KiVi: Kinesthetic-Visuospatial Integration for Dynamic and Safe Egocentric Legged Locomotion

Peizhuo Li[1*], Hongyi Li[1,2*], Yuxuan Ma[1*], Linnan Chang[1], Xinrong Yang[1], Ruiqi Yu[3], Shuhao Liao[1], Yifeng Zhang[1], Yuhong Cao[1†], Qiuguo Zhu[3], Guillaume Sartoretti[1]

***Abstract*— Vision-based locomotion has shown great promise in enabling legged robots to perceive and adapt to complex environments. However, visual information is inherently fragile, being vulnerable to occlusions, reflections, and lighting changes, which often cause instability in locomotion. Inspired by animal sensorimotor integration, we propose KiVi, a Kinesthetic-Visuospatial integration framework, where kinesthetics encodes proprioceptive sensing of body motion and visuospatial reasoning captures visual perception of surrounding terrain. Specifically, KiVi separates these pathways, leveraging proprioception as a stable backbone while selectively incorporating vision for terrain awareness and obstacle avoidance. This modality-balanced, yet integrative design, combined with memory-enhanced attention, allows the robot to robustly interpret visual cues while maintaining fallback stability through proprioception. Extensive experiments show that our method enables quadruped robots to stably traverse diverse terrains and operate reliably in unstructured outdoor environments, remaining robust to out-of-distribution(OOD) visual noise and occlusion unseen during training, thereby highlighting its effectiveness and applicability to real-world legged locomotion.**

I. Introduction

Kinesthetic sense and visuospatial perception constitute two fundamental modalities that allow legged animals to achieve effective locomotion. On the one hand, kinesthetic feedback allows animals to maintain balance and adjust body posture with precision. On the other hand, visual information facilitates navigation through complex terrains and supports pre-emptive avoidance of potential hazards. For legged robots, these mechanisms remain equally valid, but their integration has been under-studied to date. Although reinforcement learning controllers for legged robots have demonstrated effective performance without vision [1]–[4], they remain incapable of negotiating terrains that require responding in advance, such as gaps or tall obstacles. Therefore, to achieve full autonomy for robots in unstructured and dynamic environments, it is necessary to integrate exteroceptive perception into their control frameworks to significantly enhance situational awareness and long-term adaptability [5].

However, the addition of visual perception does not always perform as effectively as expected. Compared to proprioception, which provides a compact but highly informative low-dimensional representation of the robot's state, visual input is sparser in both temporal resolution and relevance to physical interactions, necessitating compression or reformatting via encoders such as CNNs or MLPs [6]. Additionally, visual sensors are highly susceptible to structured disturbances in real-world scenarios, where reflections and occlusions often lead to misinterpretation of the surrounding terrain. These issues compromise the reliability of the overall control strategy. Within reinforcement learning (RL) pipelines, the inherent disturbances of real-world visual sensors are difficult to model accurately, which can further degrade policy performance through out-of-distribution (OOD) effects [7].

Fig. 1. Robust locomotion and obstacle avoidance of Deeprobotics Lite3 across diverse terrains and under severe visual disturbances, achieved using our proposed KiVi framework.

Current vision-based controller pipelines often fuse visual and proprioceptive information, then augment the compressed information with memory mechanisms [6], [8], to finally generate a latent representation that can serve as observation for the policy network. In these approaches, although attention mechanisms have improved multimodal integration, these methods still suffer from substantial performance degradation. This is largely due to strong modality entanglement that prevents effective fallback to proprioception when vision becomes unreliable under severe disturbances or degraded signals. Consequently, robust and reliable operation under such challenging conditions remains difficult to achieve.

In nature, legged animals address this challenge by employing a robust functional division between exteroception and proprioception in complex environments: their movement primarily relies on internal observations to maintain basic stability and coordination [9], while vision serves as a complementary source [10]. When visual input is reliable, they increasingly depend on it for environmental percep-

* Equal contribution. † Corresponding author.
[1] MARMot Lab, National University of Singapore, Singapore.
[2] Center of X-Mechanics, Zhejiang University, Hangzhou, China.
[3] Robot and Robot Intelligence Lab, Zhejiang University, Hangzhou, China.

tion to avoid obstacles and maneuver adaptively to traverse complex terrains. However, when vision is unreliable, e.g. galloping in tall grass, proprioception alone can still support stable locomotion, thereby preventing a complete breakdown of movement when vision is limited [11].

Building on this insight, we introduce the **KiVi** framework, which regulates multimodal information flow through two dedicated modules. The **Kinesthetic Module** encodes proprioceptive signals into a compact latent representation that provides a stable backbone for locomotion control. In parallel, the **Visuospatial Module** integrates visual observations with proprioceptive context and employs a **MemTransformer** [12] to enhance temporal memory, enabling the system to reconstruct terrain structures and anticipate upcoming obstacles. The two latent representations are then concatenated and passed to downstream components, allowing the robot to exploit vision for adaptive terrain traversal and obstacle avoidance while relying on proprioception to sustain robust gait generation. This separation mechanism ensures that, due to the inherently different signal-to-noise ratios across modalities—where visual observations are typically noisier—the policy naturally prioritizes the more reliable proprioceptive feedback when conflicting information arises, thereby producing more robust behaviors. Meanwhile, MemTransformer accelerates training convergence while ensuring that the robot's memory mechanism enables accurate terrain representation. Extensive experiments demonstrate that KiVi achieves superior performance in diverse real-world scenarios, including stable braking at near-minimum stopping distances, dynamic gap jumping, climbing over high walls, and maintaining central alignment in narrow corridors. Moreover, the framework exhibits minimal degradation under structured disturbances (e.g., reflective surfaces or dense foliage) and continues to function reliably during complete camera occlusion, establishing a solid foundation for long-term autonomous locomotion in complex and visually challenging environments.

