ModularAgent: A Task-Aware Modular Framework for Joint Optimization of Multimodal Large Language Models and World Models

Yu-Wei Zhan[1] Xin Wang[1*] Pengzhe Mao[2] Tongtong Feng[1*] Ren Wang[1] Wenwu Zhu[1*]
[1]Department of Computer Science and Technology, BNRIST, Tsinghua University
[2]School of Software, Shandong University

Abstract

Building generalist embodied agents requires a unified system that can interpret multimodal goals, model environment dynamics, and execute reliable actions across diverse real-world tasks. Multimodal large language models (MLLMs) offer strong semantic priors and cross-modal generalization, while world models (WMs) provide actionable latent dynamics for prediction and control. Their combination holds promise for open-ended embodied intelligence, yet introduces two key challenges: (1) establishing a tight coupling between the semantic intent from MLLMs and the dynamic state representations within the WM's latent space, and (2) achieving task-aware adaptability that supports multi-task learning and cross-environment generalization. To address these limitations, we propose ModularAgent, a task-aware dynamic joint framework that enables bidirectional coupling between MLLMs and WMs. ModularAgent establishes two complementary pathways: a forward path that injects MLLM representations into the WM's latent space for semantically guided imagination, and a backward path where WM-generated feedback refines the MLLM's semantic space via dense text-conditioned rewards. This bidirectional interaction is realized through three synergistic components: Task-Aware Dynamic Joint Learning, Task-Aware Behavior Learning, and MLLM-WM Joint Optimization, which together harmonize semantic reasoning and dynamic prediction. Extensive experiments across multi-task and cross-environment settings demonstrate superior stability and generalization over state-of-the-art baselines, marking a step toward open-ended embodied learning.

1. Introduction

A generalist embodied agent aims to perform diverse tasks across real-world environments within a unified framework [14, 17, 49]. Such an agent is capable of interpreting high-level multimodal goals, grounding them into actionable representations, and executing reliable control based on learned environment dynamics. Instead of relying on task-specific pipelines, the ultimate goal is to build a single architecture that generalizes across heterogeneous tasks and environments while preserving consistent decision-making and control capabilities.

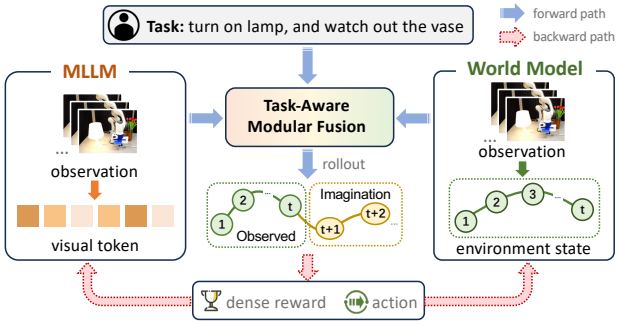

Figure 1. ModularAgent enables task-aware bidirectional coupling between MLLM and WM. In the forward path, MLLM semantics are injected into the WM via Task-Aware Modular Fusion for semantically guided imagination. In the backward path, task-conditioned imagined trajectories produce rewards and actions, which are backpropagated through the joint loss to refine MLLM.

Recent advances in Multimodal Large Language Models (MLLMs) have demonstrated strong capabilities in semantic reasoning and cross-modal understanding [2, 12, 31, 33, 34, 45, 46]. They possess rich world knowledge and strong compositional priors, enabling them to interpret high-level task instructions and align multimodal inputs within a unified semantic space. In parallel, World Models (WMs) excel at modeling environment dynamics and supporting decision-making [7, 15, 20, 23]. They learn environment latent state representations that capture temporal dependencies and enable long-horizon imagination and planning. While MLLMs provide powerful semantic understanding and generalization across modalities, they lack physical interaction capabilities; conversely, WMs offer precise predictive dynamics and control but exhibit limited semantic abstraction and weak generalization across

*Corresponding authors

tasks and environments. Therefore, integrating these two paradigms offers a promising path toward open-ended embodied intelligence, where MLLMs provide semantic intent and contextual understanding, and WMs contribute physically grounded prediction and action modeling, together forming a unified framework capable of reasoning, interaction, and adaptation in diverse real-world scenarios.

Despite recent progress, existing attempts to integrate MLLMs with world models remain limited in both scope and depth. Some studies treat the MLLM as an external tool for auxiliary functions such as high-level planning [36] or reward computation [8, 30]. While these methods leverage the semantic priors of MLLMs to assist in high-level decision-making and reward design, the MLLM and WM exhibit misaligned learning objectives and disconnected representations between the semantic and physical domains. A few recent studies, such as GenRL [32] and FOUNDER [44], attempt to address this gap by learning a connector that maps MLLM embeddings into the world model's representation space. However, these approaches still face two critical limitations. (1) They rely on one-way projection functions that transfer information solely from MLLM to the world model. The interaction with the physical environment is still handled exclusively by the world model, and no feedback from environmental dynamics is propagated back to MLLM. (2) The learned connectors are task-agnostic, applying a uniform alignment strategy across all tasks without adapting to task-specific semantics or context-dependent dynamics. The significant differences across tasks lead to task-specific parameter sensitivities, which limit the model's capability in multitask settings and generalization across tasks. These observations motivate us to move beyond static architectural designs and explore a task-aware dynamic joint mechanism that enables bidirectional coupling between MLLMs and world models.

To overcome these limitations, we propose ModularAgent, a task-aware dynamic joint framework that enables modularized coupling between MLLMs and world models. As shown in Fig. 1, the framework establishes two information pathways. In **the forward path**, the MLLM provides high-level semantic and visual representations, such as environment semantics, which are dynamically injected into the latent space of the world model through Task-Aware Fusion mechanism. This allows the world model to perform semantically guided imagination, generating trajectories that are not merely driven by physical transitions but are aligned with high-level semantic intent. In **the backward path**, the MLLM-WM joint model generates task-conditioned imagined trajectories and computes dense rewards that measure semantic–dynamic consistency. Since these rewards are differentiable in the latent space, they can be backpropagated through the MLLM–WM joint loss into the MLLM, enabling its semantic representations to self-correct according to real physical dynamics.

