VPA-WM: Vision-Priors-Aligned World Models for Robust Visual Reinforcement Learning

Xu Zhang, Sicong Liu, Liwei Guo, Chenjuan Guo,
Bin Yang, and Yang Shu*

East China Normal University, Shanghai, China
{xuzhang0109,sicongliu1014,riweguo}@gmail.com,
{cjguo,byang,yshu}@dase.ecnu.edu.cn

Abstract. Building robust world models for visual reinforcement learning remains challenging under complex visual conditions. When faced with task-irrelevant visual distractions, the representations learned by world models tend to collapse. Pre-trained vision models (PVMs) provide strong visual priors and structured representations, yet those learned for visual perception are incompatible with dynamic control in RL. In this work, we present **V**ision-**P**riors-**A**ligned **W**orld **M**odel (VPA-WM[1]): a unified framework that organically integrates PVMs and world models to achieve robust visual control. We first introduce Control-Oriented Alignment, which effectively guides PVMs to focus on control-relevant signals through two auxiliary objectives: instant action prediction and short-horizon return estimation. We then design Spatio-Temporal Collaborative Integration mechanism, which guides the world model to adaptively attend to the visual cues of PVMs through two complementary designs (spatially grounded guidance and experience query integration), thereby learning representations that are visually robust and dynamically consistent. Superior performance on visually-distracted control benchmarks demonstrates that VPA-WM effectively bridges the semantic gap between visual perception and dynamic control, and successfully leverages visual priors to enhance the robust control of world models.

Keywords: Visual Reinforcement Learning · World Model · Pre-trained Vision Model

1 Introduction

Learning predictive models of an environment's dynamics from raw pixels has become an efficient approach for learning an optimal vision-action controller. Model-based reinforcement learning (MBRL) methods train an internal world model that imagines future trajectories for policy optimization. Compared with model-free alternatives, MBRL methods enable effective policy learning with orders of magnitude fewer environment interactions [6,8–10]. And these methods

* Corresponding author.

[1] Code will be updated at https://github.com/decisionintelligence/VPA-WM.

have achieved excellent empirical results across multiple domains of modern visual reinforcement learning (VRL).

Despite these successes, building robust world models in complex and noisy visual environments remains a substantial open problem. World models are expected to abstract latent representations highly relevant to visual control tasks from raw pixels. But when pixel observations contain task-irrelevant visual distractions, the learned representations are prone to be dominated by those distractions. General reconstruction objectives encourage world models to model all pixels, reducing their ability to isolate controllable factors from distractions. Benchmarks with complex visual perturbations [21, 25] have demonstrated that even SOTA visual world models suffer severe degradation. Hence, modeling task-relevant latent dynamics under visual perturbations remains a central problem for robust VRL.

Modern large-scale pre-trained vision models (PVMs), trained on massive and diverse real-world visual datasets, have demonstrated strong generalization and transfer across domains. For example, the DINO family [1, 20] demonstrates strong object-centric representations; CLIP [23] aligns visual concepts with language-level semantics; MAE [13] learns robust structural representations by reconstructing missing visual tokens. These PVMs tend to learn stable and semantically meaningful representations that are less sensitive to incidental appearance changes (e.g., background, illumination, texture). This common capability suggests a natural and promising direction: leveraging PVMs as structured and disturbance-robust visual priors for world models, so as to mitigate the fragility of world models' representation learning under visual perturbations.

However, realizing this integration is non-trivial, as two major challenges remain to be solved. One is the mismatch between the learning objectives of PVMs and world models. The priors embedded in PVMs often emphasize semantic invariance and entity categorization, whereas the representations learned by world models focus on the dynamics and reward information required for control tasks. Utilizing these visual priors without adjustment may have potential adverse impacts on the learning of world models. Another challenge is how to integrate visual priors into the training process of world models. World models rely on smoothly evolving latent dynamics to capture agent's contingent transitions. So we have to maintain the coherence of dynamic control when introducing static visual priors into world models. Simple feature concatenation or replacement cannot achieve this. These two issues form the core difficulty of integrating pre-trained visual priors into the learning of robust world models.

To overcome these challenges, we propose a unified framework named **V**ision-**P**riors-**A**ligned **W**orld **M**odel (VPA-WM) , which dynamically integrates fine-tuned pre-trained visual priors into the world model learning process. Our approach is built on two key components. Firstly, we introduce **Control-Oriented Alignment**: instead of directly using frozen visual perception outputs, we fine-tune the PVM through auxiliary learning objectives—instant action capture and short-horizon return estimation—to enable the PVM to consider control-related factors. Secondly, we introduce a **Spatio-Temporal Collaborative Integra-**

tion mechanism that establishes complementary interactions at both spatial and temporal levels between the PVM and the world model. At the spatial level, the world model encoder receives grounded guidance from the PVM, which enables it to focus on task-relevant spatial structures while ignoring transient distractions when encoding pixels. At the temporal level, the latent dynamic model of the world model adaptively queries structured semantics from the PVM, allowing it to supplement task-required details based on its own historical experience without disrupting the dynamic coherence of the learned latent states. Furthermore, we introduce Adaptive alignment weighting to maintain a principled balance between spatial perceptual grounding and dynamic coherence during training.

