CST-WM: A Causally Structured World Model for Embodied Visual Tracking

Junyi Hu **Shuaihang Yuan** **Yi Fang***
New York University Abu Dhabi
jh10472@nyu.edu

Abstract

Embodied visual tracking requires a robot not only to react to the current view, but also to choose actions that preserve or recover future evidence of a moving target under ego-motion, occlusion, and distractors. This makes embodied visual tracking a predictive decision problem over future target observability and apparent scale. A central difficulty is a task-specific form of causal hallucination: in action-conditioned sequential prediction, a model can exploit the strong correlation between robot control and target-related observations by hallucinating a direct causal effect from the current action to target evidence, rather than allowing action to influence such evidence only through robot motion and the resulting observation change. This shortcut can yield plausible futures while giving the wrong semantics for tracking-oriented planning and target re-acquisition. We propose CST-WM, a causally structured world model for embodied visual tracking. CST-WM decomposes the latent state into a target-evidence branch, a robot branch, and an observation branch, and factorizes the transition so that direct action injection into the target-evidence branch is blocked while action remains available to robot motion and observation updates. Combined with rollout-based model-predictive control, CST-WM supports both stable following and temporary target re-acquisition within a unified planning framework. Experiments on EVT-Bench and Habitat 3.0, including standard tracking, target-loss recovery, and cross-dataset transfer, show improved following quality, distance-range control, safety, and re-acquisition over reactive and world-model baselines. Offline diagnostics further show better multi-step rollout fidelity, stronger planning-value consistency, and substantially reduced direct action leakage. These results suggest that, for embodied visual tracking, future prediction alone is not enough; the predictive structure itself must align with how target evidence enters planning.

Project page: https://junyi2005.github.io/cst-wm/

1 Introduction

Embodied visual tracking [23, 39, 40, 28] appears simple. A robot only needs to stay with a moving target. In practice, however, reliable following requires the robot to preserve target observability, regulate following distance, and remain safe while both agents move through cluttered, partially observable environments [24, 38, 34]. The hardest moments are not when the target is clearly visible, but when it is temporarily occluded, leaves the field of view, or becomes confusable with distractors [37, 19, 17]. In those cases, the robot must decide how to move before reliable target evidence is available again. The value of an action therefore depends not only on how it changes the robot state, but also on how it changes future target visibility and apparent scale. Embodied visual tracking is thus fundamentally a predictive decision problem rather than a purely reactive mapping from the current frame to the next control [10, 12, 2]. Under such conditions, spurious action–evidence correlations can distort the model's notion of target evidence during prediction.

*Corresponding author.

Preprint.

We refer to this failure mode as causal hallucination: in action-conditioned sequential prediction, a model can exploit the strong correlation between robot control and target-related observations by hallucinating a direct causal effect from the current action to target evidence, rather than allowing action to influence such evidence only through robot motion and the resulting observation change.

This viewpoint exposes a limitation of both dominant formulations. Reactive trackers [8, 40, 28, 26, 39, 23, 36, 16, 4] are inherently myopic because, once the target disappears or becomes ambiguous, they offer little basis for choosing actions that support multi-step recovery. Yet a generic action-conditioned world model [2, 11, 14, 7] is not sufficient either. In embodied tracking, the current action should influence future target evidence by moving the robot and changing what will be observed next, not by being written directly into the target-related state itself. Otherwise, the predictor can produce plausible futures while conflating robot control with target evidence, thereby inducing causal hallucination in the target-related latent update and giving the wrong semantics for re-acquisition and tracking-oriented planning. The key point here is a structural requirement on the planning representation: for tracking, target evidence should evolve through the consequences of ego-motion on subsequent observations, rather than through a direct action shortcut in the latent state. This keeps the target-evidence state tied to observability and apparent scale, instead of allowing it to absorb control information that should be mediated by robot motion. For this task, the core issue is not only whether the future is predicted, but whether the action pathway in that prediction matches the structure of tracking under ego-motion.

As illustrated in Fig. 1, our key observation is that embodied visual tracking does not require full recovery of the target's external state at every step. For planning, the controller mainly needs a compact representation of target evidence that indicates whether the target is observable, how strong that evidence is, and whether apparent scale remains compatible with a valid following distance. Based on this view, we propose CST-WM, a causally structured world model that represents the tracking state with a target-evidence branch, a robot branch, and an observation branch. Its transition is explicitly factorized so that the target-evidence branch is updated without direct action injection, while action is localized to the robot branch and only reaches future observation through the architecturally permitted route. This design lets the same model support both stable following and temporary target re-acquisition within one rollout-based planning framework.

We evaluate CST-WM on EVT-Bench [28] and Habitat 3.0 [25] across three perspectives—standard tracking, target-loss recovery, and controlled structural diagnostics. CST-WM improves following quality, distance-range control, safety, and re-acquisition over reactive and world-model baselines, and yields better multi-step rollout fidelity, stronger agreement between model-based and simulator-based ranking of candidates, and substantially reduced direct action leakage. The present evidence supports improved tracking and planning under the tested settings rather than a stronger claim about full recovery of external target state.

We summarize our contributions as follows: (i) we formulate embodied visual tracking as planning over future target evidence and propose CST-WM, a causally structured world model that separates target evidence, robot motion, and observation dynamics while blocking direct action injection into the target-evidence branch; (ii) we introduce a compact planning-oriented target-evidence representation that summarizes observability and apparent scale for following and recovery, while supporting safety-aware planning without claiming full target-state recovery or requiring privileged geometric supervision at test time; and (iii) we provide a comprehensive evaluation on EVT-Bench and Habitat 3.0 covering standard tracking, temporary target loss, rollout fidelity, planning-value consistency, and leakage diagnostics, showing that the structured transition improves both downstream control and the internal quality of imagined futures.

2 Related Work

2.1 Embodied visual tracking

Embodied visual tracking requires a robot to follow a moving target from egocentric observations while preserving visibility, regulating following distance, and remaining safe [34, 33, 9]. Detection-plus-planning systems achieve basic following but degrade when the target becomes occluded or visually ambiguous in cluttered scenes. A major recent direction is reactive policy learning: reinforcement-learning-based active trackers [8, 1, 37, 35] and offline-RL extensions such as EVT [40, 27, 20], alongside social-aware and vision-language-action variants that map rich visual context to control [26, 31, 3, 22, 36, 28, 19]. These methods strengthen robustness under occlusion and

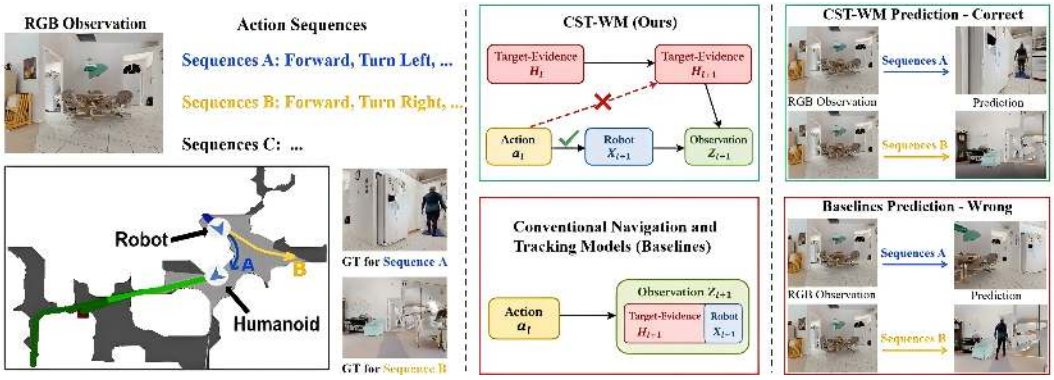

Figure 1: Causal structure of action-conditioned prediction in embodied visual tracking. Given an egocentric observation and candidate action sequences (A: forward, turn-left; B: forward, turn-right), the model must imagine which sequence preserves target observability and following distance. Conventional world models route the current action directly into the target-evidence update, producing plausible rollouts that fail to discriminate A from B. CST-WM removes this direct edge: the target-evidence branch $H_{\ell+1}$ is updated without action injection, and the action only reaches future observations through the robot branch $x_{\ell+1}$. The resulting predictions distinguish the two sequences and agree with the simulator ground truth, while baseline predictions do not.

distractors, but they still mainly map recent observations to actions rather than explicitly comparing how alternative actions affect future target observability and apparent scale across multiple steps.

