THE EVALUATOR IS A RANDOM VARIABLE: AUDITING THE FIXED FEATURE SPACE BEHIND TEXT-TO-MOTION BENCHMARKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Text-to-motion generation is benchmarked almost entirely through one artifact: a text-motion contrastive feature extractor released in 2022 whose embedding space defines FID, R-precision, matching score, Diversity, and MultiModality on HumanML3D and KIT-ML. Every confidence interval in the literature is computed over generation repeats with this evaluator held fixed, so it omits the evaluator's own training randomness. We treat the evaluator as a random variable. Retraining it from its public recipe with 30 independent seeds on KIT-ML, we score frozen pools of generations from four released state-of-the-art models under every seed and decompose metric variance into evaluator-seed, generation-seed, and batching components. The evaluator seed is the largest component for FID, R-precision, and matching score: for MoMask, the FID standard deviation across seeds is 0.05, seven times the batching noise that published intervals report and larger than the generation-seed component. Among the orderings of our pools under the released checkpoint, every pair whose seed-mean gap is within two seed standard deviations reverses under some seeds, and the gap under the released checkpoint does not predict which; a non-generative baseline that returns the nearest training motion for each test caption overtakes the strongest generator under half of the seeds, an advantage that disappears when duplicate retrieved motions are removed. Inference hyperparameters selected on one evaluator transfer to other seeds with regret no larger than the seed noise, while the identity of the optimum does not. A corruption suite shows that FID ignores a foot drift larger than the standard deviation of its own velocity features, and identity controls show that the feature-extraction route alone can move FID by more than a published state-of-the-art margin. We release the seeded evaluators, the frozen pools, and an evaluator-adjusted confidence interval so that future improvements can be reported against the noise floor of the measurement itself.

1 INTRODUCTION

Text-to-motion generation is compared on two datasets, HumanML3D (Guo et al., 2022) and KIT-ML (Plappert et al., 2016), with five numbers: FID, R-precision, matching score, Diversity, and MultiModality. All five are computed in the embedding space of a single learned model, the text-motion matching network released with the HumanML3D benchmark (Guo et al., 2022). The checkpoint, `text_mot_match`, has never been retrained by the community; the published models that followed, including T2M-GPT (Zhang et al., 2023), MDM (Tevet et al., 2023), MLD (Chen et al., 2023), MoMask (Guo et al., 2024), BAMM (Pinyoanuntapong et al., 2024), and MARDM (Meng et al., 2025), load the same file. Every reported 95% confidence interval is computed over 20 generation repeats with that file held fixed.

This paper asks a question that the protocol makes invisible by construction: how much of a reported difference between two generators is a property of the generators, and how much is a property of the particular random draw that produced the evaluator? The evaluator is a small network trained for a few epochs with a hinge contrastive loss from a random initialization and a random data order. Its feature space, and therefore every metric computed in it, is a random variable, and the intervals in the literature condition on one realization of it.

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.

We treat the evaluator as a random factor and measure it. We retrain the evaluator from its public recipe with 30 independent seeds on KIT-ML, freeze pools of generations from released checkpoints, and rescore the same pools under every seed. This yields four results.

1. **A noise floor.** Across evaluator seeds, the FID of a fixed MoMask pool has a standard deviation of 0.050 (95% interval [0.039, 0.067]), 7 times the batching standard deviation that current confidence intervals report and larger than the generation-seed component (Section 4.2). The evaluator seed is also the largest component for R-precision and matching score.
2. **Rank stability with a finite-sample bound.** We report, for every pair of pools, the standardized gap and the number of evaluator seeds that reverse the ordering given by the released checkpoint, with exact binomial intervals (Section 4.3); with N seeds and no reversals, the reversal probability is bounded by $3/N$, so a finite audit can demonstrate instability but only bound stability.
3. **Selection transfer.** Generators select checkpoints and inference hyperparameters on the fixed evaluator. We sweep MoMask's guidance scale and iteration count, select the best configuration under one evaluator, and measure the FID regret under the others (Section 4.4). The mean regret is no larger than the evaluator-seed standard deviation, the identity of the optimum is not shared, and the configuration the authors ship is not the FID optimum under any evaluator.
4. **Three validity controls.** A retrieval baseline that returns the nearest training motion for each test caption, a corruption suite that perturbs real motions in joint space and recomputes the representation, and identity controls that measure how much FID different feature-extraction routes assign to identical motions (Section 4.5).

A fifth observation was not part of the design. Our frozen pools, generated once from the released checkpoints with a single generation seed, do not reproduce the published FIDs of two of the four generators (Section 4.3). The gaps are of the same order as the evaluator-seed effect, which means generation-protocol variance is a second component that published intervals leave out. We release the retrained evaluators, the frozen pools, and the scoring code so that future comparisons can be reported against the noise floor of the measurement rather than against a single draw of it.

2 THE BENCHMARK PROTOCOL AND WHAT IT HOLDS FIXED

Representation. Motions are sequences of 251-dimensional (KIT-ML, 21 joints) or 263-dimensional (HumanML3D, 22 joints) feature vectors: root angular velocity, root planar velocity, root height, joint positions in the root frame, 6D joint rotations, joint velocities, and four binary foot-contact labels (Guo et al., 2022). Features are z-normalized with dataset statistics.

Evaluator. The evaluator has three parts: a convolutional movement encoder that maps 4-frame snippets of the first 247 (259) feature dimensions to 512-dimensional codes, a bidirectional GRU motion encoder over those codes, and a bidirectional GRU text encoder over GloVe word vectors (Pennington et al., 2014) concatenated with part-of-speech one-hot codes. The four foot-contact channels are dropped before the movement encoder, so the evaluator never sees them; foot behavior can reach it only through joint positions and velocities. The movement encoder is taken frozen from a separately trained motion autoencoder. The text and motion encoders are trained with the contrastive loss of Hadsell et al. (2006) with margin 10, positive pairs being caption-motion pairs and negatives being the same batch shifted by a random offset, with model selection by the minimum validation loss (optimizer settings in the appendix). The released KIT-ML checkpoint was selected at epoch 30 and the HumanML3D checkpoint at epoch 28.

