CALIBWM: ACTION-FIDELITY CALIBRATED UNCERTAINTY FOR FROZEN VIDEO WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Action-conditioned video world models support robot policy evaluation and imagination-based planning, but their training and evaluation focus largely on visual fidelity. Recent benchmarks show that photorealistic rollouts can misrepresent the effects of executed actions and depict failures as successes. Uncertainty estimates calibrated to visual correctness inherit this mismatch. We introduce CalibWM, a lightweight probe on a frozen action-conditioned video world model that produces a spatiotemporal uncertainty map aligned with action-fidelity failure, defined by whether a frozen inverse dynamics model can recover the executed action from generated frames. A quantile objective regularizes uncertainty to grow with rollout horizon, while a spatial term localizes the map and an alignment term trains its readout to predict action-fidelity failure. Per-horizon, segment-level split conformal calibration provides finite-sample coverage for aggregated normalized prediction-error scores under exchangeability, with unit-level implications determined by the spatial aggregator. On Bridge and DROID, CalibWM achieves AUROCs of 0.782 and 0.761, respectively, compared with 0.694 and 0.679 for the strongest baseline trained without the action-fidelity label. Its margins over the strongest control trained with that label are $+0.020$ (clustered 95% CI $[0.009, 0.031]$) on Bridge and $+0.017$ ($[0.006, 0.028]$) on DROID. On a human-annotated optimism-bias subset with uniformly high visual quality and rare failures, CalibWM achieves 0.634 AUPRC against a chance level of 0.09; an oracle visual proxy with access to the ground-truth future reaches 0.106. We establish the coverage result at each fixed horizon. Under ordering consistency, we also prove that the action-fidelity readout has AUROC no lower than a visual proxy computed from the generated rollout alone, provided that the proxy carries no additional information about action-fidelity failure given the probe's feature context.

1 INTRODUCTION

Reliable rollouts must preserve action outcomes. Action-conditioned video world models support robot policy evaluation and imagination-based planning. These applications require generated videos to preserve the consequences of executed actions, including failed grasps and misplaced objects. Visual realism alone does not establish this correspondence. MiraBench reports that post-training a 2B model on robot demonstrations raises task completion from 14.0 to 92.0 while reducing failure preservation from 48.7 to 23.1 (Yang et al., 2026). A rollout can therefore look successful while misrepresenting an action's outcome, motivating uncertainty estimates that respond to action-fidelity failures.

Visual confidence leaves an action-fidelity gap. Existing methods estimate ensemble disagreement, uncertainty for safety filtering, or uncertainty in video generation (Wan et al., 2026; Seo et al., 2025; Mei et al., 2025a). The closest system, C^3 (Mei et al., 2025b), is trained to predict whether latent prediction errors fall below a threshold, yielding spatiotemporal visual confidence. Its supervision does not directly identify whether generated transitions preserve the executed action. Its objective also leaves horizon-dependent uncertainty growth unconstrained, and pooled coverage does not characterize reliability at individual horizons. These observations motivate coupling action-fidelity supervision with an uncertainty map that supports separate calibration at each horizon.

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.

CalibWM aligns uncertainty with action fidelity. We attach a lightweight probe to a frozen video world model and train its readout to detect action-fidelity failure, defined through a frozen inverse dynamics model's ability to recover the executed action from generated frames. The probe learns a spatial error scale with horizon-aware regularization, while action-alignment supervision shapes its failure ranking. A separate segment-level conformal layer calibrates normalized prediction-error scores at each fixed horizon. This design provides an action-fidelity readout and error regions from the same uncertainty map, with auxiliary attribution learned from cached teachers.

Our contributions are:

- **Action-aligned uncertainty on frozen features.** We introduce a spatiotemporal probe that combines action-fidelity supervision, horizon-aware error estimation, and spatial localization, using fewer than 1% of the frozen backbone's parameters (§3).
- **Fixed-horizon calibration and conditional ranking guarantees.** We establish segment-level coverage for aggregated normalized prediction-error scores under exchangeability (Proposition 1). We also show that an order-consistent readout has AUROC no lower than a generated-rollout visual proxy that carries no additional failure information conditional on the probe features (Proposition 2).
- **Discrimination under matched supervision.** CalibWM improves AUROC over the strongest same-label control by 0.020 on Bridge and 0.017 on DROID (Table 1). On the human-annotated optimism subset, it reaches 0.634 AUPRC, compared with 0.106 for LPIPS with access to the ground-truth future (Figure 2c).

