CAUSALITY-INSPIRED WORLD ACTION MODEL VIA CONTRASTIVE REPRESENTATION LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

World-action models (WAMs) transfer predictive visual priors to robotic manipulation, but robust control requires invariance to task-irrelevant appearance changes that remain relevant to video prediction. Action supervision specifies the desired behavior for each observation without explicitly identifying which visually different observations share the same manipulation state. We introduce **CausalWAM**, a post-training method that supervises state correspondence through *action-preserving visual interventions*. We vary scene appearance while preserving robot–object geometry, task instructions, and demonstrated actions, changing the observation mechanism without changing the underlying manipulation state. A contrastive objective aligns alternative renderings of the same state and distinguishes different states within each task. A shared content reader connects this supervision to action learning, supplying compact tokens alongside the action expert's native visual conditioning. The same design supports interface learning over a frozen video backbone and joint video–action adaptation. With action demonstrations restricted to Clean scenes, CausalWAM-Joint increases RoboTwin 2.0 Randomized success from 1.9% for FastWAM to 34.7%. It achieves 79.4% mean success across seven unseen LIBERO-Plus perturbation types and improves real-robot success under visual disturbance from 15.6% to 46.7%. Across 50 RoboTwin tasks, CausalWAM-Interface improves cross-appearance state retrieval by 9.42 percentage points. These results demonstrate the value of intervention-defined state correspondence for adapting predictive world representations to visually robust manipulation.

1 INTRODUCTION

World-Action Models (WAMs) combine predictive visual modeling with action learning to transfer generative visual priors to robotic manipulation (Bi et al., 2025; Yuan et al., 2026b). Their benefits extend beyond explicit future imagination: FastWAM retains competitive control performance when future video generation is removed at inference time, highlighting the value of representations shaped through predictive training (Yuan et al., 2026b). High standard-benchmark success, however, does not ensure robustness to visual perturbations (Fei et al., 2026; Zhang et al., 2026b). A central challenge is how to turn these representations into ***robust control under visual-domain shift***.

World prediction rewards visual completeness, while action prediction requires selective invariance (Zhang et al., 2021; Rabinovitz et al., 2021). Predicting future pixels encourages representations to retain background, texture, and clutter even when these variations are irrelevant to manipulation (Yuan et al., 2026a). Action supervision specifies the desired action for each observation, but does not explicitly identify which visually different observations share the same manipulation state. Recent methods enrich world representations with motion, geometry, and semantic supervision (Yuan et al., 2026a; Yan et al., 2026a). Such structure specifies additional information to represent; the complementary challenge is to specify *which observations describe the same manipulation state*. Causal representation learning formalizes how known invariances across observations can constrain learned representations (von Kügelgen et al., 2021; Yao et al., 2025). This motivates explicit state-correspondence supervision alongside action learning.

To address this challenge, we introduce **CausalWAM**, a **post-training** method that learns state-discriminative representations through *action-preserving visual interventions*. We construct

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.

matched observations from recorded trajectories by varying room appearance, tabletop texture, and task-irrelevant visual clutter while preserving robot–object geometry, the instruction, and the demonstrated action. These interventions change the observation mechanism while keeping the manipulation state fixed. Different renderings of the same state provide positive pairs, and different physical states within the same task provide negatives. A supervised contrastive objective (Khosla et al., 2020) encourages agreement across appearance changes while retaining distinctions between physical configurations. The supervision is defined by the state preserved under an intervention, rather than by visual similarity.

CausalWAM couples this intervention objective with native action learning through a shared content reader. The reader extracts compact content tokens from the world representation, and a gated adapter supplies them to the action expert alongside its original visual conditioning. We instantiate this design in two variants. **CausalWAM-Interface** trains the content reader, adapter, and action expert over a frozen video backbone. **CausalWAM-Joint** additionally fine-tunes the video backbone, adapting the world representation and its action-facing readout together. Both retain the native world–action pathway and use the same correspondence supervision.

We evaluate CausalWAM on RoboTwin 2.0 (Chen et al., 2025), LIBERO (Liu et al., 2023), LIBERO-Plus (Fei et al., 2026), and three real-robot manipulation tasks. With action demonstrations restricted to Clean scenes, CausalWAM-Joint increases RoboTwin Randomized success from 1.9% for FastWAM to 34.7%. On LIBERO-Plus, it achieves 79.4% mean success across seven unseen perturbation types and 95.7% under background changes, while maintaining 97.8% success on standard LIBERO. On real robots, it raises success under appearance changes and additional distractors from 15.6% for FastWAM to 46.7%. Under Clean+Random training, CausalWAM-Interface achieves 93.62% Random success across 50 RoboTwin tasks with the video backbone frozen. Complementing these control results, CausalWAM-Interface improves mean R@1 for cross-appearance physical-state retrieval from 75.22% to 84.64% across the same 50 tasks. Together, these results demonstrate that intervention-defined state correspondence is an effective supervision signal for adapting predictive world representations to visually robust control through post-training.

Our contributions are threefold. **(1) A causality-based formulation of visual robustness.** We distinguish changes to the observation mechanism from changes to the manipulation state, and use action-preserving interventions to define representation supervision for WAMs. **(2) CausalWAM with interface and joint adaptation.** We introduce a content reader and gated action conditioning trained with contrastive and native action objectives, supporting both frozen-backbone interface learning and joint world–action adaptation. **(3) Cross-domain manipulation and representation evaluation.** Experiments on RoboTwin, LIBERO, LIBERO-Plus, and real robots demonstrate improved control under visual shift, complemented by cross-appearance physical-state retrieval.

2 Related Work

World action models. World action models couple visual prediction with action generation. UVA learns a joint video–action latent with decoupled output heads, while Motus integrates understanding, video, and action experts in a mixture-of-transformers architecture (Li et al., 2025; Bi et al., 2025). Fast-WAM separates the benefit of video–action co-training from explicit future generation at inference, while ImageWAM reuses image-editing priors to model current-to-target transformations (Yuan et al., 2026b; Zhang et al., 2026a). Such designs place the predictive representation, rather than a decoded future video, at the interface to control. Recent work enriches predictive supervision beyond RGB: DreamWAM incorporates motion, geometry, and semantics, while Flex-π jointly models RGB, pointmaps, and semantic features (Yuan et al., 2026a; Yan et al., 2026a). MT-WAM uses point-trajectory and visual-feature targets to shape a one-pass predictive representation (Yang et al., 2026b). Robust-WAM instead aligns action-stream queries with future semantic features, retaining the pretrained video-generative pathway while improving visual robustness (Yan et al., 2026b). These approaches enrich the information available for control through predictive targets or pretrained feature alignment. CausalWAM addresses the complementary problem of defining which appearance changes should leave the represented manipulation state unchanged. Its supervision comes from rendering interventions that preserve physical state and demonstrated actions, making cross-appearance state correspondence explicit. This correspondence trains a content read-

