GROVE: A Generalized Reward for Learning Open-Vocabulary Physical Skill

Jieming Cui[1,2*], Tengyu Liu[2*], Ziyu Meng[2,3], Jiale Yu[4], Ran Song[3], Wei Zhang[3],
Yixin Zhu[1,✉], Siyuan Huang[2,✉]

[1] Institute for Artificial Intelligence, Peking University [2] State Key Laboratory of General Artificial Intelligence, BIGAI
[3] School of Control Science and Engineering, Shandong University [4] Department of Automation, Tsinghua University

https://jiemingcui.github.io/grove/

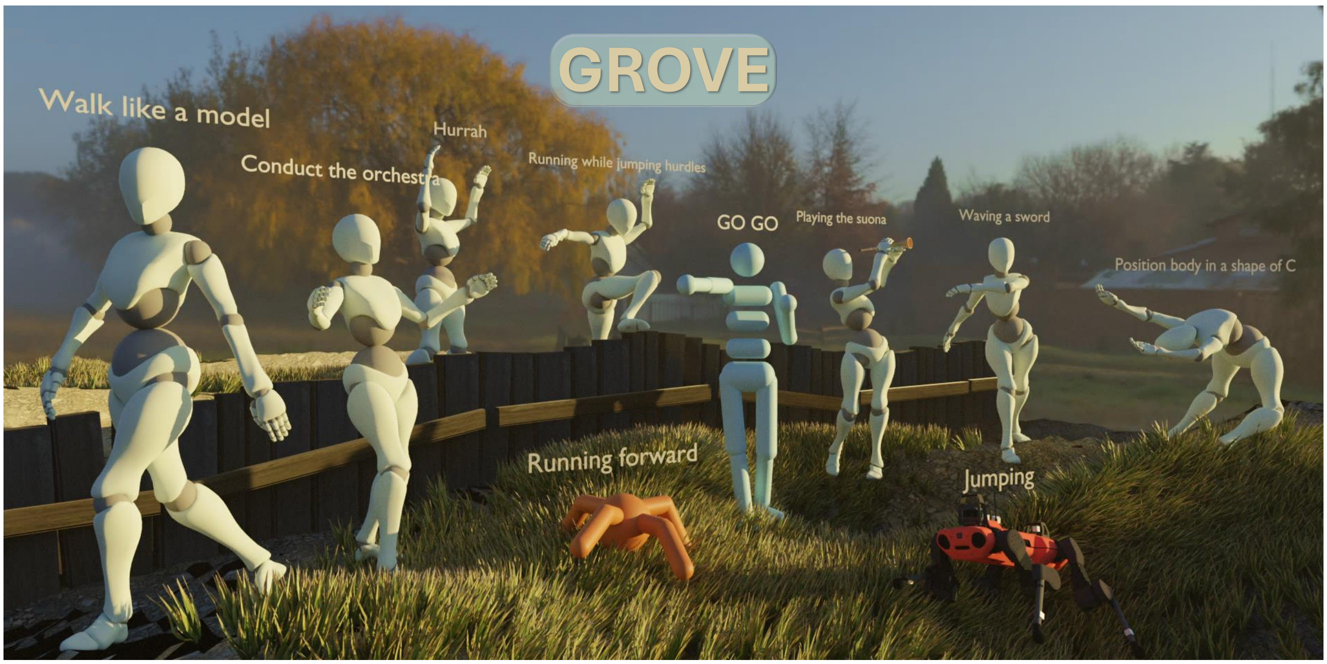


Figure 1. **Open-vocabulary physical skills learned with our generalized reward framework GROVE.** The white humanoid mannequins demonstrate diverse skills from abstract instructions (*e.g.*, "conduct the orchestra," "position body in a shape of C") using a pre-trained controller. The system also generalizes to standard RL benchmarks, including the Ant, Humanoid, and quadrupedal ANYmal, all without task-specific reward engineering.

Abstract

Learning open-vocabulary physical skills for simulated agents presents a significant challenge in Artificial Intelligence (AI). Current Reinforcement Learning (RL) approaches face critical limitations: manually designed rewards lack scalability across diverse tasks, while demonstration-based methods struggle to generalize beyond their training distribution. We introduce GROVE, a generalized reward framework that enables open-vocabulary physical skill learning without manual engineering or task-specific demonstrations. Our key insight is that Large Language Models (LLMs) and Vision Language Models (VLMs) provide complementary guidance—LLMs generate precise physical constraints capturing task requirements, while VLMs evaluate motion semantics and naturalness. Through an iterative design process, VLM-based feedback continuously refines LLM-generated constraints, creating a self-improving reward system. To bridge the domain gap between simulation and natural images, we develop Pose2CLIP, a lightweight mapper that efficiently projects agent poses directly into semantic feature space without computationally expensive rendering. Extensive experiments across diverse embodiments and learning paradigms demonstrate GROVE's effectiveness, achieving 22.2% higher motion naturalness and 25.7% better task completion scores while training 8.4× faster than previous methods. These results establish a new foundation for scalable physical skill acquisition in simulated environments.

1. Introduction

Enabling simulated agents to learn diverse physical skills from natural language instructions represents a fundamental challenge in AI for graphics and robotics. Despite significant advances in RL, developing agents that perform arbitrary physical tasks specified through natural language remains difficult due to the traditional reliance on manually engineered reward functions [43]. For example, training a humanoid to *run forward* involves a complex reward system accounting for velocity, energy consumption, survival, and more [29]. This tailored approach limits the generalizability of RL methods to open-vocabulary tasks.