To realize this bidirectional coupling, ModularAgent is composed of three key components: Task-Aware Dynamic Joint Learning, Task-Aware Behavior Learning, and MLLM-WM Joint Optimization. The Task-Aware Dynamic Joint Learning module integrates semantic representations from the MLLM and dynamic representations from the world model through task-conditioned modular fusion, where a gating mechanism adaptively balances the contributions of semantic and dynamic expert adapters at each layer. The Task-Aware Behavior Learning component constructs a shared imagination space for policy learning, in which rollouts are generated to align imagined trajectories with task semantics, and dense semantic consistency rewards guide physically plausible and semantically coherent behavior generation. Finally, the MLLM-WM Joint Optimization unifies semantic alignment, dynamic prediction, and behavior optimization under a single training paradigm, enabling gradient-level coupling between the MLLM and world model. In comprehensive experiments across multi-task and cross-environment generalization settings, our approach consistently outperforms state-of-the-art baselines, exhibiting strong adaptability and stability. To the best of our knowledge, this is the first work to establish a task-aware coupling framework between MLLMs and World Models, paving the way toward open-ended embodied decision making.

Our contributions are summarized as follows:

- We propose ModularAgent, a task-aware dynamic joint framework that enables bidirectional coupling between MLLMs and World Models.
- We introduce a Task-Aware Modular Fusion mechanism that dynamically routes information between semantic and dynamic experts under task guidance, mitigating heterogeneous task interference.
- We develop task-aware behavior learning with a joint optimization objective that aligns the MLLM's semantic space with the WM's imagination dynamics via text-conditioned dense rewards.
- Extensive experiments across multi-task and cross-environment settings show that ModularAgent outperforms state-of-the-art baselines, achieving superior stability and generalization in open-ended embodied scenarios.

2. Related Work

Multimodal Large Language Models (MLLMs). In recent years, the development of MLLMs has greatly advanced artificial intelligence in unified perception and reasoning. Representative models such as GPT-4o [1], Gemini [3], and Claude [4] have demonstrated strong multimodal capabilities in many multi-modal tasks. Meanwhile, open-source models such as LLaVA-Next [29], InternVL [9], Qwen-VL [40], and DeepSeek-VL Janus [11] have made

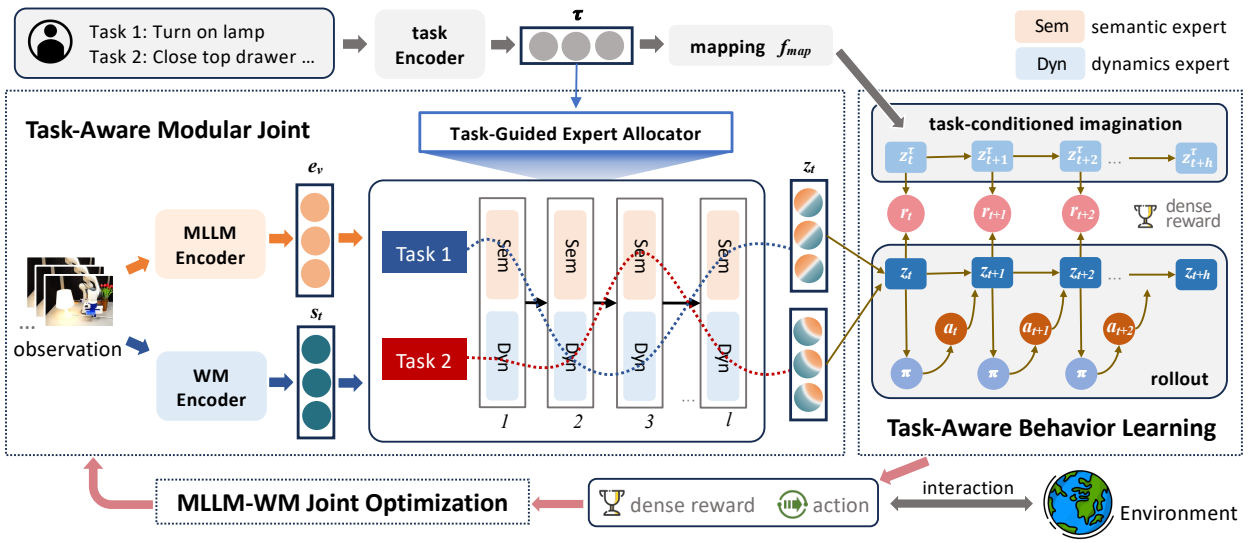

Figure 2. Overview of the proposed ModularAgent framework. It establishes bidirectional coupling between MLLM and World Model. In the forward path, semantic representations produced by MLLM are injected into the WM's latent space via the Task-Aware Modular Fusion mechanism, enabling semantically guided imagination rather than purely physics-driven rollouts. In the backward path, Task-Aware Behavior Learning leverages WM-generated imagined trajectories to compute dense text-aligned rewards, which provide gradient feedback that reshapes the MLLM's semantic space.

multimodal research more open and reproducible. However, current MLLMs lack dynamic interacting with the real physical world, making it hard to link high-level semantics with low-level control and limiting their direct application to embodied intelligence scenarios.

World Models (WMs). World models are a key component of embodied intelligence systems. Their core objective is to learn the latent dynamics of the environment and infer the next state in either a deterministic or probabilistic manner [41, 47]. Recent advances in world models can be broadly grouped into three paradigms. Recurrent State-Space Models (RSSMs) learn latent dynamics for state transitions, where models such as Dreamer [22] and PlaNet [21] perform imagination-based rollouts in latent space to guide reinforcement learning. Joint Embedding Predictive Architectures (JEPAs) emphasize semantic-level consistency across states, for example, I-JEPA modeling stable and generalizable world representations from multimodal inputs [5]. Generative video world models (e.g., Sora) [6] directly capture pixel-level dynamics through large-scale video generation, enabling visual prediction of future scenes. Despite these advances, existing world models still struggle with multi-task generalization and maintaining semantic–dynamics consistency.

Existing MLLM–World Model Integration Paradigms. In recent years, MLLMs have emerged as powerful tools for task decomposition and reward specification in RL. Owing to their strong semantic understanding and reasoning capabilities, MLLMs have been adopted as planners that decompose complex long-horizon tasks into executable subgoals and interact with world models to verify task learning effectiveness. Representative works [10, 13, 16, 48] follow this paradigm, leveraging MLLMs as high-level decision modules to augment the planning capacity of world models.