In summary, our contributions are threefold:

- We propose a unified framework, named VPA-WM, that integrates pre-trained vision priors with world model learning for robust VRL.
- We develop the Control-Oriented Alignment scheme, which effectively guides PVMs to focus on control-relevant signals. We then introduce the Spatio-Temporal Collaborative Integration mechanism that guides the world model to adaptively attend to the visual cues of PVMs, thereby learning representations that are visually robust and dynamically consistent.
- Empirical results demonstrate that our method improves sample efficiency, robustness, and generalization of agents' training on visually distracting control tasks, thereby bridging the gap between pre-trained vision priors and world models for robust visual MBRL.

2 Related Work

2.1 Visual Model-Based Reinforcement Learning

Visual Model-Based Reinforcement Learning (VMBRL) enables policy optimization from high-dimensional observations by learning latent world models, thereby significantly enhancing sample efficiency. The prevailing paradigm, established by PlaNet [8] and the Dreamer series [7, 9, 10], utilizes RSSMs to model latent dynamics via image reconstruction, allowing agents to learn from imagined trajectories. However, the objective of exact pixel-level reconstruction often conflicts with the effective learning of task-relevant dynamics, particularly in noisy or visually complex environments [22, 26]. To mitigate these limitations, subsequent research has focused on disentangling controllable states from spurious variations without relying on high-fidelity reconstruction. Approaches such as Iso-Dream [22] and Denoised MDP [29] introduce inverse dynamics and factorized latent structures to isolate task-relevant information. Similarly, RePo [34] optimizes latent representations by maximizing predictive mutual information to filter out task-irrelevant details. Notably, MInCo [26] alleviates the interference between representation learning and dynamics modeling by incorporating negative-free contrastive learning alongside time-varying KL regularization. Concurrently, other works have explored contrastive objectives such as Curled-Dreamer [17] and DreamerPro [3], as well as task-informed abstractions [5], to

further improve robustness. Diverging further from visual generation, the TD-MPC framework [11,12] advances a decoder-free paradigm that integrates latent MPC with terminal value estimation, demonstrating that implicit world models can scale effectively across diverse embodiments without the computational burden of reconstruction. Despite these advancements, existing VMBRL methods predominantly rely on learning task-agnostic features from scratch, largely overlooking the potential of leveraging structured visual priors.

2.2 Leveraging Pretrained Visual Representations

Integrating PVMs into reinforcement learning offers a promising avenue to enhance sample efficiency and generalization. Early Model-Free RL studies, such as R3M [19] and PIE-G [31], demonstrated that transferring visual priors effectively reduces overfitting to training environments. Extending this to model-based reinforcement learning, approaches like DINO-WM [33] and SAM-G [30] have leveraged features from foundation models such as DINOv2 [20] and the Segment Anything Model, SAM [18], to structure latent dynamics. However, these methods often suffer from limited adaptability due to reliance on frozen embeddings or rigid segmentation prompts, which are ill-suited for dynamic, contact-rich tasks. Furthermore, recent analyses identify a fundamental objective misalignment between visual recognition and control: PVMs often fail to encode reward-relevant signals, leading to inferior performance compared to scratch-learned representations [24]. While partial fine-tuning can improve robustness against distribution shifts [15], balancing the retention of visual priors with adaptation to task-specific dynamics remains unresolved. Consequently, a critical gap exists in unifying PVM semantics with the temporal coherence and reward structures of world models without inducing catastrophic forgetting.

3 Preliminaries

3.1 Partially Observable Markovian Decision Process (POMDP)

A Partially Observable Markov Decision Process (POMDP) can be defined as a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{Z}, P, P_0, \gamma, r)$, where $\mathcal{S}$ is the state space and $\mathcal{O}$ is the observation space. The agent uses observations $\mathcal{O}$ to sample actions from the action space $\mathcal{A}$. Every time the agent interacts with the environment, it obtains a reward $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ and the environment transforms to the next state based on the transition function $P = p(s_{t+1}|s_t, a_t)$. $\mathcal{Z} = p(o_t|s_t)$ is the observation model, which is usually used to get the posterior state distribution conditioned on the observation, and P_0 is the initial state distribution. The end goal is to train a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the sum of expected returns $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t)|s_0 = s]$, where $\gamma \in [0, 1)$ is the discount factor, $s_0 \sim P_0$, and r_t is the reward at time t.

3.2 Visual Model-Based Reinforcement Learning

Many VRL tasks can be formalized as a POMDP. In addressing these tasks, Dreamer [7] is one of the most influential and representative methods in visual MBRL. It leverages a recurrent state-space model (RSSM) [8] to tackle the challenge of partial observability. In general, the RSSM mainly consists of the following modules:

$$\begin{aligned} &\text{Representation Model:} && p_\theta(s_t|s_{t-1}, a_{t-1}, o_t) \\ &\text{Observation Model:} && q_\theta(o_t|s_t) \\ &\text{Reward Model:} && q_\theta(r_t|s_t) \\ &\text{Transition Model:} && q_\theta(s_t|s_{t-1}, a_{t-1}). \end{aligned} \tag{1}$$

Under this formulation, the dynamics modeling of the POMDP can be formulated by maximizing the evidence lower bound (ELBO) [16]. Equivalently, Dreamer can be viewed as minimizing the negative ELBO, whose loss consists of three parts: (a) observation reconstruction loss, (b) reward prediction loss, and (c) a KL regularization term that enforces posterior–prior consistency:

$$\begin{aligned} \mathcal{L}_{\mathrm{O}}^t \doteq -\ln q_\theta(o_t|s_t), \qquad \mathcal{L}_{\mathrm{R}}^t \doteq -\ln q_\theta(r_t|s_t), \\ \mathcal{L}_{\mathrm{KL}}^t \doteq \mathrm{KL}\big(p_\theta(s_t|s_{t-1}, a_{t-1}, o_t) \,\|\, q_\theta(s_t|s_{t-1}, a_{t-1})\big). \end{aligned} \tag{2}$$