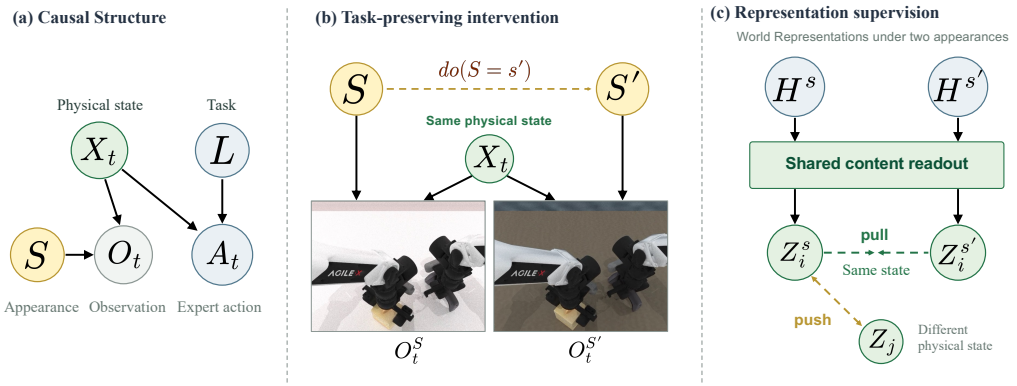


Figure 1: **Causal analysis of the world–action interface.** (a) Appearance affects observations, while physical state and task determine the expert action mechanism. (b) Re-rendering varies appearance at the same physical state. (c) Same-state pairs supervise a shared content readout; other states within the same task provide negatives.

out that conditions the action expert alongside its native visual features, retaining the world model's generative pathway.

Causal representation learning. Causal representation learning Schölkopf et al. (2016); von Kügelgen et al. (2021); Brehmer et al. (2022a); Ahuja et al. (2022); Locatello et al. (2020); van den Oord et al. (2018); Wang & Isola (2020) uses structure in how observations vary to constrain the recovery of latent factors. Intervention-paired observations enable identification of latent causal models under suitable assumptions (Brehmer et al., 2022b), while augmentation-based theory characterizes when invariant content blocks are identifiable up to invertible transformations, including with dependent latent factors (von Kügelgen et al., 2021). Control-oriented work specifies invariance through behavior or visual interventions: bisimulation-based learning defines state similarity through rewards and transitions rather than pixel reconstruction (Zhang et al., 2021), while Contrastive Domain Randomization exploits simulator access to randomize appearance independently across input and target renderings in predictive contrastive learning (Rabinovitz et al., 2021). Causal-WAM specializes this perspective to post-training generative WAMs.

3 Causal Analysis of World Action Models

Appearance matters for predicting images, but an appearance change need not change the action required at a fixed physical state and instruction. Figure 1 connects this distinction to supervision at the world–action interface.

(a) Causal structure. Let X_t be the physical scene state, L the language instruction, S the rendering appearance, O_t the complete multi-camera observation, and A_t the expert action. The state includes robot configuration, object poses, and contact. We consider appearance changes that preserve physical dynamics, task semantics, and the visual evidence needed for control. The corresponding mechanisms are

$$O_t = R(X_t, S), \qquad A_t = F_A(X_t, L, U_t^a), \tag{1}$$

where U_t^a denotes expert-action randomness and camera calibration is fixed. Thus, appearance enters rendering but not the expert action mechanism. This mechanistic distinction holds even when state and appearance are correlated in the training data.

(b) Task-preserving intervention. We implement $\mathrm{do}(S = s')$ by re-rendering recorded physical states with different background and table textures. Each pair preserves the physical state, instruction, proprioception, camera calibration, and recorded action. For fixed $X_t = x$ and $L = l$, the unchanged action mechanism gives

$$p^\star(A_t \mid x, l, \mathrm{do}(S = s)) = p^\star(A_t \mid x, l, \mathrm{do}(S = s')), \tag{2}$$

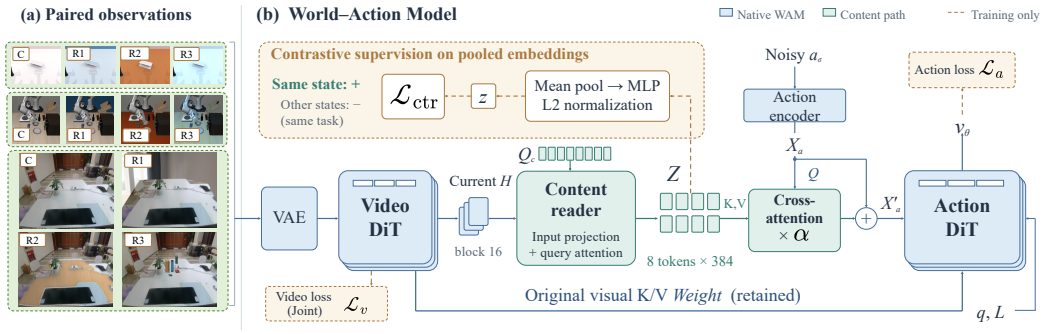


Figure 2: **CausalWAM.** (a) Matched appearances define physical-state correspondences. (b) A shared reader supplies content tokens Z for action conditioning and a normalized readout z for contrastive supervision, alongside the native world–action pathway.

where $p^\star$ is the expert action law. The observations change, but the appropriate action distribution does not. Positive pairs are two renderings of the *same camera tuple*, preserving complementary evidence across views.

(c) Representation supervision. Let H_i^s denote world features for physical state i under appearance s, and Z_i^s the shared content reader's output in Fig. 1(c). Contrastive supervision acts on a normalized projection z_i^s of these tokens: other appearances of state i are positives, and states $j \neq i$ from the same task are negatives. Under the balanced-batch and realizability conditions, the contrastive optimum aligns all renderings of each sampled state while keeping different states distinct. This provides an *appearance-stable, state-discriminative* learning target. We apply this supervision to the content readout while retaining the original world features for visual prediction and native action conditioning. Action supervision trains how the policy uses Z; contrastive supervision specifies which appearance changes its readout should ignore.

4 Method

As shown in Fig. 3, CausalWAM turns the state correspondences in Sec. 3 into an action-facing interface for world representations. In Fig. 2(a), C denotes the original Clean rendering, while $R1$–$R3$ are appearance-randomized renderings of the same recorded physical state, preserving the underlying robot–object configuration and action. A shared content reader connects two learning signals: contrastive supervision matches physical states across appearances, while action supervision determines how its tokens support manipulation. We instantiate this design with a frozen video backbone (**CausalWAM-Interface**) and joint world–action adaptation (**CausalWAM-Joint**).

4.1 Shared content reader

For state i under appearance s, a current-observation-only branch supplies VideoDiT block-16 features $H_i^s \in \mathbb{R}^{N\times D}$ across the camera tuple, conditioned on proprioception q_i and instruction L_i. An input projection and M learned queries extract content tokens:

$$\widetilde{H}_i^s = H_i^s W_{\text{in}} + b_{\text{in}}, \qquad Z_i^s = \text{MHA}_\phi(Q_c, \widetilde{H}_i^s, \widetilde{H}_i^s), \tag{3}$$

where $Q_c \in \mathbb{R}^{M\times d}$ and $Z_i^s \in \mathbb{R}^{M\times d}$. We use $M = 8$, $d = 384$, and eight attention heads. The action expert receives the full array Z_i^s; a two-layer SiLU MLP h_ϕ maps its mean-pooled readout to a unit embedding for contrastive learning:

$$z_i^s = \text{normalize}\Big(h_\phi\Big(M^{-1}\textstyle\sum_{m=1}^{M} Z_{i,m}^s\Big)\Big). \tag{4}$$

Here $\text{normalize}(u) = u/\|u\|_2$, and ϕ collects the reader and projection parameters. Action supervision trains the reader through Z; contrastive supervision trains the same reader through z.

