
Task Tokens: A Flexible Approach to Adapting Behavior Foundation Models

Ron Vainshtein¹, Zohar Rimon¹, Shie Mannor¹, Chen Tessler²

sronv@campus.technion.ac.il, zohar.rimon@campus.technion.ac.il,
ctessler@nvidia.com, shie@ee.technion.ac.il

¹Technion - Israel Institute of Technology

²NVIDIA, Israel

Abstract

Recent advancements in imitation learning have led to transformer-based behavior foundation models (BFMs) that enable multi-modal, human-like control for humanoid agents. While excelling at zero-shot generation of robust behaviors, BFMs often require meticulous prompt engineering for specific tasks, potentially yielding suboptimal results. We introduce "Task Tokens", a method to effectively tailor BFMs to specific tasks while preserving their flexibility. Our approach leverages the transformer architecture of BFMs to learn a new task-specific encoder through reinforcement learning, keeping the original BFM frozen. This allows incorporation of user-defined priors, balancing reward design and prompt engineering. By training a task encoder to map observations to tokens, used as additional BFM inputs, we guide performance improvement while maintaining the model's diverse control characteristics. We demonstrate Task Tokens' efficacy across various tasks, including out-of-distribution scenarios, and show their compatibility with other prompting modalities. Our results suggest that Task Tokens offer a promising approach for adapting BFMs to specific control tasks while retaining their generalization capabilities.

Recent advances in imitation learning have facilitated the emergence of behavior foundation models (BFMs) designed for humanoid control (Peng et al., 2022; Won et al., 2022; Luo et al., 2024a; Tessler et al., 2024). These models, trained on large-scale human demonstration data, can generate diverse, naturalistic behaviors. In this work, we focus on a specific type of BFM, which we call Goal-Conditioned Behavior Foundation Models (GC-BFMs). Methods such as Masked Trajectory Models and MaskedMimic fall into this category (Wu et al., 2023; Tessler et al., 2024). These methods use transformer architectures that process sequences of tokenized goals—high-level objectives such as “follow a path” or “reach with your right hand towards the object” are mapped to embedding tokens that condition the model’s behavior generation. Specifically, we focus on MaskedMimic, which has manifested as a particularly effective framework, demonstrating robust zero-shot generalization (ability to handle new, unseen tasks without additional training) through its token-based goal conditioning mechanism.

Despite MaskedMimic’s proficiency in generating human-like motions from high-level goals, significant challenges persist in defining precise goal specifications, or prompts, for complex tasks. We observe that different aspects of a task are better suited to different forms of specification. Consider a game character tasked with walking to an object and striking it: the approach behavior is naturally expressed through goal-based directives (e.g., orienting the character’s pelvis and head during locomotion), while the precise striking motion is more effectively defined through environment design (e.g., a reward for strike the target). The current approach of crafting precise goal descriptions (prompt engineering) often proves insufficient for tasks that combine such diverse objectives. This

creates a fundamental gap between the model’s ability to generate natural motions and the precise control needed for specialized tasks.

To address this issue, we propose Task Tokens, a novel approach that integrates goal-based control with reward-driven optimization within GC-BFMs like MaskedMimic. Our method, illustrated in Figure 1, establishes a hybrid control paradigm: users provide high-level behavioral priors via goals (e.g., "walk toward the object while facing forward"), while the system autonomously learns task-specific embeddings to optimize dense rewards (e.g., "strike the target with maximum impact"). This integration leverages the inherent tokenization framework of GC-BFMs, enabling a seamless combination of user-defined and learned conditioning tokens.

A key component of our approach is the Task Token encoder, a neural network that maps task observations to supplementary goal tokens, augmenting user-specified directives. We train this encoder using reinforcement learning, while preserving the pretrained BFM’s weights to retain its extensive behavioral knowledge. During training, the system generates behaviors from the BFM, conditioned on both user-defined goals and the emergent Task Tokens, optimizing the encoder to produce tokens that align behaviors with task-specific rewards. This strategy ensures that the resulting motions remain consistent with the natural motion manifold defined by the frozen BFM.

Our experimental evaluation demonstrates that Task Tokens effectively balance MaskedMimic’s ability to generate natural, human-like motions with the precision required for task-specific control. This hybrid framework achieves rapid convergence and high success rates, surpassing traditional hierarchical reinforcement learning methods in sample efficiency and requiring fewer learned parameters. Moreover, by adhering to the BFM’s underlying motion manifold, Task Tokens exhibit strong generalization across diverse environmental conditions, including variations in friction and gravity. These results demonstrate the potential of our method to unify goal-based and reward-driven control, enhancing behavior optimization for complex tasks.