The central challenge lies in developing generalizable reward mechanisms that can both interpret arbitrary natural language instructions and accurately evaluate complex motion patterns. While imitation learning methods offer partial generalizability [15, 16, 45], they remain fundamentally constrained by their text-motion training distributions. Recent approaches have made progress by leveraging LLMs [2, 5, 46] and VLMs [3, 35] for automatic reward design. However, they each face distinct limitations:

LLM-based approches excel at generating precise, dynamic constraints from natural language instructions [10, 26], providing detailed reward signals for specific kinematic aspects such as joint positions, velocities, and spatial relationships. However, they struggle with holistic motion assessment, focusing on individual constraints while missing stylistic qualities and natural movement patterns. This limitation frequently results in technically correct but unnatural motions requiring substantial human refinement [40, 42, 47]. Conversely, **VLM-based approaches** provide rich semantic feedback through visual evaluation [8, 37], effectively assessing whether actions "look correct" perceptually. While offering the robust evaluation of stylistic correctness and naturalness that LLMs typically miss, these approaches struggle with temporal consistency across motion sequences. They cannot enforce the precise kinematic constraints necessary for skill mastery.

We propose a generalized reward framework, `GROVE`, that leverages the complementary strengths of both sides. By combining LLM-generated constraints with VLM-based semantic evaluation, `GROVE` enables robust open-vocabulary physical skill learning. Intuitively, these two reward sources provide orthogonal constraints on the RL search space: LLM-generated rewards impose precise physical and kinematic constraints, while VLM-based rewards constrain the solutions to semantically valid and natural-looking motions. This orthogonality ensures the combined reward effectively narrows the search space to solutions satisfying physical accuracy and semantic correctness.

A significant technical challenge is bridging the visual domain gap between simplified simulation environments and the natural images on which VLMs are trained. We address this with `Pose2CLIP`, a lightweight model that efficiently maps agent poses directly to CLIP feature space without rendering. Trained on 1.7 million frames of high-quality rendered human poses, this mapping enables efficient semantic evaluation while preserving CLIP's rich representational capabilities and dramatically reducing computational requirements during training.

Extensive experiments demonstrate `GROVE`'s effectiveness across diverse embodiments and tasks in two distinct settings: (i) integrating with hierarchical controllers pretrained on human motion datasets and (ii) guiding standard RL algorithms learning from scratch; see also Fig. 1. Our evaluations show that combining LLM-generated constraints with VLM-based feedback converges 8.4x faster and produces more natural movements than using either reward source independently. User studies confirm these improvements, showing a 22.2% enhancement in perceived motion naturalness and a 25.7% increase in task completion scores compared to baseline methods.

To summarize, our contributions are four-fold:

- We introduce `GROVE`, a generalized reward framework that combines LLM-generated precise constraints with VLM-based semantic evaluation, enabling robust open-vocabulary physical skill learning without manual reward design or task-specific demonstrations.
- We develop `Pose2CLIP`, a lightweight pose-to-semantic feature mapper that bridges the domain gap between simulation and natural images, enabling efficient and effective semantic evaluation.
- We show `GROVE`'s capability to generate diverse and natural motions for arbitrary embodiments from open-vocabulary instructions. Through extensive experiments on complex humanoid tasks and standard RL benchmarks, our method achieves 25.7% higher task completion score, 22.2% improved motion naturalness, and 8.4× faster convergence compared to baselines.
- We demonstrate the generalizability of our proposed reward across multiple scenarios, from leveraging pretrained controllers to learning from scratch, and provide comprehensive ablation studies showing the complementary benefits of combining LLM and VLM rewards.

2. Related Work

Physical skill learning and reward design Research in physical skill learning has followed two primary approaches to address the challenges of reward design. Traditional RL methods rely on carefully engineered reward functions decomposed into multiple weighted components [1, 4, 18, 19, 31, 36, 41, 50]. While these approaches achieve impressive results for specific tasks, their dependence on manual tuning and domain expertise limits scalability to new behaviors.

Adversarial imitation learning offers an alternative by learning reward functions directly from demonstrations through discriminator networks [15, 16, 22, 24, 32, 33, 44, 49]. These approaches capture complex motion qualities without explicit reward engineering but are constrained by their reliance on extensive demonstration data and limited generalization beyond the training distribution. In contrast, our work leverages foundation models to provide complementary supervision signals that eliminate the need for both manual reward design and task-specific demonstrations.

Language-guided skill learning Foundation models have significant advanced zero-shot skill learning through two complementary approaches: LLM-based reward generation and VLM-based semantic evaluation.

LLM-based methods [26, 27, 48, 51] translate natural language instructions into programmatic reward functions by prompting language models to formulate physical constraints and success criteria. Recent work like Eureka [26] demonstrated that models such as GPT can generate precise rewards for controlling joint positions, velocities, and spatial relationships without human intervention. However, these approaches typically lack a holistic understanding of motion quality and cannot visually assess complex movement patterns. VLM-based methods [8, 37] use vision-language models to provide semantic feedback on generated motions. These approaches excel at determining whether actions appear natural and contextually appropriate. However, they face technical barriers including the substantial domain gap between simplified simulation renderings and natural images, challenges in maintaining temporal consistency across frames, and the inability to enforce precise physical constraints. Our approach, `Pose2CLIP`, bridges the domain gap while combining the strengths of both model types—LLMs for precise constraint formulation and VLMs for holistic semantic evaluation.

Alternative approaches to zero-shot skill acquisition Beyond reward-based approaches, recent work explores alternative paradigms for zero-shot physical skill acquisition. Representation learning methods [25, 45] focus on building task-agnostic motion embeddings from demonstration data that capture underlying motion structure. For example, MaskedMimic [45] achieves zero-shot generalization by training models to reconstruct strategically masked portions of motion sequences. However, these approaches struggle to interpret open-vocabulary language instructions.