Our main contributions can be summarized as follows:

- **Robust vision-based locomotion framework:** We propose KiVi, a framework that maintains stability under severe disturbances, supports zero-shot sim-to-real transfer on physical robots, and enables near-field obstacle avoidance and complex terrain traversal.
- **Modality-separated design:** By explicitly decoupling proprioceptive and visual pathways, our method relies on proprioception as a reliable backbone, while utilizing vision for adaptive behavior. This design ensures robust policy performance even under OOD conditions with severely corrupted or misleading visual input.
- **Application-oriented functional design:** Experiments demonstrate that our policy achieves robust performance in challenging outdoor environments, autonomously performing tasks such as obstacle avoidance, climbing, and gap jumping, thereby underscoring its effectiveness and applicability for real-world legged locomotion.

II. Related Work

Legged robots, long recognized for their ability to traverse complex terrains, have attracted sustained research attention for decades [13], [14]. However, their high degrees of freedom and inherent nonlinearity present significant challenges for controller design. Compared to traditional control methods [15]–[17], reinforcement learning (RL)-based control policies have shown greater tractability and adaptability, enabling robust locomotion on moderately challenging terrains such as gravel slopes and uneven surfaces [4], [18], [19]. Building on these advantages, recent research has increasingly focused on unlocking the full potential of quadrupedal robots, with the goal of achieving stable and dynamic mobility in highly complex and unstructured real-world environments.

A. Proprioceptive Locomotion

Proprioceptive locomotion constitutes the foundation of RL based locomotion for legged robots. These approaches typically rely exclusively on onboard sensory data such as IMU and joint encoder measurements as policy inputs, and use an actor-critic network architecture to achieve stable and reliable locomotion. In such methods, a key challenge lies in accurately estimating the state of the robot, especially its velocity, when performing complex maneuvers using only onboard sensors [20]. Early approaches introduced privileged observations for critics, allowing the actor–critic framework to evaluate the robot's state and advantage during training more accurately, making it possible to train robots that can traverse complex terrains under partial observations [21]. To further improve the real-world deployment, the incorporation of historical information to assist in state estimation is proved to be highly effective. For example, [1], [22], [23] employed a two-stage teacher-student distillation framework to leverage historical data, allowing real-world deployment in complex terrains including stairs or boulders. Alternatively, other approaches have streamlined this process into a single-stage method. For instance, the DreamWAQ [24] architecture employs supervised learning to estimate velocity and future states directly from historical data, achieving comparable performance while simplifying the training procedure.

Although these proprioception-based methods are fairly robust due to their relatively low-noise internal sensors, the lack of exteroception makes it more difficult to traverse complex terrains such as gaps, ditches, or elevated platforms (such as elevated footpaths, large steps, roadside curbs, etc.), preventing them from fully exploiting their inherent structural ability to handle challenging environments.

B. Perception-based Locomotion

To address the aforementioned limitations, many works have looked into incorporating perception such as vision into RL frameworks. Early works computed elevation maps via traditional methods as policy observations [6] or employed simplified exteroceptive sensing to achieve high-speed locomotion and obstacle avoidance [25]. To further unlock the potential of legged platforms, end-to-end learning

pipelines have been proposed. Building upon the teacher-student distillation training paradigm, algorithms such as [26]–[29] enabled robots to perform dynamic maneuvers across highly varied terrains. Other methods focused on improvements based on the DreamWAQ [24] framework. For example, PIE [30] and WMP [31] used supervised learning that integrates historical information and vision to predict the surrounding height map and the next robot observation, achieving single-stage parkour training while significantly reducing training complexity.

However, contrary to expectations, while vision equips the controller with environmental perception and anticipatory planning, it also significantly reduces the robustness of the controller under varying lighting conditions. For depth data to function effectively as an end-to-end input, it requires not only additional domain randomization during simulation training but also sufficiently favorable real-world conditions, such as the absence of severe reflections, occlusions, or poor visibility. In particular, environments with dense vegetation that partially occlude the camera can lead policies to exhibit erratic or overly aggressive actions, thereby limiting their generalizability and practical effectiveness.

III. Kinesthetic-Visuospatial Integration

To achieve robust and adaptive locomotion in legged robots, we propose **KiVi**, a bio-inspired perception–based locomotion framework that emulates key mechanisms by which animals coordinate proprioceptive and visual information. As illustrated in Fig. 2, our framework consists of two main components: an asymmetric actor–critic architecture for efficient training and a bio-inspired dual-branch estimator that assigns dedicated pathways to proprioceptive and visual inputs. The estimator produces compact latent representations that prioritize internal dynamics while selectively incorporating complementary visual cues, which are used by the actor to generate control actions, while the critic is augmented with privileged information during training. This structured design safeguards the quality of learned representations, thereby enhancing policy stability under challenging perceptual conditions, similar to how animals rely on proprioception as a stable backbone while incorporating visual feedback to navigate complex terrains.

A. Actor–Critic Architecture

We employ an **Asymmetric Actor–Critic Architecture**, in which the critic is granted access to privileged information, whereas the actor operates exclusively on proprioceptive observations together with the kinesthetic and visuospatial latent representations provided by the estimator. The asymmetric design improves value estimation and stabilizes policy updates, thereby enhancing both sample efficiency and robustness, while maintaining a deployable observation-only actor policy. By unifying perception and control within a single-stage framework, our approach avoids the error accumulation inherent to teacher–student frameworks, while directly optimizing task performance under encoder observations. We optimize the policy using proximal policy optimization (PPO) [32], resulting in an efficient and stable training process.

TABLE I
Reward Components and Weights ($dt = 0.02$).

