EMPOWERING WORLD MODELS WITH DECONFOUNDED DYNAMICS VIA MULTI-LEVEL CAUSAL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Implicit world models predict future visual representations without reconstructing images, supporting tasks such as embodied planning. However, encoding object and background information into a shared feature space can entangle task-relevant states with incidental visual context. This may bias future-state predictions and propagate errors over successive predictions. To address this problem, this paper presents a Deconfounded Causally Aligned Latent Model (DeCALM), which empowers world models with deconfounded dynamics by reducing background dependence in latent encoding. Specifically, DeCALM comprises two main modules: Background-Deconfounded Latent Encoding (BLE) and Multi-Level Causal Alignment (MCA). BLE uses object and background representations separately, approximates backdoor adjustment with a background dictionary, and constrains predictions under counterfactual background changes. MCA takes the resulting features and adaptively combines absolute-state and incremental predictions to improve future-state prediction. Experiments on three benchmark datasets show that DeCALM achieves the highest success rate among the compared methods, including a 14.58-percentage-point gain over the strongest baseline on Drawer Open. Further analyses suggest that DeCALM mitigates the ill-posedness of latent representations under visual disturbances, yielding more task-relevant features and more accurate state-change predictions at later horizons.

1 INTRODUCTION

World models learn how a scene evolves in response to an agent's actions (Ha & Schmidhuber, 2018; Hafner et al., 2020). By rolling out candidate actions before execution, they allow a planner to compare possible futures rather than react to the current observation alone (Schrittwieser et al., 2020; Hansen et al., 2024). Implicit world models make this feasible in visual environments by predicting future representations instead of reconstructing images (Zhou et al., 2025; Nam et al., 2026). Yet a useful latent state must do more than compress observations: it must retain the object changes that determine whether an action moves the scene toward the goal (Lei et al., 2025). Robustly modeling action-driven dynamics in latent space therefore remains a central challenge (Sun et al., 2024).

Recent implicit world models have made progress by choosing different units for latent prediction. DINO-WM (Zhou et al., 2025) predicts future features from a pretrained visual encoder, while Causal-JEPA (Nam et al., 2026) uses object-centric representations and masked prediction to capture interactions between objects. However, their latent states are still derived from images that contain both task-relevant objects and incidental visual context. As illustrated in Figure 1, changing the background while preserving the physical task can cause a predicted trajectory to drift away from the goal. The disturbance first biases the representation, then alters the predicted outcomes used to select an action. As predicted states are fed back into subsequent steps, this deviation can grow across the rollout.

To address visual disturbance and rollout error accumulation, we propose DeCALM, a multi-level causally aligned framework for deconfounded latent dynamics. As shown in Figure 1, DeCALM tackles these issues at both representation and transition levels. First, the Background-Deconfounded Latent Encoding module applies SCM-guided background adjustment over learned

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

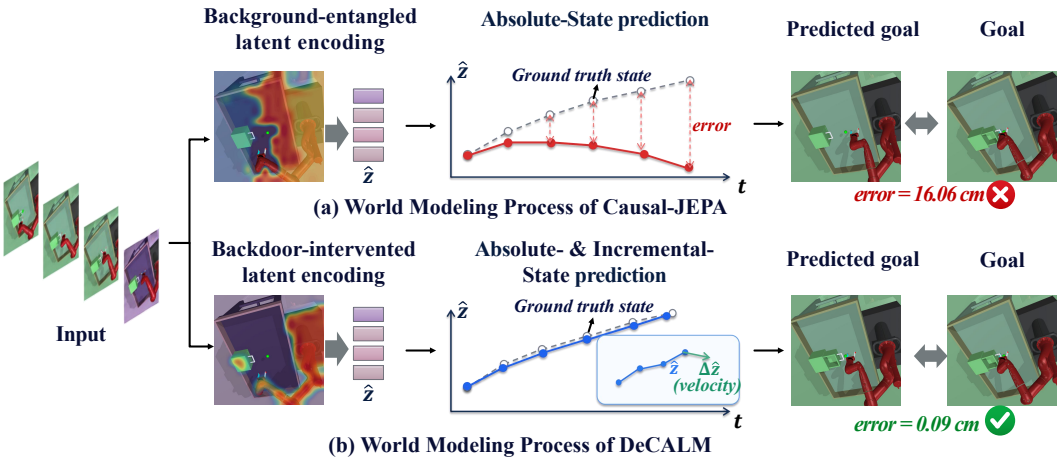


Figure 1: Causal-JEPA vs. DeCALM under visual disturbance. (a) Background-entangled latent encoding and absolute-state prediction cause Causal-JEPA to miss the goal. (b) DeCALM focuses on task-relevant regions and aligns predicted states with their changes, limiting accumulated error and reaching the goal.

background prototypes and counterfactual background substitution during training. This reduces object representations' reliance on incidental visual context while preserving task-relevant dynamics across background changes, thereby mitigating background-induced representation bias and improving future-state prediction robustness. Second, the Multi-Level Causal Alignment module jointly models the predicted next state and its incremental change from the current state, adaptively integrating both prediction paths during rollout. This better captures action-driven state transitions and preserves task-relevant object evolution over multiple steps. Together, these modules reduce background-related prediction bias and residual error propagation, yielding more reliable latent dynamics for long-horizon goal-conditioned planning.

Extensive experiments on three goal-conditioned manipulation benchmarks under diverse visual disturbances demonstrate that DeCALM produces more robust latent dynamics and more accurate long-horizon state-transition predictions. These improvements lead to consistently higher planning success rates than existing state-of-the-art world-model methods. Ablation studies and visual analyses further verify the effectiveness of the proposed components and their ability to preserve task-relevant information under visual disturbances. Our contributions are summarized as follows:

- **DeCALM: Deconfounded Causally Aligned Latent Model for Long-Horizon Embodied Planning.** We propose DeCALM, a unified framework that deconfounds implicit world models at representation and transition levels against background-confounded dynamics, enabling reliable goal-conditioned planning under visual disturbances.
- **Deconfounding Representation and Transition via BLE and MCA.** On OGBench and MetaWorld tasks, DeCALM tops success, beating Causal-JEPA by 4.00/15.62/2.60 and the best baseline by 3.34/14.58/2.60; ablations confirm BLE, MCA, CE, with lower later-horizon error, better state-change prediction, and stronger task-relevant attention.
- **Empirical Validation of Deconfounded Long-Horizon Planning.** DeCALM achieves the highest success on OGBench and MetaWorld, surpassing Causal-JEPA and the best baseline by up to 15.62/14.58 points; ablations validate BLE, MCA, and CE, with lower later-horizon error, better state-change prediction, and stronger task-relevant attention.

2 Related Work

2.1 World Models with Applications in Embodied Planning

World models support embodied planning by predicting the consequences of actions. Reward-driven methods learn policies from imagined experience, including World Models (Ha & Schmid-

huber, 2018), Dreamer (Hafner et al., 2020), IRIS (Micheli et al., 2023), DreamerV3 (Hafner et al., 2025), and R2-Dreamer (Morihira et al., 2026); MuZero (Schrittwieser et al., 2020) and TD-MPC2 (Hansen et al., 2024) instead plan with learned dynamics and values. Iso-Dream (Pan et al., 2022), AD3 (Wang et al., 2024a), HRSSM (Sun et al., 2024), PSP (Hutson et al., 2024), and SeeX (Huang et al., 2024) further address visual distractions in model-based control. Goal-driven methods compare predicted latent states with desired states: DINO-WM (Zhou et al., 2025) predicts visual features, PLDM (Sobal et al., 2025) plans with reward-free latent dynamics, C-JEPA (Nam et al., 2026) models object-level interactions, and Toso et al. (Toso et al., 2026) improve visual invariance. Both routes rely on accurate dynamics prediction, particularly under visual changes.