Code generators [12, 20] use LLMs to translate instructions directly into executable robot commands, excelling in structured environments with well-defined actions. However, they face substantial challenges with dynamic full-body control and complex physical interactions. They primarily succeed in scenarios with inherent physical stability (*e.g.*, table-mounted manipulators) and rely heavily on decomposing tasks into discrete, atomic actions. Our approach differs by creating a unified reward mechanism that handles both the precise kinematics and semantic correctness required for complex whole-body movements without requiring predefined action primitives.

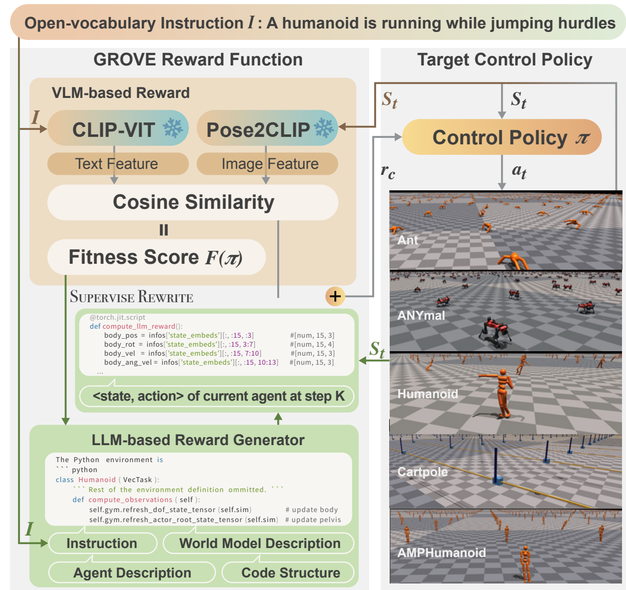


Figure 2. **The overall architecture of `GROVE`.** `GROVE` combines multiple components for zero-shot physical skill learning: a VLM-based reward for semantic evaluation, a `Pose2CLIP` mapper to bridge the simulation-to-reality gap, an LLM-based reward generator for precise constraints, and an iterative reward design process. We demonstrate `GROVE`'s effectiveness across five diverse agents: Ant, ANYmal, Humanoid, Cartpole, and AMPHumanoid. We demonstrate its effectiveness across five diverse agents: Ant, ANYmal, Humanoid, Cartpole, and AMPHumanoid.

3. The `GROVE` Framework

We introduce `GROVE`, a generalized reward framework that enables learning open-vocabulary physical skills without manual reward engineering or task-specific demonstrations. At its core, `GROVE` combines two complementary reward mechanisms: a VLM-based component that evaluates the semantic correctness of performed actions, and an LLM-based component that formulates precise physical constraints. The framework uniquely leverages a VLM to dynamically evaluate and re-sample rewards throughout the learning process (see Fig. 2), creating a synergistic effect that addresses the limitations of each approach. The complete algorithmic implementation is detailed in Algorithm 1.

3.1. Generalized Reward

We formalize `GROVE` as a Reward Design Problem (RDP) [39], building upon the framework introduced in Eureka [26]. An RDP is defined as a tuple $P = \langle \xi, \mathcal{R}, F \rangle$, where $\mathcal{R}$ is the space of all possible reward functions, and F is a fitness function that maps a policy π to a real-valued fitness score $F(\pi)$. The world model $\xi = \langle S, A, T \rangle$ en-

Algorithm 1: GROVE for physical skill learning

Input: Task instruction I, world model ξ, agent description D_{agent}, pre-trained VLM, LLM

Hyper-parameters: horzion length K

Initialize policy π; $R_L \leftarrow \text{LLM}(\xi, D_{\text{agent}}, I)$

while *not converged* **do**

  Initialize $\mathcal{B} \leftarrow \emptyset$; $F(\pi) = []$; state s

  for *step* $= 1, ..., K$ **do**

    # Standard RL operations

    $a \leftarrow \pi(s),\ \ s' \leftarrow \text{ENV}(s, a),\ \ r \leftarrow R_{\text{GROVE}}(s', a, I)$

    # Update fitness function

    $F(\pi).\text{append}(R_V(s', I))$

    if 8 consecutive drops on $\text{F}(\pi)$ & $\text{F}(\pi)_{\text{last}} < 0.1$ **then**

      # Re-generate reward function

      $R_L \leftarrow \text{LLM}(\xi, D_{\text{agent}}, I)$

    end

    # Update policy and state variable

    update $\mathcal{B}$ and π according to PPO

    $s \leftarrow s'$

  end

end

capsulates the state space S, action space A, and transition function T that governs the environment dynamics.

Given an arbitrary natural language instruction I, our key insight is that effective reward functions should combine both semantic understanding and precise physical constraints. We therefore design the GROVE reward as a weighted combination of complementary components:

$$R_{\text{GROVE}}(s, a, I) = \omega_V R_V(s, I) + \omega_L R_L(s, a; I), \quad (1)$$

where R_V represents the VLM-based semantic evaluation component, R_L denotes the LLM-generated constraint component, and weights ω_V and ω_L balance their relative importance. Critically, R_{GROVE} is dynamically refined during training according to the fitness function F, enabling continuous improvement in instruction interpretation.

VLM-based reward Our VLM-based reward leverages a frozen CLIP model [35] for semantic alignment between agent behaviors and natural language instructions. Traditionally, this would require computing CLIP similarity between rendered frames and instruction text—a computationally expensive process that often suffers from the sim-to-real gap. To address these limitations, we introduce Pose2CLIP, a specialized neural network that directly maps agent poses to the CLIP feature space without requiring explicit rendering. This design dramatically improves computational efficiency while also bridging the sim-to-real gap by training on high-fidelity Blender-rendered images that better match CLIP's training distribution of natural images. The VLM-based reward is formally computed as:

$$R_V(s, I) = \frac{\text{CLIP}(I) \cdot \texttt{Pose2CLIP}(s)}{\|\text{CLIP}(I)\|_2 \cdot \|\texttt{Pose2CLIP}(s)\|_2}. \quad (2)$$

A detailed analysis of Pose2CLIP's architecture and training methodology is provided in Sec. 3.2.

