

GRAFT: Geometry-Aware Residual Adaptation for Foundation Trackers

Liang Li¹, Guoyao Zhang¹, Xingyu Chen^{1,2}, Sikai Wu¹, Qianyang Wu¹, Yuhang Lin¹,
Renjing Xu², Mingliang Zhou^{1†}, Diyun Xiang¹
<https://graft-humanoid.github.io/>

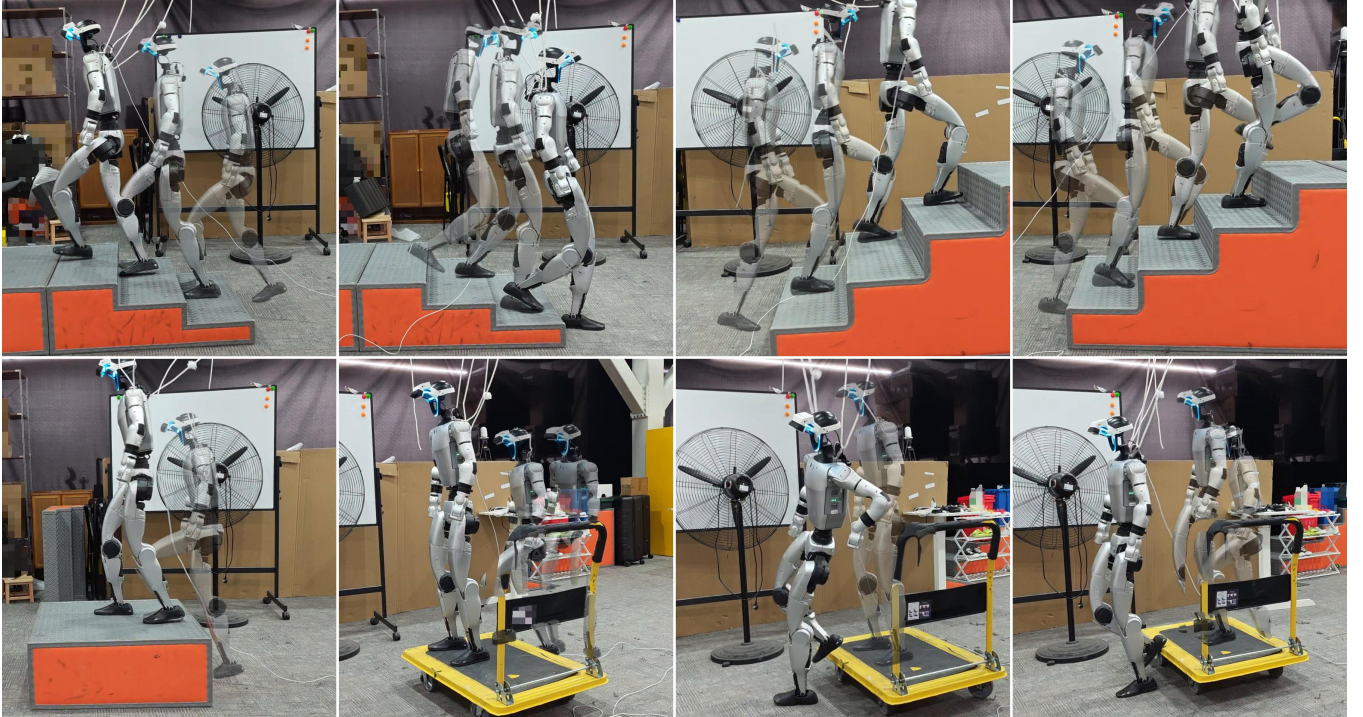


Fig. 1. GRAFT on hardware. The frozen SONIC base is driven by an unmodified reference while a geometry-aware lower-body residual adapts execution to the terrain: a 0.30 m platform, 0.10 m and 0.17 m steps, and a push-cart traversal. Frames within a scene are overlaid with increasing opacity.

Abstract—Humanoid behavior foundation models (BFMs) provide reusable whole-body motion priors: given a kinematic reference, they generate coordinated joint commands for motion replay, teleoperation, and VLA control. However, trained primarily on flat-ground motion, BFMs lack an explicit mechanism to adapt foot placement, swing clearance, and contact timing to local terrain. Many terrain-aware BFM approaches address this limitation by retraining or fine-tuning the tracker together with perception, incurring substantial data and computation costs while coupling ordinary tracking to the perception pathway. To bridge this gap, we propose GRAFT, a geometry-aware residual adaptation framework for frozen behavior trackers. GRAFT preserves the BFM parameters and motion-reference interface, while a lightweight adapter adds terrain-conditioned corrections only to lower-body execution targets. The adapter is trained with task and foot-geometry rewards, regularized to suppress unnecessary corrections on flat ground, and distilled from privileged terrain observations to onboard depth. Because the frozen BFM already provides whole-body reference tracking and nominal locomotion, GRAFT reduces terrain learning to a focused execution-adaptation problem. We instantiate the framework on four frozen trackers and evaluate stairs, gaps, and platforms in simulation. We further validate it through real-

world deployment on a Unitree G1, where the adapted SONIC tracker traverses terrain with onboard depth and unmodified references.

I. INTRODUCTION

A behavior foundation model (BFM) learns a single whole-body tracking policy from a large corpus of human motion: given a kinematic reference, it outputs joint commands [1], [2], [3]. Motion references form its sole upstream command interface. The same BFM can therefore be driven by a recorded clip, an online teleoperator, or a high-level policy such as a vision-language-action (VLA) model, without retraining for each source. This shared reference interface is what lets one released tracker serve several upstream command sources.