2 Problem formulation

Frozen world model and rollout context. Let $p_\theta(o_{t+1} \mid o_{\le t}, a_{\le t})$ be a pretrained action-conditioned video world model whose parameters θ remain frozen. Given context $(o_{\le t_0}, a_{\le t_0})$ and a future action sequence $a_{t_0:t_0+H-1}$, the model autoregressively produces $\hat{o}_{t_0+h} \sim p_\theta(\cdot \mid \hat{o}_{<t_0+h}, a_{<t_0+h})$ for $h \in \{1, \dots, H\}$.

Action-fidelity failure. Let π_{IDM} be a frozen inverse dynamics model that reads a fixed-length window of k observations, real or generated, and infers the action that produced its final transition (Baker et al., 2022); $k = 3$ by default, long enough to disambiguate velocity. Under the indexing above, the frame at time t is the consequence of a_{t-1}, so that is the action the window ending at t carries. Given a metric $d_{\mathcal{A}}$ on the action space and a threshold τ_a, the *action-fidelity failure* label at horizon h, with $t = t_0 + h$, is

$$\text{AF-fail}_h \;=\; \mathbf{1}\big[\, d_{\mathcal{A}}\big(\pi_{\mathrm{IDM}}(\hat{o}_{t-k+1:t}),\, a^\star_{t-1}\big) \;>\; \tau_a \,\big]\,. \tag{1}$$

For $h < k$, the window includes real context frames so that the label evaluates action consistency in its generated portion rather than reconstruction of the conditioning frames. Our target is a readout that ranks generated windows by this action-fidelity failure label.

Error scale and failure readout. Observations are encoded to latents $x_t = \mathrm{Enc}(o_t)$, with scalar spatial-channel units $u = (x, y, c) \in \Omega$ and $\hat{x}_h(u) = \mathrm{Enc}(\hat{o}_{t_0+h})(u)$. Writing $x^\star_h$ for the latent of the true future frame, the prediction error magnitude is $\epsilon_h(u) = |\hat{x}_h(u) - x^\star_h(u)|$. The uncertainty map $\hat{U}_h(u)$ is anchored at a conditional quantile of this magnitude:

$$\hat{U}_h(u) \;\approx\; \mathrm{Quantile}_q\big[\epsilon_h(u) \,\big|\, \Phi_h(u)\big], \qquad \Phi_h(u) = \big(z_h(u),\, e(a_{t_0+h-1}),\, \gamma(h)\big), \tag{2}$$

where $z_h(u) \in \mathbb{R}^D$ is the frozen backbone's penultimate per-token feature, $e(\cdot)$ an action embedding of a_{t_0+h-1}, the last action the frame at horizon h is conditioned on, and $\gamma(\cdot)$ a sinusoidal horizon encoding (Vaswani et al., 2017). The pinball term targets Eq. (2); the spatial and alignment terms in §3.1–3.2 further shape $\hat{U}$. The deployed output is therefore a conditional scale anchored at the q-quantile, with no requirement that it equal that quantile. The conformal layer accepts any fixed $\hat{U}$. Since $\hat{U}_h$ and ϵ_h have the same units, the nonconformity score $s_h = \epsilon_h / (\hat{U}_h + \delta)$ is dimensionless. The epistemic and aleatoric components attribute this same scalar on a common scale.

The visual-correctness label used by C^3, $\mathbf{1}[|\hat{x} - x^\star| \le \varepsilon]$ (Mei et al., 2025b), can disagree with AF-fail_h: a frame may reproduce the background accurately but mistime finger-object contact, causing π_{IDM} to recover the wrong action; a blurry frame may preserve the action semantics.