LLM-based reward For the generation of task-specific reward functions R_L, we employ *GPT-o1-preview* [2] following the approach established in Eureka [26]. However, we introduce two critical enhancements to the prompt engineering process that improve reward quality. First, we enrich the context by incorporating detailed agent specifications (D_{agent}), including precise joint nomenclature and indices that allow for more targeted reward formulation. Second, we explicitly guide the model with the instruction that auxiliary mechanisms already handle fundamental aspects like stability and locomotion, directing it to "*focus solely on capturing the essence of the task.*" This strategic constraint prevents the model from generating overly complex rewards addressing already-solved problems. The generation and application of the LLM-based reward can be formalized as:

$$R_L(\cdot, \cdot\,; I) = \text{LLM}(\xi, D_{\text{agent}}, I), \quad (3a)$$

$$r_L = R_L(s, a; I), \quad (3b)$$

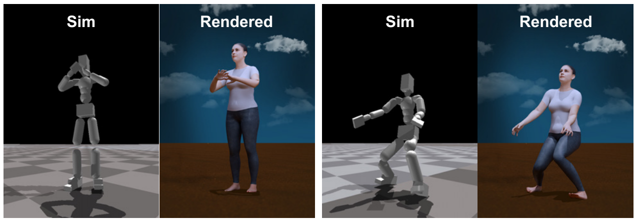
where Eq. (3a) is the generation of the reward function R_L, and Eq. (3b) its evaluation on specific state-action pairs.

RDP and fitness function While LLM-based rewards offer detailed temporal awareness and precise constraint specification, they can suffer from instability and occasional semantic drift [17, 26]. To mitigate this limitation, we implement a quality control mechanism using our VLM-based reward R_V as a fitness evaluator for R_L. Specifically, we employ an adaptive rejection sampling strategy: the LLM-based reward function is regenerated whenever we detect a consistent performance decline, defined as 8 consecutive steps where the average R_V across all parallel environments decreases relative to the previous step and the final average falls below a threshold (0.1). This verification ensures that optimization toward R_L consistently results in visually and semantically correct behaviors, preventing the learning process from diverging toward unintended solutions that might satisfy the letter but not the spirit of the instruction.

3.2. Pose2CLIP

A critical challenge in using VLMs for embodied control lies in the substantial domain gap between simulator-rendered images and the natural images on which models like CLIP were trained. This discrepancy undermines the effectiveness of direct CLIP-based reward computation. While high-fidelity rendering could theoretically bridge this gap, it introduces prohibitive computational overhead that makes real-time RL impractical.

To address this challenge, we introduce Pose2CLIP, a lightweight neural network that directly maps agent pose representations to CLIP's feature space without requiring explicit rendering. This design serves dual purposes: it eliminates the computational burden of high-fidelity ren-



(a) clap (b) sit down

Figure 3. **Comparison of simulator vs. Blender-rendered images for Pose2CLIP training.** The left shows raw simulation, while the right the Blender renderings. Pose2CLIP maps agent poses directly to CLIP feature space, thus bridging the sim-to-real gap and improving reward quality for (a) "clap" and (b) "sit down."

dering (achieving a three-to-six-fold acceleration in our experiments) while preserving CLIP's semantic understanding capabilities. Though our implementation focuses on humanoid skill learning—arguably the most challenging embodiment scenario—the method generalizes to any agent with a fixed state space representation.

For training Pose2CLIP, we construct a diverse dataset combining SMPL [23] poses from the AMASS training split [28] and the Motion-X dataset [21]. Each pose $\theta \in \mathbb{R}^{J\times 3}$ represents the joint-wise Euler angles of the humanoid. To ensure comprehensive coverage of the pose distribution encountered during policy learning, we continuously augment this dataset with policy rollouts throughout the training process, ultimately yielding a refined dataset of 1.7 million frames after downsampling.

The generation of ground truth CLIP features follows a carefully designed process. Using Blender with the SMPL-X Add-on [30], we render each pose with high-quality textures and realistic lighting. To minimize occlusion effects and ensure viewpoint robustness, we render each pose from five strategic angles (front, side, oblique, rear side, and rear) at 45° intervals (see Fig. 3). These multi-view images are then processed through a pre-trained CLIP-ViT-B/32 model [7] to obtain image feature embeddings, with features from different viewpoints of the same pose being mean-pooled to create a single representation.

The Pose2CLIP is intentionally lightweight—a two-layer MLP optimized to minimize MSE reconstruction loss between predicted and ground truth CLIP features. To counter the inherent data imbalance problem, where common poses dominate the dataset while important but rare poses are underrepresented, we implement a balanced sampling strategy. First, we apply k-means++ [6] to cluster the dataset into 500 semantically meaningful groups. Next, we employ a two-stage uniform sampling approach during training: we first sample uniformly across clusters to ensure the representation of all pose types, then sample uniformly within the selected clusters. This hierarchical sampling strategy ensures comprehensive coverage of the entire pose space, enhancing model generalization.

Table 1. **Quantitative evaluation of open-vocabulary humanoid skill synthesis.** We compare GROVE against state-of-the-art approaches in five metrics: task completion (C), motion Naturalness (N), motion smoothness (S), physical realism (P), and CLIP text-image similarity (CLIP_S). Our approach integrating CALM outperforms alternatives in task completion while maintaining competitive or superior performance across other metrics.

