INVARIANT IMAGINATION: LEARNING WHAT SHOULD NOT CHANGE THE FUTURE IN JEPA WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

JEPA-based world models learn latent dynamics for planning, yet accurate prediction under clean observations may conceal failures when visual evidence or observation history is degraded. We quantify this failure with the *predictive brittleness gap*: the excess error incurred when a perturbed, rather than clean, context is used to predict the same clean future representation. On PushT, this gap grows with observation noise and accompanies sharp declines in planning success; for LeWM, success falls from 96.0% to 12.0% at $\sigma_o = 0.05$. This finding motivates a training principle: degraded contexts that retain predictive information about a recorded future should still be trained to predict that future. Following this principle, We introduce `FAPI` (*Future-Anchored Predictive Invariance*), which perturbs the observation-derived context while retaining the recorded action context and clean future target. Its transformations include visual corruption and occlusion and, for history-based predictors, masking of preceding observations. Training these contexts against a shared future anchor encourages agreement between their predictions. We evaluate `FAPI` on LeWM, Sub-JEPA, V-JEPA, and I-JEPA, FAPI improves predictive robustness, planning reliability across four control benchmarks, Something-Something-v2, and ImageNet. For instance, under Gaussian noise, `FAPI+LeWM` improves mean planning success from 49.5% to 61.5% and reduces the mean brittleness gap by 24.6%; `FAPI+Sub-JEPA` improves success from 40.0% to 57.0% with a 43.1% gap reduction.

1 INTRODUCTION

World models enable agents to anticipate future environment states, evaluate candidate actions, and perform planning through learned dynamics (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020; 2025; Hansen et al., 2022; 2024). Rather than reconstructing observations at the pixel level, Joint-Embedding Predictive Architectures (JEPAs) learn by predicting abstract representations of missing or future content (LeCun, 2022; Baevski et al., 2022; Assran et al., 2023; Bardes et al., 2024). This principle has progressed from image-level masked prediction in I-JEPA to spatiotemporal representation learning in V-JEPA and embodied prediction and planning in V-JEPA 2 (Assran et al., 2023; Bardes et al., 2024; Assran et al., 2025). Recent methods, including PLDM, DINO-WM, JEPA-WM, and LeWorldModel (LeWM), have further demonstrated the potential of JEPA-style latent dynamics for navigation and manipulation (Sobal et al., 2025; Zhou et al., 2025; Terver et al., 2026; Maes et al., 2026). In particular, LeWM jointly optimizes a visual encoder and an action-conditioned latent predictor directly from raw pixels, providing a compact and stable end-to-end foundation for JEPA-based world modeling.

However, existing JEPA world models are predominantly trained and evaluated using clean observations and complete histories. Visual corruption and occlusion can remove task-relevant evidence, while missing observations can weaken temporal reasoning (Hendrycks & Dietterich, 2019; Hendrycks et al., 2020; Chua et al., 2018; Janner et al., 2019; Lambert et al., 2020). Consequently, strong clean prediction and planning performance alone do not guarantee reliable latent imagination.

This predictive brittleness gap is the central motivation of our work. We propose `FAPI`, a *Future-Anchored Predictive Invariance* framework that learns to predict a clean future representation from multiple perturbed contexts. By preserving the clean future target as a shared predictive anchor, `FAPI` encourages stable future prediction without directly enforcing similarity between context rep-

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.

resentations or introducing an explicit consistency loss. We apply FAPI to LeWM (Maes et al., 2026), Sub-JEPA (Zhao et al., 2026), V-JEPA (Bardes et al., 2024), and I-JEPA (Assran et al., 2023), denoting the corresponding combinations as FAPI+LeWM, FAPI+Sub−JEPA, FAPI+V−JEPA, and FAPI+I−JEPA, respectively. Our main contributions are summarized as follows:

- ⋆ We identify and systematically characterize the *predictive brittleness gap* in JEPA-based world models, showing that strong prediction and planning performance under clean conditions does not guarantee reliable behavior when observations or temporal histories are degraded.
- ⋆ We introduce FAPI, a general future-anchored predictive-invariance framework that uses diverse perturbed contexts to predict a shared clean future representation. Its asymmetric context–target design improves predictive stability without requiring an additional consistency objective or modifying the clean target pathway.
- ⋆ We demonstrate the generality of FAPI by applying it to LeWM, Sub-JEPA, V-JEPA, and I-JEPA. Extensive experiments across four control benchmarks, Something-Something-v2, and ImageNet demonstrate improved predictive robustness, planning reliability, and clean–corrupt consistency across world-modeling, video, and image domains.

2 Predictive Brittleness in World-Model Planning

Standard evaluation of world models focuses primarily on prediction accuracy under clean observations. For planning, however, accuracy alone may be insufficient: when changes to the observation context do not alter the underlying future, a desirable world model should preserve its prediction of that future. Standard JEPA training does not explicitly enforce this property. This raises a natural question: *how sensitive are learned future predictions to changes in the observation context that should not change the future?* We characterize this sensitivity as *predictive brittleness*.

To quantify this sensitivity, let $\mathcal{R}*\mathrm{clean}^{\mathrm{MSE}}(\Theta)$ and $\mathcal{R}*\xi^{\mathrm{MSE}}(\Theta)$ denote the prediction risks under clean and perturbed contexts, respectively, where both predictions are evaluated against the same clean future representation. We define the corresponding *predictive brittleness gap* as

$$\mathcal{G}_{\xi}^{\mathrm{MSE}}(\Theta) = \mathcal{R}_{\xi}^{\mathrm{MSE}}(\Theta) - \mathcal{R}_{\mathrm{clean}}^{\mathrm{MSE}}(\Theta). \tag{1}$$

Throughout this analysis, predictive risk is measured using the per-dimension mean squared error: $\mathcal{R}^{\mathrm{MSE}} = \frac{|\hat{z}-z^*|_2^2}{d}$, where $\hat{z}$ denotes the predicted future representation, z^* is the corresponding clean future target, and d is the latent feature dimension ($d = 192$ in our implementation). By evaluating clean and perturbed contexts against the same future target, $\mathcal{G} * \xi^{\mathrm{MSE}}(\Theta)$ isolates the additional prediction error induced by the context perturbation. Importantly, minimizing $\mathcal{R} * \mathrm{clean}^{\mathrm{MSE}}(\Theta)$ alone places no explicit constraint on this quantity.

We study this behavior through a controlled experiment on PushT. Gaussian noise is independently applied to every frame in the observation history:

$$\widetilde{o}_{t-k} = \mathrm{clip}\left(o_{t-k} + \epsilon_{t-k},\, 0,\, 1\right), \qquad \epsilon_{t-k} \sim \mathcal{N}(0, \sigma_o^2 I), \tag{2}$$

where k indexes the observation history and $\sigma_o \in 0, 0.01, 0.03, 0.05, 0.10$ controls the corruption severity. Only the observation context is perturbed: action inputs, environment dynamics, and all other evaluation conditions remain unchanged. All models are evaluated over the same 50 episodes using an identical planning configuration.

① **The Predictive Brittleness Gap.** Figure 1 shows that predictive brittleness emerges consistently across the four world models. As observation noise increases, $\mathcal{G}_{\xi}^{\mathrm{MSE}}(\Theta)$ increases monotonically for every model, indicating progressively larger excess prediction errors relative to clean-context evaluation. The magnitude of this increase, however, differs substantially across models. These differences are largely hidden by clean-condition evaluation, revealing that clean predictive accuracy and predictive stability capture distinct properties of a learned world model.

② **Predictive Brittleness Matters for Planning.** The differences in predictive brittleness are also reflected in downstream planning performance. Across models and corruption levels, increases in $\mathcal{G}_{\xi}^{\mathrm{MSE}}(\Theta)$ are accompanied by substantial drops in planning success. Clean performance alone provides little indication of this vulnerability. LeWM, for example, achieves the highest clean success

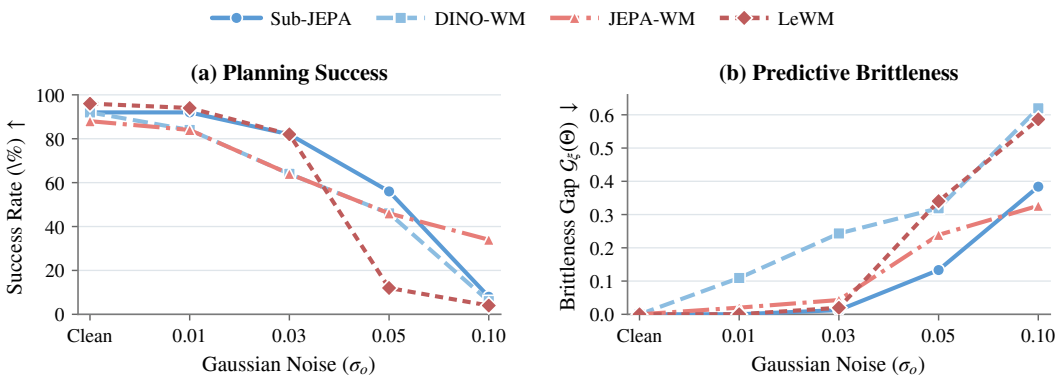


Figure 1: PushT planning success and predictive brittleness under increasing Gaussian image noise. Planning success rate and the five-step predictive brittleness gap are averaged over the same 50 evaluation episodes.

rate (96.0%), yet its success drops to 12% at $\sigma_o = 0.05$, where its brittleness gap reaches 0.340. Under the same perturbation, Sub-JEPA exhibits a considerably smaller gap (0.133) and retains 56.0% success. The pattern remains visible under stronger corruption: at $\sigma_o = 0.10$, JEPA-WM exhibits the smallest brittleness gap (0.326) and retains the highest planning success rate (34.0%).

