Learning Robust Perceptive Locomotion for Quadrupedal Robots in the Wild

Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, Marco Hutter

Science Robotics, 2022 · ETH Zurich · KAIST · Intel

MOAI Seminar

발표자 : 진동환

Introduction

Why Physical AI?

*"Multimodal LLMs can see and reason — why not let them **move the robot directly**?"*

① Speed

The locomotion policy acts at **50 Hz** (every 20 ms). LLM inference takes seconds — too slow for a robot that is falling.

② Data

The internet has sentences like "take stairs carefully", but almost no **contact forces or joint torques**. Riding a bike cannot be learned from a book.

③ Embodiment

Output is not words but **continuous joint targets** in a closed loop, and every body has different dynamics.

- **Division of labor**: LLM / VLA = cerebrum (*where* to go) · locomotion controller = cerebellum (*how* to step).

How Physical AI Learns: Current Trends

Paradigm	How it learns	Typical task	Examples
Sim RL + Sim-to-Real	Trial and error in massively parallel simulation, then transfer to the real robot	Locomotion (this paper)	ANYmal line of work, Isaac Lab, MuJoCo Playground
Imitation learning	Copy demonstrations from human teleoperation	Manipulation	ALOHA, Diffusion Policy
Foundation models (VLA)	Internet-scale pretraining + robot data; language → action	General-purpose manipulation	RT-2, π0, Gemini Robotics, GR00T

- Infrastructure underneath all three: **GPU-parallel simulators** (revisited at the end; MuJoCo playground)

What is Locomotion?

- **Locomotion**: the ability of a robot to **move its whole body** from one place to another.
- Legs reach where wheels cannot: stairs, steep slopes, gaps, rubble.
- Why legged locomotion is hard:
  - **Hybrid dynamics**: contacts make and break every step
  - **Unknown terrain**: never perfectly known in advance
  - **Balance**: must be kept at every instant



Fig. 1 — Robust locomotion in the wild

Locomotion Through the Eras

Era	Approach	Examples	Limitation
1980s–2000s	Physics models + hand-written rules	Raibert hopper, ASIMO (ZMP), BigDog	Fragile outside the rules
2010s	Optimization / MPC, map-based footstep planning	MIT Cheetah, foothold optimization	Assumes an accurate map
2019–2021	Simulation RL + sim-to-real	Hwangbo 2019, Lee 2020 (blind), RMA 2021	Mostly blind (no vision)
2022	Vision + body sense, fused by learning	This paper (Miki et al.)	Today's topic
2023–	Parkour, raw-vision end-to-end, humanoids	ANYmal Parkour, Extreme Parkour, …	Covered at the end

- This paper marks the turn from **blind** RL controllers to **perceptive** ones.

The Dilemma Before 2022

Exteroception (vision)

Pro: sees terrain **before contact** → plans the gait ahead → **fast**

Con: fooled by snow, grass, water, reflections → **falls**

Proprioception only (blind)

Pro: never misled by sensors → **robust**

Con: must physically feel the terrain first → **slow**

This paper: learn when to trust each sense → fast and robust

Default controller of Team CERBERUS, **winner of the DARPA Subterranean Challenge Final (2021)**: 4 ANYmals explored 1,700+ m without a single fall.

Why RMA · Perceptive Locomotion · MuJoCo Playground?

*"The robot cannot see the true state of the world — so **teach it with the simulator's answers**, and let it **infer from past experience** in the real world."*

RMA (2021)

Adapt by feeling: infer hidden environment factors from proprioceptive history

Perceptive Locomotion (2022)

See + feel: fuse vision and body sense, and learn how much to trust vision

MuJoCo Playground (2025)

The factory: train such policies in minutes on a single GPU

• Common thread: the real problem is a **POMDP**; simulation provides **privileged information** (friction, contact forces, exact terrain) that the real robot never sees.

RMA Recap & Connection

• **RMA**: ① train a base policy with an encoded environment vector *z* (privileged) → ② train an adaptation module that estimates *z* from recent states and actions only.

Aspect	RMA (2021)	Perceptive Locomotion (2022)
Robot	Unitree A1	ANYmal C
Input senses	Proprioception only	Proprioception + exteroception (terrain heights)
What is estimated	Environment latent z	Terrain + contacts, friction, external force (belief state)
History encoder	1D CNN over a recent window	GRU accumulating the whole history
Stage-2 objective	Regress estimated z to true z	Imitate teacher action + reconstruct true world state
Policy body in stage 2	Frozen	Trained jointly (initialized from teacher)

MuJoCo Playground: Where It Fits

- Open-source robot-learning framework built on **MJX** (MuJoCo on GPU via JAX) — Google DeepMind, UC Berkeley et al., 2025
- *"pip install playground"* → train policies **in minutes on a single GPU**
- Quadrupeds, humanoids, dexterous hands, arms; zero-shot sim-to-real from both **state and pixel** inputs (batch renderer)

RMA	**This paper**	**MuJoCo Playground**
what to learn: adapting by feel	what to learn: seeing + feeling	how fast, and for whom

Method : Perceptive Locomotion

Problem: Exteroception is Unreliable in the Wild



Fig. 3 — Perception challenges in the field

- **Missing** map information
  reflective ground (B), deep snow (C), occlusion (G)

- **Misleading** map information
  overhanging branches (D), soft vegetation (E), pose-estimation drift (F)

- Prior work: **trust the map** blindly (MIT Cheetah), or **ignore vision** altogether (RMA)

→ Need a controller that knows **when to trust vision**

Key Contributions

- **Fusing exteroception and proprioception** → a controller **faster and more robust** than either sense alone

  Like relying more on touch when the lights go out

- **Attention-based recurrent belief encoder** → learns **when and how much to trust each sense**, without hand-crafted heuristics

  Trained end-to-end in simulation

- **Extensive validation in the wild**: alpine, forest, underground, urban, over multiple seasons — **zero falls**