2.2 Causality in World Models

Causal world models characterize how states interact and evolve. CBM (Wang et al., 2024b) learns reusable causal dynamics for task-specific abstractions; OOCDM (Yu et al., 2024) shares mechanisms across objects, while Hwang et al. (Hwang et al., 2024) model context-specific dependencies. CSR (Yang et al., 2025) adapts causal representations to changing environments, and WM3C (Wang & Huang, 2025) composes causal dynamics for unseen tasks. ECL (Cao et al., 2025) and Causality-Seeking Agents (Zhao et al., 2025) refine causal models through exploration. For object-centric modeling, SPARTAN (Lei et al., 2025) learns sparse interaction structures, whereas STICA (Nishimoto & Matsubara, 2026) couples an object-centric world model with causality-aware control. Murillo et al. (Murillo-González et al., 2025) account for uncertainty in causal structure, and Exo-Predicator (Liang et al., 2026) learns abstract causal processes. These methods use causal structure to organize or adapt dynamics. DeCALM instead addresses background-induced bias in visual latent prediction through approximate backdoor adjustment and counterfactual background substitution.

2.3 Visual Disturbances and Robust World Models

Visual disturbances can draw learned representations toward task-irrelevant scene information. PSP (Hutson et al., 2024) emphasizes policy-relevant regions during visual reconstruction. HRSSM (Sun et al., 2024) combines masking, latent consistency, and bisimulation to learn robust representations. Toso et al. (Toso et al., 2026) regularize predictive representations under distractions, while MInCo (Sun et al., 2026) uses contrastive objectives to retain control-relevant information. LAOM (Nikulin et al., 2025) uses limited action supervision to improve latent-action learning with distractors. Other methods separate task dynamics from visual variation. Iso-Dream (Pan et al., 2022) distinguishes controllable, noncontrollable, and static components. AD3 (Wang et al., 2024a) models agent-driven and distractor-driven changes separately. SeeX (Huang et al., 2024) separates endogenous from exogenous information for exploration, while Klepach et al. (Klepach et al., 2026) select action-relevant object slots. These methods improve robustness by emphasizing or separating task-related information. DeCALM further adjusts background-dependent predictions and tests their consistency under counterfactual background changes.

3 Problem Formulation

We consider visual world modeling under partial observability. Given an observation O_t and action A_t, the visual dynamics are described by

$$P(O_{t+1} \mid O_{\leq t}, A_{\leq t}). \tag{1}$$

Learning uses an offline dataset $\mathcal{D} = \{(O^{(n)}_{1:T}, A^{(n)}_{1:T-1})\}_{n=1}^{N}$ of N trajectories of length T. Given an observation history and goal image O_g, a planner evaluates candidate actions through predicted future states.

Existing causal world models organize dependencies among state factors to learn reusable dynamics (Wang et al., 2024b; Yu et al., 2024). Such structure, however, does not by itself prevent irrelevant visual backgrounds from affecting latent predictions.

In our object-centric setting, we represent objects, backgrounds, and the next-state target as

$$S_t = f_S(O_t), \qquad C_t = f_C(O_t, U), \qquad Y^{\mathrm{gt}} = f_S(O_{t+1}), \tag{2}$$

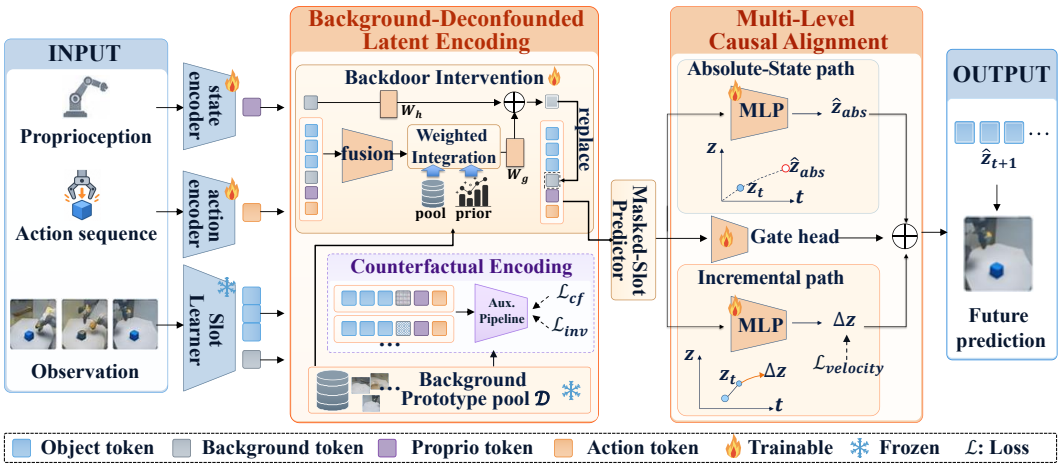


Figure 2: Illustration of the framework of DeCALM. DeCALM proposes a *Background-Deconfounded Latent Encoding* module to mitigate representation bias caused by visual disturbances through backdoor intervention and counterfactual consistency. To alleviate error accumulation during multi-step prediction, DeCALM proposes a *Multi-Level Causal Alignment* module to align state and state-change predictions and adaptively fuse the absolute-state and incremental paths.

where f_S and f_C are representation mappings, and U denotes background variation in the causal model. DeCALM predicts the next object representation Y from object and background information conditioned on candidate actions, while reducing background-induced bias and retaining object-state changes. The predicted state is compared with the goal representation $f_S(O_g)$ for planning.

4 METHOD

Figure 2 demonstrates DeCALM's Deconfounded Causally Aligned Latent Model fuses BLE (backdoor/counterfactual background-bias mitigation) and MCA (error-reducing state/change alignment), yielding background-reduced action-driven transitions for reliable long-horizon embodied planning.

4.1 BACKGROUND-DECONFOUNDED LATENT ENCODING

Based on the structural causal model (SCM) presented in Appendix A, we introduce the *Background-Deconfounded Latent Encoding* module to mitigate background bias in dynamics learning. Specifically, we first employ backdoor intervention to account for the observable confounder via a background dictionary. We then apply counterfactual encoding to regularize the predicted object dynamics under background changes. Collectively, these components reduce the influence of background correlations on the representations used for future prediction.

4.1.1 BACKDOOR INTERVENTION

To mitigate representation bias caused by visual disturbances, we apply backdoor adjustment to reduce the influence of the background confounder on future-state prediction. As illustrated by the SCM in Appendix, the observational distribution conditioned on the observation O_t and action $A_t = a$ can be expressed as:

$$P(Y \mid O_t, A_t = a) = \sum_{k=1}^{K} P\big(Y \mid S_t = f_S(O_t), \\ C_t = f_C(O_t, u_k),\ A_t = a\big)\, P(U = u_k \mid O_t). \tag{3}$$

where f_S and f_C denote the object and background representation mappings, respectively. The weight $P(U = u_k \mid O_t)$ reflects the association between the observation and the background confounder, allowing background cues to bias the prediction. To mitigate this bias, we stratify the

observable confounder U and average the conditional predictions according to the background proportions in the training data. Under the proposed SCM and assuming sufficient observational support across these strata, backdoor adjustment gives:

$$P(Y \mid \mathrm{do}(O_t), A_t = a) = \sum_{k=1}^{K} P\big(Y \mid S_t = f_S(O_t), \\ C_t = f_C(O_t, u_k),\ A_t = a\big)\, P(U = u_k), \tag{4}$$

where, $\mathrm{do}(O_t)$ denotes intervening on the observation while retaining the action as a conditioning input. This intervention removes the edge $U \to O_t$ and blocks the backdoor path $O_t \leftarrow U \to C_t \to Y$. The adjustment replaces the observation-dependent weights $P(U = u_k \mid O_t)$ with the background prior $P(U = u_k)$, reducing bias from the association between observations and backgrounds. We approximate this adjustment through the background dictionary described below.