4.2 INTERVENTION-DEFINED CONTRASTIVE SUPERVISION

The paired stream $\mathcal{D}_p$ contains K renderings of each recorded state, varying background and table appearance while preserving the physical state, instruction, proprioception, camera calibration, and timestamp. Each rendering is a complete camera tuple, and copies of a physical trajectory remain in the same data split. We use $K = 4$ in the main RoboTwin Interface configuration. For anchor $a = (i, r)$, positives $P(a)$ are the other appearances of state i; negatives $N(a)$ are appearances of other sampled states within the same task. We use supervised contrastive learning (Khosla et al., 2020):

$$\mathcal{L}_{\text{ctr}} = -\frac{1}{|\mathcal{B}|} \sum_{a \in \mathcal{B}} \frac{1}{|P(a)|} \sum_{p \in P(a)} \log \frac{\exp(z_a^\top z_p/\tau)}{\sum_{j \in P(a) \cup N(a)} \exp(z_a^\top z_j/\tau)}, \qquad \tau = 0.07. \tag{5}$$

The anchor set $\mathcal{B}$ contains samples with both positives and negatives; self-pairs and cross-task samples are excluded from the denominator. Thus, state correspondence rather than task identity defines the discrimination target. Appendix B.2 characterizes the resulting appearance-alignment and state-separation geometry.

4.3 CONTENT-CONDITIONED ACTION GENERATION

Let X_a be the action encoder output for a noisy action chunk. A cross-attention adapter with parameters η supplies content before the ActionDiT stack:

$$X'_a = X_a + \alpha \, \text{MHA}_\eta(X_a, Z, Z), \tag{6}$$

where α is a scalar controlling the contribution of the content residual. In CausalWAM-Interface, we parameterize $\alpha = \tanh(g)$ with a learnable scalar g initialized to zero, so the adapter initially leaves the original action features unchanged. In CausalWAM-Joint, we use a fixed scaling coefficient $\alpha = 0.1$. The attention output matches the action width. Native visual key/value conditioning W remains available throughout the action expert, so Z augments rather than replaces the world–action pathway. This gives action queries access to a readout trained for cross-appearance state correspondence while retaining the original visual evidence.

4.4 TRAINING AND INFERENCE

Original demonstrations $\mathcal{D}_a$ supply native action and video targets; re-rendered appearance variants enter the paired stream $\mathcal{D}_p$ for contrastive supervision. Writing $\mathcal{L}_a$ for the native action loss and $\mathcal{L}_v$ for Joint's video-prediction loss, the objectives are

$$\begin{aligned} \mathcal{L}_{\text{Interface}} &= \mathcal{L}_a(\mathcal{D}_a) + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}(\mathcal{D}_p), \\ \mathcal{L}_{\text{Joint}} &= \mathcal{L}_a(\mathcal{D}_a) + \lambda_v \mathcal{L}_v(\mathcal{D}_a) + \lambda_{\text{ctr}} \mathcal{L}_{\text{ctr}}(\mathcal{D}_p). \end{aligned} \tag{7}$$

We retain the backbone's native noise schedules and loss weighting, with $\lambda_v = 1$ and $\lambda_{\text{ctr}} = 0.5$ in the main configurations.

CausalWAM-Interface. We initialize from a post-trained FastWAM policy. In the RoboTwin setting with Clean+Random demonstrations, this is the released policy for that regime. VideoDiT and the proprioceptive encoder are frozen, while the entire ActionDiT, reader, and adapter are trained. Contrastive gradients update the reader and projection, stopping at the frozen world features. The residual scale is learnable, $\alpha = \tanh(g)$ with $g_{\text{init}} = 0$, preserving the native action computation at initialization.

CausalWAM-Joint. We use Fast-WAM-Joint with VideoDiT initialized from Wan2.2-TI2V-5B, without loading a RoboTwin policy. VideoDiT, the entire ActionDiT, the reader, adapter, and proprioceptive encoder are trained together. Contrastive gradients additionally reach the VideoDiT path through block 16 and its proprioceptive conditioning; the native objectives train the full video and action experts. Joint uses a fixed residual scale $\alpha = 0.1$. The VAE and T5 encoder remain frozen in both variants.

Table 1: **Standard RoboTwin 2.0.** Mean success rate (%) over 50 tasks, with 100 trials per task and condition. Baseline values are published results from the cited papers. CausalWAM-Interface uses Clean+Random action demonstrations; Table 3 evaluates transfer from Clean demonstrations.

Method	Clean ↑	Random ↑	Avg. ↑
Vision-language-action policies			
π_0 (Black et al., 2024)	65.92	58.40	62.16
X-VLA (Zheng et al., 2025)	72.90	72.80	72.85
$\pi_{0.5}$ (Physical Intelligence et al., 2025)	82.74	76.76	79.75
ABot-M0 (Yang et al., 2026a)	81.20	80.40	80.80
World-action models			
GigaWorld-Policy (Ye et al., 2026)	86.36	85.04	85.70
Motus (Bi et al., 2025)	88.66	87.02	87.84
X-WAM (Guo et al., 2026)	89.80	90.70	90.25
Fast-WAM (reported) (Yuan et al., 2026b)	91.88	91.78	91.83
LingBot-VA (Li et al., 2026)	92.90	91.50	92.20
Faster-WAM (Zhao et al., 2026)	92.78	92.26	92.52
SelfWAM (Pan et al., 2026)	92.16	93.08	92.62
AHA-WAM (Cai et al., 2026)	93.40	92.20	92.80
ImageWAM (Zhang et al., 2026a)	93.20	93.56	93.38
CausalWAM-Interface	**93.24**	**93.62**	**93.43**

Table 2: **Standard LIBERO and LIBERO-Plus robustness.** Success rate (%). LIBERO averages the four canonical suites; LIBERO-Plus evaluates the same LIBERO-trained policy under seven unseen perturbation types.

	LIBERO	LIBERO-Plus							
Method	**Avg.**	**Cam.**	**Robot**	**Lang.**	**Light**	**BG**	**Noise**	**Layout**	**Avg.**
Fast-WAM (Yuan et al., 2026b)	97.6	16.4	44.5	68.9	78.2	53.7	37.7	60.7	51.5
Faster-WAM† (Zhao et al., 2026)	99.0	53.8	71.6	94.7	96.3	61.3	63.6	79.1	73.6
ImageWAM† (Zhang et al., 2026a)	98.4	80.8	50.3	91.4	98.1	85.5	93.8	80.5	83.1
CausalWAM-Interface	**98.7**	**60.4**	**49.5**	**65.9**	**97.7**	**91.1**	**83.9**	**79.0**	**75.4**
CausalWAM-Joint	**97.8**	**72.1**	**46.3**	**76.9**	**97.8**	**95.7**	**86.8**	**80.5**	**79.4**

Inference. Paired views and the contrastive projection h_ϕ are unnecessary. Content tokens Z are computed from the current observation and reused during sampling. CausalWAM-Interface retains action-only sampling with cached native visual conditioning. CausalWAM-Joint retains native joint video–action denoising, with its added content branch conditioned only on the current observation.