  Including an hour-long hike in the Alps at human hiking pace

Method Overview

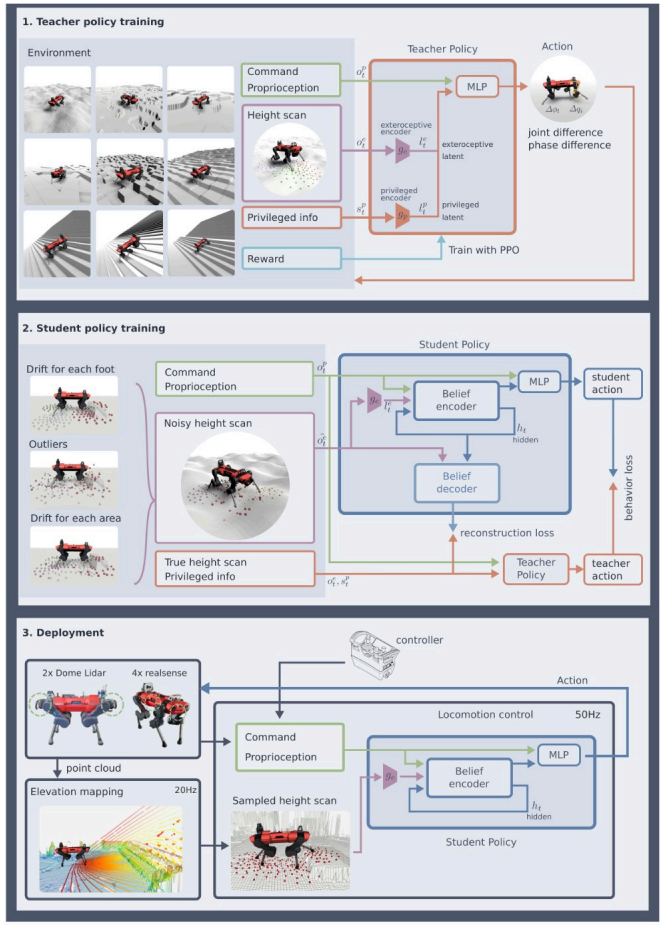


Fig. 5 — Training and deployment

① Teacher policy training

RL with full **privileged information**: noiseless terrain, contacts, friction, disturbances

② Student policy training (supervised)

Only onboard-available info + **noisy height scan**; imitate the teacher's action and reconstruct the true world state

③ Deployment (zero-shot)

ANYmal C with a robot-centric elevation map, no fine-tuning

Simulator: RaiSim · 1,000 parallel envs (teacher), 300 (student) · policy at 50 Hz

Key Terms: Policy, World State, Belief

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Policy

reward on t-th step

total sum of future reward

γ (=0.996)

↳ discount factor: 지금의 reward가 먼 reward보다 중요

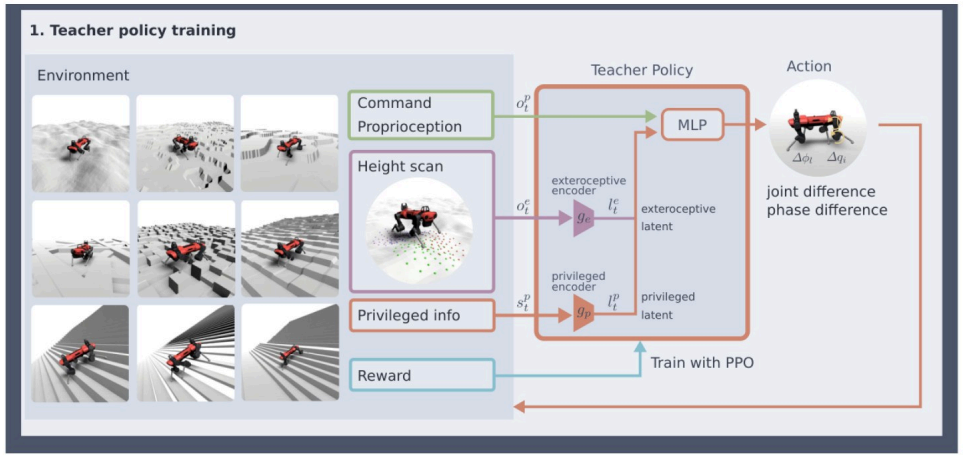
cf) $\frac{1}{1-\gamma} \approx 250$ step정도 앞까지를 의미 있게 고려함.

보통은 5초 먼저 내다보게 됨.

Term	Meaning in this paper
Policy π(a | o)	Function from observation to action; teacher and student are separate policies, each outputting joint targets
World state s_t	The true full state: exact terrain, friction, external forces — only the teacher sees it
Belief state b_t	The student's estimate of s_t from its observation history (“what I believe is there”)

- o_t ≠ s_t → **POMDP**. Recurrent RL from scratch is slow → split into **RL (teacher, easy MDP) + supervised learning (student)**.

Teacher Policy: Learning with the Answer Sheet



$$l_t^e = g_e(o_t^e)$$: exteroceptive latent

$$l_t^{priv} = g_p(s_t^p)$$: privileged latent.

$$a_t \sim \mathcal{N}\left(\pi_\theta(o_t^p,\, l_t^e,\, l_t^{priv}),\ \sigma I\right)$$

Input	Contents	Dim.
Proprio. o^p	Command, body orientation and velocity, joint states + history, CPG phase	133
Extero. o^e	52 noiseless height samples around each foot × 4 feet	208
Privileged s^p	Contacts, contact forces and normals, friction, external force/torque, airtime	50

- MLP (256-160-128), trained with **RL**
- Task: **track a random velocity command** (v_x, v_y, ω)

Action Space: CPG-Inspired Residuals

$$q_i^{target} = \mathrm{IK}\big(\mathbf{p}(\phi_l + \Delta\phi_l + \Delta\phi_0)\big) + \Delta q_i$$

Inverse kinetic : trajectory → joint target.

- Each leg *l* keeps a **phase variable φ_l** (a metronome) that drives a nominal foot trajectory **p(φ)**: swing-up → swing-down → stance (0.2 m foot lift)
- The policy outputs only **16 numbers**: phase offsets **Δφ_l** (4) + joint residuals **Δq_i** (12)
- Analytic IK → **12 joint position targets** (rad) → tracked by joint PD control, at 50 Hz
- Why? A built-in walking rhythm makes RL **much easier to train**

+cf) IK가 궁금하신 분들을 위해...

$D = \sqrt{x^2+z^2-d^2}$

$q_1 = \tan2(x,-z) - \tan2(d,D)$ → $C^2 = x^2+D^2$

$q_3 = \pm \arccos((C^2-L_1^2-L_2^2)/2L_1L_2)$ → $q_2 = \tan2(x,D) - \tan2(L_2\sin q_3, L_1+L_2\cos q_3)$