Metrics. The test set is shuffled and cut into batches of 32. In each batch, R-precision top-k is the fraction of captions whose own motion embedding is among the k nearest of the 32 motion embeddings, and the matching score is the mean Euclidean distance between paired caption and motion embeddings. FID is the Fréchet distance (Heusel et al., 2017) between Gaussians fitted to the motion embeddings of real and generated motions over all batches. Diversity is the mean distance between 300 random pairs of generated embeddings. Published numbers are means over 20 repeats of generation and batching, with a $\pm$ interval of 1.96 standard errors.

What the interval conditions on. The repeats resample generation noise and batch composition. They do not resample the evaluator. Any variance introduced by the evaluator's initialization, data order, or early-stopping epoch is absent from every interval in the literature, and the differences between published methods (on KIT-ML, MoMask 0.204 versus BAMM 0.183 versus T2M-GPT 0.514 FID, from each paper's main table) are read against intervals a few hundredths wide.

3 TREATING THE EVALUATOR AS A RANDOM VARIABLE

Retraining. We reimplement the training loop of the released code (loss, optimizer, schedule, selection rule) and train $N = 30$ evaluators that differ only in the random seed controlling initialization, data order, and the caption and crop drawn for each training motion. The movement encoder is held at its released weights for all seeds, so "evaluator seed" throughout this paper means the seed of the text and motion encoders with the movement encoder frozen, and the variance we report is a lower bound on the variance of the full recipe (Section 6). Each seed trains for 40 epochs on one consumer GPU in about 5 minutes; 1 of the 30 seeds reached its best validation loss at the final epoch.

Frozen pools. A pool is one generated motion per test motion, produced once from a released checkpoint with a fixed generation seed and stored in the dataset's native representation. The caption, the crop of the real motion, and the length are frozen per row, so every evaluator scores exactly the same motions. We build pools for MoMask, BAMM, T2M-GPT, and MDM from their released KIT-ML checkpoints and their repositories' own sampling code; MLD and MARDM have no released KIT-ML checkpoint. Two non-generative pools complete the set: the real test motions themselves (the "Real" row of every results table) and a retrieval pool described below.

Scoring and variance decomposition. Each pool is scored under each evaluator with the standard protocol, repeated five times over batch composition. For MoMask we also draw five generation seeds. This gives a three-way layout (evaluator seed $\times$ generation seed $\times$ batching repeat) from which we estimate the standard deviation attributable to each factor as the square root of the between-level variance of level means, with the remaining factors averaged out, and report chi-square intervals on the evaluator-seed and generation-seed standard deviations.

Rank statistics. For each pair of pools (a, b) and each seed s, let $\Delta_s = m_s(a) - m_s(b)$ be the difference in a metric m. We report $z = \bar{\Delta}/\mathrm{sd}(\Delta_s)$, the number of seeds whose sign of Δ_s differs from the sign under the released checkpoint with its Clopper-Pearson interval, and Kendall's τ between each seed's ranking of all pools and the released ranking. When no reversal is observed among N seeds, the rule of three (Hanley & Lippman-Hand, 1983) gives a 95% upper bound of $3/N$ on the reversal probability; with $N = 30$ this bound is 0.10.

Selection transfer. Released generators choose checkpoints and inference hyperparameters by their score under the fixed evaluator; MoMask, for instance, ships a residual model and a quantizer named `net_best_fid`. To measure what such selection is worth under a different draw of the evaluator, we sweep MoMask's classifier-free guidance scale over $\{1, 2, 3, 4, 6, 8\}$ and its number of masked-generation iterations over $\{5, 10, 18, 49\}$, producing 24 pools. For a selecting evaluator e, let c_e^* be the configuration with the lowest FID under e. The regret of that choice under a target evaluator t is $\mathrm{FID}_t(c_e^*) - \min_c \mathrm{FID}_t(c)$. We compare the regret of selecting on the released checkpoint, on a retrained seed, of using the published default, and of choosing a configuration at random.

Retrieval baseline. For each test caption we retrieve the training caption nearest in the released text encoder's embedding space and return that training motion verbatim, at its own length. This is not a generator; it is a probe of how much of the score is available without generalization. Its R-precision under the released evaluator is circular, since the retrieval used the same text space that scores it; under a retrained seed the text space differs. FID depends only on motion embeddings, but the retrieved motions were still chosen by the released text encoder and are training-distribution samples, so we also report a duplicate-free variant matched in size to a subsampled generator pool.

Corruption suite with identity controls. We decode each real test motion to joint positions with the released inverse mapping, perturb the joints, and re-encode the perturbed joints with the dataset's

Real test motions	released	seed min	seed median	seed max
R-precision top-1	0.427	0.364	0.422	0.457
top-2	0.661	0.593	0.639	0.684
top-3	0.784	0.727	0.768	0.805
matching score	2.785	2.594	2.790	3.037

Table 1: Discriminativeness of the released evaluator and of the 30 retrained seeds, measured on real test motions (the quantity the evaluator is trained on). Higher R-precision and lower matching score are better.

own feature pipeline. Corruptions are Gaussian jitter on all joints (0.25% to 4% of body height), a slow horizontal drift of the four foot joints during detected ground contact (amplitude 0.5% to 8% of body height, two sinusoidal periods over each clip's contact frames), temporal resampling by factors 0.5 to 2, truncation to 35% to 75% of the frames, and a caption permutation that swaps the motions of a fraction of the rows among themselves (a caption-motion mismatch with physically valid motion and an unchanged marginal distribution). A control for the foot drift applies the same Gaussian jitter to the four foot joints only, at per-joint magnitudes from 0.25% to 8% of body height. The identity control re-encodes the unperturbed joints through the same pipeline, which re-canonicalizes floor height and initial heading and therefore moves absolute joint positions by 7% of body height on average while leaving the motion itself unchanged. Two further identity controls route the real motions through other feature extractors shipped with published code: the alternative extraction routine in the same file as the pipeline, which omits its skeleton, floor, and heading canonicalization, and the 64-dimensional representation used by MARDM (Meng et al., 2025) with its prescribed conversion back to the 251-dimensional layout.