In terms of reward modeling, the MLLM evaluates the performance of the world model from a visual perspective. Several works [25, 27, 37, 43] compute the semantic similarity between agent states and task descriptions to generate dense reward signals from visual observations, while others [28] leverage MLLMs to provide preference-based feedback for reward model training. These representative studies have shown that it is possible to learn effective rewards without manual design or explicit fine-tuning. However, both the planner-based and reward-based approaches treat MLLMs as external tools that extend only partial capabilities of the world models (e.g., planning or reward generation), without establishing true bidirectional communication between world interaction and MLLM.

To bridge this gap, recent works such as GenRL [32] and FOUNDER [44] attempt to couple MLLMs and WMs within the latent state space by learning a connector that maps foundational MLLM representations into the WM's latent dynamics, thus injecting high-level semantic priors into the world model. Nonetheless, these methods maintain a unidirectional information flow, from MLLM to WM. In contrast, our method achieves bidirectional alignment and

joint optimization between MLLMs and WMs at the latent representation level, enabling embodied agents to perform unified semantic–dynamics reasoning and achieve adaptive decision-making.

3. Methodology

We propose ModularAgent, a task-aware dynamic joint framework that enables modularized coupling between MLLMs and world models. As illustrated in Fig. 2, the framework is composed of three components: Task-Aware Dynamic Joint Learning, Task-Aware Behavior Learning, and MLLM-WM Joint Optimization. We next detail the three core modules and describe how they interact within the complete pipeline of our method.

3.1. Task-Aware Dynamic Joint Learning

We introduce a Dynamic Joint Architecture that seamlessly integrates the MLLM with the World Model into a unified framework. The architecture employs a task-aware modular dynamic strategy to achieve robust generalization across diverse tasks, effectively alleviating architectural conflicts in multi-task learning.

Dynamic Joint Architecture. Given a task instruction y, a video observation sequence $x_{1:t}$, and previous actions a_{t-1}, our framework comprises a Task Encoder, an MLLM Encoder, an RSSM-based World Model, and a Task-Aware Modular Fusion (TAMF) module:

$$\begin{aligned} Task\ Encoder &: \tau = f_{\text{task}}(y), \\ MLLM\ Encoder &: e_v = f_{\text{mllm}}(x_{1:t}), \\ World\ Model\ Encoder &: q_\phi(s_t|x_t), \\ &\quad p_\theta(s_t|s_{t-1}, a_{t-1}), \\ &\quad s_t = f_{\text{wm}}(x_t, s_{t-1}, a_{t-1}), \\ Modular\ Fusion &: z_t = f_{\text{mod}}(e_v, s_t, \tau), \end{aligned} \tag{1}$$

The task encoder transforms textual task descriptions into a global task embedding $\boldsymbol{\tau}$ through f_{task}, which encodes both semantic intent and task identity. This embedding serves as a dynamic routing signal for the subsequent modular layers. The MLLM encoder extracts high-level semantic representations, where f_{mllm} encodes temporally ordered visual tokens to produce semantically consistent visual features. The world model adopts a Recurrent State-Space Model (RSSM) architecture to capture latent dynamics in a temporally coherent form. Its encoder maps the observation x_t to a latent state s_t following the recurrent transition, $s_t = f_{\text{wm}}(x_t, s_{t-1}, a_{t-1})$, which integrates the current observation, previous latent state, and executed action to summarize both observed environmental transitions. The modular fusion jointly encodes semantic and dynamic representations under the guidance of $\boldsymbol{\tau}$ and produces a unified latent representation z_t, forming a task-conditioned feature space for imagination and behavior generation.

Task-Aware Modular Fusion. To bridge the semantic reasoning capability of the MLLM and the dynamic modeling capacity of the world model, we design a Task-Aware Modular Fusion (TAMF) module that dynamically fuses embeddings from MLLM and world model under task guidance. The module consists of L stacked layers, where each layer contains two expert adapters specialized for semantic alignment and dynamics alignment, respectively.

Given the embedding from the MLLM, $\mathbf{e}_v \in \mathbb{R}^{d_m}$, and the latent state from the world model, $\mathbf{s}_t \in \mathbb{R}^{d_s}$, we first compute an initial fused representation:

$$\mathbf{z}^{(0)} = h_{\text{fuse}}([\,\mathbf{e}_v\,;\,\mathbf{s}_t\,])\,, \tag{2}$$

where h_{fuse} is a lightweight projection layer.

Each modular layer $\ell \in \{1, \ldots, L\}$ takes the previous representation $\mathbf{z}^{(\ell-1)}$ and the task embedding $\boldsymbol{\tau}$ as input. A gating controller computes the task-conditioned routing probability, $p^{(\ell)} = f_{gate}(\tau)$, which determines the contribution of each expert branch. Then, the output of layer l in the proposed TAMF is formulated as:

$$\mathbf{z}^{(\ell)} = \mathbf{z}^{(\ell-1)} + (1-p^{(\ell)})\,\mathcal{A}^{(\ell)}_{\text{sem}}(\mathbf{z}^{(\ell-1)}) + p^{(\ell)}\,\mathcal{A}^{(\ell)}_{\text{dyn}}(\mathbf{z}^{(\ell-1)}), \tag{3}$$

where $\mathcal{A}^{(\ell)}_{\text{sem}}$ expert focuses on semantic adaptation to align textual and visual representations; $\mathcal{A}^{(\ell)}_{\text{dyn}}$ expert focuses on dynamics adaptation to integrate physical state transitions.

We draw inspiration from prior work on adapter-based architectures [24, 38], each expert adapter adopt a lightweight adaptation block structured as Pre-LayerNorm $\rightarrow$ GEGLU $\rightarrow$ Linear $\rightarrow$ LayerScale $\rightarrow$ Residual. We incorporate residual connections to mitigate training instability and ensure stable optimization. After L layers of iterative fusion, the model produces a unified latent representation:

$$\mathbf{z}^{(L)} = f_{\text{TAMF}}(\mathbf{e}_v, \mathbf{s}_t, \boldsymbol{\tau}), \tag{4}$$

which jointly encodes multimodal semantics and world dynamics, serving as a compact foundation for reconstruction, prediction, and policy optimization.