Reward & Curriculum

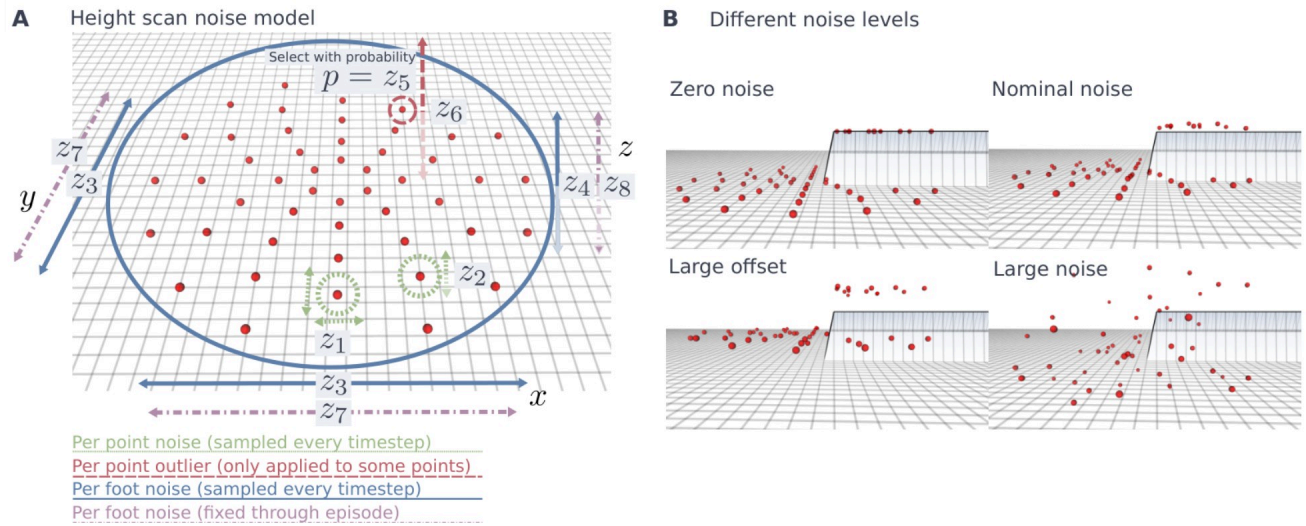
$$r_{lv} = \begin{cases} 1.0, & \text{if } v_{des} \cdot v > |v_{des}| \\ \exp\big(-(v_{des} \cdot v - |v_{des}|)^2\big), & \text{otherwise} \end{cases}$$

- **Positive**: follow the commanded linear and yaw velocity; penalize velocity orthogonal to the command
- **Shaping penalties**: body motion, foot clearance, shank/knee collision, joint velocity and acceleration, torque, foot slip, smoothness
- **Termination**: body–ground collision, large tilt, exceeding the torque limit
- **Terrain curriculum**: a particle filter keeps terrain **hard but achievable**; penalty and randomization weights ramp up with:

$$c_{k+1} = (c_k)^d, \qquad 0 < d < 1 \;\; (d = 0.98)$$

- 학습 초반→약한 penalty
- 점점 강해짐

Student Training: Learning with Noisy Eyes

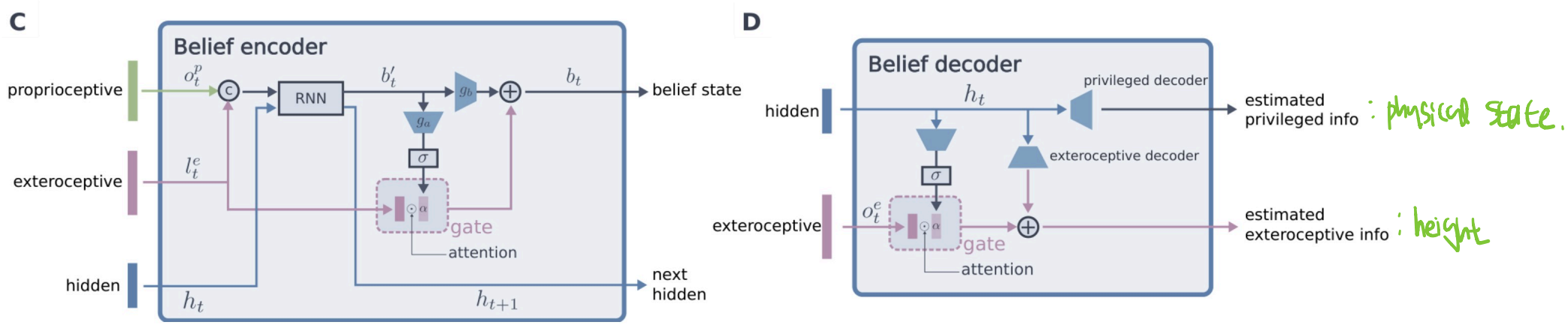


$$h_p = h(x_p, y_p) + \underbrace{\epsilon_{pz}}_{\text{per point}} + \underbrace{\epsilon_{fz}}_{\text{per foot}} + \underbrace{w_z}_{\text{per episode}} + \underbrace{\epsilon_{outlier}}_{\text{outlier}}$$

height

- Map condition picked per episode: **nominal** 60% · **large offset** 30% (pose drift, soft terrain) · **large noise** 10% (occlusion, sensor failure); lateral point shifts too, and noise grows with a curriculum
- Only the teacher sees clean data — the student is trained on **deliberately corrupted** inputs

Belief State Encoder: Gated Fusion of Two Senses



$$b'_t,\ h_{t+1} = \mathrm{GRU}\left(o^p_t,\ l^e_t,\ h_t\right)$$

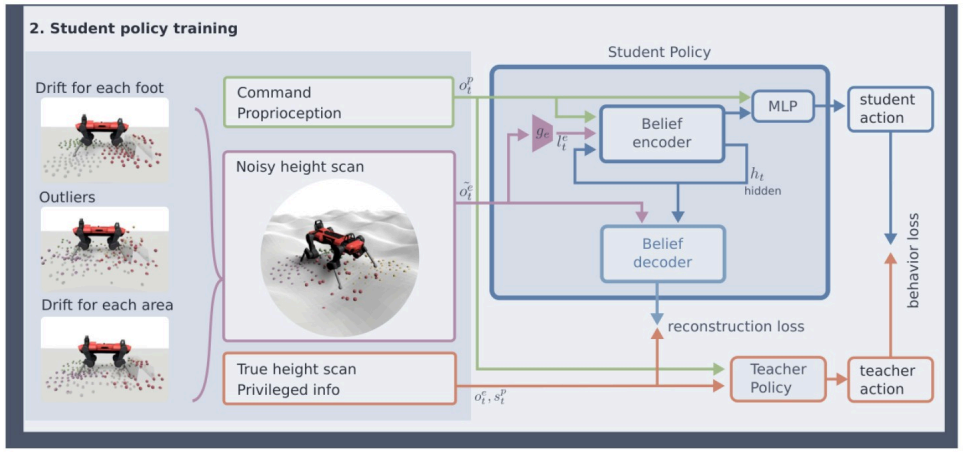
$$\alpha_t = \sigma\left(g_a(b'_t)\right)$$

$$b_t = g_b(b'_t) + l^e_t \odot \alpha_t$$

- **GRU** fuses current senses with memory → intermediate belief b′_t
- **Gate α_t ∈ (0,1)[96]**: **how much exteroception passes**, per feature
- Senses agree → gate opens (use vision); disagree → gate closes (fall back on proprioception)
- Decoder (**training only**) reconstructs clean heights + privileged info