These results highlight a limitation of evaluating and training world models solely through clean-context prediction accuracy. For planning, it is also important that imagined futures remain stable under changes in the observation context that should not alter the future itself.

3 FAPI: Future-Anchored Predictive Invariance

The previous section shows that clean prediction accuracy alone does not guarantee stable imagined futures under degraded observations. This motivates a simple principle: *different contexts associated with the same recorded future should remain consistent in what they predict about that future*. We introduce *Future-Anchored Predictive Invariance* (FAPI) to enforce this principle. FAPI does not require the conditioning representations themselves to be invariant since different degraded contexts may preserve different amounts of information and therefore need not admit identical representations. Instead, FAPI anchors their predictions to a shared clean future, so that what remains stable is their implication for the future.

3.1 Future-Anchored Predictive Learning

Predictive setting. We consider one-step latent prediction in a JEPA-based world model. For a trajectory sample $q_t = (O_t^{(H)}, A_t^{(H)}, o_{t+1}) \sim \mathcal{D}$, let $O_t^{(H)} = [o_{t-H+1}, \dots, o_t]$ and $A_t^{(H)} = [a_{t-H+1}, \dots, a_t]$. Encoding each observation as $z_\tau = E_\psi(o_\tau)$ gives the latent context $Z_t^{(H)} = [z_{t-H+1}, \dots, z_t]$. We write the predictor as

$$F_\Theta(Z, A) := P_\theta(Z, \phi_a(A)), \qquad \widehat{z}_{t+1} = F_\Theta\left(Z_t^{(H)}, A_t^{(H)}\right). \tag{3}$$

The standard single-state predictor is recovered when $H = 1$.

Future-anchored context construction. For each recorded transition, we retain the clean future representation $z^\star_{t+1} = E_\psi(o_{t+1})$ as the *future anchor*. A context transformation $\mathcal{T}_\xi$ acts only on the observation-derived conditioning context and produces $\widetilde{Z}_t^{(H),\xi}$; the corresponding action context and future anchor remain unchanged. The transformed training tuple is therefore

$$\Gamma_\xi(q_t) = \left(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)}, z^\star_{t+1}\right). \tag{4}$$

This construction is a training intervention rather than a counterfactual claim about the environment. We modify the evidence presented to the predictor while retaining the recorded future associated with the original trajectory. In particular, FAPI is applied only to transformations that retain sufficient information for predicting that future. Transformations that destroy essentially all predictive information are excluded. We select the transformation regime using the severity–recoverability diagnostic described in the experimental setup; a formal predictive-admissibility condition is provided in Appendix C.

Training objective. Given a transformed context, the resulting future prediction is $\widehat{z}_{t+1}^{\xi} = F_\Theta(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)})$. Crucially, FAPI transforms only the conditioning context while keeping the prediction target fixed. It therefore minimizes

$$\mathcal{L}_{\mathrm{FAPI}} = \mathbb{E}_{q_t\sim\mathcal{D}}\mathbb{E}_{\xi\sim p(\xi)}\left[\ell\Big(F_\Theta\Big(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)}\Big), z_{t+1}^\star\Big)\right]. \tag{5}$$

The shared target is what distinguishes FAPI from representation-level invariance. For two transformations ξ and ξ' of the same trajectory, FAPI does not require $\widetilde{Z}_t^{(H),\xi} = \widetilde{Z}_t^{(H),\xi'}$. Instead, both contexts are trained to recover the same future anchor $z_{t+1}^\star$. Thus, invariance is encouraged at the level of the *predicted future*, rather than at the level of the conditioning representation.

Why does a shared future anchor induce predictive invariance? To connect the training objective to the predictive brittleness studied in Section 2, consider two transformed contexts derived from the same trajectory. We measure their predictive drift using the same per-dimension squared error:

$$\Delta_\Theta(q_t;\xi,\xi') = \frac{1}{d}\left\|F_\Theta\Big(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)}\Big) - F_\Theta\Big(\widetilde{Z}_t^{(H),\xi'}, A_t^{(H)}\Big)\right\|_2^2. \tag{6}$$

When one of the transformations is the identity, this quantity is exactly the clean-to-degraded predictive drift underlying the predictive brittleness gap. The following statement shows that the shared anchor controls this drift without any pairwise term, and also identifies the part of the drift that no predictor can remove.

Proposition 3.1 (Shared-anchor control of predictive drift)**.** *Let ℓ in Eq. equation 5 be the per-dimension squared error $\ell(u,z) = \frac{1}{d}\|u-z\|_2^2$, let the target encoder E_ψ be held fixed, and let $\xi,\xi' \sim p(\xi)$ be drawn independently. Then, for any predictor F_Θ,*

$$\mathbb{E}_{q_t,\xi,\xi'}\left[\Delta_\Theta(q_t;\xi,\xi')\right] \;\le\; 4\,\mathcal{L}_{\mathrm{FAPI}}(\Theta). \tag{7}$$

Moreover, writing $\sigma_\xi^2 := \frac{1}{d}\,\mathbb{E}_{q_t}\Big[\operatorname{tr}\operatorname{Cov}\Big(z_{t+1}^\star \,\Big|\, \widetilde{Z}_t^{(H),\xi}, A_t^{(H)}\Big)\Big]$ for the irreducible predictive uncertainty of the future anchor given the transformed context,

$$\mathcal{L}_{\mathrm{FAPI}}(\Theta) \;\ge\; \mathbb{E}_{\xi\sim p(\xi)}\left[\sigma_\xi^2\right], \tag{8}$$

with equality if and only if F_Θ coincides almost surely with the conditional mean $\mathbb{E}[\, z_{t+1}^\star \mid \widetilde{Z}_t^{(H),\xi}, A_t^{(H)}]$.

Proposition 3.1 has two consequences. Eq. equation 7 shows that decreasing the prediction error to the shared future anchor bounds the expected disagreement between predictions from independently transformed contexts. An explicit pairwise consistency objective is therefore unnecessary: the anchor already acts as the common reference through which any two transformed contexts are compared. Eq. equation 8 then qualifies when this control is meaningful. If a transformation family retains little predictive information, $\mathbb{E}_\xi[\sigma_\xi^2]$ is large, the FAPI loss cannot be driven below it, and Eq. equation 7 no longer certifies small drift. The admissibility condition in Appendix C requires transformed contexts to retain predictive value relative to an observation-null baseline, but does not by itself guarantee a small absolute prediction error or drift bound.

3.2 Context Transformations for Sequential Prediction

FAPI is agnostic to the particular transformation family. In sequential world models, however, the conditioning context contains two distinct sources of predictive evidence: the visual content available at each time step and the temporal history available to the predictor. We therefore consider two

complementary transformation families. *Visual-content transformations* alter what is observed at each time step, whereas *history-availability transformations* alter which preceding states are available to the predictor. The latter applies only when $H > 1$.

Visual-content transformations. We first perturb the visual evidence available at individual time steps. For each $\tau = t - H + 1, \ldots, t$, a visual transformation is applied before encoding:

$$z_\tau^{\xi_{\text{vis}}} = E_\psi\left(\mathcal{T}_{\text{vis}}^{\xi_{\text{vis}}}(o_\tau)\right). \tag{9}$$

Collecting the transformed states gives

$$Z_t^{(H),\text{vis}} = \left[z_{t-H+1}^{\xi_{\text{vis}}}, \ldots, z_t^{\xi_{\text{vis}}}\right]. \tag{10}$$

We use transformations that degrade the visual evidence available to the predictor, including global appearance corruption and localized occlusion. These transformations change *what* is observed while preserving which temporal positions remain available.

