Author translation of the published TOK paper (original in Turkish): https://doi.org/10.5281/zenodo.24052125.
Please cite the published version.

An End-to-End Evaluation of Text-to-Motion Generation, Retargeting, and Tracking in Humanoid Robot Simulation

Kadir Yavuz Kurt, Hakan Temeltaş

Department of Control and Automation Engineering
Istanbul Technical University, Istanbul, Türkiye
{kurtk19, temeltash}@itu.edu.tr

Abstract

Text-driven diffusion models can synthesize realistic human motions from natural language; whether these motions are executable on humanoid robots, however, depends on physical criteria that generation metrics do not capture. The outputs of three diffusion models (MDM, ReMoDiffuse, MotionDiffuse) are retargeted to the Unitree G1 humanoid platform with two retargeting methods (GMR, PyRoki) and tracked in a physics simulation of this platform by BeyondMimic-based reinforcement learning policies. Across 360 retargeting evaluations and 16 tracking policies, GMR consistently outperforms PyRoki on foot slip and stance clearance, and this advantage carries over to tracking error and joint jerk in a paired comparison on ReMoDiffuse references. The best-performing generative model varies across action classes, and in an 18-cell correlation analysis no generation metric emerges as a reliable predictor of retargeting quality after multiple-comparison correction. The findings indicate that evaluating text-to-motion models for robotics must include the retargeting and tracking stages.

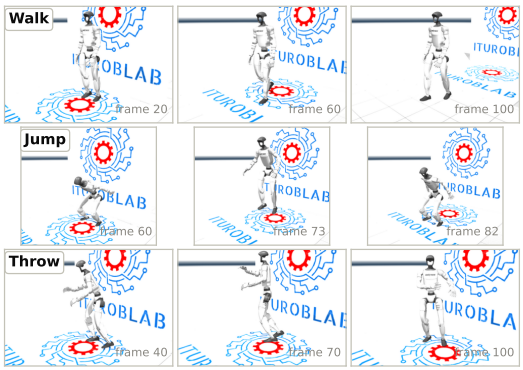


Figure 1: Frames from ReMoDiffuse motions retargeted to Unitree G1 with GMR: representative sequences of the walking, jumping, and throwing classes (time progresses from left to right).

1. Introduction

The ability of humanoid robots to take on tasks in environments designed for humans depends on generating and executing human-like, physically consistent whole-body motions [1]. Text-driven diffusion models have reached notable maturity in generating realistic human motions from natural-language descriptions [2, 3, 4], giving rise to a scalable source of reference motions as an alternative to costly motion capture recordings.

Two fundamental obstacles remain, however. First, generated motions are defined on a human skeleton and, because of the embodiment gap, can be transferred to a robot only through retargeting. Second, the metrics commonly used in generation, such as FID (Fréchet Inception Distance), diversity, and text–motion alignment, measure distributional and semantic quality; they do not cover physical executability criteria such as foot slip, ground penetration, and dynamic consistency [5]. Moreover, artifacts carried from retargeting into the reference trajectories are known to directly degrade the performance of tracking policies [6]. This gives rise to the research question of this paper: *do motion generation metrics predict whether generated motions are executable on a humanoid robot?*

In this paper, this question is investigated through an end-to-end pipeline: (i) motion generation from HumanML3D [7] text descriptions with three widely used diffusion models (MDM [2], ReMoDiffuse [4], MotionDiffuse [3]); (ii) retargeting of the generated motions to the Unitree G1 morphology with GMR [6] and PyRoki [8]; (iii) tracking of the retargeted trajectories in a physics simulation of the Unitree G1 platform with BeyondMimic-based [9] reinforcement learning (RL) policies. Representative frames from the pipeline output are shown in Fig. 1.

The main contributions of the paper are:

- An end-to-end evaluation of generated motions in humanoid robot simulation through a generation–retargeting–RL tracking pipeline;
- A per-action-class comparison of GMR and PyRoki using six error metrics and contact quality, with implications for selecting RL references;
- A quantification, through correlation analysis, of the relationship between generation metrics and executability on the robot.