Training Objective: Imitate + Reconstruct

$$\mathcal{L} = \underbrace{\left\| a_t^{student} - a_t^{teacher} \right\|_2^2}_{\text{behavior cloning } \mathcal{L}_{bc}} + 0.5 \underbrace{\left\| \hat{o}_t^e - o_t^e \right\|_2^2 + \left\| \hat{s}_t^p - s_t^p \right\|_2^2}_{\text{reconstruction } \mathcal{L}_{re}}$$



- **L_bc**: match the teacher's *action* in the same state
- **L_re**: rebuild the simulator's *true world state* → forces b_t to be informative
- Data from **student rollouts** (DAgger-style) → learns to recover from its own mistakes
- Encoder, decoder, MLP trained **jointly**; MLP initialized from teacher; GRU via TBPTT (10 steps)

Does the Gate Help? — Ablation

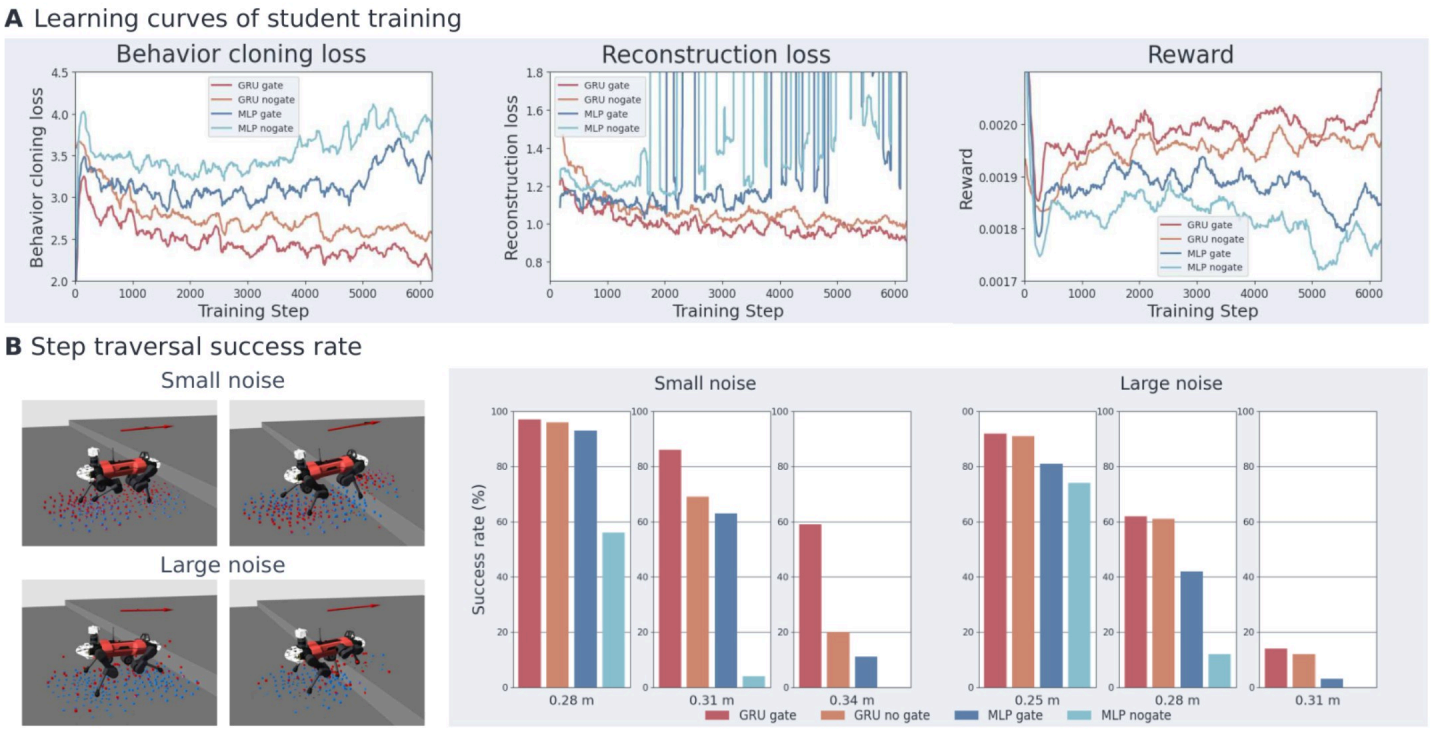


Fig. S2 — Belief encoder ablation

- **GRU > MLP**: memory is essential; MLP fails to reconstruct privileged info
- **Gate helps** both GRU and MLP — most under **small noise**
- Under large noise, gate ≈ no gate → the gate **opens when vision is reliable, closes when not**

Deployment: Zero-shot Sim-to-Real

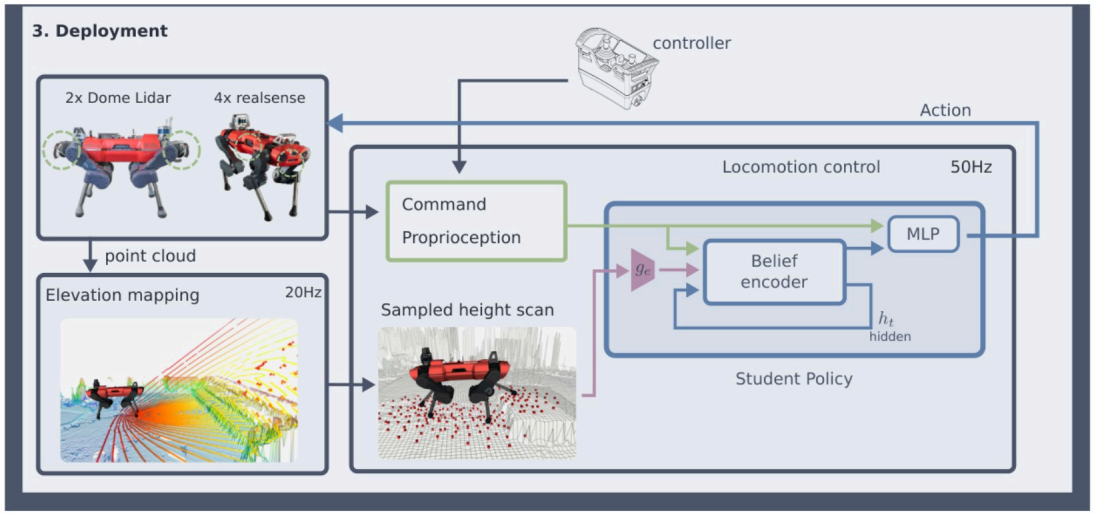


Fig. 5 — Deployment

- 2× dome LiDAR **or** 4× depth cameras → point cloud
- GPU elevation map at **20 Hz**; policy samples heights at **50 Hz**
- Missing cells → random values (already seen in training)
- Elevation map = **sensor-unrelated** layer: no retraining when swapping sensors

↳ Lidar든 RealSense든 point cloud 거쳐 elevation map로 바뀌기 때문임.