Reward Terms	**Equation** (r_i)	**Weight** (w_i)
Locomotion Objectives		
$r_{\text{tracking},xy}$	$\phi\big(v_{xy} - v_{xy}^{cmd}\big)$	$3dt$
$r_{\text{tracking},yaw}$	$\phi\big(\omega_{yaw} - \omega_{yaw}^{cmd}\big)$	$1.5dt$
Behavioral Constraints and Penalties		
$r_{\text{velocity},z}$	v_z^2	$-0.1dt$
$r_{\text{ang. vel.},xy}$	$\boldsymbol{\omega}_{xy}^2$	$-0.05dt$
$r_{\text{joint acc.}}$	$\ddot{\boldsymbol{\theta}}^2$	$-2.5 \times 10^{-7}dt$
$r_{\text{joint power}}$	$\lvert\boldsymbol{\tau}\rvert\lvert\dot{\boldsymbol{\theta}}\rvert$	$-2 \times 10^{-5}dt$
$r_{\text{joint torque}}$	$\boldsymbol{\tau}^2$	$-1 \times 10^{-5}dt$
$r_{\text{power dist.}}$	$\text{var}\,\lvert\boldsymbol{\tau} \cdot \dot{\boldsymbol{\theta}}\rvert$	$-2 \times 10^{-7}dt$
$r_{\text{collision}}$	$-n_{\text{collision}}$	$-1dt$
$r_{\text{action rate}}$	$(\mathbf{a}_t - \mathbf{a}_{t-1})^2$	$-0.01dt$
$r_{\text{smoothness}}$	$(\mathbf{a}_t - 2\mathbf{a}_{t-1} + \mathbf{a}_{t-2})^2$	$-0.01dt$

1) Actor Network: Our actor network is provided with a 45-dimensional proprioceptive observation $\mathbf{o}_t$, a 31-dimensional Kinesthetic latent embedding, and a 20-dimensional Visuospatial latent embedding. The proprioceptive observation $\mathbf{o}_t$ is defined as:

$$\mathbf{o}_t = \begin{bmatrix}\boldsymbol{\omega}_t & \mathbf{g_t} & \mathbf{c_t} & \boldsymbol{\theta}_t & \dot{\boldsymbol{\theta}}_t & \mathbf{a}_{t-1}\end{bmatrix}^T, \tag{1}$$

where $\boldsymbol{\omega}_t$ denotes the body angular velocity, $\mathbf{g_t}$ is the gravity direction vector expressed in the body frame, and $\mathbf{c_t}$ represents the velocity command. $\boldsymbol{\theta}_t$ and $\dot{\boldsymbol{\theta}}_t$ correspond to joint positions and joint velocities, respectively, while $\mathbf{a}_{t-1}$ denotes the action produced by the actor network in the previous timestep.

2) Critic Network: The critic network leverages privileged observations obtained in the simulation environment to accurately estimate the state value, thereby facilitating the update of actor parameters. The input to the critic network at time step t is defined as:

$$\mathbf{s}_t = \begin{bmatrix}\mathbf{v}_t & \mathbf{o}_t & \mathbf{f}_t^{xy,z} & \mathbf{m}_t^b\end{bmatrix}^T, \tag{2}$$

where $\mathbf{v}_t$ denotes the base velocity, $\mathbf{f}_t^{xy,z}$ represents eight contact forces of the four feet in both x–y plane and z direction, and $\mathbf{m}_t^b$ represents the local height scans around the base.

3) Reward: To demonstrate the effectiveness of our framework, we adopt a commonly used reward design and parameter settings for legged locomotion control tasks [23], [24] without extensive task-specific customization or fine-tuning. The details of each reward terms are comprehensively summarized in Table I.

4) Action: Our framework employs low-level position control, where the actor network outputs a 12-dimensional vector $\mathbf{a}_t$ at each timestep, which is added to the predefined default joint positions $\boldsymbol{\theta}^{\text{default}}$ to obtain the desired joint positions for all joints $\boldsymbol{\theta}_t^{\text{target}}$, as shown below

$$\boldsymbol{\theta}_t^{\text{target}} = \boldsymbol{\theta}^{\text{default}} + \mathbf{a}_t. \tag{3}$$

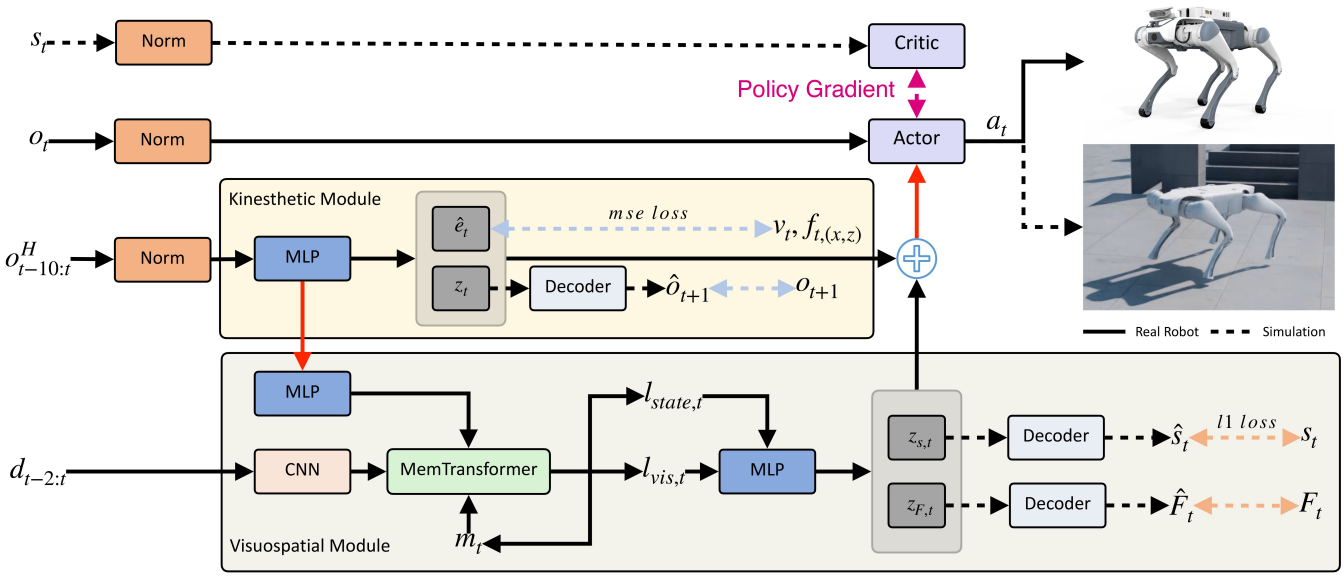


Fig. 2. Overview of the **KiVi** framework. Our bio-inspired dual-branch estimator consists of the **Kinesthetic Module** (highlighted in yellow) and the **Visuospatial Module** (highlighted in gray), focusing on proprioceptive information and the integration of visual inputs, respectively. Solid lines indicate components that are deployed on the real robot, while dashed lines denote parts used only during simulation training. Red lines represent gradient blocking between modules during training.