2.2 World models

World models instead learn latent dynamics from observation histories and support planning by rolling out futures before action execution. From early latent-dynamics methods [10, 12] to the Dreamer family and its embodied extensions [11, 13, 14, 32], and to broader generative or driving environments [7, 29, 30, 18, 15], this paradigm enables comparing multiple candidate futures before acting. The most directly relevant prior work in embodied navigation is Navigation World Models [2], which shows that diffusion-based world modeling can support egocentric planning. Yet a generic action-conditioned transition can write the current action directly into the target-evidence branch, producing a tracking-specific form of causal hallucination in which futures look plausible while assigning the wrong semantics to target evidence for re-acquisition. We therefore decompose dynamics into target-evidence, robot, and observation branches: the target-evidence branch excludes direct action injection while action remains available to motion and observation updates, preserving the predictive advantage of world models while aligning imagined futures with how target evidence should enter tracking-oriented planning.

3 Method

3.1 Problem Definition

We consider an embodied visual tracking setting in which, at each timestamp ℓ, the robot receives an egocentric RGB observation $O_\ell \in \mathbb{R}^{H\times W\times 3}$ and selects the next action $a_\ell = (v_\ell, \omega_\ell) \in \mathcal{A} \subseteq \mathbb{R}^3$, with linear velocity $v_\ell \in \mathbb{R}^2$ and angular velocity $\omega_\ell \in \mathbb{R}$. Starting from any position in a scene $E \in \mathcal{E}_{\text{scene}}$ and without privileged target states, the task is successful if the agent locates the moving target and maintains an appropriate following distance (1–3 m in our protocol) while facing toward it. Letting $a_{0:T-1}$ and $O_{0:T}$ denote the corresponding action and observation sequences, the tracking objective selects actions that preserve or recover future target evidence under partial observability. Unlike static-goal navigation, the future utility of an action depends on how it changes both the robot state and future target observability and apparent scale, making embodied tracking a predictive decision problem rather than a purely reactive frame-to-control mapping.

Our framework, CST-WM, uses a single world model to predict the evolving tracking state across different agent initializations. The central modeling challenge is to extract tracking-relevant information from egocentric observations *and* to ensure that the action pathway inside prediction matches the structure required by tracking under ego-motion: a generic action-conditioned transition is not sufficient because it may spuriously inject the current action into the target-evidence representation instead of letting action influence future evidence through robot motion and the resulting observation

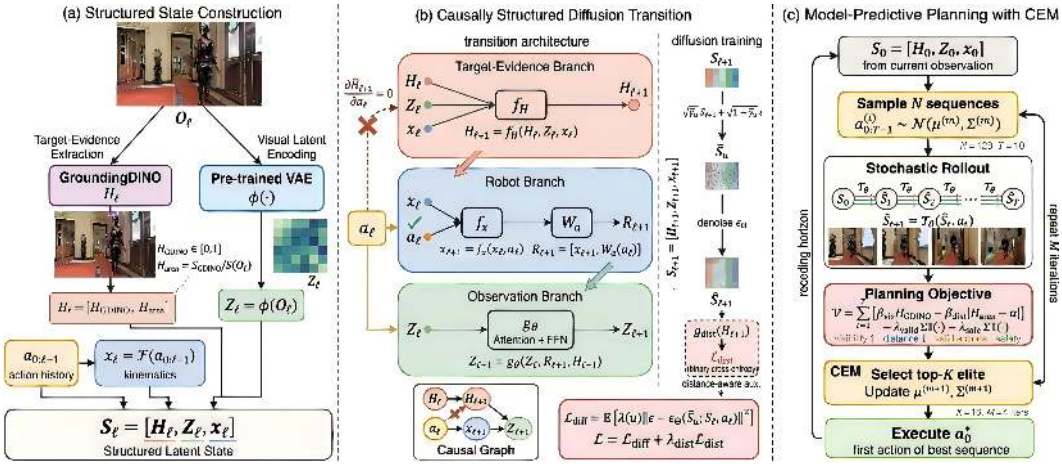

Figure 2: Overall pipeline of CST-WM. (a) *Structured state construction.* The egocentric observation O_ℓ yields a target-evidence token H_ℓ (GroundingDINO confidence and normalized target area) and a visual latent Z_ℓ (frozen VAE); the robot state x_ℓ is integrated from the action history. They form $S_\ell = [H_\ell, Z_\ell, x_\ell]$. (b) *Causally structured diffusion transition.* The three branches are denoised in a fixed order under architectural masking: $H_{\ell+1}$ is updated without reading a_ℓ; $x_{\ell+1}$ absorbs the current action and forms the sole action carrier $R_{\ell+1}$; the observation branch then fuses both to produce $Z_{\ell+1}$. The causal graph contains no direct edge from a_ℓ to $H_{\ell+1}$. (c) *Model-predictive planning with CEM.* Starting from S_0, N candidate sequences are rolled out, scored by the planning objective, and used to update the sampling distribution; after M iterations the first action of the best top-K sequence is executed in receding-horizon form.

change. The architecture, illustrated in Figure 2, enforces this structural constraint on the planning representation.

3.2 Causally Structured World Model

3.2.1 Information Extraction and Causal Consistency

At timestamp ℓ we extract a target-evidence representation H_ℓ from O_ℓ. We use GroundingDINO [21] to obtain an open-vocabulary detection confidence $H_{\text{GDINO}}(O_\ell) \in [0,1]$ and the corresponding bounding box, then normalize the box area S_{GDINO} by the image area $S(O_\ell)$ to give $H_{\text{area}}(O_\ell) = S_{\text{GDINO}}/S(O_\ell) \in [0,1]$, and collect the two cues into

$$H_\ell = [\, H_\ell^{\text{vis}},\, H_\ell^{\text{area}} \,] = [\, H_{\text{GDINO}}(O_\ell),\, H_{\text{area}}(O_\ell) \,] \in [0,1]^2. \tag{1}$$

Here H_ℓ^{vis} reflects whether the target is found and H_ℓ^{area} provides a qualitative estimate of following distance through apparent scale. H_ℓ is an observation-derived target-evidence representation rather than a full external state estimate; it summarizes target observability and apparent scale for tracking-oriented planning, avoiding privileged geometric supervision at inference time. We further encode the observation with a pre-trained VAE [5], $Z_\ell = \phi(O_\ell)$, and recover the robot state $x_\ell \in \mathbb{R}^3$ from the action history under the kinematic update $x_\ell = \mathcal{F}(a_{0:\ell-1})$, giving the structured state $S_\ell = [H_\ell, Z_\ell, x_\ell]$.