Results

A Hike in the Alps (Etzel, Switzerland)

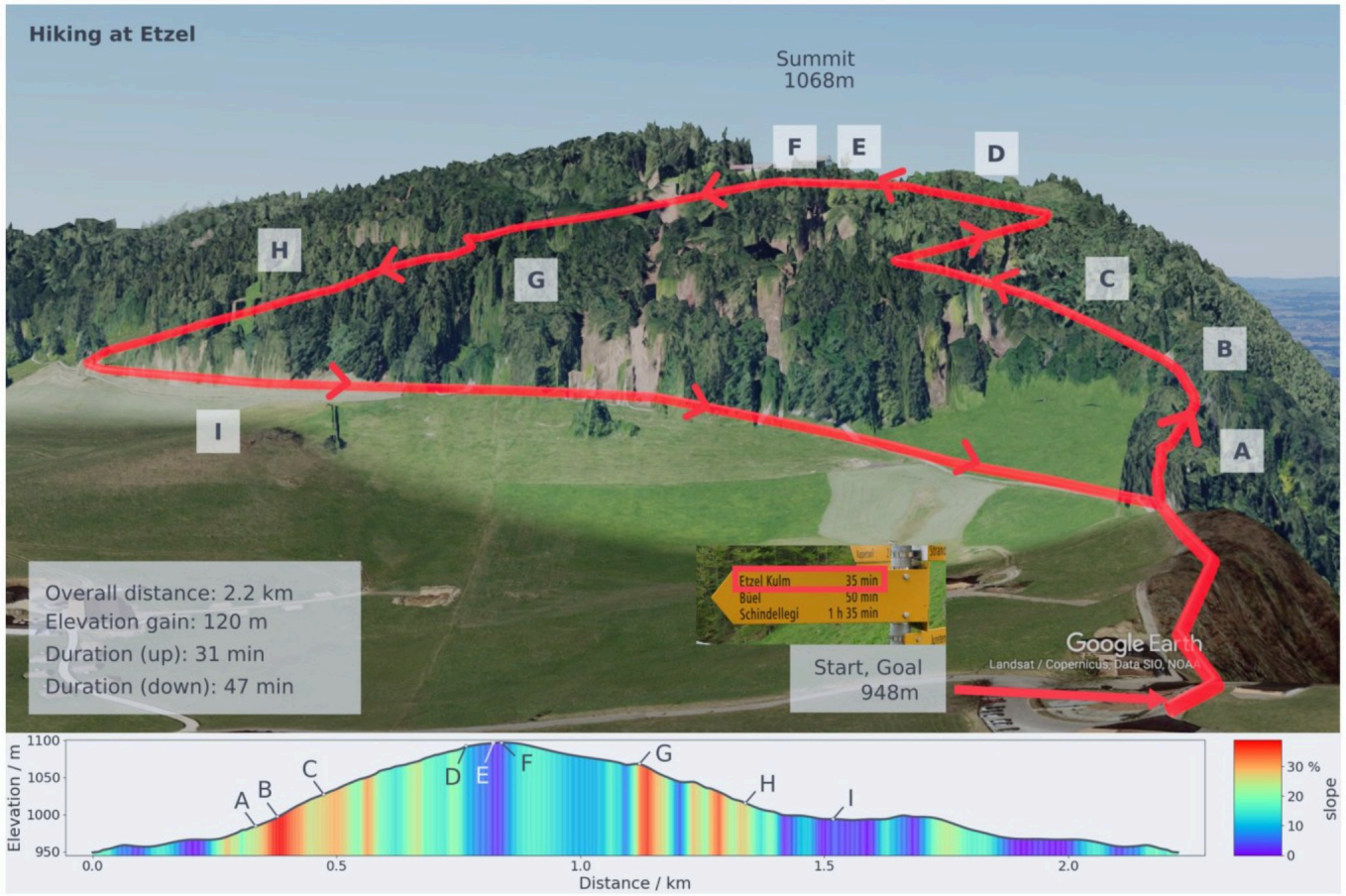


Fig. 2 — Hiking route rated “difficult”