Author translation of the published TOK paper (original in Turkish): https://doi.org/10.5281/zenodo.24052125.
Please cite the published version.

2. Related Work

Text-driven human motion generation. MotionDiffuse is the first large-scale text-driven motion diffusion model [3]. MDM, which trains a lightweight transformer-based architecture with geometric losses, has become the field's widely used baseline [2]. ReMoDiffuse augments the generation process with retrieval from a motion database to improve distributional realism [4]. These models are evaluated largely on the HumanML3D dataset and its accompanying R-Precision, FID, MM-Dist, and diversity metrics [7]. In our previous work, the three models were compared per action class using physical plausibility criteria such as mean jerk and foot slip in addition to semantic metrics [5]; that evaluation, however, remained at the kinematic level on the human skeleton.

Retargeting to humanoid robots. Optimization-based kinematic methods are widely used to transfer human motion to robot morphologies. PyRoki is a modular toolkit that unifies inverse kinematics, trajectory optimization, and retargeting within a common differentiable cost framework [8]. GMR offers a retargeting pipeline specialized for humanoid motion tracking; by accounting for body scaling, contact consistency, and velocity limits, it produces reference trajectories suitable for RL tracking [6]. The same study showed experimentally that retargeting quality is the dominant factor determining tracking performance. In this paper, the two approaches are compared on motions generated from text. To our knowledge, this is the first systematic study to compare two retargeting methods on generative model outputs rather than motion capture recordings; existing text-to-robot pipelines do not evaluate the retargeting stage comparatively.

Whole-body tracking with reinforcement learning. In physics-based character animation, DeepMimic introduced training components that remain standard today, such as tracking rewards, reference state initialization, and early termination [10]. With the transfer of this approach to humanoid robots and the spread of massively parallel simulation, whole-body tracking policies have achieved agile skills on real hardware; progress in the field is surveyed in [1]. BeyondMimic is a recent representative that combines high-fidelity tracking with guided diffusion control [9]. This paper adopts BeyondMimic's tracking architecture; unlike common practice in the literature, however, the references are not motion capture recordings but retargeted outputs of text-driven diffusion models. The performance interplay of the generation, retargeting, and tracking stages is thereby evaluated as a whole.

3. Method and Experimental Design

The evaluation pipeline follows the three stages described in Section 1; performance is measured both with stage-specific metrics and with end-to-end tracking performance.

3.1. Text-Driven Reference Motion Generation

Reference motions are synthesized with MDM [2], ReMoDiffuse [4], and MotionDiffuse [3] using prompts derived from text descriptions in the HumanML3D [7] dataset. Generation is configured to cover six action classes: walking, running, jumping, picking up, swinging, and throwing. The kinematic and semantic evaluation of these models' generation quality was presented in our previous work [5]; in this paper, the generation stage is treated as the input source of the pipeline.

3.2. Retargeting to the Humanoid Robot

The 3D human motion sequences obtained from the generative models are transferred to the 29-degree-of-freedom Unitree G1 humanoid robot morphology with the two methods introduced in Section 2 (PyRoki [8] and GMR [6]). PyRoki was run with the toolkit's example humanoid retargeting setup (keypoint matching, scale regularization, alignment, foot-contact and slip-suppression costs, smoothness/posture regularization, hard joint limits; 3.5 cm sole offset in the foot-contact cost), and GMR with the default settings of its specialized pipeline, which tracks the height-scaled human trajectory.