4 Results

All experiments use KIT-ML (734 usable test motions after the standard length filter, of which 704 are scored per batching repeat), the released KIT-ML evaluator, and 30 retrained evaluators. Every number below is read from the analysis file written by the scoring code and included in the supplement.

4.1 The retrained evaluators are as discriminative as the released one

Table 1 reports the R-precision top-1 of real test motions under each evaluator, the quantity the evaluator is trained to maximize. The retrained seeds span 0.364 to 0.457 against 0.427 for the released checkpoint. The released checkpoint is an ordinary draw.

4.2 Evaluator seed is the largest variance component of FID, R-precision, and matching score

Figure 1 and Table 2 give, for each pool, the released score and the distribution over 30 seeds; for the MoMask pool, FID is 0.185 under the released checkpoint and ranges from 0.207 to 0.462 across seeds.

Absolute FIDs are not comparable across feature spaces, so Table 3 decomposes the variance of each metric for MoMask over evaluator seed, generation seed, and batching. The evaluator seed is the largest component for FID (0.050, interval [0.039, 0.067], against 0.032, interval [0.019, 0.093], for generation seed and 0.007 for batching), for R-precision top-1 (0.026, 0.007, 0.013), and for the matching score (0.114, 0.025, 0.012); for Diversity it exceeds batching by a smaller factor, and we do not audit MultiModality. Generation noise, the only source the protocol resamples, is the smaller of the two model-dependent sources for FID, although the interval on the generation-seed component, estimated from five seeds, is wide. The comparison that matters for a ranking is the difference between two pools: across the pairs in Table 4, the seed standard deviation of a FID difference is 0.060 to 0.185, which is 6 to 19 times the batching standard deviation of that difference.

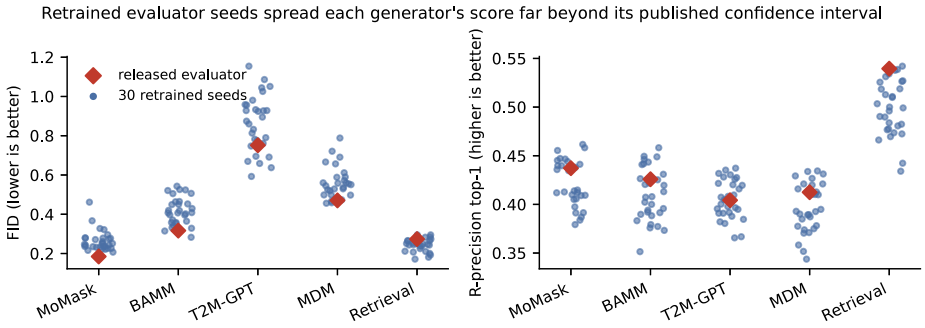

Figure 1: Score of each frozen pool under the released evaluator (diamond) and under 30 retrained evaluators (dots). Left: FID. Right: R-precision top-1. The retrieval pool's R-precision top-1 is circular under the released checkpoint (Section 4.5).

	FID				R-precision top-1		
Pool	published	released	seeds (mean ± sd)	seeds [min, max]	published	released	seeds (mean ± sd)
MoMask	0.204	0.185	0.263 ± 0.052	[0.207, 0.462]	0.433	0.437	0.422 ± 0.024
BAMM	0.183	0.317	0.417 ± 0.069	[0.283, 0.543]	0.438	0.426	0.410 ± 0.028
T2M-GPT	0.514	0.751	0.858 ± 0.144	[0.593, 1.154]	0.416	0.404	0.405 ± 0.020
MDM	0.497	0.471	0.554 ± 0.083	[0.456, 0.788]	0.396	0.412	0.397 ± 0.025
Retrieval	n/a	0.274	0.246 ± 0.033	[0.171, 0.295]	n/a	0.539	0.499 ± 0.028

Table 2: Frozen KIT-ML pools scored under the released evaluator and under 30 retrained seeds. "Published" is the value in the main results table of each method's paper, obtained under the released evaluator with the authors' own generation protocol; "released" is our frozen pool under the same checkpoint. The seed columns describe one generation seed per pool across evaluator seeds; Table 3 separates the evaluator-seed, generation-seed, and batching components for MoMask. The retrieval pool's R-precision is circular (Section 4.5).

4.3 Which orderings survive a change of seed, and which orderings we are testing

Under the released checkpoint our frozen pools (Table 2) reproduce the published FID of MoMask and MDM to within a few hundredths but score BAMM and T2M-GPT well above their published values. Each pool is a single generation seed drawn with the released checkpoint's own sampling code, without the 20-repeat protocol or the checkpoint and sampling choices behind a headline number; the supplement records every choice. We do not treat the published numbers as wrong but treat the gap as a second omitted variance component: the generation protocol, like the evaluator, is fixed in published intervals, and its effect here is of the same order as the evaluator seed. Consequently the orderings tested below are the orderings of our pools under the released checkpoint, which put the retrieval pool second, and not the published ordering, in which BAMM leads MoMask.

Table 4 reports, for each pair of pools, the FID gap under the released checkpoint, the mean and standard deviation of the gap across seeds, the standardized gap z, and the number of reversing seeds with its interval. Kendall's τ between each seed's ranking and the released ranking averages 0.88 for FID and 0.70 for R-precision top-1. Of the 4 adjacent pairs in the released-checkpoint ordering of our pools, 3 reverse under at least one seed: MoMask versus Retrieval in 15 of 30 seeds (95% interval [0.313, 0.687]); BAMM versus MDM in 2 of 30 seeds (95% interval [0.008, 0.221]); T2M-GPT versus MDM in 1 of 30 seeds (95% interval [0.001, 0.172]). The remaining pair reverses in none, which bounds its reversal probability below 0.10 at 95% confidence. Among the published generators alone (MoMask $<$ BAMM $<$ MDM $<$ T2M-GPT), 2 of 3 adjacent pairs reverse, on 2 and 1 seeds of 30; MoMask versus BAMM never does. The predictor of a reversal is the seed-mean gap relative to its seed standard deviation, not the gap under the released checkpoint: 3 of the 3 pairs with $|z| < 2$ reverse at least once and 0 of the 7 pairs with $|z| \geq 2$ do, while the pair with the smallest released gap, BAMM versus Retrieval at 0.043, reverses in 0 seeds because its seed-mean