The desired joint positions are then fed into a low-level PD controller to compute the target joint torques $\boldsymbol{\tau}_t$, which are defined as:

$$\boldsymbol{\tau}_t = K_p \left(\boldsymbol{\theta}_t^{\text{target}} - \boldsymbol{\theta}_t\right) - K_d \, \dot{\boldsymbol{\theta}}_t, \tag{4}$$

where the stiffness K_p and the damping K_d are set to 30.0 and 1.0, respectively.

B. Bio-inspired Dual-branch Estimator

To enable robust multimodal perception, we design a **bio-inspired dual-branch estimator** that emulates the complementary roles of proprioception and vision in animal locomotion. Biological systems rely on proprioception as a dependable backbone for balance and coordination, while selectively incorporating visual information to perceive external obstacles and terrain in a dynamic and adaptive manner. Inspired by this principle, our estimator separates sensory processing into two dedicated pathways.

The **Kinesthetic Module** processes a short sequence of history proprioceptive observations to infer internal dynamics, producing a latent representation that captures the current and the next state of the robot. In parallel, our **Visuospatial Module** integrates egocentric depth images with temporal context to generate a compact terrain embedding, representing surrounding elevations and local foothold height scan.

The two latent representations are then concatenated and passed to the actor network, providing a structured and task-aligned perceptual input. This design reduces the difficulty of representation learning, mitigates the impact of visual disturbances by relying on solid proprioception, and enhances stability and adaptability across diverse environments.

1) Kinesthetic Module: Our Kinesthetic Module leverages short histories of proprioceptive observations $\boldsymbol{o}_{t-10:t}^{H}$ to infer both explicit and implicit dynamic representations of the robot. The explicit outputs $\hat{\boldsymbol{e}}_t \in \mathbb{R}^{11}$ include the estimated base linear velocity and the estimated foot contact forces in the x and z directions. These quantities are supervised with an MSE loss against privileged ground truth values $\boldsymbol{e}_t$ in simulation, providing reliable estimation of the robot's state and contact conditions that are otherwise difficult to measure directly with onboard sensors. In parallel, our module generates an implicit latent vector $\boldsymbol{z}_t \in \mathbb{R}^{20}$ that predicts the next-step observation $\hat{\boldsymbol{o}}_{t+1}$ in a VAE-style manner, encouraging the network to capture latent dynamic priors beyond the explicitly supervised targets.

By jointly capturing the explicit estimate $\hat{\boldsymbol{e}}_t$ and the implicit prediction $\boldsymbol{z}_t$, our Kinesthetic Module provides the policy with stable and temporally consistent internal state representations. This design not only strengthens short-term dynamics modeling but also enables robust self-assessment under noisy proprioceptive readings or rapidly changing contact conditions, thereby enhancing both the robustness and generalization of the learned locomotion policy.

2) Visuospatial Module: Our Visuospatial Module focuses on fusing visual and proprioceptive information to infer key terrain-related features, such as surrounding elevations and foot-level height map, which cannot be directly derived from proprioception alone. This process is challenging as the robot cannot always directly observe the ground beneath its feet, making effective multimodal integration and temporal memory essential for reliable terrain understanding.

To address these challenges, our module takes as input the past 10 steps of proprioceptive observations $\boldsymbol{o}_{t-10:t}^{H}$ and the past 2 frames of egocentric depth images $\boldsymbol{d}_{t-2:t}$. The proprioceptive sequence is normalized and encoded by an MLP into a single token, while the depth sequence is processed by a convolutional neural network (CNN) to produce 16 visual tokens. These tokens are then fused within a memory-augmented Transformer (**memTransformer**). At each timestep, the Transformer receives a total of 20 tokens:

1 proprioceptive token $l_{state,t} \in R^{32}$, 16 visual tokens $l_{vis,t} \in R^{16\times32}$, and 3 memory tokens $m_t \in R^{3\times32}$ inherited from the previous timestep. Our memTransformer consists of two stacked Transformer layers with a single attention head (i.e., $N_{\text{head}} = 1$) and a feedforward dimension of 128. After attention processing, the 20 output tokens are handled asymmetrically. The updated proprioceptive token is directly propagated to downstream policy layers. The 16 visual tokens are average-pooled into a single visual embedding before being forwarded. In contrast, the 3 memory tokens are not passed to the policy head; instead, they are cached and reused as memory inputs at the next timestep. This design allows the model to maintain a compact, content-addressable memory buffer while keeping the policy input dimension fixed. Compared with recurrent models such as GRUs, this mechanism provides a more flexible balance between parameter efficiency and representational capacity. A small number of persistent memory tokens can selectively accumulate long-horizon contextual information without enforcing strict sequential compression as in recurrent hidden states. Moreover, by avoiding recursive hidden-state updates and reducing long backpropagation chains through time, our memTransformer improves training stability and sample efficiency, leading to faster convergence in long-horizon tasks.