These corpora, however, are recorded almost entirely on flat, empty ground, and the environment enters neither the policy’s observation nor its reward. A BFM therefore learns how to approximate a reference kinematically, not how to place its feet on the surface it is actually crossing. Changing the reference does not supply what is missing. Given a clip

¹Xiaomi Robotics Lab. ²HKUST(Guangzhou). [†]Corresponding author.

of a person climbing stairs, the policy reproduces the motion in the air but does not actually climb the stairs, because the foot placement, swing clearance and contact timing the steps require were never part of what it was trained to produce.

Current approaches give BFM scene-interaction ability by pairing motion with reconstructed scene geometry and retraining [4], or by fine-tuning a pretrained model together with a terrain encoder [5], [6]. These approaches are effective, but they change the deployed tracker and couple ordinary tracking to the terrain-perception pathway. Constructing terrain-paired motion data requires annotation and processing at scale, together with the compute to retrain on it. Fine-tuning a behavior prior on a narrow new task can also degrade the reusable prior, which motivates residual adaptation on a frozen backbone elsewhere [7].

To address this problem, we propose GRAFT. The frozen BFM processes the unmodified reference, while a separately trained adapter adds geometry-conditioned corrections to the lower-body execution targets; the upper-body targets remain those of the frozen tracker, since they are what a teleoperator or a VLA is commanding. A deployment gate controls whether the correction is applied at all.

The key idea is to separate reusable whole-body motion generation from terrain execution adaptation. The BFM retains responsibility for reference tracking, nominal locomotion, and whole-body coordination; the adapter learns only the terrain-dependent deviation from this nominal execution. We train it with foot-clearance and placement penalties and suppress its output on flat ground, then distill a teacher trained under privileged terrain observations into a student that reads onboard depth. This decomposition turns terrain adaptation into a focused execution problem without rewriting the upstream reference or modifying the released tracker.

Our contributions are:

- We introduce GRAFT, a terrain-adaptation framework that adds geometry-conditioned lower-body corrections to a frozen behavior tracker while preserving its parameters and its motion-reference interface, with a gate that can be switched off at deployment; mid-run disabling and re-enabling are verified in simulation.
- We propose GRAFT as a general framework for adapting frozen behavior foundation trackers to terrain interaction. We instantiate the framework on four distinct trackers, training a separate adapter for each, and show that the same residual structure and reward family acquire terrain traversal while keeping flat-ground corrections small.
- We validate GRAFT through real-world deployment on a Unitree G1. The SONIC adapter is distilled from privileged terrain observations to onboard depth, and the resulting tracker traverses terrain with unmodified references; a separate composite-course simulation verifies mid-run residual disabling and re-enabling.

II. RELATED WORK

A. Whole-body motion tracking and behavior foundation models

Whole-body motion tracking has expanded from single-clip imitation to learning from large, heterogeneous motion corpora [11]. Behavior foundation models encode that motion diversity in a reusable prior. The bases we adapt in this work span the two dominant forms this takes. SONIC exposes a latent command interface to upstream modules and is trained to track references supplied at run time, so a planner, a teleoperator or a VLA can drive the same controller [1]. Humanoid-GPT instead distills a generalist tracker for zero-shot use on unseen motion, without a task-specific latent [2]. HoloMotion [3] and further work on prompting and scaling these priors [12], [13] follow the same interface. Other approaches reach hardware-ready tracking through minimal reward formulations [14], real-time retargeting for teleoperation [15], and alternative representations of motion priors [16]. Accepting a reference from any upstream source is what makes these models reusable as an interface rather than as a skill. Their training distribution, however, is free-space motion on flat ground in empty scenes [4]: the policy observes proprioception and a reference window, and the environment enters neither the observation nor the reward. However far the corpus is scaled, foot placement and contact behavior on terrain do not emerge from it.

B. Terrain-aware whole-body control

Terrain geometry was addressed earlier along the perceptive-locomotion line: elevation-map and depth-based policies traverse stairs, gaps and parkour terrain [8], [17], [18], [19], with separate lines on surefooted placement [20], on robustness when exteroception degrades [21], and on sparse footholds [22]; related parkour controllers exist on quadrupeds [23]. Their command channel, however, is a velocity direction or a navigation goal, which specifies where the base should go rather than what the whole body should do [4]. Three routes bring terrain into a reference-driven BFM. GigaBrain-WBC recovers the geometry each motion must have been performed on and trains a world-model tracker on the resulting terrain-paired corpus. Terrain-interactive motion becomes trackable, but the deployed policy reads only proprioception, the previous action and a latent command [4]. GRAIL specializes a pretrained BFM for two task families, and its terrain-aware branch fine-tunes the controller together with a height-map encoder [5]. Perceptive BFM is the closest to our setting: it keeps the raw kinematic reference as the deployment command and injects terrain features from a torso-depth height grid through an identity-initialized residual path. The tracking prior is fine-tuned with PPO at the same time, the residual acts on the whole-body actor mean, terrain supervision needs references rewritten to conform to the terrain, and the identity gate is a training initialization rather than a deployment switch [6]. These three routes each pay a price in how perception and tracking are coupled. The deployed policy of GigaBrain-

TABLE I

CAPABILITY COMPARISON, GROUPED BY THE LINES OF WORK DISCUSSED ABOVE. WHOLE-BODY REFERENCE MEANS THE DEPLOYMENT COMMAND IS A WHOLE-BODY MOTION REFERENCE RATHER THAN A ROOT VELOCITY OR A NAVIGATION GOAL; TERRAIN SENSING MEANS THAT THE DEPLOYED POLICY CONSUMES ROBOT-SIDE EXTEROCEPTION; BASE INTACT MEANS THAT NO BASE PARAMETER IS MODIFIED; PLUG-AND-PLAY MEANS THAT THE RESIDUAL CAN BE DISABLED AT DEPLOYMENT, RETURNING THE BASE ACTION; MULTI-BASE MEANS THAT ONE PROCEDURE IS INSTANTIATED ON TWO OR MORE DISTINCT BASE MODELS. “—” MARKS A CELL THAT DOES NOT APPLY, FOR INSTANCE THE ADAPTATION COLUMNS OF A COMPLETE POLICY.