1 Related Work

Humanoid Control: Humanoid control is a challenging domain spanning both robotics and computer animation, with the shared goal of generating realistic, robust, and human-like behaviors. In the animation community, physics simulation is used to ensure the generated motions are realistic and enable the characters to react to dynamic changes in the environment. To achieve this, they typically leverage imitation learning methods combined with motion capture data to learn and generate human-like behaviors in new and unseen scenarios (Peng et al., 2018; 2021; Luo et al., 2023; Tessler et al., 2023; Gao et al., 2025). Similar approaches are observed in the robotics community, with the addition of sim-to-real adaptation used to ensure the controller can overcome the imperfect modeling of the world by the simulator (Viceconte et al., 2022; Lu et al., 2024; Ji et al., 2024).

Our work builds on these foundations by preserving the natural, human-like motion qualities while enabling precise task-specific control.

Behavior Foundation Models: Recent advances in reinforcement learning have led to the development of Behavior Foundation Models (BFMs) that can generate diverse behaviors for embodied agents. PSM (Agarwal et al., 2024) and FB representations Touati & Ollivier (2021) provide a framework for learning policies conditioned on a target stationary distribution. These models perform remarkably well when the requested behavior can be represented by a stationary distribution (for example, using a reward-weighted combination of data samples). However, covering the entire space of solutions in high dimensional control tasks remains a challenge for these models. Methods such as Adversarial Skill Embeddings (Peng et al., 2022, ASE) and PULSE (Luo et al., 2024a) overcome this limitation by constraining the policy to reproducing human demonstrations. First, they compress a large repertoire of human reference motions into a latent generative policy. Then using reinforcement learning they train a hierarchical controller (Sutton et al., 1999) to pick the latent (motion to perform) at each step and solve new and unseen tasks.

In this work, we focus on Goal Conditioned Behavior Foundation Models (Chen et al., 2021; Zitkovich et al., 2023; Tessler et al., 2024). In contrast to the aforementioned methods, GC-BFMs can solve new and unseen tasks without specific training by directly mapping from goals to actions. Their mode of operation can be seen as a form of inpainting, where the model attempts to reproduce the most likely outcome given the training data for any provided objective. However, this strength is also a limitation, as these models struggle when presented with out-of-distribution constraints, such as those defined manually by a user or task. Our Task Tokens approach addresses this limitation by providing a mechanism to incorporate task-specific optimization while preserving the model’s ability to generate natural, human-like behaviors.

2 Preliminaries

Our proposed method leverages MaskedMimic to effectively solve a specific distribution of humanoid tasks by learning a “task encoder” with reinforcement learning.

2.1 Reinforcement Learning

A Markov Decision Process (Puterman, 2014, MDP) models sequential decision making as a tuple $M = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$. At each time step t the agent observes a state $s_t \in \mathcal{S}$ and predicts an action $a_t \in \mathcal{A}$. As a result, the environment transitions to a new state s_{t+1} based on the transition kernel P and the agent is provided a reward $r_t \sim R(s_t, a_t)$. The objective is to learn a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that maximizes the expected discounted cumulative reward $\pi^* = \text{argmax}_{\pi \in \Pi} \mathbb{E} \pi [\sum_t \gamma^t r_t]$.

2.2 MaskedMimic

MaskedMimic presents a unified framework for humanoid control, extending goal-conditioned reinforcement learning through imitation learning. Goal-Conditioned Reinforcement Learning (GCRL) involves augmenting the state space with a goal g , allowing a policy $\pi(s|s, g)$ to map states and desired goals to appropriate actions, effectively enabling a single policy to solve multiple tasks. Unlike traditional GCRL approaches that learn from reward signals, MaskedMimic learns directly from demonstration data through online distillation (Ross et al., 2011, DAgger). By combining a transformer architecture with random masking on future goals represented as input tokens, MaskedMimic learns to reproduce human-like behaviors from various modalities, such as future joint positions, textual commands, and objects for interaction. When trained on vast amounts of human motion capture data, this goal-conditioned approach allows MaskedMimic to generalize to new objectives without additional training, all while preserving the natural, human-like qualities of the training data. This combination of architecture and control scheme makes it an ideal foundation for our Task Tokens method, which further enhances its capabilities by learning task-specific tokens to optimize for downstream tasks.

3 Method

BFRMs excel at producing a wide range of natural-looking motions, but optimizing them for specific tasks presents significant challenges. Downstream applications often require specialized behaviors that fall outside the common distribution of motions. Traditional approaches to achieve such behaviors involve either time-consuming “prompt engineering” (Tessler et al., 2024) or fine-tuning procedures that risk compromising the rich prior knowledge encoded in the BFM.

3.1 Task Tokens

The transformer-based architecture of MaskedMimic provides a natural mechanism for integrating new task-specific information. By design, transformers process sequences of tokens allowing for flexible input composition. This enables us to seamlessly incorporate additional tokens without modifying the core network structure.

