

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

Research Motivation

Hey robot, please cheer me up!

Cheer up -> dancing -> dynamically moves all joints -> wide angles and fast movement

Online LLM query

1. Unexpected outputs
2. Potentially dangerous motions
3. Slow query

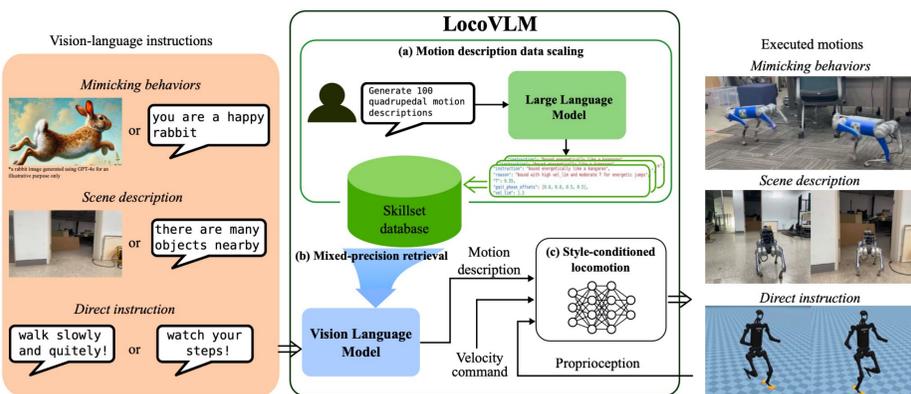
LocoVLM

1. Controlled and filtered behaviors
2. Safe and bounded to know behaviors
3. Real-time

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.



Methodology

Scaling-up skill database

The methodology involves scaling up a skill database. It shows an 'Embedding space' where different skills are represented. Examples include 'run & catch that mouse!', 'shh! someone is sleeping, move quietly', 'you are a kangaroo!', and 'show me how a rabbit jumps'. A 'Skill database' is used to retrieve skills based on these queries. The process involves 'Reasoning' and 'Motion descriptor'.

Legend for pie charts:

- Red: Reasoning Motion descriptor
- Green: Reasoning Motion descriptor
- Blue: Reasoning Motion descriptor

Category	Baseline (a)	LocoVLM without prompted reasoning (b)	LocoVLM with prompted reasoning (c)
Rotary gallop	9.0%	6.0%	12.3%
Trot	32.3%	44.0%	53.7%
Bound	4.0%	9.3%	10.7%
Pace	5.0%	12.7%	10.0%
Pronk	4.0%	2.3%	8.0%
Others	45.7%	25.7%	5.3%

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

```

1: Input:  $\mathcal{I}_{query}$ ,  $\mathcal{D}$ ,  $f_{BLIP}$ ,  $f_{ITM}$ ,  $K$ 
2:  $\mathbf{I}^K \leftarrow \arg \max_{\mathcal{I} \in \mathcal{D}} \text{cosim}(f_{BLIP}(\mathcal{I}_{query}), f_{BLIP}(\mathbf{I}))$ 
3:  $p_1(\mathbf{I}^K) \leftarrow \text{softmax}(\text{cosim}(f_{BLIP}(\mathcal{I}_{query}), f_{BLIP}(\mathbf{I}^k)))$ 
4: for  $\mathcal{I}^k \in \mathbf{I}^K$  do
5:    $p_2(\mathcal{I}^k) \leftarrow \text{softmax}(f_{ITM}(\mathcal{I}_{query}, \mathcal{I}^k))$ 
6: end for
7:  $\mathcal{I}^* \leftarrow \arg \max_{\mathcal{I}^k} (p_1(\mathbf{I}^k) + p_2(\mathbf{I}^k))$ 
8: Output:  $\mathcal{I}^*$ 

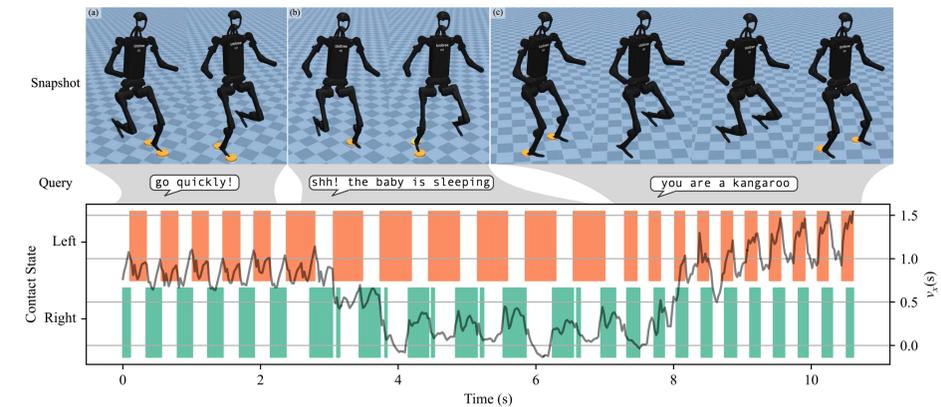
```

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

The figure shows 'Responding to robot-centric vision'. It includes 'Image queries' of a snowy environment. Below the images are 'Retrieved motion descriptions' for different instructions. For example, the instruction 'traipse lightly like a deer' results in a gait phase offset of [0.0, 0.5, 0.5, 0.0] and a velocity limit of 0.5. The instruction 'skulk with stealth like a lynx' results in a gait phase offset of [0.0, 0.5, 0.5, 0.0] and a velocity limit of 0.2.

- 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