Task-Guided Expert Allocator. To enable task-aware feature routing, we introduce a lightweight gating network that transforms the task embedding $\boldsymbol{\tau}$ into a continuous gate $p \in (0, 1)$, which controls the activation of semantic and dynamic experts at each layer. Formally, given the task embedding $\boldsymbol{\tau} \in \mathbb{R}^{d_\tau}$, the gating function is defined as:

$$p = \sigma(W_2\,\text{GELU}(W_1\,\text{LN}(\boldsymbol{\tau}))\,), \tag{5}$$

where $\text{LN}(\cdot)$ denotes Layer Normalization, and $\sigma(\cdot)$ is the sigmoid function. The output p serves as a selection coefficient between the semantic adapter $\mathcal{A}_{\text{sem}}$ and the dynamics

adapter $\mathcal{A}_{\text{dyn}}$, allowing the model to modulate fusion behavior according to the task semantics. For different tasks, the relative weighting between the semantic and dynamics branches is adaptively adjusted, allowing the model to share low-level representations across tasks while preserving task-specific expressiveness. Such dynamic routing mitigates gradient conflicts and interference across heterogeneous tasks, thereby enhancing both performance and generalization in multi-task scenarios.

In addition, unlike conventional single-layer multi-expert designs, the proposed Task-Aware Modular Fusion (TAMF) adopts a layer-wise dual-expert architecture, where each layer contains independent semantic and dynamics adapters, enabling localized routing and adaptive feature blending. This design brings three major benefits: first, independent gating at each layer decomposes gradient propagation into multiple local decisions, leading to clearer gradient signals and more stable convergence; second, each layer allows progressive balancing between high-level semantic abstraction and low-level physical reasoning; and third, the layer-wise routing mechanism provides inherent robustness, an inaccurate gate in one layer affects only local computation without corrupting the overall feature flow.

3.2. Task-Aware Behavior Learning

Building upon the fused latent representation generated by TAMF, this section focuses on how the model leverages z_t to learn task-consistent behavior policies.

Rollout. To enable the fused latent representation $\mathbf{z}_t$ to drive behavior generation, we construct a task-conditioned rollout mechanism within the shared imagination space. At each timestep t, the fused latent representation is computed according to Eq. 4: $\mathbf{z}_t = f_{\text{TAMF}}(\mathbf{e}_v, \mathbf{s}_t, \boldsymbol{\tau})$, where $\mathbf{e}_v$ denotes the multimodal embedding from the MLLM, $\mathbf{s}_t$ is the latent state inferred by the RSSM-based world model $q_\phi(\mathbf{s}_t|\mathbf{x}_t)$, and $\boldsymbol{\tau}$ represents the task embedding produced by the task encoder $f_{\text{task}}(\mathcal{T})$.

Within this latent space, the imagination process unfolds according to the learned dynamics:

$$\begin{aligned} \mathbf{a}_t &\sim \pi_\psi(\mathbf{a}_t|\mathbf{z}_t), \\ \tilde{\mathbf{z}}_{t+1} &\sim p_\theta(\mathbf{z}_{t+1}|\mathbf{z}_t, \mathbf{a}_t), \end{aligned} \tag{6}$$

where π_ψ is the policy network and p_θ is the transition function. Iteratively applying this process for a horizon H yields a sequence of imagined trajectories $\{\tilde{\mathbf{z}}_{t+h}, \mathbf{a}_{t+h}\}_{h=1}^{H}$. Unlike conventional model-based rollouts, our imagination process explicitly incorporates visual information from the MLLM and implicitly depends on the task embedding $\boldsymbol{\tau}$. This design enables the world model to generate task-aligned trajectories that remain semantically consistent throughout the entire temporal evolution.

Task-Conditioned Reward. In behavior modeling, the reward function plays a pivotal role in guiding the agent toward right actions. Unlike traditional reinforcement learning where rewards are manually defined based on environmental outcomes, our framework learns dense semantic rewards directly within the latent imagination space.

Given a task embedding $\boldsymbol{\tau}$, we first map it into the latent state space of the world model through a lightweight projection:

$$\mathbf{z}_t^\tau = f_{\text{map}}(\boldsymbol{\tau}), \tag{7}$$

where f_{map} ensures that the task embedding and the world model state $\mathbf{z}_t$ share the same representational manifold. Conditioned on the current task embedding $\mathbf{z}_t^\tau$ and executed action $\mathbf{a}_t$, the text imagination module predicts the next task-aligned latent state as:

$$\tilde{\mathbf{z}}_{t+1}^\tau \sim \tilde{p}_\psi(\mathbf{z}_{t+1}^\tau \mid \mathbf{z}_t^\tau, \mathbf{a}_t), \tag{8}$$

$\tilde{\mathbf{z}}_{t+1}^\tau$ represents the expected state transition under the given task instruction. This formulation transforms a static task goal into a temporally evolving target trajectory. By iteratively applying this transition over a horizon H, we obtain a sequence of task-conditioned imagined trajectories $\{\tilde{\mathbf{z}}_{t+h}^\tau, \mathbf{a}_{t+h}\}_{h=1}^{H}$.

During training, we encourage the world model to align its predicted latent dynamics with those imagined from task semantics. This is achieved by minimizing the Kullback–Leibler divergence between the world-model rollout distribution and the task-conditioned imagination distribution:

$$D_{\text{KL}}\big(p_\theta(\mathbf{z}_{t+1} \mid \mathbf{z}_t, \mathbf{a}_t) \,\|\, \tilde{p}_\psi(\tilde{\mathbf{z}}_{t+1}^\tau \mid f_{map}(\tau), \mathbf{a}_t)\big)\,, \tag{9}$$

where p_θ denotes the transition dynamics of the world model and $\tilde{p}_\psi$ represents the text-imagination model. This alignment enforces that the imagined trajectories are both physically consistent and semantically coherent with the given instruction.

