AnySkill: Learning Open-Vocabulary Physical Skill for Interactive Agents

Jieming Cui[1,2,*], Tengyu Liu[2,*], Nian Liu[2], Yaodong Yang[1], Yixin Zhu[1,2,✉], Siyuan Huang[2,✉]

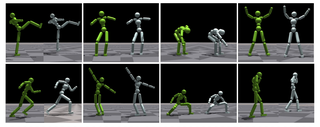
[1] Institute for Artificial Intelligence, Peking University, [2] State Key Laboratory of General Artificial Intelligence, BIGAI

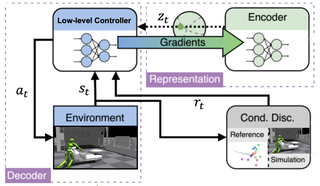
CVPR SEATTLE, WA JUNE 17-21, 2024

https://anyskill.github.io/

Low-level Algorithm

- ✓ **93** modified reference motions
- ✓ **Semantic** latent space





$$\mathcal{L}_{\mathcal{D}} = -\mathbb{E}_{M\in\mathcal{M}}\Big(\mathbb{E}_{d^{\pi}(s,s'|z)}\left[\log\left(1-\mathcal{D}(s,s'|z)\right)\right] + \mathbb{E}_{d^{\mathcal{M}}(s,s')}\left[\log\mathcal{D}(s,s'|z)+\log\left(1-\mathcal{D}(s,s'|z'\sim\mathcal{Z})\right)\right] + w_{\text{gp}}\mathbb{E}_{d^{\mathcal{M}}(s,s')}\left[||\nabla_{\theta}\mathcal{D}(\theta)|_{\theta=(s,s'|z)}||^2\right]\Big|\hat{z}=\text{sg}(E(M))\Big)$$

$$r^{L}(s,s',z) = -\log\left(1-\mathcal{D}\left(s,s'|z\right)\right)$$

High-level Algorithm

- ✓ **Only** image-based reward
- ✓ **Agent-centered** image rendering
- ✓ **LLM** enhances open-vocabulary capabilities.

Input: Reference motion dataset $\mathcal{M}$, frozen low-level controller π^L, frozen motion encoder E, simulation environment ENV, renderer image $\mathcal{I}$, CLIP feature of the description text f_d

$\mathcal{Z} = E(\mathcal{M})$ initialize motion latent space
while *not converged* **do**
  $\mathcal{B} \leftarrow \emptyset$; $p \leftarrow 0$ initialize
  for *horizon_length* $= 1, ..., n$ **do**
    sample $\hat{z}$ from $\mathcal{Z}$
    if *horizon_length* $= 1$ **then**
      $s \leftarrow$ initialize; $z \leftarrow \hat{z}$
    else
      $s \leftarrow \text{ENV}(s, a)$; $z \leftarrow \pi^H(s)$
    end
    for *llc_steps* $= 1, ..., t$ **do**
      $s \leftarrow \text{ENV}(s, \pi^L(s, z))$ step simulation
      $r^H \leftarrow$ calculate reward with Eq. (3)
      if HEAD_HEIGHT < 0.15 **then**
        $s, p \leftarrow 0$ reset agent and counter
      end
      if similarity is less than last step **then**
        $p \leftarrow p + 1$ increment counter
        if $p \geqslant 8$ **then**
          $p \leftarrow 0$ reset counter
          reset s with 80% probability
        end
      end
    end
    update $\mathcal{B}$ and π^H according to PPO
  end
end

ANY Text Description

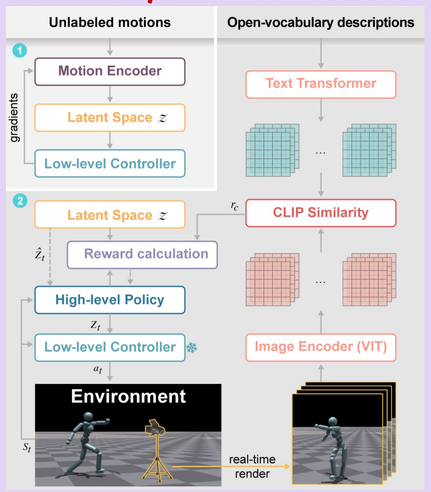
a. Sit down
b. Dance
c. Kick
d. Jump rope

Iterate text enhancement

1. Describe one specific human pose.
2. Ensure the verb and noun relate to a human limb.
3. Make the description fluent, concise, and unambiguous. ...