2.2 km · +120 m

inclines up to 38%, roots, wet rock

Summit: 31 min

official sign for humans: 35 min

Full loop: 78 min

hiking planner estimate: 76 min

Zero falls

only help: reattach a shoe, swap batteries

Quantitative Comparison vs. Blind Baseline

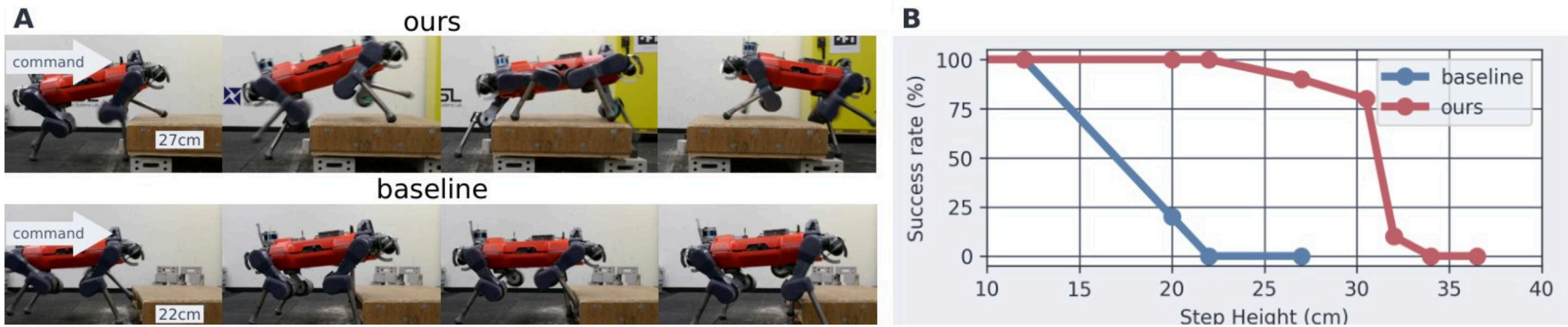
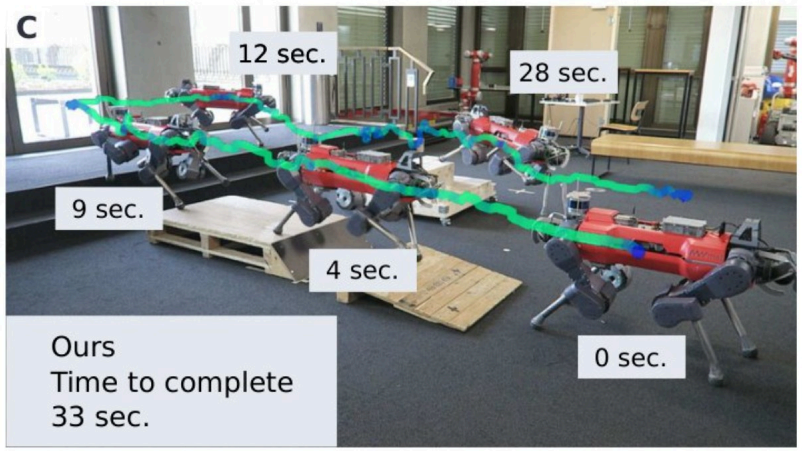
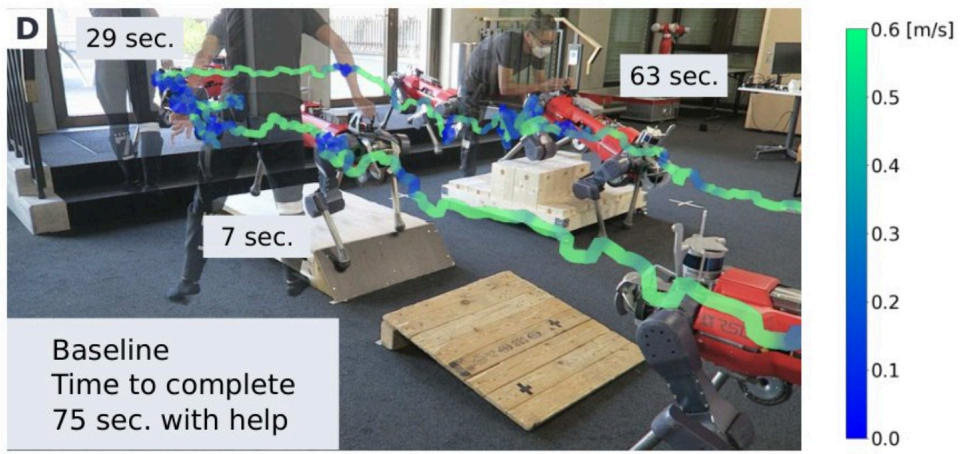


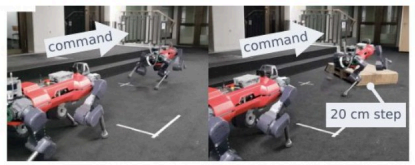
Fig. 4A–B — Step traversal (10 trials per height)

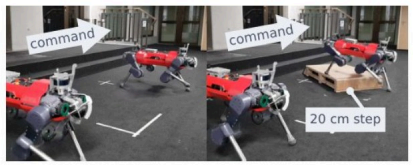
Metric	Ours	Proprioceptive baseline
Step height, reliable traversal	up to 30.5 cm	fails from 20 cm
Max speed on flat ground	1.2 m/s	0.6 m/s
Turning speed	3 rad/s (5×)	0.6 rad/s

