World Model Predictive Control with Multimodal Adaptation: Towards Artificial Dynamics Intelligence for General Robotics

Baha Zarrouki[1,2] and Johannes Betz[2]

***Abstract*—Traditional robotic control systems rely on modular sense-plan-act architectures or end-to-end deep learning, both of which struggle with seamless task transfer, real-time adaptation, and computational efficiency. In this work, we propose World Model Predictive Control (WMPC), a novel framework that integrates World Models with Model Predictive Control (MPC) principles. By leveraging pre-trained differentiable world models to predict system dynamics and optimize control actions, WMPC eliminates the need for extensive policy training, unlike reinforcement learning (RL)-based world models. Our approach unifies multimodal state representations, task-specific cost learning, and constraint-aware optimization within a receding horizon framework. Inspired by human motor control and learning, it integrates general scene understanding and basic dynamics estimation while fine-tuning actions through rapid interaction with the environment. Furthermore, elastic model updating balances short-term corrections—instantaneous reactive adjustments to new dynamics—with long-term knowledge retention, enabling memory-augmented fine-tuning and improved general skill proficiency.**

I. Introduction

Human motor control exhibits remarkable adaptability, seamlessly integrating sensory feedback to estimate physical properties (e.g., mass, friction) and adjust forces in real time [1]. Human sensorimotor control [2] exhibits three key properties current robots lack:

1) Cross-Modal Fusion: Seamless integration of visual, tactile, and inertial cues
2) Temporal Hierarchy: Short-term adjustments (100-500ms) nested within long-term skill refinement
3) Physics-Guided Learning: Priors from biomechanical constraints and Newtonian dynamics

Robotic control has traditionally relied on two main paradigms: modular sense-plan-act architectures and end-to-end learning approaches. While the former ensures transparency and stability, it suffers from high development costs, limited scalability, and computational inefficiencies [3]. Conversely, end-to-end learning methods, such as Vision-Language-Action (VLA) models [4], world model-based reinforcement learning [5], and deep latent feature-based control [6], demonstrate strong task acquisition but remain data-hungry and struggle with generalization [7] and real-time adaptability [8], [9].

[1] Chair of Automotive Technology, Technical University of Munich
[2] Professorship of Autonomous Vehicle Systems, TUM School of Engineering and Design, Technical University of Munich, 85748 Garching, Germany; Munich Institute of Robotics and Machine Intelligence (MIRMI), {baha.zarrouki, johannes.betz}@tum.de

Model-based control methods, particularly Model Predictive Control (MPC), have been instrumental in bridging these gaps. Traditional MPC relies on handcrafted dynamics models [10], but recent advances incorporate learned dynamics, as seen in latent-space planning approaches such as PlaNet [11] and DeepMPC [6]. However, these methods often require task-specific training and suffer from loss of physical interpretability [12].

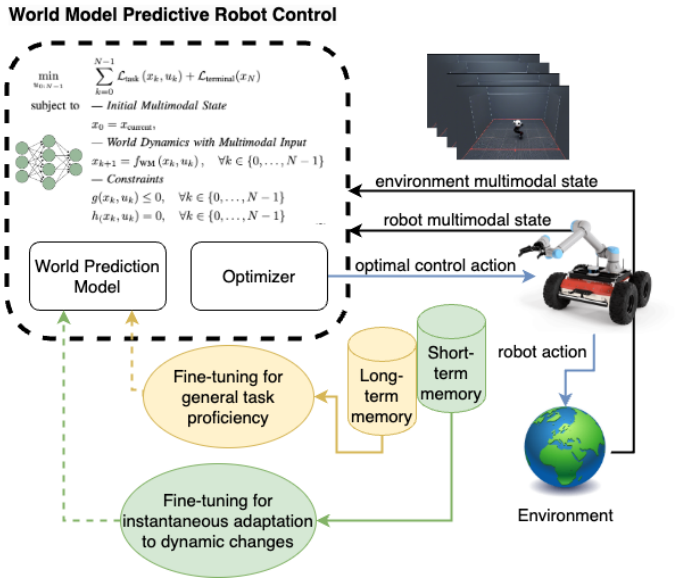

Fig. 1. World Model Predictive Control with Multimodal Adaptation.

Foundation models have recently emerged as a promising direction for robotic control, with architectures like CLIP [4], DINO [13], and OpenVLA [14] enabling multimodal state estimation. Despite their advantages [15], these models remain constrained by static representations, a lack of continuous adaptation mechanisms, and a disconnect between semantic understanding and physical constraints [16].