When computing rewards for policy optimization, we measure the semantic consistency between the imagined state and the task-imagined reference as:

$$r_t = \text{Sim}(\mathbf{z}_t, \tilde{\mathbf{z}}_t^\tau)\,, \tag{10}$$

where $\text{Sim}(\cdot)$ denotes cosine similarity. This dense reward provides a smooth learning signal, guiding the policy toward behaviors that align with task semantics while maintaining physical plausibility.

3.3. MLLM-WM Joint Optimization

We jointly optimize the World Model and MLLM through a unified objective that enables bidirectional coupling between semantic reasoning and dynamic prediction. The overall training objective is formulated as:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{WM}}\mathcal{L}_{\text{WM}} + \lambda_{\text{MLLM}}\mathcal{L}_{\text{MLLM}} + \lambda_{\text{JBO}}\mathcal{L}_{\text{JBO}}, \tag{11}$$

where $\mathcal{L}_{\text{WM}}$ focuses on reconstructing and predicting environmental dynamics, $\mathcal{L}_{\text{MLLM}}$ ensures semantic-level representation alignment, and Joint Behavior Optimization Loss $\mathcal{L}_{\text{JBO}}$ enforces task-aware joint optimization between semantic cues and dynamic rollouts.

World Model Loss. The world model learns to capture latent environment dynamics through a combination of prior–posterior consistency and observation reconstruction. Following the recurrent state-space modeling (RSSM) formulation [22, 32], the overall loss is defined as:

$$\mathcal{L}_{\text{WM}} = \sum_t \underbrace{D_{\text{KL}}[q_\phi(\mathbf{z}_t \mid x_t) \,\|\, p_\theta(\mathbf{z}_t \mid \mathbf{z}_{t-1}, a_{t-1})]}_{\text{dynamics loss}} - \underbrace{\mathbb{E}_{q_\phi(\mathbf{z}_t|x_t)}[\log p_\theta(x_t \mid \mathbf{z}_t)]}_{\text{reconstruction loss}}. \tag{12}$$

Here, p_θ and q_ϕ denote the prior and posterior latent distributions of the RSSM, respectively. The first term enforces temporal consistency by aligning the predicted prior with the inferred posterior, while the second term reconstructs multimodal observations from latent states. Together, these objectives enable the world model to learn a compact and predictive imagination space.

Multimodal Semantic Loss. The MLLM branch is trained to enhance controllable semantic representations through semantic reconstruction and cross-modal alignment:

$$\mathcal{L}_{\text{MLLM}} = \underbrace{\| \mathbf{e}_v - f_{dec}(z_t) \|_2^2}_{\text{reconstruction loss}} + \underbrace{\| \mathbf{e}_v - f_\psi(\tau) \|_2^2}_{\text{alignment loss}}, \tag{13}$$

where $\mathbf{e}_v$ denote the visual embeddings, f_ψ is the representation aligner that maps linguistic features into the visual space, and f_{dec} serves as the decoder that maps $\mathbf{z}_t$ back into the visual space. This loss encourages visual reconstruction and cross-modal alignment.

Joint Behavior Optimization Loss. To align the agent's imagined behavior with task semantics, according to Eq. 9, we define a task-conditioned joint behavior optimization objective:

$$\mathcal{L}_{\text{JBO}} = -\mathbb{E}_t\left[w_{t+h} \cdot \text{Sim}\left(\mathbf{z}_{t+h}, \tilde{\mathbf{z}}^\tau_{t+h}\right)\right], \quad w_{t+h} = \gamma^{t+h}, \tag{14}$$

where $\mathbf{z}_{t+h}$ denotes the latent trajectory generated by the world model, and $\tilde{\mathbf{z}}^\tau_{t+h}$ represents the text-conditioned imagination trajectory. The similarity function $\text{Sim}(\cdot)$ is implemented as negative KL divergence. Following DreamerV3 [22], we use a discount weight w_{t+h} to down-weight long-horizon imagination, thereby prioritizing reliable near-term predictions and reducing uncertainty accumulation. This loss bridges semantic and physical spaces, encouraging the model to generate behaviors that are both dynamically coherent and semantically aligned with the task.

4. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of our method. Specifically, we aim to address three key questions: (1) how well the model performs across multiple tasks within a single training environment, (2) how effectively it generalizes to unseen environments when performing the same task, and (3) how each component contributes to the overall performance improvement.

4.1. Experimental Setup

Experimental Environments. Our evaluation covers four locomotion control environments (Cheetah, Walker, Quadruped, and Stickman), all implemented on the DeepMind Control Suite [39] frameworks. Following GenRL, we construct offline datasets collected using the Plan2Explore strategy [35] and replay buffers from reinforcement learning agents trained on domain-specific tasks, encompassing diverse semantic instructions and action trajectories. Since existing work [32] does not include the Quadruped environment, we additionally collect and preprocess it. Detailed task definitions and task prompts are provided in the Appendix.

Baseline. We compare our proposed approach against both model-free and model-based reinforcement learning methods. For the model-free baselines, we adopt three representative algorithms: the off-policy RL method TD3 [19], the advantage-weighted behavior cloning method IQL [26], and the behavior-regularized approach TD3+BC [18]. For the model-based baselines, we focus on three recent methods that also explore integrating MLLMs with World Model, such as WM-CLIP, GenRL [32], and FOUNDER [44]. WM-CLIP is a variant of GenRL that learns a reversed connector, mapping latent states from the WM to the MLLM embedding space. In contrast, GenRL and FOUNDER employ a forward connector that maps representations from the MLLM to the WM latent space, allowing semantic priors to guide imagination and planning within the world model. Additionally, FOUNDER introduces a Temporal Distance Predictor to estimate temporally consistent reward signals during behavior learning.

Experimental details. In our experimental setup, we pretrain the world model and its associated components for 100K gradient steps, followed by another 50K updates during the behavior learning phase. To ensure a fair comparison, our method and all model-based baselines employ the

Table 1. Performance comparison between our method and all baselines on the DMC. Reported scores are the mean episodic rewards over 10 random seeds (± standard error), normalized using min–max scaling where the random policy corresponds to the minimum and the expert policy to the maximum.