5 Experiments

We evaluate whether intervention-defined state correspondence improves manipulation under visual variation while preserving task performance. Our experiments examine benchmark success, cross-appearance state retrieval, design ablations, and real-world manipulation. CausalWAM-Interface trains the content reader, adapter, and action expert over a frozen video backbone; CausalWAM-Joint additionally adapts the video backbone.

5.1 Main Results

Standard benchmark performance. On the 50-task RoboTwin 2.0 benchmark, we report mean success over 100 trials per task and evaluation condition. Under Clean+Random training, CausalWAM-Interface achieves 93.24% on Clean and 93.62% on Random, with an overall average of 93.43% (Table 1). This places the adapted policy among the strongest methods in the comparison while keeping the video backbone frozen. On standard LIBERO, CausalWAM-Interface and CausalWAM-Joint achieve 98.7% and 97.8%, respectively, averaged over the four canonical suites (Table 2). Thus, learning an appearance-consistent action-facing representation is compatible with high success on the standard manipulation benchmarks.

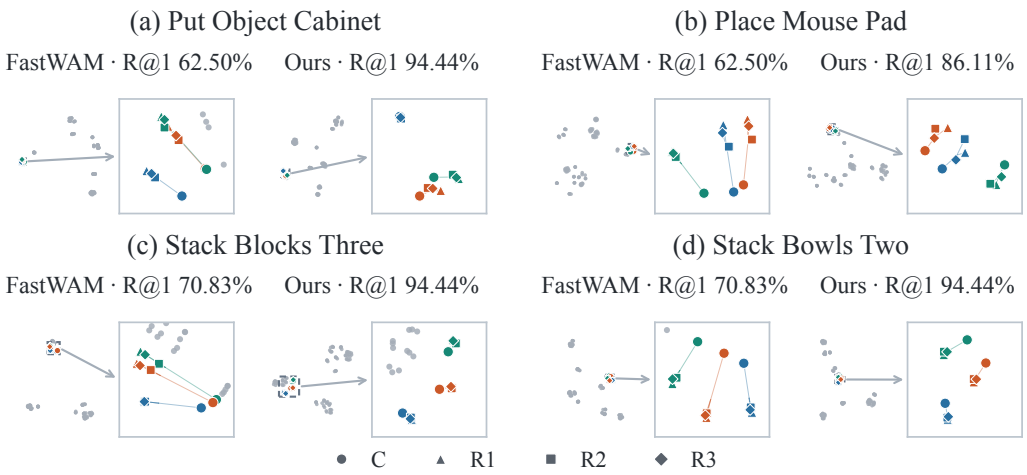


Figure 3: **Cross-appearance state correspondence.** We compare pooled FastWAM features and CausalWAM-Interface content representations on four selected RoboTwin validation tasks. Each small plot contains 96 observations: 24 physical states under C/R1/R2/R3. Colors mark three highlighted states, shapes indicate appearances, and other states are gray. Insets enlarge the marked neighborhoods. in the original feature space.

Robustness to unseen perturbations. LIBERO-Plus evaluates the same LIBERO-trained policies under seven unseen perturbation types. CausalWAM-Interface and CausalWAM-Joint attain average success rates of 75.4% and 79.4%, respectively (Table 2). Both variants perform particularly well under background changes, reaching 91.1% and 95.7%. This result directly addresses the appearance variation targeted by the intervention objective, while the full benchmark evaluates robustness across additional perturbation families.

Transfer from Clean Demonstrations. With action demonstrations restricted to Clean, Randomized success increases from 1.9% for FastWAM to 7.5% for CausalWAM-Interface and 34.7% for CausalWAM-Joint (Tab. 3). The Joint variant improves Randomized success by 32.8 percentage points while also increasing Clean success from 77.8% to 82.8%. Its Clean-to-Randomized gap decreases from 75.9 to 48.1 percentage points. These results show that joint adaptation substantially improves transfer from Clean demonstrations to randomized visual conditions without sacrificing source-domain performance.

Table 3: **Transfer from Clean demonstrations on RoboTwin 2.0.** Action supervision uses Clean demonstrations; CausalWAM additionally uses paired appearance supervision. Δ_{OOD} denotes the Clean-to-Randomized success gap in percentage points.

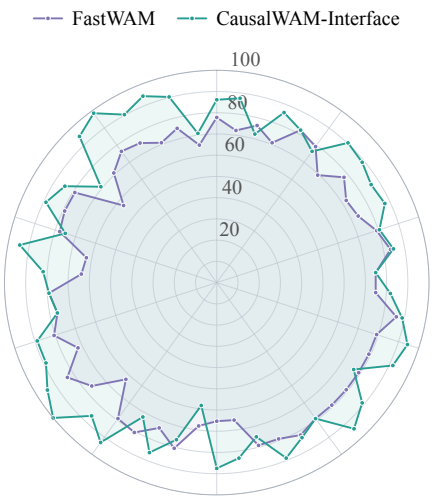
Method	Clean	Randomized	$\Delta_{\mathrm{OOD}} \downarrow$
FastWAM	77.8%	1.9%	75.9
CausalWAM-Interface	80.3%	7.5%	72.8
CausalWAM-Joint	**82.8%**	**34.7%**	**48.1**

5.2 Representation Analysis

Cross-appearance physical-state retrieval. The intervention pairs define a direct test of the desired representation: observations of the same physical state should retrieve one another despite their different appearances. We compare pooled FastWAM visual features with the content representations of CausalWAM-Interface. R@1 measures whether the nearest retrieved observation under a different rendering has the same recorded physical-state identity. Retrieval is evaluated within each task in the original feature space, testing appearance consistency together with discrimination between physical states.

Across 50 RoboTwin tasks, CausalWAM-Interface improves mean R@1 from 75.22% to 84.64%, a gain of 9.42 percentage points (Fig. 4). It improves retrieval on 37 tasks and matches FastWAM on five. For example, R@1 on Put Object Cabinet increases from 62.50% to 94.44%; Stack Blocks

(a) Per-task R@1



(b) Retrieval summary

Method	R@1 ↑	Δ (pp)	W/T/L
FastWAM	75.22	–	–
Ours	**84.64**	**+9.42**	**37/5/8**

Largest per-task improvements

Task	FastWAM	Ours	Δ
Put Object Cabinet	62.50	**94.44**	**+31.94**
Place Mouse Pad	62.50	**86.11**	**+23.61**
Place Object Basket	76.39	**100.00**	**+23.61**
Stack Blocks Three	70.83	**94.44**	**+23.61**
Stack Bowls Two	70.83	**94.44**	**+23.61**

Figure 4: **Cross-appearance physical-state retrieval on 50 RoboTwin tasks.** (a) Per-task R@1. (b) Mean R@1 and the largest per-task gains. W/T/L counts tasks with higher/equal/lower R@1 relative to FastWAM. CausalWAM-Interface improves mean R@1 by 9.42 percentage points. Full results appear in Appendix Table 5.

Table 4: **Ablation of CausalWAM-Interface.** Average success rate (%) on 50 RoboTwin tasks, varying (a) contrastive weight, (b) feature extraction layer, and (c) paired appearances. The main configuration is highlighted.