Recent work has sought to address these limitations through gradient-based planning [17], diffusion-based world modeling [18], [19], and hybrid approaches combining policy learning with gradient-based MPC [17]. Notably, DIAMOND [18] introduces diffusion models for world modeling, mitigating the limitations of discrete latent representations by preserving critical visual details. Additionally, methods such as RoboGen [20] leverage large-scale synthetic data generation for training generalist robotic policies, though they remain dependent on task-specific policy learning.

Our proposed framework, World Model Predictive Control (WMPC), advances beyond existing approaches by integrating pre-trained Large Multimodal Foundation Models (LMFMs) into a receding horizon optimization structure. Unlike PETS [21], which requires full model retraining, WMPC employs elastic model updating to selectively adapt task-relevant parameters while retaining foundational knowledge. Furthermore, it incorporates constraint-aware optimization [22] and multimodal state propagation [23], enabling scalable, sample-efficient, and dynamically adaptable robotic control.

By synthesizing insights from foundation models, model predictive control, and learned world models, WMPRC represents a paradigm shift towards generalizable, interpretable, and real-time adaptable robotic autonomy.

II. Proposed Methodology

Our framework integrates multimodal state representations, world model predictions, and receding horizon optimization to solve an Optimal Robot Control Problem. The core idea is to leverage pre-trained foundation models for world prediction and fine-tune them dynamically for task-specific requirements.

A. System Architecture

WMPC integrates three key components (Fig.1):

1) A pre-trained world model for state prediction
2) An optimal control framework for action selection
3) Adaptive memory systems for short-term and long-term learning

B. Mathematical Formulation

The core optimization problem is formulated as:

$$\begin{aligned} \min_{u_{0:N-1}} \quad & \sum_{k=0}^{N-1} \mathcal{L}_{\text{task}}\left(x_k, u_k\right) + \mathcal{L}_{\text{terminal}}(x_N) \\ \text{subject to} \quad & \text{— } \textit{Initial Multimodal State} \\ & x_0 = x_{\text{current}}, \\ & \text{— } \textit{World Dynamics with Multimodal Input} \\ & x_{k+1} = f_{\text{WM}}\left(x_k, u_k\right), \quad \forall k \in \{0, \ldots, N-1\} \\ & \text{— } \textit{Constraints} \\ & g(x_k, u_k) \leq 0, \quad \forall k \in \{0, \ldots, N-1\} \\ & h_(x_k, u_k) = 0, \quad \forall k \in \{0, \ldots, N-1\} \end{aligned} \tag{1}$$

where x_k is the multimodal state vector defined as:

$$x_k = \begin{bmatrix} x_k^{\text{visual}} \\ x_k^{\text{dynamics}} \\ x_k^{\text{other}} \end{bmatrix} \tag{2}$$

Here, f_{WM} represents the world model dynamics, $\mathcal{L}_{\text{task}}$ is the task-specific loss, and g, h define constraints. The objective is to minimize the cumulative task-specific loss $\mathcal{L}$ while adhering to physical and operational constraints. This framework employs a receding horizon implementation, where the optimal control problem is solved iteratively at each time step using updated multimodal sensor data of the robot and of the environment.

Pre-trained LMFMs, such as DINO [13], [24] or CLIP [4], provide a strong foundation for multimodal understanding. These models are fine-tuned dynamically using short-term memory to adapt to task-specific requirements.

C. Cost Function and Task Handling

The framework supports both analytic and learned cost formulations:

- Analytic Formulation: For well-defined tasks, $\mathcal{L}_{\text{task}}$ combines classical objectives:

$$\mathcal{L}_{\text{task}}(x_k, u_k) = \sum_{i=1}^{M} w_i \mathcal{L}_i(x_k, u_k) \tag{3}$$

  where w_i are task-specific weights and $\mathcal{L}_i$ represents individual objective terms such as:
  - Task achievement error: $\|x_k - x_{\text{desired}}\|_Q$
  - Control effort: $\|u_k\|_R$
  - Safety margins: $\mathcal{L}_{\text{safety}}(x_k)$
  - Task-Specific Terms: $\phi(x_k, u_k)$
- Learned Formulation: For complex tasks, we parameterize the loss using a neural network $\mathcal{N}_\theta$ with latent task variables z:

$$\mathcal{L}_{\text{task}}(x, u) = \mathcal{N}_\theta \left(\underbrace{[x; u]}_{\text{Current State-Action}}, \underbrace{z}_{\text{Latent Task Rep.}} \right) \tag{4}$$