Task	IQL	TD3+BC	TD3	WM-CLIP	GenRL	FOUNDER	ModularAgent
walker stand	0.66 ± 0.05	0.64 ± 0.03	1.01 ± 0.00	0.94 ± 0.01	1.02 ± 0.00	1.01 ± 0.02	1.03 ± 0.02
walker run	0.29 ± 0.02	0.24 ± 0.02	0.35 ± 0.01	0.70 ± 0.01	0.77 ± 0.02	0.78 ± 0.04	0.87 ± 0.02
walker walk	0.40 ± 0.03	0.44 ± 0.03	0.88 ± 0.02	0.91 ± 0.02	1.01 ± 0.00	0.94 ± 0.04	1.03 ± 0.01
cheetah run	0.15 ± 0.02	-0.01 ± 0.00	0.37 ± 0.01	0.56 ± 0.03	0.74 ± 0.01	0.81 ± 0.02	0.79 ± 0.02
quadruped stand	0.52 ± 0.06	0.43 ± 0.05	0.61 ± 0.05	0.97 ± 0.00	0.97 ± 0.00	0.98 ± 0.01	1.00 ± 0.01
quadruped run	0.38 ± 0.03	0.25 ± 0.02	0.26 ± 0.01	0.61 ± 0.02	0.86 ± 0.02	0.94 ± 0.03	0.95 ± 0.02
quadruped walk	0.32 ± 0.02	0.28 ± 0.04	0.28 ± 0.02	0.92 ± 0.01	0.93 ± 0.01	0.90 ± 0.05	0.99 ± 0.03
stickman stand	0.43 ± 0.04	0.45 ± 0.05	0.08 ± 0.02	0.32 ± 0.01	0.70 ± 0.02	0.91 ± 0.04	0.95 ± 0.03
stickman walk	0.51 ± 0.02	0.46 ± 0.03	0.41 ± 0.02	0.65 ± 0.05	0.83 ± 0.01	0.91 ± 0.03	0.95 ± 0.03
stickman run	0.23 ± 0.02	0.19 ± 0.02	0.21 ± 0.00	0.35 ± 0.01	0.35 ± 0.01	0.48 ± 0.02	0.49 ± 0.02
overall	0.39 ± 0.03	0.34 ± 0.03	0.45 ± 0.02	0.69 ± 0.02	0.82 ± 0.01	0.87 ± 0.03	0.91 ± 0.02

same video–language backbone, InternVideo2 [42]. The visual observations are rendered at a resolution of 64×64, with a batch size of 64 and a sequence length of 32. Additional hyperparameter and training details are provided in the supplementary material.

4.2. Task Solving on DMC

We evaluate our method on DMC to assess its capability in multi-task understanding and execution within a single environment. For DMC, we select three representative locomotion tasks including Stand, Walk, and Run, and test them across multiple embodiments.

As shown in Table 1, model-free baselines generally perform worse. Since they do not explicitly model environmental dynamics and instead learn policies or value functions directly from experience data, these methods struggle with long-term temporal consistency.

Compared with other model-based methods, our approach shows clear advantages. ModularAgent achieves the highest scores in 9 out of 10 tasks, demonstrating the strongest overall performance. WM-CLIP learns a one-way mapping from world model representations to MLLM embeddings, while GenRL and FOUNDER adopt the reverse direction from MLLM to WM. However, both directions in isolation fail to capture the complementary nature of semantic reasoning and physical dynamics. Our joint modular fusion enables bidirectional coupling within a unified latent space, integrating the semantic abstraction capability of MLLMs with the actionable physical modeling of WMs, resulting in more accurate decision-making.

Notably, ModularAgent maintains strong performance on tasks that are already well-solved by existing methods, while delivering substantial improvements on more challenging tasks within the same environment. For example, in the Walker domain, ModularAgent preserves competitive performance on stand and walk, while boosting the run task from 0.78 to 0.87. These results indicate that the proposed task-aware modular fusion effectively alleviates task-specific architectural conflicts, enabling the model to retain task-unique information while preserving overall stability across diverse tasks.

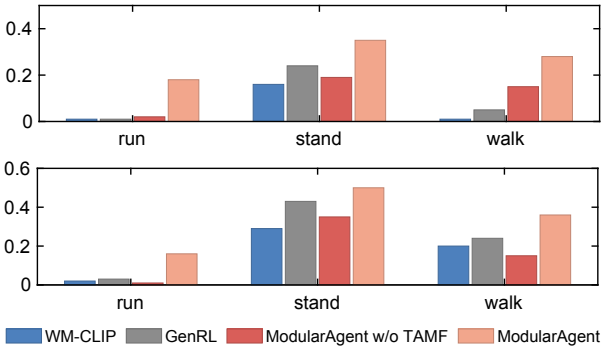

Figure 3. Cross-environment generalization of agents trained in the Walker domain. The top row shows evaluation results in the Quadruped environment, while the bottom row presents results in the Stickman environment.

4.3. Cross-domain Task Solving

We further evaluate the cross-domain generalization capability of our method using three representative tasks: Run, Stand, and Walk. For each experiment, one environment is used as the source domain for training, and the remaining two environments serve as target domains for transfer evaluation. This configuration produces six domain combinations and eighteen task settings, enabling a comprehensive assessment. Results are shown in Fig. 3, and additional comparisons are provided in the supplementary materials.

Since the official implementation of FOUNDER has not been released, we compare our method with two accessible model-based baselines, WM-CLIP and GenRL. The results show that our approach achieves the best performance on all

Table 2. The effectiveness of components.

