DAWN: Noise-Robust Quadruped Parkour via Depth-Denoising World Models

Anonymous Author(s)

Abstract— **Vision-based legged locomotion methods assume clean depth at training time and rely on hand-tuned post-processing filters at deployment. However, filter parameters are rarely disclosed, hindering reproducibility, and performance degrades substantially when depth noise is left unaddressed. Building noise robustness directly into the learning pipeline would eliminate this dependency. While such robustness has been explored for proprioceptive inputs, analogous approaches for depth perception remain largely absent in legged locomotion. We propose DAWN (Denoising and Alignment in World models for Noise-robustness), a noise-robust perception framework for legged locomotion, which builds noise robustness directly into a world model via two modifications: (1) feeding noisy depth to the encoder while keeping clean depth as the reconstruction target, forcing the model to implicitly denoise its input; and (2) applying contrastive learning to align the latent states of noisy and clean depth. Importantly, DAWN is agnostic to the specific noise model, requiring no assumptions about the noise distribution at deployment. Furthermore, it incurs no additional inference cost over existing world model-based methods. Without any manual filter calibration—relying solely on the learned noise model—DAWN achieves zero-shot quadruped parkour on a Unitree Go1: traversing stairs up to 18 cm, clearing gaps up to 70 cm, and mounting steps up to 45 cm from raw depth observations. Ablation studies show that denoising and contrastive alignment contribute at complementary levels—reconstruction and representation, respectively—and yield additive gains when combined. Videos and code are available at:** `https://dawn-parkour.github.io/`

I. INTRODUCTION

Reinforcement learning (RL) has enabled quadruped robots to traverse diverse terrains through sim-to-real transfer [1], [2]. A blind policy with only proprioceptive input can traverse moderate terrains such as slopes and stairs [3], [4], but fails on obstacles that require perceiving terrain geometry in advance, such as gaps and steps [5], [6]. Depth cameras provide direct geometric information about such terrains and have become the primary sensor for vision-based locomotion [6], [5]. Recent depth-based methods have demonstrated increasingly agile parkour on low-cost quadrupeds, from climbing obstacles up to 0.55 m and leaping gaps up to 0.85 m to jumping obstacles exceeding twice the robot height [7], [8], [9]. These and most other depth-based locomotion methods rely on Intel RealSense D435/D435i cameras [7], [8], [9], [10], [11]. However, the handling of sensor noise at deployment is rarely discussed.

Despite this progress, most vision-based locomotion methods assume clean depth during training and defer noise handling to post-processing filters at deployment [9], [8], [7], [5]. Even when depth degradation is acknowledged, common remedies remain outside the main learning pipeline—proprioceptive fallback [6] or a separate learned reconstruction module [10]—suggesting that the standard training procedure alone does not produce noise-robust policies. Moreover, while post-processing filters are widely used at inference [7], [8], [9], [11], their specific parameters are seldom reported. The librealsense2 pipeline for the D435i [12] exposes six filter types with 12–24 interdependent parameters. Their optimal values vary with illumination, surface material, and scene depth, making consistent reproduction across environments difficult.

The severity of this issue has been quantified: Sun et al. [13] reported a 56 percentage-point (p.p.) improvement in success rate when depth augmentations were applied; separately, and Sun et al. [14] showed that physically-informed noise synthesis reduces terrain reconstruction error by 27%. However, the approaches explored so far operate at the input level: domain randomization [15], hand-crafted augmentation [13], [14], or separate denoising modules [10]. Hand-crafted filters cannot address scene-dependent noise with fixed parameters [16], prior depth denoising methods exhibit mismatches with current stereo sensors [17]. These input-level augmentations provide limited representational capacity [18]. Meanwhile, world model-based denoising has been applied only to proprioception [19], [20], where noise is low-dimensional and approximately i.i.d. Depth images, by contrast, exhibit spatially structured noise whose statistics vary with scene geometry and illumination [21], [12]—a fundamentally different regime that prior approaches have not addressed.

In this work, we propose DAWN (**D**enoising and **A**lignment in **W**orld models for **N**oise-robustness), a perception framework for legged locomotion that embeds noise robustness into the latent space of a world model through two modifications to the Recurrent State-Space Model (RSSM) architecture of Lai et al. [9]. Specifically, DAWN introduces an input–target mismatch in the RSSM, where the encoder receives noisy depth but the decoder is supervised with clean depth, and applies SimCLR [22]-based contrastive alignment between the latent representations of noisy and clean depth. Unlike fixed-parameter filters that require scene-specific tuning, our RSSM encoder learns a nonlinear mapping that discards noise conditioned on scene context, trained end-to-end with the locomotion policy. Critically, both modifications apply only at training time, adding no inference cost over the base world model. Notably, DAWN is agnostic to the specific noise model, requiring no assumptions about the noise distribution at deployment.

In simulation, DAWN achieves 96.9% average success rate across stairs, gaps, and steps from raw depth observations,

improving over World Model-based Perception (WMP) [9], the depth-based baseline (91.4%) by 5.5 p.p. and closing 77% of the gap to the clean-depth oracle (98.5%). When noise intensity is doubled, DAWN degrades by only 6.5 p.p. compared to 17.5 for the baseline, indicating robustness to out-of-distribution noise levels. Deployed on a real Unitree Go1 without any manual filter calibration—relying solely on the learned noise model to bridge the sim-to-real gap—DAWN accomplishes zero-shot quadruped parkour: traversing stairs up to 18 cm, clearing gaps up to 70 cm, and mounting steps up to 45 cm, with consistent performance across both indoor and outdoor environments.

Below, we summarize our main contributions:

1) **DAWN**, a noise-robust perception framework that builds depth denoising directly into the world model eliminating environment-dependent filter tuning at deployment.
2) **Zero-shot sim-to-real quadruped parkour** on a Unitree Go1 **without manual filter calibration**, traversing stairs 18 cm, gaps 70 cm, steps 45 cm, from raw depth observations.
3) **Systematic ablations** confirming that RSSM denoising and contrastive alignment operate at complementary levels—reconstruction and representation—and yield additive gains when combined.