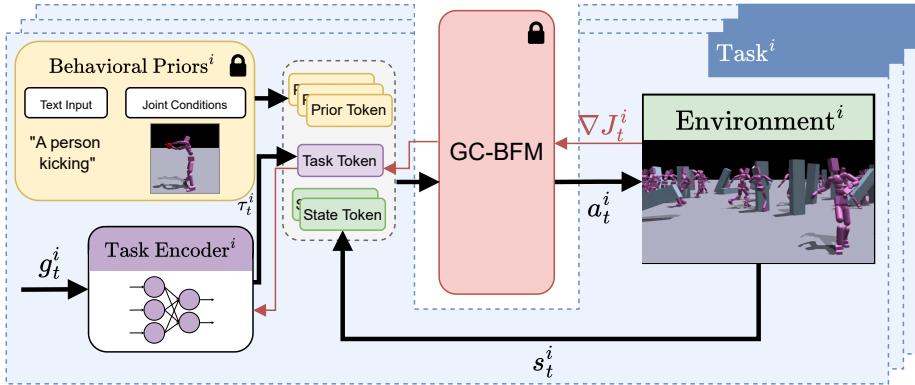


Figure 1: **Task Tokens:** Our approach combines three input sources: (1) Prior Tokens: optional tokens enabling user-defined behavioral priors from text prompts or joint conditions, (2) Task Token: generated by our learned Task Encoder that processes the current goal observation g_t^i , and (3) State Token: representing the current environment state s_t^i . The prior and state tokens are generated using the pre-trained encoders from the GC-BFM model. The frozen GC-BFM integrates these inputs to produce natural, task-optimized actions a_t^i . During training, the policy gradient objective is computed with respect to the BFM’s actions, with gradients flowing through the frozen GC-BFM and back to the Task Encoder, enabling task-specific optimization without modifying the foundation model’s parameters.

Our method, Task Tokens (Figure 1), leverages the tokenized nature of the BFM’s objectives. We propose to train a dedicated task encoder to produce specialized token representations for each new task. This task token encapsulates the unique requirements and constraints of the target behavior, providing a concise yet informative signal that can guide the foundation model toward task-specific outputs while preserving its general behavioral priors.

3.2 Task Encoder

The Task Encoder processes observations that define the current task goal g_t^i , represented in the agent’s egocentric reference frame and predicts a Task Token $\tau_t^i \in \mathbb{R}^{512}$. These observations vary by task – for instance, in a steering task, they include the target direction of movement $\in \mathbb{R}^2$, facing direction $\in \mathbb{R}^2$, and desired speed $\in \mathbb{R}$, resulting in $g_t^i \in \mathbb{R}^5$. As MaskedMimic is trained to reach future-pose goals, the task encoder is also provided with proprioceptive information. This aligns the encoder with the pre-trained representations, ensuring it can provide meaningful target objectives (see ablation studies in Appendix C).

We implement the task encoder as a feed-forward neural network. Its output—the Task Token—is concatenated with tokens from other encoders in the BFM’s input space. This creates a token “sentence” where the task encoder’s outputs represent specialized “words” that guide the model towards achieving the specific task while maintaining natural motion.

3.3 Training

To optimize the Task Encoder for new downstream tasks, we use Proximal Policy Optimization (Schulman et al., 2017, PPO). During training, the BFM predicts action probabilities based on the combined input tokens (including the learned task token). We compute the PPO objective with respect to the task-specific reward and the BFM’s action probabilities. This approach ensures the BFM provides meaningful gradients for updating the task encoder parameters, while the BFM itself remains frozen. This design choice is fundamental – while fine-tuning the entire model might yield

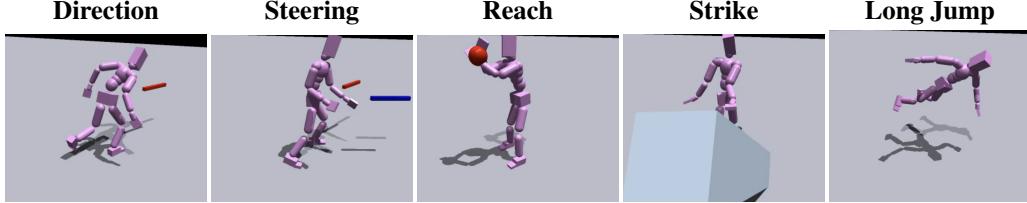


Figure 2: **Multi-task adaptation.** Task Tokens is an effective approach to adapt BFM to new downstream tasks, while preserving its prior knowledge. Task Tokens can be used alongside other prompting modalities to generate personalized and robust motions to solve new tasks.

higher task-specific returns, it would compromise the BFM’s prior knowledge, resulting in less natural and robust motions.

By utilizing MaskedMimic’s token-based architecture, Task Tokens requires only ~200k parameters for each additional task, making it highly parameter-efficient.