The overall world model objective can be written as:

$$\mathcal{L}_{\mathrm{Dreamer}} \doteq \mathbb{E}_p \left[\sum_t \left(\mathcal{L}_{\mathrm{O}}^t + \mathcal{L}_{\mathrm{R}}^t + \lambda_t \cdot \mathcal{L}_{\mathrm{KL}}^t \right) \right], \tag{3}$$

where λ_t is a (typically constant) coefficient in Dreamer; in our Method section we reuse $\mathcal{L}_{\mathrm{R}}$ and $\mathcal{L}_{\mathrm{KL}}$ to keep notation consistent. Dreamer utilizes the learned dynamics model to imagine multiple trajectories and trains the policy by the actor-critic method on these synthetic transition data:

$$\begin{aligned} &\text{Action Model:} && a_\tau \sim q_\phi(a_\tau|s_\tau) \\ &\text{Value Model:} && v_\psi \approx \mathbb{E}[q(\cdot|s_\tau) \textstyle\sum_{\tau=t}^{t+H} \gamma^{\tau-t} r_\tau]. \end{aligned} \tag{4}$$

Specifically, during policy learning, the dynamics model remains frozen. Dreamer imagines the latent dynamics of a fixed H horizon in the latent space. The value model approximates the truncated λ-return [27] through a regression loss, and the action model learns to maximize the sum of values over the future H steps.

4 Method

4.1 Overview of VPA-WM

In our VPA-WM framework, shown in Fig. 1, we integrate structured visual priors from fine-tuned PVM into the world model to maintain robust dynamic modeling from noisy visual observations. Concretely, VPA-WM consists of two collaborative components: (1) Control-Oriented Alignment (COA) and (2) Spatio-Temporal Collaborative Integration (STCI). During training, PVM first undergoes lightweight auxiliary objectives in COA to endow its purely visual perception features with dynamic control-related semantics. Then, the adapted PVM

visual priors are integrated into the world model learning process at both spatial and temporal levels. The world model receives instantaneous visual perception cues at the feature level, and adaptively queries the required structured information from PVM features based on its historical modeling experience.

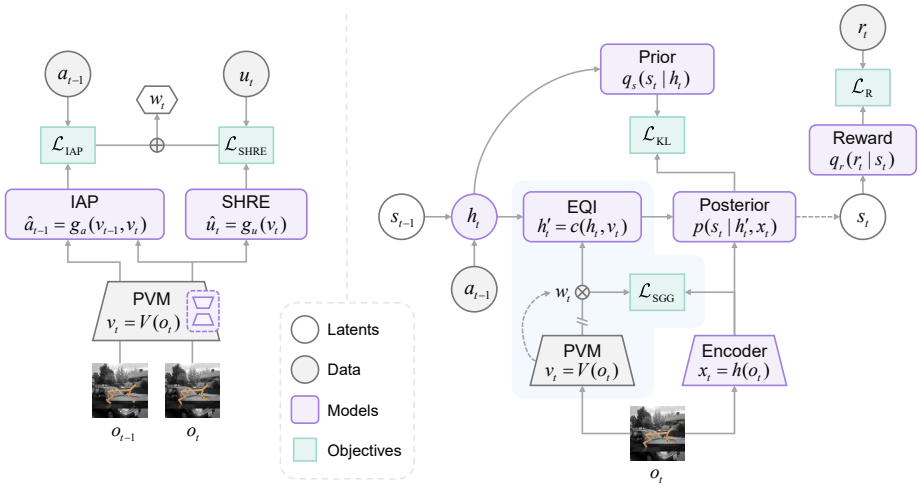

Fig. 1: The VPA-WM framework. (a) COA adapts a pre-trained vision model with lightweight IAP and SHRE objectives, producing control-relevant features v_t for robust dynamics modeling under visual distractions. (b) STCI injects vision priors via SGG and EQI, balanced by an adaptive weight w_t.

For the convenience of describing the proposed method, we first introduce the symbols used. Let o_t be the raw environmental observation at time t, $v_t = V(o_t)$ denote the structured perceptual features extracted from o_t by PVM, and $x_t = h(o_t)$ denote the instantaneous encoding generated by the encoder of the world model. The method includes the following main components:

(1) Control-Oriented Alignment of PVM:

$$\begin{aligned} &\text{Instant Action Prediction:} && \hat{a}_{t-1} = g_a(v_{t-1}, v_t) \\ &\text{Short-Horizon Return Estimation:} && \hat{u}_t = g_u(v_t) \end{aligned} \tag{5}$$

(2) VPA-WM:

$$\begin{aligned} &\text{Recurrent model:} && h_t = f_\theta(h_{t-1}, s_{t-1}, a_{t-1}) \\ &\text{Interaction model:} && h'_t = c_\theta(h_t, v_t) \\ &\text{Representation model:} && s_t \sim p_\theta(s_t \mid h'_t, x_t) \\ &\text{Transition predictor:} && \hat{s}_t \sim q_\theta(\hat{s}_t \mid h_t) \\ &\text{Reward predictor:} && \hat{r}_t \sim q_\theta(\hat{r}_t \mid h_t, s_t) \end{aligned} \tag{6}$$

4.2 Control-Oriented Alignment (COA)

Despite the structured visual priors provided by PVMs, there exists a significant misalignment between their visual semantic representations and reinforcement learning control tasks, which manifests as: (1) an inherent discrepancy between the pre-training objectives of PVMs and the dynamics learning objectives of world models; (2) PVM features failing to perceive fine-grained action-induced state changes; (3) the semantic clustering of PVMs being unable to capture reward-driven inter-entity correlations. To bridge these misalignments, we propose a control-oriented alignment mechanism. This mechanism keeps the backbone network of PVMs frozen and performs lightweight parameter updates through two complementary auxiliary objectives, rather than simply discarding or overwriting pre-trained knowledge. Instant Action Prediction captures action signals, while Short-Horizon Return Estimation captures reward signals; together, they endow PVMs with control-relevant semantics.