These results suggest that noise robustness for depth-based legged locomotion can be achieved through learned representation rather than hand-engineered filtering.

II. RELATED WORK

We review four lines of work relevant to DAWN: visual legged locomotion, depth sim-to-real transfer, world models for locomotion, and contrastive learning for locomotion.

A. Visual Legged Locomotion

Sim-to-real RL [1], [2] combined with privileged learning [23], [3] has become the dominant paradigm for legged locomotion. Policies with only proprioceptive input can traverse moderate terrains such as slopes and stairs [3], but fail on obstacles that require perceiving terrain geometry in advance, such as gaps and steps [5], [6]. Depth cameras provide direct geometric information for such terrains, and vision-based parkour has progressed from climbing obstacles up to $1.5\times$ the robot height [7] to jumps exceeding $2\times$ the robot height [8]. Lai et al. [9] introduced a world model-based approach that bypasses the teacher–student distillation pipeline, learning perception and policy end-to-end via the RSSM. These methods commonly train with clean depth and apply hand-tuned post-processing filters at deployment. Because optimal filter parameters vary with illumination, surface material, and distance distribution, a configuration tuned for one environment does not transfer reliably to another.

B. Depth Sim-to-Real Transfer

Domain randomization [15] is widely used for sim-to-real transfer, but even large-scale automatic domain randomization does not fully close all sim-to-real gaps [24]. Keselman et al. [12] reported the official characteristics of the RealSense D400 series, and Ahn et al. [21] provided an empirical noise model of the D435. Sweeney et al. [16] analyzed scene-dependent filter behavior, and Hu et al. [17] examined the applicability of prior depth denoising methods to current stereo sensors. Liu et al. [25] proposed a systematic depth simulation pipeline, and Sun et al. [14] and Zhuang et al. [11] introduced physically-informed noise synthesis. Sun et al. [13] reported a 56 p.p. success rate gain through eight depth augmentations, quantifying the severity of depth noise. Hoeller et al. [10] addressed noisy depth with a learned terrain reconstruction module that fuses six depth cameras and LiDAR. This approach, however, requires auxiliary hardware and is optimized independently of the downstream policy, preventing end-to-end learning.

Existing work focuses on input-level noise modeling or separate denoising modules, both of which require scene-specific filters or auxiliary components at deployment. Repurposing the reconstruction objective of a world model to remove filter dependency entirely has not been explored.

C. World Models for Locomotion

Following Ha and Schmidhuber [26], Hafner et al. [27] introduced the RSSM, which evolved through subsequent iterations [28], [29], [30]. Wu et al. [31] demonstrated learning locomotion from scratch on a real robot using a world model.

Gu et al. [19] applied noisy-to-clean reconstruction of proprioception to achieve noise-robust locomotion, and Sun et al. [20] introduced gradient cutoff to protect denoising quality. These works, however, address only proprioception, whose noise is low-dimensional and approximately i.i.d. Depth noise is fundamentally different: it concentrates at depth discontinuities and grows nonlinearly with distance [21], [12], requiring architectural considerations for processing spatially structured features.

D. Contrastive Learning for Locomotion

Contrastive representation learning originated with van den Oord et al. [32] and advanced in the vision domain through He et al. [33] and Chen et al. [22]. In RL, Srinivas et al. [34] demonstrated improved sample efficiency for pixel-based control. For locomotion, Long et al. [35] applied prototypical representation alignment, Mousa et al. [36] used contrastive triplet loss for teacher–student alignment, and Lu et al. [37] improved sim-to-real transfer for humanoid locomotion. These approaches target proprioceptive or general visual representations; enforcing invariance to depth noise via contrastive learning has not been explored.

III. METHOD

DAWN builds upon the RSSM architecture of Lai et al. [9] and introduces two modifications for depth noise robustness: (1) a denoising reconstruction objective that feeds noisy depth to the encoder while keeping clean depth as the reconstruction target, and (2) a contrastive loss that aligns latent states from noisy and clean observations. An overview is shown in Fig. 1.

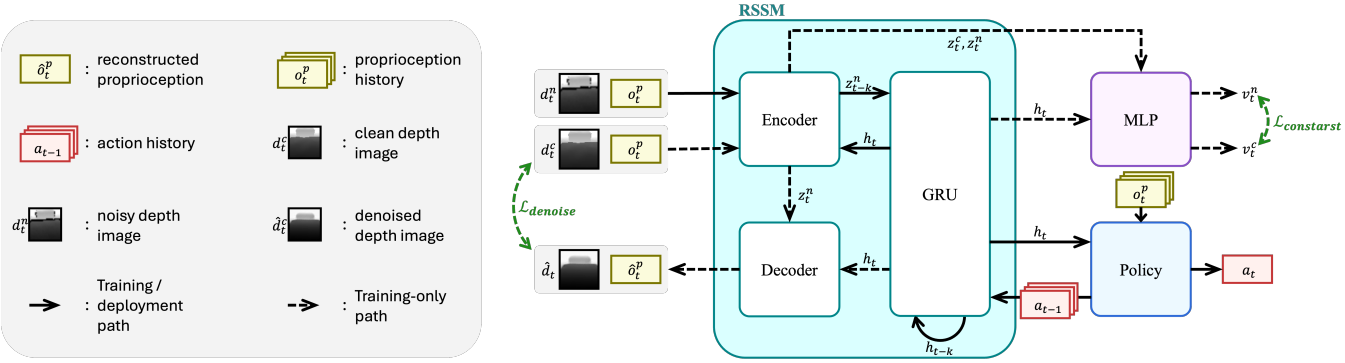

Fig. 1: Overview of the DAWN framework. Solid arrows indicate the training and deployment path; dashed arrows indicate the training-only path. DAWN's two modifications are highlighted in green: the denoising reconstruction loss $\mathcal{L}_{\text{denoise}}$ between the decoder output and the clean depth target, and the contrastive loss $\mathcal{L}_{\text{contrast}}$ applied to the projection head (MLP). At deployment, the training-only components are removed, and only the solid-arrow path is executed.