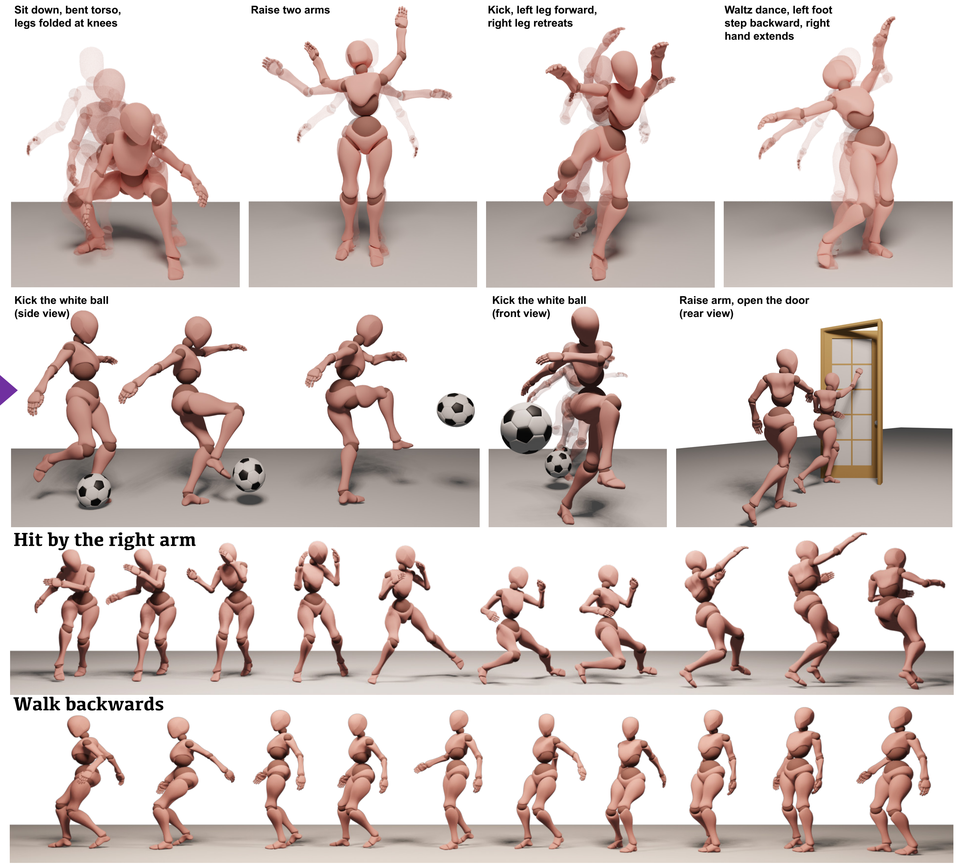
a. Sit down, bent torso, legs folded at knees
b. Dance, left foot step backward, right hand extends
c. Kick, left leg forward, right leg retreats
d. Jump rope, legs off the ground, wave hands

Model w/o manual reward

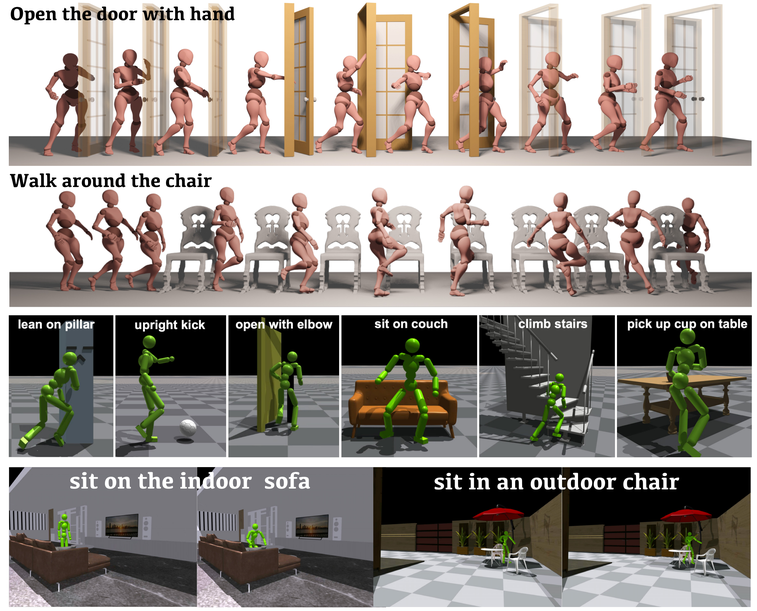


Contributions

- ✓ AnySkill is a **hierarchical approach** designed for learning open-vocabulary physical skills.
- ✓ We create flexible, generalizable, **image-based** rewards, **eliminating** the need for **manual** design.
- ✓ Our method outperforms existing approaches and enables agents to **interact smoothly with dynamic objects** in various contexts.



Interaction with object & scene



Quantitative Results

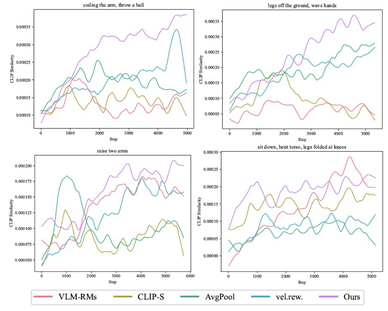


Table 1. **Quantitative evaluation of high-level policy.**

	Success ↑	Natural ↑	Smooth↑	Physics↑	CLIP_S↑
AvatarCLIP [11]	4.29	4.74	5.79	5.74	21.11
MotionCLIP [43]	3.16	4.93	5.72	5.83	21.16
Ours (w/o ET)	5.05	4.88	5.68	5.31	21.89
Ours (w/o text-enhance)	3.06	4.48	5.19	5.96	20.76
Ours (w/ VideoCLIP [56])	2.37	4.90	5.65	6.41	21.35
Ours (full)	**6.16**	**6.23**	**6.51**	**6.93**	**24.18**

Table 2. **Quantitative evaluation of interaction motions.**

	Success ↑	Natural ↑	Smooth↑	Physics↑	CLIP_S↑
Interaction w. object	5.42 -0.74	5.62 -0.61	5.34 -1.17	5.45 -1.48	24.49 +0.35
Interaction w. scene	4.53 -1.63	4.47 -1.76	5.01 -1.50	5.41 -1.52	22.41 -1.73

Table 3. **Comparisons of the reward design.**

	Success↑	Natural↑	Smooth↑	Physics↑	CLIP_S↑
VLM-RMs [37]	3.15	4.36	5.35	5.17	19.46
CLIP-S [69]	3.80	5.41	5.98	6.21	19.78
AvgPool [56]	5.09	5.96	**6.55**	6.70	20.25
+ vel. rew. [32]	2.73	4.42	5.35	5.22	18.39
Ours	**6.16**	**6.23**	6.51	**6.93**	**24.18**