A central modeling requirement is that the target-evidence branch should not receive direct action injection: robot action can influence future target evidence *indirectly* by changing ego-motion and thereby the subsequent observation, but, for the observation-derived representation used here, the current action should not be written directly into its update. We therefore impose the architectural non-leakage constraint

$$p(H_{\ell+1} \mid H_\ell, Z_\ell, x_\ell, a_\ell) = p(H_{\ell+1} \mid H_\ell, Z_\ell, x_\ell), \quad \frac{\partial \widehat{H}_{\ell+1}}{\partial a_\ell} = 0. \tag{2}$$

This is a structural requirement on the planning representation rather than a strong physical claim: it keeps the target-evidence branch tied to observability and apparent scale, instead of letting it absorb control information that should be mediated by robot motion. Under this constraint, the one-step transition factorizes into target-evidence, robot, and observation components,

$$p(H_{\ell+1}, x_{\ell+1}, Z_{\ell+1} | H_\ell, Z_\ell, x_\ell, a_\ell) = p(H_{\ell+1}|H_\ell, Z_\ell, x_\ell)\, p(x_{\ell+1}|x_\ell, a_\ell)\, p(Z_{\ell+1}|Z_\ell, H_{\ell+1}, x_{\ell+1}) \tag{3}$$

so action is excluded only from the direct update of the target-evidence branch and is allowed to affect future observations through the robot branch.

3.2.2 Single-Layer Causal Architecture

Each layer updates the three branches in the fixed order target-evidence, robot, observation, with strict masking that prevents any action-carrying representation from entering the target-evidence update. Concretely, $H_{\ell+1} = f_H(H_\ell, Z_\ell, x_\ell)$, $x_{\ell+1} = f_x(x_\ell, a_\ell)$ and $R_{\ell+1} = [\, x_{\ell+1}, W_a(a_\ell)\,]$, where $R_{\ell+1}$ is the sole representation that carries current-action information into the observation branch, $Z_{\ell+1} = g_\theta(Z_\ell; R_{\ell+1}, H_{\ell+1})$, with g_θ implemented by attention and feed-forward operators. The full transition operator is $\mathcal{T}_\theta : (H_\ell, x_\ell, Z_\ell, a_\ell) \mapsto (H_{\ell+1}, x_{\ell+1}, Z_{\ell+1})$ with computational graph $a_\ell \to x_{\ell+1} \to Z_{\ell+1}$, $H_\ell \to H_{\ell+1} \to Z_{\ell+1}$ and no direct edge from a_ℓ to $H_{\ell+1}$. The zero Jacobian property of the target-evidence predictor with respect to the current action then serves as a diagnostic that the architectural constraint is realized in practice.

3.2.3 Diffusion Training

Training uses transitions $(O_\ell, a_\ell, O_{\ell+1})$ from EVT-Bench [28] and Habitat 3.0 [25]; data filtering and the recovery protocol are described in Section 4.1. The model predicts the next structured state $S_{\ell+1} = [H_{\ell+1}, Z_{\ell+1}, x_{\ell+1}]$ from S_ℓ and a_ℓ. We corrupt $S_{\ell+1}$ with Gaussian noise under a DDPM scheduler, $\tilde{S}_u = \sqrt{\bar{\gamma}_u}\, S_{\ell+1} + \sqrt{1-\bar{\gamma}_u}\, \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, I)$, and train the denoising network ϵ_Θ with

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{\ell,u,\varepsilon}\Big[\lambda(u)\big\|\varepsilon - \epsilon_\Theta(\tilde{S}_u; S_\ell, a_\ell)\big\|_2^2\Big]. \tag{4}$$

The three branches are denoised under a single objective over the shared next state, while causal masking determines the permitted dependencies inside the transition architecture. We additionally attach a lightweight prediction head g_{dist} to the target-evidence branch and supervise it with a binary in-range distance label $y^{\text{dist}}_{\ell+1} = \mathbb{I}(d^- \le d_{\ell+1} \le d^+)$ derived from simulator-only relative distance $d_{\ell+1}$, giving the auxiliary distance-aware loss

$$\mathcal{L}_{\text{dist}} = -\mathbb{E}_\ell\big[\, y^{\text{dist}}_{\ell+1} \log g_{\text{dist}}(H_{\ell+1}) + (1 - y^{\text{dist}}_{\ell+1}) \log(1 - g_{\text{dist}}(H_{\ell+1}))\,\big]. \tag{5}$$

The simulator distance enters only through this offline supervision and is never available at test time. The final objective is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}}$ with $\lambda_{\text{dist}} = 0.2$ by default; the auxiliary term is active only in distance-aware variants and removing it leaves the architecture and the diffusion objective unchanged. For inference-time rollout, $\widehat{S}_{\ell+1} = \mathcal{T}_\theta(S_\ell, a_\ell)$ is applied recursively over a candidate action sequence of length T.

3.3 Tracking Task Planning with World Models

With a trained CST-WM, we plan by simulating candidate action sequences under the learned dynamics. Given $S_0 = [H_0, Z_0, x_0]$, the planner seeks actions that preserve a valid following configuration when the target is visible, and that recover target evidence while preserving safety when the target is temporarily out of view, handling stable following and target re-acquisition within one rollout-based procedure. We optimize action sequences with the Cross-Entropy Method [6], using horizon $T = 10$, candidate budget $N = 128$, elite size $K = 16$ and $M = 4$ distribution-update iterations: at iteration m, N sequences $a^{(i)}_{0:T-1} \sim \mathcal{N}(\mu^{(m)}, \Sigma^{(m)})$ are sampled, rolled out and rescored, and the indices $\mathcal{E}^{(m)}$ of the top-K elites update the sampling distribution by

$$\mu^{(m+1)} = \tfrac{1}{K}\sum_{i\in\mathcal{E}^{(m)}} a^{(i)}_{0:T-1}, \quad \Sigma^{(m+1)} = \tfrac{1}{K}\sum_{i\in\mathcal{E}^{(m)}} \big(a^{(i)}_{0:T-1} - \mu^{(m+1)}\big)\big(a^{(i)}_{0:T-1} - \mu^{(m+1)}\big)^\top. \tag{6}$$

The inference-time planning value is accumulated over the full rollout:

$$\mathcal{V}(S_0, a_{0:T-1}) = \sum_{t=1}^{T}\big[\beta_{\text{vis}}\widehat{H}^{\text{vis}}_t - \beta_{\text{dist}}\big|\,\widehat{H}^{\text{area}}_t - \alpha\,\big|\big] - \lambda_{\text{valid}}\sum_{t=0}^{T-1}\mathbb{I}(a_t \notin \mathcal{A}_{\text{valid}}) - \lambda_{\text{safe}}\sum_{t=1}^{T}\mathbb{I}(\widehat{S}_t \notin \mathcal{O}_{\text{safe}}). \tag{7}$$

The predicted components $\widehat{H}^{\text{vis}}_t$ and $\widehat{H}^{\text{area}}_t$ are read directly from the rolled-out target-evidence vector $\widehat{H}_t$, so the planner does not need to decode a full image at every horizon: $\widehat{H}^{\text{vis}}_t$ captures whether the target is observed and $\widehat{H}^{\text{area}}_t$ acts as a scale-based proxy for distance regulation. We use $\beta_{\text{vis}} = \beta_{\text{dist}} = \lambda_{\text{valid}} = \lambda_{\text{safe}} = 1.0$ and reference scale $\alpha = 0.5$. The admissible action set enforces platform velocity and smoothness limits,

$$\mathcal{A}_{\text{valid}} = \big\{\, a_t = [v_t, \omega_t] \;\big|\; \|v_t\|_\infty \le v_{\max},\ |\omega_t| \le \omega_{\max},\ \|a_t - a_{t-1}\|_\infty \le \delta_{\max} \,\big\}, \tag{8}$$

and the safe rollout set is defined on predicted states through a minimum clearance proxy implied by the rollout, $\mathcal{O}_{\text{safe}} = \{\, \widehat{S}_t \mid d_{\text{clr}}(\widehat{S}_t) \geq d_{\text{safe}} \,\}$ with $d_{\text{safe}} = 0.20\,\text{m}$, and the motion limits are matched to the simulator bounds used by all baselines. The planning problem $a^{\star}_{0:T-1} = \arg\max_{a_{0:T-1}} \mathbb{E}[\mathcal{V}(S_0, a_{0:T-1})]$ is solved with CEM, and the first action of the best sequence is executed before replanning in receding-horizon form. Tables 1 and 2 summarize the training and planning procedures end to end.