Retargeting performance is evaluated quantitatively with six basic error metrics:

- Mean body position error in global coordinates: $E_{\text{g-mpbpe}} = \mathbb{E}\left[\|\boldsymbol{p}_t - \boldsymbol{p}_t^{\text{ref}}\|_2\right]$
- Root-relative mean body position error: $E_{\text{mpbpe}} = \mathbb{E}\left[\|(\boldsymbol{p}_t - \boldsymbol{p}_{\text{root},t}) - (\boldsymbol{p}_t^{\text{ref}} - \boldsymbol{p}_{\text{root},t}^{\text{ref}})\|_2\right]$
- DOF mismatch (joint coordinate difference between methods): $E_{\text{dof}} = \mathbb{E}\left[\|\boldsymbol{q}_t^{\text{GMR}} - \boldsymbol{q}_t^{\text{PyRoki}}\|_2\right]$
- DOF velocity mismatch (joint velocity difference between methods): $E_{\text{dof-v}} = \mathbb{E}\left[\|\Delta\boldsymbol{q}_t^{\text{GMR}} - \Delta\boldsymbol{q}_t^{\text{PyRoki}}\|_2\right]$
- Mean body velocity error: $E_{\text{mpbve}} = \mathbb{E}\left[\|\Delta\boldsymbol{p}_t - \Delta\boldsymbol{p}_t^{\text{ref}}\|_2\right]$
- Mean body acceleration error: $E_{\text{mpbae}} = \mathbb{E}\left[\|\Delta^2\boldsymbol{p}_t - \Delta^2\boldsymbol{p}_t^{\text{ref}}\|_2\right]$

Here, $\boldsymbol{p}_t$ denotes the body (link) positions in the retargeted motion, $\boldsymbol{q}_t^{\text{GMR}}$ and $\boldsymbol{q}_t^{\text{PyRoki}}$ the robot joint coordinates produced by the two methods, and Δ the discrete time difference. Sequences are processed at 20 frames/s (the HumanML3D standard); frame-based units correspond to a 50 ms step. Since no neutral reference robot joint trajectory exists, the joint-space measures (E_{dof}, $E_{\text{dof-v}}$) are defined not as the error of a single method but as the mismatch between the two methods. In addition to kinematic errors, contact quality is measured by foot slip (horizontal displacement of the foot during the contact phase), ground penetration depth, and stance clearance (height of the sole above the ground during the contact phase). Contact phases are labeled identically for both methods from the foot height in the human reference (SMPL foot height < 0.12 m); sole height is measured from the lowest point of the foot's collision geometry. Two MDM walking samples in which no contact phase could be detected were excluded from the metrics that require a contact phase (foot slip, stance clearance).

3.3. Whole-Body Tracking with Reinforcement Learning

To track the retargeted trajectories under dynamic constraints, the tracking framework of BeyondMimic [9] is adopted. The reward design follows the tracking components that have become standard since physics-based character animation [10]: in

Author translation of the published TOK paper (original in Turkish): https://doi.org/10.5281/zenodo.24052125.
Please cite the published version.

addition to joint position and velocity rewards and body position/orientation tracking rewards, regularization terms such as action-rate and joint-limit-violation penalties are used. Deviations in base height/orientation are penalized, and the episode is terminated early upon a fall. Policies are optimized with PPO and GAE. The observation–action structure and reward weights are inherited from BeyondMimic's open-source implementation; the retargeted references were resampled to 50 frames/s for training.

The learning environment is scaled to 4096 parallel environments in GPU-accelerated physics simulation (Isaac Sim). Training is run uninterrupted for 2000 iterations per reference motion on NVIDIA RTX 5070 Ti hardware.

3.4. Experimental Setup and Evaluation Criteria

The experiment matrix is organized along two axes. On the kinematic axis, the combinations of the three generative models and the two retargeting methods are evaluated over six action classes: for each model, ten samples per class are generated from three text prompts, and each of the resulting 180 motions is transferred with both methods (360 evaluations). On the dynamic axis, a separate tracking policy is trained for each combination for the selected motion texts, so that the effects of the generative model and of the retargeting method can be separated. Tracking performance is reported in terms of anchor-relative tracking error, contact quality (slip, penetration, contact timing), smoothness (joint jerk, action rate), and physical stability (fall rate, fraction of time the center of mass remains inside the support polygon), as well as the convergence behavior of the episode reward components.