4 Results

We evaluate the effectiveness of our Task Tokens approach through a comprehensive set of experiments. We examine four critical aspects of our method to validate its performance and applicability. First, we assess the capability of Task Tokens to effectively adapt MaskedMimic for various downstream applications, demonstrating significant improvements in task-specific performance (Section 4.1). Second, we analyze whether the resulting controller preserves the robustness characteristics inherent to the original Behavioral Frequency Modulation (BFM) framework, confirming that stability under variable conditions remains consistent (Section 4.2). Third, we investigate the natural and human-like quality of the generated motions through a human study (Section 4.3). Finally, we explore the synergy of Task Tokens and other prompting modalities, combining effects that further demonstrate the versatility of our method (Section 4.4). These experiments collectively demonstrate that our approach successfully balances task-specific adaptation with the preservation of desirable properties from the foundation model.

We provide accompanying video visualizations for all experiments in sites.google.com/view/task-tokens.

Tasks: We evaluate our approach on a diverse set of humanoid control tasks, all simulated in Isaac Gym (Makoviychuk et al., 2021). For all experiments, we simulate the SMPL humanoid Loper et al. (2015) which consists of 69 degrees of freedom. We focus on the following tasks: **Reach**, the agent must reach a randomly placed goal with its right hand; **Direction**, the agent must walk in a randomly chosen direction; **Steering**, this task combines walking and orienting toward (look-at) random directions; **Strike**, here the agent must reach and strike a target placed at a random location; and **Long Jump**, based on the SMPL-Olympics benchmark (Luo et al., 2024b), the objective is to run and jump as far as possible from a target location. Sample images are shown in Figure 2, full technical details can be found in Appendix A.

Baselines and Evaluation We compare our Task Tokens approach against several competitive baselines: **Pure RL**, a policy trained directly using PPO without leveraging any foundation model; **MaskedMimic Fine-Tune**, using the reward signal to optimize all of the MaskedMimic model without freezing; **MaskedMimic (J.C. only)**, the original MaskedMimic model using only joint conditioning (J.C.) as the prompting mechanism. We use the J.C. defined in the original MaskedMimic for the Reach, Direction, and Steering tasks. In addition, we compare against two state-of-the-art humanoid control baselines: **PULSE**, a hierarchical approach that re-uses a latent space of skills from motion capture data; and **AMP**, which uses a discriminator to ensure motion quality while

Table 1: **Downstream tasks adaptation.** We compare the success rates on the various tasks. While the reward provides a proxy for the policy to learn, the success metric measures the actual success on the task. For example, in Strike whether the target object is knocked down.

Method	Reach	Direction	Steering	Long Jump	Strike
Task Tokens (ours)	94.88 ± 1.99	99.26 ± 0.79	88.69 ± 4.04	99.75 ± 0.57	76.61 ± 3.49
MaskedMimic (J.C. only)	24.77	2.19	3.83	-	-
MaskedMimic Fine-Tune	93.70 ± 4.59	99.10 ± 1.29	87.44 ± 6.79	47.36 ± 54.78	83.07 ± 5.71
PULSE	83.96 ± 2.20	97.60 ± 0.62	40.72 ± 7.64	99.37 ± 1.40	83.18 ± 2.67
AMP	57.14 ± 4.80	5.14 ± 0.68	4.28 ± 1.42	76.59 ± 43.42	52.21 ± 47.58
PPO	89.90 ± 3.25	97.74 ± 1.40	32.64 ± 40.21	61.91 ± 52.26	81.36 ± 1.41

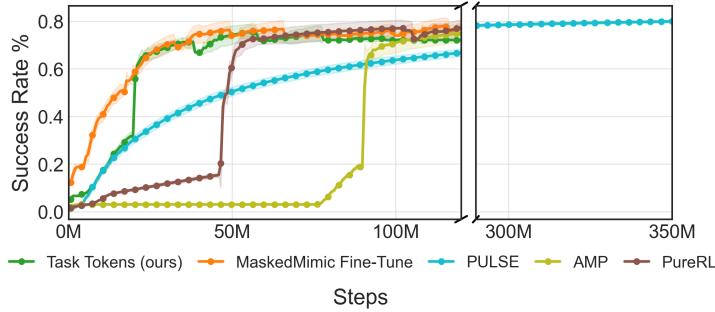


Figure 3: **Convergence curves for Strike.** Task Tokens is sample efficient, adapting to new tasks in under 50M steps.

optimizing for task performance. **Task Tokens** is used with joint conditioning on the relative tasks. Long Jump and Strike pose a great challenge in this sense, thus J.C. is not available for them neither in Task Tokens nor in MaskedMimic.

In all experiments, we report the mean and standard deviation of the success rate over 5 seeds. Success rate definitions for each task are listed in Appendix A, and details on the training and evaluation setups are available in Appendix B.

4.1 Task Adaptation

We first show that we can use Task Tokens to effectively adapt MaskedMimic to downstream tasks. For each downstream task, we train a unique task encoder. In the Reach, Direction, and Steering environments, we also use the joint conditioning presented in MaskedMimic (we test the effect of this choice in Appendix C). Visualizations of the resulting motions can be seen in Figure 2.