A. Preliminaries: World Model Learning

We define the observation at timestep t as $x_t = (d_t, o_t^{\text{p}})$, where d_t is the depth image and o_t^{p} is the proprioception. The depth image is updated every k steps, and a_t denotes the joint position target action.

Lai et al. [9] propose World Model-based Perception (WMP), an end-to-end framework that adopts an RSSM following Hafner et al. [29]. The RSSM consists of four components parameterized by ϕ:

$$\text{Sequence model:} \quad h_t = f_\phi(h_{t-k},\, z_{t-k},\, a_{t-k:t-1}) \tag{1}$$

$$\text{Encoder:} \quad z_t \sim q_\phi(\cdot \mid h_t,\, x_t) \tag{2}$$

$$\text{Dynamics predictor:} \quad \hat{z}_t \sim p_\phi(\cdot \mid h_t) \tag{3}$$

$$\text{Decoder:} \quad \hat{x}_t \sim p_\phi(\cdot \mid h_t,\, z_t) \tag{4}$$

where h_t is a deterministic state computed by a Gated Recurrent Unit (GRU) [38]-based sequence model (1), z_t is the stochastic state incorporating the current observation (2), $\hat{z}_t$ is the prior predicted without observation access (3), and $\hat{x}_t$ is the reconstructed observation (4).

These components are jointly optimized by minimizing:

$$\mathcal{L}_{\text{WMP}} = \mathbb{E}\Big[\sum_t \underbrace{-\ln p_\phi(x_t \mid z_t,\, h_t)}_{\text{reconstruction}} + \beta\, \mathrm{KL}\big[q_\phi(z_t \mid h_t,\, x_t) \,\|\, p_\phi(z_t \mid h_t)\big]\Big], \tag{5}$$

where β is a hyperparameter. The reconstruction term encourages z_t to retain sufficient information about x_t, while the KL term regularizes the posterior toward the prior. The policy is trained with Proximal Policy Optimization [39] using the deterministic state h_t as input.

In WMP, both the encoder input and the reconstruction target are the *same* clean observation x_t. Depth noise at deployment is handled by post-processing filters applied to raw sensor readings. DAWN removes this filter dependency by modifying how the RSSM is trained, without changing its architecture.

B. Denoising Reconstruction Objective

We denote clean and noisy depth as d_t^{c} and d_t^{n}, respectively, and write the corresponding observations as $x_t^{\text{c}} = (d_t^{\text{c}},\, o_t^{\text{p}})$ and $x_t^{\text{n}} = (d_t^{\text{n}},\, o_t^{\text{p}})$. The noisy depth d_t^{n} is generated by a noise model.

In WMP's standard training, the encoder input and the decoder target are identical:

$$z_t \sim q_\phi(\cdot \mid h_t,\, x_t^{\text{c}}), \qquad \hat{x}_t \approx x_t^{\text{c}}. \tag{6}$$

DAWN replaces the encoder input with the noisy observation while keeping the clean observation as the reconstruction target:

$$z_t^{\text{n}} \sim q_\phi(\cdot \mid h_t,\, x_t^{\text{n}}), \qquad \hat{x}_t \approx x_t^{\text{c}}. \tag{7}$$

Comparing (6) and (7), the only change is in the encoder input: from x_t^{c} to x_t^{n}. The decoder must still reconstruct the clean observation, which forces the encoder to map noisy input to a latent state from which clean depth can be recovered.

This input–target mismatch turns the standard reconstruction loss into a denoising objective. Substituting (7) into the WMP loss (5) yields the DAWN reconstruction loss:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}\Big[\sum_t \underbrace{-\ln p_\phi(x_t^{\text{c}} \mid z_t^{\text{n}},\, h_t)}_{\text{denoising}} + \beta\, \mathrm{KL}\big[q_\phi(z_t^{\text{n}} \mid h_t,\, x_t^{\text{n}}) \,\|\, p_\phi(\hat{z}_t \mid h_t)\big]\Big]. \tag{8}$$

The KL term constrains the information capacity of z_t^{n}. Since reconstructing x_t^{c} does not require any information about the noise ϵ in x_t^{n}, the encoder discards noise-specific features under this capacity constraint and retains only terrain-relevant geometry. This behavior follows directly from the information bottleneck principle [40]: when the target is x_t^{c}, the mutual information $I(z_t^{\text{n}};\, \epsilon)$ does not contribute to reducing the reconstruction loss and is suppressed by the KL penalty.

This learned compression provides a practical advantage over handcrafted post-processing filters. Filter pipelines operate with a fixed set of parameters that require scene-specific tuning, whereas the encoder learns a nonlinear mapping trained end-to-end with the policy and conditioned on scene context. Because the denoising objective already drives the encoder to retain only task-relevant geometry, no additional filter stage is required at deployment.

TABLE I: Comparison of WMP and DAWN training configurations. The RSSM architecture and policy training are identical; only the encoder input and loss terms differ.

	WMP	DAWN
Encoder input	x_t^{c}	x_t^{n}
Decoder target	x_t^{c}	x_t^{c}
Reconstruction loss	$\mathcal{L}_{\mathrm{WMP}}$	$\mathcal{L}_{\mathrm{denoise}}$
Contrastive loss	–	$\mathcal{L}_{\mathrm{contrast}}$
Post-processing filter	Required	Not required

C. Contrastive Latent Alignment

The denoising objective encourages noise-invariant reconstruction but does not explicitly constrain the latent space structure. We add a contrastive loss to directly align the encoder outputs from clean and noisy observations.

We write $z_t^{\mathrm{c}} \sim q_\phi(\cdot \mid h_t, x_t^{\mathrm{c}})$ and $z_t^{\mathrm{n}} \sim q_\phi(\cdot \mid h_t, x_t^{\mathrm{n}})$ for the posterior states encoded from clean and noisy inputs at the same timestep. A projection head g_ψ maps these posteriors to a contrastive embedding space: $v_t^{\mathrm{c}} = g_\psi(h_t,\, z_t^{\mathrm{c}})$ and $v_t^{\mathrm{n}} = g_\psi(h_t,\, z_t^{\mathrm{n}})$.