Table 1: Retargeting metrics (all classes)

Metric	GMR	PyRoki	p	$\lvert r\rvert$
g-MPBPE* [mm]	**83.1**±18.2	136.8±79.9	<0.001	0.91
MPBPE [mm]	**82.2**±16.8	88.9±21.7	<0.001	0.44
MPBVE* [mm/frame]	**3.3**±3.4	7.2±4.4	<0.001	0.93
MPBAE* [mm/frame^2]	**2.1**±3.5	5.0±4.0	<0.001	0.89
Foot slip [mm]	**9.5**±10.7	11.8±12.7	<0.001	0.93
Ground penetration [mm]	1.9±5.1	3.1±8.0	0.60	0.05
Stance clearance [mm]	**16.7**±14.7	32.6±19.0	<0.001	0.70

Values are mean±std; p: Wilcoxon signed-rank test (the penetration difference is not significant); $|r|$: matched-pairs rank-biserial effect size. Starred (*) metrics are computed against the GMR-scaled human reference. Because the distributions are right-skewed, the medians are lower (e.g., foot slip 5.5 vs. 7.4 mm); the 95% confidence intervals of the median differences exclude zero for all metrics except penetration.

4. Results

4.1. Retargeting Comparison

Since the same motion is transferred with both methods, the comparison uses a paired design. Significance was assessed with the Wilcoxon signed-rank test on the per-pair error differences (n=180 pairs; n=178 for the foot slip and clearance metrics, which require a contact phase, Section 3.2).

Table 1 gives the aggregate results. GMR produced lower mean error on all discriminative metrics: foot slip is 20% lower (26% in terms of medians), and stance clearance is roughly halved (both $p < 0.001$, $|r| \geq 0.70$). Although the means appear to differ for ground penetration, the paired differences are inconsistent (p=0.60); the methods should therefore be distinguished by slip and clearance rather than by penetration. The small difference in root-relative MPBPE (8%) indicates that both methods preserve the local pose comparably well; the distinction emerges in global trajectory consistency and contact quality. Because the global and derivative metrics (g-MPBPE,

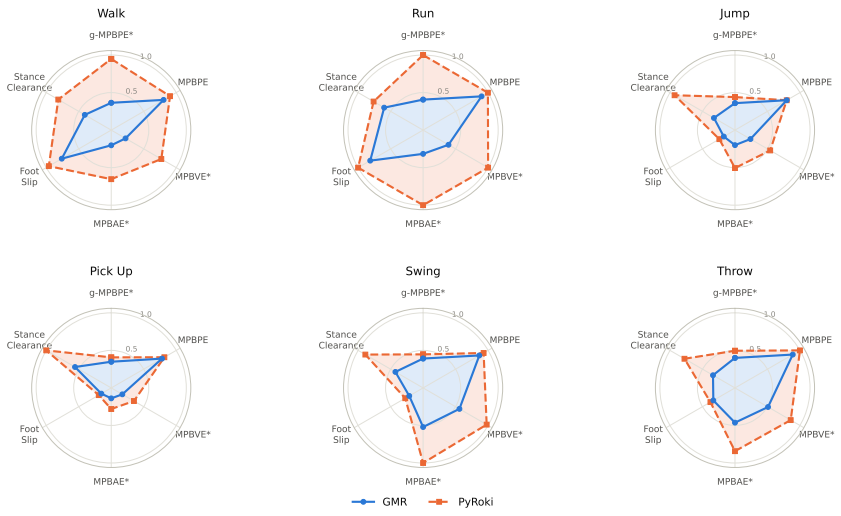


Figure 2: Normalized per-action-class error profiles of GMR and PyRoki (smaller area is better). Starred (*) metrics are computed against the GMR-scaled human reference. The DOF mismatch measures (E_{dof}, $E_{\text{dof-v}}$) are symmetric because they quantify the difference between the two methods; since they cannot be attributed to a single method, they are not used as axes in the method comparison.

Author translation of the published TOK paper (original in Turkish): `https://doi.org/10.5281/zenodo.24052125`.
Please cite the published version.