We present the numerical results in Table 1. The results show that Task Tokens obtains a high score across the majority of environments, with PULSE, MaskedMimic Fine-Tune and PureRL obtaining higher scores on the Strike task. Although fine-tuning leads to comparable results, similarly to PureRL, it lacks in human-like motion quality. These claims are demonstrated in Section 4.3. Moreover, in Figure 3, we present the evaluated success rate during training. Here, we observe that Task Tokens converges within approximately 50 million steps, while PULSE reaches the same performance around 300 million steps. To achieve these results, Task Tokens requires training an encoder with ~200k parameters, whereas PULSE and MaskedMimic Fine-Tune require 9.3M and 25M parameters, higher by factors of $\times 46.5$ and $\times 125$ respectively.

These results show that Task Tokens can effectively and efficiently be used to adapt BFM, like MaskedMimic, to new and previously unseen tasks.

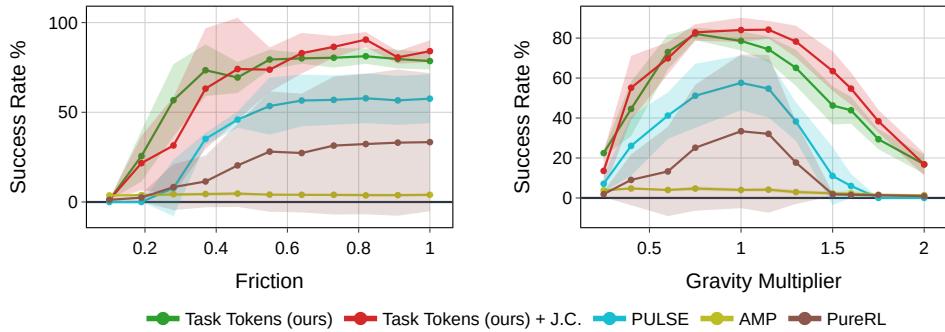


Figure 4: **Out-of-distribution perturbations.** We test the success rate on the steering task when changing the ground friction (on the left) and gravity (on the right). Task tokens (both with and without J.C.) exhibit improved robustness.

4.2 OOD Generalization

The premise of using MaskedMimic is that it has been pre-trained on vast amounts of data and scenarios, which in turn should result in more robust behavior to new and unseen tasks. To test this, we compare on out-of-distribution (OOD) perturbations, not seen during training. We consider changes in both gravity and ground friction.

Indeed, the results, Figure 4, show that by utilizing a BFM, Task Tokens demonstrates improved robustness to new and unseen scenarios. First, it performs better on the baseline task (no perturbations). Then, this increased performance is also maintained as the perturbation increases. Notably, Task Tokens exhibits an order of magnitude higher rate of success in very low friction scenarios (e.g., $\times 0.4$) and very large gravity ($\times 1.5$).

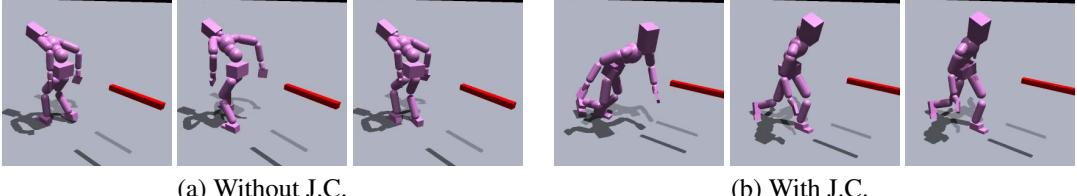
4.3 Human Study

Table 2: **Human study, Task Tokens win rate.** We report the percentage by which Task Tokens was deemed more human-like. Higher values means Task Tokens was deemed more human-like.

Wins v.s. Algorithm	Direction	Steering	Reach	Strike	Long Jump
MaskedMimic (J.C. only)	95.45%	74.74%	52.51%	-	-
MaskedMimic F.T.	99.48%	90.40%	84.57%	85.37%	94.26%
MaskedMimic F.T. + J.C.	96.09%	89.26%	81.99%	-	-
PULSE	14.80%	46.44%	36.28%	24.19%	39.13%
AMP	92.38%	83.75%	70.09%	67.88%	96.17%
PPO	99.29%	92.57%	88.62%	81.98%	93.82%

In many scenarios, such as animation, it is of interest to adapt and generate new behaviors (solutions to tasks) without compromising on the motion quality. We evaluate the realism of the generated motions by performing a comprehensive human study. In our study, we presented ~ 100 anonymous participants with video triplets. The participants were required to choose the motion that looked more human-looking. We provide additional details in Appendix D.

Table 2 describes the percentage of times Task Tokens outperformed each alternative. The results show that alternative approaches to adapting the BFM, such as fine-tuning, lead to dramatically reduced motion quality. This motivates reusing a frozen BFM by only training a new input con-



(a) Without J.C.