Background Confounder Dictionary. To represent the observable confounder U, we construct a background dictionary from the training observations. We collect their background representations into a feature pool $\mathcal{M}_{\mathrm{bg}} = \{C_i\}_{i=1}^{N_{\mathrm{bg}}}$ and apply k-means clustering to the normalized features. Each cluster is represented by a background prototype c_k corresponding to the stratum $U = u_k$, with its empirical prior determined by the cluster frequency:

$$\mathcal{D}_{\mathrm{bg}} = \{c_k\}_{k=1}^{K}, \qquad \pi_k = \frac{n_k}{N_{\mathrm{bg}}}, \qquad N_{\mathrm{bg}} = \sum_{k=1}^{K} n_k, \tag{5}$$

where K is the number of prototypes and n_k is the number of background features assigned to the k-th cluster. The dictionary summarizes the background variation observed during training. In our implementation, the empirical prior π_k estimates $P(U = u_k)$ in Equation (4). The prototypes and their priors remain fixed during dynamics learning and planning.

Instantiation of Backdoor Intervention. Directly averaging predictions over all background prototypes would require repeated evaluations of the dynamics model. We therefore approximate the adjustment in representation space by integrating the prototypes before future prediction.

We first obtain a joint representation $H_t = \psi(\bar{S}_t, C_t, e_a(A_t))$, where $\bar{S}_t$ is the mean object representation, C_t is the background representation, e_a encodes the action A_t, and ψ is a learned fusion function. In practice, proprioception P_t is additionally encoded by e_p and incorporated into ψ. The contribution of each prototype is determined by its interaction with this joint representation:

$$\alpha_{t,k} = \frac{\exp\Big(\langle W_Q H_t, W_K c_k\rangle / \sqrt{d}\Big)}{\sum_{j=1}^{K} \exp\Big(\langle W_Q H_t, W_K c_j\rangle / \sqrt{d}\Big)}, \tag{6}$$

where W_Q and W_K are learnable projections and d is the projected feature dimension. We combine these interaction weights with the background prior to obtain

$$G_t = \sum_{k=1}^{K} \alpha_{t,k}\, \pi_k\, c_k, \tag{7}$$

where $\alpha_{t,k}$ measures the relevance of each prototype to the current input, while π_k accounts for its prevalence in the training data. The integrated feature is then fused with the observed background representation:

$$\tilde{C}_t = \mathrm{LN}(W_h C_t + \gamma W_g G_t)\,, \tag{8}$$

where W_h and W_g are learnable projections, γ controls the contribution of the integrated feature, and LN denotes layer normalization. The resulting $\tilde{C}_t$ provides adjusted background context for the latent predictor. Replacing C_t with $\tilde{C}_t$ yields the adjusted input sequence, which is processed by the transformer-based masked-slot predictor of Causal-JEPA (Nam et al., 2026) for token interaction and slot prediction. We then extract the future object-slot outputs as R_t for the subsequent Multi-Level Causal Alignment module.

4.1.2 COUNTERFACTUAL ENCODING

To further reduce reliance on background correlations, we introduce counterfactual encoding during training. We sample a prototype c_k according to the prior π_k and substitute it for the background representation C_t throughout the observation window, while keeping the object representations, proprioceptive states, and actions fixed. The substituted input produces a counterfactual prediction Y^{cf} through the same prediction network. For background changes that preserve the task dynamics, the predicted object transitions should remain consistent. We therefore supervise the counterfactual prediction using the future target $Y^{\mathrm{gt}} = f_S(O_{t+1})$ and the original prediction Y:

$$\mathcal{L}_{\mathrm{cf}} = \left\| Y^{\mathrm{cf}} - Y^{\mathrm{gt}} \right\|_2^2, \tag{9}$$

$$\mathcal{L}_{\mathrm{inv}} = \left\| Y^{\mathrm{cf}} - \mathrm{sg}[Y] \right\|_2^2, \tag{10}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. The first loss supervises future prediction under the substituted background, while the second encourages agreement between the original and counterfactual predictions.

4.2 MULTI-LEVEL CAUSAL ALIGNMENT

This module combines state-level prediction and transition-level supervision through dual-path causal gating. Building on the future object-slot features produced by the masked-slot predictor, it estimates absolute future states and relative state changes, and adaptively combines them to mitigate error accumulation during multi-step prediction.

Dual-Path Prediction. The future object-slot features R_t obtained above are passed to two prediction heads:

$$Y^{\mathrm{abs}} = f_{\mathrm{abs}}(R_t), \qquad \Delta Y_t = f_{\Delta}(R_t), \tag{11}$$

where f_{abs} predicts the absolute next state and f_{Δ} predicts the latent state increment. Using the current object representation $S_t = f_S(O_t)$ as the reference, the incremental path reconstructs the next state as $Y^{\Delta} = S_t + \Delta Y_t$.

Adaptive Gating. We compute a gate from R_t to combine the two estimates:

$$\begin{aligned} g_t &= \sigma\big(f_{\mathrm{gate}}(R_t)\big), \\ Y &= g_t \odot Y^{\mathrm{abs}} + (1 - g_t) \odot Y^{\Delta}, \end{aligned} \tag{12}$$

where f_{gate} is a learnable gating head, σ is the sigmoid function and the gate weights are broadcast over the corresponding representations. During training, the next observation provides the target object representation $Y^{\mathrm{gt}} = f_S(O_{t+1})$. The future-state prediction loss supervises the fused estimate Y, while the incremental path is additionally supervised by

$$\mathcal{L}_{\mathrm{velocity}} = \mathrm{MSE}\big(\Delta Y_t,\, Y^{\mathrm{gt}} - S_t\big)\,. \tag{13}$$

This objective aligns the predicted increment with the change between consecutive object representations, complementing state-level supervision. During autoregressive rollout, the fused prediction is fed back as the current object representation for the next step.

4.3 TRAINING STRATEGY

The training objective combines masked-history reconstruction and future object prediction with the causal objectives introduced above. The two basic losses are defined as

$$\mathcal{L}_{\mathrm{maskhist}} = \mathrm{MSE}\Big(\hat{S}_{\mathrm{hist}}[\mathcal{M}], S_{\mathrm{hist}}[\mathcal{M}]\Big)\,, \tag{14}$$

$$\mathcal{L}_{\mathrm{future}} = \mathrm{MSE}\big(Y, Y^{\mathrm{gt}}\big)\,, \tag{15}$$

where S_{hist} collects the historical object representations, $\hat{S}_{\mathrm{hist}}$ denotes their reconstructions, and $\mathcal{M}$ indexes the masked historical object slots. Combining these losses with the counterfactual and transition-level objectives gives

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{maskhist}} + \mathcal{L}_{\mathrm{future}} + \lambda_{\mathrm{cf}}\mathcal{L}_{\mathrm{cf}} + \lambda_{\mathrm{inv}}\mathcal{L}_{\mathrm{inv}} + \lambda_{\mathrm{vel}}\mathcal{L}_{\mathrm{velocity}}, \tag{16}$$

Table 1: Performance comparison of baselines and DeCALM. Best results are highlighted in bold.