History-availability transformations. For $H > 1$, predictive information also comes from states preceding the current observation. We therefore additionally perturb the availability of this historical evidence. Writing

$$Z_t^{(H),\text{vis}} = \left[Z_{<t}^{(H-1),\text{vis}}, z_t^{\xi_{\text{vis}}}\right],$$

we sample a binary history mask $m^{\text{hist}} = [m_{t-H+1}, \ldots, m_{t-1}] \in \{0,1\}^{H-1}$, where $m_\tau = 0$ indicates that the corresponding historical state is unavailable. Let $\mathcal{M}_{m^{\text{hist}}}$ denote the resulting masking operator. Combining the visual and history-availability transformations gives

$$\widetilde{Z}_t^{(H),\xi} = \left[\mathcal{M}_{m^{\text{hist}}}\left(Z_{<t}^{(H-1),\text{vis}}\right), z_t^{\xi_{\text{vis}}}\right], \qquad \xi = \left(\xi_{\text{vis}}, m^{\text{hist}}\right). \tag{11}$$

The action history $A_t^{(H)}$ remains unchanged. When $H = 1$, no preceding history exists, so the history-availability transformation disappears and FAPI reduces naturally to transforming only the current visual observation.

Overall, `FAPI` leaves the prediction architecture, prediction horizon, and action-conditioning pathway unchanged. It modifies only the observation-derived evidence presented during training while preserving the clean future as a shared predictive anchor. In this way, `FAPI` directly targets the failure identified in Section 2: learning not only to predict the future, but also which changes in the conditioning context should *not* change that prediction.

4 Experiments

Datasets and Metrics. We evaluate `FAPI` across four representative JEPA-style architectures covering world-model control, video prediction, and image prediction. For world-model control, we use PushT, OGB-Cube, TwoRoom, and DMC, which respectively evaluate planar manipulation, three-dimensional object interaction, visual navigation, and continuous control. Planning performance is measured by task success rate. We additionally report the per-dimension mean squared error (MSE) between the predicted representation and the corresponding clean future target, together with the predictive brittleness gap, defined as the increase in predictive risk from clean to degraded contexts. For video-level evaluation, we use Something-Something-v2 and consider video corruption, history masking, and their combination. For image-level evaluation, we use ImageNet with Gaussian visual corruption. In both settings, we report prediction error and clean–corrupt consistency using cosine similarity and relative ℓ_2 distance. Detailed task configurations, corruption levels, and evaluation protocols are provided in Appendix B.

Baselines. We apply `FAPI` to pretrained models for world-model control, video prediction, and image prediction. For world-model control, we fine-tune LeWorldModel and Sub-JEPA using future-anchored context degradation, yielding `FAPI+LeWM` and `FAPI+Sub-JEPA`, respectively. For video and image prediction, we apply the same fine-tuning framework to V-JEPA and I-JEPA, yielding `FAPI+V-JEPA` and `FAPI+I-JEPA`, respectively. For each of these two prediction settings, we include the frozen pretrained model and a fine-tuned baseline trained without context

Table 1: Planning success (Succ., %, ↑) and predictive brittleness gap $\mathcal{G}_{\xi}^{\mathrm{MSE}}(\Theta)$ (Gap, ↓) under clean contexts and fixed degradation settings. Gaussian noise uses $\sigma_o = 0.10$, patch masking uses $r_p = 0.30$, and history masking uses $r_t = 0.30$. Mean denotes the arithmetic mean across the four benchmarks. Success and gap values are reported to one and three decimal places, respectively.

Condition	Model	PushT		OGB-Cube		TwoRoom		DMC		Mean	
		Succ.	Gap	Succ.	Gap	Succ.	Gap	Succ.	Gap	Succ.	Gap
Clean	LeWM	96.0	-	74.0	-	88.0	-	86.0	-	86.0	-
	FAPI+LeWM	**98.0**	-	**78.0**	-	**96.0**	-	**88.0**	-	**90.0**	-
Gaussian	LeWM	4.0	0.586	46.0	0.328	80.0	0.344	68.0	0.007	49.5	0.316
	FAPI+LeWM	**16.0**	**0.465**	**52.0**	**0.294**	**92.0**	**0.195**	**86.0**	**0.000**	**61.5**	**0.239**
Patch	LeWM	2.0	0.868	42.0	0.580	80.0	0.664	**30.0**	0.726	38.5	0.710
	FAPI+LeWM	**6.0**	**0.803**	**46.0**	**0.499**	**84.0**	**0.645**	**30.0**	**0.718**	**41.5**	**0.666**
History	LeWM	90.0	0.010	66.0	0.282	82.0	0.125	**82.0**	**0.005**	80.0	0.105
	FAPI+LeWM	**96.0**	**0.004**	**72.0**	**0.154**	**92.0**	**0.009**	**82.0**	0.005	**85.5**	**0.043**

Note: Boldface comparisons and relative gap reductions use values before rounding. A displayed gap of 0.000 denotes a positive value below 0.0005. The two DMC history gaps both round to 0.005, but LeWM has the lower value before rounding.

degradation to distinguish the effect of future-anchored context degradation from that of additional adaptation alone.

Implementation Details. All FAPI-based models are initialized from the corresponding pretrained checkpoints. We adapt FAPI+LeWM, FAPI+Sub-JEPA, and FAPI+V-JEPA for 5 epochs, and FAPI+I-JEPA for 10 epochs, using AdamW. The fine-tuned and FAPI-based models use identical checkpoint initialization, training data, optimizer, learning-rate schedule, number of update steps, and adaptation duration. Their only difference is whether degradation is applied to the predictive context while the future target remains clean. Complete optimization settings and architecture-specific details are reported in Appendix B.

4.1 Robustness to Degraded Predictive Context

We evaluate whether future-anchored adaptation improves downstream planning under degraded predictive contexts. Experiments are conducted on PushT, OGB-Cube, TwoRoom, and DMC using matched evaluation and planning settings. Success rate is computed over 50 evaluation episodes for each setting. Full configurations and severity sweeps are provided in Appendix F.

Visual-Content Robustness. ❶ *Perturbation-based adaptation also improves clean planning.* As shown in Table 1, FAPI+LeWM improves clean success on all four benchmarks, raising the mean from 86.0% to 90.0%. This gain indicates that adaptation to degraded contexts does not come at the expense of performance on clean observations. It also suggests that the learned improvement extends beyond the perturbed inputs used during adaptation.

❷ *Visual robustness gains depend on the perturbation and task.* Under Gaussian noise, FAPI+LeWM improves success on every benchmark, with a mean gain of 12.0 percentage points; DMC benefits most (+18.0 points). Under patch masking, the mean gain is 3.0 points, and DMC shows no success improvement. The lower gaps under both perturbations support improved predictive stability, but the smaller planning gains under patch masking indicate that missing visual content remains harder to address at the evaluated settings.

History-Availability Robustness. ❶ *Adaptation improves planning with incomplete history.* Under history masking, FAPI+LeWM raises mean success from 80.0% to 85.5%, improving PushT, OGB-Cube, and TwoRoom. Its success remains within 2.0–6.0 percentage points of its clean result on every benchmark when the latest frame is retained. These results show that the adaptation benefits planning even when earlier observations are unavailable. ❷ *The benefit is consistent across conditions, but its size varies by task.* TwoRoom gains 8.0 points under clean, Gaussian, and patch conditions and 10.0 points under history masking. In contrast, DMC gains 18.0 points under Gaus-

sian noise but none under patch or history masking. Thus, the method's advantage is broad across the evaluated conditions, while its magnitude depends on which information the task needs. The reduced mean history gap provides supporting evidence of predictive stability, without implying a planning gain on DMC.

Generalization to Out-of-Distribution Context Degradations. ❶ *Robustness gains transfer to unseen degradation operators.* Without further adaptation, `FAPI+LeWM` outperforms LeWM in all 12 TwoRoom settings in Table 2. The gains span visual smoothing, contiguous occlusion, and repeated observations, supporting transfer beyond the specific noise and masking operators used during training. Mean success improves by 9.5, 7.0, and 5.0 percentage points under blur, block cutout, and context repetition, respectively. ❷ *Transfer benefits persist across degradation levels.* Under blur, the success advantage grows from 8.0 points at L_1 to 12.0 points at L_4. For block cutout and context repetition, gains vary across levels but remain positive throughout. These patterns show that the benefits are not confined to a single evaluation setting, with the largest average success gain observed under blur. Full settings appear in Appendix G.

Table 2: Planning success rates (%) on TwoRoom under out-of-distribution context degradations, evaluated over 50 episodes per setting. During adaptation, `FAPI+LeWM` uses Gaussian context noise ($\sigma_o = 0.05$, application probability 1.0), random patch masking (16×16 patches, masking ratio $r_p = 0.10$, zero filling), and random history masking (masking probability $r_t = 0.05$, latest frame retained), with clean future targets. Gaussian blur, block cutout, and context repetition are excluded from adaptation and evaluated without further training. Mean averages over four severity levels. Severity definitions are provided in Table 11.