MPBVE, MPBAE) are computed against the height-scaled human reference used by GMR, they carry a structural scale bias against PyRoki. Method selection was therefore based on the scale-independent contact metrics.

Fig. 2 shows the error profiles per action class: the GMR polygon lies entirely inside the PyRoki polygon in five of the six classes (including the biased axes). The only exception is the jumping class: owing to the weight of the flight phase, the two methods are on par in root-relative pose (78.0 mm), and in penetration, which is not among the radar axes, PyRoki takes the lead (0.9 mm vs. 2.1 mm for GMR). Fig. 3 breaks down contact quality: PyRoki's contact geometry is inconsistent—in some motions the foot loses contact with the ground (high clearance, "floating" behavior), whereas in others deeper penetration is observed; from a reinforcement learning perspective, this inconsistency weakens the contact timing information.

The DOF mismatch E_{dof} defined between the two methods (Section 3.2) is reported as a single value per model: MDM 229.1, MotionDiffuse 191.5, ReMoDiffuse 158.0 milliradians. This ordering shows that the two methods diverge most on MDM motions, which is consistent with the interpretation that generation noise increases retargeting ambiguity.

4.2. Comparison Along the Generative Model Axis

Along the generative model axis, the ranking is consistent across all metrics (Fig. 4): ReMoDiffuse yields the most compatible and MDM the weakest transfer profile (e.g., g-MPBPE 103.5 / 108.0 / 118.4 mm; foot slip 9.7 / 10.7 / 11.6 mm; penetration 1.7 / 1.9 / 3.8 mm for ReMoDiffuse / MotionDiffuse / MDM, respectively). Whereas the best distributional realism (FID) in our previous work belonged to MotionDiffuse [5], ReMoDiffuse leads on all metrics in terms of compatibility with transfer to the robot. This finding is the first indication that generation metrics alone do not predict transferability to the robot; a systematic correlation analysis is presented in Section 4.4. MDM's high DOF velocity mismatch is consistent with the high generative diversity and jerk profile observed for the same model in the previous work.

4.3. Whole-Body Tracking Performance with RL

On the dynamic axis, 16 tracking policies were trained. Because of the training budget, four classes representing distinct contact and dynamic regimes were selected: walking (periodic single support), jumping (flight phase), swinging (rhythmic double support), and throwing (arm motion with weight transfer). Instead of a full factorial design (4 classes $\times$ 3 models $\times$ 2 methods = 24 runs), the two factors were isolated separately: the generative model effect was measured on the three models using references from GMR, which was found superior on the kinematic axis (12 runs); the retargeting effect was measured through a paired comparison between the GMR and PyRoki references of ReMoDiffuse, which yielded the most compatible transfer profile (4 additional runs). In each cell, the representative motion was chosen as the sample closest to the median root-relative error; the within-class MPBPE spread of the candidate pools is narrow in 14 of the 16 cells (IQR $\leq$ 12 mm) and markedly wider in two cells (throwing–MotionDiffuse, swinging–MDM). In these two cells, a different representative could change the cell value considerably; however, since the best models of the corresponding classes come from narrow-pool cells (MotionDiffuse for swinging, ReMoDiffuse for throwing), the sensitivity of the class winners to this choice is limited. The $\pm$std in Table 2 is computed over 50 evaluation episodes of the single representative motion of each cell. Training took 35.7 minutes per run on average (32.9–42.9), and all runs converged. Evaluation was performed with 50 episodes per policy, with external pushes and observation noise disabled (variability across episodes stems only from initial state randomization); tracking error is reported using BeyondMimic's definition of anchor-relative mean body position error (MPJPE; a different measure from the MPBPE of the retargeting stage) [9].