Instant Action Prediction (IAP) Effective control representations must be highly sensitive to action-induced state transitions while remaining invariant to task-irrelevant visual perturbations. We introduce instant action prediction: given consecutive observations o_{t-1}, o_t, PVM generates features v_{t-1}, v_t. A lightweight instant action prediction auxiliary model g_a predicts the intermediate action: $\hat{a}_t = g_a(v_t, v_{t+1})$, We optimize the mean squared error:

$$\mathcal{L}_{\text{IAP}}^t \doteq \mathbb{E}\big[\|\hat{a}_t - a_t\|^2\big], \tag{7}$$

This objective guides PVM to encode visual changes caused by actions, making it focus on the dynamic features of controllable objects and thereby naturally suppressing background distractions. When PVM features sufficiently capture environmental dynamics, instant action prediction can reconstruct actions with high accuracy. Notably, this method eliminates the need to explicitly handle complex transformations in high-dimensional observation spaces, making it particularly suitable for guiding the redirection of pre-trained visual features.

Short-Horizon Return Estimation (SHRE) The essential goal of optimizing RL tasks is to pursue higher reward values. Therefore, we design a short-horizon return estimation auxiliary task to inject reward signals required for policy optimization into PVMs. For a time horizon n and discount factor γ, the short-horizon return is defined as:

$$u_t \;=\; \sum_{k=0}^{n-1} \gamma^k r_{t+k}. \tag{8}$$

A lightweight return predictor $g_u(v_t)$ maps the PVM feature at time t to an estimate: $\hat{u}_t = g_u(v_t)$. This prediction is optimized via mean squared error:

$$\mathcal{L}_{\text{SHRE}}^t \doteq \mathbb{E}\big[\|\hat{u}_t - u_t\|^2\big]. \tag{9}$$

This objective guides PVMs to encode state attributes closely related to the reward function, thereby addressing the fundamental disconnection between PVM representations and RL task objectives. By focusing on short-term rather than long-term returns, we avoid the high variance issue in value function estimation while ensuring PVM features capture fine-grained visual cues related to immediate rewards. A moderate value of n (typically $n = 3 \sim 5$) achieves the best balance between stable training and capturing task-related features.

Low-Rank Adaptation The optimization objective of Control-Oriented Alignment is defined as:

$$\mathcal{L}_{\text{COA}}^{t} \doteq \mathcal{L}_{\text{IAP}}^{t} + \mathcal{L}_{\text{SHRE}}^{t}. \tag{10}$$

To ensure effective adaptation of PVM and avoid catastrophic forgetting, we inject Low-Rank Adaptation (LoRA) [14] *only* into the query and value projection layers within the attention mechanism of PVM. Meanwhile, a two-stage decoupled strategy is adopted for Control-Oriented Alignment and world model training: the LoRA parameters of PVM do not receive gradients from the world model's dynamics learning, and are optimized exclusively through $\mathcal{L}_{\text{COA}}^{t}$. This decoupling ensures that PVM learns to focus on control-related semantics while retaining its robust visual capabilities. Experimental results in Section 5 demonstrate that this design significantly enhances the applicability of PVM's representations to visual control tasks, providing high-quality visual priors for Spatio-Temporal Collaborative Integration.

4.3 Spatio-Temporal Collaborative Integration (STCI)

Although Control-Oriented Alignment successfully bridges the semantic-dynamics gap between PVM visual semantics and RL control objectives, effectively integrating structured visual priors without sacrificing temporal coherence remains a challenge. World models rely on smoothly evolving latent trajectories for multi-step prediction and planning, yet the rigid injection of visual features at each step can disrupt temporal coherence, leading to degraded prediction performance. We propose a Spatio-Temporal Collaborative Integration mechanism — a hierarchical integration strategy that establishes complementary interactions in both spatial and temporal dimensions. As shown in Fig. 1.(b), this mechanism constructs two collaborative yet functionally distinct interaction levels: Spatially Grounded Guidance (SGG) provides instantaneous geometric and semantic anchors during the encoding phase of the world model, enabling the encoder to quickly focus on task-relevant regions; Experience Query Integration (EQI) allows the RSSM to adaptively extract the required fine-grained visual evidence based on its recurrent evolutionary experience. This dual-level design ensures that the world model can fully leverage the strong visual capabilities of PVM while maintaining the temporal stability required for long-term prediction.