Level	Gaussian blur		Block cutout		Context repetition	
	LeWM	`FAPI+LeWM`	LeWM	`FAPI+LeWM`	LeWM	`FAPI+LeWM`
L_1	86.0	**94.0**	88.0	**92.0**	88.0	**94.0**
L_2	84.0	**92.0**	76.0	**88.0**	88.0	**94.0**
L_3	78.0	**88.0**	78.0	**84.0**	86.0	**90.0**
L_4	72.0	**84.0**	66.0	**72.0**	82.0	**86.0**
Mean	80.0	**89.5**	77.0	**84.0**	86.0	**91.0**

4.2 Generalization Beyond LeWorldModel

We evaluate `FAPI` beyond LeWorldModel on action-conditioned Sub-JEPA and action-free V-JEPA and I-JEPA, perturbing prediction contexts while retaining clean targets. For V-JEPA and I-JEPA, standard and future-anchored adaptation use identical initialization, data, optimization settings, and adaptation duration.

Combination of `FAPI` and Sub-JEPA. ❶ *Benefits extend to a second world-model architecture under both clean and degraded contexts.* Table 3 shows that `FAPI+Sub-JEPA` improves planning success across all four benchmarks and all evaluated conditions. Despite adaptation with perturbed contexts, mean clean success also increases from 80.5% to 89.5%. Together with the LeWM results, this pattern supports the applicability of `FAPI` across the two evaluated architectures. ❷ *Planning gains vary across perturbations and tasks.* Mean success gains are 17.0 percentage points under Gaussian noise, 8.5 under patch masking, and 10.0 under history masking. DMC shows the largest gains under both Gaussian noise (+26.0 points) and patch masking (+12.0 points), whereas history-masking gains are more evenly distributed across tasks (+6.0–12.0 points). Notably, PushT reaches 100.0% success under both clean and history-masked contexts. Lower mean gaps across all three degradations provide supporting evidence of improved predictive stability, although individual gap changes need not track success gains, as observed for PushT under history masking.

Combination of `FAPI` and V-JEPA. ❶ *FAPI improves prediction under both clean and degraded video contexts.* On SSv2, we evaluate Gaussian video noise ($\sigma_o = 0.02$), history masking ($p_t = 0.05$, latest frame retained), and their combination. Table 4 shows that `FAPI+V-JEPA` achieves the lowest clean and mean corrupted losses (0.485/0.487). Standard fine-tuning improves feature consistency over the frozen baseline but increases prediction loss; FAPI improves both, demonstrating benefits beyond standard adaptation. ❷ *Prediction quality is maintained alongside*

Table 3: Planning success (Succ., %, ↑) and predictive brittleness gap (Gap, ↓) for Sub-JEPA and FAPI+Sub-JEPA. Gaussian noise uses $\sigma_o = 0.10$, patch masking uses $r_p = 0.30$, and history masking uses $r_t = 0.30$. Mean denotes the arithmetic mean across the four benchmarks. Success and gap values are reported to one and three decimal places, respectively; relative reductions are calculated before rounding. Bold values indicate the better result within each condition and benchmark, including the mean.

Condition	Model	PushT		OGB-Cube		TwoRoom		DMC		Mean	
		Succ.	Gap	Succ.	Gap	Succ.	Gap	Succ.	Gap	Succ.	Gap
Clean	Sub-JEPA	92.0	–	72.0	–	96.0	–	62.0	–	80.5	–
	FAPI+Sub-JEPA	**100.0**	–	**80.0**	–	**98.0**	–	**80.0**	–	**89.5**	–
Gaussian	Sub-JEPA	8.0	0.384	44.0	0.492	56.0	0.703	52.0	0.048	40.0	0.407
	FAPI+Sub-JEPA	**14.0**	**0.268**	**62.0**	**0.166**	**74.0**	**0.466**	**78.0**	**0.026**	**57.0**	**0.232**
Patch	Sub-JEPA	6.0	0.369	42.0	0.490	38.0	0.798	18.0	0.236	26.0	0.473
	FAPI+Sub-JEPA	**10.0**	**0.299**	**50.0**	**0.371**	**48.0**	**0.707**	**30.0**	**0.059**	**34.5**	**0.359**
History	Sub-JEPA	90.0	**0.003**	64.0	0.116	90.0	0.010	60.0	0.025	76.0	0.039
	FAPI+Sub-JEPA	**100.0**	0.007	**76.0**	**0.082**	**96.0**	**0.003**	**72.0**	**0.006**	**86.0**	**0.025**

Table 4: Context-degradation robustness on SSv2 with V-JEPA and ImageNet with I-JEPA. Frozen denotes evaluation without fine-tuning; Fine-tuned denotes standard adaptation. SSv2 corrupted results average over Gaussian video noise and history masking. ImageNet uses Gaussian context-image noise ($\sigma_o = 0.20$) with a clean target pathway. Cosine similarity and relative ℓ_2 distance measure clean–corrupt feature consistency. All values are reported to three decimal places. Bold indicates the best result within each JEPA setting.

Model	Prediction loss ↓		Cosine similarity ↑	Relative ℓ_2 ↓
	Clean	Corrupted		
SSv2 / V-JEPA				
Frozen V-JEPA	0.546	0.570	0.929	0.320
Fine-tuned V-JEPA	0.588	0.594	0.977	0.143
FAPI+V-JEPA	**0.485**	**0.487**	**0.994**	**0.062**
ImageNet / I-JEPA				
Frozen I-JEPA	0.191	0.223	0.861	0.493
Fine-tuned I-JEPA	0.193	0.223	0.867	0.481
FAPI+I-JEPA	**0.187**	**0.212**	**0.888**	**0.437**

greater feature consistency. The clean-to-corrupted loss increase is only 0.002 for FAPI, compared with 0.024 for Frozen V-JEPA and 0.006 for Fine-tuned V-JEPA. FAPI also achieves the highest cosine similarity (0.994) and lowest relative ℓ_2 distance (0.062). Together, these results support reduced sensitivity to context degradation while improving clean-context prediction.

Combination of FAPI and I-JEPA. ❶ *Benefits extend to image-based context–target prediction.* On ImageNet, Gaussian noise ($\sigma_o = 0.20$) is applied only to the context image, with the target pathway kept clean. FAPI+I-JEPA reduces both clean and corrupted losses to $0.187/0.212$, outperforming the frozen and fine-tuned baselines. The improvement over Frozen I-JEPA is larger under corrupted contexts (0.011) than clean contexts (0.004), supporting robustness gains alongside improved clean prediction. ❷ *FAPI provides gains where standard adaptation offers limited improvement.* Standard fine-tuning leaves corrupted loss unchanged at 0.223 and only slightly improves feature consistency. FAPI lowers this loss while increasing cosine similarity to 0.888 and reducing relative ℓ_2 distance to 0.437. These complementary improvements extend the observed benefits beyond video prediction to the evaluated image-noise setting.

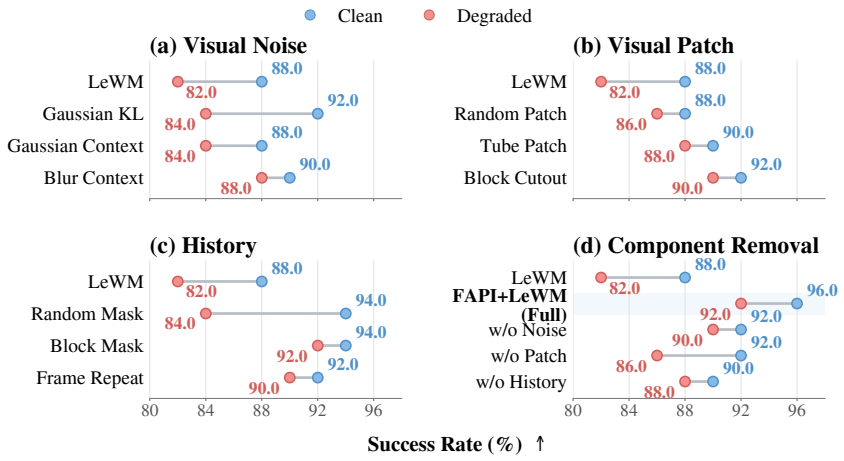


Figure 2: Ablation study on TwoRoom, reporting planning success (%). Panels (a)–(c) compare transformation designs under clean conditions and the respective single degradation family: visual noise, patch masking, and history masking. Panel (d) evaluates component removal under clean conditions and combined degradation using Gaussian image noise ($\sigma_o = 0.05$), patch masking ($r_p = 0.10$), and history masking ($r_t = 0.05$). Component removal applies during adaptation; the combined evaluation condition remains fixed.