Method	Whole-body reference	Terrain sensing	Base intact	Plug-and-play	Multi-base
<i>Behavior foundation models</i>					
SONIC [1]	✓	×	—	—	—
Humanoid-GPT [2]	✓	×	—	—	—
<i>Perceptive locomotion</i>					
DPL [8]	×	✓	—	—	—
<i>Terrain-aware BFM</i>					
GigaBrain-WBC [4]	✓	×	—	—	—
GRAIL, terrain branch [5]	✓	✓	×	×	×
Perceptive BFM [6]	✓	✓	×	×	×
<i>Residual adaptation of a frozen policy</i>					
ResMimic [9]	✓	×	✓	✓	×
MOSAIC [7]	✓	×	✓	✓	×
Cerebellar residual [10]	×	×	✓	✓	×
GRAFT (ours)	✓	✓	✓	✓	✓

WBC consumes no exteroception, so its tracking does not depend on perception, at the cost of never measuring the geometry in front of it. GRAIL and Perceptive BFM bring perception in, but the terrain skill and the base weights sit in the same parameters, so a degradation in perception costs more than terrain traversal. These routes motivate a complementary question: how much terrain capability can be obtained without modifying the foundation tracker or adding a task-specific perception-and-control architecture? GRAFT studies this adaptation-side trade-off. Its central claim is that a frozen whole-body prior can reduce the task-specific structure required for terrain adaptation, while preserving the upstream reference interface.

C. Residual and parameter-efficient adaptation of a frozen policy

Learning a residual correction on top of a fixed policy has developed as a separate line. ResMimic conditions whole-body corrections on object and contact state for locomanipulation [9], MOSAIC uses a residual adapter to close an interface gap, explicitly motivated by avoiding the forgetting that continual fine-tuning causes [7], and a cerebellum-inspired controller supplies inference-time corrections for recovery from actuator and dynamics faults [10]. In all of them the frozen policy remains runnable once the residual is taken away, and ResMimic evaluates that base-only mode as a baseline. GRAFT differs in what conditions the correction and in what it writes: local terrain geometry rather than object, interface or fault state, and the lower body rather than the whole body. We additionally evaluate disabling and re-enabling the adapter within a single continuous simulation run, which none of them reports.

Related approaches project commands onto a learned competence manifold [24], modify the weights themselves [25], [26], adapt a frozen backbone through learned masks [27],

or absorb dynamics shifts on a fixed controller [28]. Table I places GRAFT against these lines: it keeps the frozen-base, additive-residual structure, conditions it on terrain geometry, and treats plug-and-play as a deployment property.

III. METHOD

GRAFT separates whole-body motion generation from terrain execution adaptation. The frozen BFM provides a reusable motion prior, including reference tracking, nominal locomotion, and whole-body coordination. A lightweight residual policy then learns the terrain-dependent deviation from this nominal execution. At each control step the frozen BFM π maps the kinematic reference and proprioception to nominal whole-body joint-position targets a_t^0 . A residual policy f_ϕ runs at the same control rate and produces corrections in the same units, and the combined targets are executed by a joint-level PD controller,

$$\Delta a_t = m \odot f_\phi(o_t^{\text{res}}), \quad a_t = a_t^0 + \Delta a_t, \quad (1)$$

where the mask m confines the correction to the lower-body joints. The output of f_ϕ carries a fixed per-joint scale and is expressed in the same units as a_t^0 , so Δa_t is the correction actually added. Channels outside the mask are identical to the base output, and zeroing the residual makes the executed target equal π ’s output at every step. Only ϕ is trained; π stays frozen. The trainable problem is therefore focused on terrain-dependent execution rather than relearning the nominal gait or whole-body reference-following behavior.

The adapter observes proprioception, the nominal base action, its own previous correction and a local terrain observation. It does not receive the reference: what the upstream source asks for reaches the adapter only through that base action, leaving the upstream reference interface unchanged.

At deployment a gate (Fig. 2) selects whether the correction is applied; disabling it leaves π ’s output unchanged,