D. Memory-Augmented Adaptive Fine-Tuning

To balance immediate adaptation with long-term stability, we employ a dual-time scale learning strategy that leverages structured memory and fine-tuning. The adaptation process is driven by two key components:

- **Online Adaptation:** A short-term memory buffer (τ_{STM}) captures recent trajectory segments, enabling rapid updates via context-aware gradient descent. This ensures responsiveness to instantaneous changes in system dynamics.
- **Offline Enhancement:** A long-term memory (τ_{LTM}) accumulates historical knowledge, facilitating periodic fine-tuning for improved generalization and robustness over time.

To integrate both adaptation scales, we apply a dual-rate optimization scheme:

$$\theta_{t+1} = \theta_t - \eta_1 \nabla \mathcal{L}_{\text{short}} - \eta_2 \nabla \mathcal{L}_{\text{long}} \tag{5}$$

where η_1 governs rapid corrections based on short-term memory that optimizes immediate predictions, while η_2 ensures gradual refinement from accumulated experience.

III. Challenges

Our approach faces several challenges, including the complexity of integrating multimodal data into a unified state representation, ensuring real-time adaptability in dynamic environments, and fine-tuning pre-trained models for task-specific requirements. Extracting meaningful latent variables

from high-dimensional sensory inputs can be computationally demanding.

REFERENCES

[1] Reza Shadmehr and Sandro Mussa-Ivaldi. *Biological Learning and Control: How the Brain Builds Representations, Predicts Events, and Makes Decisions*. MIT Press, Cambridge, MA, 2010.

[2] Daniel M Wolpert, Jörn Diedrichsen, and J Randall Flanagan. Principles of sensorimotor learning. *Nature Reviews Neuroscience*, 12(12):739–751, 2011.

[3] Bruno Siciliano and Oussama Khatib. *Springer handbook of robotics*. Springer, 2016.

[4] 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. *International Conference on Machine Learning (ICML)*, 2021.

[5] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[6] Ian Lenz, Ross A Knepper, and Ashutosh Saxena. Deepmpc: Learning deep latent features for model predictive control. In *Robotics: Science and Systems*, volume 10, page 25. Rome, Italy, 2015.

[7] Oussama Khatib, Luis Sentis, and Jaeheung Park. A unified approach to whole-body humanoid robot control with multiple constraints and contacts. *European Robotics Symposium*, pages 303–312, 2019.

[8] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *International Conference on Learning Representations (ICLR)*, 2020.

[9] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, pages 2555–2565, 2019.

[10] James B. Rawlings, David Q. Mayne, and Moritz M. Diehl. *Model Predictive Control: Theory, Computation, and Design*. Nob Hill Publishing, 2017.

[11] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International Conference on Machine Learning (ICML)*, 2019.

[12] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Deepmpc: Learning deep latent features for model predictive control. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2405–2415, 2018.

[13] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *IEEE International Conference on Computer Vision (ICCV)*, 2021.

[14] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, et al. Openvla: An open-source vision-language-action model. *arXiv preprint arXiv:2406.09246*, 2024.

[15] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[16] Anthony Brohan, Noah Brown, Julian Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, et al. Rt-1: Robotics transformer for real-world control at scale. *arXiv preprint arXiv:2212.06817*, 2022.

[17] SV Jyothir, Siddhartha Jalagam, Yann LeCun, and Vlad Sobal. Gradient-based planning with world models. *arXiv preprint arXiv:2312.17227*, 2023.

[18] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari. *arXiv preprint arXiv:2405.12399*, 2024.

[19] Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

[20] Shengyi Yan, Steven James, Yuchen Zhang, Xingyu Tan, Zhou Li, Jie Li, and Cewu Lu. Robogen: Towards unleashing infinite data for automated robot learning. *arXiv preprint arXiv:2312.17227*, 2023.

[21] Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[22] Anusha Nagabandi, Ignasi Clavera, Simin Liu, Ronald S. Fearing, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Learning to adapt in dynamic, real-world environments through meta-reinforcement learning. *International Conference on Machine Learning (ICML)*, 2019.

[23] Arunkumar Byravan and Dieter Fox. Se3-pose-nets: Structured deep dynamics models for visuomotor control. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

[24] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.