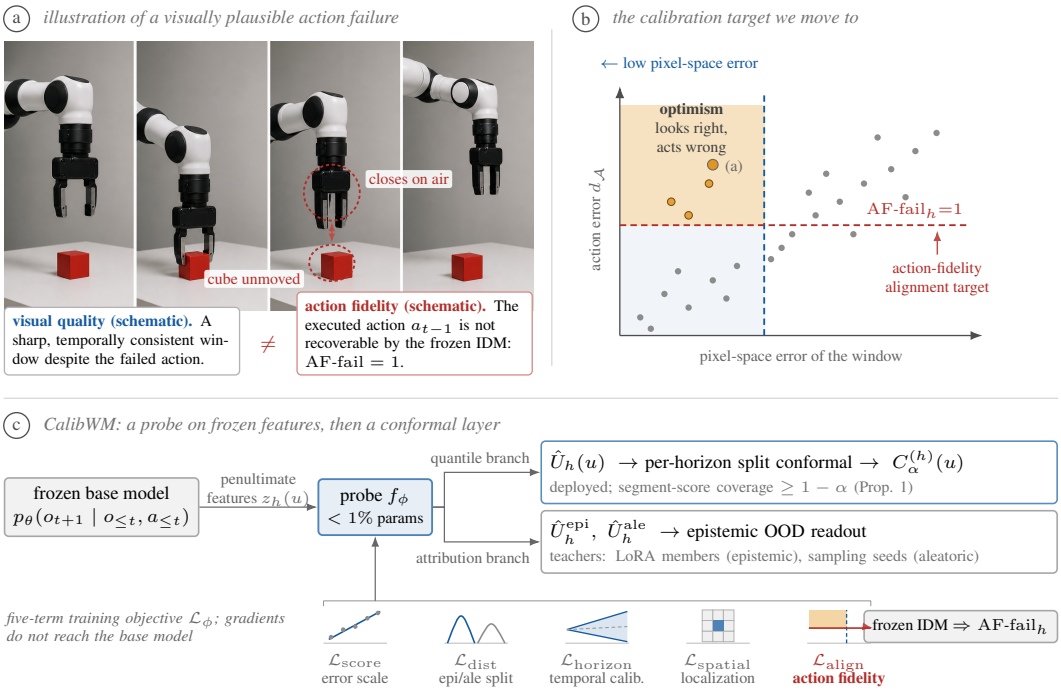

Figure 1: Action-fidelity failure and the CalibWM architecture. **(a)** A rendered illustration, not a model sample: the gripper closes on air while the cube remains stationary, depicting a visually clear rollout from which a frozen IDM cannot recover the executed action. **(b)** Blue and red boundaries mark visual-error and action-fidelity thresholds; the amber quadrant contains low-visual-error, high-action-error windows. **(c)** A probe with fewer than 1% of the frozen backbone's parameters reads penultimate features. Its quantile branch is deployed and conformalized at segment level per horizon under Assumption 1; its attribution branch uses two teachers. The red alignment term supervises the deployed readout with the IDM's action-fidelity label.

3 CalibWM

CalibWM trains an action-fidelity readout on a spatial uncertainty map from frozen features (Figure 1). Action supervision shapes its failure ranking (§3.1); horizon and spatial objectives learn its error scale (§3.2); segment-level conformal calibration provides fixed-horizon prediction-error regions (§3.3).

3.1 Frozen-feature probe and action-fidelity readout

The backbone is frozen and its features are detached, so no gradient reaches θ. For each unit u and horizon h the probe f_ϕ consumes $\Phi_h(u)$ and emits two branches with distinct responsibilities:

$$\underbrace{\hat{U}_h(u) = f_\phi^{\text{q}}(\Phi_h(u))}_{\text{quantile branch: deployed, conformalized}} \quad , \quad \underbrace{\big(\hat{U}_h^{\text{epi}}(u), \hat{U}_h^{\text{ale}}(u)\big) = f_\phi^{\text{d}}(\Phi_h(u))}_{\text{attribution branch: distilled only}} . \tag{3}$$

The quantile branch learns $\hat{U}_h$ without reconstructing it from the attribution components; only this branch is conformalized and deployed. The attribution branch receives scoring and distillation supervision and influences $\hat{U}$ only through the shared trunk, as an auxiliary task (Appendix C).

The probe is a 2-layer MLP with hidden width 512, shared weights across tokens, and non-negative softplus outputs in the units of ϵ. Its roughly 0.8M parameters are below 1% of the 154M-parameter denoiser; Appendix E compares a 2-layer transformer with $d_{\text{model}} = 384$. An upsampling decoder gives the pixel heatmap $\hat{U}_h^{\text{px}}(x, y)$, and object masks $\{\mathcal{M}_o\}$ define $U_{\text{obj}}(o, h) = \max_{u \in \mathcal{M}_o} \hat{U}_h(u)$.