λ	Clean	Random
0	92.46	92.18
0.1	92.91	93.08
0.5	**93.24**	**93.62**
1.0	92.78	93.11

(a) Contrastive weight.

Layer	Clean	Random
4	92.37	92.54
8	92.86	93.03
16	**93.24**	**93.62**
24	92.74	92.97

(b) Feature layer.

Paired views	Clean	Random
C+R1	92.63	92.71
C+R1+R2	92.96	93.20
C+R1+R2+R3	**93.24**	**93.62**

(c) Paired appearances.

Three and Stack Bowls Two each improve from 70.83% to 94.44%. These results show that frozen world features admit a readout that more reliably preserves physical-state identity across appearance changes. Full per-task results are provided in Appendix Table 5.

Visualizing state correspondence. Figure 3 visualizes four selected validation tasks. Each task contains 24 physical states rendered under C/R1/R2/R3, yielding 96 observations. Colors identify highlighted states and marker shapes identify rendering variants. The local views illustrate appearance variants grouping more consistently by physical state in the learned representation. The t-SNE plots visualize these correspondences, while original-space R@1 provides their quantitative evaluation.

5.3 Ablation Studies

Table 4 studies the contrastive weight, feature extraction layer, and number of paired appearances for CausalWAM-Interface on the 50-task RoboTwin benchmark.

Contribution of correspondence supervision. The $\lambda = 0$ configuration retains the content reader and action adapter but removes the contrastive objective. Adding correspondence supervision with $\lambda = 0.5$ increases Clean success from 92.46% to 93.24% and Random success from 92.18% to 93.62%. The larger gain on Random, 1.44 percentage points, supports the contribution of explicit state correspondences beyond the added interface capacity. Among the tested weights, $\lambda = 0.5$ gives the highest success in both conditions.

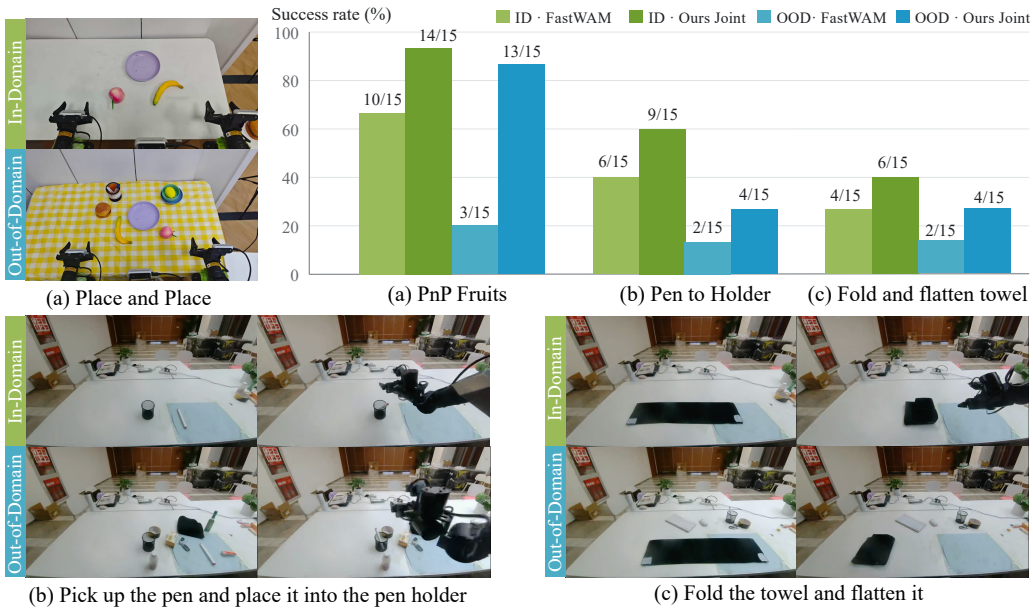


Figure 5: **Real-world manipulation under visual disturbance.** Representative scenes and task success for fruit placement, pen placement, and towel folding and flattening. FastWAM and CausalWAM-Joint are each evaluated for 15 trials per task and condition. OOD scenes introduce appearance changes and additional distractors. Bars show success rates; labels give successful trials out of 15.

Feature extraction layer. Reading from block 16 gives the best results among blocks 4, 8, 16, and 24, reaching 93.24% on Clean and 93.62% on Random. This comparison supports the intermediate-layer readout used in the main configuration.

Number of paired appearances. Random success increases from 92.71% with C+R1 to 93.20% with C+R1+R2 and 93.62% with C+R1+R2+R3. Clean success follows the same ordering, increasing from 92.63% to 92.96% and 93.24%. The trend supports using multiple matched appearances to supervise the representation of each physical state.

5.4 Real-World Manipulation

We evaluate FastWAM and CausalWAM-Joint on three real-robot tasks spanning rigid-object placement and deformable manipulation, with 15 trials per task under in-domain (ID) and visually perturbed out-of-domain (OOD) conditions (Fig. 5). CausalWAM-Joint improves aggregate success from 44.4% to 64.4% in ID scenes and, more substantially, from 15.6% to 46.7% under OOD visual shift, a gain of 31.1 percentage points. Improvements occur across all three tasks, with OOD fruit placement increasing from 3/15 to 13/15. These results show that the learned world–action interface transfers to physical manipulation, with its strongest benefit under visual disturbance.

6 Conclusion

We study visual robustness in World-Action Models through the world–action interface. CausalWAM uses action-preserving visual interventions to make physical-state correspondence explicit, learning an action-facing content representation while retaining the native generative pathway. Across RoboTwin 2.0, LIBERO-Plus, representation analysis, and real-robot experiments, this design improves robustness under visual shift; with Clean-only action supervision, CausalWAM-Joint raises RoboTwin Randomized success from 1.9% to 34.7%. These results show that world models can remain visually expressive while the action interface learns the invariances required for robust manipulation.

AI USE STATEMENT

In this work, we used generative AI tools to assist with checking parts of the system implementation and evaluation pipelines, refining experimental methodology and evaluation protocols, and polishing the manuscript and figures. All AI-assisted outputs were reviewed and verified by the authors against the underlying code, experiments, results, and source literature. The authors made all final decisions regarding the methodology, experimental design, interpretation of results, and scientific claims, and take full responsibility for the final content of this work.

REFERENCES

Kartik Ahuja, Jason S. Hartford, and Yoshua Bengio. Weakly supervised representation learning with sparse perturbations. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Hongzhe Bi, Hengkai Tan, Shenghao Xie, Zeyuan Wang, Shuhe Huang, Haitian Liu, Ruowen Zhao, Yao Feng, Chendong Xiang, Yinze Rong, Hongyan Zhao, Hanyu Liu, Zhizhong Su, Lei Ma, Hang Su, and Jun Zhu. Motus: A unified latent action world model. *arXiv preprint arXiv:2512.13030*, 2025. URL https://arxiv.org/abs/2512.13030.

Kevin Black, Noah Brown, Danny Driess, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint arXiv:2410.24164*, 2024.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022a.

Johann Brehmer, Pim de Haan, Phillip Lippe, and Taco S. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/fa567e2b2c870f8f09a87b6e73370869-Abstract-Conference.html.