Table 2: RL tracking performance: anchor-relative MPJPE [mm]

		GMR		PyRoki
Action	MDM	MotionD.	ReMoD.	ReMoD.
Walk	38.1±2.2	56.1±1.7	**28.2**±1.6	52.1±20.7
Jump	**35.0**±3.6	44.3±3.1	40.4±3.4	47.4±3.4
Swing	46.5±12.0	**31.0**±3.5	35.3±4.0	58.5±3.5
Throw	50.9±20.0	51.1±4.0	**38.9**±3.3	57.6±5.4

50 episodes per policy, mean±std. The first three columns were trained with GMR references and the last column with PyRoki references; bold values mark the lowest error per class.

Table 2 summarizes the results. The retargeting effect carries over directly to the dynamic axis: with references from the same generative model (ReMoDiffuse), GMR yielded lower

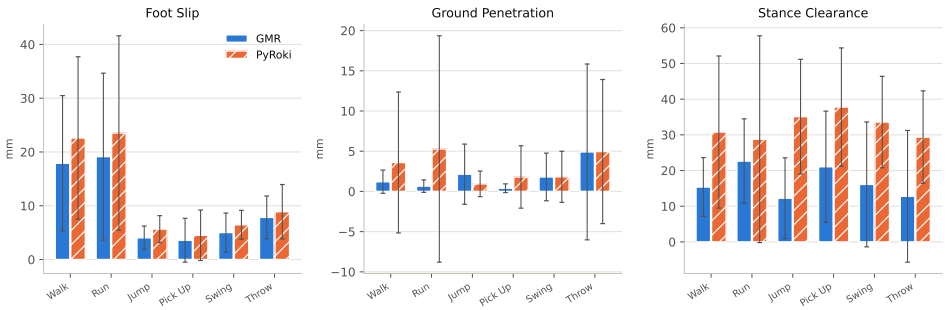


Figure 3: Distribution of contact quality metrics by action class and retargeting method (mean±std; lower is better).

tracking error than PyRoki in all four classes (e.g., 28.2 vs. 52.1 mm for walking) and reduced joint jerk by a factor of 1.4–2.0 (class median 397 vs. 649 rad/s^3). Fig. 5 shows that this gap persists throughout training. Fall rates are low for all combinations (at most 8%, walking–ReMoDiffuse–PyRoki). These results are consistent with the interpretation that the kinematic ranking in Section 4.1 carries over to trackability: a contact-consistent reference yields a smoother policy with less corrective burden. Note that references hovering above the ground impose a structural floor on the position error; the differences in jerk and fall rate are independent of this floor effect. The fraction of time the center of mass remains inside the support polygon also favors GMR in all four classes (e.g., 53% vs. 36% for walking).

The generative model effect, in contrast, is class-dependent: ReMoDiffuse gave the lowest error for walking and throwing, MotionDiffuse for swinging, and MDM for jumping. The overall ranking on the kinematic axis is only partially preserved on the dynamic axis, which suggests that the generative model may need to be chosen according to the target action class (an observation limited by a single representative motion per cell). Regarding robustness, when the two best policies were initialized from different phases of the reference (0–75%), the fall rate did not exceed 2% and the tracking error increased by at most 6.6 mm. Contact timing agreement was additionally measured to be low for the PyRoki combinations; however, this value is partly a measurement artifact, because hovering references produce sparse contact labels, and it was not used in the method comparison. To verify that the results do not depend on a single training seed, the four ReMoDiffuse–GMR runs were repeated with three seeds each; the inter-seed range of the mean tracking error over the last 100 iterations remained below 2.3 mm (6% relative) in all classes, and the variation in mean reward below 3%. The PyRoki runs, however, are limited to a single training seed: although the narrow inter-seed range observed on the GMR side is small compared with the 7–24 mm differences between methods, it cannot be assumed that this band also holds for PyRoki. In particular, the walking cell, which shows high variability across episodes (±20.7 mm) and an 8% fall rate, is where seed sensitivity may be highest; the GMR–PyRoki comparison in Table 2 should be read with this asymmetry in mind.