Method	walker stand	walker run	walker walk
$base$	0.89 ± 0.03	0.67 ± 0.01	0.86 ± 0.03
$+L_{MLLM}$	0.96 ± 0.02	0.76 ± 0.02	0.90 ± 0.03
$+TAMF$	1.00 ± 0.03	0.83 ± 0.01	1.00 ± 0.01
$+TARO$	1.03 ± 0.02	0.87 ± 0.02	1.02 ± 0.01

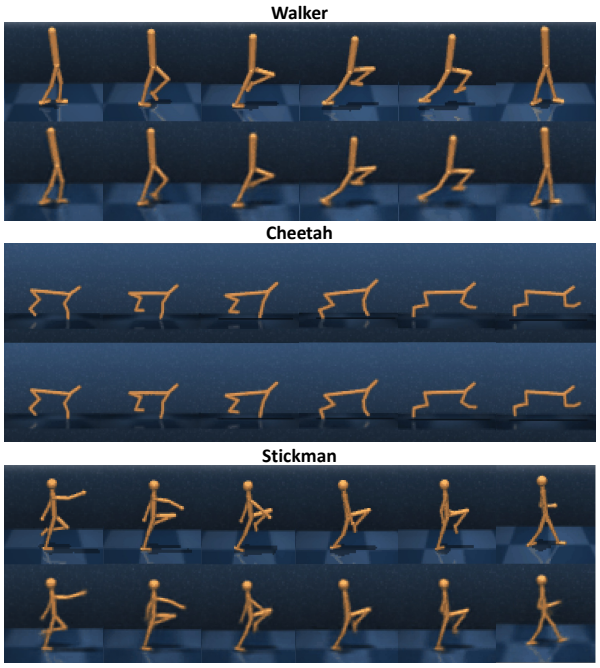

Figure 4. Visualization of task-conditioned imagined trajectories. For each environment, the top row shows real observations and the bottom row shows reconstructions decoded from the task-conditioned imagination trajectories.

of tasks, demonstrating superior cross-domain adaptability.

These performance gains can be attributed to the TAMF module, which effectively fuses semantic and physical representations through adaptive modular routing. In addition, the modular routing mechanism enables stable sharing of low-level parameters while allowing higher-level components to adapt to task-specific requirements. As shown in the figure, the superior performance of ModularAgent over the without-TAMF variant further supports this observation.

4.4. Effectiveness of each Component.

As shown in Table 2, we conduct a comprehensive ablation study to examine the contribution of each key component in our framework across three tasks in the Walker environment: Stand, Walk, and Run.

- Base Model. The Base Model adopts a GenRL-style mapping from the MLLM representation to the world model's latent space, where the fused feature is obtained by direct summation of the two embeddings. Trained solely under the world model loss $\mathcal{L}_{\text{WM}}$, it achieves average scores of 0.89, 0.67, and 0.86 on the three tasks, respectively, serving as a minimal baseline.
- + MLLM Loss. Introducing the multimodal reconstruction loss $\mathcal{L}_{\text{MLLM}}$ improves semantic alignment between MLLM and WM representations. The scores increase to 0.96, 0.76, and 0.90, showing that language–vision alignment benefits task performance, especially in semantically guided behaviors.
- + Task-Aware Modular Fusion. By adding the proposed TAMF, performance further improves to 1.00, 0.83, and 1.00. This enhancement demonstrates that dynamically routing semantic and physical features per task effectively mitigates task interference.
- + MLLM-WM Joint Optimization. Finally, incorporating the MLLM-WM joint optimization achieves the best results, 1.03, 0.87, and 1.02. The dense, text-conditioned reward provides stable supervision in the imagination space, allowing the model to refine its behavior generation through semantic–dynamic alignment.

4.5. Effectiveness of behavior learning.

To evaluate the effectiveness of task-aware behavior learning, we visualize the decoded task-conditioned imagination trajectories, as shown in Fig. 4. For each environment, the first row displays the real observations, while the second row shows the reconstructed results from the imagined trajectories. We observe that, after behavior learning, the task-conditioned imagination trajectories closely approximate the real observations. This provides strong evidence that using these task-conditioned imagined trajectories as reference targets during reward computation is effective.

5. Conclusions

In this paper, we propose ModularAgent, a task-aware dynamic joint framework that enables bidirectional coupling between MLLMs and World Models. ModularAgent establishes two complementary pathways: a forward path, where MLLM-derived visual representations are dynamically injected into the WM's latent space, and a backward path, where WM-generated rollouts and text-conditioned rewards provide feedback to refine the MLLM's semantic space. To realize this coupling, ModularAgent integrates three key components: Task-Aware Dynamic Joint Learning, which adaptively fuses semantic and dynamic representations under task guidance; Task-Aware Behavior Learning, which aligns imagined trajectories with task semantics through dense text-conditioned rewards; and Joint Optimization Objectives, which unify semantic reasoning and physical prediction under a single training paradigm. Extensive experiments across multi-task and cross-environment settings demonstrate that ModularAgent achieves superior performance compared to state-of-the-art baselines.

Acknowledgements

This work was supported in part by the China Postdoctoral Science Foundation No. 2025M771576 and the Beijing National Research Center for Information Science and Technology under Grant No. BNR2023TD03006.

References

[1] 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

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Igor Barr, Yana Hasson, Arthur Mensch, Katie Millican, Malcolm Reynolds, Dieter Ruesch, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, 2022. 1

[3] Rohan Anil et al. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[4] Anthropic. The claude 3 model family: Opus, sonnet, haiku (model card), 2024. Model card / technical report. 2

[5] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 3

[6] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024. 3

[7] Tim Brooks, Saining Xie, et al. Genie: Generative interactive environments. In *Advances in Neural Information Processing Systems*, 2024. 1

[8] Huayu Chen, Guande He, Lifan Yuan, Ganqu Cui, Hang Su, and Jun Zhu. Noise contrastive alignment of language models with explicit rewards. *Advances in Neural Information Processing Systems*, 37:117784–117812, 2024. 2

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 2

[10] Xiaowei Chi, Chun-Kai Fan, Hengyuan Zhang, Xingqun Qi, Rongyu Zhang, Anthony Chen, Chi-Min Chan, Wei Xue, Qifeng Liu, Shanghang Zhang, et al. Empowering world models with reflection for embodied video prediction. In *Proceedings of the International Conference on Machine Learning*. 3

[11] DeepSeek-AI. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 2

[12] Danny Driess, Fei Xia, Pratyusha Srinivasan, Tianhe Yu, Jonathan Gehring, C. Karen Zhao, Xinying Chen, Parth Raghunathan, Marcin Andrychowicz, Julian Ibarz, et al. Palm-e: An embodied multimodal language model. In *Proceedings of the International Conference on Machine Learning*, 2023. 1

[13] Lutfi Eren Erdogan, Hiroki Furuta, Sehoon Kim, Nicholas Lee, Suhong Moon, Gopala Anumanchipalli, Kurt Keutzer, and Amir Gholami. Plan-and-act: Improving planning of agents for long-horizon tasks. In *Proceedings of the International Conference on Machine Learning*, 2025. 3