Table 1: Training procedure for CST-WM.

Input: transition dataset $(O_\ell, a_\ell, O_{\ell+1})$, model parameters Θ, diffusion scheduler, optimizer.
Output: trained world model $\Theta^{\star}$.
1. Encode O_ℓ with the visual encoder to obtain the observation latent.
2. Construct the target-evidence branch from the detector score and normalized target area.
3. Construct the robot branch from the robot state and the action.
4. Form the structured state at time ℓ and the corresponding next state at time $\ell + 1$.
5. Sample a diffusion step and corrupt the next structured state with Gaussian noise.
6. Predict the diffusion noise with the masked structured transition network.
7. Compute the denoising objective and, when enabled, the auxiliary distance-aware objective.
8. Update model parameters with AdamW.
9. Repeat until convergence and keep the validation-best checkpoint.

Table 2: MPC planning procedure for CST-WM with the Cross-Entropy Method.

Input: current observation O_0, current latent state S_0, planning hyperparameters T, N, K, M.
Output: next control action $a_0^{\star}$.
1. Initialize the Gaussian sampling distribution over action sequences.
2. For each CEM iteration, sample N candidate action sequences.
3. Roll out the structured world model over horizon T for each candidate.
4. Score each candidate with the planning value in Eq. 7.
5. Select the top-K elite sequences and update the sampling distribution by Eq. 6.
6. After the final iteration, choose the best sequence and execute its first action.
7. Replan in receding-horizon form at the next step.

4 Experiments

We evaluate CST-WM from three complementary perspectives: (i) embodied visual tracking when the target is initially visible; (ii) robustness under temporary target loss from occlusion, viewpoint change, drift, or distractors, and the ability to re-acquire the target; and (iii) the contribution of the structural design, target-evidence proxy, and training objectives to rollout quality, planning consistency, and downstream control, assessed through controlled diagnostics and ablations.

4.1 Experimental Setting

Benchmarks and tasks. We evaluate on EVT-Bench [28] and Habitat 3.0 [25]. In both, the robot receives only egocentric observations and platform actions at test time; no privileged humanoid state, future trajectory, or oracle distance is available. EVT-Bench targets embodied visual tracking in dynamic indoor scenes, with the robot initialized near the target in a standard following configuration. Habitat 3.0 introduces a more demanding humanoid–robot setting with richer scene geometry, stronger embodiment constraints, and more frequent occlusion and reappearance events. On Habitat 3.0 we evaluate two protocols: (i) standard tracking from initially visible targets, and (ii) a target-loss protocol in which the target becomes temporarily unavailable due to short occlusion, long occlusion, out-of-field-of-view drift, or distractor crossing. Target-loss episodes are produced directly by simulator rollout rather than by post hoc frame removal. To probe generalization, we also report cross-dataset transfer by training on EVT-Bench and evaluating directly on Habitat 3.0 under the same tracking protocol.

Data construction and splits. For each benchmark, we construct one-step transition tuples $(O_\ell, a_\ell, O_{\ell+1})$ with simulator-side metadata used only for offline analysis and auxiliary supervision. Observations and actions are sampled at the simulator control frequency and only temporally consecutive pairs are retained. We filter out corrupted frames, invalid robot states, reset fragments, terminal-only fragments without effective motion, and transitions violating the platform action con-

straints. Train, validation, and test splits are scene-disjoint, and every learned model is trained with three random seeds with mean results reported.

Metrics. On EVT-Bench we report Success Rate (SR), Tracking Rate (TR), and Collision Rate (CR), which measure task completion, tracking continuity, and execution safety. On Habitat 3.0 standard tracking we report Following Rate (F), Distance-Range Success (DRS), Collision Rate (CR), and Episode Success (ES): F measures the temporal consistency of target following, DRS the fraction of time spent within the desired following interval, CR safety, and ES whether the robot completes the episode with valid following behavior. For target-loss recovery we report Re-acquisition Success (Re-acq.), Time to Re-acquire (TTR), Post-Reacquisition Following Rate (Post-F), and Recovery Episode Success (Rec-ES). Model quality is additionally assessed offline by two planning-oriented diagnostics, whose protocols are given in Section 4.4: *multi-step rollout fidelity*, measured by latent prediction error and target-visibility AUROC against simulator futures, and *planning-value consistency*, measured by Spearman's ρ and Kendall's τ against the simulator-induced ranking of candidate action sequences.

Baselines. On EVT-Bench, we compare against Uni-NaVid [36], TrackVLA [28], and TrackVLA++ (single view) [19]. On Habitat 3.0, we compare against the Habitat 3.0 baseline [25], SDA-S2 [26] where applicable, and an adapted NWM [2]. Our primary comparison is structural: we retrain NWM on the same trajectory data and replace its original selection objective with the planning value in Eq. 7. Since the tracking setting provides no goal image, both the adapted NWM and our method use the same GroundingDINO-based target-evidence signal and share the observation interface, planning horizon, candidate budget, and action bounds at inference time. The only remaining difference is the latent transition design: the adapted NWM and the TrackVLA-style predictor use non-factorized transitions without action masking, whereas our method uses the proposed factorized transition with masked target-evidence modeling. This matched protocol isolates the contribution of the structural design.

Implementation protocol. Unless otherwise specified, all compared models are trained on the same benchmark-specific transition data under the same scene-disjoint split. For controlled architecture ablations, the diffusion backbone, training schedule, and inference-time planner are held fixed so that performance differences can be attributed to the transition structure and its objectives. The main model uses a pre-trained visual VAE encoder, a structured transition transformer with shared latent denoising over the three branches, and DDPM training with DDIM inference. Checkpoints are selected on the validation split by the mean of F and DRS, with ties (equal to three decimal places) broken by lower CR. In cross-dataset transfer the training domain is fixed and only the evaluation environment changes; in offline analyses all methods share the same starting states and candidate action sequences. Dataset scale, architecture, optimization, planner, and compute configuration are collected in Table 3, and $\mathcal{A}_{\text{valid}}$ and $\mathcal{O}_{\text{safe}}$ follow Section 3.3.

4.2 Quantitative Comparison

Table 4 reports the main quantitative results in three settings: in-domain evaluation on EVT-Bench, in-domain evaluation on Habitat 3.0 under initially visible target episodes, and cross-dataset transfer from EVT-Bench to Habitat 3.0. On EVT-Bench, our method obtains the best SR, TR, and CR, surpassing TrackVLA++ and the adapted NWM. On Habitat 3.0, the gain is consistent across F, DRS, and ES, indicating that the learned representation supports both stable following and reliable distance control. Under cross-dataset transfer, our method remains superior to all transferable baselines, suggesting that the structured dynamics generalize beyond a single benchmark.