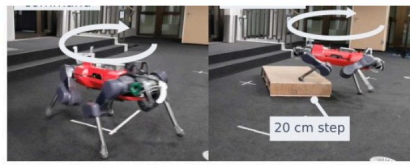
Obstacle Course & Speed

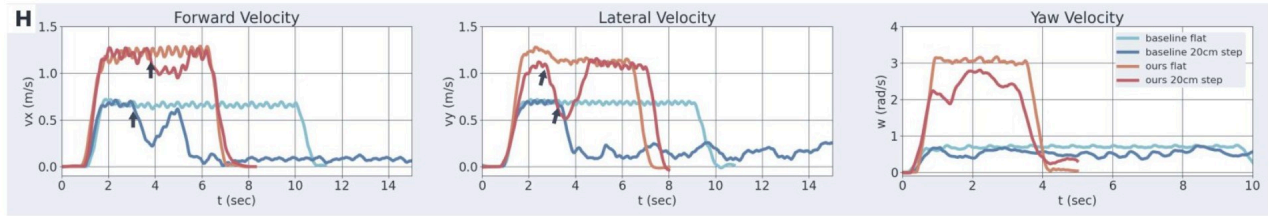






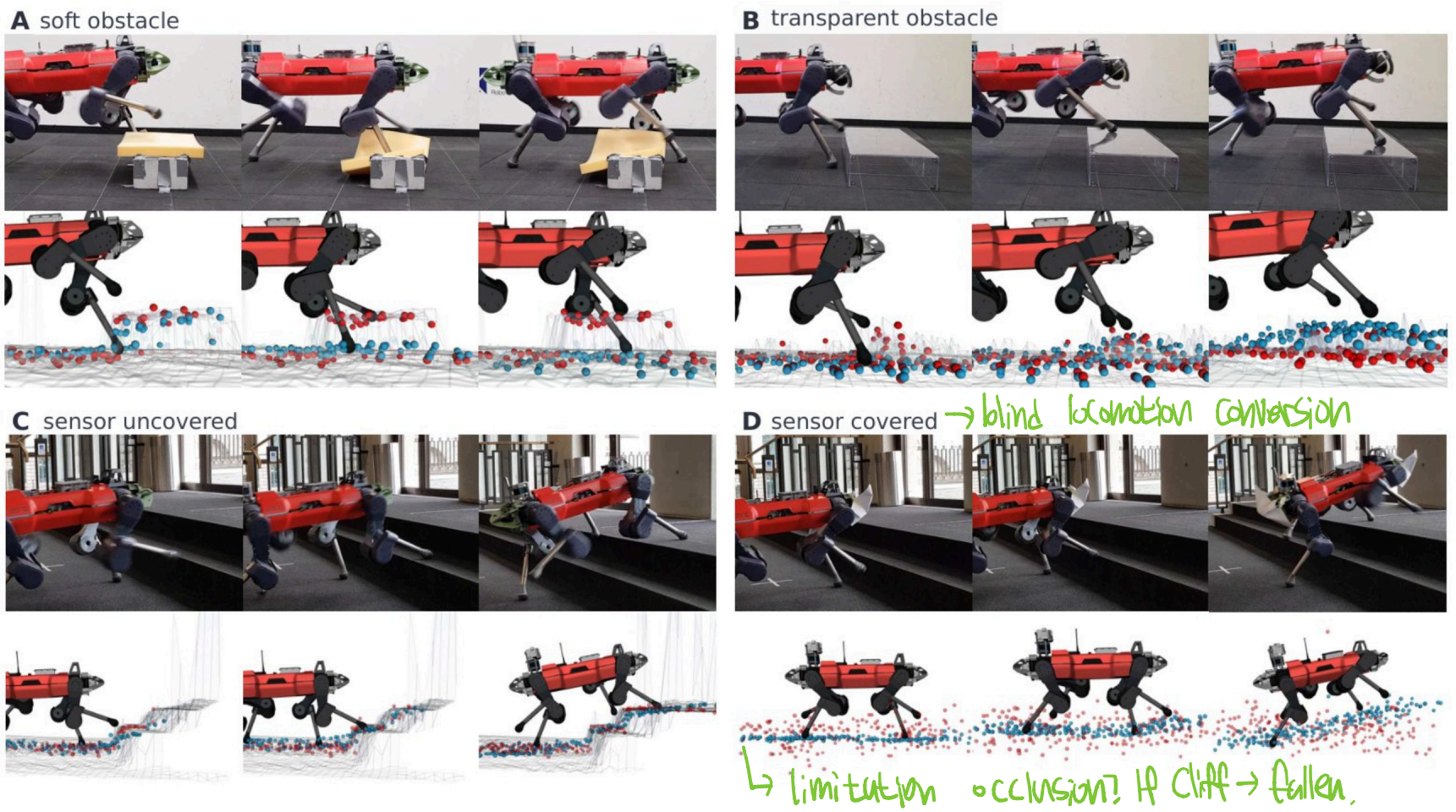






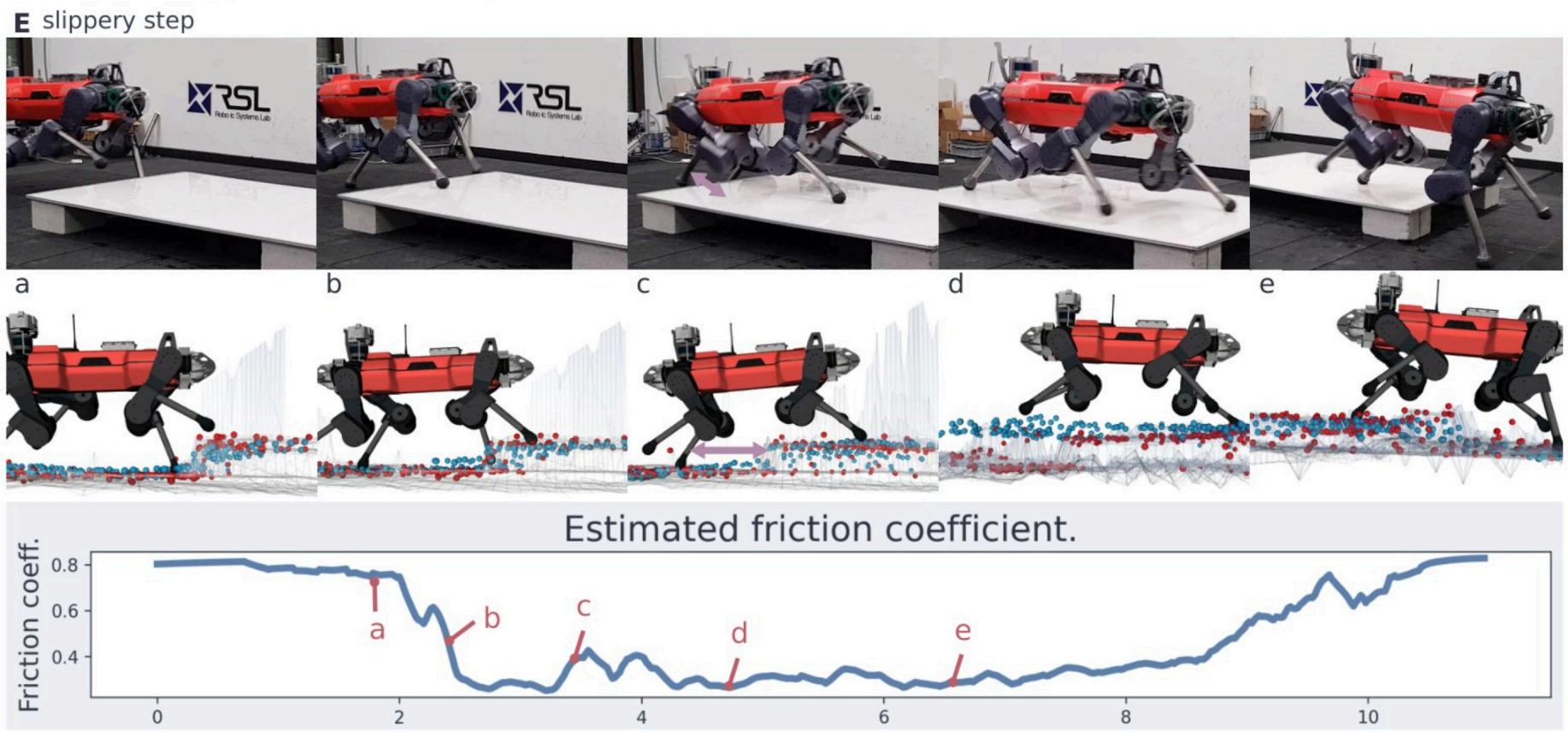
Obstacle course: ours **33 s, no help** · baseline 75 s, lifted and pushed at all three obstacles

Looking Inside the Belief State



Red: height samples given as input · **Blue**: robot's belief, decoded — (A) soft (B) transparent (C) sensors on (D) sensors covered

Slippery Surface: Friction Is Inferred Online



Detects **low friction** → steps faster; while pose drift corrupts the map, it **falls back on proprioception** until the map recovers

Discussion

Limitations & What Came Next

Limitation	Why	Follow-up direction
Still falls back to proprioception under depth noise / occlusion	Fallback is safe but slow; cannot prepare for what it cannot see	Richer inputs and explicit uncertainty
May step off an occluded cliff or stepping stone	Uncertainty used only implicitly; assumes continuous ground	Explicit uncertainty, learned occlusion; precise footholds via MPC + RL (DTC, 2024)
Branches / low ceilings seen as walls	2.5D height map cannot represent overhangs	Raw depth end-to-end; 3D volumetric memory
Loses material and texture; pose errors leak into the map	Geometry-only map; separate classical pose estimator	Feed raw sensor data into the network
No non-walking skills (stuck leg, high ledges)	Gait generator and rewards are built for walking	Parkour (ANYmal Parkour, Extreme Parkour)
Two-stage cost and imitation gap	Some teacher info can never be inferred	Single-stage: asymmetric critic, DreamWaQ, world models (WMP)

Intuition: Answer Sheet vs. Foggy Night

Teacher

A student solving problems with the **answer sheet open**: knows the exact terrain and how slippery the ground is, and first learns *"this is how to walk"*.