	C↑	N↑	S↓	P↑	CLIP_S↑
AvatarCLIP [11]	1.621	4.793	0.411	5.821	22.105
TMR [34]	3.793	5.966	**0.374**	7.684	18.885
MoMask [9]	3.621	6.724	0.492	7.863	22.372
MotionGPT [14]	5.207	6.552	0.885	7.000	23.142
AnySkill [8]	6.108	5.938	0.486	8.168	23.925
Ours (+ CALM) [44]	**7.924**	**6.793**	0.488	**8.452**	**28.998**

4. Experiments

Our experimental evaluation is structured in three parts: Sec. 4.1 assesses GROVE's capability to generate open-vocabulary humanoid skills, Sec. 4.2 evaluates its performance on standard RL benchmarks across different embodiments, and Sec. 4.3 provides an ablation study demonstrating the contribution of each component.

4.1. Open-Vocabulary Humanoid Skill Acquisition

Generating open-vocabulary physical skills for humanoids presents significant challenges in motion synthesis. To fully leverage GROVE's capabilities, we employ a hierarchical control approach using a pre-trained low-level controller from CALM [44], defined as $a = \pi_L(z, s)$, which generates actions a based on the current state s and a latent code z. We then train a high-level policy $z = \pi_H(s; I)$ that optimizes the GROVE reward for instruction I. This approach enables complex movements through instruction-guided coordination of motion primitives.

We conduct both qualitative and quantitative evaluations using five representative open-vocabulary instructions. For qualitative assessment, the resulting motions are visualized in Fig. 4, with additional sequences available on our project website. These demonstrations illustrate our method's ability to generate naturalistic motions that accurately reflect diverse textual instructions.

For quantitative assessment, we employ computational metrics and human judgment. The computational metrics include **Smoothness** (average change in acceleration across joints [52]) and **CLIP similarity** (cosine similarity between rendered frames and instruction text). For human evaluation, 30 participants rated each motion sequence on three dimensions (0-10 scale): **Task completion** (how effectively the motion fulfills the instruction), **Motion naturalness** (biological plausibility), and **Physics** (adherence to physical constraints). Participants underwent calibration with reference examples to ensure rating consistency.

Figure 4. **Open-vocabulary humanoid skill synthesis using GROVE.** Our framework successfully generates physically plausible motions for diverse, previously **unseen** instructions. Each row demonstrates a sequence of frames depicting a different skill: (a) "running while jumping hurdle" showcases dynamic locomotion with obstacle navigation, (b) "conduct the orchestra" demonstrates expressive arm gestures, (c) "position body in a shape of 'C'" illustrates the understanding of abstract concepts, (d) "playing the suona" captures realistic hand positioning, and (e) "walking like a model" exhibits stylized locomotion with characteristic posture and gait patterns.

We benchmark GROVE against state-of-the-art approaches in different paradigms: (i) **Motion retrieval**: TMR [34] uses extensive motion data sets, establishing a standard for naturalness. (ii) **Data-driven generation**: MotionGPT [14] and MoMask [9] represent cutting-edge approaches trained on text-motion paired datasets. (iii) **Zero-shot skill learning**: AnySkill [8] and AvatarCLIP [11] generate motions from unseen text using VLM supervision.

Tab. 1 reveals several insights: (i) TMR achieves high naturalness but lower task completion, indicating our instructions extend beyond existing text-motion dataset distributions. (ii) MoMask and MotionGPT demonstrate strong naturalness but lower task completion, reinforcing the limitations of purely data-driven approaches. (iii) AnySkill, relying exclusively on VLM-based rewards, achieves the highest task completion among baselines but exhibits reduced naturalness, highlighting the inadequacy of single-modal rewards. (iv) GROVE outperforms all baselines in task completion while maintaining naturalness and smoothness, validating our multi-modal reward framework.

Notably, our evaluation presents a more stringent challenge as GROVE operates within physical simulation constraints, while most comparison methods (except AnySkill) produce motions without such constraints.

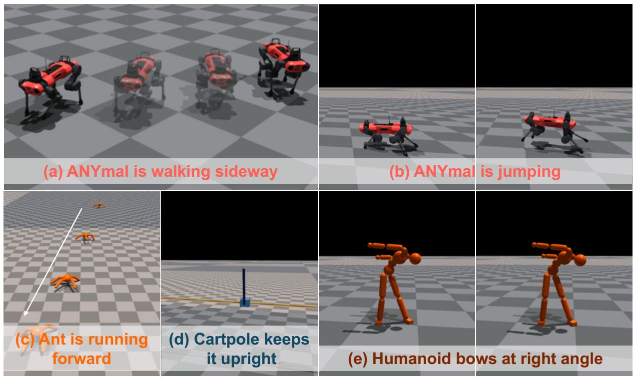


Figure 5. **Qualitative results of GROVE on standard RL benchmarks, requiring to learn diverse control policies across different embodiments and task complexities.** Tasks shown include: ANYmal quadruped (a) moving sideways and (b) jumping, (c) Ant locomotion, (d) Cartpole balancing, (e) and Humanoid performing a bowing motion at a specific angle. Each task presents distinct control challenges requiring different physical capabilities. Additional benchmarks and videos are available on the project website.

4.2. Standard RL Benchmarks

To establish GROVE's generalizability beyond humanoid control, we evaluate its performance across five diverse RL tasks representing varying embodiments and complexity: Cartpole balancing (low-dimensional control), Ant walking (locomotion), ANYmal moving sideways (quadrupedal movement), ANYmal jumping (dynamic maneuver), and Humanoid bowing (articulated coordination), all illustrated in Fig. 5. For all experiments, we use three-layer MLP policies trained with PPO in IsaacGym.