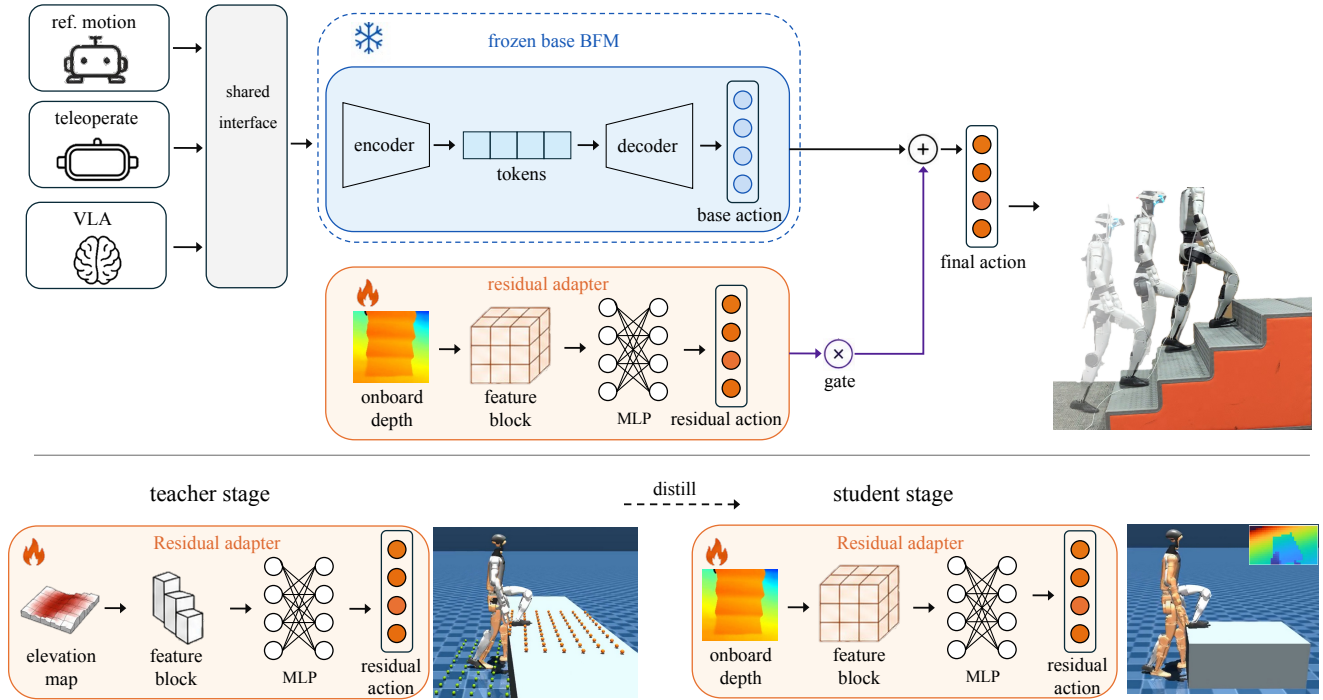


Fig. 2. GRAFT. Top: the frozen BFM maps the unmodified reference, from a recorded clip, a teleoperator or a VLA, to a nominal whole-body action, and the adapter reads onboard depth and adds a correction on the lower-body channels only. Bottom: the adapter is first trained against a privileged elevation map and then distilled into a student that reads onboard depth. The node marked *gate* closes the residual pathway at deployment, after which the executed target is the frozen BFM’s own output. The adapter’s proprioception, base-action and residual-history inputs are omitted for clarity.

and the gate is set rather than learned. The lower-body mask prevents direct writes to the upper-body targets. It is not an assumption that legs and arms are dynamically independent — the effect of a leg correction on upper-body tracking is measured in Sec. IV-D.

A. Terrain rewards

The corrections terrain requires are about where and when a foot arrives, so the reward acts on the feet. Two terms carry most of the terrain behavior and a third suppresses the residual where the ground is flat. They are optimized together with the task reward r_{task} of the underlying course, which rewards commanded forward progress and staying upright, and with the regularizers r_{reg} of the base task,

$$r_t = r_{\text{task},t} + r_{\text{reg},t} + \lambda_{\text{clr}} r_{\text{clr},t} + \lambda_{\text{fp}} r_{\text{fp},t} + \lambda_{\text{flat}} r_{\text{flat},t}, \quad (2)$$

evaluated along rollouts of the combined controller. The three terrain terms are non-positive penalties with non-negative weights. The adaptation-specific recipe is thus limited to foot clearance, foot placement, and flat-ground residual suppression, while the retained base regularizers preserve the released controller’s motion and stability conventions. Forward progress and termination are scored whether or not the adapter acts, which is what gives it an incentive to intervene once the terrain makes the base action fail; the terrain penalties then shape the clearance and placement that traversal requires.

Swing clearance. Clearance is scored over the swing rather than at a single instant. At each swing step the terrain observation gives the surface height under the foot, and any deficit against that surface plus a margin δ is accumulated over the swing and charged at the corresponding touchdown,

$$r_{\text{clr}}^{(i,k)} = - \sum_{\tau \in \mathcal{S}_{i,k}} [h_{i,\tau}^{\text{surf}} + \delta - h_{i,\tau}^{\text{foot}}]_+. \quad (3)$$

Here $\mathcal{S}_{i,k}$ is the set of control steps in swing k of foot i , $[\cdot]_+$ is the positive part, and foot and terrain heights share one vertical reference. Scoring the whole swing, rather than the instant of contact, is what encourages the foot to lift ahead of a step instead of into it.

Edges and overhang. The sole of foot i is sampled at a set of footprint points $\mathcal{P}_i$. Only a foot that carries load is charged: $c_{i,t}$ indicates contact, and each of that foot’s sample points is penalized when it lies on a visible height discontinuity, indicated by $\mathbf{1}_{\text{edge}}$, and when it overhangs the surface below it beyond a threshold, indicated by $\mathbf{1}_{\text{over}}$,

$$r_{\text{fp},t} = - \sum_i \frac{c_{i,t}}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} (\mathbf{1}_{\text{edge}}(p,t) + \mathbf{1}_{\text{over}}(p,t)). \quad (4)$$

Flat-ground suppression. A flatness weight is formed from the same terrain observation over two regions, one under the feet and one ahead of the robot,

$$g_t = \exp\left(-\frac{\text{Var}(h_t^u) + \text{Var}(h_t^f) + (\bar{h}_t^u - \bar{h}_t^f)^2}{\sigma_h^2}\right), \quad (5)$$

where h_t^u and h_t^f are the elevation samples under the feet and ahead of the robot and σ_h is a shared length scale. The two variance terms measure flatness within each region; the mean-height difference separates locally flat regions that sit at different elevations, such as the ground and a stair tread. The residual is then suppressed in proportion to this weight, through a bounded surrogate of the weighted residual magnitude,

$$\rho_t = \frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} |\Delta a_{t,j}|, \quad r_{\text{flat},t} = -g_t(1 - e^{-\rho_t/\sigma_r}), \quad (6)$$

with ρ_t the mean absolute correction over the writable lower-body channels $\mathcal{J}$ and σ_r its scale. The penalty is bounded between $-g_t$ and zero, which keeps a large residual early in training from dominating the terrain terms. This term shapes behavior during training only; the deployed adapter does not depend on it.

B. Teacher-student training

Stage I: privileged teacher. The adapter is trained with PPO while the base is frozen, reading a privileged simulated elevation sample as its terrain observation. Episodes terminate when the robot completes the course or meets a failure condition.