[14] Tongtong Feng, Xin Wang, Yu-Gang Jiang, and Wenwu Zhu. Embodied ai: From llms to world models. *arXiv preprint arXiv:2509.20021*, 2025. 1

[15] Tongtong Feng, Xin Wang, Zekai Zhou, Ren Wang, Yuwei Zhan, Guangyao Li, Qing Li, and Wenwu Zhu. Evoagent: Agent autonomous evolution with continual world model for long-horizon tasks. *arXiv e-prints*, pages arXiv–2502, 2025. 1

[16] Tongtong Feng, Xin Wang, Feilin Han, Leping Zhang, and Wenwu Zhu. U2udata+: A scalable swarm uavs autonomous flight dataset for embodied long-horizon tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1792–1800, 2026. 3

[17] Tongtong Feng, Xin Wang, and Wenwu Zhu. Self-evolving embodied ai. *arXiv preprint arXiv:2602.04411*, 2026. 1

[18] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021. 6

[19] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *Proceedings of the International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018. 6

[20] David Ha and Jürgen Schmidhuber. World models. In *Advances in Neural Information Processing Systems*, 2018. 1

[21] Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*, pages 2555–2565, 2019. 3

[22] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, et al. Mastering diverse control tasks through world models. *Nature*, 640: 647–653, 2025. 3, 6

[23] Nicklas Hansen, Wei Suo, Michael Laskin, Pieter Abbeel, and Vikash Kumar. Temporal difference model predictive control. In *Proceedings of the International Conference on Machine Learning*, 2022. 1

[24] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bryan Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *Proceedings of the International Conference on Machine Learning*, 2019. 4

[25] Martin Klissarov, Pierluca D'Oro, Shagun Sodhani, Roberta Raileanu, Pierre-Luc Bacon, Pascal Vincent, Amy Zhang, and Mikael Henaff. Motif: Intrinsic motivation from artificial intelligence feedback. *arXiv preprint arXiv:2310.00166*, 2023. 3

[26] Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *arXiv preprint arXiv:2110.06169*, 2021. 6

[27] Minae Kwon, Sang Michael Xie, Kalesha Bullard, and Dorsa Sadigh. Reward design with language models. *arXiv preprint arXiv:2303.00001*, 2023. 3

[28] Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. Rlaif vs. rlhf: scaling reinforcement learning from human feedback with ai feedback. In *Proceedings of the International Conference on Machine Learning*, 2024. 3

[29] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024. 2

[30] Hao Li, Xue Yang, Zhaokai Wang, Xizhou Zhu, Jie Zhou, Yu Qiao, Xiaogang Wang, Hongsheng Li, Lewei Lu, and Jifeng Dai. Auto mc-reward: Automated dense reward design with large language models for minecraft. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16426–16435, 2024. 2

[31] Haotian Liu, Chunyuan Li, Qiang Xu, and Yong Jae Li. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, 2023. 1

[32] Pietro Mazzaglia, Tim Verbelen, Bart Dhoedt, Aaron Courville, and Sai Rajeswar. Genrl: Multimodal-foundation world models for generalization in embodied agents. *Advances in neural information processing systems*, 37:27529–27555, 2024. 2, 3, 6

[33] Zheyun Qin, Xiankai Lu, Xiushan Nie, Dongfang Liu, Yilong Yin, and Wenguan Wang. Coarse-to-fine video instance segmentation with factorized conditional appearance flows. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1192–1208, 2023. 1

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 1

[35] Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the International Conference on Machine Learning*, pages 8583–8592. PMLR, 2020. 6

[36] Chan Hee Song, Jiaman Wu, Clayton Washington, Brian M Sadler, Wei-Lun Chao, and Yu Su. Llm-planner: Few-shot grounded planning for embodied agents with large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2998–3009, 2023. 2

[37] Sumedh Sontakke, Jesse Zhang, Séb Arnold, Karl Pertsch, Erdem Bıyık, Dorsa Sadigh, Chelsea Finn, and Laurent Itti. Roboclip: One demonstration is enough to learn robot policies. *Advances in Neural Information Processing Systems*, 36:55681–55693, 2023. 3

[38] Yi-Lin Sung, Arsha Nagrani, Anurag Arnab, Shangbang Li, and Cordelia Schmid. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 4

[39] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. Deepmind control suite. *arXiv preprint arXiv:1801.00690*, 2018. 6

[40] Qwen Team. Qwen2-vl: Enhancing vision-language models with stronger reasoning and understanding. *arXiv preprint arXiv:2407.10671*, 2024. 2

[41] Ren Wang, Xin Wang, Tongtong Feng, Xinyue Gong, Guangyao Li, Yu-Wei Zhan, Qing Li, and Wenwu Zhu. Improving compositional generalization in cross-embodiment learning via mixture of disentangled prototypes. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 7162–7171, 2025. 3

[42] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *Proceedings of the European Conference on Computer Vision*, pages 396–416. Springer, 2024. 7

[43] Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. Rl-vlm-f: Reinforcement learning from vision language foundation model feedback. *arXiv preprint arXiv:2402.03681*, 2024. 3

[44] Yucen Wang, Rui Yu, Shenghua Wan, Le Gan, and De-Chuan Zhan. Founder: Grounding foundation models in world models for open-ended embodied decision making. In *Proceedings of the International Conference on Machine Learning*, 2025. 2, 3, 6

[45] Yu-Wei Zhan, Fan Liu, Xin Luo, Xin-Shun Xu, Liqiang Nie, and Mohan Kankanhalli. Enhancing hoi detection with contextual cues from large vision-language models. In *Proceedings of the ACM International Conference on Multimedia*, pages 8557–8566, 2025. 1

[46] Peng-Fei Zhang, Zi Huang, and Guangdong Bai. Universal adversarial perturbations for vision-language pre-trained models. In *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 862–871, 2024. 1

[47] Peng-Fei Zhang, Ying Cheng, Xiaofan Sun, Shijie Wang, Lei Zhu, and Heng Tao Shen. A step toward world models: A survey on robotic manipulation. *arXiv preprint arXiv:2511.02097*, 2025. 3

[48] 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. 3

[49] 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*, pages 2165–2183. PMLR, 2023. 1