4.4. Correlation Analysis

The relationship between generation metrics and transfer quality was quantified with the following setup. On the generation side, the FID, MM-Dist, diversity, and MultiModality metrics from the evaluation set of the previous work [5] are used, together with the foot slip measured on the human skeleton. On the transfer side, seven metrics are used: two position errors ($E_{\text{g-mpbpe}}$, E_{mpbpe}) and three contact metrics (foot slip, ground penetration, stance clearance) of the GMR transfer, together with the two inter-method mismatch measures (E_{dof}, $E_{\text{dof-v}}$). Each metric was aggregated by within-class averaging into 3 models $\times$ 6 classes $=$ 18 cells; relationships were measured with Spearman rank correlation over the full family of 35 pairs formed by the 5 generation $\times$ 7 transfer metrics and corrected with the Holm and Benjamini–Hochberg (BH) procedures. Reliable predictiveness was defined as significance at the α=0.05 level after correction. Because the DOF mismatch measures are defined between the two methods and are symmetric, they should be interpreted as indicators of retargeting ambiguity, whereas the GMR position and contact metrics indicate method-specific transfer quality.

By this criterion, none of the examined generation metrics is a reliable predictor of transfer quality: although a raw $p < 0.05$ is observed for 10 of the 35 pairs, no pair remains significant after correction (smallest BH-adjusted p=0.078); the 18-cell sample already limits statistical power. At the exploratory level, the strongest relationship lies not in distributional realism but in generation consistency: as MultiModality, which measures the diversity among generations for the same prompt, increases, the DOF mismatch also increases (E_{dof}; ρ=0.65). FID is only moderately related to foot slip on the robot side (ρ=0.57) and is unrelated to the DOF mismatch ($\rho = -0.33$). The main relationships have similar direction and magnitude on the PyRoki transfer side as well (e.g., FID–foot slip ρ=0.62); the findings are therefore not specific to a single retargeter. Foot slip on the human skeleton is informative within a class, correctly pointing to the best model in five of the six classes, but it can invert the ranking at the model level: ReMoDiffuse, which slips the most on the human side, is the model that slips the least after transfer. A possible mechanism for this inversion is that re-

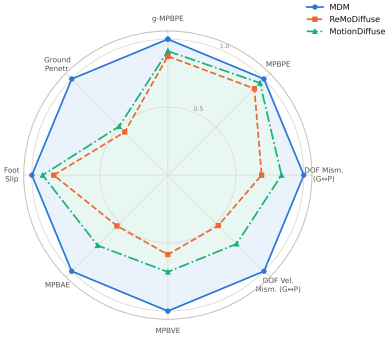


Figure 4: Retargeting profiles by generative model (average of GMR and PyRoki; normalized; smaller area is better).

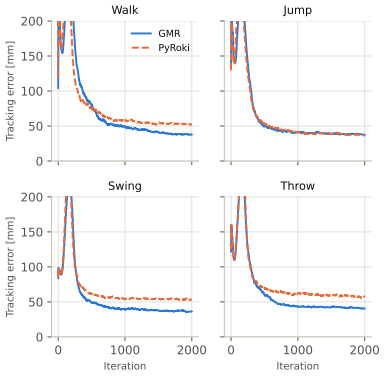


Figure 5: Online tracking error during training (ReMoDiffuse references; 25-iteration moving average).

Author translation of the published TOK paper (original in Turkish): https://doi.org/10.5281/zenodo.24052125.
Please cite the published version.

targeting acts as a contact-enforcing filter: since both methods pin the foot to the ground during the contact phase, absolute slip is largely removed in transfer, and the remaining slip becomes sensitive to the joint-velocity consistency of the motion. The fact that ReMoDiffuse, whose retrieval-based generation yields velocity-consistent sequences, has the lowest DOF mismatch (158 mrad), together with MDM's known high-jerk profile [5], is consistent with this explanation; definitive confirmation requires per-motion paired human–robot slip measurements.