4.3 Ablation

❶ *Transformation design improves robustness beyond gains in clean performance.* Figure 2(a)–(c) shows that Blur Context, Block Cutout, and Block Mask achieve the highest degraded success within their respective groups. Notably, Block Mask improves history-degraded success from 84.0% to 92.0% over Random Mask while preserving the same clean success of 94.0%. Similarly, Blur Context outperforms Gaussian KL under degradation (88.0% versus 84.0%), despite lower clean success. These comparisons indicate that transformation design matters specifically for robustness, beyond its effect on clean planning. ❷ *All three adaptation components contribute to the complete model's performance.* In Figure 2(d), the complete `FAPI+LeWM` achieves the highest success under both clean (96.0%) and combined-degradation (92.0%) conditions. Removing any component reduces success in both settings, supporting the value of their joint use. Under the fixed combined evaluation, removing patch masking causes the largest decrease (6.0 percentage points), followed by history masking (4.0) and noise (2.0). Thus, patch masking contributes most to combined-degradation success in this ablation, while the other components provide additional benefits. Variant definitions are provided in Appendix H.

5 Conclusion

We identify predictive brittleness as a limitation of JEPA-based world models: strong clean performance does not guarantee reliable prediction and planning under degraded observations. This motivates a training principle: observation-derived contexts that retain predictive information about the same recorded future should remain predictive of that future. We introduce `FAPI`, which trains degraded contexts against shared clean target representations, without requiring an explicit consistency objective. Experiments with LeWorldModel and Sub-JEPA across four control benchmarks show improved mean planning success under clean, Gaussian-noise, patch-masking, and history-masking conditions, together with reduced mean predictive brittleness gaps. Extensions to V-JEPA on Something-Something-v2 and I-JEPA on ImageNet further yield lower prediction error and stronger clean–corrupt consistency. These findings support future anchoring as a transferable approach to improving JEPA prediction reliability, while severe information loss and the imperfect correspondence between predictive brittleness and planning success remain open challenges.

5.1 Ethics Statement

This work does not involve the collection of new human-participant data, user studies, or deployment in real-world systems. Our experiments use established benchmark datasets and simulated control environments. ImageNet and Something-Something-v2 are used solely to evaluate learned representations; we do not attempt to identify individuals, infer sensitive attributes, or create new annotations about people. Third-party datasets, code, and pretrained checkpoints are used in accordance with their applicable terms and licences. All control experiments are conducted in simulation.

Reproducibility statement

We provide a mathematical specification of `FAPI`, its training objective, context transformations, and theoretical properties in the main paper. The appendix reports the datasets and evaluation protocols, checkpoint initialization, optimization settings, learning rates, batch sizes, adaptation durations, perturbation operators, severity levels, evaluation metrics, and definitions of all ablation variants. It also provides complete downstream severity sweeps, held-out corruption experiments, predictive-recoverability analyses, and the proof of Proposition 3.1.

AI use statement

Large language models were used solely to refine the wording of selected sentences and to check formatting during manuscript preparation. All ideas, methods, experiments, analyses, and conclusions are the authors' own, and the authors take full responsibility for the content of this paper.

References

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

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. doi: 10.48550/arXiv.2506.09985. URL https://arxiv.org/abs/2506.09985.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022. URL https://proceedings.mlr.press/v162/baevski22a.html.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://proceedings.neurips.cc/paper/2018/hash/3de568f8597b94bda53149c7d7f5958c-Abstract.html.

David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2450–2462. Curran Associates, Inc., 2018.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020.

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

Nicklas Hansen, Xiaolong Wang, and Hao Su. Temporal difference learning for model predictive control. In *Proceedings of the International Conference on Machine Learning*, 2022.

Nicklas Hansen, Hao Su, and Xiaolong Wang. TD-MPC2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, 2024.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HJz6tiCqYm.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1gmrxHFvB.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5faf461eff3099671ad63c6f3f094f7f-Abstract.html.

Nathan Lambert, Brandon Amos, Omry Yadan, and Roberto Calandra. Objective mismatch in model-based reinforcement learning. In *Proceedings of the 2nd Conference on Learning for Dynamics and Control*, volume 120 of *Proceedings of Machine Learning Research*, pp. 761–770. PMLR, 2020. URL https://proceedings.mlr.press/v120/lambert20a.html.

Yann LeCun. A path towards autonomous machine intelligence. *OpenReview.net*, 2022. URL https://openreview.net/forum?id=BZ5a1r-kVsf. Version 0.9.2.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Vlad Sobal et al. Learning from reward-free offline data: A case for planning with latent dynamics models. *arXiv preprint arXiv:2502.14819*, 2025.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann Le Cun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026.

Kai Zhao, Dongliang Nie, Yuchen Lin, Zhehan Luo, Yixiao Gu, Deng-Ping Fan, and Dan Zeng. Sub-JEPA: Subspace gaussian regularization for stable end-to-end world models. *arXiv preprint arXiv:2605.09241*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. In *Proceedings of the International Conference on Machine Learning*, 2025.

APPENDIX

A The Use of Large Language Models (LLMs) 13

B Detailed Experimental Setup 13

B.1 Perturbation Operators and Severity Levels 13

B.2 Evaluation Protocol 13

B.3 Adaptation and Implementation Details 14

C Formalizing Predictive Admissibility 14

D Predictive Recoverability under Degraded Contexts 14

E Strongest-Severity Downstream Comparison 14

E.1 Matched-Backbone Comparison 14

E.2 Effect of `FAPI` Adaptation on LeWM 16

F Complete Downstream Severity Sweeps 16

F.1 Gaussian Noise 16

F.2 Patch Masking 16

F.3 History Masking 17

G Implementation of Out-of-Distribution Corruptions 17

H Definitions of Ablation Variants 18

I V-JEPA Robustness Across Combined-Degradation Severities 19

J I-JEPA Gaussian-Noise Severity-Sweep Evaluation 19

K Proof for the Proposition 3.1 20

A THE USE OF LARGE LANGUAGE MODELS (LLMS)

Large language models (LLMs) were not used as part of the core methodology, experiments, or original research contributions. They were only used as an assistive tool for language editing and formatting.

B DETAILED EXPERIMENTAL SETUP

World-model control. We evaluate downstream planning performance on PushT, OGB-Cube, TwoRoom, and DMC, covering planar object manipulation, three-dimensional object interaction, visual navigation, and continuous control, respectively. We compare `FAPI+LeWM` and `FAPI+Sub-JEPA` with their corresponding pretrained baselines, LeWM and Sub-JEPA, under matched evaluation and planning settings. For each benchmark, both adapted models are initialized from their respective pretrained checkpoints and fine-tuned for 5 epochs using AdamW with a learning rate of 5×10^{-5}.

Video-level evaluation. We evaluate Frozen V-JEPA, Fine-tuned V-JEPA, and `FAPI+V-JEPA` on Something-Something-v2. Frozen V-JEPA denotes the pretrained V-JEPA model evaluated without additional adaptation, whereas Fine-tuned V-JEPA is adapted using the standard fine-tuning procedure without context perturbations. The evaluated corruption conditions include Gaussian video noise, history masking, and their combination. Fine-tuned V-JEPA and `FAPI+V-JEPA` are initialized from the same pretrained V-JEPA checkpoint and adapted for 5 epochs with a batch size of 4. AdamW is used with a configured learning rate of 6.25×10^{-4}, while the learning-rate scheduler begins from 2×10^{-4}. Training is performed using bfloat16 precision.

Image-level evaluation. Frozen I-JEPA, Fine-tuned I-JEPA, and `FAPI+I-JEPA` are initialized from the same pretrained I-JEPA ViT-H/14 checkpoint. Frozen I-JEPA is evaluated without additional adaptation, whereas Fine-tuned I-JEPA and `FAPI+I-JEPA` are adapted on the ImageNet training split and evaluated on the ImageNet validation split. We consider only visual context degradation, implemented using Gaussian image noise, while the target image and target-encoder pathway remain clean. Both adapted models are trained for 10 epochs using the same data, optimizer, learning-rate schedule, and number of update steps. AdamW is used with a peak learning rate of 5×10^{-5}, an initial learning rate of 1×10^{-5}, and a final learning rate of 1×10^{-6}. Training uses two GPUs with a batch size of 96 per GPU. The only difference between Fine-tuned I-JEPA and `FAPI+I-JEPA` is whether Gaussian visual degradation is applied to the predictive context during adaptation.

For both the video- and image-level evaluations, the Fine-tuned and `FAPI`-based models use the same checkpoint initialization, training data, optimizer, learning-rate schedule, and adaptation duration. Their only difference is that context perturbations are disabled during standard fine-tuning and enabled when applying `FAPI`, while the predictive target remains clean in both settings.