Jisong Cai, Long Ling, Shiwei Chu, et al. Aha-wam: Asynchronous horizon-adaptive world-action modeling with observation-guided context routing. *arXiv preprint arXiv:2606.09811*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Qiwei Liang, Zixuan Li, Xianliang Lin, Yiheng Ge, Zhenyu Gu, Weiliang Deng, Yubin Guo, Tian Nian, Xuanbing Xie, Qiangyu Chen, Kailun Su, Tianling Xu, Guodong Liu, Mengkang Hu, Huan-ang Gao, Kaixuan Wang, Zhixuan Liang, Yusen Qin, Xiaokang Yang, Ping Luo, and Yao Mu. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *CoRR*, abs/2506.18088, 2025. doi: 10.48550/ARXIV.2506.18088. URL https://doi.org/10.48550/arXiv.2506.18088.

Senyu Fei, Siyin Wang, Junhao Shi, Zihao Dai, Jikun Cai, Pengfang Qian, Li Ji, Xinzhe He, Shiduo Zhang, Zhaoye Fei, Jinlan Fu, Jingjing Gong, and Xipeng Qiu. Libero-plus: A progressive robustness benchmark for visual-language-action models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2026, Denver, CO, USA, June 3-7, 2026*, pp. 38574–38583. Computer Vision Foundation, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Fei_LIBERO-Plus_A_Progressive_Robustness_Benchmark_for_Visual-Language-Action_Models_CVPR_2026_paper.html.

Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. Dissecting supervised contrastive learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 3821–3830. PMLR, 2021. URL https://proceedings.mlr.press/v139/graf21a.html.

Jun Guo, Qiwei Li, Peiyan Li, et al. Unified 4d world action modeling from video priors with asynchronous denoising. *arXiv preprint arXiv:2604.26694*, 2026.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

Lin Li, Qihang Zhang, Yiming Luo, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified Video Action Model. In *Robotics: Science and Systems*, 2025. URL https://arxiv.org/abs/2503.00200.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: benchmarking knowledge transfer for lifelong robot learning. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/8c3c666820ea055a77726d66fc7d447f-Abstract-Datasets_and_Benchmarks.html.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020*, volume 119, pp. 6348–6359. PMLR, 2020.

Bikang Pan, Fan Liu, Haotao Lu, et al. Selfwam: A self-grounded unified world action model for fast robot control. *arXiv preprint arXiv:2608.00725*, 2026.

Physical Intelligence, Kevin Black, Noah Brown, et al. $\pi_{0.5}$: A vision-language-action model with open-world generalization. *arXiv preprint arXiv:2504.16054*, 2025.

Carmel Rabinovitz, Niko Grupen, and Aviv Tamar. Unsupervised feature learning for manipulation with contrastive domain randomization. In *IEEE International Conference on Robotics and Automation*, 2021. URL https://arxiv.org/abs/2103.11144.

Bernhard Schölkopf, David W. Hogg, Dun Wang, Daniel Foreman-Mackey, Dominik Janzing, Carl-Johann Simon-Gabriel, and Jonas Peters. Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. USA*, 113:7391–7398, 2016.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 16451–16467, 2021.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/8929c70f8d710e412d38da624b21c3c8-Abstract.html.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119, pp. 9929–9939, 2020.

Ge Yan, Jinghao Liu, Yuzhi Fan, Lei Cai, Minwen Liao, Jesse Zhang, and Dieter Fox. Flex-π: A multi-stream world-action model with compute flexibility. *arXiv preprint arXiv:2608.10860*, 2026a. URL https://arxiv.org/abs/2608.10860.

Haodong Yan, Junfeng Li, Junjie He, Zhide Zhong, MingMing Yu, Wenxuan Song, Jiaguan Zhu, Yangyang Zheng, Yuqiao Du, Jiadi You, Yingjie Cai, Xu Yan, Guanyi Zhao, Bingbing Liu, and Haoang Li. Robust-WAM: Bridging generative pretraining and semantic foresight in world-action models. *arXiv preprint arXiv:2608.05903*, 2026b. URL https://arxiv.org/abs/2608.05903.

Yandan Yang, Shuang Zeng, Tong Lin, et al. Abot-m0: Vla foundation model for robotic manipulation with action manifold learning. *arXiv preprint arXiv:2602.11236*, 2026a.

Yiguang Yang, Jiankun Peng, Xiaoming Wang, Yiran Zhang, and Zhibo Fang. MT-WAM: Reorienting the one-pass predictive representation toward action generation. *arXiv preprint arXiv:2609.21474*, 2026b. URL https://arxiv.org/abs/2609.21474.

Dingling Yao, Dario Rancati, Riccardo Cadei, Marco Fumero, and Francesco Locatello. Unifying causal representation learning with the invariance principle. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2409.02772.

Angen Ye, Boyuan Wang, Chaojun Ni, et al. Gigaworld-policy: An efficient action-centered world–action model. *arXiv preprint arXiv:2603.17240*, 2026.

Shanglin Yuan, Weiheng Zhao, Xin Shi, Haoyi Jiang, Xianda Guo, Liu Liu, Wenyu Liu, Wei Sui, and Xinggang Wang. DreamWAM: Beyond RGB future prediction for world action models. *arXiv preprint arXiv:2608.04996*, 2026a. URL https://arxiv.org/abs/2608.04996.

Tianyuan Yuan, Zibin Dong, Yicheng Liu, and Hang Zhao. Fast-WAM: Do world action models need test-time future imagination? *arXiv preprint arXiv:2603.16666*, 2026b. URL https://arxiv.org/abs/2603.16666.

Amy Zhang, Rowan McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2006.10742.

Yuyang Zhang, Wenyao Zhang, Zekun Qi, He Zhang, Haitao Lin, Jingbo Zhang, Yao Mu, Xiaokang Yang, Wenjun Zeng, and Xin Jin. ImageWAM: Do world action models really need video generation, or just image editing? *arXiv preprint arXiv:2606.19531*, 2026a. URL https://arxiv.org/abs/2606.19531.

Zhanguang Zhang, Zhiyuan Li, Behnam Rahmati, et al. Do world action models generalize better than vlas? a robustness study. *arXiv preprint arXiv:2603.22078*, 2026b.

Weiheng Zhao, Haoyi Jiang, Xin Shi, et al. Faster-wam: Efficient inference-time future conditioning for robust world action models. *arXiv preprint arXiv:2608.04404*, 2026.

Jinliang Zheng, Jianxiong Li, Zhihao Wang, et al. X-vla: Soft-prompted transformer as scalable cross-embodiment vision-language-action model. *arXiv preprint arXiv:2510.10274*, 2025.

A Appendix

B Intervention-Defined Representation Learning

CausalWAM uses rendering interventions to define action-preserving physical-state correspondences across appearances. Following Fig. 1, we establish this supervision relation, characterize the representation geometry it induces, and connect it to learning the world–action interface.

B.1 Action-preserving rendering interventions

Let X_t denote the physical state, L the instruction, and S the intervened appearance variable. For V cameras with fixed calibrations $\kappa_1, \ldots, \kappa_V$, the observation and expert-action mechanisms are

$$\begin{aligned} O_t^s &= \big(R_v(X_t, \kappa_v, s)\big)_{v=1}^{V}, \quad q_t = Q(X_t), \\ A_t &= F_A(X_t, L, U_t^a), \end{aligned} \tag{8}$$

where q_t is proprioception and U_t^a captures expert-action randomness. Each R_v renders the physical evidence visible to its camera; the complete tuple O_t^s supplies their complementary observations.