From N parallel environments within a batch, the clean and noisy embeddings of the same scene i, $(v_i^{\mathrm{c}},\, v_i^{\mathrm{n}})$, form a positive pair, while those from different scenes $i \neq j$ form negative pairs. The Normalized Temperature-scaled Cross Entropy (NT-Xent) loss is:

$$\mathcal{L}_{\mathrm{contrast}} = -\log \frac{\exp(\mathrm{sim}(v_i^{\mathrm{c}},\, v_i^{\mathrm{n}})/\tau)}{\displaystyle\sum_{j=1}^{2N} \mathbb{1}_{[j\neq i]}\, \exp(\mathrm{sim}(v_i^{\mathrm{c}},\, v_j)/\tau)}, \quad (9)$$

where v_i^{c} is the anchor, $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity, $\mathbb{1}_{[j\neq i]}$ is an indicator function excluding the anchor, $\{v_j\}_{j=1}^{2N} = \{v_1^{\mathrm{c}}, \ldots, v_N^{\mathrm{c}},\, v_1^{\mathrm{n}}, \ldots, v_N^{\mathrm{n}}\}$ is the set of all embeddings in the batch, and τ is the temperature parameter.

Following Chen et al. [22], the contrastive loss is applied to the projected embeddings v_t rather than directly to z_t, preventing the encoder from collapsing toward the contrastive objective and preserving diverse information useful for policy learning. At deployment, the projection head g_ψ is removed. The contrastive loss serves only as a training-time regularizer that shapes the latent geometry; the inference pipeline remains identical to WMP.

D. Total Training Objective

The full DAWN loss combines the denoising reconstruction objective (8) with the contrastive alignment loss (9):

$$\mathcal{L}_{\mathrm{DAWN}} = \mathcal{L}_{\mathrm{denoise}} + \lambda\, \mathcal{L}_{\mathrm{contrast}}, \quad (10)$$

where λ controls the relative weight of the contrastive term. Setting $\lambda = 0$ and replacing x_t^{n} with x_t^{c} in (8) recovers the original WMP loss (5).

Table I summarizes the differences between WMP and DAWN. The RSSM architecture and the policy training procedure remain unchanged; DAWN modifies only the training signals.

IV. EXPERIMENTAL RESULTS

A. Experimental Setup

Simulation. Training is conducted in IsaacLab [41] with 4,096 parallel environments. The control frequency is 50 Hz, and depth images at 64×64 resolution are updated every $k = 5$ steps (0.1 s). All experiments are repeated with three seeds, and 100 episodes per condition are evaluated, reporting mean±standard deviation.

Real-world. A Unitree Go1 is equipped with an Intel RealSense D435i depth camera and an Nvidia Jetson NX. Policies trained in simulation are transferred zero-shot without additional fine-tuning. We evaluate the success rate over 10 trials per difficulty level across both indoor and outdoor environments.

Depth Noise Simulation. DAWN is agnostic to the specific noise model. We use a D435i-characteristic model [21] capturing three phenomena, validated in §IV-F. We denote the intermediate depth image as $\tilde{d}$. The noise model applies the following filters to d_t^{c} to produce the final noisy depth d_t^{n}:

- *Gaussian sensor noise* adds zero-mean noise:

$$\tilde{d} = d_t^{\mathrm{c}} + \eta, \quad \eta \sim \mathcal{N}(0,\, \sigma^2), \quad \sigma = 0.01\,\mathrm{m}. \quad (11)$$

- *Edge dropout* simulates stereo matching failure at depth discontinuities with quadratic probability:

$$P_{\mathrm{drop}}(x,y) = P_{\max} \cdot \mathrm{clip}\left(\frac{G - G_{\mathrm{th}}}{G_{\mathrm{sat}} - G_{\mathrm{th}}},\, 0,\, 1\right)^2, \quad (12)$$

  where $G = \|\nabla \tilde{d}\|$ is the spatial gradient magnitude, $G_{\mathrm{th}} = 0.04$ is the minimum gradient for edge detection, $G_{\mathrm{sat}} = 0.20$ is the gradient at which the dropout probability saturates, and $P_{\max} = 0.60$ is the upper bound on dropout probability. Dropped pixels are filled by 3×3 max pooling.
- *Far particle noise* models infrared (IR)–ambient light interference:

$$P_{\mathrm{particle}}(x,y) = r_p \cdot \mathrm{clip}\left(\frac{\tilde{d} - d_{\mathrm{th}}}{0.2},\, 0,\, 1\right), \quad (13)$$

  where $r_p = 0.003$ is the maximum particle rate and $d_{\mathrm{th}} = 0.3$ is the depth threshold beyond which particles are applied.

All RealSense built-in filters are disabled.

Terrains. Four terrain types with five difficulty levels each: Slope (6–22°), Stair (6–18 cm), Gap (10–90 cm), and Step (10–54 cm). Slope is included only in the baseline comparison (Table II) and excluded from ablation and noise sweep analyses, as its lack of depth discontinuities yields similar performance across all methods.

Compared Methods. Table III summarizes the compared methods. WMP w/ N adds noise training to WMP. DAWN w/o C and DAWN w/o D are ablation variants that remove contrastive learning and RSSM denoising, respectively, yielding denoising-only and contrastive-only variants.

Metrics. We report Success Rate (SR, %) and velocity Tracking Error (TE, m/s). SR measures the percentage of

TABLE II: Baseline comparison (noise $\times 1.0$). SR (%, ↑) and TE (m/s, ↓) are reported. Averaged over all difficulty levels, 3 seeds $\times$ 100 episodes.