Spatially Grounded Guidance (SGG) In the early stages of world model training, the instantaneous encoding x_t generated by the convolutional encoder

$h(\cdot)$ is susceptible to transient visual distractions. To provide a robust geometric and semantic foundation, we introduce a Spatially Grounded Guidance mechanism: first, we align the dimensions of x_t with the PVM global features v_t through a lightweight projection function π, then optimize the following objective:

$$\mathcal{L}^t_{\text{SGG}} \doteq \mathbb{E}\big[\|\pi(x_t) - v_t\|^2\big]. \tag{11}$$

This loss function forces the encoder to focus on spatial structures that are verified as semantically stable and task-relevant in PVM, such as object boundaries and relative positions, rather than being affected by distractions like background textures. This guidance is equivalent to injecting a spatial regularization prior into the encoder's learning process, significantly reducing the dimensionality of its solution space and improving learning efficiency. Experiments show that under strong visual distractions, SGG enables the encoder to quickly establish invariance to irrelevant variations while maintaining high sensitivity to task-critical objects, laying a stable perceptual foundation for subsequent dynamic modeling.

Experience Query Integration (EQI) SGG addresses the robustness of initial feature extraction. However, the long-term prediction capability of the world model still depends on the accurate capture of dynamics by its recurrent state h_t. To this end, we design an integration mechanism based on a query-response paradigm. Specifically, we treat PVM's patch tokens $v_t^{\text{patch}} \in \mathbb{R}^{P\times D}$ (where P is the number of patches and D is the feature dimension) as a structured knowledge base, enabling the recurrent state h_t to query this knowledge base through cross-attention to adaptively supplement detailed information relevant to the current task:

$$\begin{aligned} h'_t &= h_t + \alpha \cdot \text{CrossAttn}(Q, K, V), \\ Q = W_q h_t, \quad K &= W_k v_t^{\text{patch}}, \quad V = W_v v_t^{\text{patch}}, \end{aligned} \tag{12}$$

where $\alpha \propto w_t$ (Eq.13) is the fusion coefficient, and W_q, W_k, W_v are learnable projection matrices. This design ensures that the world model only integrates visual evidence of detailed information most relevant to the long-term dynamic optimization objective. Unlike traditional feature concatenation or simple weighted fusion, this mechanism preserves the integrity of the RSSM's autoregressive transition structure while endowing it with the ability to selectively utilize structured visual evidence, enhancing the richness and robustness of state representations.

Adaptive Perception-Dynamics Balance To prevent training collapse caused by an imbalance between the injection of visual priors and dynamics learning during training, we propose an adaptive perception-dynamics balance strategy that calculates adaptive weights based on the Exponential Moving Average (EMA) of auxiliary losses in Control-Oriented Alignment:

$$w_t \doteq \text{clip}\left(\frac{w_{\text{base}}}{\bar{\mathcal{L}}^t_{\text{IAP}} + \bar{\mathcal{L}}^t_{\text{SHRE}}},\ w_{\min},\ w_{\max}\right). \tag{13}$$

where $\bar{\mathcal{L}}^t_{\text{IAP}}$ and $\bar{\mathcal{L}}^t_{\text{SHRE}}$ are the EMAs of $\mathcal{L}^t_{\text{IAP}}$ and $\mathcal{L}^t_{\text{SHRE}}$, respectively, and $\text{clip}(\cdot)$ is the clipping function. This design enables the world model to automatically adjust its degree of reliance on PVM representations according to the progress of COA, ensuring that it can efficiently utilize high-quality visual cue information. This adaptive mechanism ensures the stability of the training process while achieving the optimal trade-off between perceptual robustness and dynamic accuracy.

Optimization Objective of VPA-WM Integrating the above components, the complete optimization objective of VPA-WM combines reward prediction, dynamics regularization, spatially grounded guidance, and temporal evidence integration:

$$\mathcal{L}^t_{\text{VPA}-\text{WM}} \doteq \mathcal{L}^t_{\text{R}} + \lambda_t \cdot \mathcal{L}^t_{\text{KL}} + w_t \cdot \mathcal{L}^t_{\text{SGG}}, \tag{14}$$

where λ_t is a time-varying coefficient that increases with training steps t to stabilize KL divergence optimization, we refer to the design of MInCo [26] and adjust hyperparameters more suitable for VPA-WM. Reward prediction loss $\mathcal{L}^t_{\text{R}}$ and KL divergence $\mathcal{L}^t_{\text{KL}}$ follow the settings of RePo [34], optimizing reward accuracy and posterior-prior consistency, respectively. We also discard the reconstruction loss to prevent the model from reconstructing irrelevant visual noise.

5 Experiments

5.1 Experimental Setup

Environment Configuration We select two types of representative visually distracting environments, covering high-dimensional continuous control scenarios and fine-grained navigation scenarios, to comprehensively evaluate the method's robust visual control capabilities. All experiments use 64×64 image observations.

- **Distracted DeepMind Control Suite (Distracted DMC)** [32]: A variant of the DeepMind Control Suite [28] that preserves the original physical dynamics while replacing the static background with dynamic natural videos.
- **Textured Maze2D**: A texture-augmented variant of the Maze2D navigation task from D4RL [4]. The background is replaced with complex textures.

Baseline Methods We select three representative visual MBRL works as baselines, covering different robustness optimization strategies:

- **Dreamer** [7], a foundational visual MBRL algorithm that leverages latent dynamics to train policies.
- **RePo** [34], a decoder-free variation of Dreamer, which learns robust representations by maximizing the predictability of rewards and dynamics while constraining the information flow from observations.

– **MInCo** [26], which employs contrastive representation learning [2] and a time-varying dynamics (TVD) weighting to mitigate information conflicts and enhance representation robustness, and is one of the current SOTA methods for robust MBRL.

Framework Instantiation To verify the universality of the VPA-WM framework to PVM, we instantiated two different types of PVM in the experiment: DINOv2 [20] (denoted as **DA-WM**) and CLIP [23] (denoted as **CA-WM**). The world model backbone uses RePo [34]. More detailed model architectures and hyperparameter settings are provided in the supplementary materials.