B.1 PERTURBATION OPERATORS AND SEVERITY LEVELS

For world-model control, we consider visual corruption, visual occlusion, and history perturbation. Each perturbation family contains multiple operators or sampling strategies and is evaluated at progressively increasing severity levels. Visual perturbations are applied to the observation context before encoding, and history perturbations modify the available history.

Only the selected perturbation family is varied in each individual severity sweep, while all remaining evaluation settings are fixed. The specific perturbation operators, application probabilities, and severity values are described in the following appendix subsections.

B.2 EVALUATION PROTOCOL

For downstream control, we report the planning success rate over 50 evaluation episodes for each model and perturbation setting. All compared models use the same task initialization, planning configuration, perturbation severity, and success criterion.

For V-JEPA and I-JEPA, we report prediction loss under clean and perturbed conditions. We additionally measure perturbation-induced degradation and clean–corrupt feature consistency where applicable. Lower prediction loss and degradation indicate more stable latent prediction, whereas higher feature consistency indicates better preservation of representations under perturbation.

B.3 Adaptation and Implementation Details

All FAPI-based models are initialized from pretrained checkpoints rather than trained from scratch. For each control benchmark, FAPI+LeWM and FAPI+Sub-JEPA are initialized from the corresponding LeWM and Sub-JEPA checkpoints, respectively. Both models are adapted for 5 epochs using AdamW with a learning rate of 5×10^{-5}. FAPI+V-JEPA is initialized from the official V-JEPA checkpoint and adapted for 5 epochs, while FAPI+I-JEPA is initialized from the official I-JEPA ViT-H/14 checkpoint and adapted for 10 epochs.

During adaptation, transformations are applied only to the observation-derived predictive context, while target representations are computed from the corresponding clean inputs. The prediction architecture and, where applicable, action-conditioning pathway remain unchanged. For V-JEPA and I-JEPA, standard-adaptation baselines use identical initialization, training data, optimization settings, and adaptation duration, but omit context degradation. Thus, these comparisons isolate the effect of future-anchored context degradation from additional clean adaptation.

C Formalizing Predictive Admissibility

Let $\mathcal{F}$ denote a fixed hypothesis class. We define the optimal predictive risk under transformation ξ as

$$\mathcal{R}_\xi^\star = \inf_{f \in \mathcal{F}} \mathbb{E}\left[\ell\left(f(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)}), z_{t+1}^\star\right)\right].$$

A transformation regime is predictively admissible if

$$\mathcal{R}_\xi^\star \leq \mathcal{R}_{\text{null}}^\star - \gamma, \qquad \gamma > 0.$$

D Predictive Recoverability under Degraded Contexts

We report the complete predictive-recoverability results corresponding to Section 4.1. This analysis evaluates one-step future-latent prediction under the strongest severity of each context degradation. The degraded-context risk, $\mathcal{R}_\xi$, is compared with the observation-null risk, $\mathcal{R}_{\text{null}}$, obtained by replacing all observation-derived latent tokens in the initial predictive context with zeros while retaining the corresponding planned action sequence and clean future target.

The recoverability margin is defined as $\Delta_{\text{rec}} = \mathcal{R}_{\text{null}} - \mathcal{R}_\xi$. A positive margin indicates that the degraded context retains useful future-predictive information beyond that available from the planned action sequence alone.

Table 5 shows that FAPI+LeWM achieves a positive recoverability margin with a 95% confidence interval excluding zero in all 12 evaluated task–degradation combinations. LeWM satisfies the same criterion in 11 combinations, with the exception of patch masking on TwoRoom. Moreover, FAPI+LeWM produces a larger recoverability margin than LeWM in 11 of the 12 combinations; the only exception is history masking on PushT, where the two models achieve comparable margins. These results indicate that the selected degradation severities generally preserve useful predictive information rather than reducing the context to an effectively observation-null input.

E Strongest-Severity Downstream Comparison

E.1 Matched-Backbone Comparison

Table 6 reports the complete numerical results corresponding to Table 1. We compare FAPI+LeWM with its matched LeWM backbone under the strongest evaluated severity of each perturbation family. Success rate (%) is reported, and Gain denotes the signed difference in success rate,

Table 5: Predictive recoverability under the strongest context degradations. GN: Gaussian noise; PM: patch masking; HM: history masking. Risks denote one-step future-latent prediction MSE. The recoverability margin is defined as $\Delta_{\text{rec}} = \mathcal{R}_{\text{null}} - \mathcal{R}_{\xi}$ and is reported with a paired-bootstrap 95% confidence interval. Margins and confidence intervals are calculated using unrounded values.

Task	Deg.	Model	Sev.	$\mathcal{R}_{\xi} \downarrow$	$\mathcal{R}_{\text{null}} \downarrow$	Δ_{rec} [95% CI] $\uparrow$
OGB-Cube	GN	FAPI+LeWM	0.10	0.655	0.992	0.337 [0.214, 0.456]
OGB-Cube	GN	LeWM	0.10	0.694	0.999	0.304 [0.172, 0.434]
OGB-Cube	PM	FAPI+LeWM	0.30	0.542	1.019	0.477 [0.418, 0.534]
OGB-Cube	PM	LeWM	0.30	0.607	1.047	0.440 [0.387, 0.489]
OGB-Cube	HM	FAPI+LeWM	0.30	0.193	1.019	0.826 [0.760, 0.888]
OGB-Cube	HM	LeWM	0.30	0.233	0.997	0.764 [0.662, 0.850]
PushT	GN	FAPI+LeWM	0.10	0.533	1.001	0.468 [0.384, 0.546]
PushT	GN	LeWM	0.10	0.589	1.045	0.456 [0.353, 0.554]
PushT	PM	FAPI+LeWM	0.30	0.377	1.008	0.631 [0.582, 0.675]
PushT	PM	LeWM	0.30	0.659	1.120	0.461 [0.393, 0.528]
PushT	HM	FAPI+LeWM	0.30	0.012	1.005	0.993 [0.981, 1.005]
PushT	HM	LeWM	0.30	0.018	1.018	1.000 [0.978, 1.023]
DMC	GN	FAPI+LeWM	0.10	0.320	1.003	0.683 [0.568, 0.788]
DMC	GN	LeWM	0.10	0.473	1.097	0.624 [0.518, 0.725]
DMC	PM	FAPI+LeWM	0.30	1.079	1.279	0.201 [0.126, 0.272]
DMC	PM	LeWM	0.30	1.019	1.142	0.123 [0.030, 0.212]
DMC	HM	FAPI+LeWM	0.30	0.273	1.003	0.730 [0.617, 0.832]
DMC	HM	LeWM	0.30	0.424	1.061	0.637 [0.518, 0.749]
TwoRoom	GN	FAPI+LeWM	0.10	0.397	1.174	0.777 [0.693, 0.861]
TwoRoom	GN	LeWM	0.10	0.767	1.171	0.404 [0.304, 0.505]
TwoRoom	PM	FAPI+LeWM	0.30	0.945	1.206	0.261 [0.161, 0.360]
TwoRoom	PM	LeWM	0.30	1.097	1.123	0.026 [-0.089, 0.143]
TwoRoom	HM	FAPI+LeWM	0.30	0.099	1.187	1.088 [1.017, 1.155]
TwoRoom	HM	LeWM	0.30	0.167	1.081	0.914 [0.829, 0.992]

Table 6: Complete matched-backbone robustness results under the strongest evaluated severity of each perturbation family. Success rate (%) is reported. Gain denotes the signed difference (FAPI+LeWM minus LeWM) in percentage points. Averages are computed equally across PushT, OGB-Cube, TwoRoom, and DMC.

Perturbation	Severity	Dataset	LeWM	FAPI+LeWM	Gain
Gaussian Noise	$\sigma_o = 0.10$	PushT	4.0	**16.0**	+12.0
		OGB-Cube	46.0	**52.0**	+6.0
		TwoRoom	80.0	**92.0**	+12.0
		DMC	68.0	**86.0**	+18.0
		Average	49.5	**61.5**	**+12.0**
Patch Masking	$r_p = 0.30$	PushT	2.0	**6.0**	+4.0
		OGB-Cube	42.0	**46.0**	+4.0
		TwoRoom	80.0	**84.0**	+4.0
		DMC	**30.0**	**30.0**	0.0
		Average	38.5	**41.5**	**+3.0**
History masking	$r_t = 0.30$	PushT	90.0	**96.0**	+6.0
		OGB-Cube	66.0	**72.0**	+6.0
		TwoRoom	82.0	**92.0**	+10.0
		DMC	**82.0**	**82.0**	0.0
		Average	80.0	**85.5**	**+5.5**

$\text{Success}_{\text{FAPI+LeWM}} - \text{Success}_{\text{LeWM}}$. Averages are computed equally across PushT, OGB-Cube, TwoRoom, and DMC.