Stage II: onboard-depth student. We freeze the teacher and train a student to predict its correction from a rendered onboard-like depth image. The student acts in the environment and is supervised by the teacher under a squared-error loss. At deployment the rendered depth is replaced by the onboard stream, while the residual-action output stays the same. We evaluate whether the student retains the teacher’s near-zero flat-ground correction without enforcing it during distillation.

C. Training terrain and references

Training uses procedurally generated terrain with a difficulty curriculum. The reference motions are drawn from sources whose upper-body motion varies in style and origin, which is what leads the adapter to compensate for the balance effects of a commanded arm motion instead of rewriting it. Domain randomization covers the usual dynamics and observation noise. Instantiating the same procedure on another BFM requires matching the interface it writes against: joint ordering, action units and default posture. What is shared across bases is the residual placement and the objective family; the interface mapping and the base-specific regularizers are adapted, and one adapter is trained per base, so what is reused is the procedure and not the weights.

IV. EXPERIMENTS

We evaluate whether a frozen behavior foundation model can acquire terrain competence and what this adaptation costs:

Q1 Does a frozen whole-body motion prior simplify multi-terrain adaptation, and where should the correction be injected — on the executed action, on the base’s latent command, or in its weights?

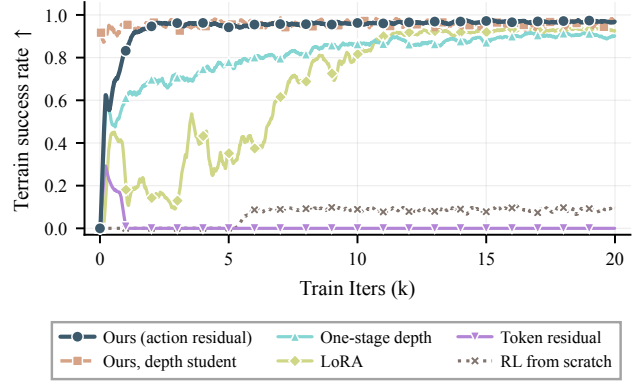


Fig. 3. Training success on the same frozen SONIC base. Success is averaged over the four obstacle terrains; Table II reports the corresponding hard-terrain evaluation.

TABLE II

COMPARISON OF TERRAIN ADAPTATION APPROACHES ON HARD TERRAIN. EACH SUCCESS CELL IS COMPUTED OVER TEN SESSIONS ($n = 10$ PER CELL). FLAT RESIDUAL IS REPORTED IN RAW ACTION UNITS.

Adaptation	Success				Flat resid.
	Up stairs	Down stairs	Gap	Platform	
Action residual (ours)	1.00	1.00	1.00	0.90	0.0520 ± 0.0019
Depth student (ours)	1.00	0.80	1.00	1.00	0.4491 ± 0.3071
LoRA (decoder)	1.00	0.90	1.00	0.60	0.6929 ± 0.7173
One-stage depth	0.40	0.90	0.70	0.20	1.5930 ± 0.0212
Motion-token residual	0.00	0.00	0.00	0.00	4.7025 ± 1.2187
RL from scratch	0.00	0.10	0.00	0.00	—

Q2 Does the same adaptation procedure apply to different frozen bases?

Q3 How does the adaptation affect flat-ground actions and upper-body tracking, and can it be disabled during a run?

Q4 Does the adapted system work on hardware?

A. Setup

All simulated experiments use a Unitree G1 model. Adapters are trained in Isaac Lab with 4096 parallel environments for 20k iterations at 50 Hz on procedurally generated courses with a difficulty curriculum. We evaluate in MuJoCo on fixed terrains: up stairs, down stairs, a gap, a platform, and flat ground. The reported obstacle results use the hard terrain setting, corresponding to the maximum difficulty used during training. Reference motions are drawn from the SONIC training dataset and are not modified for terrain. Success is the fraction of ten evaluation sessions per cell that reach the terrain end without falling and remain there. We also report the mean L_2 norm of the applied lower-body correction in raw action units.

B. Comparing terrain adaptation approaches

We compare six terrain adaptation approaches with the same simulator, reward specification, and iteration budget. The action residual adds a lower-body correction to the

executed target; the depth student is its two-stage, onboard-depth version. The motion-token residual perturbs the latent command, LoRA adapts decoder weights, one-stage depth skips the privileged teacher, and RL from scratch uses no base. LoRA represents the fine-tuned terrain-aware BFM route [5], [6].

(Q1) A residual on the executed action adapts fastest. As shown in Fig. 3, the action residual reaches 0.947 within 2k iterations and 0.969 by 20k, showing a clear optimization advantage under the same training setup. LoRA reaches 0.925 by 20k and achieves comparable hard-terrain success, but changes flat-ground actions more: 0.6929 ± 0.7173 versus 0.0520 ± 0.0019 for the action residual (Table II). Updating decoder weights therefore reaches similar terrain performance at the cost of larger and more variable nominal-motion changes.

Perturbing the latent command did not produce foothold behavior. The motion-token residual fails on all four hard obstacle terrains and also disrupts flat-ground behavior, producing the largest flat residual at 4.7025 ± 1.2187 . Because the token is a compressed representation of the reference motion and its decoder is trained for flat-ground tracking, perturbing the token changes the decoded motion without providing a mechanism for terrain interaction or foothold adjustment.

A frozen motion prior simplifies terrain adaptation. The BFM already supplies locomotion, reference tracking, and whole-body coordination, so the adapter only has to learn terrain-dependent execution. Under the same evaluation, the action residual achieves 39/40 hard-terrain successes, whereas RL from scratch achieves only 1/40. This gap supports the view that the frozen prior changes the optimization problem from learning locomotion and terrain interaction jointly to learning a focused execution correction.