	Slope (6–22°)		Stair (6–18 cm)		Gap (10–90 cm)		Step (10–54 cm)	
Method	SR	TE	SR	TE	SR	TE	SR	TE
Blind	98.9±3.8	0.025±.004	74.6±44.5	0.064±.066	26.2±44.0	0.154±.073	36.2±43.4	0.936±.071
EP	100.0±0.0	0.011±.003	82.2±35.4	0.031±.042	83.5±35.9	0.072±.048	92.6±23.2	0.038±.031
WMP	100.0±0.0	0.010±.003	89.7±30.4	0.026±.038	88.6±31.8	**0.058**±.034	95.9±19.8	0.026±.018
WMP w/ N	99.9±3.2	0.009±.003	94.5±22.8	0.023±.036	93.7±24.3	0.067±.029	95.3±21.2	**0.025**±.021
DAWN	**99.9**±2.5	**0.008**±.002	**96.6**±18.2	**0.015**±.020	**97.2**±16.6	0.065±.031	**97.0**±17.0	0.034±.020
Oracle	100.0±0.0	0.008±.002	98.2±13.3	0.013±.015	98.8±10.9	0.054±.028	98.5±12.2	0.020±.014

episodes in which the robot successfully traverses the terrain without falling. TE is the mean squared error between the commanded velocity and the measured velocity capturing how precisely the robot tracks the desired motion.

B. Baseline Comparison

Table II presents baseline comparison results under the noise $\times 1.0$ condition. Results are averaged across all difficulty levels and reported as mean±standard deviation over three seeds.

Under this condition, WMP degrades most on Stair (89.7%) and Gap (88.6%), where boundary artifacts remove depth at the geometric features the policy relies on. Stair presents consecutive edges subject to cumulative dropout. For gap, accurate width estimation requires both boundaries to remain visible. Step (95.9%) involves a single depth discontinuity. Even when edge dropout partially corrupts the boundary, the height change can still be inferred from depth values on either side. Slope yields $\sim$100% SR for all methods including Blind (98.9%), indicating that proprioception alone suffices for this terrain and depth noise has negligible impact on performance.

Across Stair/Gap/Step, DAWN reaches 96.9% average SR, improving over WMP (91.4%) by 5.5 p.p. WMP w/ N (94.5% on Stair, 93.7% on Gap) improves over WMP through noise exposure but remains 2.4 p.p. below DAWN, confirming that domain randomization alone is insufficient. For TE, DAWN records 0.015 m/s on Stair—42% lower than WMP's 0.026—reflecting more precise foot placement from denoised depth. Blind fails on Gap (26.2%) and Step (36.2%), confirming that depth information is essential for obstacle traversal.

TABLE III: Summary of compared methods. Noise: depth noise applied during training. Denoise: reconstruction target set to clean depth. Contrastive: contrastive learning applied.

Method	Noise	Denoise	Contrastive
Blind	–	–	–
EP [8]	×	×	×
WMP [9]	×	×	×
WMP w/ N	✓	×	×
DAWN w/o C	✓	✓	×
DAWN w/o D	✓	×	✓
DAWN	✓	✓	✓
Oracle	–	–	–

TABLE IV: Noise scale parameters. $\times 1.0$ is the default setting.

Parameter	$\times 1.0$	$\times 1.5$	$\times 2.0$
Gaussian σ (m)	**0.01**	0.015	0.02
Edge $P_{\max}$	**0.60**	0.80	0.90
Particle r_p	**0.003**	0.0045	0.006

C. Ablation Study

Fig. 2 (left) compares the average SR across Stair/Gap/Step by difficulty level for five methods (WMP, WMP w/ N, DAWN w/o C, DAWN w/o D, DAWN).

All methods exceed 97% SR at low difficulty (levels 1–3) but diverge sharply from level 4 onward, as larger obstacles produce wider dropout regions that make depth quality increasingly critical. At difficulty 5, DAWN obtains 88.2% SR, compared to DAWN w/o C 82.4%, DAWN w/o D 81.3%, WMP w/ N 76.7%, and WMP 74.2%. These results indicate that both modifications are essential for noise robustness, as removing either one leads to a notable performance drop.

DAWN w/o C (denoising only) and DAWN w/o D (contrastive only) yield similar performance ($\sim$82% vs. $\sim$81%), yet neither alone reaches DAWN's 88.2%. DAWN w/o C influences the encoder only indirectly through the decoder-to-latent gradient and lacks a direct noise-robustness constraint. DAWN w/o D constrains the encoder directly, but with a noisy reconstruction target the latent space still encodes noise artifacts that conflict with alignment. At the highest difficulty, DAWN w/o C and DAWN w/o D achieve 5.8 and 6.9 p.p. lower SR than DAWN, respectively. These results confirm that a complementary gain arises when both components operate together. This arises because denoising structures the latent space toward noise-free geometry that facilitates contrastive alignment, while contrastive learning stabilizes encoder output for more reliable decoder reconstruction.

D. Noise Robustness Analysis

Fig. 2 (center) shows average SR at the highest difficulty across Stair/Gap/Step—Stairs 18 cm, Gap 90 cm, Step 54 cm—as the noise scale varies from $\times 1.0$ to $\times 2.0$. Table IV lists the parameter values for each scale, where $\times 1.0$ is the default setting.

