.30 ± 1.40	8.31 ± 0.02
Matched random removal	26.75 ± 1.71	46.87 ± 0.03	10.48 ± 1.37	10.45 ± 0.07	30.70 ± 0.95	51.61 ± 1.51	11.01 ± 0.22	9.65 ± 0.01
MVA-SWI isolation	19.71 ± 1.79	37.33 ± 0.10	2.46 ± 0.56	12.15 ± 0.11	23.87 ± 1.00	42.58 ± 1.16	3.33 ± 0.78	11.24 ± 0.01
CAugBERT ∪ CLaMP	16.53 ± 2.10	33.53 ± 0.14	1.29 ± 0.26	13.18 ± 0.12	20.80 ± 0.82	38.36 ± 1.00	2.13 ± 0.70	12.21 ± 0.02
Label-isolated reference	19.69 ± 1.65	37.67 ± 0.29	2.41 ± 0.49	11.97 ± 0.12	23.64 ± 0.95	42.83 ± 1.30	3.22 ± 0.90	11.11 ± 0.02

Melody similarity uses the highest active voice as a proxy. Melody and rhythm similarities compare with the true held-out continuations. A long-passage match requires at least 16 notes over four beats with at most two pitch and two inter-onset-interval discrepancies, under a rule fixed before evaluation. These measures quantify continuation agreement and possible exposure, rather than general musical quality.

FMD-CLaMP1 is Fréchet Music Distance on FP32 768-dimensional projected CLaMP embeddings, using fixed all-note ABC conversion and a common natural-continuation reference. For empirical means μ and covariances Σ of generated (g) and reference (r) embeddings,

$$D = \|\mu_g - \mu_r\|_2^2 + \operatorname{tr}\left(\Sigma_g + \Sigma_r - 2\left(\Sigma_g^{1/2}\Sigma_r\Sigma_g^{1/2}\right)^{1/2}\right).$$

Automatic similarity and distance differences do not establish a perceptual quality difference; the manuscript reports listening results separately.