The depth student outperforms direct depth training. The depth student reaches 0.963 in training and achieves 38/40 successes on hard terrain, although its failures are distributed differently across terrain types. One-stage depth reaches 0.901 by 20k iterations but obtains only 22/40 hard-terrain successes and produces a larger flat residual of 1.5930 ± 0.0212 , compared with 0.4491 ± 0.3071 for the distilled student. The result suggests that privileged-to-depth distillation improves both optimization and execution stability.

C. Transfer across frozen bases

(Q2) The same adaptation procedure applies to different frozen bases. Each base receives a separate adapter with the same residual placement and objective family; only interface mappings and base-specific regularizers are adapted. Fig. 4 and Table III show the resulting training and hard-terrain success.

The training curves show that the same residual recipe converges on all four bases, but at different rates and final levels. SONIC converges fastest and reaches the highest training success, Humanoid-GPT follows a similar trend, while HoloMotion and especially ScaleBFM converge more

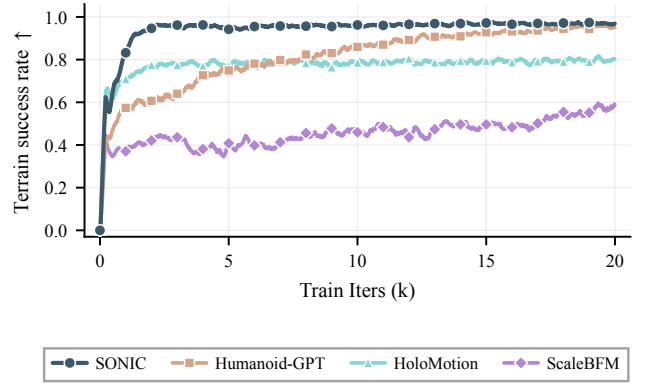


Fig. 4. Training success for the same adaptation procedure on four frozen bases. Each base has its own adapter; the curves are not a ranking of the trackers. Table III reports the corresponding hard-terrain evaluation.

TABLE III

CROSS-BASE TERRAIN ADAPTATION ON HARD TERRAIN. EACH SUCCESS CELL REPORTS BASE/GRAFT PERFORMANCE OVER TEN SESSIONS ($n = 10$ PER CELL); FLAT RESIDUAL IS REPORTED FOR GRAFT IN RAW ACTION UNITS. H-GPT AND HOLOMOT. DENOTE HUMANOID-GPT AND HOLOMOTION, RESPECTIVELY. ALL TRACKERS ARE DRIVEN BY REFERENCE MOTIONS FROM THE SONIC TRAINING DATASET.

Base	Success (base/GRAFT)				Flat resid. GRAFT
	Up stairs	Down stairs	Gap	Platform	
SONIC	0.00/1.00	0.30/1.00	0.00/1.00	0.00/0.90	0.0520 ± 0.0019
H-GPT	0.00/1.00	0.20/0.90	0.10/1.00	0.00/0.50	0.0781 ± 0.0021
HoloMot.	0.00/1.00	0.80/0.90	0.00/1.00	0.00/1.00	0.0648 ± 0.0009
ScaleBFM	0.00/0.30	0.10/0.00	0.10/0.90	0.00/0.80	0.0928 ± 0.0024

slowly or plateau lower. This confirms that the adaptation procedure can be instantiated across bases, while the optimization difficulty remains base-dependent.

The hard-terrain evaluation is consistent with this base dependence. All trackers are driven by reference motions from the SONIC training dataset. Without GRAFT, the four bases obtain 3/40, 3/40, 8/40, and 2/40 successes, respectively. With GRAFT, these totals increase to 39/40, 34/40, 39/40, and 20/40. The gains are substantial overall but are not uniform across terrain cells: ScaleBFM, for example, decreases from 1/10 to 0/10 on down stairs, while its main gains occur on the gap and platform.

D. Adaptation cost and removal

(Q3) The correction follows terrain geometry. Fig. 5 shows that, across the composite course, the mean residual is 1.529 on obstacle segments versus 0.408 on flat segments; after removing approach and recovery margins, the values are 1.492 versus 0.184. It peaks at 3.488 on the gap. The depth student shows the same terrain-dependent pattern, while the four-base hard-terrain flat residuals range from 0.0520 ± 0.0019 to 0.0928 ± 0.0024 (Table III).

Upper-body tracking is largely preserved. Arm error remains comparable between the adapter-active and adapter-disabled intervals. Waist error rises from 1.18° on flat ground

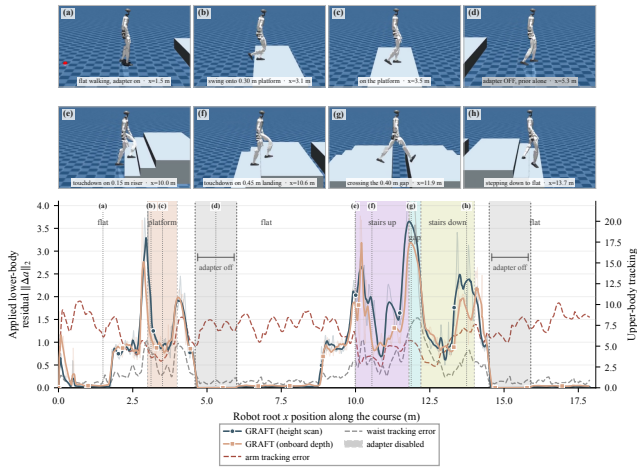


Fig. 5. Composite-course simulation trace. The residual follows terrain geometry while arm and waist tracking remain visible; grey intervals indicate simulation-only residual disabling.

to 3.17° on obstacles, reflecting torso compensation for larger leg motion.

In the composite-course simulation, the module can be disabled mid-run. On flat ground, disabling the gate set the correction to zero while walking continued; re-enabling restored the prior residual level without a transient. This directly demonstrates the switchable base pathway.