Our evaluation combines human assessment with quantitative metrics. Following the protocol from Sec. 4.1, human evaluators rated task completion across all embodiments, consistently confirming GROVE's effectiveness. For quantitative comparison, we measure learning efficiency using the reward distance metric from [37], capturing both maximum performance and convergence speed (see Appendix C.2).

Compared to a baseline that directly optimizes expert rewards, GROVE demonstrates superior training efficiency in three of four tasks, as shown in Tab. 2. The exception is the jumping task, where the baseline's advantage stems from directly optimizing the evaluation metric itself—a favorable scenario given this task's multiple success criteria.

Table 2. **Quantitative evaluation on standard RL benchmarks.** We compare GROVE against direct reward optimization across different agents and tasks. C represents the human-evaluated completion score, and RD (Reward Distance) learning efficiency. GROVE outperforms direct optimization in three of four tasks.

Agent	Task	C ↑	RD (direct) ↓	RD (GROVE) ↓
Ant	running forward	8.897	103.81	92.45
Cartpole	keep upright	8.862	20.72	20.04
ANYmal	jump up	7.241	57.29	60.84
ANYmal	walk sideways	8.414	39.71	37.51

Table 3. **Comparison with VLM-RM on humanoid bowing task.** T represents training time in minutes, and C human-evaluated completion score. GROVE outperforms VLM-RM while requiring only 11.4% of the training time.

	Task	Textures	T↓	C↑
VLM-RM [37]	Bow	Original	411min	1.655
VLM-RM [37]	Bow	Improved	411min	3.483
GROVE	Bow	Original	**47min**	**6.276**

A particularly revealing comparison is with VLM-RM, which enhances agent textures specifically to improve VLM-based reward effectiveness. As Tab. 3 demonstrates, our method outperforms VLM-RM while using unmodified textures and only requiring 11.4% of their training time. This efficiency differential highlights the advantage of our LLM-based reward formulation over approaches that rely on visual enhancements to the reward model.

These results collectively demonstrate GROVE's robust generalizability across diverse embodied control tasks—from simple to complex embodiments and from standard locomotion to specialized movements. The consistent performance advantages suggest that GROVE's multi-modal reward structure effectively captures the essential characteristics of diverse tasks while providing smoother learning gradients than specialized reward engineering approaches.

4.3. Ablations

To analyze the contribution of each component in our approach, we conduct a comprehensive ablation study with seven model configurations, each isolating or combining different aspects of our full system: (i) **VLM only** uses CLIP features extracted directly from simulation environment images, resembling the AnySkill [8] baseline. (ii) **LLM only w/o RDP** employs the LLM-generated reward function without dynamic refinement, similar to Eureka [26]. (iii) **LLM only (with RDP)** extends this by periodically regenerating the reward function based on fitness scores. (iv) **VLM + LLM** combines both reward modalities with RDP-based resampling of LLM rewards. (v)

Table 4. **Quantitative ablation study results.** We evaluate seven ablative model configurations across multiple metrics: completion quality (C), motion naturalness (N), movement smoothness (S), CLIP similarity (CLIP_S), and training time (T).

Ablations	C ↑	N↑	S↓	CLIP_S↑	T↓
VLM only [8]	6.108	5.938	0.486	23.925	59min
VLM + LLM	6.954	6.292	**0.457**	24.088	47min
LLM only w/o RDP [26]	6.622	5.907	0.582	22.977	20min
LLM only with RDP	6.650	6.177	0.599	24.099	19min
Pose2CLIP only	6.785	5.954	0.553	24.547	16min
Pose2CLIP + LLM w/o RDP	7.269	6.738	0.475	25.331	**7min**
Pose2CLIP + LLM	**7.924**	**6.793**	0.488	**28.998**	**7min**

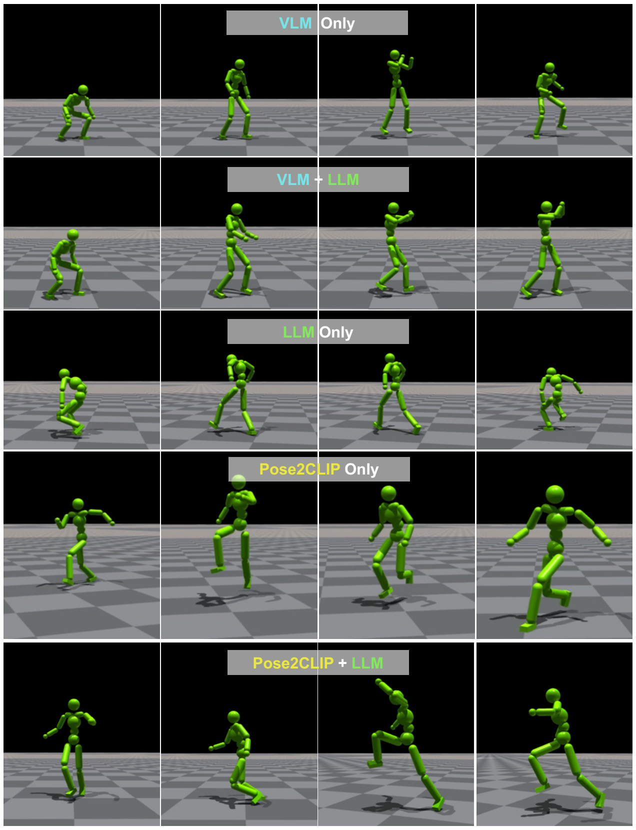


Figure 6. **Qualitative ablation study results.** We compare motion sequences generated by different model configurations for the instruction "*running while jumping hurdles*." The visualization reveals clear performance differences: VLM-Only and LLM-Only approaches produce running motions with minimal vertical displacement, while `Pose2CLIP` configurations show distinct jumping actions with proper elevation and form. The full `Pose2CLIP` + LLM model generates the most natural hurdle-jumping motion, with appropriate preparation, elevation, and landing phases.