Through this mechanism, our module implicitly predicts the global surrounding elevation $\hat{s}_t$ and the local terrain heights around each foot $\hat{F}_t$, and produces latent embeddings $z_{s,t} \in \mathbb{R}^{12}$ and $z_{F,t} \in \mathbb{R}^8$ that encode both coarse terrain structure and fine-grained foothold surroundings. The predictions $\hat{s}_t$ and $\hat{F}_t$ are supervised with a stricter L_1 loss against privileged terrain information available in simulation. By jointly leveraging visual and proprioceptive cues and maintaining temporal memory, our Visuospatial Module provides stable and informative terrain embeddings, complementing the Kinesthetic Module and enabling reliable locomotion control even under partial or noisy visual observations.

The outputs of our Kinesthetic and Visuospatial Modules, namely $\hat{\boldsymbol{e}}_t$, $\boldsymbol{z}_t$, $\boldsymbol{z}_{s,t}$, and $\boldsymbol{z}_{f,t}$, are concatenated into a single vector and fed into the actor network. The overall loss of the estimator is defined as:

$$\mathcal{L} = D_{\mathrm{KL}}\big(q(\boldsymbol{z}_t \mid \boldsymbol{o}^H_{t-10:t}, \boldsymbol{d}_{t-2:t}) \,\|\, p(\boldsymbol{z}_t)\big) + \mathrm{MSE}(\hat{\boldsymbol{e}}_t, \boldsymbol{e}_t) + \mathrm{MSE}(\hat{\boldsymbol{o}}_{t+1}, \boldsymbol{o}_{t+1}) + \ell_1(\hat{\boldsymbol{s}}_t, \boldsymbol{s}_t) + \ell_1(\hat{\boldsymbol{F}}_t, \boldsymbol{F}_t), \quad (5)$$

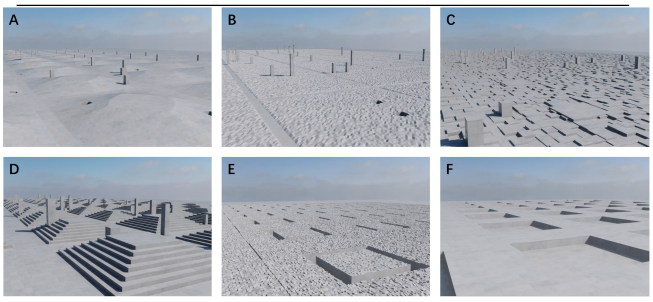
where $q(\boldsymbol{z}_t \mid \boldsymbol{o}^H_{t-10:t}, \boldsymbol{d}_{t-2:t})$ denotes the posterior distribution of $\boldsymbol{z}_t$ conditioned on $\boldsymbol{o}^H_{t-10:t}$ and $\boldsymbol{d}_{t-2:t}$, while $p(\boldsymbol{z}_t)$ represents the prior distribution of $\boldsymbol{z}_t$ which is typically parameterized as a standard normal distribution.

C. Training details

1) Simulation Platform: Leveraging NVIDIA Isaac Sim and Isaac Lab, we train 4096 agents in parallel in simulation; our final policy can be directly deployed on Deeprobotics Lite3 robot. Training converges after approximately 12,000 iterations, about 8 hours on a single NVIDIA RTX 4090.

TABLE II
DOMAIN RANDOMIZATION RANGES.

Parameter	Randomization range	Unit
Payload	$[-1, 3]$	kg
K_p factor	$[0.9, 1.1]$	Nm/rad
K_d factor	$[0.9, 1.1]$	Nms/rad
Center of mass shift	$[-50, 50]$	mm
Static friction coefficient	$[0.5, 1.25]$	-
Dynamic friction coefficient	$[0.3, 1.1]$	-
Initial joint positions	$[0.5, 1.5]$	rad
System delay	$[0, 20]$	ms
Camera shaking	$[-2, 2]$	deg



G	A	B	C	D	E	F
Terrain Type	Slope	Random Rough	Boxes	Stair	Gap	High Wall
Range ($°/m$)	$[0, 20]$	$[0.02, 0.10]$	$[0.05, 0.20]$	$[0.05, 0.23]$	$[0.05, 0.70]$	$[0.05, 0.55]$
Parameter	slope inclination	noise amplitude	obstacle height	step height	gap width	step height
Proportion	0.2	0.1	0.1	0.2	0.2	0.2

Fig. 3. Simulated terrain types A–F used during training, each representing a distinct terrain challenge. As the training difficulty increases, each terrain randomly generates 0–5 obstacles. Figure G lists the control parameters and their respective ranges for each terrain type, which are used to procedurally generate diverse terrain instances.

2) Training Terrain and Curriculum: In simulation, we train the policy simultaneously on six types of terrains: stairs, platforms, random rough, slope, gaps, and high walls. Among them, stairs, platforms, random rough and slope are native terrains provided by Isaac Lab, while gaps and high walls are custom-designed terrains. To prevent policy collapse in the early stages of training due to overly challenging terrains, we adopt a curriculum [4] that progressively increases difficulty as the policy improves. The specific terrain and the curriculum settings are shown in Fig. 3.

3) Domain Randomization: We also use Domain Randomization to enhance policy robustness and facilitate smooth sim-to-real transfer. Our system design ensures that strong performance can be achieved by relying only on standard domain randomization techniques. On the proprioceptive side, we follow a common setting that includes the randomization of the payload, the center of mass, the PD gains (K_p, K_d) and the ground friction coefficient. On the visual side, aside from standard sensor noise, we introduce camera shaking to randomize camera position, simulating sensor vibrations during deployment. The specific randomization settings are summarized in Table II.

IV. EXPERIMENTS

To systematically evaluate the robustness of our **KiVi** framework, we conduct comparative studies, including both ablation experiments and comparisons against representative baselines. These evaluations comprehensively cover algorithmic and functional aspects, providing a holistic understanding of system performance:

- **KiVi w/o Kin.**: A variant where the velocity v_t and the predicted observation $\hat{o}_{t+1}$ are produced from a fused visual–proprioceptive representation, rather than through explicitly separated pathways. This configuration is structurally similar to the representation design adopted in PIE [30], and serves as a controlled baseline to evaluate the contribution of modality separation to robustness.
- **KiVi w/o Memory**: A variant where the MemTransformer is replaced by a standard Transformer, to study the effect of temporal memory in long-horizon tasks.
- **KiVi GRU**: A variant in which the MemTransformer is replaced by a standard Transformer followed by a GRU module. This configuration is designed to assess the effectiveness of the MemTransformer architecture compared to a conventional Transformer+GRU temporal modeling scheme.
- **Himloco [33]**: A blind locomotion policy relying solely on proprioception, which we include to highlight the role of vision in locomotion.