Metric (MoMask)	grand mean	sd: evaluator seed [95% interval]	sd: generation seed [95% interval]	sd: batching
FID	0.260	0.0496 [0.0395, 0.0667]	0.0325 [0.0194, 0.0933]	0.0070
R-precision top-1	0.423	0.0256 [0.0204, 0.0344]	0.0066 [0.0040, 0.0191]	0.0131
R-precision top-3	0.770	0.0199 [0.0158, 0.0267]	0.0054 [0.0032, 0.0155]	0.0096
matching score	2.882	0.1139 [0.0907, 0.1531]	0.0252 [0.0151, 0.0724]	0.0116
Diversity	10.296	0.5318 [0.4235, 0.7149]	0.0230 [0.0138, 0.0661]	0.3418

Table 3: Three-way variance decomposition over 30 evaluator seeds, 5 generation seeds, and 5 batching repeats. Each column is the standard deviation attributable to one factor with the others averaged out; brackets are chi-square intervals. The batching column is pooled over all evaluator and generation seeds and its interval is narrower than the printed precision.

A	B	ΔFID released	ΔFID seeds	z	reversals (95% interval)	adjacent
BAMM	Retrieval	+0.043	$+0.171 \pm 0.082$	2.1	0/30 [0.00, 0.12]	yes
MoMask	Retrieval	-0.089	$+0.017 \pm 0.064$	0.3	15/30 [0.31, 0.69]	yes
BAMM	MDM	-0.154	-0.137 ± 0.105	-1.3	2/30 [0.01, 0.22]	yes
T2M-GPT	MDM	+0.281	$+0.304 \pm 0.152$	2.0	1/30 [0.00, 0.17]	yes
MoMask	BAMM	-0.132	-0.153 ± 0.060	-2.5	0/30 [0.00, 0.12]	
MDM	Retrieval	+0.197	$+0.308 \pm 0.081$	3.8	0/30 [0.00, 0.12]	
MoMask	MDM	-0.286	-0.291 ± 0.095	-3.1	0/30 [0.00, 0.12]	
BAMM	T2M-GPT	-0.435	-0.441 ± 0.185	-2.4	0/30 [0.00, 0.12]	
T2M-GPT	Retrieval	+0.478	$+0.612 \pm 0.134$	4.6	0/30 [0.00, 0.12]	
MoMask	T2M-GPT	-0.567	-0.594 ± 0.147	-4.1	0/30 [0.00, 0.12]	

Table 4: Pairwise FID differences (A minus B) under the released evaluator and across 30 seeds. z is the mean seed difference divided by its standard deviation; reversals counts the seeds whose sign disagrees with the released checkpoint, with a Clopper-Pearson interval; "adjacent" marks neighbors in the released ordering (MoMask $<$ Retrieval $<$ BAMM $<$ MDM $<$ T2M-GPT). Every pool is one generation seed; the MoMask pool is generation seed 0, and its released FID spans 0.165 to 0.225 over the five generation seeds of Table 3.

gap is 0.171 ($z = 2.1$). A gap of a few hundredths under the released checkpoint therefore neither implies nor rules out a reversal under other draws of the evaluator; only the spread across draws, which no published interval reports, decides it.

4.4 Selection on one evaluator transfers, with regret no larger than the seed noise

Figure 2 and Table 5 report the sweep. Under the released evaluator the best configuration is guidance 2, 49 iterations; the configuration the authors ship is guidance 2, 10 iterations, which scores 0.185 against 0.169 for the optimum under the same checkpoint, so the shipped default was not chosen by FID under the released evaluator either. Selecting on the released checkpoint and evaluating on the retrained seeds incurs a mean regret of 0.020 FID; selecting on one retrained seed and evaluating on another incurs 0.023; selecting on a retrained seed and evaluating on the released checkpoint incurs 0.038; the published default incurs 0.048; a random configuration incurs 0.324. Only 3 configurations are ever selected across the 30 seeds and the released checkpoint: guidance 2 with 49 iterations for 16 of 31 evaluators; guidance 1 with 49 iterations for 11 of 31 evaluators; guidance 1 with 18 iterations for 4 of 31 evaluators. All three use low guidance and many iterations, and the configuration selected by one evaluator is the optimum under another in 39% of evaluator pairs. Selection on any single evaluator transfers to the others with a mean regret under half the evaluator-seed standard deviation of FID (the worst case over held-out evaluators is 0.108, about 2 times that deviation), far below the regret of a random configuration; what does not transfer is the identity of the optimum: 15 of 31 evaluators select a configuration other than the modal one. The default's regret is 0.015 under the released checkpoint, 12% of that checkpoint's gap between MoMask and BAMM (0.132), and 0.048 averaged over retrained seeds, 31% of the seed-mean gap (0.153); each regret is compared within its own feature space, since retrained spaces inflate every

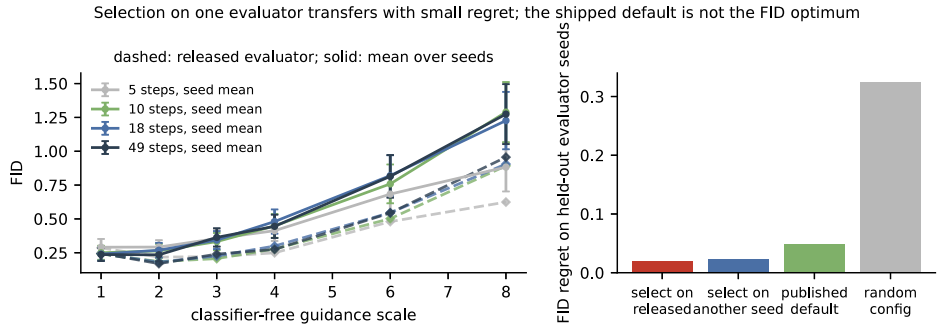

Figure 2: Left: MoMask FID as a function of guidance scale for each iteration count, under the released evaluator (dashed) and averaged over seeds with ± 1 standard deviation (solid). Right: FID regret of four selection rules on held-out evaluator seeds.