Intervention conditions. The admissible interventions change background and table appearance while preserving physical dynamics, task semantics, and the visual evidence needed for control. Rendering copies are generated from one recorded simulator state with fixed instruction, proprioception, camera calibration, and timestamp. Robot configuration, object poses, and contact state agree across copies, as do task-defining attributes such as a target color named by the instruction.

At $X_t = x$ and $L = l$, replacing S changes the observation mechanism while leaving F_A and the distribution of U_t^a unchanged. Coupling two renderings with the same expert randomness gives identical potential actions, $A_t^s = F_A(x, l, U_t^a) = A_t^{s'}$. Marginalizing U_t^a establishes the expert-action equality in Eq. (2). This equality follows from the intervention mechanism even when physical state and appearance are statistically dependent in the training data. It justifies assigning the same demonstrated action to the different renderings of a recorded state.

State correspondence with complementary views. The positive relation matches appearance versions of the *same complete camera tuple*. An external view may provide global object poses while a wrist view provides local gripper–object geometry. Re-rendering the tuple at a fixed state preserves these complementary measurements, so the shared reader receives their combined physical evidence across appearances.

B.2 Geometry of intervention-defined contrastive supervision

Consider a balanced batch of $B \geq 2$ distinguishable physical-state identities from one task, each with $K \geq 2$ rendering copies. Write $z_{ir} = z_i^{s_r} \in \mathbb{R}^{d_z}$ for the unit-normalized readout in Eq. (4), where d_z is its embedding dimension. An anchor (i, r) has the other $K-1$ copies of state i as positives and all $K(B-1)$ copies of the other states as negatives. This construction varies appearance within a positive group and physical configuration across negative groups. Since the denominator in Eq. (5) excludes cross-task samples, the analysis applies to each task's contribution separately.

Proposition 1 (Appearance alignment and state separation)**.** *For the balanced batch above, let $\tau > 0$ and optimize Eq. (5) over freely realizable unit embeddings in $\mathbb{R}^{d_z}$ with $d_z \geq B - 1$. Its minimum is*

$$\mathcal{L}_{\mathrm{ctr}}^{\min} = \log\left[(K-1) + K(B-1)\exp\left(-\frac{B}{(B-1)\tau}\right)\right]. \tag{9}$$

The minimum is attained if and only if every rendering of state i maps to a common unit vector μ_i, and the state vectors satisfy

$$z_{ir} = \mu_i \quad \textit{for all } i, r, \qquad \mu_i^\top \mu_j = -\frac{1}{B-1} \quad \textit{for } i \neq j. \tag{10}$$

Thus, the B state centers form a unit regular simplex.

This specializes the supervised-contrastive geometry of Graf et al. (2021) to intervention-defined state groups. The proof uses the positive averaging and denominator in Eq. (5).

Proof. For anchor (i, r), define its mean positive and negative similarities as

$$u_{ir} = \frac{1}{K-1} \sum_{p \neq r} z_{ir}^\top z_{ip}, \qquad v_{ir} = \frac{1}{K(B-1)} \sum_{j \neq i} \sum_{p=1}^{K} z_{ir}^\top z_{jp}. \tag{11}$$

Applying Jensen's inequality to the positive and negative exponential sums in the denominator bounds the anchor loss by

$$\ell_{ir} \geq f(v_{ir} - u_{ir}), \qquad f(x) = \log\Big[(K-1) + K(B-1)e^{x/\tau}\Big]. \tag{12}$$

The function f is strictly increasing and strictly convex. Denoting averages over the BK anchors by $\overline{u}$ and $\overline{v}$, a second application of Jensen's inequality gives $\mathcal{L}_{\mathrm{ctr}} \geq f(\overline{v} - \overline{u})$.

Let $\mu_i = K^{-1} \sum_r z_{ir}$ be the mean embedding of state i. Since each embedding has unit norm, $\|\mu_i\| \leq 1$. Expanding the inner products in Eq. (11) yields

$$\begin{aligned} \overline{v} - \overline{u} = &-\frac{B}{B-1} + \frac{\left\|\sum_i \mu_i\right\|^2}{B(B-1)} \\ &+ \left(\frac{1}{B-1} + \frac{K}{K-1}\right)\left(1 - \frac{1}{B}\sum_i \|\mu_i\|^2\right) \geq -\frac{B}{B-1}. \end{aligned} \tag{13}$$

Both correction terms are nonnegative. Monotonicity of f then gives the lower bound in Eq. (9).

Equality in Eq. (13) requires $\sum_i \mu_i = 0$ and $\|\mu_i\| = 1$ for every state. An average of unit vectors has unit norm exactly when its constituent vectors coincide, so $z_{ir} = \mu_i$ for every rendering r. Equality in the negative-sample Jensen step further requires $\mu_i^\top \mu_j$ to be constant over $j \neq i$ for each i. Taking the inner product of μ_i with $\sum_j \mu_j = 0$ fixes this constant at $-1/(B-1)$, giving Eq. (10). Conversely, a unit regular simplex with B vertices exists in dimension $B-1$. Assigning all copies of each state to its corresponding vertex attains equality throughout, proving the stated minimum and its characterization. □

Meaning of the optimum. At the optimum, the readout is constant across each sampled state's rendering copies and distinct across states, with squared distance $\|\mu_i - \mu_j\|^2 = 2B/(B-1)$. The grouping gives the contrastive geometry its physical meaning: the same manipulation state is identified across appearances, while different sampled configurations are separated. With all negatives drawn from the same task, the discriminative target is the physical configuration.

B.3 Connecting state correspondence to action learning

CausalWAM applies this learning target to the normalized readout z, while the action expert attends to the full content-token array Z through the adapter in Sec. 4.3. Both objectives train the shared content reader: the contrastive loss supplies cross-appearance state correspondences through z, and the action loss trains the reader and its action readout through Z to predict demonstrated controls. The paired stream defines which appearance changes should be ignored; the demonstration stream determines how the resulting interface is used for manipulation.

This division follows the information requirements in Eq. (8). Changing a table texture changes the visual prediction target while preserving the expert action target at a matched physical state. The content reader provides a dedicated location for intervention supervision, alongside the original world features and native visual conditioning W. The architecture therefore retains the world model's visual pathway while learning an action-facing readout with explicit physical-state correspondence.

The proposition characterizes the representation target; the shared reader connects it to action supervision. Cross-appearance retrieval in Sec. 5.2 evaluates the learned state correspondence, and closed-loop experiments in Sec. 5 evaluate its contribution to manipulation.

Table 5: **Per-task comparison between FastWAM and our method.** We report R@1 (%) and the absolute improvement in percentage points (pp).

