LocoVLM

Versatile Legged Locomotion Adaptation through Vision-Language Grounding

I Made Aswin Nahrendra anahrendra@kaist.ac.kr

Seunghyun Lee kevin9709@kaist.ac.kr

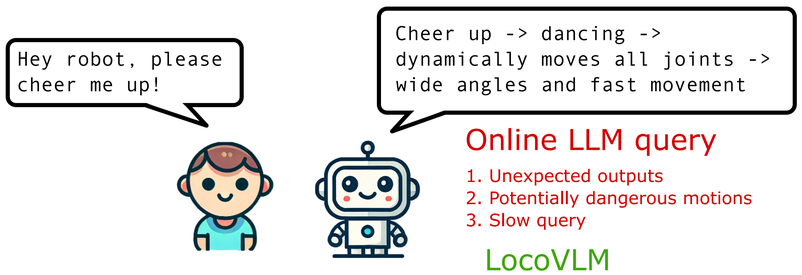
Dongkyu Lee dklee@urobotics.ai

Hyun Myung* hmyung@kaist.ac.kr

URBAN ROBOTICS LAB
https://urobot.kaist.ac.kr
School of Electrical Engineering, KAIST, Republic of Korea

ICRA IEEE INTERNATIONAL CONFERENCE ON ROBOTICS AND AUTOMATION
ATLANTA 2025

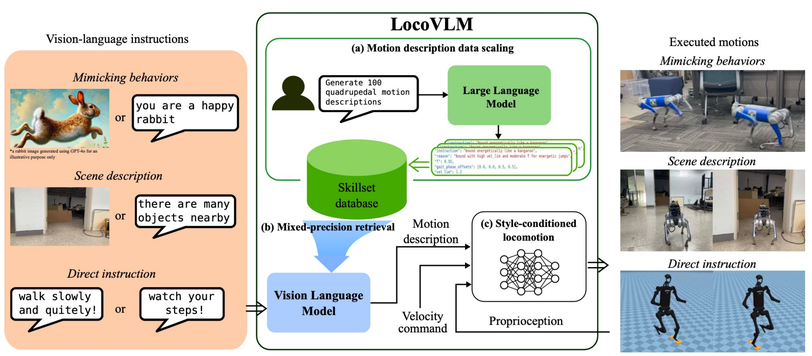
Research Motivation



How can we instruct robot behaviors on-the-fly without compromising safety?

Overview

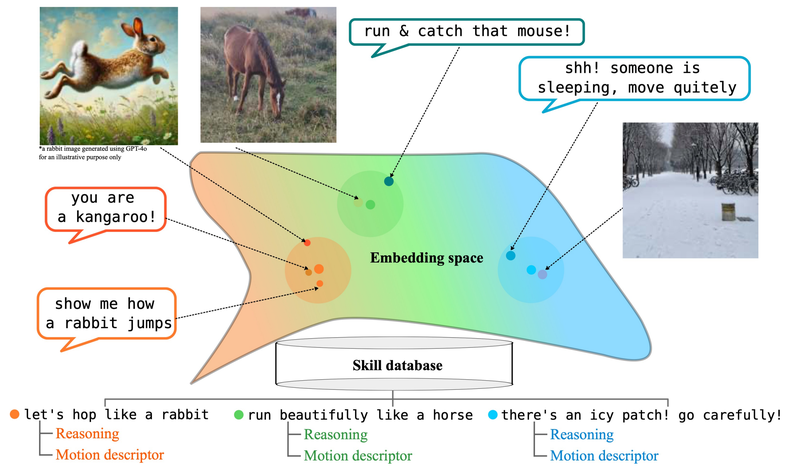
- A **style-conditioned locomotion** framework with a compliant gait tracking that enhances safety and robustness of gait tracking in cluttered terrains.
- **Prevents uncontrolled LLM inference** during deployment to ensure safe and robust locomotion adaptation through vision and language.