All experiments were conducted on the **DeepRobotics Lite3** quadruped equipped with a RealSense D435i depth camera. Image preprocessing and policy inference were run onboard on a Jetson Orin NX Super. The system operated with depth acquisition at 10 Hz, policy inference at 50 Hz, and a low-level PD controller at 200 Hz.

A. Simulation Experiments

We first evaluate the robustness under severe visual disturbances in simulation. **KiVi**, **KiVi w/o Kin.**, and **Himloco** were deployed on random rough terrains with corrupted depth inputs, including high-intensity Gaussian noise, large random occlusions, and severe camera jitter. These disturbances are considerably stronger than those encountered during training.

During the experiments, we recorded both the total motor power of joints and the variance of the power distribution across joints (Fig. 4). As expected, the blind policy **Himloco** was unaffected by visual corruption, showing consistently low mean power and minimal variance, indicating well-balanced actuation. In contrast, **KiVi w/o Kin.** exhibited large power fluctuations and sharp peaks, reflecting severe joint oscillations with high torques and velocities. This unstable behavior leads to higher energy consumption, motor overheating, and increased mechanical wear. In comparison, the full **KiVi** framework maintained stable operation with energy consumption close to Himloco, demonstrating its robustness and suitability for long-term deployment in disturbed environments.

B. Hardware Experiments

We further evaluated all methods on the robot in outdoor scenarios (Fig. 5 and Fig. 7). The experimental results are summarized in Table III.

1) Terrain Traversability: In comparative studies, **KiVi** and **KiVi w/o Kin.** showed nearly identical performance: both traversed uneven terrains and avoided pedestrians at near-limit distances (~ 3 cm), comparable to **ABS** [25] on flat ground, while exhibiting more stable control on stairs.

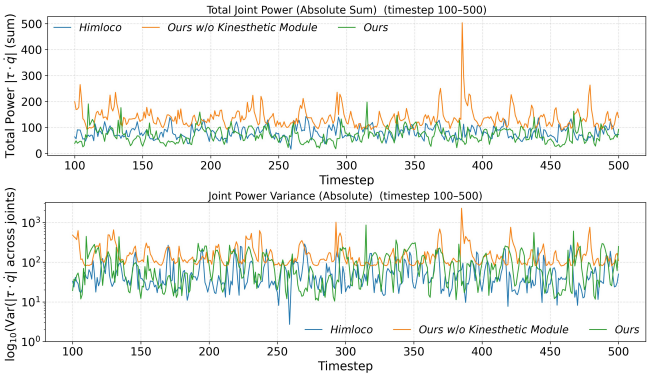


Fig. 4. Total joint power and power variance across all joints for KiVi, KiVi w/o Kin., and Himloco on simulated rough terrains under severe visual disturbances.

TABLE III

PERFORMANCE COMPARISON ON-ROBOT BETWEEN OUR METHOD AND THE BASELINES ACROSS ROBUSTNESS TESTS (5 TRIALS PER TEST).

	High Platform	Obstacle Avoidance	Tall Grass	Block Camera
KiVi	**5/5**	**5/5**	**5/5**	**5/5**
KiVi w/o Kin. (PIE)	4/5	**5/5**	3/5	0/5
KiVi GRU	4/5	**5/5**	4/5	3/5
KiVi GRU (24000 iters)	**5/5**	**5/5**	**5/5**	**5/5**
KiVi w/o Memory	3/5	2/5	2/5	0/5
Himloco	0/5	0/5	**5/5**	/

KiVi GRU, which replaces the memTransformer with a Transformer+GRU structure, achieved comparable success rates in obstacle avoidance but showed slightly reduced robustness on high platforms and tall grass under the same training budget. Notably, when trained for extended iterations (24000 iterations), KiVi GRU eventually reached performance parity with **KiVi**. This suggests that while recurrent compression can ultimately capture sufficient temporal context, our proposed memory-token mechanism facilitates faster convergence and more efficient temporal credit assignment.

KiVi w/o Memory, however, lacks temporal context, which results in more conservative behaviors and reduced stability. **Himloco**, which does not use vision, responds only after physical contact occurs. As a result, it frequently stumbles at the edge of platforms or collides with obstacles. It also fails to traverse challenging terrains such as gaps and high steps. These results highlight the importance of both multimodal fusion and temporal memory for safe operation in dynamic outdoor environments.

2) Visual Robustness: We then tested environments with strong visual disturbances, including abrupt illumination changes, tall grass, reflective surfaces, and complete camera occlusion. Random illumination noise had minor overall impact, but structured disturbances, such as tall grass, provided persistent, misleading semantic cues that directly conflicted with proprioceptive feedback, creating challenging OOD scenarios unseen in training. Unlike random noise, which is largely filtered in network representations, structured noise forced inconsistent interpretations of the terrain, greatly