Model	Success rate (%)		
	OGBench-Cube	MW-Drawer-Open	MW-Faucet-Open
IRIS (ICLR'23)	80.67	18.23	49.48
TD-MPC2 (ICLR'24)	79.33	10.94	65.10
DINO-WM (ICML'25)	85.33	38.02	46.88
DreamerV3 (Nature'25)	80.67	23.44	50.52
Causal-JEPA (ICML'26)	84.67	36.98	65.63
DeCALM	**88.67**	**52.60**	**68.23**

where λ_{cf}, λ_{inv}, and λ_{vel} control the contributions of the corresponding losses.

We initialize from a pretrained Causal-JEPA model and keep the DINO encoder, slot-attention module, and pretrained dynamics backbone frozen. We adopt a two-stage fine-tuning strategy. First, we optimize Background-Deconfounded Latent Encoding using the prediction and counterfactual objectives, without the velocity loss. We then freeze this module and optimize Multi-Level Causal Alignment with the full objective.

Inference. At inference time, the model performs background adjustment and dual-path fusion following Equation (12) conditioned on candidate actions. The resulting predictions are fed back autoregressively to generate future object-state trajectories for downstream planning.

5 Experiments

5.1 Experimental Setup

Datasets. We evaluate DeCALM on three vision-based embodied planning tasks from OGBench (Park et al., 2025) and MetaWorld (Yu et al., 2020). In OGBench *cube-single*, a gripper moves a cube to a target configuration. The MetaWorld *Drawer Open* and *Faucet Open* tasks require pulling a drawer and turning a faucet, respectively. All three tasks involve visual disturbances in background appearance, camera viewpoint, and lighting. These conditions remain fixed within each OGBench episode, whereas the MetaWorld tasks include visual changes within an episode.

Metrics. Following prior work on vision-based planning (Zhou et al., 2025; Nam et al., 2026), we adopt task success rate (SR) as the evaluation metric, defined as the percentage of episodes that reach the task goal within the step limit.

Implementation Details. Following Causal-JEPA (Nam et al., 2026), DeCALM uses a frozen DINOv2 ViT-S/14 (Oquab et al., 2024) and slot-attention module (Locatello et al., 2020) to obtain object and background slots. We build the background dictionary from training slots and fine-tune the proposed modules from a pretrained Causal-JEPA predictor with AdamW, using seed 42 for the final fine-tuning stage. At inference, CEM evaluates predicted latent trajectories with 300 candidates; 32 training-trajectory sequences initialize the first plan of each episode. Within each task, CEM-based methods use the same candidate initialization and planning budget. We report success rates averaged over evaluation seeds 42, 123, and 3407. Further implementation and baseline-reproduction details are provided in Appendix C.

5.2 Performance Comparison

Table 1 shows that DeCALM achieves the highest success rate on all tasks among IRIS, TD-MPC2, DINO-WM, DreamerV3, and Causal-JEPA: Drawer Open is the hardest (all baselines below 40%, four exceed 80% on Cube, and greater sensitivity to long-horizon prediction errors); compared with Causal-JEPA, it improves by 4.00/15.62/2.60 percentage points on Cube/Drawer/Faucet, and surpasses the strongest baseline by 3.34/14.58/2.60 percentage points, with the largest margin on Drawer; subsequent ablations verify each component.

Table 2: Ablation study of modules and loss functions on MetaWorld-Drawer Open and OGBench-Cube (success rate, %). DeCALM consists of Causal-JEPA, BD, MCA, and CE. BD (backdoor intervention) and CE (counterfactual encoding) belong to Background-Deconfounded Latent Encoding (BLE); MCA denotes Multi-Level Causal Alignment. In (b), each loss is removed from the full model. Best results are highlighted in bold.

(a) Ablation Study of Modules

Model	Drawer	Cube
Causal-JEPA	36.98	84.67
Causal-JEPA + BD	45.31	87.33
Causal-JEPA + BD + MCA	50.00	88.00
DeCALM (+ CE)	**52.60**	**88.67**

(b) Ablation Study of Loss Functions

Model	Drawer	Cube
DeCALM	**52.60**	**88.67**
$-\mathcal{L}_{\text{velocity}}$	51.56	86.67
$-\mathcal{L}_{\text{cf}}$	52.08	88.00
$-\mathcal{L}_{\text{inv}}$	48.96	87.33

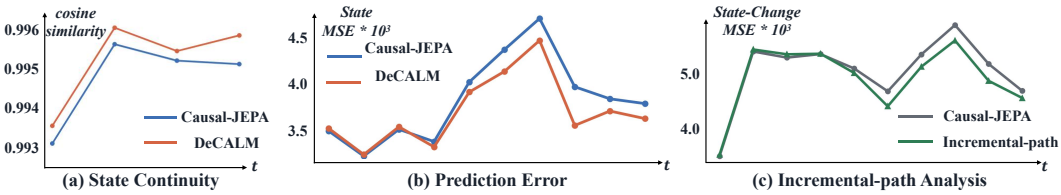


Figure 3: In-depth analyses on OGBench-Cube. (a) Cosine similarity between adjacent predicted states along independently planned trajectories. (b) State-feature MSE over evaluation trajectories. (c) State-change MSE over evaluation trajectories, comparing Causal-JEPA with DeCALM's forced incremental path. MSE values are multiplied by 10^3.

5.3 Ablation Study

In this section, we examine DeCALM components on MetaWorld-Drawer Open and OGBench-Cube in Table 2. Panel (a) successively adds BD, MCA, and CE to Causal-JEPA; panel (b) removes each training loss from the full model. BD gives the largest stepwise gain on both tasks (+8.33 percentage points on Drawer Open and +2.66 on Cube), supporting addressing background-related representation bias before predicting future states. MCA atop BD adds +4.69 points on Drawer Open and +0.67 on Cube; its larger Drawer Open effect is consistent with the value of aligning state changes during extended manipulation. CE further improves both tasks, enabling full DeCALM to achieve its highest success rates of 52.60% and 88.67%, respectively, indicating counterfactual encoding complements background adjustment and multi-level alignment. Removing any of the three losses reduces success, with the largest decreases from removing $\mathcal{L}_{\text{inv}}$ on Drawer Open and $\mathcal{L}_{\text{velocity}}$ on Cube, showing these objectives contribute differently across tasks.

5.4 In-depth Analysis

State continuity. As shown in Figure 3(a), DeCALM has higher adjacent-state cosine similarity at all measured intervals along independently planned trajectories. The smaller directional changes between successive latent states suggest greater temporal continuity in its predictions.

Prediction error. We compare predicted states with observed states across the rollout in Figure 3(b). DeCALM has slightly higher MSE at the earliest measured horizons, but lower MSE thereafter, with a clearer gap at later steps. This indicates lower prediction error over longer horizons.

Incremental-path analysis. To examine state-change prediction, we evaluate DeCALM with its incremental path selected and compare adjacent-state changes with those observed. In Figure 3(c), this path is comparable to Causal-JEPA at early steps and has lower change MSE from the fifth measured step onward. This indicates that the added state-change path more faithfully captures how the latent state evolves at later horizons.