5.2 Result Analysis

Distracted DMC We select 6 continuous control tasks across 4 domains in Distracted DMC and conduct 5 rounds of random tests. The average learning speed and asymptotic performance of DA-WM and CA-WM are significantly superior to the baseline methods in most tasks, as shown in Fig. 2. Specifically, Dreamer relies on pixel reconstruction and is therefore easily dominated by distracting textures and motions; RePo removes reconstruction to suppress interference but may discard task-relevant details; MInCo mitigates information conflict via contrastive learning, yet purely visual invariance is often insufficient in hard control tasks. These results suggest that leveraging PVM priors can substantially improve interference resistance while preserving control-relevant dynamics.

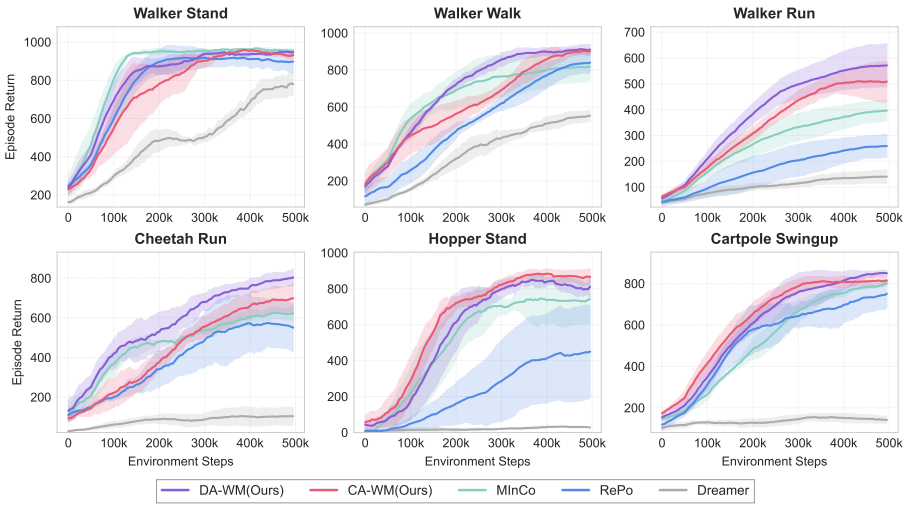

Fig. 2: Experimental results on the Distracted DMC environments. VPA-WM achieves SOTA on almost all tasks, demonstrating that through our design, PVMs can significantly enhance the robust control capabilities of the world model.

Textured Maze2D To further test our method's robust control capabilities at a finer granularity, we select 2 mazes of different difficulties and 3 different textures, and perform 5 rounds of random tests. DA-WM and CA-WM outperform baseline methods in most tasks (Table 1). Textured Maze2D places more emphasis on preserving fine-grained geometric structure and global navigational consistency under appearance shifts. By injecting structured priors and aligning them with control signals, our representations retain maze topology and goal-relevant spatial cues rather than overfitting to texture statistics, which improves both sample efficiency and final performance.

Table 1: Experimental results on the Textured Maze2D environments. Under extremely challenging visual conditions, the fine-grained control performance of VPA-WM is still excellent, indicating the visual perception advantage of PVM is fully used.

Method	Metal		Wood		Jeans	
	Umaze	Large	Umaze	Large	Umaze	Large
Dreamer	69±26	47±28	93±24	176±59	37±16	41±23
RePo	91±18	7±8	96±14	16±9	94±16	8±5
MInCo	102±16	387±57	103±11	239±70	102±15	**380**±55
DA-WM	**107**±25	349±70	**106**±19	297±52	104±16	376±61
CA-WM	104±22	**387**±53	104±13	**328**±67	**108**±25	359±78

Standard DMC To verify the robustness of the representations learned by VPA-WM against distracting environments, we additionally conducted experiments on the standard DMC environments. As shown in Table 2, although VPA-WM is trained without a reconstruction objective, its performance on standard DMC tasks is comparable to that of Dreamer. When transferring from standard DMC to distracted DMC, VPA-WM exhibits almost no performance degradation and still maintains a leading performance. This clearly demonstrates that VPA-WM can effectively ignore task-irrelevant distractions and focus on controlling the task-relevant entities.

Visualization Analysis Since our framework doesn't utilize reconstruction loss, we additionally train a pixel reconstruction decoder for analyzing representation quality without gradient backpropagation. The visualization results are shown in Fig. 3. In Distracted DMC, Dreamer can hardly distinguish the task's main body from the distracting background, while RePo and MInCo suppress part of the background but often exhibit poor action coherence on the main body. DA-WM and CA-WM not only clearly identify the controllable entity but maintain coherent dynamics. In Textured Maze2D, our methods still reconstruct relatively clear structural layouts under harsh textures, indicating that the learned latent states encode stable geometry rather than superficial appearance.

Table 2: Comparison with baselines on Standard DMC and Distracted DMC. In this table, we use W.S. to denote Walker Stand, W.W. to denote Walker Walk, and so on.