Selecting evaluator	selected configuration	mean regret	max regret	argmin agreement
released	guidance 2, 49 iterations	0.020	0.106	50%
seed0	guidance 1, 49 iterations	0.025	0.108	33%
seed1	guidance 1, 49 iterations	0.025	0.108	33%
seed2	guidance 2, 49 iterations	0.020	0.106	50%
seed3	guidance 2, 49 iterations	0.020	0.106	50%
seed4	guidance 2, 49 iterations	0.020	0.106	50%
published default	guidance 2, 10 iterations	0.048	n/a	n/a
random configuration	n/a	0.324	n/a	n/a

Table 5: Selection transfer on the MoMask sweep (6 guidance scales $\times$ 4 iteration counts, 24 configurations). Regret is the FID excess of the selected configuration over the best configuration under a held-out evaluator, averaged (mean) or maximized (max) over held-out evaluators; the first five seeds are shown, all 30 are in Table 7.

FID. A speed-versus-quality choice made by the authors is therefore worth between a tenth and a third of the margin between the two best methods on our pools.

4.5 Controls

Retrieval. The retrieval pool scores FID 0.274 under the released evaluator and 0.171 to 0.295 across seeds (Table 2). Its R-precision top-1 under the released evaluator is 0.539, above the real-motion value of 0.427, the circularity the design predicts. On FID, retrieval is better than MoMask under 50% of retrained seeds (and worse under the released checkpoint); averaged over seeds it scores 0.246 against 0.263 for MoMask. The 734 test captions retrieve only 438 distinct training motions, and FID is biased by sample size (Chong & Forsyth, 2020). A size-matched control that keeps the duplicates in proportion (random 438-row subsamples of the retrieval pool against random subsamples of the MoMask pool, ten draws each) gives 0.398 (standard deviation 0.088 over draws) against 0.283 (0.038) under the released checkpoint and 0.370 against 0.386 averaged over seeds, with retrieval ahead under 53% of seeds, so sample size does not explain the result. Removing the duplicates instead, so that each of the 438 retrieved motions appears once, raises the retrieval pool's FID to 0.665 (0.653 over seeds) against 0.357 (0.393 over seeds, standard deviation 0.119 over draws) for a size-matched MoMask subsample, and retrieval is then ahead under 0% of seeds. The advantage therefore needs the duplicates. It is consistent with a weighting effect, in which repeating a popular training motion once per test caption that names it matches the caption-frequency structure of the test set, which FID rewards; we do not measure how many repeated retrievals correspond to repeated test motions, so the control establishes the dependence and not the mechanism. The retrieved motions also keep their own length (mean 68 frames against 76 for the test crops), a distribution difference we do not correct. Since 34% of KIT-ML test captions appear verbatim in the training split, excluding each test motion's mirror twin from the index changes no retrieval.

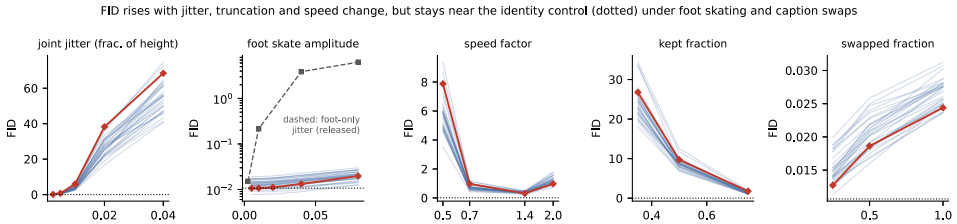

Figure 3: FID of real test motions after joint-space corruption and re-encoding through the dataset's feature pipeline, under the released evaluator (red) and each retrained seed (blue). The dotted line is the identity control under the released evaluator (unperturbed joints re-encoded the same way; 0.008 to 0.019 across seeds).

Corruption suite. Figure 3 plots FID against severity for each corruption under every evaluator. Joint jitter moves FID monotonically from 0.13 at 0.25% of body height to 68.5 at 4%, and truncation from 1.8 at 75% of the frames to 26.8 at 35%. Foot skating during ground contact of up to 8% of body height raises FID only from 0.011 to 0.020, against an identity control of 0.011, and leaves R-precision top-1 unchanged (0.426 to 0.423). Speed warps are asymmetric: halving speed costs FID 7.9, doubling it 0.98. Permuting motions among captions leaves FID between 0.013 and 0.024 at every fraction, within 0.014 of the identity control, while R-precision top-1 falls from 0.328 to 0.028: the distribution-level metric cannot see a caption-motion mismatch, by construction. The foot drift is large in the representation: its injected per-frame foot speed at the largest amplitude is 34.4 units (2.2 at the smallest), 83 times the natural foot speed during labeled contact (0.4 units), above the contact threshold of 0.22 units per frame, and larger than the standard deviation of the velocity features (29 units), so the pipeline relabels the sliding frames as non-contact and the velocity entries of four joints move by about one standard deviation. Two measurements locate the blindness. The drift barely moves the embedding: at the largest amplitude the mean displacement of a motion's 512-dimensional embedding is 0.6% of the typical distance between two real motions under the released evaluator (0.6% over seeds), against 4.1% (3.6%) for the smallest all-joint jitter. And the same Gaussian jitter restricted to the four foot joints scores FID 0.015 at 0.25% of body height, 3.9 at 4%, and 6.3 at 8%, against 0.130 and 68.5 for all joints at 0.25% and 4% and 0.020 for the largest drift. Confining a perturbation to four joints lowers FID by an order of magnitude or more at matched magnitude, but it does not explain the drift: foot-only jitter at the drift's amplitude scores 319 times the drift's FID, and even the 1% level, whose mean per-frame foot speed of 19 units is below the drift's 34.4, scores 0.215. What the evaluator misses is not foot motion but smooth, slow foot motion: a sliding foot moves like a plausible foot, and the relation that makes it a violation, motion during ground contact, lives in the contact channels the evaluator drops.