5.5 Case Study

Planning under Background Disturbance. As shown in Figure 4, C-JEPA approaches the goal with a clean history but deviates markedly after a background-only change, whereas DeCALM re-

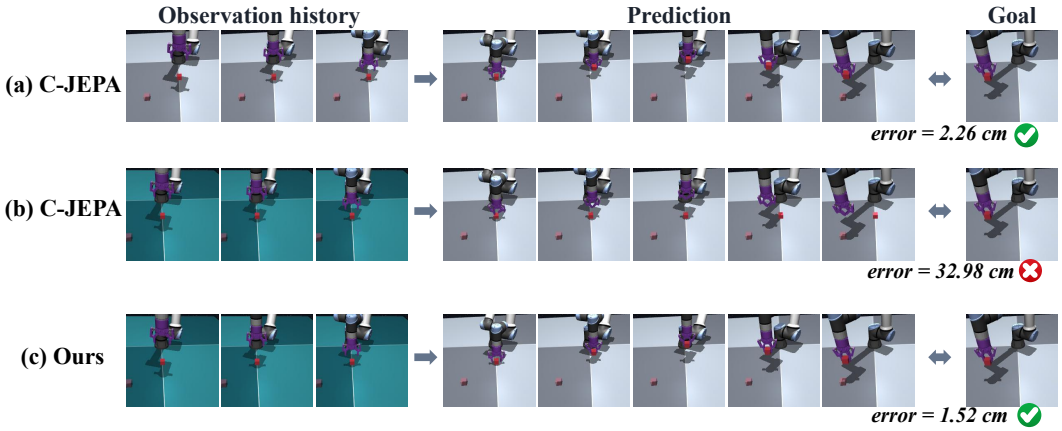


Figure 4: Background-disturbance case on OGBench-Cube. Rows show C-JEPA with clean history, C-JEPA with perturbed history, and DeCALM with the same perturbation. The middle frames show the execution of selected actions; errors are final cube-to-goal distances.

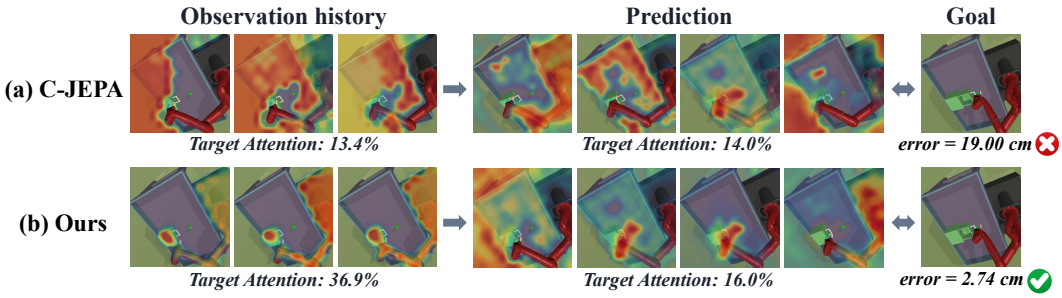


Figure 5: Spatial heatmaps for a selected MetaWorld-Drawer Open case. Rows compare C-JEPA and DeCALM. "Target Attention" reports the mean projected heatmap mass over the robot–drawer region. Error values denote the final handle-to-target distances.

mains close to the goal under the same disturbance. This case therefore demonstrates that DeCALM enables more robust prediction-guided planning under background interference, better preserving goal-directed behavior when the visual context is altered.

Visualization of Task-Relevant Regions. We compare the spatial heatmaps of C-JEPA and DeCALM for a selected Drawer Open case in Figure 5. C-JEPA exhibits broad responses across the scene, whereas DeCALM focuses more on the robot–drawer interaction, especially within the observation history, with a target-region heatmap mass of 36.9% versus 13.4%. This suggests that DeCALM attends more to task-relevant regions during latent prediction.

6 Conclusion

We propose DeCALM (Deconfounded Causally Aligned Latent Model) for deconfounded embodied planning under visual disturbances. Background-Deconfounded Latent Encoding reduces background-cue reliance via backdoor/counterfactual encoding, while Multi-Level Causal Alignment coordinates absolute-state/state-change predictions to limit rollout error accumulation. DeCALM achieves the highest success on three goal-conditioned manipulation tasks; ablations show stepwise gains, and visual analyses indicate lower later-horizon error and stronger task-relevant responses. A limitation is reliance on a training background prototype dictionary, which may degrade out-of-distribution performance, though evaluated disturbances are covered. Future work includes online adaptation, expandable dictionaries, and complex object interactions.

AI USE STATEMENT

Generative AI tools were used to assist in drafting and polishing portions of the manuscript, translating author-provided notes into English, organizing related work, and preparing figures and tables. They also provided feedback on the structural causal formulation and experimental design, assisted in developing experimental code and the mathematical arguments and proofs in the appendix, and helped articulate the interpretation of results. No AI-generated training or test data were used. The authors made the final methodological and experimental decisions, reviewed the AI-assisted text, code, derivations, figures, and analyses, and take full responsibility for the content of this submission.

REPRODUCIBILITY STATEMENT

We have made efforts to support reproduction of the reported results. The Method section specifies the structural causal model, background dictionary, counterfactual objectives, dual-path predictor, and training strategy. The Experiments section describes the OGBench and MetaWorld tasks, disturbed evaluation, success-rate metric, and planning protocol; CEM-based comparisons use the same retrieved action candidates and planning budget. Appendix B states the assumptions and derivations for the theoretical analysis, while Appendix D provides additional visual examples. The benchmark sources are cited in the paper. These descriptions document the principal methodological and evaluation choices without claiming that implementation code has been publicly released.

REFERENCES

Hongye Cao, Fan Feng, Meng Fang, Shaokang Dong, Tianpei Yang, Jing Huo, and Yang Gao. Towards empowerment gain through causal structure learning in model-based reinforcement learning. In *ICLR*, 2025.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *NeurIPS*, volume 31, 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *ICLR*, 2020. URL https://arxiv.org/abs/1912.01603.

Danijar Hafner, Jurgis Pašukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640:647–653, 2025. doi: 10.1038/s41586-025-08744-2.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *ICLR*, 2024.

Kaichen Huang, Shenghua Wan, Minghao Shao, Hai-Hang Sun, Le Gan, Shuai Feng, and De-Chuan Zhan. Leveraging separated world model for exploration in visually distracted environments. In *NeurIPS*, volume 37, pp. 82350–82374, 2024.

Miles Hutson, Isaac Kauvar, and Nick Haber. Policy-shaped prediction: Avoiding distractions in model-based reinforcement learning. In *NeurIPS*, volume 37, pp. 13124–13148, 2024.

Inwoo Hwang, Yunhyeok Kwak, Suhyung Choi, Byoung-Tak Zhang, and Sanghack Lee. Fine-grained causal dynamics learning with quantization for improving robustness in reinforcement learning. In *ICML*, volume 235, pp. 20842–20870, 2024.

Albina Klepach, Alexander Nikulin, Ilya Zisman, Denis Tarasov, Alexander Derevyagin, Andrei Polubarov, Nikita Lyubaykin, Igor Kiselev, and Vladislav Kurenkov. Object-centric latent action learning. In *AAAI*, volume 40, pp. 22626–22634, 2026. doi: 10.1609/aaai.v40i27.39423.