Table 7: Downstream success rates of LeWM and FAPI+LeWM under the strongest evaluated context degradations. Success rate (%) is reported. Bold values indicate the higher success rate in each row; ties are bolded for both models.

Dataset	Perturbation	LeWM	FAPI+LeWM
PushT	Gaussian noise	4.0	**16.0**
	Patch masking	2.0	**6.0**
	History masking	90.0	**96.0**
OGB-Cube	Gaussian noise	46.0	**52.0**
	Patch masking	42.0	**46.0**
	History masking	66.0	**72.0**
TwoRoom	Gaussian noise	80.0	**92.0**
	Patch masking	80.0	**84.0**
	History masking	82.0	**92.0**
DMC	Gaussian noise	68.0	**86.0**
	Patch masking	**30.0**	**30.0**
	History masking	**82.0**	**82.0**

E.2 Effect of FAPI Adaptation on LeWM

Table 7 reports the success rates of LeWM and FAPI+LeWM under the strongest evaluated perturbation severities: Gaussian image noise with $\sigma_o = 0.10$, patch masking with $r_p = 0.30$, and history masking with $r_t = 0.30$. This comparison evaluates the effect of FAPI adaptation on the same world-model backbone.

FAPI+LeWM achieves higher success rates than LeWM in 10 of the 12 dataset–perturbation settings and matches it in the remaining two. Averaged across the four benchmarks, success rates increase from 49.5% to 61.5% under Gaussian noise, from 38.5% to 41.5% under patch masking, and from 80.0% to 85.5% under history masking. These correspond to gains of 12.0, 3.0, and 5.5 percentage points, respectively. The largest observed gain is on DMC under Gaussian noise, where success increases from 68.0% to 86.0%. However, success under severe patch masking remains low on PushT and unchanged on DMC, indicating that substantial visual occlusion remains challenging after adaptation.

F Complete Downstream Severity Sweeps

This section reports the complete perturbation-severity sweeps corresponding to the strongest-severity results summarized in Tables 6 and 7. Success rate (%) is reported for all settings. We compare FAPI+LeWM with LeWM and representative JEPA-based world models whenever the corresponding evaluation results are available.

F.1 Gaussian Noise

Gaussian noise is applied to input images, with the noise standard deviation σ_o varied from 0 to 0.10. Complete results are reported in Table 8.

As shown in Table 8, FAPI+LeWM outperforms LeWM across all four benchmarks at the highest noise severity and achieves particularly strong results on TwoRoom and DMC.

F.2 Patch Masking

Patch masking is evaluated by removing spatial image patches, with the masking ratio r_p varied from 0 to 0.30. The results are summarized in Table 9.

Table 8: Complete Gaussian-noise results under increasing noise severity. Success rate (%) is reported.

Dataset	Model	$\sigma_o = 0$	$\sigma_o = 0.01$	$\sigma_o = 0.03$	$\sigma_o = 0.05$	$\sigma_o = 0.10$
PushT	LeWM	96.0	94.0	82.0	12.0	4.0
	FAPI+LeWM	98.0	96.0	90.0	68.0	16.0
OGB-Cube	LeWM	74.0	72.0	70.0	68.0	46.0
	FAPI+LeWM	78.0	76.0	72.0	70.0	**52.0**
TwoRoom	LeWM	88.0	86.0	82.0	80.0	80.0
	FAPI+LeWM	96.0	94.0	94.0	92.0	**92.0**
DMC	LeWM	86.0	78.0	76.0	68.0	68.0
	FAPI+LeWM	88.0	88.0	88.0	86.0	**86.0**

Table 9: Complete patch-masking results under increasing masking ratios. Success rate (%) is reported.

Dataset	Model	$r_p = 0$	$r_p = 0.05$	$r_p = 0.10$	$r_p = 0.20$	$r_p = 0.30$
PushT	LeWM	96.0	16.0	4.0	4.0	2.0
	FAPI+LeWM	98.0	20.0	10.0	6.0	**6.0**
OGB-Cube	LeWM	74.0	66.0	52.0	40.0	42.0
	FAPI+LeWM	78.0	70.0	54.0	46.0	**46.0**
TwoRoom	LeWM	88.0	84.0	82.0	82.0	80.0
	FAPI+LeWM	96.0	94.0	88.0	84.0	**84.0**
DMC	LeWM	86.0	72.0	54.0	30.0	**30.0**
	FAPI+LeWM	88.0	72.0	54.0	30.0	**30.0**

Patch masking substantially degrades performance on PushT and DMC, indicating a strong dependence on localized visual evidence. FAPI+LeWM outperforms LeWM on PushT and OGB-Cube and matches it on DMC at the strongest masking ratio. These results indicate that removing localized visual information is more challenging than global visual corruption.

F.3 History Masking

History masking is applied to the observation-history context, with the masking ratio r_t varied from 0 to 0.30. Complete results are reported in Table 10.

As shown in Table 10, FAPI+LeWM maintains competitive performance under increasing history masking. At $r_t = 0.30$, it achieves 96.0% success on PushT, compared with 90.0% for LeWM.

G Implementation of Out-of-Distribution Corruptions

We evaluate three held-out context-degradation operators on TwoRoom: Gaussian blur, block cutout, and context repetition. These operators are not used during FAPI adaptation. Each operator is evaluated independently at four severity levels, with all other perturbations disabled. LeWM and FAPI+LeWM use the same 50 initial states and random seed 42. The checkpoint of each model is held fixed across all operators and severity levels, without additional training.

Table 11 lists the evaluated parameters for 224×224 input observations. Levels L_1–L_4 represent increasing severity within each operator; they do not imply equivalent difficulty across different operators.

Gaussian blur. Gaussian blur is applied to context observations before visual encoding using channel-wise convolution with a $k \times k$ Gaussian kernel and standard deviation σ_b. The application probability is fixed at 1.0.

Table 10: Complete history-masking results under increasing masking ratios. Success rate (%) is reported.

Dataset	Model	$r_t = 0$	$r_t = 0.05$	$r_t = 0.10$	$r_t = 0.20$	$r_t = 0.30$
PushT	LeWM	96.0	92.0	92.0	90.0	90.0
	FAPI+LeWM	98.0	98.0	98.0	98.0	**96.0**
OGB-Cube	LeWM	74.0	70.0	70.0	66.0	66.0
	FAPI+LeWM	78.0	76.0	74.0	74.0	**72.0**
TwoRoom	LeWM	88.0	84.0	82.0	82.0	82.0
	FAPI+LeWM	96.0	94.0	92.0	92.0	**92.0**
DMC	LeWM	86.0	84.0	82.0	82.0	**82.0**
	FAPI+LeWM	88.0	88.0	86.0	84.0	**82.0**

Table 11: Severity settings for out-of-distribution context degradations. Cutout percentages indicate the approximate fraction of image area covered by one square block.

Level	Blur (k, σ_b)	Cutout block size	Cutout area	Repetition p_r
L_1	$(3, 0.5)$	50×50	5.0%	0.05
L_2	$(5, 1.0)$	71×71	10.0%	0.10
L_3	$(9, 2.0)$	100×100	19.9%	0.20
L_4	$(13, 3.0)$	123×123	30.2%	0.30

Block cutout. Block cutout is applied before visual encoding and fills one randomly positioned square region in each context frame with zeros. The application probability is 1.0, and block locations are sampled independently across frames. In addition to the four-level sweep, we retain the original 32×32 cutout setting (approximately 2.0% of the image area) as a separate reference. This additional setting is excluded from averages over L1–L4.

Context repetition. Context repetition is applied after encoding. With probability p_r, an eligible intermediate context latent and its corresponding action embedding are replaced with their preceding values. The first and latest context steps are preserved. Replacements are taken from an unmodified copy of the original history to avoid recursive repetition.

Only the context pathway is perturbed; the target observation and target-encoder pathway remain clean. These operators are excluded from adaptation of the checkpoints evaluated in Table 2; separate models trained with these operators are considered only in the ablation study.

H Definitions of Ablation Variants

This section defines the implementation variants evaluated in Figure 2. Panels (a)–(c) examine alternative visual-noise, visual-occlusion, and history-degradation designs used during adaptation. Panel (d) examines component removal. The variant names describe adaptation procedures, whereas Clean and Full stress denote evaluation conditions.