`Pose2CLIP` only substitutes standard CLIP features with our `Pose2CLIP` features in the VLM-only setting. (vi) **`Pose2CLIP` + LLM w/o RDP** combines `Pose2CLIP` with a static LLM-based reward generated once at training initialization. (vii) **`Pose2CLIP` + LLM** is our full model.

Qualitative results in Fig. 6 and quantitative evaluations in Tab. 4 reveal three key insights. (i) `Pose2CLIP` feature extraction improves task completion and reduces convergence time compared to direct CLIP features, underscoring the importance of high-quality visual representations for physically embodied tasks. (ii) LLM-only and `Pose2CLIP`-only configurations show complementary strengths—the former producing more natural motions, while the latter better adheres to instructional details—validating our hypothesis that these reward modalities capture different yet essential aspects of performance. (iii) Our full model consistently outperforms all variants across metrics, confirming the synergistic interaction between components and demonstrating that each element effectively addresses limitations of the others.

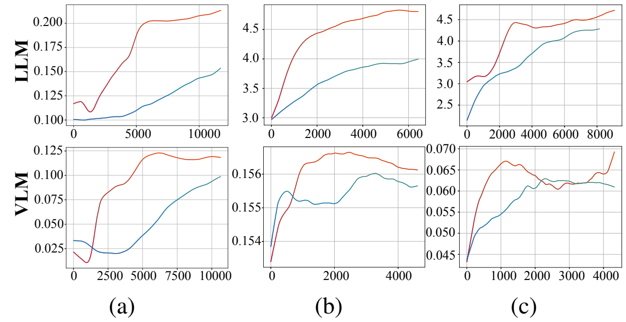


Figure 7. **Reward convergence comparison across tasks.** We compare reward trajectories between our full model (red) and single-reward baselines (blue) across three different tasks. Tasks: (a) arms folded over check; (b) position body in a shape of 'C'; and (c) playing the suona.

The reward trajectories in Fig. 7 further support these findings by comparing our full model (red) against single-reward baselines (blue) across three tasks. The top row displays the LLM reward component, while the bottom row shows the VLM component, with each baseline trained using only its corresponding reward type. These trajectories reveal a consistent pattern: the integration of multi-modal rewards accelerates the convergence of both reward functions compared to their uni-modal counterparts. This acceleration provides compelling evidence that LLM-based and VLM-based rewards offer complementary guidance signals, enabling more efficient skill acquisition.

5. Conclusion

We present `GROVE`, a generalized framework for learning open-vocabulary physical skills that integrates VLM-based reward computation, LLM-based reward generation, pose-to-semantic feature mapping, and dynamic reward programming. Our approach demonstrates significant advantages, improving task completion by 25.7% while converging 8.4× faster than baselines, with comparable motion quality to specialized generative models. Experiments across standard RL benchmarks confirm our method's generalizability, while ablation studies validate that our multimodal approach provides complementary guidance that accelerates skill acquisition. These results establish `GROVE` as an efficient framework for open-vocabulary skill learning in robotic control and motion synthesis.

Acknowledgment This work is supported in part by the National Natural Science Foundation of China (62376031), the Beijing Nova Program, the State Key Lab of General AI at Peking University, the PKU-BingJi Joint Laboratory for Artificial Intelligence, and the National Comprehensive Experimental Base for Governance of Intelligent Society, Wuhan East Lake High-Tech Development Zone.

References

[1] Farzad Abdolhosseini, Hung Yu Ling, Zhaoming Xie, Xue Bin Peng, and Michiel Van de Panne. On learning symmetric locomotion. In *Proceedings of the ACM SIGGRAPH Conference on Motion, Interaction and Games*, 2019. 2

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 4

[3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[4] OpenAI: Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, et al. Learning dexterous in-hand manipulation. *International Journal of Robotics Research (IJRR)*, 39(1):3–20, 2020. 2

[5] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*, 2023. 2

[6] David Arthur and Sergei Vassilvitskii. k-means++: the advantages of careful seeding. In *ACM-SIAM Symposium on Discrete Algorithms*, 2007. 5

[7] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 5

[8] Jieming Cui, Tengyu Liu, Nian Liu, Yaodong Yang, Yixin Zhu, and Siyuan Huang. Anyskill: Learning open-vocabulary physical skill for interactive agents. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 5, 6, 7, A1, A8

[9] Chuan Guo, Yuxuan Mu, Muhammad Gohar Javed, Sen Wang, and Li Cheng. Momask: Generative masked modeling of 3d human motions. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 5, 6, A1

[10] Xu Han, Qiannan Yang, Xianda Chen, Zhenghan Cai, Xiaowen Chu, and Meixin Zhu. Autoreward: Closed-loop reward design with large language models for autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[11] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: zero-shot text-driven generation and animation of 3d avatars. *ACM Transactions on Graphics (TOG)*, 41(4):1–19, 2022. 5, 6, A1

[12] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning (CoRL)*, 2023. 3

[13] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. A1

[14] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 5, 6

[15] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Padl: Language-directed physics-based character control. In *SIGGRAPH Asia 2022 Conference Papers*, 2022. 2, 3

[16] Jordan Juravsky, Yunrong Guo, Sanja Fidler, and Xue Bin Peng. Superpadl: Scaling language-directed physics-based control with progressive supervised distillation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024. 2, 3

[17] Puhao Li, Tengyu Liu, Yuyang Li, Muzhi Han, Haoran Geng, Shu Wang, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Ag2manip: Learning novel manipulation skills with agent-agnostic visual and action representations. In *International Conference on Intelligent Robots and Systems (IROS)*, 2024. 4

[18] Zhongyu Li, Xuxin Cheng, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for robust parameterized locomotion control of bipedal robots. In *International Conference on Robotics and Automation (ICRA)*, 2021. 2