Identity controls. Re-encoding unperturbed joints through the dataset's own pipeline scores FID 0.011 under the released evaluator and 0.008 to 0.019 across seeds, so the suite's floor is small. Routing the same joints through the alternative extraction routine in the same file scores 0.096, which is 52% of MoMask's released FID. The conversion prescribed for MARDM's 64-dimensional representation coincides with the identity control by construction: the joint recovery reads only those 64 dimensions and the prescribed route then applies the dataset pipeline, so the two pools are bit-identical and score the same 0.011. A heading-fixed conversion of the same representation, which re-derives rotations and velocities without the pipeline's heading re-estimation, scores 0.061 (0.054 over seeds). No audited generator uses the alternative routine or the heading-fixed conversion, so these are bounds on what a pipeline choice can cost rather than errors in any published number. The bound is not small: a difference of 0.096 on identical motions exceeds two of the three adjacent published gaps on this benchmark.

5 Related Work

Critiques of motion metrics. Voas et al. (2023) showed that no standard motion metric correlates with human judgment at the sample level and proposed MoBERT; Wang et al. (2025) learned a critic from 52k human preference pairs and reported that FID disagrees with user studies; Chen et al.

(2026) attribute FID's bias to its Gaussian assumption and propose optimal-transport and MMD replacements; Kucherenko et al. (2024) rank-correlated a Fréchet gesture distance against human ratings. All of these hold the evaluator fixed. We do not propose a replacement metric; we measure the variance of the incumbent.

Retraining the evaluator. Meng et al. (2025) retrained the text-motion matching network on a de-redundanced representation and showed that gaps between methods compress under the new evaluator while the ordering is preserved. That is a single alternative evaluator; our contribution is the distribution over seeds of the standard recipe, the variance decomposition against generation noise and batching, and the selection-transfer experiment. Their compression is consistent with our seed means, under which the MoMask-retrieval gap shrinks and reverses. Petrovich et al. (2023) suggested their retrieval model as an alternative evaluator; our seeds are drop-in replacements for the standard checkpoint rather than new architectures.

Corruption studies. Maiorca et al. (2023) validated a Fréchet motion distance against graded foot skating and over-smoothing in a learned feature space, and Rekik et al. (2026) rated six distortion types at five severities against kinematic metrics. Neither tests the HumanML3D evaluator, and neither includes an identity control that isolates the extractor's conventions.

Fréchet distances in learned spaces. Stein et al. (2023) showed that Fréchet distances in learned feature spaces mis-rank image generators, Kynkäänniemi et al. (2023) traced FID's sensitivity to the class structure of its feature network, Chong & Forsyth (2020) characterized FID's sample-size bias, and Jayasumana et al. (2024) proposed a distance without the Gaussian assumption. Bouthillier et al. (2021) decomposed benchmark variance in supervised learning across seeds and splits. Our work is the analogous accounting for a benchmark whose feature network is small enough to retrain many times.

Retrieval and duplication. Liu et al. (2026) quantified caption duplication in HumanML3D and KIT-ML for retrieval. No prior work reports a retrieve-from-train pool under the generation metrics.

6 Limitations

The audit is on KIT-ML, the smaller of the two benchmarks (734 test motions), whose small test set likely inflates evaluator variance relative to HumanML3D; every conclusion is for KIT-ML. HumanML3D requires the Archive of Motion Capture as Surface Shapes (AMASS) (Mahmood et al., 2019) body models, which we did not obtain before the deadline; the recipe is identical and costs about 5 minutes per seed, so the replication is cheap and its absence is the paper's main gap. The movement encoder is shared across seeds because it is a separately trained autoencoder in the released recipe; retraining it per seed would only enlarge the evaluator variance. The generation-seed component rests on five seeds and its interval is wide. Our pools are single generation seeds and do not reproduce two published FIDs; we report the gap rather than tune the pools toward the published values. The selection-transfer experiment sweeps inference hyperparameters of one generator; checkpoint selection during training would require retraining generators.

7 Conclusion

A benchmark that computes every metric in one learned feature space inherits that space's randomness, and the intervals in the text-to-motion literature omit it. With 30 retrained evaluators on KIT-ML: the evaluator seed is the largest variance component of FID, R-precision, and matching score; orderings whose seed-mean gap is within two seed standard deviations reverse, including a retrieval pool overtaking the strongest generator, and the released-checkpoint gap does not predict which; hyperparameter selection transfers with regret no larger than the seed noise while the optimum's identity does not; and the feature-extraction route alone can move FID by more than a published state-of-the-art margin. None of this needs a new metric, only the variance of the one in use, and we release the seeds, the pools, and a scoring script that reports it.

Reproducibility statement

Section 2 specifies the evaluator recipe and Section 3 the retraining, pooling, scoring, and analysis procedures. The supplementary material contains the retraining, pooling, scoring, corruption, and analysis code, per-seed training logs, the per-pool and per-evaluator score files, and the analysis file from which every number in the paper is derived. The retrained evaluator checkpoints (about 245 MB each) and the frozen pools (about 145 MB each) exceed the supplement size limit; they are regenerated deterministically by the included scripts and will be released with the camera-ready version.

AI use statement

Large language model tools were used as coding and writing assistants for this submission: to draft and refactor the evaluation, retraining, and analysis scripts, to search and summarize related work, and to edit the prose. All experimental design decisions, the interpretation of results, and the final text were reviewed and are the responsibility of the authors. No large language model was used as an evaluator or judge in any experiment; every reported number is produced by the deterministic scoring code in the supplementary material.

References

Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gaël Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. In *Proceedings of Machine Learning and Systems (MLSys)*, volume 3, 2021.

Wenshuo Chen, Haozhe Jia, Kuimou Yu, Songning Lai, Lei Wang, and Yutao Yue. Towards better evaluation metrics for text-to-motion generation. In *Companion Proceedings of the ACM Web Conference 2026 (WWW Companion)*, pp. 1301–1311, 2026. doi: 10.1145/3774905.3794682.