4.3 Target Loss and Re-acquisition Analysis

Temporary target loss is the main long-horizon challenge in practical embodied tracking. Such failures arise from short occlusion, long occlusion, out-of-field-of-view drift, and distractor crossing. We therefore evaluate recovery separately from standard tracking. The results are reported in Table 5. Our method consistently performs best in this setting. It improves re-acquisition success, reduces recovery time, and yields stronger post-reacquisition following quality. The advantage is particularly clear under long occlusion and distractor crossing, where effective recovery requires stronger long-horizon reasoning and better separation between target evidence and action-conditioned observation change. Figure 3 shows qualitative recovery episodes on EVT-Bench.

4.4 Offline World-Model Evaluation

4.4.1 Multi-Step Rollout Fidelity

Online tracking metrics do not directly measure the quality of long-horizon model rollout. We therefore evaluate offline rollout fidelity under shared starting states and shared future action se-

Table 3: Implementation details: dataset scale, architecture, optimization, planner, and compute.

Item	Details
Data	EVT-Bench / Habitat 3.0; scene-disjoint splits. Train scenes: 64 / 72; val: 8 / 9; test: 8 / 9. Trajectories: 48,000 / 61,000. One-step clips: 1.92M / 2.44M. Filtering removes corrupted frames, invalid robot states, reset fragments, terminal-only collision fragments, and action-invalid transitions.
Observation/action	Observations sampled at simulator control frequency; consecutive frames without temporal skipping; action stream at the simulator control rate.
Model	VAE [5] encoder, latent dim 256; target-evidence dim 32; robot-state dim 32; transition backbone: 6 transformer blocks, hidden 512, 8 heads; action embedding 128; token order: target-evidence, robot, observation.
Diffusion training	Linear DDPM, 1000 train steps, 20 DDIM inference steps; AdamW, lr 10^{-4}, wd 10^{-4}, batch 64, 80 epochs, grad-clip 1.0; 3 seeds; auxiliary distance-aware loss weight $\lambda_{\text{dist}} = 0.2$.
Planning	T=10, N=128, K=16, M=4; β_{vis}=β_{dist}=λ_{valid}=λ_{safe}=1.0; α=0.5; d_{safe}=0.20 m; one stochastic rollout per candidate.
Compute	GPUs: 4×NVIDIA RTX 4090 (24 GB each). CPU: AMD EPYC 7763 64-core, 128 threads; system memory: 256 GB DDR4. Storage: 8 TB NVMe SSD. Total project compute: ∼320 GPU-hours for the main models reported in the paper, plus an additional ∼190 GPU-hours of preliminary, ablation, and failed runs not included in the tables. Average wall-clock training time per main model: 19.5 h (DDPM, 80 epochs). Inference: ∼150 ms per planning step at T=10, N=128, K=16, M=4 (batched on 1×RTX 4090). Mixed-precision training (fp16) with deterministic evaluation seeds; PyTorch 2.2 / CUDA 12.1. Random seeds per main model: 3. Evaluation episodes per model: 500 starting states × 32 candidate sequences for offline diagnostics, plus standard EVT-Bench / Habitat 3.0 splits for downstream metrics.

Table 4: Main quantitative comparison. First block: in-domain EVT-Bench. Second block: standard tracking on Habitat 3.0 with initially visible targets. Third block: cross-dataset transfer by training on EVT-Bench and testing on Habitat 3.0. Best in **bold**, second in underline.

EVT-Bench				Habitat 3.0, standard tracking					Cross-dataset transfer to Habitat 3.0				
Method	SR↑	TR↑	CR↓	Method	F↑	DRS↑	CR↓	ES↑	Method	F↑	DRS↑	CR↓	ES↑
Uni-NaVid [36]	25.7	39.5	41.9	Habitat 3.0 [25]	0.29	0.47	0.48	0.40	Uni-NaVid [36]	0.40	0.56	0.39	0.45
TrackVLA [28]	85.1	78.6	1.65	Adapted NWM [2]	0.41	0.61	0.33	0.49	TrackVLA [28]	0.38	0.54	0.41	0.43
TrackVLA++ [19]	86.0	81.0	2.10	SDA-S2 [26]	0.39	0.63	0.57	0.43	Adapted NWM [2]	0.43	0.58	0.35	0.47
Ours	**88.7**	**83.4**	**1.41**	Ours	**0.53**	**0.70**	**0.27**	**0.61**	Ours	**0.48**	**0.65**	**0.30**	**0.54**

quences. Specifically, we use 500 starting states and 32 shared action sequences per state, and compare model-predicted futures with simulator futures at horizons $\{1, 5, 10, 20, 40\}$.

As shown in Figure 5, our method achieves lower latent prediction error and higher target-visibility AUROC across all horizons. The gap between methods widens with horizon length, indicating that the structured transition accumulates less error over long rollouts. Since planning is performed by ranking predicted futures, improved rollout fidelity directly benefits downstream decision making, especially when the target becomes temporarily unobservable.

4.4.2 Planning-Value Consistency

A world model may produce plausible predictions while still failing to preserve the relative ordering of candidate action sequences. Since our controller relies on comparative planning, we evaluate the agreement between model-based ranking and simulator-based ranking. For each starting state, we score the same shared action candidates using the inference-time reward in Eq. 7, and compare the resulting ranking with the ranking induced by simulator outcomes. Table 6 reports Spearman's ρ and Kendall's τ. Our method achieves the highest rank correlation, indicating that the proposed structured dynamics improve comparative evaluation of future behaviors. This property is particularly important during recovery, where the planner must distinguish between futures that restore target tracking and futures that continue to drift.

4.5 Structural and Invariance Analysis

We test whether instantaneous action information improperly enters the target-evidence pathway through three diagnostics.

Jacobian-based leakage. Table 7 reports the distributional leakage statistics over scenes, episodes, and time steps. Table 8 complements it with the per-setting summary $|\partial H_{\ell+1}/\partial a_\ell|$, averaged over

Table 5: Tracking recovery after temporary target loss on EVT-Bench. Re-acq.: Re-acquisition Success; TTR: Time to Re-acquire (steps); Post-F: Post-Reacquisition Following Rate; Rec-ES: Recovery Episode Success. Best in **bold**, second in underline.

	Short occlusion		Long occlusion		Out-of-FOV loss		Distractor crossing	
Method	Re-acq.↑	TTR↓	Re-acq.↑	TTR↓	Re-acq.↑	TTR↓	Re-acq.↑	TTR↓
TrackVLA [28]	0.71	8.9	0.48	14.6	0.52	12.8	0.43	15.2
Adapted NWM [2]	0.75	8.1	0.54	13.2	0.57	11.7	0.49	14.0
Ours	**0.84**	**6.4**	**0.69**	**9.8**	**0.73**	**8.7**	**0.65**	**10.6**
Method	Post-F↑	Rec-ES↑	Post-F↑	Rec-ES↑	Post-F↑	Rec-ES↑	Post-F↑	Rec-ES↑
TrackVLA [28]	0.58	0.47	0.41	0.29	0.45	0.33	0.37	0.24
Adapted NWM [2]	0.62	0.50	0.45	0.34	0.49	0.37	0.41	0.29
Ours	**0.71**	**0.61**	**0.57**	**0.47**	**0.60**	**0.50**	**0.54**	**0.43**

Figure 3: Qualitative recovery episodes after temporary target loss on EVT-Bench.

Figure 4: GroundingDINO-based target-evidence score vs. relative robot–human distance on EVT-Bench.

rollout states, for four design variants (A: no masking, single-branch baseline; B: shallow masking; C: partial masking; D: deeper masking). Leakage falls as masking deepens, 0.09 (C) $\rightarrow$ 0.04 (D) $\rightarrow$ 0.00 (ours), and full architectural masking drives the term to zero; shallow masking (B, 0.37) is worse than the unmasked baseline (A, 0.19), indicating that a partially blocked pathway is not by itself sufficient.