Action-fidelity alignment. The alignment term supervises the deployed uncertainty readout against action-fidelity failure. Let g be a spatial or object-level aggregator. We use an object max by default because contact failures are typically driven by a single object. The readout is $r = \sigma(w\, g(\hat{U}_h) + b)$, with learned $w > 0$ and b, and its objective combines a proper binary loss with a pairwise ranking surrogate for AUROC:

$$\mathcal{L}_{\text{align}} = \text{BCE}\big(r,\ \text{AF-fail}_h\big) + \zeta\, \mathbb{E}_{(i^+,i^-)}\Big[\text{softplus}\big(g(\hat{U}^{(i^-)}) - g(\hat{U}^{(i^+)})\big)\Big], \tag{4}$$

Here (i^+, i^-) ranges over pairs of failing and non-failing samples. Because $\hat{U}$ is non-negative and g is a max, $\sigma(g(\hat{U}_h))$ cannot produce probabilities below $\frac{1}{2}$; the learned affine layer makes these probabilities attainable. The terms in §3.2 fit error magnitudes, while Eq. (4) relates the map to AF-fail_h. Under the weak-label route, the BCE target is the inverse dynamics model's exceedance probability p, weighted by its certainty $|2p - 1|$ to reduce the contribution of ambiguous frames and their label noise.

3.2 Learning horizon-aware spatial uncertainty

Only ϕ is optimized. The total loss is

$$\mathcal{L}_\phi = \mathcal{L}_{\text{score}} + \lambda_{\text{dist}}\mathcal{L}_{\text{dist}} + \lambda_h\mathcal{L}_{\text{horizon}} + \lambda_s\mathcal{L}_{\text{spatial}} + \lambda_a\mathcal{L}_{\text{align}}, \tag{5}$$

The horizon and spatial terms shape the deployed map; scoring and distillation supervise the auxiliary attribution branch, while alignment trains the readout in §3.1.

Horizon calibration. With the pinball loss $\rho_q(r) = \max\{qr, (q-1)r\}$ (Koenker & Bassett, 1978), the quantile branch is trained directly to be the q-quantile of ϵ_h, with a monotone-in-horizon regularizer:

$$\mathcal{L}_{\text{horizon}} = \frac{1}{|\mathcal{B}|}\sum \rho_q\big(\epsilon_h(u) - \hat{U}_h(u)\big) \ + \ \eta \cdot \frac{1}{H-1}\sum_{h=1}^{H-1} \mathbb{E}_u\Big[\text{ReLU}\big(\hat{U}_h(u) - \hat{U}_{h+1}(u)\big)\Big]. \tag{6}$$

A quantile rather than a mean is essential here. Rollout errors are right-skewed, since most frames are accurate and a few optimistic frames are badly wrong, so mean regression predicts far below the tail and produces exactly the overconfidence on dangerous frames we are trying to remove.

Spatial localization. To keep $\hat{U}$ from degenerating into a spatially uniform scalar we regress the pixel heatmap onto the pixel-space error with a Huber loss, and add a total variation penalty (Rudin et al., 1992) that encourages piecewise-smooth maps with sharp boundaries:

$$\mathcal{L}_{\text{spatial}} = \frac{1}{|\mathcal{B}|}\sum \text{Huber}_\beta\big(\hat{U}_h^{\text{px}} - \epsilon_h^{\text{px}}\big) + \eta_{\text{tv}} \sum_{x,y}\Big(\big|\partial_x \hat{U}_h^{\text{px}}\big| + \big|\partial_y \hat{U}_h^{\text{px}}\big|\Big). \tag{7}$$

The heatmap is indexed by pixel position, so its target aggregates the colour channels, $\epsilon_h^{\text{px}}(x,y) = \|\hat{o}_h(x,y,\cdot) - o_h^\star(x,y,\cdot)\|$, and it is computed directly in pixel space rather than by decoding the latent error, because the decoder does not preserve error magnitude.

Offline teachers and single-pass deployment. Unhatted U denotes a sampled teacher quantity and hatted U its prediction. Teachers are cached offline from $K = 4$ parameter-perturbed members and $L = 8$ diffusion seeds per member, separating parameter variation from sampling noise. Appendix E.1 gives the variance decomposition and finite-sampling corrections. Deployment needs one forward pass without sampling, at the inference cost of a pixel-confidence head; Appendix E gives the caching details.