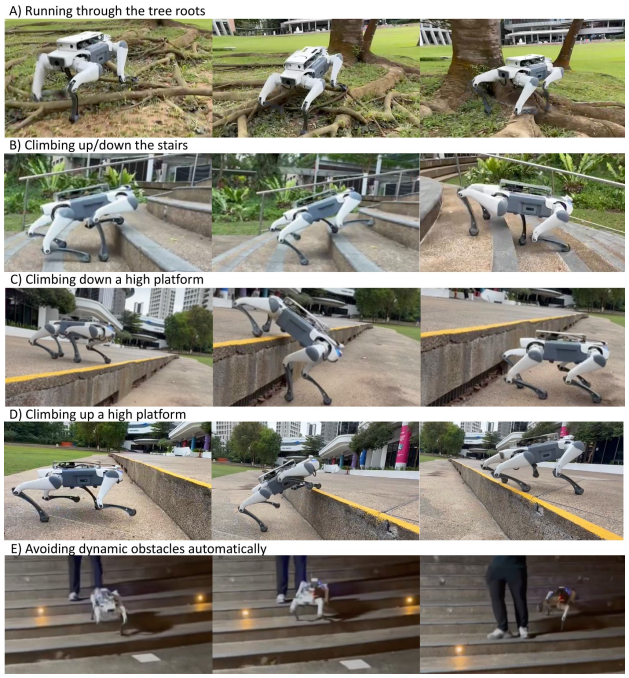


Fig. 5. Outdoor hardware experiments under low visual disturbances, including tree roots, staircases, elevated platforms, and dynamic pedestrians. With only a constant forward velocity command of $[1.0, 0, 0]$, the robot traversed all terrains and avoided obstacles.

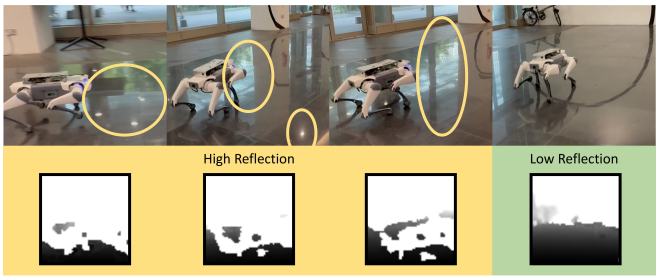


Fig. 6. Locomotion on reflective surfaces. Strong reflections (yellow) distort the depth image, causing the sensor to misinterpret the ground as distant empty space, while the green region indicates weaker artifacts. Despite these distortions, KiVi maintains stable locomotion.

increasing locomotion difficulty.

In these settings, **KiVi** consistently completed all tasks with only slight reductions in gait stability. For example, in tall grass (Fig. 7A), the robot exited the disturbed area and immediately adjusted its gait to climb onto an elevated platform. Finally, when vision was completely occluded (Fig. 7B), the policy gracefully fell back to proprioception, generating conservative yet stable motions resembling blind locomotion. This demonstrates the value of modality separation: proprioception provides a reliable backbone for safe fallback, even when vision becomes entirely unreliable/unavailable. By contrast, **KiVi w/o Kin.** suffered from unstable behaviors due to entangled representations, occasionally producing abrupt pitching. We believe these findings highlight that explicit modality separation not only enhances robustness to corrupted vision but also enables graceful degradation under extreme conditions, whereas mixed representations remain vulnerable to OOD conflicts.

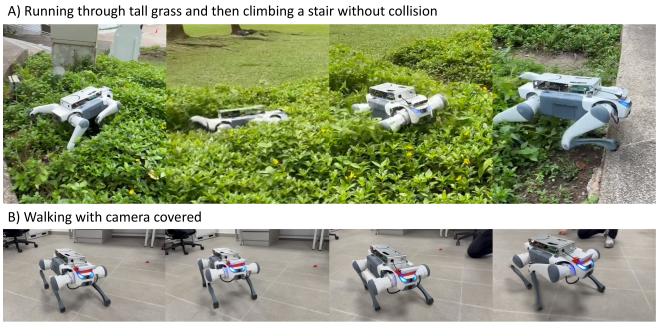


Fig. 7. Performance of KiVi under strong visual disturbances. (**A**) Traversing tall grass and adapting gait to climb a platform. (**B**) Camera fully covered, where the policy seamlessly falls back to proprioception for stable locomotion.

V. Conclusion

In this work, we introduced **KiVi**, a framework for robust quadruped locomotion in visually challenging environments. By explicitly separating proprioceptive and visual pathways and enhancing multi-modal fusion with a memory-augmented transformer, KiVi maintains stable locomotion by relying on proprioception as a backbone, while selectively exploiting visual cues to enhance terrain awareness and anticipatory obstacle avoidance whenever visual perception is reliable. Our results show that this design achieves strong resilience even under unstructured visual noise, such as tall grass, reflective surfaces, and camera occlusion, where fused-latent or vision-dominant baselines often fail. At the same time, our framework exhibits a graceful degradation property: when vision becomes unreliable, control naturally falls back to proprioception without catastrophic failure.

Despite these advantages, KiVi has certain limitations. The current memory module captures only short-term temporal information, which may be insufficient for tasks requiring longer persistence; for instance, when the robot pauses in front of an elevated platform, the memory may decay and misinterpret the obstacle. Extending the memory mechanism or integrating hierarchical reasoning could address this issue. Looking ahead, we believe that KiVi's core principles, including modality separation, conditional integration of visual input, and memory-enhanced fusion, provide a solid foundation for robust multimodal control, with potential to extend beyond remote controlled locomotion to long-term autonomous deployment in real-world applications.

Acknowledgment

We used ChatGPT to assist in the linguistic refinement of the Introduction and Conclusion sections. The generated text was critically reviewed and revised by the authors to ensure alignment with the research findings and academic standards.

We thank Dr. Chao Li of DEEP Robotics for his valuable assistance with the real-world experiments and for providing hardware support.

This work was supported in part by the Singapore Ministry of Education (MOE), the National University of Singapore under its Robotics Grand Challenge, the "Leading Goose" R&D Program of Zhejiang under Grant 2023C01177, the National Key R&D Program of China under Grant

2022YFB4701502, and the 2035 Key Technological Innovation Program of Ningbo City under Grant 2024Z300.

REFERENCES

[1] 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.

[2] G. B. Margolis and P. Agrawal, "Walk these ways: Tuning robot control for generalization with multiplicity of behavior," in *Conference on Robot Learning*, pp. 22–31, PMLR, 2023.