Task	FastWAM	Ours	△ (pp)
Adjust Bottle	77.78%	86.11%	+8.33
Beat Block Hammer	72.22%	87.50%	+15.28
Blocks Ranking RGB	76.39%	72.22%	-4.17
Blocks Ranking Size	70.83%	86.11%	+15.28
Click Alarmclock	81.94%	81.94%	0.00
Click Bell	79.17%	76.39%	-2.78
Dump Bin Bigbin	69.44%	90.28%	+20.83
Grab Roller	77.78%	88.89%	+11.11
Handover Block	72.22%	86.11%	+13.89
Handover Mic	73.61%	87.50%	+13.89
Hanging Mug	79.17%	80.56%	+1.39
Lift Pot	83.33%	84.72%	+1.39
Move Can Pot	75.00%	75.00%	0.00
Move Pillbottle Pad	75.00%	81.94%	+6.94
Move Playingcard Away	86.11%	88.89%	+2.78
Move Stapler Pad	79.17%	94.44%	+15.28
Open Laptop	79.17%	91.67%	+12.50
Open Microwave	79.17%	76.39%	-2.78
Pick Diverse Bottles	79.17%	88.89%	+9.72
Pick Dual Bottles	79.17%	94.44%	+15.28
Place A2B Left	79.17%	79.17%	0.00
Place A2B Right	81.94%	83.33%	+1.39
Place Bread Basket	79.17%	88.89%	+9.72
Place Bread Skillet	79.17%	75.00%	-4.17
Place Burger Fries	65.28%	83.33%	+18.06
Place Can Basket	65.28%	87.50%	+22.22
Place Cans Plasticbox	68.06%	58.33%	-9.72
Place Container Plate	80.56%	76.39%	-4.17
Place Dual Shoes	73.61%	86.11%	+12.50
Place Empty Cup	80.56%	72.22%	-8.33
Place Fan	79.17%	93.06%	+13.89
Place Mouse Pad	62.50%	86.11%	+23.61
Place Object Basket	76.39%	100.00%	+23.61
Place Object Scale	83.33%	94.44%	+11.11
Place Object Stand	72.22%	88.89%	+16.67
Place Phone Stand	80.56%	88.89%	+8.33
Place Shoe	76.39%	76.39%	0.00
Press Stapler	79.17%	79.17%	0.00
Put Bottles Dustbin	63.89%	81.94%	+18.06
Put Object Cabinet	62.50%	94.44%	+31.94
Rotate Qrcode	77.78%	75.00%	-2.78
Scan Object	79.17%	88.89%	+9.72
Shake Bottle	79.17%	84.72%	+5.56
Shake Bottle Horizontally	56.94%	70.83%	+13.89
Stack Blocks Three	70.83%	94.44%	+23.61
Stack Blocks Two	76.39%	98.61%	+22.22
Stack Bowls Three	75.00%	90.28%	+15.28
Stack Bowls Two	70.83%	94.44%	+23.61
Stamp Seal	75.00%	90.28%	+15.28
Turn Switch	65.28%	70.83%	+5.56
Average	**75.22%**	**84.64%**	**+9.42**

Table 6: **Per-task success rates on RoboTwin under clean and randomized evaluation settings.** All values are success rates (%).

Task	**Ours**		Fast-WAM		ImageWAM		LingBot-VA		$\pi_{0.5}$		Motus	
	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.	Clean	Rand.
Adjust Bottle	100	100	100	100	100	99	90	94	100	99	89	93
Beat Block Hammer	98	98	99	97	98	99	96	98	96	93	95	88
Blocks Ranking RGB	100	100	100	100	96	99	99	98	92	85	99	97
Blocks Ranking Size	98	98	94	98	96	100	94	96	49	26	75	63
Click Alarmclock	100	100	100	100	98	100	99	100	98	89	100	100
Click Bell	100	100	100	100	100	100	100	100	99	66	100	100
Dump Bin Bigbin	96	98	97	96	96	90	89	96	92	97	95	91
Grab Roller	100	100	100	100	100	100	100	100	100	100	100	100
Handover Block	96	85	95	81	96	95	99	78	66	57	86	73
Handover Mic	99	100	99	100	100	100	94	96	98	97	78	63
Hanging Mug	75	70	58	62	74	84	40	28	18	17	38	38
Lift Pot	100	100	100	100	100	100	100	99	96	85	96	99
Move Can Pot	93	93	90	88	96	98	94	97	51	55	34	74
Move Pillbottle Pad	99	100	100	99	98	100	99	99	84	61	93	96
Move Playingcard Away	100	100	100	100	100	99	100	99	96	84	100	96
Move Stapler Pad	75	73	77	64	67	60	91	79	56	42	83	85
Open Laptop	98	99	98	100	98	98	92	94	90	96	95	91
Open Microwave	64	60	62	45	97	94	82	86	34	77	95	91
Pick Diverse Bottles	85	86	80	85	84	88	89	82	81	71	90	91
Pick Dual Bottles	95	96	100	96	96	98	100	99	93	63	96	90
Place A2B Left	97	97	95	93	95	93	97	93	87	82	88	79
Place A2B Right	98	98	93	99	96	94	97	95	87	84	91	87
Place Bread Basket	93	95	91	93	96	92	97	95	77	64	91	94
Place Bread Skillet	95	95	90	93	90	89	95	90	85	66	86	83
Place Burger Fries	95	98	96	99	95	100	97	95	94	87	98	98
Place Can Basket	66	76	71	69	74	72	81	84	62	62	81	76
Place Cans Plasticbox	99	99	99	96	99	97	100	99	94	84	98	94
Place Container Plate	98	100	96	100	98	98	99	97	99	95	98	99
Place Dual Shoes	93	90	94	88	81	81	94	89	75	75	93	87
Place Empty Cup	100	100	100	100	100	100	100	100	100	99	99	98
Place Fan	96	96	96	96	95	94	99	93	87	85	91	87
Place Mouse Pad	85	91	83	89	84	93	93	96	60	39	66	68
Place Object Basket	88	92	89	88	86	83	91	88	80	76	81	87
Place Object Scale	91	97	90	97	97	96	96	95	86	80	88	85
Place Object Stand	97	95	90	94	98	98	99	96	91	85	98	97
Place Phone Stand	100	99	97	99	100	100	97	97	81	81	87	86
Place Shoe	98	98	96	99	97	95	98	98	92	93	99	97
Press Stapler	92	95	90	97	97	100	85	82	87	83	93	98
Put Bottles Dustbin	96	94	95	90	97	91	87	91	84	79	81	79
Put Object Cabinet	96	90	94	89	91	89	85	87	80	79	88	71
Rotate QRcode	92	91	93	89	87	92	96	91	89	87	89	73
Scan Object	88	93	89	92	94	90	96	91	72	65	67	66
Shake Bottle	100	100	100	100	100	100	100	97	99	97	100	97
Shake Bottle Horizontally	100	100	100	100	100	100	100	99	99	99	100	98
Stack Blocks Three	96	97	95	97	96	97	99	98	91	76	91	95
Stack Blocks Two	100	100	100	100	99	100	100	98	97	100	100	98
Stack Bowls Three	81	84	80	81	78	83	86	83	77	71	79	87
Stack Bowls Two	95	100	92	98	94	97	94	98	95	96	98	98
Stamp Seal	92	95	90	94	79	84	96	97	79	55	93	92
Turn Switch	74	70	61	59	77	79	44	45	62	54	84	78
Average	**93.24**	**93.62**	91.88	91.78	93.20	93.56	92.90	91.50	82.74	76.76	88.66	87.02