Proper scoring on a non-negative magnitude. For the shifted error magnitude $\epsilon_h(u) + \delta$, we use a log-normal likelihood with mean $\hat{U}_h^{\text{ale}}(u) + \delta$ and fixed shape ς:

$$\mathcal{L}_{\text{score}} = \frac{1}{|\mathcal{B}|}\sum \left[\frac{1}{2\varsigma^2}\Big(\log\tfrac{\epsilon_h(u)+\delta}{\hat{U}_h^{\text{ale}}(u)+\delta} + \tfrac{\varsigma^2}{2}\Big)^2 + \log(\epsilon_h(u)+\delta) + \log\varsigma + \tfrac{1}{2}\log 2\pi\right]. \tag{8}$$

The aleatoric teacher in Eq. (17) is rescaled to the same mean-magnitude target. We fix ς; Appendix E.4 explains this parameterization and Appendix E reports its sensitivity.

Attribution distillation. $\mathcal{L}_{\text{dist}} = \text{SmoothL1}(\hat{U}_h^{\text{epi}}, U_h^{\text{epi}}) + \text{SmoothL1}(\hat{U}_h^{\text{ale}}, U_h^{\text{ale}})$ against the two rescaled teachers of Eq. (17). This is the only supervision $\hat{U}^{\text{epi}}$ receives.

3.3 Segment-level conformal calibration

Given a nonconformity score S and a calibration set of n exchangeable draws, split conformal prediction (Papadopoulos et al., 2002; Vovk et al., 2005; Lei et al., 2018) takes $\hat{Q}_{1-\alpha}$ as the $\lceil (1-\alpha)(n+1) \rceil$-th smallest calibration score and guarantees $\mathbb{P}(S^{\text{test}} \le \hat{Q}_{1-\alpha}) \ge 1-\alpha$, with a matching upper bound $1-\alpha+\frac{1}{n+1}$ when the $n+1$ scores are almost surely distinct (Romano et al., 2019). This distribution-free guarantee requires exchangeability. Treating temporally dependent frames as exchangeable calibration samples can violate that requirement. The segment sampling and horizon stratification below are designed to support it.

One calibration score per segment. We add a split conformal layer to the learned error scale, using rollout segments as the statistical units.

The unit-level score is $s_h(u) = \epsilon_h(u)/(\hat{U}_h(u)+\delta)$. Each segment contributes one scalar at each horizon, $S_h = \Psi(\{s_h(u)\}_{u\in\Omega})$, with $\Psi = \text{Quantile}_{0.95}$ by default or $\max_{u\in\Omega}$. Treating the tens of thousands of correlated units within a frame as independent samples would overstate the number of independent calibration samples and yield intervals that are too narrow. At horizon h, the threshold is the $\lceil (1-\alpha)(n_h+1) \rceil$-th smallest of the n_h calibration scores $\{S_h^{(i)}\}$, giving the region below.

$$C_\alpha^{(h)}(u) = \big[\hat{x}_h(u) \pm \hat{Q}_{1-\alpha}^{(h)}\big(\hat{U}_h(u)+\delta\big)\big]. \tag{9}$$

The aggregator Ψ constructs coverage regions from normalized conformal scores. The aggregator g in Eq. (4) constructs the action-fidelity readout from the raw uncertainty map. These outputs answer distinct reliability questions. E1 asks whether error regions reach their nominal coverage, while E2 asks whether r ranks action-fidelity failures above non-failures. The conformal threshold is fitted from prediction-error scores without action-fidelity labels and changes the region in Eq. (9). The readout uses the raw uncertainty map and receives action-fidelity supervision during probe training. We therefore assess coverage and discrimination separately, using the conformal-layer ablation to distinguish the effect of calibrating error regions from learning the action-fidelity ranking (§4.2).

Horizon stratification compares segment scores at the same rollout horizon. Exchangeability within each stratum remains an assumption about the sampled trajectories and score function; it is not created by stratification alone. For sequential rollouts where this assumption fails, adaptive conformal inference (Gibbs & Candes, 2021) and sequential predictive conformal inference (Xu & Xie, 2023) are the standard fallbacks; both trade the fixed-horizon finite-sample bound for a long-run time-average or an asymptotic conditional guarantee, and we report CalibWM under the latter alongside the stratified version.

3.4 Coverage and ranking guarantees

We establish per-horizon marginal coverage and AUROC dominance under the assumptions below; formal statements and proofs appear in Appendix A. For coverage, Assumption 1 requires the calibration and test segment scores to be jointly exchangeable at a fixed h. Our protocol draws segments from mutually independent trajectories instead of adjacent slices of one long rollout and fits f_ϕ on a split disjoint from $\mathcal{D}_{\text{cal}}$.