Xin Chen, Biao Jiang, Wen Liu, Zilong Huang, Bin Fu, Tao Chen, and Gang Yu. Executing your commands via motion diffusion in latent space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18000–18010, 2023. arXiv:2212.04048.

Min Jin Chong and David Forsyth. Effectively unbiased FID and Inception score and where to find them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6069–6078, 2020. doi: 10.1109/CVPR42600.2020.00611. arXiv:1911.07023.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5152–5161, 2022.

Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. MoMask: Generative masked modeling of 3d human motions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1900–1910, 2024. doi: 10.1109/CVPR52733.2024.00186. arXiv:2312.00063.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pp. 1735–1742, 2006. doi: 10.1109/CVPR.2006.100.

James A. Hanley and Abby Lippman-Hand. If nothing goes wrong, is everything all right? interpreting zero numerators. *JAMA*, 249(13):1743–1745, 1983. doi: 10.1001/jama.1983.03330370053031.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017. arXiv:1706.08500.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking FID: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9307–9315, 2024. arXiv:2401.09603.

Taras Kucherenko, Pieter Wolfert, Youngwoo Yoon, Carla Viegas, Teodor Nikolov, Mihail Tsakov, and Gustav Eje Henter. Evaluating gesture generation in a large-scale open challenge: The GENEA challenge 2022. *ACM Transactions on Graphics*, 43(3):37:1–37:28, 2024. doi: 10.1145/3656374. arXiv:2303.08737.

Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The role of ImageNet classes in Fréchet inception distance. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2203.06026.

Fulong Liu, Liang Xu, Chengqun Yang, Yuhao Zhang, Yichao Yan, and Xiaokang Yang. MRBench: A comprehensive benchmark for human motion-text retrieval. *arXiv preprint arXiv:2608.07993*, 2026. arXiv:2608.07993.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5441–5450, 2019. doi: 10.1109/ICCV.2019.00554. arXiv:1904.03278.

Antoine Maiorca, Hugo Bohy, Youngwoo Yoon, and Thierry Dutoit. Objective evaluation metric for motion generative models: Validating Fréchet motion distance on foot skating and over-smoothing artifacts. In *Proceedings of the 16th ACM SIGGRAPH Conference on Motion, Interaction and Games (MIG)*, pp. 1–11, 2023. doi: 10.1145/3623264.3624443.

Zichong Meng, Yiming Xie, Xiaogang Peng, Zeyu Han, and Huaizu Jiang. Rethinking diffusion for text-driven human motion generation: Redundant representations, evaluation, and masked autoregression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 27859–27871, 2025. doi: 10.1109/CVPR52734.2025.02594. arXiv:2411.16575.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014. doi: 10.3115/v1/D14-1162.

Mathis Petrovich, Michael J. Black, and Gül Varol. TMR: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9488–9497, 2023. arXiv:2305.00976.

Ekkasit Pinyoanuntapong, Muhammad Usama Saleem, Pu Wang, Minwoo Lee, Srijan Das, and Chen Chen. BAMM: Bidirectional autoregressive motion model. In *Computer Vision – ECCV 2024*, volume 15073 of *Lecture Notes in Computer Science*, pp. 172–190. Springer, 2024. doi: 10.1007/978-3-031-72633-0_10. arXiv:2403.19435.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028.

Rim Rekik, Stefanie Wuhrer, Ludovic Hoyet, Katja Zibrek, and Anne-Hélène Olivier. Quality assessment of 3d human animation: Subjective and objective evaluation. *IEEE Transactions on Visualization and Computer Graphics*, 32(2):1780–1792, 2026. doi: 10.1109/TVCG.2025.3631385. arXiv:2505.23301.

George Stein, Jesse C. Cresswell, Rasa Hosseinzadeh, Yi Sui, Brendan Leigh Ross, Valentin Villecroze, Zhaoyan Liu, Anthony L. Caterini, J. Eric T. Taylor, and Gabriel Loaiza-Ganem. Exposing flaws of generative model evaluation metrics and their unfair treatment of diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. arXiv:2306.04675.

Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H. Bermano. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2209.14916.

Jordan Voas, Yili Wang, Qixing Huang, and Raymond Mooney. What is the best automated metric for text to motion generation? In *SIGGRAPH Asia 2023 Conference Papers*, pp. 1–11, 2023. doi: 10.1145/3610548.3618185. arXiv:2309.10248.

Haoru Wang, Wentao Zhu, Luyi Miao, Yishu Xu, Feng Gao, Qi Tian, and Yizhou Wang. Aligning human motion generation with human perceptions. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.02272.

Jianrong Zhang, Yangsong Zhang, Xiaodong Cun, Yong Zhang, Hongwei Zhao, Hongtao Lu, Xi Shen, and Ying Shan. T2M-GPT: Generating human motion from textual descriptions with discrete representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14730–14740, 2023. doi: 10.1109/CVPR52729.2023.01415. arXiv:2301.06052.

A Additional details

A.1 Evaluator retraining details

The evaluator recipe follows the released training script. Text encoder: GloVe-300 word vectors concatenated with a 15-way part-of-speech one-hot code, bidirectional GRU with hidden size 512 and a 512-dimensional output projection. Motion encoder: bidirectional GRU with hidden size 1024 over the 512-dimensional snippet codes produced by the frozen movement encoder (a two-layer 1D convolutional network over 4-frame windows of the first 247 feature dimensions). Loss: the contrastive loss of Hadsell et al. (2006) with margin 10; each batch contributes one positive term over caption-motion pairs and one negative term over the batch cyclically shifted by a random offset. Optimizer: Adam, learning rate 10^{-4}, batch size 32, gradient norm clipped at 0.5, 40 epochs, the checkpoint with the lowest validation loss retained. The seed controls parameter initialization, data order, and the random caption and crop selected for each training motion. Each seed trains in about 5 minutes on one RTX 2060.