[3] B. Hu, S. Shao, Z. Cao, Q. Xiao, Q. Li, and C. Ma, "Learning a faster locomotion gait for a quadruped robot with model-free deep reinforcement learning," in *2019 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pp. 1097–1102, IEEE, 2019.

[4] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning*, pp. 91–100, PMLR, 2022.

[5] J. Long, J. Ren, M. Shi, Z. Wang, T. Huang, P. Luo, and J. Pang, "Learning humanoid locomotion with perceptive internal model," 2024.

[6] T. Miki, J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter, "Learning robust perceptive locomotion for quadrupedal robots in the wild," *Science robotics*, vol. 7, no. 62, p. eabk2822, 2022.

[7] W. Yu and et al., "Visual-locomotion: Learning to walk on complex terrains with vision," in *Proceedings of CoRL*, vol. 164 of *PMLR*, pp. 992–1003, 2022.

[8] S. Li, S. Luo, J. Wu, and Q. Zhu, "Move: Multi-skill omnidirectional legged locomotion with limited view in 3d environments," *arXiv preprint arXiv:2412.03353*, 2024.

[9] H.-J. Moon, H.-S. Kim, J.-H. Park, and J.-H. Park, "Proprioception, the regulator of motor function," *Brain and NeuroRehabilitation*, vol. 14, no. 3, p. e26, 2021.

[10] J. S. Matthis, J. L. Yates, and M. M. Hayhoe, "The many roles of vision during walking," *Experimental Brain Research*, vol. 238, no. 12, pp. 2851–2864, 2020.

[11] T. Akay and A. J. Murray, "Relative contribution of proprioceptive and central inputs to locomotion: lessons from invertebrates and vertebrates," *Frontiers in Neural Circuits*, vol. 15, p. 635952, 2021.

[12] M. S. Burtsev, Y. Kuratov, A. Peganov, and G. V. Sapunov, "Memory transformer," 2021.

[13] D. J. Hyun, "High speed trot-running: Implementation of a hierarchical controller on the mit cheetah," tech. rep., MIT, 2014.

[14] M. Hutter, C. Gehring, D. Jud, A. Lauber, D. Bellicoso, V. Tsounis, C. Gramazio, A. Yapici, A. Fedoseev, R. Siegwart, and M. Hutter, "Anymal – a highly mobile and dynamic quadrupedal robot," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 38–45, 2016.

[15] J. Di Carlo, P. M. Wensing, B. Katz, G. Bledt, and S. Kim, "Dynamic locomotion in the mit cheetah 3 through convex model-predictive control," in *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pp. 1–9, IEEE, 2018.

[16] D. Kim, J. Di Carlo, B. Katz, G. Bledt, and S. Kim, "Highly dynamic quadruped locomotion via whole-body impulse control and model predictive control," *arXiv preprint arXiv:1909.06586*, 2019.

[17] Y. Ding, A. Pandala, and H.-W. Park, "Real-time model predictive control for versatile dynamic motions in quadrupedal robots," in *2019 International Conference on Robotics and Automation (ICRA)*, pp. 8484–8490, IEEE, 2019.

[18] X. B. Peng, E. Coumans, T. Zhang, T.-W. Lee, J. Tan, and S. Levine, "Learning Agile Robotic Locomotion Skills by Imitating Animals," in *Proceedings of Robotics: Science and Systems*, (Corvallis, Oregon, USA), July 2020.

[19] P. Li, H. Li, G. Sun, J. Cheng, X. Yang, G. Bellegarda, M. Shafiee, Y. Cao, A. Ijspeert, and G. Sartoretti, "Sata: Safe and adaptive torque-based locomotion policies inspired by animal learning," *arXiv preprint arXiv:2502.12674*, 2025. Equal contribution: P. Li and H. Li; Corresponding author: Y. Cao.

[20] Z. Wang, W. Wei, R. Yu, J. Wu, and Q. Zhu, "Toward understanding key estimation in learning robust humanoid locomotion," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 11232–11239, 2024.

[21] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," 2022.

[22] I. Radosavovic, T. Xiao, B. Zhang, T. Darrell, J. Malik, and K. Sreenath, "Real-world humanoid locomotion with reinforcement learning," *Science Robotics*, vol. 9, no. 89, p. eadi9579, 2024.

[23] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," *arXiv preprint arXiv:2107.04034*, 2021.

[24] I. M. A. Nahrendra, B. Yu, and H. Myung, "Dreamwaq: Learning robust quadrupedal locomotion with implicit terrain imagination via deep reinforcement learning," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5078–5084, IEEE, 2023.

[25] T. He, C. Zhang, W. Xiao, G. He, C. Liu, and G. Shi, "Agile but safe: Learning collision-free high-speed legged locomotion," *arXiv preprint arXiv:2401.17583*, 2024.

[26] A. Agarwal, A. Kumar, D. Pathak, and J. Malik, "Legged locomotion in challenging terrains using egocentric vision," in *Conference on Robot Learning (CoRL)*, 2022.

[27] Z. Zhuang, K. Caluwaerts, A. Iscen, and J. Tan, "Robot parkour learning," in *Conference on Robot Learning (CoRL)*, 2023.

[28] X. Cheng, Z. Luo, Z. Fu, and D. Pathak, "Extreme parkour with legged robots," in *Conference on Robot Learning (CoRL)*, 2023.

[29] D. Hoeller, L. Wellhausen, J. Carius, and M. Hutter, "Anymal parkour: Learning perception-enhanced locomotion skills," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[30] S. Luo, S. Li, R. Yu, Z. Wang, J. Wu, and Q. Zhu, "Pie: Parkour with implicit-explicit learning framework for legged robots," *IEEE Robotics and Automation Letters*, 2024.

[31] 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 *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 11531–11537, 2025.

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

[33] J. Long, Z. Wang, Q. Li, J. Gao, L. Cao, and J. Pang, "Hybrid internal model: Learning agile legged locomotion with simulated robot response," *arXiv preprint arXiv:2312.11460*, 2023.