Proposition 1 (Per-horizon finite-sample marginal coverage)**.** *Let $n_h \ge (1-\alpha)/\alpha$, so that $\lceil (1-\alpha)(n_h+1) \rceil \le n_h$ and the threshold is finite. Under Assumption 1, for each fixed h, the split conformal threshold $\hat{Q}_{1-\alpha}^{(h)}$ satisfies $\mathbb{P}\big(S_h^{\text{test}} \le \hat{Q}_{1-\alpha}^{(h)}\big) \ge 1-\alpha$. If in addition the n_h+1 scores are almost surely distinct, then $\mathbb{P}\big(S_h^{\text{test}} \le \hat{Q}_{1-\alpha}^{(h)}\big) \le 1-\alpha+\frac{1}{n_h+1}$.*

Proposition 1 covers a segment score. The aggregator Ψ determines the implication for individual units without requiring a particular dependence structure among them. With $\Psi = \max_u$, the event contains every unit simultaneously, at the cost of wider intervals. The default $\Psi = \text{Quantile}_{0.95}$

gives a tolerance guarantee and implies neither simultaneous nor per-unit marginal coverage. Separately calibrated thresholds also preserve the bound under a mixture of test horizons selected independently of the calibration draws, which supports the pooled coverage results. The bound does not extend to a single threshold fitted to a bucket of horizons or to distribution shift. Remark 1 gives the details; those settings are reported without a finite-sample guarantee.

The second result compares action-fidelity ranking with a visual proxy. Write $Y := \text{AF-fail}$, let Φ be the readout's feature context, and let $r^\star(\Phi) = \mathbb{P}(Y = 1 \mid \Phi)$. Assumption 2 requires the deployed readout to preserve the ranking of $r^\star$, which suffices for AUROC without requiring $r = r^\star$. Since r is a monotone function of the scalar $g(\hat{U}_h)$, this condition requires that scalar to retain the information needed for $r^\star$. Assumption 3 requires the quality proxy q to be a garbling of Φ, that is, $q \perp Y \mid \Phi$.

Proposition 2 (AUROC dominance under sufficiency)**.** *Under Assumptions 2 and 3,* $\mathrm{AUROC}(r \to Y) \geq \mathrm{AUROC}(q \to Y)$. *Equality holds whenever q is almost surely a strictly* increasing *transformation of $r^\star(\Phi)$.*

The proof combines Neyman–Pearson optimality of posterior ranking (Neyman & Pearson, 1992) with Blackwell's comparison of experiments (Blackwell, 1953; Torgersen, 1991). Its scope is visual proxies computable from the generated rollout alone that satisfy Assumption 3, including the label-free C^3 subpatch-confidence head (Remark 2). Oracle PSNR, SSIM, and LPIPS use $x^\star$, which is absent from Φ; the proposition provides no guarantee for the rows marked † in Table 1. Those comparisons are empirical. We evaluate the nonnegative AUROC ordering against label-free C^3 using a clustered 95% interval for ΔAUROC, with an interval strictly below zero as the specified falsification criterion. A separate interval against the strongest label-using control measures the gain beyond access to the action-fidelity label.

4 Experiments

4.1 Setup

Backbone and datasets. All experiments use C^3's action-conditioned latent DiT and Stable Diffusion VAE tokenizer (Mei et al., 2025b), giving both methods identical frozen features. Transfer to other backbones such as Cosmos (NVIDIA et al., 2025) is untested. We use BridgeData V2 (Walke et al., 2023) and DROID (Khazatsky et al., 2024) through Open X-Embodiment (O'Neill et al., 2024) in RLDS format (Ramos et al., 2021).

Trajectory-level splits and calibration. We hash *trajectory* identifiers to assign splits and draw at most one segment per calibration trajectory to remove within-rollout dependence (Assumption 1). Identical distribution across trajectories remains an assumption, checked by comparing score distributions across task clusters. We use $H = 16$, $\alpha = 0.1$, and 500 calibration segments per (domain, horizon) stratum. Under the i.i.d., continuous-score reference model, conditional coverage has standard deviation near 0.013 across calibration draws (Remark 3). Calibration uses the full cache because it needs only ϵ_h; evaluation of AF-fail uses the smaller replay-labelled subset.