Standard DMC	W.S.	W.W.	W.R.	C.R.	H.S.	C.S.
Dreamer	**960**±15	**950**±10	471±96	**892**±37	682±192	797±25
RePo	928±40	457±106	287±69	598±133	390±53	698±62
MInCo	952±11	929±10	392±266	605±185	774±63	832±18
DA-WM	940±24	934±18	**557**±41	789±46	805±27	**862**±6
CA-WM	929±31	935±54	522±49	713±75	**829**±55	846±47
Distracted DMC	W.S.	W.W.	W.R.	C.R.	H.S.	C.S.
Dreamer	780±59	553±34	141±26	104±46	28±2	141±18
RePo	898±55	840±54	259±45	551±121	450±260	751±67
MInCo	**962**±7	817±78	396±41	622±36	742±131	800±7
DA-WM	947±15	**910**±31	**572**±84	**804**±43	812±61	**850**±15
CA-WM	931±19	902±15	509±77	699±78	**866**±45	814±20

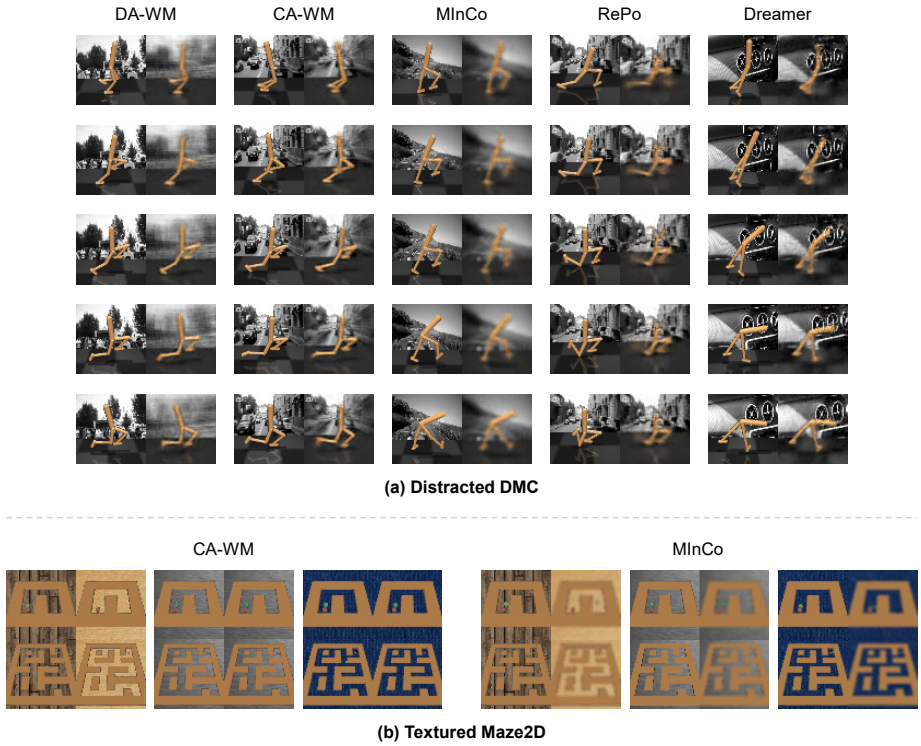

Fig. 3: Visual reconstruction in Distracted DMC and Textured Maze2D environments. VPA-WM can not only clearly identify the task subject, but also achieve accurate, precise and coherent control over it.

5.3 Ablation Studies

We designed two groups of ablation variants (7 in total), Group 1 verifies the necessity of using Control-Oriented Alignment (COA) to fine-tune PVM, and Group 2 verifies the reasonableness and effectiveness of using Spatio-Temporal Collaborative Integration (STCI) to fuse PVM features.

Group 1: (1) **rCOA**: Remove Control-Oriented Alignment, directly use frozen DINO features. (2) **rIAP**: Fine-tune DINO but remove Instant Action Prediction. (3)**rSHRE**: Fine-tune DINO but remove Short-Horizon Return Estimation.

Group 2: (4) **rSGG**: Remove Spatially Grounded Guidance. (5) **rEQI**: Remove Experience Query Integration. (6) **Mask**: Remove STCI and overlay the masks generated by DINO onto the original image as the input for the world model. (7) **rp.Enc**: Replace the world model encoder with DINO directly.

Table 3: Ablation experiment results. Confirmed the necessity and rationality of each module in the VPA-WM framework.

Task	DA-WM	Ablation Group 1			Ablation Group 2			
		rCOA	rIAP	rSHRE	rSGG	rEQI	Mask	rp.Enc
Walker Run	572±84	424±97	219±114	450±103	161±89	478±55	146±72	168±55
Cheetah Run	804±43	753±66	609±101	402±98	154±64	540±54	293±90	349±125

We conducted experiments on the walker run and cheetah run tasks in the Distracted DMC environment, and the results are shown in Table 3. The results of Group 1 demonstrate that COA, including the design of its two complementary components, is critical. It is difficult for PVM without COA fine-tuning to provide the best visual prior to the world model. Fine-tuning a PVM using only one component tends to cause its features to overfit a single dynamic or reward signal, resulting in imbalanced training of the world model. The results of Group 2 shows that STCI is also a necessary condition for VPA-WM to have top-level robust control capabilities. Without spatial hint, the world model will become blind during autonomous query. If the world model is only provided with spatial cues and is not allowed to query autonomously, it will not reach its full potential. However, if the PVM feature is introduced forcefully like directly masking and replacing the encoder, it will be counterproductive, causing the world model to lose focus on the control task. This also verifies that our design of treating PVM's visual prior as an auxiliary signal is reasonable and superior.