Anson Lei, Bernhard Schölkopf, and Ingmar Posner. SPARTAN: A sparse transformer world model attending to what matters. In *NeurIPS*, volume 38, 2025. doi: 10.52202/085713-5154.

Yichao Liang, Dat Nguyen, Cambridge Yang, Tianyang Li, Joshua B. Tenenbaum, Carl Edward Rasmussen, Adrian Weller, Zenna Tavares, Tom Silver, and Kevin Ellis. ExoPredicator: Learning abstract models of dynamic worlds for robot planning. In *ICLR*, 2026.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *NeurIPS*, 2020.

Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *ICLR*, 2023. URL https://openreview.net/forum?id=vhFu1Acb0xb.

Naoki Morihira, Amal Nahar, Kartik Bharadwaj, Yasuhiro Kato, Akinobu Hayashi, and Tatsuya Harada. R2-Dreamer: Redundancy-reduced world models without decoders or augmentation. In *ICLR*, 2026. URL https://openreview.net/forum?id=Je2QqXrcQq.

Alejandro Murillo-González, Junhong Xu, and Lantao Liu. Learning causal structure distributions for robust planning. *IEEE Robotics and Automation Letters*, 10(10):9916–9923, 2025. doi: 10.1109/LRA.2025.3598663.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *ICML*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.11389.

Alexander Nikulin, Ilya Zisman, Denis Tarasov, Nikita Lyubaykin, Andrei Polubarov, Igor Kiselev, and Vladislav Kurenkov. Latent action learning requires supervision in the presence of distractors. In *ICML*, volume 267, pp. 46427–46447, 2025.

Yosuke Nishimoto and Takashi Matsubara. Object-centric world models for causality-aware reinforcement learning. In *AAAI*, volume 40, pp. 24585–24593, 2026. doi: 10.1609/aaai.v40i29.39642.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Minting Pan, Xiangming Zhu, Yunbo Wang, and Xiaokang Yang. Iso-Dream: Isolating and leveraging noncontrollable visual dynamics in world models. In *NeurIPS*, volume 35, pp. 23178–23191, 2022.

Seohong Park, Kevin Frans, Benjamin Eysenbach, and Sergey Levine. OGBench: Benchmarking offline goal-conditioned RL. In *ICLR*, 2025. URL https://openreview.net/pdf?id=M992mjgKzI.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, Timothy Lillicrap, and David Silver. Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, 588:604–609, 2020. doi: 10.1038/s41586-020-03051-4.

Uladzislau Sobal, Wancong Zhang, Kyunghyun Cho, Randall Balestriero, Tim G. J. Rudner, and Yann LeCun. Learning from reward-free offline data: A case for planning with latent dynamics models. In *NeurIPS*, volume 38, 2025. doi: 10.52202/085713-1465.

Ruixiang Sun, Hongyu Zang, Xin Li, and Riashat Islam. Learning latent dynamic robust representations for world models. In *ICML*, volume 235, pp. 47234–47260, 2024.

Shiguang Sun, Hanbo Zhang, Zeyang Liu, Xinrui Yang, Lipeng Wan, Xingyu Chen, and Xuguang Lan. MInCo: Mitigating conflicting objectives in distracted visual model-based reinforcement learning. *Knowledge-Based Systems*, 350:116309, 2026. doi: 10.1016/j.knosys.2026.116309.

Leonardo F. Toso, Davit Shadunts, Yunyang Lu, Nihal Sharma, Donglin Zhan, Nam H. Nguyen, and James Anderson. Learning invariant visual representations for planning with joint-embedding predictive world models. *arXiv preprint arXiv:2602.18639*, 2026. URL https://arxiv.org/abs/2602.18639.

Xinyue Wang and Biwei Huang. Modeling unseen environments with language-guided composable causal components in reinforcement learning. In *ICLR*, 2025.

Yucen Wang, Shenghua Wan, Le Gan, Shuai Feng, and De-Chuan Zhan. AD3: Implicit action is the key for world models to distinguish the diverse visual distractors. In *ICML*, volume 235, pp. 51546–51568, 2024a.

Zizhao Wang, Caroline Wang, Xuesu Xiao, Yuke Zhu, and Peter Stone. Building minimal and reusable causal state abstractions for reinforcement learning. In *AAAI*, volume 38, pp. 15778–15786, 2024b. doi: 10.1609/aaai.v38i14.29507.

Yupei Yang, Biwei Huang, Fan Feng, Xinyue Wang, Shikui Tu, and Lei Xu. Towards generalizable reinforcement learning via causality-guided self-adaptive representations. In *ICLR*, 2025.

Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *CoRL*, 2020.

Zhongwei Yu, Jingqing Ruan, and Dengpeng Xing. Learning causal dynamics models in object-oriented environments. In *ICML*, volume 235, pp. 57597–57638, 2024.

Zhiyu Zhao, Haoxuan Li, Haifeng Zhang, Jun Wang, Francesco Faccio, Jürgen Schmidhuber, and Mengyue Yang. Curious causality-seeking agents in open-ended worlds. In *NeurIPS*, volume 38, 2025.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *ICML*, 2025.

A Causal View of Vision-Based Embodied Planning

We use a structural causal model to model the variable relationships in latent dynamics for embodied planning. As illustrated in Figure 6, it is a directed acyclic graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where nodes denote variables and edges denote their dependencies. The SCM includes six variables: observation O_t, background and object representations C_t and S_t, observable background confounder U, action A_t, and next-state prediction Y. The relationships in $\mathcal{G}$ are as follows:

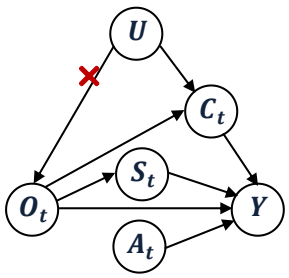


Figure 6: Structural causal model of latent dynamics. The observation O_t provides background and object representations C_t and S_t, which, together with action A_t, contribute to predicting Y. The observable background confounder U induces the backdoor path $O_t \leftarrow U \rightarrow C_t \rightarrow Y$, motivating causal intervention for learning deconfounded dynamics.

$\boldsymbol{U \rightarrow C_t \rightarrow Y}$: U represents the observable background confounder, which we model using a background dictionary. The link $U \rightarrow C_t$ describes how background bias enters the representation, further influencing the prediction of Y through $C_t \rightarrow Y$.

$\boldsymbol{O_t \rightarrow C_t \rightarrow Y,\ O_t \rightarrow S_t \rightarrow Y,\ A_t \rightarrow Y}$: C_t and S_t represent contextual and object information extracted from the observation. These pathways describe the estimation of Y based on background context, the current object state, and the applied action.

$\boldsymbol{O_t \leftarrow U \rightarrow C_t \rightarrow Y}$: We use backdoor adjustment to mitigate the influence of the background confounder U along this path, which introduces spurious associations between background cues and future-state predictions. Counterfactual consistency further discourages reliance on these cues when object representations and actions remain unchanged.

Training an implicit world model focuses on learning action-driven state transitions. It aims to predict Y from task-relevant information while reducing spurious correlations induced by U. This addresses a source of representation bias that can distort action selection and accumulate over multiple steps in embodied planning.

B Theoretical Analysis of Proposed DeCALM

We follow the notation of Section 4: S_t is the current object representation, $Y^{\mathrm{gt}} = f_S(O_{t+1})$ is the next-state target, Y^{abs} and $Y^{\Delta} = S_t + \Delta Y_t$ are the two path predictions, and Y is their gated fusion. All norms below are taken over the object-slot representation.