5. Conclusion

In this paper, motions generated by text-driven diffusion models were evaluated end to end, together with the retargeting and RL tracking stages, in a simulation of the Unitree G1 platform. On the kinematic axis, GMR was found to be consistently superior to PyRoki in foot slip and stance clearance; in the paired comparison with ReMoDiffuse references, this advantage carried over to the dynamic axis, yielding lower tracking error in all four classes and reducing joint jerk by a factor of 1.4–2.0. Along the generative model axis, ReMoDiffuse yielded the most compatible transfer profile, whereas on the dynamic axis the ranking turned out to be class-dependent.

Do motion generation metrics predict executability on the robot? The correlation analysis (Section 4.4) answers this question in the negative for the examined setup: in the 35-pair family, no generation metric emerges as a reliable predictor of transfer quality after multiple-comparison correction. Even the strongest raw relationship comes not from distributional realism (FID) but from the MultiModality metric, which reflects generation consistency; foot slip on the human skeleton, while informative within a class, can invert the ranking at the model level. Generation metrics therefore cannot substitute for an assessment of transferability to the robot; when text-to-motion models are used in robotics, the evaluation must also cover the retargeting and dynamic tracking stages.

The findings should be read within the following limits: the evaluation uses a single robot morphology and is carried out only in physics simulation; the generative models are 2022–2023-era models that do not account for physical constraints; the dynamic axis is limited to four classes and a single motion per cell; since the PyRoki runs were conducted with a single seed, the dynamic comparison has not been validated against seed variability (multi-seed validation in at least one class is a priority); and the semantic fidelity of the executed motion to the text was not evaluated. The simulation-to-reality (sim-to-real) gap also conditions the interpretation: because of differences in actuator dynamics and in contact and friction modeling, the absolute error and fall values are expected to increase on hardware, whereas the rankings between methods and between models are expected to remain more robust, since they stem from reference trajectory quality; moreover, a concrete transfer path exists, as the underlying tracking architecture [9] and GMR [6] have been validated on real G1 hardware. Future work will cover the transfer of this pipeline to real hardware and the evaluation, within the same framework, of models that account for physical constraints at the generation stage.

6. Acknowledgment

This work was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under project no. 126E386.

7. References

[1] Zhaoyuan Gu, Junheng Li, Wenlan Shen, Wenhao Yu, Zhaoming Xie, Stephen McCrory, Xianyi Cheng, Abdulaziz Shamsah, Robert Griffin, C Karen Liu, et al., "Humanoid locomotion and manipulation: Current progress and challenges in control, planning, and learning," *IEEE/ASME Transactions on Mechatronics*, 2025.

[2] Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit Haim Bermano, "Human motion diffusion model," in *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

[3] Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu, "Motiondiffuse: Text-driven human motion generation with diffusion model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 6, pp. 4115–4128, 2024.

[4] Mingyuan Zhang, Xinying Guo, Liang Pan, Zhongang Cai, Fangzhou Hong, Huirong Li, Lei Yang, and Ziwei Liu, "Remodiffuse: Retrieval-augmented motion diffusion model," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023, pp. 364–373.

[5] Ali Ihsan Ozcetin, Burak Tantay, Kadir Yavuz Kurt, and Hakan Temeltas, "A comparative study of text-driven diffusion models for generative human-like motion," in *2026 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2026, pp. 1140–1145.

[6] Joao Pedro Araujo, Yanjie Ze, Pei Xu, Jiajun Wu, and C. Karen Liu, "Retargeting matters: General motion retargeting for humanoid motion tracking," *arXiv preprint arXiv:2510.02252*, 2025.

[7] 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*, 2022, pp. 5142–5151.

[8] Chung Min Kim, Brent Yi, Hongsuk Choi, Yi Ma, Ken Goldberg, and Angjoo Kanazawa, "Pyroki: A modular toolkit for robot kinematic optimization," in *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.

[9] Qiayuan Liao, Takara E Truong, Xiaoyu Huang, Yuman Gao, Guy Tevet, Koushil Sreenath, and C Karen Liu, "Beyondmimic: From motion tracking to versatile humanoid control via guided diffusion," *arXiv preprint arXiv:2508.08241*, 2025.

[10] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions on Graphics*, vol. 37, no. 4, pp. 1–14, 2018.