Student

A hiker on a **foggy night**: only blurry eyes and feet. Recalls the last few steps (**GRU**), decides how far to trust the eyes (**gate**), and imitates what the answer-sheet student would do.

"Is this hard rock or soft snow?" — estimated **before the foot lands**

ANYmal's Day ①: Training

STAGE 1 · TEACHER

1,000 ANYmals fall at once

Born in simulation with the answer sheet. Points for tracking the command, penalties for stumbling. Walk well → the curriculum serves harder terrain.

STAGE 2 · STUDENT

Answer sheet gone, blurry glasses on

Heights jitter, go blank, or shift wholesale. At every step the teacher says “I'd do this” and “the ground really looked like that”.

STAGE 3 · DEPLOY

Dropped on a mountain

No extra training. Straight to the Alps, caves and snowy stairs.

ANYmal's Day ②: One Second on Snowy Stairs

Time	What happens inside ANYmal	Gate α
0.00 s	Map shows a 17 cm step ahead; eyes and feet have agreed so far → lifts the front foot early	Open (trust vision)
0.10 s	Foot keeps sinking below the mapped floor — the map saw the snow surface, not the ground	Open
0.14 s	Late contact enters GRU memory: “eyes and feet disagree” → belief revised downward	Partly closing
0.20 s	A slip → pose estimate drifts, whole map shifts → robust proprioceptive gait, faster steps	Mostly closed
0.60 s+	Map stabilizes, senses agree again → back to full speed	Re-opening

Illustrative scenario combining behaviors observed in the paper — none of it is a hand-written “snow rule”; all learned.

Latest Trend: End-to-End & Single-Stage

① End-to-end input

Skip the hand-designed elevation map; feed **raw depth or RGB pixels** straight into the policy.

Keeps what the map threw away: space under branches, surface material.

e.g., Agarwal et al. 2022 (depth, still teacher–student)

② No privileged teacher

Learn in **one stage** from observations only — as animals do, without an answer sheet.

Privileged info only for the critic, concurrent state estimation, or a **world model** predicting future perception.

e.g., DreamWaQ; WMP (Lai et al., 2024)

Direction: **fewer hand-designed intermediates, fewer answer sheets** — now spreading to humanoids

Conclusion

Take-home Messages

1. Fast physical control is handled not by LLMs but by controllers trained with **simulation RL** — the two **divide the work**.

2. Privileged teacher → noisy-eyed student + an attention gate gives a walker that decides when to trust vision: **fast and robust**.

3. The field is moving toward **fewer intermediates and answer sheets**, and **GPU parallelism** makes "training in minutes" the norm.

Thank you

Q & A

Appendix: Success Rates in Simulation

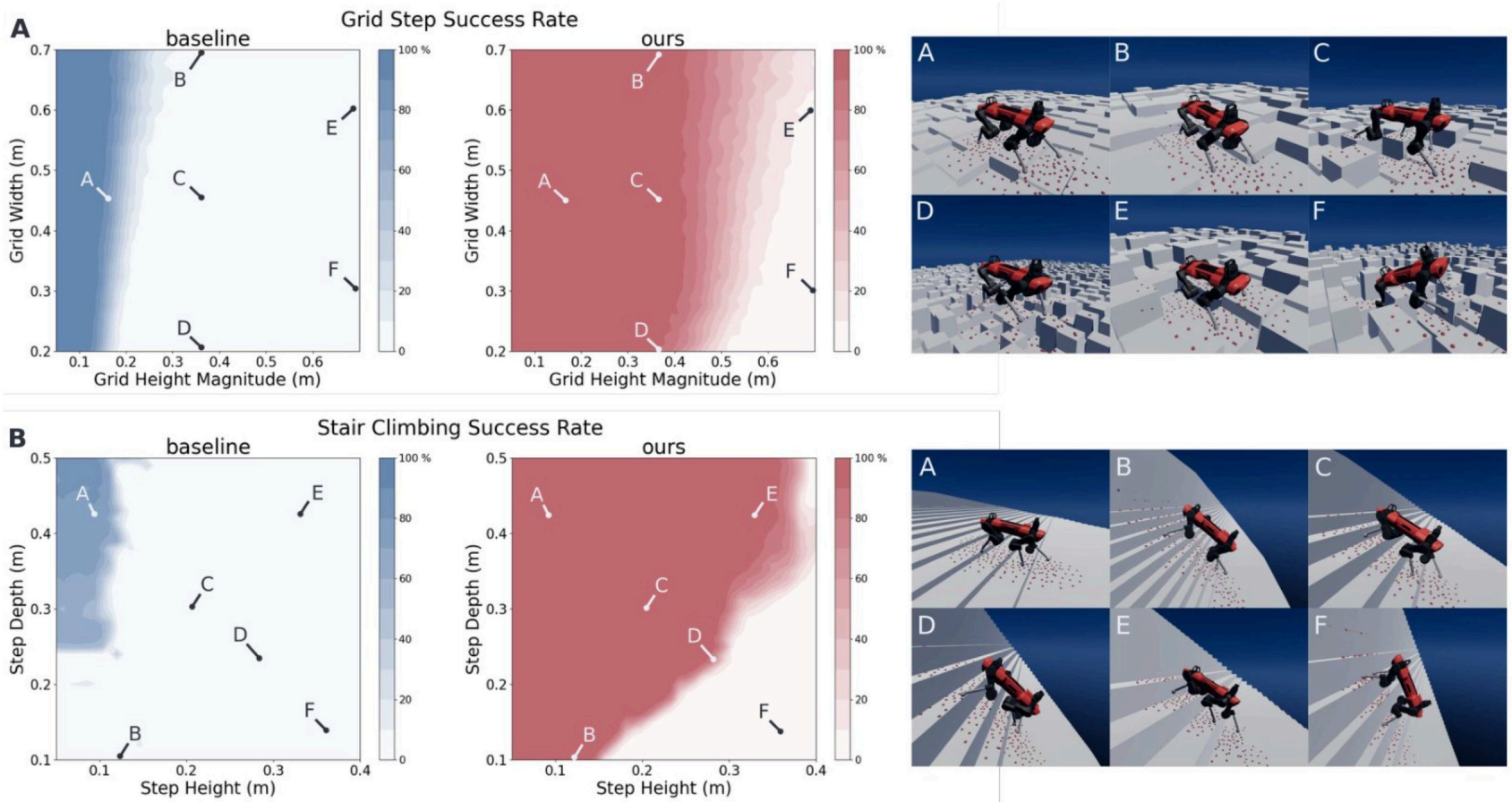


Fig. S1 — 300 trials per setting over 41 × 41 terrain parameters; ours (red) vs. proprioceptive baseline (blue)