B.1 Visual Disturbance and Background Adjustment

Consider an observation o with object representation $s = f_S(o)$, a candidate action a, and the background strata $U = u_1, \ldots, u_K$. We introduce two auxiliary quantities for the analysis. $\mu_k(s, a)$ denotes the expected next-state target for object representation s and action a within stratum u_k; $q_k(o)$ denotes the probability that observation o belongs to that stratum. Formally,

$$\mu_k(s,a) = \mathbb{E}\big[Y^{\mathrm{gt}} \mid S_t = s, A_t = a, U = u_k\big], \qquad q_k(o) = P(U = u_k \mid O_t = o). \tag{17}$$

Suppose that S_t retains the task-relevant information needed for predicting Y^{gt} within each stratum, the action-conditioned query satisfies $U \perp A_t \mid O_t$, and the relevant strata have observational support. The observational and background-standardized prediction targets are then

$$m_{\mathrm{obs}}(o,a) = \sum_{k=1}^{K} q_k(o)\mu_k(s,a), \qquad m_{\mathrm{adj}}(s,a) = \sum_{k=1}^{K} \pi_k \mu_k(s,a), \tag{18}$$

where $\pi_k = P(U = u_k)$ is the background prior. Their difference is

$$m_{\mathrm{obs}}(o,a) - m_{\mathrm{adj}}(s,a) = \sum_{k=1}^{K} \big(q_k(o) - \pi_k\big)\mu_k(s,a). \tag{19}$$

Thus, if a visual disturbance changes the background weights $q_k(o)$ while the task-relevant state and action remain fixed, the observational prediction can change for reasons unrelated to the intended state transition. Standardizing the weights to π_k removes this particular source of variation in the idealized target. The dictionary integration in Equations (7) and (8) approximates this operation in feature space.

B.2 Absolute-State Prediction and Error Propagation

Let G denote the goal representation and let $\ell(\cdot, G)$ be a goal cost that is L-Lipschitz in its predicted state. For a set of candidate action sequences $\mathcal{A}_{\text{cand}}$, write $Y_H^{\text{abs}}(a)$ for the terminal state predicted by an absolute-only rollout under candidate a, and $Y_H^{\text{gt}}(a)$ for its corresponding true latent state. Let $\hat{a}$ minimize the predicted cost and $a^\star$ minimize the true cost. If

$$\varepsilon_{\text{abs}} = \max_{a \in \mathcal{A}_{\text{cand}}} \left\| Y_H^{\text{abs}}(a) - Y_H^{\text{gt}}(a) \right\|, \tag{20}$$

then

$$0 \le \ell\big(Y_H^{\text{gt}}(\hat{a}), G\big) - \ell\big(Y_H^{\text{gt}}(a^\star), G\big) \le 2L\varepsilon_{\text{abs}}. \tag{21}$$

Indeed, the cost assigned to any candidate changes by at most $L\varepsilon_{\text{abs}}$ when its predicted terminal state is replaced by the true one. Applying this observation to $\hat{a}$ and $a^\star$ proves the bound. Hence state-prediction errors can alter the ranking of candidate actions, although a ranking change is not inevitable.

To examine how such errors reach later prediction steps, let $\mathcal{H}_h$ contain the object-slot history supplied to the predictor at rollout step h. Let T_{abs} denote one absolute-path prediction followed by the history update, and let $\widehat{\mathcal{H}}_h$ and $\mathcal{H}_h^{\text{gt}}$ be the predicted and reference histories. Assume T_{abs} is ρ-Lipschitz in this history. Defining

$$E_h = \left\| \widehat{\mathcal{H}}_h - \mathcal{H}_h^{\text{gt}} \right\|, \qquad \varepsilon_h = \left\| T_{\text{abs}}(\mathcal{H}_h^{\text{gt}}, A_{t+h}) - \mathcal{H}_{h+1}^{\text{gt}} \right\|, \tag{22}$$

the triangle inequality gives

$$E_{h+1} \le \rho E_h + \varepsilon_h, \qquad E_H \le \rho^H E_0 + \sum_{h=0}^{H-1} \rho^{H-1-h} \varepsilon_h. \tag{23}$$

The first term carries forward an existing history error; the second collects errors introduced at subsequent steps. For example, when $\rho = 1$ and every $\varepsilon_h \le \varepsilon$, the bound becomes $E_H \le E_0 + H\varepsilon$.

B.3 State-Change Prediction and Multi-Level Alignment

The incremental path has a direct next-state interpretation. Because $Y^\Delta = S_t + \Delta Y_t$, its supervision satisfies

$$\mathcal{L}_{\text{velocity}} = \text{MSE}\big(\Delta Y_t, Y^{\text{gt}} - S_t\big) = \text{MSE}\big(Y^\Delta, Y^{\text{gt}}\big). \tag{24}$$

The equality follows by adding S_t to both arguments of the first MSE. Thus, the velocity loss directly supervises a second estimate of the next object state, expressed through its change from the current state.

During a rollout, let $\hat{S}_{t+h}$ be the current predicted object state and S_{t+h} its reference counterpart. Define

$$e_h = \hat{S}_{t+h} - S_{t+h}, \quad \alpha_h = Y_{t+h}^{\text{abs}} - S_{t+h+1}, \quad \delta_h = \Delta Y_{t+h} - (S_{t+h+1} - S_{t+h}). \tag{25}$$

Applying the fusion rule in Equation (12) yields the exact error recursion

$$e_{h+1} = g_{t+h} \odot \alpha_h + (1 - g_{t+h}) \odot (e_h + \delta_h). \tag{26}$$

The absolute path supplies a fresh next-state estimate, whereas the incremental path carries the current state forward and adds a predicted change.

A sufficient condition makes the potential effect on accumulated error explicit. Suppose every gate coordinate is at least $\gamma > 0$ and, along the evaluated rollout, $\|\alpha_h\| \le \varepsilon_{\text{abs}}$ and $\|\delta_h\| \le \varepsilon_\Delta$. Then

$$\|e_H\| \le (1-\gamma)^H \|e_0\| + \big(\varepsilon_{\text{abs}} + (1-\gamma)\varepsilon_\Delta\big) \sum_{j=0}^{H-1} (1-\gamma)^j. \tag{27}$$

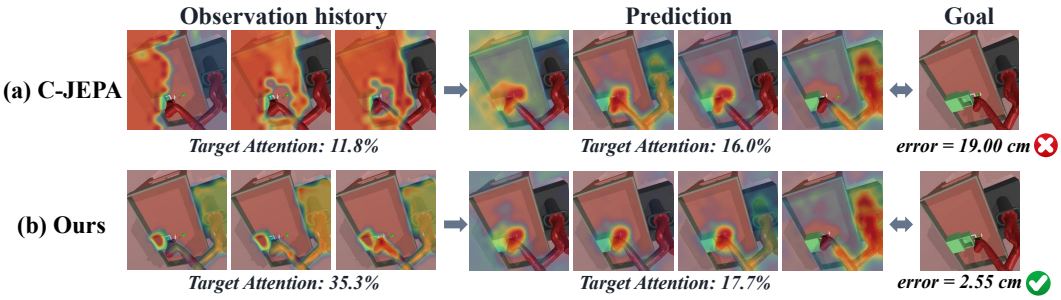


Figure 7: Spatial heatmaps in an additional MetaWorld-Drawer Open case. Rows compare C-JEPA and DeCALM. "Target Attention" denotes the projected heatmap mass on the robot–drawer region. Errors are final handle-to-target distances.