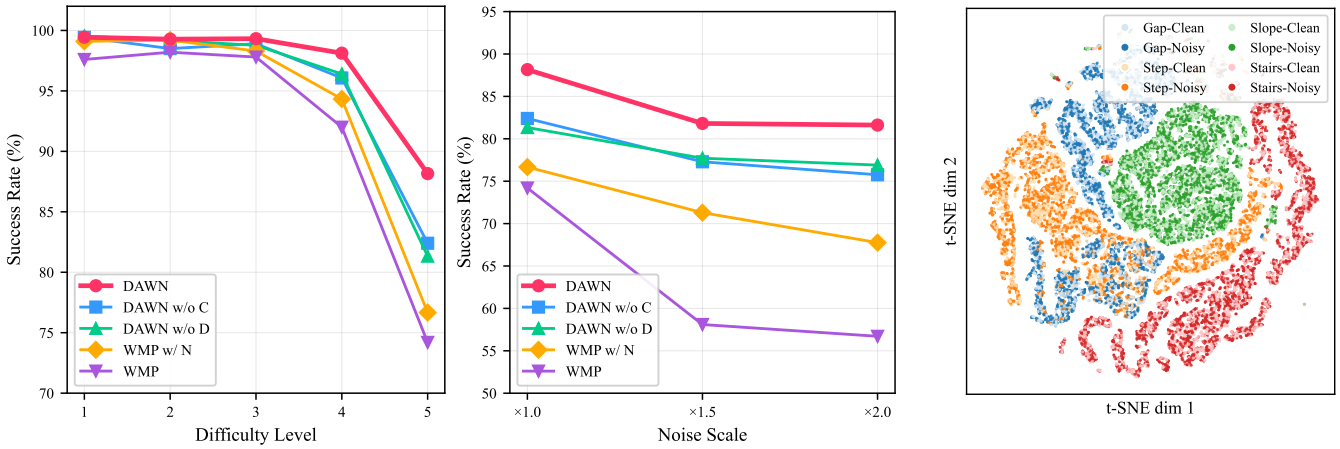

Fig. 2: Simulation analysis. Ablation study showing average SR over Stairs/Gap/Step by difficulty level (noise ×1.0), mean of 3 seeds (left). Noise robustness analysis showing average SR over Stairs/Gap/Step at the highest difficulty as the noise scale increases from ×1.0 to ×2.0, averaged over 3 seeds (center). t-SNE visualization of encoder latent states; colors indicate terrain type, shading indicates input condition (light: clean, dark: noisy) (right).

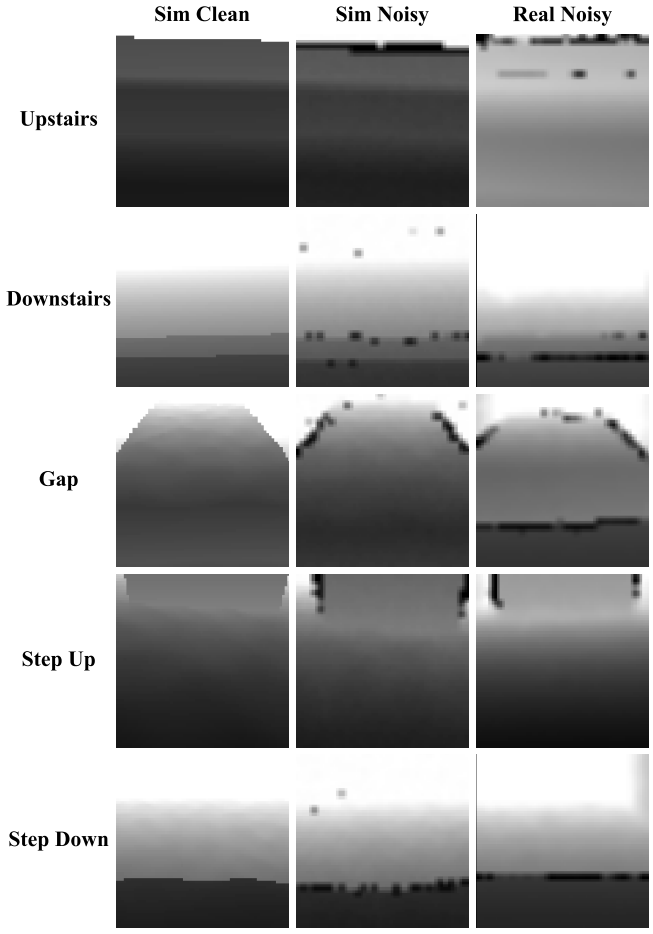

Fig. 3: Comparison of simulated clean depth, simulated noisy depth (×1.0), and real D435i depth across five terrains (Upstairs, Downstairs, Gap, Step Up, Step Down). The simulated noise patterns closely match real sensor output.

From ×1.0 to ×2.0, WMP drops 17.5 pp (74.2% to 56.7%), whereas DAWN degrades by 6.5 pp (88.2% to 81.6%)—less than half the rate. WMP w/ N degrades 8.9 p.p. (from 76.7% to 67.8%), confirming that noise exposure during training improves robustness but remains substantially below DAWN. The DAWN–WMP gap widens from 14.0 p.p. at ×1.0 to 24.9 p.p. at ×2.0, demonstrating that DAWN's architectural modifications provide increasing advantage as noise intensifies. Similarly, the gap between DAWN and WMP w/ N grows from 11.5 p.p. to 13.9 p.p., showing that domain randomization alone cannot match architectural-level robustness beyond the training distribution.

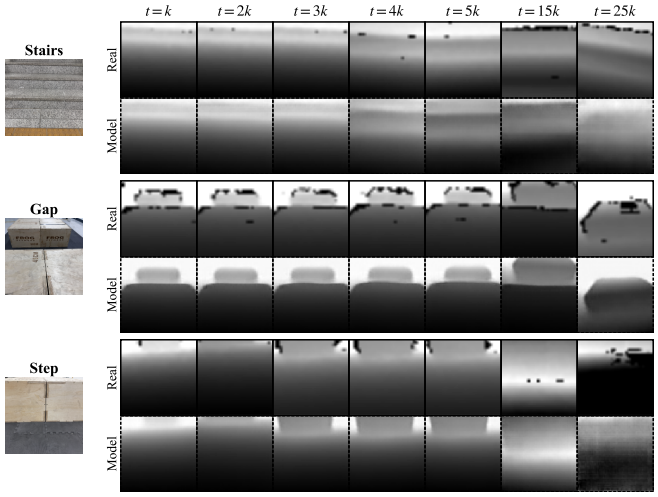

Fig. 4: Real-world depth reconstruction across three terrains (Stairs, Gap, Step). Each terrain shows seven timesteps ($t = k$ to $t = 5k$, $t = 15k$, $t = 25k$); top row: real D435i depth; bottom row (dashed border): DAWN's decoder output.