Visual-noise variants. **Gaussian KL** applies Gaussian noise to input images and introduces an auxiliary Kullback–Leibler (KL) divergence regularizer between noisy and clean representations, with the clean branch serving as a stop-gradient reference. The noisy representations are used only for this auxiliary regularizer and are not passed to the predictor. In contrast, **Gaussian Context** feeds representations of noise-corrupted context images directly into the predictor, while retaining clean prediction targets. **Blur Context** replaces additive Gaussian noise with Gaussian blurring of context images before encoding and prediction, again retaining clean targets. Thus, Gaussian KL tests auxiliary representation consistency, whereas Gaussian Context and Blur Context test prediction from degraded visual context.

Visual-occlusion variants. **Random Patch** independently samples spatial patch masks for each context frame, allowing occluded locations to vary across frames. **Tube Patch** shares the same spatial mask across all context frames within a sample, producing persistent occlusion at fixed image locations. **Block Cutout** masks a contiguous image region rather than spatially dispersed patches. These variants differ in the spatial structure and cross-frame persistence of the occlusion.

History-degradation variants. **Random Mask**, referred to as random history masking, randomly masks individual preceding-frame representations while retaining the latest frame. **Block Mask**, referred to as block history masking, masks a contiguous segment of the preceding history while retaining the latest frame. **Frame Repeat** replaces selected historical representations with those from the preceding frame, simulating stale observations. These variants examine scattered missing observations, consecutive missing observations, and repeated observations, respectively.

Component-removal variants. **LeWM** denotes the base LeWorldModel, and **Full** denotes the complete `FAPI+LeWM` adaptation configuration. **w/o Noise**, **w/o Patch**, and **w/o History** denote removal of the visual-noise, patch-masking, and history-degradation components, respectively, during adaptation. These labels describe which adaptation component is removed, rather than which corruption is omitted during evaluation.

I V-JEPA Robustness Across Combined-Degradation Severities

We compare Frozen V-JEPA, Fine-tuned V-JEPA, and `FAPI+V-JEPA` under jointly applied Gaussian video noise and history masking. Table 12 specifies the parameter pair used at each of five severity levels. L_0 denotes the clean condition, while L_4 denotes the strongest evaluated combination.

Table 12: Parameters for the V-JEPA combined-degradation severity sweep. Gaussian video noise and history masking are applied simultaneously at each nonzero level.

Perturbation	Parameter	L_0	L_1	L_2	L_3	L_4
Gaussian video noise	σ_o	0.00	0.02	0.05	0.10	0.15
History masking	p_t	0.00	0.05	0.10	0.20	0.30

Table 13: V-JEPA prediction loss under combined Gaussian video noise and history masking, using the severity levels in Table 12. Lower values are better; bold indicates the lowest loss at each level.

Model	L_0	L_1	L_2	L_3	L_4
Frozen V-JEPA	0.546	0.558	0.573	0.586	0.596
Fine-tuned V-JEPA	0.588	0.590	0.593	0.598	0.601
`FAPI+V-JEPA`	**0.485**	**0.484**	**0.485**	**0.486**	**0.486**

`FAPI+V-JEPA` achieves the lowest prediction loss at every severity level and remains stable as combined degradation increases. As shown in Table 13, its loss increases from 0.485 at L_0 to 0.486 at L_4, compared with 0.546 to 0.596 for Frozen V-JEPA and 0.588 to 0.601 for Fine-tuned V-JEPA. The corresponding increases are 0.001, 0.050, and 0.013, respectively. These results support improved robustness to combined visual and temporal context degradation beyond standard fine-tuning, alongside lower clean-context prediction loss.

J I-JEPA Gaussian-Noise Severity-Sweep Evaluation

We evaluate frozen I-JEPA, fine-tuned I-JEPA, and `FAPI+I-JEPA` under progressively increasing Gaussian image noise. This analysis extends the fixed-severity visual-corruption comparison in Table 4 by examining how prediction loss changes as visual context degradation becomes more severe. Noise is applied only to the context image, while the target image and target-encoder pathway remain clean.

Corruption-induced degradation is measured as the increase in prediction loss between the clean condition and the maximum evaluated Gaussian-noise severity:

$$\Delta_{\text{deg}} = \mathcal{L}_{\max} - \mathcal{L}_{\text{base}}, \tag{12}$$

where $\mathcal{L}_{\text{base}}$ and $\mathcal{L}_{\max}$ denote the prediction losses under clean and maximum-noise conditions, respectively.

All loss differences and relative reductions below are calculated from values before rounding.

The relative reduction in maximum-severity loss achieved by `FAPI+I-JEPA` over frozen I-JEPA is defined as

$$R_{\max} = \frac{\mathcal{L}_{\max}^{\text{Official}} - \mathcal{L}_{\max}^{\text{FAPI}}}{\mathcal{L}_{\max}^{\text{Official}}} \times 100\%. \tag{13}$$

The corresponding degradation reduction is

$$R_{\text{deg}} = \frac{\Delta_{\text{deg}}^{\text{Official}} - \Delta_{\text{deg}}^{\text{FAPI}}}{\Delta_{\text{deg}}^{\text{Official}}} \times 100\%. \tag{14}$$

Gaussian noise sampled from $\mathcal{N}(0, \sigma_o^2)$ is added to the context image, while the target pathway remains clean. The noise standard deviation is varied over $\sigma_o \in \{0.00, 0.05, 0.10, 0.20\}$.

Table 14: I-JEPA prediction loss under increasing Gaussian image-noise severity. Column headers denote the noise standard deviation σ_o. Lower values are better.

Model	0.00	0.05	0.10	0.20
Frozen I-JEPA	0.191	0.196	0.207	0.223
Fine-tuned I-JEPA	0.193	0.197	0.206	0.223
`FAPI+I-JEPA`	**0.187**	**0.190**	**0.197**	**0.212**

As shown in Table 14, `FAPI+I-JEPA` achieves the lowest prediction loss at every noise level. At the maximum severity of $\sigma_o = 0.20$, its loss is 0.212, compared with 0.223 for frozen I-JEPA and 0.223 for Standard I-JEPA. Relative to frozen I-JEPA, this corresponds to a 4.9% reduction in maximum-severity prediction loss.

Across the complete sweep, the prediction loss of frozen I-JEPA increases by 0.032, from 0.191 to 0.223. In contrast, `FAPI+I-JEPA` increases by only 0.025, from 0.187 to 0.212, corresponding to a 22.7% reduction in noise-induced degradation. Fine-tuned I-JEPA remains close to the official checkpoint under strong noise, indicating that continued clean adaptation alone provides limited improvement in predictive stability.

K Proof for the Proposition 3.1

Proof. Write $Y = z_{t+1}^{\star}$ and $\widehat{Y}_{\xi} = F_{\Theta}(\widetilde{Z}_t^{(H),\xi}, A_t^{(H)})$. Because both predictions use the same target Y,

$$\begin{aligned}\Delta_{\Theta}(q_t; \xi, \xi') &= \frac{1}{d}\big\|(\widehat{Y}_{\xi} - Y) - (\widehat{Y}_{\xi'} - Y)\big\|_2^2 \\ &\le \frac{2}{d}\|\widehat{Y}_{\xi} - Y\|_2^2 + \frac{2}{d}\|\widehat{Y}_{\xi'} - Y\|_2^2. \end{aligned} \tag{15}$$

Taking expectations, and using the fact that ξ and ξ' have the same marginal distribution, gives $\mathbb{E}_{q_t,\xi,\xi'}[\Delta_{\Theta}(q_t; \xi, \xi')] \le 4\mathcal{L}_{\text{FAPI}}(\Theta)$.

For the second claim, fix ξ and let $X_\xi = (\widetilde{Z}_t^{(H),\xi}, A_t^{(H)})$ and $m_\xi(X_\xi) = \mathbb{E}[Y \mid X_\xi]$. The orthogonality property of conditional expectation gives

$$\mathbb{E}_{q_t}\left[\frac{1}{d}\|F_\Theta(X_\xi) - Y\|_2^2\right] = \mathbb{E}_{q_t}\left[\frac{1}{d}\|F_\Theta(X_\xi) - m_\xi(X_\xi)\|_2^2\right] + \frac{1}{d}\mathbb{E}_{q_t}[\operatorname{tr}\operatorname{Cov}(Y \mid X_\xi)]\,. \tag{16}$$

The second term is σ_ξ^2. Averaging over ξ therefore yields

$$\mathcal{L}_{\mathrm{FAPI}}(\Theta) = \mathbb{E}_\xi[\sigma_\xi^2] + \mathbb{E}_{\xi,q_t}\left[\frac{1}{d}\|F_\Theta(X_\xi) - m_\xi(X_\xi)\|_2^2\right]. \tag{17}$$

The remaining term is nonnegative and vanishes if and only if $F_\Theta(X_\xi) = \mathbb{E}[Y \mid X_\xi]$ almost surely for $p(\xi)$-almost every ξ. □