Interventional invariance under action swap. Starting from the same latent state, we replace a_ℓ with alternative actions sampled from $\mathcal{A}_{\text{valid}}$ while holding other factors fixed, and measure the variance of the predicted target-evidence together with the pairwise 1D Wasserstein distance between the resulting predictive distributions. Table 10 shows $0.006/0.011$ for our model against $0.084/0.127$ for the adapted NWM and $0.146/0.214$ for a leaky baseline, confirming that the target-evidence pathway is not driven by direct action injection.

Controlled robot-motion-only check. With the humanoid held stationary while the robot executes varied motion (Table 9), target-evidence stability (TE-stab., the inverse of variation in $\widehat{H}$ under fixed humanoid state) and observation sensitivity (Obs-sens., normalized variation in $\widehat{Z}$ under viewpoint change) reach $0.98/0.84$ for our method versus $0.61/0.78$ for the adapted NWM. The pathway behaviour is therefore selective rather than an indiscriminate insensitivity to action.

Together these diagnostics support that the proposed factorization suppresses spurious action leakage while preserving the action sensitivity required in the observation pathway.

4.6 Real-world Tracking

Since the proposed method combines model rollout with trajectory optimization, we characterize the tradeoff between performance and computation across planning configurations. As expected, increasing planning horizon and candidate count improves recovery but raises per-step latency. Figure 7 visualizes this tradeoff: the default configuration captures most of the performance gain while avoiding the substantial cost of the largest setting, occupying a favorable position on the Pareto front. Figure 6 shows qualitative real-world tracking in different environments; a video of the same deployment is available on the project page.

4.7 Target-Evidence Proxy and Distance-Aware Validation

Figure 4 shows the relation between the GroundingDINO-based target-evidence score and relative robot–human distance: the score increases monotonically over most of the trajectory with mild saturation only at very close range, consistent with its intended use as a target observability-distance proxy for planning rather than as a metric-distance estimator. Table 11 compares four variants trained

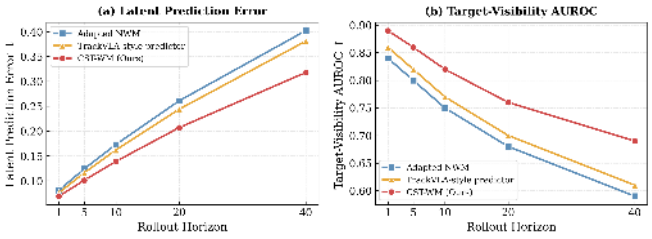

Figure 5: Offline rollout fidelity vs. planning horizon. (a) Latent prediction error (lower is better). (b) Target-visibility AUROC (higher is better). The gap widens at longer horizons.

Table 6: Planning-value consistency: rank correlation with simulator-induced ranking. Best in **bold**, second in underline.

Method	$\rho\uparrow$	$\tau\uparrow$
Adapted NWM	0.47	0.33
TrackVLA-style	0.52	0.37
Ours	**0.68**	**0.51**

Table 7: Jacobian leakage distribution. Lower values indicate better suppression of direct action leakage into the target-evidence branch. Best in **bold**, second in underline (leakage columns only).

Method	All states↓	In-view↓	Out-of-view↓	Leakage–performance corr.
Adapted NWM	0.112	0.096	0.131	−0.41
Leaky variant B	0.331	0.304	0.357	−0.53
Leaky variant C	0.086	0.071	0.102	−0.36
Ours	**0.001**	**0.001**	**0.002**	−0.04

Table 8: Jacobian leakage summary. Lower values indicate weaker direct action leakage into the target-evidence branch.

Setting	A	B	C	D	Ours
$\lvert\partial H_{\ell+1}/\partial a_\ell\rvert$	0.19	0.37	0.09	0.04	**0.00**

Table 9: Robot-motion-only check on Habitat 3.0.

Method	TE-stab.↑	Obs-sens.↑
Adapted NWM	0.61	0.78
Ours	**0.98**	**0.84**

Table 10: Interventional analysis under action swap. Lower action-induced variance and lower pairwise distance indicate weaker direct action leakage. Best in **bold**, second in underline.

Method	Variance across actions↓	Pairwise Wasserstein distance↓	Interpretation
Adapted NWM	0.084	0.127	noticeable action dependence
Leaky baseline	0.146	0.214	strong branch drift
Ours	**0.006**	**0.011**	stable under intervention

under the same protocol – detector score only, the proxy without distance-aware loss, a direct-distance oracle from the simulator, and the full model: the full model substantially improves over weaker proxy variants and approaches the oracle-based result, supporting the claim that the proxy and distance-aware objective provide practically useful distance-related evidence for planning-oriented tracking without requiring an additional distance detector at test time. Table 12 extends this validation across humanoids that differ in body scale, height, appearance, clothing, and motion style. Adding the distance-aware loss raises the monotonic Spearman correlation between proxy score and true relative distance from 0.71 to 0.83 and improves all four downstream metrics, supporting practical robustness of the proxy across humanoids rather than full identity-invariant geometric recovery.

4.8 Ablation Study

We organize the ablation study around four questions: whether action masking is necessary for suppressing leakage into the target-evidence branch, whether branch factorization is necessary for planning-oriented tracking, whether the distance-aware objective is necessary for stable distance control, and whether stochastic rollout is useful for recovery after temporary loss. Figure 8 visualizes the main ablation results on Habitat 3.0 as a radar chart and Table 13 gives the condition-wise quantitative breakdown: removing action masking causes the largest degradation in re-acquisition and worsens the leakage diagnostic; the single-branch alternative reduces F and planning-value consistency; removing the distance-aware objective mainly harms DRS; the deterministic predictor weakens long-horizon recovery. Table 14 reports additional architectural variants (partial masking, shuffled update order, deterministic predictor). Table 15 reports per-seed stability of the full model under three independent random initializations: F, DRS, and CR vary by at most 0.02, much smaller than the gap to the strongest baseline in Table 4, so the reported gain is not seed-specific.

Figure 6: Qualitative results for real-world tracking in different environments.

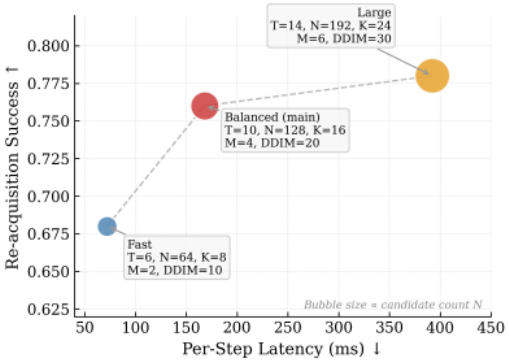

Figure 7: Performance–latency Pareto tradeoff for three planning configurations. Horizontal axis: per-step latency (ms); vertical axis: Re-acquisition Success. Bubble size is proportional to candidate count N.

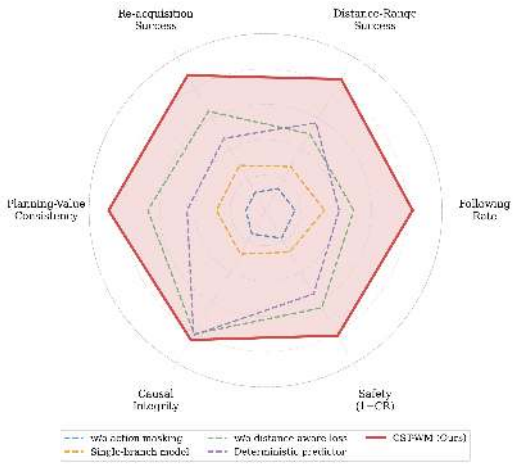

Figure 8: Radar chart of the ablation study on Habitat 3.0. Each axis represents a normalized evaluation metric (higher is better; Leak and CR are inverted). The full model (red, shaded) consistently encloses the ablated variants.