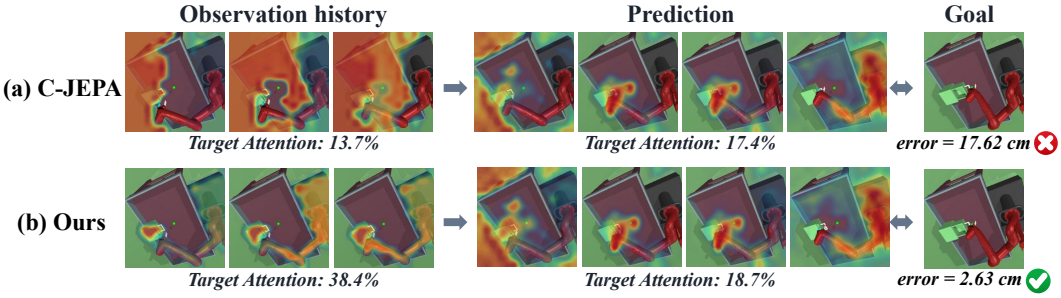


Figure 8: Spatial heatmaps in another MetaWorld-Drawer Open case. Rows compare C-JEPA and DeCALM. "Target Attention" denotes the projected heatmap mass on the robot–drawer region. Errors are final handle-to-target distances.

This follows by taking norms in Equation (26) and applying the resulting one-step inequality repeatedly. Under the stated conditions, the portion of the previous-state error carried explicitly by the incremental path is attenuated by $1 - \gamma$ at each step.

The two paths can also reduce a background-dependent error when their errors differ. As a simple example, suppose the reference transition is $S_{t+h+1} = S_{t+h} + d(A_{t+h})$, an absolute predictor returns $\hat{S}_{t+h} + d(A_{t+h}) + b(U)$, and the incremental path predicts $d(A_{t+h})$ exactly. Starting from $e_0 = 0$, an absolute-only rollout has $e_H = Hb(U)$, whereas a constant scalar gate $g \in [0, 1)$ gives $e_H = gHb(U)$.

C Additional Implementation Details

DeCALM uses four 128-dimensional object slots and one explicit background slot. The background dictionaries are constructed only from training slots, with eight prototypes for OGBench-Cube and two for each MetaWorld task. Final fine-tuning uses AdamW for five epochs with a batch size of 64 per GPU. The counterfactual substitution probability is 0.3, with $\lambda_{\text{cf}} = 1.0$ and $\lambda_{\text{inv}} = 0.1$. At the first planning call, 32 of the 300 CEM candidates are initialized from training action sequences.

For baseline reproduction, IRIS (Micheli et al., 2023) uses VQ image tokens and an action codebook derived from training data. DINO-WM (Zhou et al., 2025) retains frozen DINOv2 features, whereas Causal-JEPA (Nam et al., 2026) uses object-centric slots. DreamerV3 (Hafner et al., 2025) is adapted to reward-free goal-image MPC; TD-MPC2 (Hansen et al., 2024) retains reward/Q learning and its MPPI controller. The methods use the corresponding offline task data and the same registered evaluation cases, while retaining their method-specific encoders and control objectives.

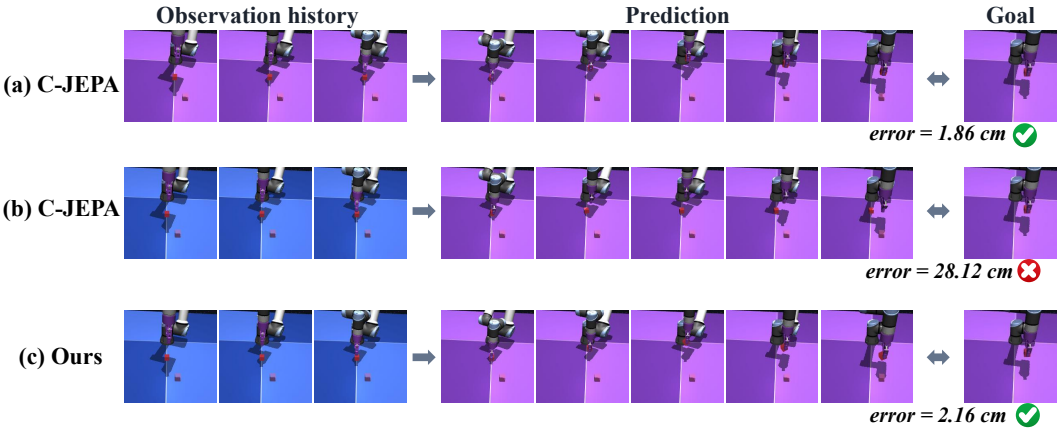


Figure 9: Additional background-disturbance case on OGBench-Cube. Rows show C-JEPA with clean history, C-JEPA with perturbed history, and DeCALM with the same perturbation. The middle frames show predicted future states; errors are terminal cube-to-goal distances.

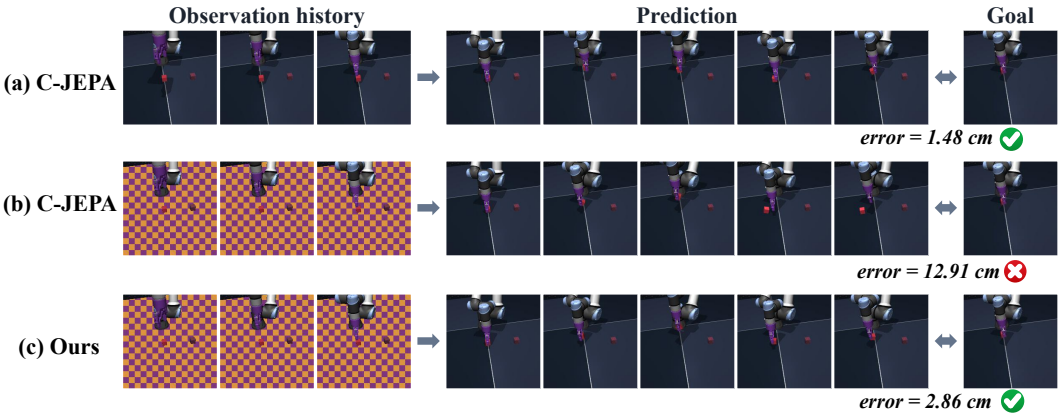


Figure 10: Another background-disturbance case on OGBench-Cube. Rows show C-JEPA with clean history, C-JEPA with a checkerboard background in its history, and DeCALM with the same perturbation. The middle frames show predicted future states; errors are terminal cube-to-goal distances.

D More Visualization Results

Additional Heatmap Cases. Figures 7 and 8 show two additional Drawer Open cases. Compared with C-JEPA, DeCALM places more heatmap mass on the robot–drawer region in the observation history (35.3% versus 11.8%, and 38.4% versus 13.7%). In these cases, DeCALM also ends closer to the target, providing further examples of its task-focused responses.

Future Prediction under Background Disturbance. As shown in Figures 9 and 10, C-JEPA's terminal prediction errors are 1.86 and 1.48 cm with clean observation histories, but rise to 28.12 and 12.91 cm when only the backgrounds change. Under the same respective perturbations, DeCALM's errors are 2.16 and 2.86 cm. Because the object configurations and goals remain unchanged within each case, these contrasts reflect sensitivity to visual context rather than changes in task dynamics. In both cases, DeCALM maintains predictions closer to the goal under background disturbance.