(b) With J.C.

Figure 5: Multi-modal prompting. When trained on the direction task, the policy often learns to walk backwards. Task Tokens enables adding human-defined-priors through additional tokens. By combining orientation priors, the BFM is instructed to face the movement direction.

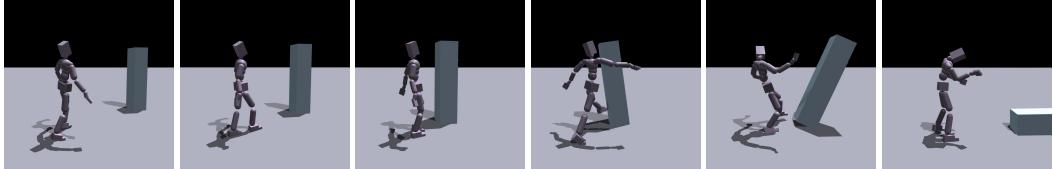


Figure 6: Incorporating user priors: Joint and text-based goals can be provided alongside the task-specific task tokens. Here, the user-defined objectives guide the motion style, ensuring the character faces the object while walking upright, and strikes it using a kick.

troller. In addition, we observe that despite Task Tokens resulting in faster convergence and better performance, PULSE scores higher in terms of human-likeness of the motion.

4.4 Multi-modal Prompting

While some objectives are easy to define through target goals, others are easier to define through rewards. Here, we show how Task Tokens can be trained alongside human-constructed priors (tokens) to achieve more desirable behaviors. We showcase two scenarios, one in the Direction task and another in Strike.

The Direction task provides a reward for moving in the right direction but does not consider the humanoid’s orientation. As a result, the policy may converge to walking backwards. While this behavior achieves high reward and success metrics, it is an unwanted behavior. In Figure 5 we show that by combining human-designed priors, providing a target height and orientation for the head, the training converges to an upright forward-moving motion.

An additional challenge is Strike. In this task, the agent needs to hit a target. An emergent behavior is walking backwards toward the target and then performing a “whirlwind” motion, where the agent swirls in circles to hit the target with its hand. In Figure 6 we showcase a combination of 2 prior modalities. First, conditioning on the orientation (similar to the Direction task) the agent is instructed to face the target during the locomotion phase. Then, once close to the target, the agent is guided to strike the target with its foot using a textual objective “a person performs a kick”.

5 Summary

This work introduces Task Tokens, a novel approach that enhances Goal-Conditioned Behavior Foundation Models (GC-BFMs), such as MaskedMimic, by integrating goal-based control with reward-driven optimization. Our method enables a hybrid control paradigm: users provide high-level behavioral priors, while the system learns task-specific embeddings to optimize dense rewards. A Task Encoder, trained using reinforcement learning, maps task observations to goal tokens, augmenting user directives while preserving the BFM’s natural motion capabilities. Experimental evaluations show that Task Tokens effectively balances motion realism with task precision, achieving

rapid convergence, high success rates, and superior generalization compared to existing methods. Human studies confirm that Task Tokens generates more human-like movements. However, while Task Tokens offer a parameter-efficient adaptation strategy, its reliance on the quality of the underlying BFM and its current restriction to simulated environments presents limitations.

References

- Siddhant Agarwal, Harshit Sikchi, Peter Stone, and Amy Zhang. Proto successor measure: Representing the space of all possible solutions of reinforcement learning. *arXiv preprint arXiv:2411.19418*, 2024.
- Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.
- Jiawei Gao, Ziqin Wang, Zeqi Xiao, Jingbo Wang, Tai Wang, Jinkun Cao, Xiaolin Hu, Si Liu, Jifeng Dai, and Jiangmiao Pang. Coohoi: Learning cooperative human-object interaction with manipulated object dynamics. *Advances in Neural Information Processing Systems*, 37:79741–79763, 2025.
- Mazeyu Ji, Xuanbin Peng, Fangchen Liu, Jialong Li, Ge Yang, Xuxin Cheng, and Xiaolong Wang. Exbody2: Advanced expressive humanoid whole-body control. *arXiv preprint arXiv:2412.13196*, 2024.
- Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.
- Chenhao Lu, Xuxin Cheng, Jialong Li, Shiqi Yang, Mazeyu Ji, Chengjing Yuan, Ge Yang, Sha Yi, and Xiaolong Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. *arXiv preprint arXiv:2412.07773*, 2024.
- Zhengyi Luo, Jinkun Cao, Kris Kitani, Weipeng Xu, et al. Perpetual humanoid control for real-time simulated avatars. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10895–10904, 2023.
- Zhengyi Luo, Jinkun Cao, Josh Merel, Alexander Winkler, Jing Huang, Kris M. Kitani, and Weipeng Xu. Universal humanoid motion representations for physics-based control. In *The Twelfth International Conference on Learning Representations*, 2024a. URL <https://openreview.net/forum?id=OrOd8Px002>.
- Zhengyi Luo, Jiashun Wang, Kangni Liu, Haotian Zhang, Chen Tessler, Jingbo Wang, Ye Yuan, Jinkun Cao, Zihui Lin, Fengyi Wang, et al. Smplolympics: Sports environments for physically simulated humanoids. *arXiv preprint arXiv:2407.00187*, 2024b.
- Viktor Makoviychuk, Lukasz Wawrzyniak, Yunrong Guo, Michelle Lu, Kier Storey, Miles Macklin, David Hoeller, Nikita Rudin, Arthur Allshire, Ankur Handa, and Gavriel State. Isaac gym: High performance GPU based physics simulation for robot learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. URL https://openreview.net/forum?id=fgFBtYgJQX_.
- 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.
- 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.