5 Limitations

Detector dependence. The target-evidence branch is constructed from GroundingDINO [21] confidence and normalized box area. When detection becomes unstable – under heavy motion blur, low light, extreme viewpoint, or out-of-distribution appearance – the planning signal can also become unstable. The proposed factorization does not, by itself, recover from a fundamentally unreliable target detector, and a stronger or task-tuned detector would likely improve performance further.

Proxy is not metric distance. The apparent-area cue H_ℓ^{area} is an observability-distance proxy rather than a geometrically calibrated distance estimate. It is most informative within the typical 1–3 m following range used in our protocol; at very long range it becomes near-saturated near zero, and at extreme close range it saturates from above. Distance regulation outside the calibrated range is therefore not guaranteed.

Simulation-to-real gap. The main quantitative conclusions are derived from EVT-Bench [28] and Habitat 3.0 [25]. Although Section 4.6 reports qualitative real-world tracking, we do not report large-scale physical-robot statistics. Detector behavior, motion statistics, lighting, and recovery

Table 11: Validation of the target-evidence proxy and the distance-aware objective on Habitat 3.0. DRS: Distance-Range Success; Re-acq.: Re-acquisition Success after temporary loss.

Training signal and scoring design	F↑	DRS↑	Re-acq.↑	CR↓
Detector score only	0.46	0.60	0.66	0.32
Current proxy without distance-aware loss	0.49	0.65	0.72	0.29
Direct-distance oracle	0.54	0.72	0.78	0.26
Current proxy with distance-aware loss	0.53 (↓0.01)	0.70 (↓0.02)	0.76 (↓0.01)	0.27 (↑0.01)

Table 12: Cross-humanoid distance-aware tracking validation on Habitat 3.0. Mono. ρ: Spearman correlation between proxy score and true relative distance trend after sign inversion. Best in **bold**.

Method	Mono. ρ↑	F↑	DRS↑	Re-acq.↑	CR↓
Current proxy without distance-aware loss	0.71	0.47	0.62	0.68	0.31
Current proxy with distance-aware loss	**0.83**	**0.52**	**0.69**	**0.75**	**0.28**

Table 13: Condition-wise ablation breakdown on Habitat 3.0. Best in **bold**, second in underline.

Variant	F↑	DRS↑	Re-acq.↑	Plan. ρ ↑
W/o masking	0.47	0.63	0.61	0.55
W/o factorization	0.49	0.66	0.64	0.58
W/o dist.-aware loss	0.49	0.65	0.72	0.64
Deterministic rollout	0.50	0.67	0.68	0.60
Full model	**0.53**	**0.70**	**0.76**	**0.68**

Table 14: Additional architecture ablations on Habitat 3.0.

Method	F↑	Re-acq.↑	PV↑	CR↓
Partial masking	0.46	0.68	0.62	0.30
Shuffled order	0.46	0.65	0.58	0.32
Deterministic	0.48	0.69	0.60	0.30
Ours	**0.53**	**0.76**	**0.68**	**0.27**

Table 15: Per-seed breakdown of the full model on Habitat 3.0 standard tracking.

Seed	F↑	DRS↑	CR↓
1	0.52	0.69	0.28
2	**0.54**	**0.71**	**0.26**
3	0.53	0.70	0.27

patterns can shift in physical deployment, and the present evidence does not establish robustness under uncontrolled real-world distribution shift.

Multi-target identity ambiguity. The target-evidence branch summarizes a single "the target is observable" signal and does not explicitly maintain persistent target identity under prolonged distractor ambiguity. Under crowded scenes with multiple visually similar humans, the controller may occasionally lock onto the wrong individual; identity-aware re-acquisition is left to future work.

Privileged signals at training time only. The auxiliary distance-aware loss $\mathcal{L}_{\text{dist}}$ uses simulator-only relative distance for offline supervision. While no privileged signal is used at test time, the training-time signal is unavailable in real-data-only deployments; the calibration evidence in Table 11 should be read as a lower bound on the achievable proxy quality without privileged supervision.

Compute and scaling. Training a single CST-WM main model takes ∼19.5 wall-clock hours on 4×RTX 4090, and the full reported set used ∼320 GPU-hours plus an additional ∼190 GPU-hours of preliminary and ablation runs. The diffusion-based rollout is also more expensive at inference time than reactive policies (Figure 7). Practitioners with strict latency budgets should expect either to use the "Fast" configuration or to distill the rollout.

Scope of structural claims. The non-leakage constraint $\partial \widehat{H}_{\ell+1}/\partial a_\ell = 0$ is a structural requirement on the planning representation, not a strong physical claim about the world. The empirical gains support that the architectural choice is beneficial under the tested settings; they do not, by themselves, establish a stronger claim about full causal recovery of external humanoid dynamics.

6 Conclusion

We presented CST-WM, a causally structured world model for embodied visual tracking. CST-WM decomposes the latent state into target-evidence, robot, and observation branches and factorizes the transition so that direct action injection into the target-evidence branch is blocked while action remains available to motion and observation updates, directly targeting the task-specific causal hallucination in action-conditioned prediction. Combined with rollout-based MPC, CST-WM supports stable

following and target re-acquisition within a single planning framework. On EVT-Bench and Habitat 3.0 it improves following quality, distance-range control, safety, and re-acquisition, while offline diagnostics show better rollout fidelity, stronger planning-value consistency, and substantially reduced direct action leakage. These results suggest that, for embodied visual tracking, the predictive structure itself must align with how target evidence enters planning.

References

[1] Bajcsy, A., Loquercio, A., Kumar, A., Malik, J.: Learning vision-based pursuit-evasion robot policies. In: 2024 IEEE International Conference on Robotics and Automation (ICRA). pp. 9197–9204. IEEE (2024)

[2] Bar, A., Zhou, G., Tran, D., Darrell, T., LeCun, Y.: Navigation world models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15791–15801 (2025)

[3] Bayraktar, E.: Retrackvlm: Transformer-enhanced multi-object tracking with cross-modal embeddings and zero-shot re-identification integration. Applied Sciences **15**(4) (2025)

[4] Black, K., Brown, N., Driess, D., Esmail, A., Equi, M., Finn, C., Fusai, N., Groom, L., Hausman, K., Ichter, B., et al.: π_0: A vision-language-action flow model for general robot control. arXiv preprint arXiv:2410.24164 (2024)

[5] Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023)

[6] de Boer, P.T., Kroese, D.P., Mannor, S., Rubinstein, R.Y.: A tutorial on the cross-entropy method. Annals of Operations Research **134**(1), 19–67 (2005)

[7] Bruce, J., Dennis, M., Edwards, A., Parker-Holder, J., Shi, Y., Hughes, E., Lai, M., Mavalankar, A., Steigerwald, R., Apps, C., et al.: Genie: Generative interactive environments. arXiv preprint arXiv:2402.15391 (2024)

[8] Devo, A., Dionigi, A., Costante, G.: Enhancing continuous control of mobile robots for end-to-end visual active tracking. Robotics and Autonomous Systems **142**, 103799 (2021)

[9] Francis, A., Pérez-d'Arpino, C., Li, C., Xia, F., Alahi, A., Alami, R., Bera, A., Biswas, A., Biswas, J., Chandra, R., Chiang, H.T.L., Everett, M., Ha, S., Hart, J., Karnan, H., Lee, T.W.E., Manso, L.J., Mirksy, R., Pirk, S., Singamaneni, P.T., Stone, P., Taylor, A.V., Trautman, P., Tsoi, N., Vázquez, M., Xiao, X., Xu, P., Yokoyama, N., Toshev, A., Martín-Martín, R.: Principles and guidelines for evaluating social robot navigation algorithms. ACM Transactions on Human-Robot Interaction **14**(2), 1–65 (2025)