E. Hardware

(Q4) The adapted system works on the robot. With onboard depth, the adapted SONIC crosses a 0.30 m platform and a 0.10 m step using the same references, whereas the frozen base treads in place before both (Fig. 6).

The same policy also handles a 0.30 m platform, 0.10 m and 0.17 m steps, and a push-cart scene (Fig. 1), using the released base and unmodified references. The push-cart scene is outside the procedural terrain distribution used during training and serves as an additional qualitative deployment test.

V. CONCLUSION

We presented GRAFT as a separation between a reusable whole-body motion prior and a lightweight terrain-execution adapter. The frozen BFM provides reference tracking, nominal locomotion, and whole-body coordination, allowing the adapter to acquire terrain-dependent foot placement with a compact task-and-foot reward recipe. Across four frozen trackers, this decomposition enabled traversal of stairs, gaps, and platforms that the original policies largely failed to cross, while keeping corrections small on flat ground. In a composite-course simulation, the SONIC adapter could also be disabled and re-enabled during flat-ground walking. Distilling this adapter from privileged terrain observations to onboard depth enabled real-world deployment on a Unitree G1 with unmodified references.

These results show that a frozen whole-body prior can turn multi-terrain learning from a complete locomotion-learning

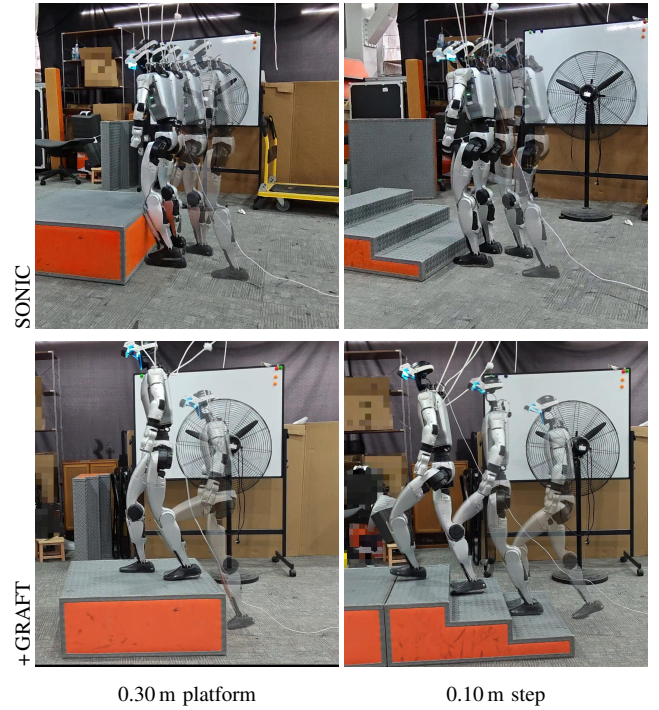


Fig. 6. Hardware comparison with the same base and references. GRAFT raises the foot and crosses both obstacles; the frozen base alone treads in place.

problem into a lightweight execution-adaptation problem. The current evaluation uses recorded motion references and focuses on terrain adaptations achievable through leg corrections. A next step is to evaluate tracking under live teleoperation and VLA commands, particularly when changes in upper-body motion require coordinated adjustments of balance and foot placement.

REFERENCES

- [1] Z. Luo, Y. Yuan, T. Wang, C. Li, F. Castañeda, S. Chen, Z.-A. Cao, J. Li, D. Minor, Q. Ben *et al.*, “SONIC: Supersizing motion tracking for natural humanoid whole-body control,” *Science Robotics*, vol. 11, no. 117, p. eaed4592, 2026.
- [2] Z. Qi, X. Chen, D. Liu, C. Lin *et al.*, “Humanoid-GPT: Scaling data and structure for zero-shot motion tracking,” *arXiv preprint arXiv:2606.03985*, 2026.
- [3] M. Chen, K. Wang, B. Zhang, X. Ma *et al.*, “HoloMotion-1 technical report,” *arXiv preprint arXiv:2605.15336*, 2026.
- [4] Z. Cheng, T. Tang, J. Lan, X. Chen, Y. Gong, Z. Liu, C. Wu, Y. Mao, Z. Deng, M. Ma, H. Xi, Y. Liu, Y. Wu, X. Wang, Y. Wang, Y. Ye, G. Huang, X. Jin, Z. Zhu, and J. Lu, “GigaBrain-WBC-0.5: A behavior world model for robust whole-body control with environment interaction,” *arXiv preprint arXiv:2608.18234*, 2026.
- [5] T. Xie, H. Zhang, J. Park, Z. Wang, B. Wen, J. Li, X. Li, Q. Ben, H. Weng, Y. Ye, D. Minor, T. Wang, C. Jiang, S. Fidler, J. Kautz, L. Fan, Y. Zhu, Z. Luo, U. Iqbal, and Y. Yuan, “GRAIL: Generating humanoid loco-manipulation from 3d assets and video priors,” *arXiv preprint arXiv:2606.05160*, 2026.
- [6] Z. Wang, Y. Li, T. Ma, Q. Zhang, Y. Fan, H. Xu, S. Yang, and J. Liang, “Perceptive behavior foundation model: Adapting human motion priors to robot-centric terrain,” *arXiv preprint arXiv:2606.08059*, 2026.
- [7] Z. Sun, B.-S. Huang, Y. Peng, X. Li, J. Ma, Y. Sun, Z. Li, H. Jiang, B. Gao, Z. Bing, X. Wang, and A. Knoll, “MOSAIC: Bridging the sim-to-real gap in generalist humanoid motion tracking and teleoperation with rapid residual adaptation,” *arXiv preprint arXiv:2602.08594*, 2026.