6 Conclusion

In this paper, we propose the VPA-WM framework: integrating large-scale pretrained visual priors into visual MBRL for robust visual action control. We

develop the Control-Oriented Alignment scheme to steer pre-trained perceptual priors toward task-relevant subspaces. We then introduce the Spatio-Temporal Collaborative Integration with adaptive weighting, enabling the model to exploit perceptual priors for stable long-horizon prediction. VPA-WM significantly improves sample efficiency, robustness, and generalization of MBRL methods on visually distracting control tasks. In future work, we will focus on integrating VLMs into our method for a more stable and consistent dynamic modeling.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (62406112, 62372179). Yang Shu is the corresponding author of the work.

References

1. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Int. Conf. Comput. Vis. (2021)
2. Chen, X., He, K.: Exploring simple siamese representation learning. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
3. Deng, F., Jang, I., Ahn, S.: DreamerPro: Reconstruction-free model-based reinforcement learning with prototypical representations. In: Int. Conf. Mach. Learn. pp. 4956–4975 (2022)
4. Fu, J., Kumar, A., Nachum, O., Tucker, G., Levine, S.: D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint (2020)
5. Fu, X., Yang, G., Agrawal, P., Jaakkola, T.S.: Learning task informed abstractions. In: Int. Conf. Mach. Learn. pp. 3480–3491 (2021)
6. Ha, D., Schmidhuber, J.: World models. arXiv preprint (2018)
7. Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: Int. Conf. Learn. Represent. (2020)
8. Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels pp. 2555–2565 (2019)
9. Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. In: Int. Conf. Learn. Represent. (2020)
10. Hafner, D., Pasukonis, J., Ba, J., Norouzi, M.: Mastering diverse domains through world models. arXiv preprint (2023)
11. Hansen, N., Su, H., Wang, X.: TD-MPC2: Scalable, robust world models for continuous control (2024)
12. Hansen, N., Wang, X., Su, H.: Temporal difference learning for model predictive control (2022)
13. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
14. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-rank adaptation of large language models. arXiv preprint (2021)
15. Jones, S., Zhou, L., Pattinson, S.W.: Pre-trained visual representations generalize where it matters in model-based reinforcement learning. arXiv preprint (2025)

16. Jordan, M.I., Ghahramani, Z., Jaakkola, T.S., Saul, L.K.: An introduction to variational methods for graphical models. In: Jordan, M.I. (ed.) Learning in Graphical Models, NATO ASI Series, vol. 89, pp. 105–161. Springer Netherlands (1998). https://doi.org/10.1007/978-94-011-5014-9_5
17. Kich, V.A., Bottega, J.A., Steinmetz, R., Grando, R.B., Yorozu, A., Ohya, A.: CURLing the dream: Contrastive representations for world modeling in reinforcement learning (2024)
18. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (October 2023)
19. Nair, S., Rajeswaran, A., Kumar, V., Finn, C., Gupta, A.: R3M: A universal visual representation for robot manipulation. arXiv preprint (2022)
20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint (2023)
21. Ortiz, J., Gupta, S., Schmid, C., Laptev, I.: DMC-VB: A benchmark for visual representation learning in control. In: Adv. Neural Inform. Process. Syst. (2024)
22. Pan, M., Zhu, X., Wang, Y., Yang, X.: Iso-dream: Isolating and leveraging noncontrollable visual dynamics in world models. In: Adv. Neural Inform. Process. Syst. (2022)
23. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
24. Schneider, M., Krug, R., Vaskevicius, N., Palmieri, L., Boedecker, J.: The surprising ineffectiveness of pre-trained visual representations for model-based reinforcement learning. arXiv preprint (2025)
25. Stone, A., Ramirez, O.M., Brohan, A., et al.: The distracting control suite: A challenging benchmark for reinforcement learning from pixels. In: Proc. Conf. Robot Learn. (2021)
26. Sun, S., Zhang, H., Liu, Z., Yang, X., Wan, L., Yan, B., Chen, X., Lan, X.: MInCo: Mitigating information conflicts for robust visual model-based reinforcement learning (2025)
27. Sutton, R.S., Barto, A.G.: Reinforcement learning - an introduction, 2nd Edition. MIT Press (2018)
28. Tassa, Y., Doron, Y., Muldal, A., Erez, T., Li, Y., de Las Casas, D., Budden, D., Abdolmaleki, A., Merel, J., Lefrancq, A., Lillicrap, T.P., Riedmiller, M.A.: DeepMind control suite. arXiv preprint (2018)
29. Wang, T., Du, S.S., Torralba, A., Isola, P., Zhang, A., Tian, Y.: Denoised mdps: Learning world models better than the world itself. In: Int. Conf. Mach. Learn. pp. 22591–22612 (2022)
30. Wang, Z., Ze, Y., Sun, Y., Yuan, Z., Xu, H.: Generalizable visual reinforcement learning with segment anything model. arXiv preprint (2023)
31. Yuan, Z., Xue, Z., Yuan, B., Wang, X., Wu, Y., Gao, Y., Xu, H.: Pre-trained image encoder for generalizable visual reinforcement learning. In: Adv. Neural Inform. Process. Syst. vol. 35, pp. 13022–13037 (2022)
32. Zhang, A., Wu, Y., Pineau, J.: Natural environment benchmarks for reinforcement learning. arXiv preprint (2018)
33. Zhou, G., Pan, H., LeCun, Y., Pinto, L.: DINO-WM: World models on pre-trained visual features enable zero-shot planning. arXiv preprint (2025)

34. Zhu, C., Simchowitz, M., Gadipudi, S., Gupta, A.: RePo: Resilient model-based reinforcement learning by regularizing posterior predictability. In: Adv. Neural Inform. Process. Syst. (2023)