[10] Ha, D., Schmidhuber, J.: Recurrent world models facilitate policy evolution. In: Advances in Neural Information Processing Systems (NeurIPS) (2018)

[11] Hafner, D., Lillicrap, T., Ba, J., Norouzi, M.: Dream to control: Learning behaviors by latent imagination. In: International Conference on Learning Representations (ICLR) (2020)

[12] Hafner, D., Lillicrap, T., Fischer, I., Villegas, R., Ha, D., Lee, H., Davidson, J.: Learning latent dynamics for planning from pixels. In: International Conference on Machine Learning (ICML) (2019)

[13] Hafner, D., Lillicrap, T., Norouzi, M., Ba, J.: Mastering atari with discrete world models. arXiv preprint arXiv:2010.02193 (2020)

[14] Hafner, D., Pasukonis, J., Ba, J., Lillicrap, T.: Mastering diverse domains through world models. arXiv preprint arXiv:2301.04104 (2023)

[15] Hu, A., Russell, L., Yeo, H., Murez, Z., Fedoseev, G., Kendall, A., Shotton, J., Corrado, G.: Gaia-1: A generative world model for autonomous driving. arXiv preprint arXiv:2309.17080 (2023)

[16] Kim, M.J., Pertsch, K., Karamcheti, S., Xiao, T., Balakrishna, A., Nair, S., Rafailov, R., Foster, E., Lam, G., Sanketi, P., Vuong, Q., Kollar, T., Burchfiel, B., Tedrake, R., Sadigh, D., Levine, S., Liang, P., Finn, C.: Openvla: An open-source vision-language-action model. arXiv preprint arXiv:2406.09246 (2024)

[17] Li, J., Xu, J., Zhong, F., Kong, X., Qiao, Y., Wang, Y.: Pose-assisted multi-camera collaboration for active object tracking. Proceedings of the AAAI Conference on Artificial Intelligence **34**(01), 759–766 (2020)

[18] Li, X., Wu, C., Yang, Z., Xu, Z., Liang, D., Zhang, Y., Wan, J., Wang, J.: Driverse: Navigation world model for driving simulation via multimodal trajectory prompting and motion alignment. arXiv preprint arXiv:2504.18576 (2025), baidu Inc.

[19] Liu, J., Qi, Y., Zhang, J., Li, M., Wang, S., Wu, K., Ye, H., Zhang, H., Chen, Z., Zhong, F., Zhang, Z., Wang, H.: Trackvla++: Unleashing reasoning and memory capabilities in vla models for embodied visual tracking. arXiv preprint arXiv:2510.07134 (2025)

[20] Liu, J., Ma, Y., Hao, J., Hu, Y., Zheng, Y., Lv, T., Fan, C.: A trajectory perspective on the role of data sampling techniques in offline reinforcement learning. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS '24). IFAAMAS, Auckland, New Zealand (2024)

[21] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

[22] Liu, X., Zou, Z., Hao, J.: Adaptive text feature updating for visual-language tracking. In: Pattern Recognition. pp. 366–381. Springer Nature Switzerland, Cham (2025)

[23] Luo, W., Sun, P., Zhong, F., Liu, W., Zhang, T., Wang, Y.: End-to-end active object tracking via reinforcement learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML). pp. 3286–3295 (2018)

[24] Maalouf, A., Jadhav, N., Jatavallabhula, K.M., Chahine, M., Vogt, D.M., Wood, R.J., Torralba, A., Rus, D.: Follow anything: Open-set detection, tracking, and following in real-time. IEEE Robotics and Automation Letters **9**(4), 3283–3290 (2024)

[25] Puig, X., Undersander, E., Szot, A., Dallaire Cote, M., Yang, T.Y., Partsey, R., Desai, R., Clegg, A.W., Hlavac, M., Min, S.Y., Vondruš, V., Gervet, T., Berges, V.P., Turner, J.M., Maksymets, O., Kira, Z., Kalakrishnan, M., Malik, J., Chaplot, D.S., Jain, U., Batra, D., Rai, A., Mottaghi, R.: Habitat 3.0: A co-habitat for humans, avatars and robots. arXiv preprint arXiv:2310.13724 (2023)

[26] Scofano, L., Sampieri, A., Campari, T., Sacco, V., Spinelli, I., Ballan, L., Galasso, F.: Following the human thread in social navigation. In: International Conference on Learning Representations (ICLR) (2025)

[27] Shah, D., Bhorkar, A., Leen, H., Kostrikov, I., Rhinehart, N., Levine, S.: Offline reinforcement learning for visual navigation. In: 6th Annual Conference on Robot Learning (2022)

[28] Wang, S., Zhang, J., Li, M., Liu, J., Li, A., Wu, K., Zhong, F., Yu, J., Zhang, Z., Wang, H.: Trackvla: Embodied visual tracking in the wild. arXiv preprint arXiv:2505.23189 (2025)

[29] Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint arXiv:2309.09777 (2023)

[30] Wang, Y., He, J., Fan, L., Li, H., Chen, Y., Zhang, Z.: Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)

[31] Wu, K., Xu, S., Chen, H., Wang, C., Li, Z., Wang, Y., Zhong, F.: Vlm can be a good assistant: Enhancing embodied visual tracking with self-improving vision-language models (2025)

[32] Wu, P., Escontrela, A., Hafner, D., Goldberg, K., Abbeel, P.: Daydreamer: World models for physical robot learning. In: Proceedings of the 6th Conference on Robot Learning (CoRL) (2023)

[33] Ye, H., Cai, K., Zhan, Y., Xia, B., Ajoudani, A., Zhang, H.: Rpf-search: Field-based search for robot person following in unknown dynamic environments. arXiv preprint arXiv:2503.02188 (2025)

[34] Ye, H., Zhao, J., Zhan, Y., Chen, W., He, L., Zhang, H.: Person re-identification for robot person following with online continual learning. IEEE Robotics and Automation Letters (2024)

[35] Zeng, K.H., Zhang, Z., Ehsani, K., Hendrix, R., Salvador, J., Herrasti, A., Girshick, R., Kembhavi, A., Weihs, L.: Poliformer: Scaling on-policy rl with transformers results in masterful navigators. CoRL (2024)

[36] Zhang, J., Wang, K., Wang, S., Li, M., Liu, H., Wei, S., Wang, Z., Zhang, Z., Wang, H.: Uni-navid: A video-based vision-language-action model for unifying embodied navigation tasks. arXiv preprint arXiv:2412.06224 (2024)

[37] Zhong, F., Sun, P., Luo, W., Yan, T., Wang, Y.: Towards distraction-robust active visual tracking. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 12782–12792. PMLR (2021)

[38] Zhong, F., Bi, X., Zhang, Y., Zhang, W., Wang, Y.: RSPT: Reconstruct surroundings and predict trajectories for generalizable active object tracking. In: Proceedings of the AAAI Conference on Artificial Intelligence (2023)

[39] Zhong, F., Sun, P., Luo, W., Yan, T., Wang, Y.: AD-VAT: An asymmetric dueling mechanism for learning visual active tracking. In: International Conference on Learning Representations (ICLR) (2019)

[40] Zhong, F., Wu, K., Ci, H., Wang, C., Chen, H.: Empowering embodied visual tracking with visual foundation models and offline rl. In: European Conference on Computer Vision (2024)