Xue Bin Peng, Yunrong Guo, Lina Halper, Sergey Levine, and Sanja Fidler. Ase: Large-scale reusable adversarial skill embeddings for physically simulated characters. *ACM Trans. Graph.*, 41(4), July 2022.

Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

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

Richard S Sutton, Doina Precup, and Satinder Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial intelligence*, 112(1-2):181–211, 1999.

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*, pp. 1–9, 2023.

Chen Tessler, Yunrong Guo, Ofir Nabati, Gal Chechik, and Xue Bin Peng. Maskedmimic: Unified physics-based character control through masked motion inpainting, 2024. URL <https://arxiv.org/abs/2409.14393>.

Ahmed Touati and Yann Ollivier. Learning one representation to optimize all rewards. *Advances in Neural Information Processing Systems*, 34:13–23, 2021.

Paolo Maria Viceconte, Raffaello Camoriano, Giulio Romualdi, Diego Ferigo, Stefano Dafarra, Silvio Traversaro, Giuseppe Oriolo, Lorenzo Rosasco, and Daniele Pucci. Adherent: Learning human-like trajectory generators for whole-body control of humanoid robots. *IEEE Robotics and Automation Letters*, 7(2):2779–2786, 2022.

Jungdam Won, Deepak Gopinath, and Jessica Hodgins. Physics-based character controllers using conditional vaes. *ACM Transactions on Graphics (TOG)*, 41(4):1–12, 2022.

Philipp Wu, Arjun Majumdar, Kevin Stone, Yixin Lin, Igor Mordatch, Pieter Abbeel, and Aravind Rajeswaran. Masked trajectory models for prediction, representation, and control. In *International Conference on Machine Learning*, pp. 37607–37623. PMLR, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Environments Technical Details

The controllers operate at 30 Hz, and the simulation runs at 120 Hz.

The tasks are designed to test the versatility and adaptability of the models across a range of real-world scenarios, each adding layers of complexity to the control problem.

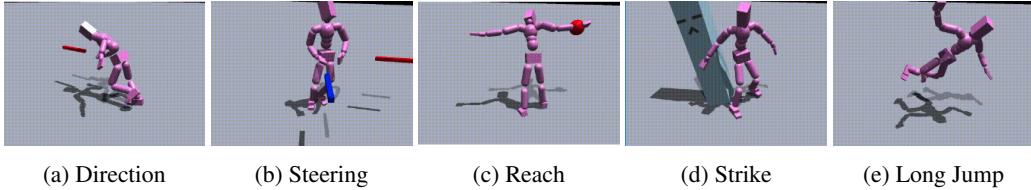


Figure 7: Visual samples from the tasks.

Direction: This task involves directing the character to move in a specific direction. We test the model’s ability to control basic locomotion and alignment with a target direction. We measure success by the humanoid’s speed in the target direction not deviating from the target speed by more than 20% in the measurement period.

Steering: This task requires the humanoid to move in a specific direction while also ensuring that it faces some orientation with its pelvis. This tests more nuanced proceeding motion of the model, creating more diverse scenarios. Success is defined by the character not deviating from the target direction speed by more than 20% while also not deviating from the direction of facing by a sum greater than 45°.

Reach: For this scenario, we task the humanoid with reaching a specified coordinate with the right hand. This requires precision of movement to achieve the specific target. The success is measured by reaching a distance within 20cm between the right hand’s position and the target position.

Strike: Here we challenge the model to make the character walk toward a target and, once within range, perform an action to knock down the target. This task tests the model’s ability to handle both locomotion and more intricate, task-oriented behaviors, involving precise timing and spatial awareness. Success is then defined by the target falling to its side in some orientation and not deviating from it by more than ~78°.

Long Jump The character is tasked with committing a run in a meter-wide corridor, then jumping over a line after 20 meters, not touching the ground after crossing the jump start line. Success is defined by achieving a jumping distance greater than 1.5 meters.