[19] Zhongyu Li, Xue Bin Peng, Pieter Abbeel, Sergey Levine, Glen Berseth, and Koushil Sreenath. Reinforcement learning for versatile, dynamic, and robust bipedal locomotion control. *International Journal of Robotics Research (IJRR)*, page 02783649241285161, 2024. 2

[20] Jacky Liang, Wenlong Huang, Fei Xia, Peng Xu, Karol Hausman, Brian Ichter, Pete Florence, and Andy Zeng. Code as policies: Language model programs for embodied control. In *International Conference on Robotics and Automation (ICRA)*, 2023. 3

[21] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 5

[22] Nian Liu, Libin Liu, Zilong Zhang, Zi Wang, Hongzhao Xie, Tengyu Liu, Xinyi Tong, Yaodong Yang, and Zhaofeng He. Learning uniformly distributed embedding clusters of stylistic skills for physically simulated characters, 2024. 3

[23] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 5

[24] Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 3

[25] Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. *arXiv preprint arXiv:2310.04582*, 2023. 3

[26] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi

Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 4, 7, A6

[27] Yecheng Jason Ma, William Liang, Hungju Wang, Sam Wang, Yuke Zhu, Linxi Fan, Osbert Bastani, and Dinesh Jayaraman. Dreureka: Language model guided sim-to-real transfer. In *Robotics: Science and Systems (RSS)*, 2024. 3

[28] Naureen Mahmood, Nima Ghorbani, Nikolaus F Troje, Gerard Pons-Moll, and Michael J Black. Amass: Archive of motion capture as surface shapes. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2019. 5

[29] Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, et al. Isaac gym: High performance gpu-based physics simulation for robot learning. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 2

[30] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[31] Xue Bin Peng, Pieter Abbeel, Sergey Levine, and Michiel Van de Panne. Deepmimic: Example-guided deep reinforcement learning of physics-based character skills. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 2

[32] Xue Bin Peng, Ze Ma, Pieter Abbeel, Sergey Levine, and Angjoo Kanazawa. Amp: Adversarial motion priors for stylized physics-based character control. *ACM Transactions on Graphics (TOG)*, 40(4):1–20, 2021. 3

[33] Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Transactions on Graphics (TOG)*, 41(4):1–17, 2022. 3

[34] Mathis Petrovich, Michael J Black, and Gül Varol. Tmr: Text-to-motion retrieval using contrastive 3d human motion synthesis. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2023. 5, 6, A1

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of International Conference on Machine Learning (ICML)*, 2021. 2, 4

[36] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017. 2

[37] Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 2, 3, 7, A8

[38] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. A10

[39] Satinder Singh, Richard L Lewis, and Andrew G Barto. Where do rewards come from. In *Annual Meeting of the Cognitive Science Society (CogSci)*, 2009. 3, A6

[40] Guijin Son, Hyunwoo Ko, Hoyoung Lee, Yewon Kim, and Seunghyeok Hong. Llm-as-a-judge & reward model: What they can and cannot do. *arXiv preprint arXiv:2409.11239*, 2024. 2

[41] Seungmoon Song, Łukasz Kidziński, Xue Bin Peng, Carmichael Ong, Jennifer Hicks, Sergey Levine, Christopher G Atkeson, and Scott L Delp. Deep reinforcement learning for modeling human locomotion control in neuromechanical simulation. *Journal of Neuroengineering and Rehabilitation*, 18:1–17, 2021. 2

[42] Shengjie Sun, Runze Liu, Jiafei Lyu, Jing-Wen Yang, Liangpeng Zhang, and Xiu Li. A large language model-driven reward design framework via dynamic feedback for reinforcement learning. *arXiv preprint arXiv:2410.14660*, 2024. 2

[43] Richard S Sutton. *Reinforcement learning: An introduction*. A Bradford Book, 2018. 2

[44] Chen Tessler, Yoni Kasten, Yunrong Guo, Shie Mannor, Gal Chechik, and Xue Bin Peng. Calm: Conditional adversarial latent models for directable virtual characters. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 3, 5

[45] Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting. *ACM Transactions on Graphics (TOG)*, 43(6):1–21, 2024. 2, 3

[46] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 2

[47] Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 2024. 2

[48] Tianbao Xie, Siheng Zhao, Chen Henry Wu, Yitao Liu, Qian Luo, Victor Zhong, Yanchao Yang, and Tao Yu. Text2reward: Reward shaping with language models for reinforcement learning. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2024. 3

[49] Heyuan Yao, Zhenhua Song, Baoquan Chen, and Libin Liu. Controlvae: Model-based learning of generative controllers for physics-based characters. *ACM Transactions on Graphics (TOG)*, 41(6):1–16, 2022. 3

[50] Wenhao Yu, Greg Turk, and C Karen Liu. Learning symmetric and low-energy locomotion. *ACM Transactions on Graphics (TOG)*, 37(4):1–12, 2018. 2

[51] Wenhao Yu, Nimrod Gileadi, Chuyuan Fu, Sean Kirmani, Kuang-Huei Lee, Montse Gonzalez Arenas, Hao-Tien Lewis Chiang, Tom Erez, Leonard Hasenclever, Jan Humplik, Brian Ichter, Ted Xiao, Peng Xu, Andy Zeng, Tingnan Zhang, Nicolas Heess, Dorsa Sadigh, Jie Tan, Yuval Tassa, and Fei Xia. Language to rewards for robotic skill synthesis. In *Conference on Robot Learning (CoRL)*, 2023. 3

[52] Siwei Zhang, Yan Zhang, Federica Bogo, Marc Pollefeys, and Siyu Tang. Learning motion priors for 4d human body capture in 3d scenes. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 5