- [8] J. Sun, G. Han, P. Sun, W. Zhao, J. Cao, J. Wang, Q. Zhang, and Y. Guo, "DPL: Depth-only perceptive humanoid locomotion via realistic depth synthesis and cross-attention terrain reconstruction," *IEEE Robotics and Automation Letters*, vol. 11, no. 8, pp. 9407–9414, 2026.
- [9] S. Zhao, Y. Ze, Y. Wang, C. K. Liu, P. Abbeel, G. Shi, and R. Duan, "ResMimic: From general motion tracking to humanoid whole-body loco-manipulation via residual learning," *arXiv preprint arXiv:2510.05070*, 2025.
- [10] N. Jayasinghe, D. Gontero, S. T. Brown, V. K. Sangwan, M. C. Hersam, and A. R. Trivedi, "Cerebellar-inspired residual control for fault recovery: From inference-time adaptation to structural consolidation," *arXiv preprint arXiv:2602.07227*, 2026.
- [11] Z. Chen, M. Ji, X. Cheng, X. Peng, X. B. Peng, and X. Wang, "GMT: General motion tracking for humanoid whole-body control," *arXiv preprint arXiv:2506.14770*, 2025.
- [12] Y. Li, Z. Luo, T. Zhang, C. Dai, A. Kanervisto, A. Tirinzoni, H. Weng, K. Kitani, M. Guzek, A. Touati, A. Lazaric, M. Pirotta, and G. Shi, "BFM-Zero: A promptable behavioral foundation model for humanoid control using unsupervised reinforcement learning," in *Int. Conf. Learning Representations (ICLR)*, 2026.
- [13] W. Zeng, K. Yin, X. Niu, S. Lu, W. Zhong, J. Chen, F. Jia, X. Chen, Z. Wang, F. Xu, M. Zhou, K. Li, W. Zhang, H. Wang, L. Yi, D. Lin, J. Pang, and J. Wang, "Scaling behavior foundation model for humanoid robots," *arXiv preprint arXiv:2607.15163*, 2026.
- [14] Q. Liao, T. E. Truong, X. Huang, Y. Gao, G. Tevet, K. Sreenath, and C. K. Liu, "BeyondMimic: From motion tracking to versatile humanoid control via guided diffusion," *Science Robotics*, vol. 11, no. 117, p. eadx8924, 2026.
- [15] Y. Ze, Z. Chen, J. P. Araújo, Z.-A. Cao, X. B. Peng, J. Wu, and C. K. Liu, "TWIST: Teleoperated whole-body imitation system," *arXiv preprint arXiv:2505.02833*, 2025.
- [16] Z. Zhang, C. Li, T. Miki, and M. Hutter, "Motion priors reimaged: Adapting flat-terrain skills for complex quadruped mobility," in *Proceedings of the 9th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, vol. 305, 2025, pp. 3762–3777.
- [17] W. Sun, B. Cao, L. Chen, Y. Su, Y. Liu, Z. Xie *et al.*, "Learning perceptive humanoid locomotion over challenging terrain," in *IEEE/RSJ Int. Conf. Intelligent Robots and Systems (IROS)*, 2025, pp. 6571–6578.
- [18] Z. Wu, X. Huang, L. Yang, Y. Zhang, X. Chen, P. Abbeel, R. Duan, A. Kanazawa, C. Sferrazza, G. Shi, and C. K. Liu, "Perceptive humanoid parkour: Chaining dynamic human skills via motion matching," in *Robotics: Science and Systems XXII*, 2026.
- [19] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *IEEE Int. Conf. Robotics and Automation (ICRA)*, 2024, pp. 11 443–11 450.
- [20] R. Yu, Y. Wang, Y. Hao, J. Wu, and Q. Zhu, "SSR: Scaling surefooted and symmetric humanoid traversal to the open world," *arXiv preprint arXiv:2605.30770*, 2026.
- [21] Y. Zhang, Y. Seo, J. Chen, Y. Yuan, K. Sreenath, P. Abbeel, C. Sferrazza, K. Liu, R. Duan, and G. Shi, "RPL: Learning robust humanoid perceptive locomotion on challenging terrains," *arXiv preprint arXiv:2602.03002*, 2026.
- [22] H. Wang, Z. Wang, J. Ren, Q. Ben, T. Huang, W. Zhang, and J. Pang, "BeamDojo: Learning agile humanoid locomotion on sparse footholds," in *Robotics: Science and Systems XXI*, 2025.
- [23] L. Wang, K. Yao, Y. Liu, W. Qin, J. Wu, Z. Sun, and Q. Zhu, "PUMA: Perception-driven unified foothold prior for mobility augmented quadruped parkour," *IEEE Robotics and Automation Letters*, vol. 11, no. 9, pp. 10 290–10 297, 2026.
- [24] Z. Cheng, H. Wei, H. Yin, X. Xu, B. Yu, J. Zhou, and J. Lu, "CMP: Robust whole-body tracking for loco-manipulation via competence manifold projection," *arXiv preprint arXiv:2604.07457*, 2026.
- [25] S. V. Lyngset, T. V. Raanaas, G. Sveipe, E. M. Nilsen, J. Torresen, K. O. Ellefsen, and T. Lømo, "Memory-efficient policy libraries with low-rank adaptation in reinforcement learning," *arXiv preprint arXiv:2606.25700*, 2026.
- [26] E. Daneshmand, S. Omar, G. Berseth, M. Khadiv, and H.-C. Lin, "SLOWRL: Safe low-rank adaptation reinforcement learning for locomotion," *arXiv preprint arXiv:2603.17092*, 2026.
- [27] S. Park, E. Lee, K. B. Lee, and S.-H. Lee, "MaskAdapt: Learning flexible motion adaptation via mask-invariant prior for physics-based characters," in *IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2026, pp. 2285–2294.
- [28] N. Jayasinghe, D. Gontero, F. Migliarba, S. T. Brown, V. K. Sangwan, M. C. Hersam, and A. R. Trivedi, "Residual control for fast recovery from dynamics shifts," *arXiv preprint arXiv:2603.07775*, 2026.