B Training and Evaluation Details

All experiments were trained in parallel on 1024 environments for 4000 epochs resulting in 120M frames. PULSE was also trained on 120M frames but on 128 environments in parallel. In the Task Tokens main results, we used a Task Encoder that is an MLP of size [512, 512, 512] and an MLP critic that’s the size of [1024, 1024, 1024]. We also concatenated to the Task Encoder’s input the current positions of the head and pelvis joints. This seemed to produce slightly more human-like motions. In all experiments, we report the mean and standard deviation of the success rate over 5 seeds. Each experiment seed is the last model checkpoint achieved during training. Success rate results in training and evaluation might differ since we do not use the same episode termination rules.

C Ablation Study

We experimented with several variables, constructing the Task Encoder. The results are shown in Table 3.

C.1 MaskedMimic Adaptation Paradigm

Table 3: **Algorithm training scheme ablations.** While fine-tuning the whole MaskedMimic model can produce performing results, we've shown it lacks the human-like abilities of Task Tokens.

Method	Reach	Direction	Steering	Long Jump	Strike
Task Tokens (ours)	95.37 ± 1.80	96.89 ± 4.33	83.66 ± 5.66	99.75 ± 0.57	76.61 ± 3.49
Task Tokens (ours) + J.C.	94.88 ± 1.99	99.26 ± 0.79	88.69 ± 4.04	-	-
MaskedMimic (J.C. only)	24.77	2.19	3.83	-	-
MaskedMimic F.T.	93.70 ± 4.59	99.10 ± 1.29	87.44 ± 6.79	47.36 ± 54.78	83.07 ± 5.71
MaskedMimic F.T. + J.C.	92.88 ± 3.42	98.86 ± 0.32	96.41 ± 4.94	-	-

Results in Table 3 demonstrate superior performance when using J.C. when available.

C.2 Task Encoder Architecture

We further present some architectural changes made to the Tak Encoder and their effect on output performance, listed in Table 4. Bigger MLP denotes using [512, 512, 512] size MLP encoder versus [256, 256] and Using Current Pose denotes whether the current positions of the head and pelvis are concatenated to the input alongside the task goal. When using joint conditioning, the performance stays high for every choice except when using a smaller encoder with current pose information. This result replicates when not using joint conditioning.

Table 4: **Task Encoder architectural ablations.**

Method	Bigger MLP	Using Current Pose	Steering Success Rate
Task Tokens (ours) + J.C.	True	True	87.77 ± 7.14
Task Tokens (ours) + J.C.	True	False	87.58 ± 7.02
Task Tokens (ours) + J.C.	False	False	86.88 ± 6.65
Task Tokens (ours)	False	False	84.28 ± 7.72
Task Tokens (ours)	True	False	83.30 ± 10.06
Task Tokens (ours) + J.C.	False	True	79.47 ± 4.71
Task Tokens (ours)	True	True	78.59 ± 8.93
Task Tokens (ours)	False	True	66.31 ± 13.11

D Human Study Technical Details

To assess the quality of the motions generated by our method we conducted a human study. The participants were met with this description before filling out the form:

In this study, you will watch three short videos side by side each time.

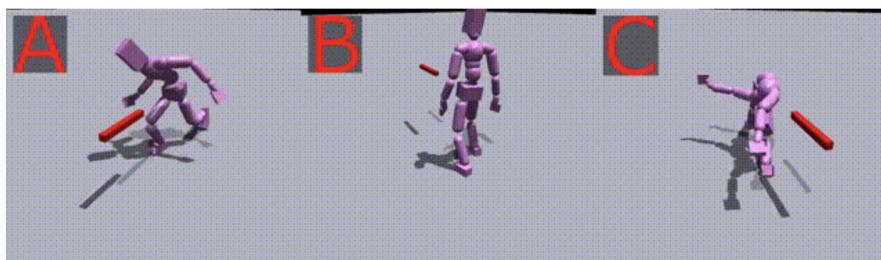
These videos show different ways a character moves to complete a task. Your job is to decide which movement looks the most human-like.

Each video is labeled A, B, or C — the labels are shuffled every time. Just pick the one you think does the best job.

Don't worry — there's no right or wrong answer! We just want your opinion.

We used Google Forms to create 3 forms, each containing 40 questions - 8 questions for each of the tasks listed in Appendix A. In each question, we showed 3 videos side-by-side of 3 randomly sampled sequences generated by the algorithms. We ensured that Task Tokens was presented every time. Joint conditioning was used when applicable, i.e. for Direction, Steering and Reach tasks. The participants were asked to choose which algorithm looks most human-like for the task described in the question. An example

Q 1: Which example looks more human-like for task **walking in red direction**? *



- A
- B
- C

To ensure no bias, we shuffled the order of algorithms in every question and captured the videos from similar angles. The number of participants was 96: 20, 24, 52 for the forms respectively. We analyzed the winning rate for each environment as winning percentage $_{env}^A = \frac{\# \text{Task Tokens chosen}}{\# \text{Task Tokens chosen} + \# \text{Algorithm A chosen}}$