E. Latent Space Analysis

Fig. 2 (right) shows a t-SNE visualization of the concatenated (h_t, z_t). For each terrain type, the clean and noisy clusters overlap, indicating that the encoder maps both inputs to the same states. This means the encoder has learned to discard depth noise and extract only terrain geometry, achieving the noise invariance that DAWN targets. Mean-

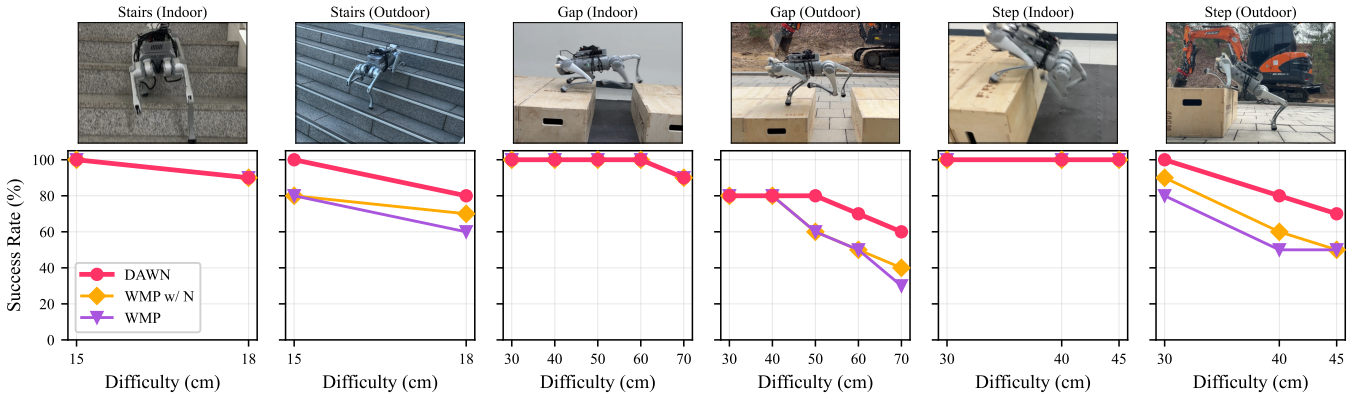

Fig. 5: Real-world experimental results. Three methods (DAWN, WMP w/ N, WMP) × three terrains (Stairs, Gap, Step) × two environments (Indoor, Outdoor). SR (%) over 10 trials per difficulty level.

while, different terrain types form well-separated clusters, confirming that the encoder preserves terrain discriminability.

F. Noise Model Validation

Fig. 3 compares simulated clean depth, simulated noisy depth ($\times 1.0$), and real D435i depth across five terrain views. The simulated noise visually matches real sensor output, particularly the missing pixels at depth discontinuities and far particle artifacts at distance, validating that the noise model captures the dominant characteristics of the D435i.

G. Real-World Depth Reconstruction

Fig. 4 shows DAWN's RSSM decoder reconstructing clean depth from real D435i input across consecutive timesteps on Stair, Gap, and Step. This confirms that the deterministic states h_t has learned to discard sensor noise. Since the policy conditions in h_t, it operates on noisy observations without requiring external filters.

H. Real-World Deployment

Fig. 5 presents real-robot results on a Unitree Go1 in indoor and outdoor environments, with SR measured over 10 trials per difficulty level.

In indoor environments, all three methods achieve comparable performance across terrains, reaching 100% SR at most difficulty levels and dropping uniformly to 90% only at the highest difficulty (Stairs 18 cm and Gap 70 cm). Step indoor shows no degradation, with all methods maintaining 100%.

The performance gap emerges outdoors, where sunlight-induced IR interference and surface material variation amplify depth corruption. On Stair at 18 cm, DAWN achieves 80% vs. WMP w/ N 70% and WMP 60%. On Gap at 70 cm, DAWN reaches 60%, WMP w/ N 40%, and WMP 30%. On Step at 45 cm, DAWN records 70% vs. WMP w/ N and WMP both at 50%. These results confirm that DAWN's noise-robust representations absorb environmental variation without filter tuning, whereas methods trained with clean depth degrade more sharply under outdoor noise conditions.

Fig. 6 shows outdoor deployment, further validating that the learned noise model generalizes beyond the controlled indoor setting. Despite IR interference from sunlight and diverse surface materials that intensify depth corruption, the policy maintains consistent behavior across all terrain geometries, without any environment-specific recalibration.

Fig. 6: Outdoor deployment snapshots: curb climbing on stone terrain, stair descending, slope descending, gravel walking, grass walking, and slope ascending.

V. CONCLUSION

We presented DAWN, a noise-robust perception framework for quadruped parkour that builds depth noise robustness into the world model's existing mechanisms, requiring no external modules or environment-specific tuning. DAWN passes noisy depth to the RSSM encoder while keeping clean depth as the reconstruction target, forcing the model to implicitly denoise its input. A contrastive loss further aligns noisy and clean representations at the encoder level. Together, these modifications remove the need for environment-dependent filter tuning at deployment while adding no inference overhead. In simulation, DAWN achieved 96.9% average success rate across stairs, gaps, and steps—close to the clean-depth oracle—and these improvements transferred consistently to a real Unitree Go1 via zero-shot deployment. Ablation confirmed that the two modifications serve complementary roles: denoising restructures the representation space toward clean geometry, while contrastive learning enforces noise-robust encoding, yielding a compound gain when combined. The current noise model targets the D435i,

but the framework is sensor-agnostic. Extending it to other depth sensors is a natural direction for future work.

REFERENCES

[1] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter, "Learning agile and dynamic motor skills for legged robots," *Sci. Robot.*, vol. 4, no. 26, 2019.

[2] J. Tan, T. Zhang, E. Coumans, A. Iscen, Y. Bai, D. Hafner, S. Bohez, and V. Vanhoucke, "Sim-to-real: Learning agile locomotion for quadruped robots," in *Proc. Robot.: Sci. Syst. (RSS)*, 2018.