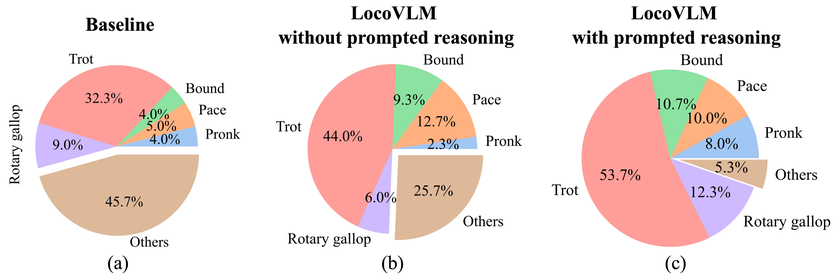


https://locovlm.github.io https://urobot.kaist.ac.kr

Methodology

Scaling-up skill database





The proposed method reduces vague and unsafe locomotion gaits typically generated by online LLM queries

Mixed-precision retrieval

- LocoVLM employs BLIP-2 VLM to perceive the vision-language input and retrieve the closest skill from the database with an accuracy of **up to 87%**.

Algorithm 1 Mixed-Precision Retrieval

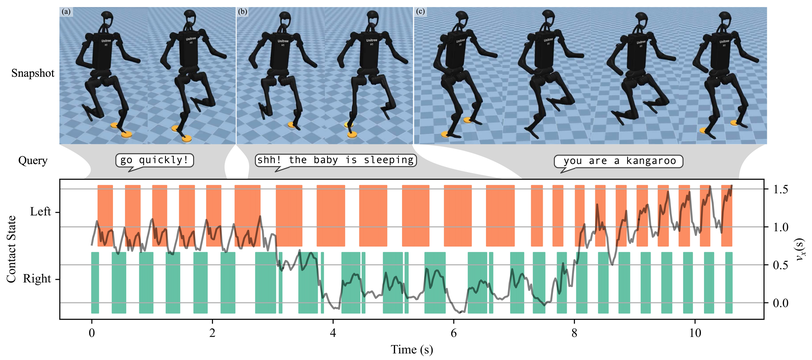
Input: $\mathcal{I}_{\text{query}}$, $\mathcal{D}$, f_{BLIP}, f_{ITM}, K
$\mathbf{I}^K \leftarrow \arg\max_{\mathcal{I} \in \mathcal{D}} \text{cossim}\left(f_{\text{BLIP}}(\mathcal{I}_{\text{query}}), f_{\text{BLIP}}(\mathbf{I})\right)$
$p_1(\mathbf{I}^K) \leftarrow \text{softmax}\left(\text{cossim}\left(f_{\text{BLIP}}(\mathcal{I}_{\text{query}}), f_{\text{BLIP}}(\mathbf{I}^K)\right)\right)$
for $\mathcal{I}^k \in \mathbf{I}^K$ **do**
  $p_2(\mathcal{I}_k) \leftarrow \text{softmax}\left(f_{\text{ITM}}(\mathcal{I}_{\text{query}}, \mathcal{I}_k)\right)$
end for
$\mathcal{I}^* \leftarrow \arg\max_{\mathcal{I}_k} \left(p_1(\mathbf{I}^K) + p_2(\mathbf{I}^k)\right)$
Output: $\mathcal{I}^*$

RETRIEVAL ACCURACY OF LOCOVLM USING DIFFERENT RETRIEVAL METRICS FOR 100 INSTRUCTIONS FROM THE DATABASE.

Retrieval Metric	Text as String	Text as Image	Average
Cosine similarity	21/100	30/100	20.5%
Top-K similarity	27/100	48/100	37.5%
Top-K to ITM	51/100	57/100	54.0%
Mixed-precision	**72/100**	**87/100**	**79.5%**

Experimental Results

Real-time language instruction following



- LocoVLM adapts the robot's locomotion either **using direct or vague language instructions**.

Responding to robot-centric vision

- LocoVLM guides the quadruped robot to **skulk like a lynx** in a snowy environment, increasing safety against slippery snow

Conclusion

- LLM-constructed database allows us to filter out potentially dangerous behaviors
- An offline VLM model is able to accurately match real-time instructions to behaviors in a static database