A.2 Pool construction

Each pool row fixes one test motion, one of its captions, and one crop drawn with a fixed data seed, following the standard evaluation loader (minimum length 24 frames, maximum 196, lengths rounded down to a multiple of 4). Generators receive the caption string and, where they require it, the crop length. Generated motions are stored in the dataset's native feature space and z-normalized with the evaluator's statistics at scoring time. The retrieval pool embeds every training caption with the released text encoder, retrieves the nearest one for each test caption by Euclidean distance, and returns the corresponding training motion at its own length; a variant that excludes the mirror twin of the test motion from the index produced identical retrievals on KIT-ML, because the nearest training caption is almost always a verbatim duplicate attached to an unrelated motion.

A.3 Scoring protocol

For each evaluator and pool, the rows are shuffled and cut into batches of 32 (dropping the remainder); R-precision and matching score are accumulated per batch; FID, Diversity, and the real-motion statistics are computed over the retained rows. Five batching repeats use five fixed permutations. Generation seeds for MoMask are separate pools.

A.4 Corruption suite

Joint positions are recovered with the released inverse mapping and re-encoded with the released feature extractor after each corruption. Jitter adds isotropic Gaussian noise to every joint. Foot skate adds a sinusoidal horizontal displacement to the four foot joints on frames whose stored contact labels are active, so the corrupted motion slides while the pose is intact. Speed warp resamples the sequence in time by linear interpolation. Truncation keeps a prefix of the frames. The caption permutation swaps the motions of a random subset of rows among themselves by a cyclic shift, so the marginal motion distribution is unchanged. The foot-only jitter control adds the same isotropic

Gaussian noise to the four foot joints alone. Body height is the median over rows of the vertical joint extent. Embedding displacement is the mean over rows of the Euclidean distance between the corrupted and identity-control motion embeddings, divided by the mean distance between two random real motions under the same evaluator.

A.5 Full MoMask sweep (24 configurations)

Table 6 lists every configuration of the sweep in Section 4.4 under the released evaluator and across seeds.

Configuration	FID released	FID seeds (mean $\pm$ sd)
guidance 1, 10 iterations	0.241	0.250 ± 0.052
guidance 1, 18 iterations	0.243	0.241 ± 0.049
guidance 1, 49 iterations	0.245	0.237 ± 0.048
guidance 1, 5 iterations	0.287	0.291 ± 0.061
guidance 2, 10 iterations	0.185	0.263 ± 0.052
guidance 2, 18 iterations	0.181	0.269 ± 0.054
guidance 2, 49 iterations	0.169	0.235 ± 0.047
guidance 2, 5 iterations	0.218	0.292 ± 0.051
guidance 3, 10 iterations	0.207	0.332 ± 0.064
guidance 3, 18 iterations	0.226	0.343 ± 0.066
guidance 3, 49 iterations	0.240	0.364 ± 0.068
guidance 3, 5 iterations	0.222	0.355 ± 0.061
guidance 4, 10 iterations	0.278	0.448 ± 0.088
guidance 4, 18 iterations	0.300	0.480 ± 0.090
guidance 4, 49 iterations	0.274	0.445 ± 0.086
guidance 4, 5 iterations	0.249	0.413 ± 0.076
guidance 6, 10 iterations	0.502	0.759 ± 0.143
guidance 6, 18 iterations	0.546	0.820 ± 0.150
guidance 6, 49 iterations	0.544	0.814 ± 0.158
guidance 6, 5 iterations	0.480	0.683 ± 0.132
guidance 8, 10 iterations	0.892	1.289 ± 0.222
guidance 8, 18 iterations	0.905	1.227 ± 0.212
guidance 8, 49 iterations	0.956	1.274 ± 0.222
guidance 8, 5 iterations	0.624	0.883 ± 0.180

Table 6: MoMask inference sweep on KIT-ML under the released evaluator and across seeds.

A.6 Selection transfer under every evaluator

Selecting evaluator	selected configuration	mean regret	max regret	argmin agreement
released	guidance 2, 49 iterations	0.020	0.106	50%
seed0	guidance 1, 49 iterations	0.025	0.108	33%
seed1	guidance 1, 49 iterations	0.025	0.108	33%
seed2	guidance 2, 49 iterations	0.020	0.106	50%
seed3	guidance 2, 49 iterations	0.020	0.106	50%
seed4	guidance 2, 49 iterations	0.020	0.106	50%
seed5	guidance 2, 49 iterations	0.020	0.106	50%
seed6	guidance 1, 18 iterations	0.029	0.098	10%
seed7	guidance 1, 49 iterations	0.025	0.108	33%
seed8	guidance 1, 49 iterations	0.025	0.108	33%
seed9	guidance 1, 49 iterations	0.025	0.108	33%
seed10	guidance 1, 49 iterations	0.025	0.108	33%
seed11	guidance 1, 18 iterations	0.029	0.098	10%
seed12	guidance 1, 49 iterations	0.025	0.108	33%
seed13	guidance 2, 49 iterations	0.020	0.106	50%
seed14	guidance 2, 49 iterations	0.020	0.106	50%
seed15	guidance 1, 49 iterations	0.025	0.108	33%
seed16	guidance 1, 49 iterations	0.025	0.108	33%
seed17	guidance 2, 49 iterations	0.020	0.106	50%
seed18	guidance 2, 49 iterations	0.020	0.106	50%
seed19	guidance 2, 49 iterations	0.020	0.106	50%
seed20	guidance 2, 49 iterations	0.020	0.106	50%
seed21	guidance 2, 49 iterations	0.020	0.106	50%
seed22	guidance 2, 49 iterations	0.020	0.106	50%
seed23	guidance 1, 18 iterations	0.029	0.098	10%
seed24	guidance 2, 49 iterations	0.020	0.106	50%
seed25	guidance 2, 49 iterations	0.020	0.106	50%
seed26	guidance 2, 49 iterations	0.020	0.106	50%
seed27	guidance 1, 49 iterations	0.025	0.108	33%
seed28	guidance 1, 49 iterations	0.025	0.108	33%
seed29	guidance 1, 18 iterations	0.029	0.098	10%

Table 7: Selection transfer under every evaluator (released checkpoint and all 30 retrained seeds); the columns are those of Table 5.