[3] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "RMA: Rapid motor adaptation for legged robots," in *Proc. Robot.: Sci. Syst. (RSS)*, 2021.

[4] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning quadrupedal locomotion over challenging terrain," *Science Robotics*, vol. 5, no. 47, p. eabc5986, 2020.

[5] A. Agarwal, A. Kumar, J. Malik, and D. Pathak, "Legged locomotion in challenging terrains using egocentric vision," in *Proc. Conf. Robot Learn. (CoRL)*, 2022.

[6] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Sci. Robot.*, vol. 7, no. 62, 2022.

[7] Z. Zhuang, Z. Fu, J. Wang, C. Atkeson, S. Schwertfeger, C. Finn, and H. Zhao, "Robot parkour learning," in *Proc. Conf. Robot Learn. (CoRL)*, 2023.

[8] X. Cheng, K. Shi, A. Agarwal, and D. Pathak, "Extreme parkour with legged robots," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024.

[9] H. Lai, J. Cao, J. Xu, H. Wu, Y. Lin, T. Kong, Y. Yu, and W. Zhang, "World-model-based perception for visual legged locomotion," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2025.

[10] D. Hoeller, N. Rudin, D. Sako, and M. Hutter, "ANYmal parkour: Learning agile navigation for quadrupedal robots," *Sci. Robot.*, vol. 9, no. 88, p. eadi7566, 2024.

[11] Z. Zhuang, S. Yao, and H. Zhao, "Humanoid parkour learning," in *Proc. Conf. Robot Learn. (CoRL)*, 2024.

[12] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, "Intel RealSense stereoscopic depth cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR-W)*, 2017.

[13] W. Sun, Y. Su, L. Huang, A. Zhang, D. Wei, M. San, D. Tian, E. Cao, F. Yan, E. Xie, and Z. Xie, "Now you see that: Learning end-to-end humanoid locomotion from raw pixels," *arXiv preprint arXiv:2602.06382*, 2026.

[14] J. Sun, G. Han, P. Sun, W. Zhao, J. Cao, J. Wang, Y. Guo, and Q. Zhang, "DPL: Depth-only perceptive humanoid locomotion," *arXiv preprint arXiv:2510.07152*, 2025.

[15] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2017.

[16] C. Sweeney, G. Izatt, and R. Tedrake, "A supervised approach to predicting noise in depth images," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2019.

[17] J. Hu, C. Bao, M. Ozay, C. Fan, Q. Gao, H. Liu, and T. L. Lam, "Deep depth completion from extremely sparse data: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 7, pp. 8244–8264, 2023.

[18] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.

[19] X. Gu, Y.-J. Wang, X. Zhu, C. Shi, Y. Guo, Y. Liu, and J. Chen, "Advancing humanoid locomotion: Mastering challenging terrains with denoising world model learning," in *Proc. Robot.: Sci. Syst. (RSS)*, 2024.

[20] W. Sun, L. Chen, Y. Su, B. Cao, Y. Liu, and Z. Xie, "Learning humanoid locomotion with world model reconstruction," *arXiv preprint arXiv:2502.16230*, 2025.

[21] M. S. Ahn, H. Chae, D. Noh, H. Nam, and D. Hong, "Analysis and noise modeling of the Intel RealSense D435 for mobile robots," in *Proc. Int. Conf. Ubiquitous Robots (UR)*, 2019, pp. 707–711.

[22] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

[23] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl, "Learning by cheating," in *Proc. Conf. Robot Learn. (CoRL)*, 2019, pp. 66–75.

[24] OpenAI, I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, J. Schneider, N. Tezak, J. Tworek, P. Welinder, L. Weng, Q. Yuan, W. Zaremba, and L. Zhang, "Solving Rubik's Cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.

[25] X. Liu, C. Zhang, G. Wang, R. Zhang, and X. Ji, "RaSim: A range-aware high-fidelity RGB-D data simulation pipeline for real-world applications," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, 2024, pp. 17 057–17 064.

[26] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[27] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson, "Learning latent dynamics for planning from pixels," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 2555–2565.

[28] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020.

[29] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering Atari with discrete world models," in *Proc. International Conference on Learning Representations (ICLR)*, 2021.

[30] D. Hafner, J. Pasukonis, J. Ba, and T. Lillicrap, "Mastering diverse domains through world models," *Nature*, vol. 640, no. 8059, pp. 647–653, 2025.

[31] P. Wu, A. Escontrela, D. Hafner, K. Goldberg, and P. Abbeel, "DayDreamer: World models for physical robot learning," in *Proc. Conf. Robot Learn. (CoRL)*, 2022.

[32] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[33] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[34] A. Srinivas, M. Laskin, and P. Abbeel, "CURL: Contrastive unsupervised representations for reinforcement learning," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2020.

[35] J. Long, Z. Wang, Q. Li, L. Cao, J. Gao, and J. Pang, "HIM: Hybrid internal model for agile legged locomotion," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2024.

[36] A. Mousa, N. Karavis, M. Caprio, W. Pan, and R. Allmendinger, "TAR: Teacher-aligned representations via contrastive learning for quadrupedal locomotion," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, 2025, pp. 11 669–11 676.

[37] Y. Lu, R. Yang, Q. Kou, M. Chen, T. Fan, P. Cui, Y. Dong, and P. Lu, "Contrastive representation learning for robust sim-to-real transfer of adaptive humanoid locomotion," *arXiv preprint arXiv:2509.12858*, 2025.

[38] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[39] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[40] N. Tishby, F. C. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, 2000.

[41] M. Mittal, P. Roth, J. Tigue, A. Richard, O. Zhang, P. Du, A. Serrano-Muñoz, X. Yao, R. Zurbrugg, N. Rudin, *et al.*, "Isaac Lab: A GPU-accelerated simulation framework for multi-modal robot learning," *arXiv preprint arXiv:2511.04831*, 2025.