Training and evaluation label sources. Weak labels from Eq. (1) train the probe on every cached rollout; their reliability can change under the IDM's distribution shift on generated frames. Replay labels support the primary evaluation: because Bridge and DROID lack task outcomes, we execute $a^\star_{t:t+H}$ in reconstructed scenes and evaluate the task's success predicate, yielding $2{,}140$ Bridge and $1{,}875$ DROID segments for E2 (Appendix C.3). Cohen's κ measures agreement between weak and replay labels. MiraBench's human annotations (Yang et al., 2026) validate τ_a and support a separate benchmark on its own generators' videos; they are not assigned to our cache. We fix τ_a in distribution and retain it on every shifted cache.

Matched baselines and label controls. Methods use identical test segments and output nonnegative uncertainty per segment, horizon, and spatial position (Appendix C). Oracle proxies (†) use the ground-truth future: negated PSNR and SSIM and unmodified LPIPS assign larger scores to greater visual error (Zhang et al., 2018; Wang et al., 2004). E2 evaluates their action-fidelity ranking; corpus-level FVD (Unterthiner et al., 2018) cannot rank segments. Generic baselines are deep

ensembles (Lakshminarayanan et al., 2017) and MC-dropout (Gal & Ghahramani, 2016). We reproduce C^3's continuous-scale binary (CS-BC) and fixed-scale classification (FSC) heads (Mei et al., 2025b); its multi-class head is anti-calibrated in the source report (Appendix C.4). HAUWM (Wan et al., 2026), UNISafe (Seo et al., 2025), and S-QUBED (Mei et al., 2025a) retain their source components within the same probe, cache, and budget (Appendix C). Two controls isolate design gains beyond label access: a classifier pooling $z_h(u)$ without quantile, spatial, or conformal components, and C^3 with $\mathcal{L}_{\text{align}}$. Both use the same features, labels, and budget.

Metrics and statistical comparisons. We report AUROC and AUPRC of r against AF-fail_h, with the failure rate as the random-ranking AUPRC reference. Calibration metrics are ECE, empirical versus nominal coverage at $\alpha \in \{0.05, 0.1, 0.15, 0.2\}$, and interval width. Temporal calibration uses per-horizon coverage and $\rho = \text{slope}(\hat{U}_h)/\text{slope}(\epsilon_h^{(q)})$, where $\epsilon_h^{(q)}$ is the true error's q-quantile. We also report Spearman correlation and Shepherd's π (Schwarzkopf et al., 2012), used by C^3. Comparisons use trajectory-clustered bootstrap intervals and paired permutation tests; every quoted interval is clustered. DeLong's test (DeLong et al., 1988; Sun & Xu, 2014) is an independence-assuming reference. Holm–Bonferroni (Holm, 1979) controls family-wise error. Results average three training seeds on the shared cache; Table 3 records the prespecified measurements and falsification criteria.

4.2 Results

CalibWM improves ranking under matched supervision. In E2 (Table 1), CalibWM improves Bridge AUROC from 0.762 for C^3 with $\mathcal{L}_{\text{align}}$ to 0.782, a gain of $+0.020$ (trajectory-clustered 95% CI $[0.009, 0.031]$). On DROID, it improves from 0.744 to 0.761, a gain of $+0.017$ with CI $[0.006, 0.028]$. These comparisons hold features, labels, and budget fixed. On Bridge, the plain supervised probe reaches 0.751 and the strongest method trained without AF-fail reaches 0.694. Oracle visual proxies also score below CalibWM despite access to the ground-truth future. Removing alignment brings CalibWM close to label-free C^3 (FSC); the remaining design choices are evaluated in Table 2.

CalibWM detects failures in visually convincing rollouts. On the human-annotated optimism subset (Figure 2(c)), CalibWM reaches 0.634 AUPRC versus 0.408 for C^3 and 0.106 for oracle LPIPS†. This subset selects high-visual-quality videos and has a failure rate of 0.09, the random-ranking AUPRC reference. Without alignment, CalibWM reaches 0.421, supporting action-fidelity supervision's role in detecting these failures. The optimism and Bridge replay subsets have different failure rates (0.09 and 0.31); their respective AUPRC lifts over random ranking are approximately sevenfold and 1.9-fold.

Calibration recovers nominal segment-level coverage. In E1 (Figure 2(a)), both conformalized variants track the nominal level across the grid; three uncalibrated methods under-cover by four to eight percentage points. Under Proposition 1's assumptions, horizon-stratified calibration gives marginal coverage in $[0.900, 0.902]$ at $n_h = 500$, independent of $\hat{U}$'s shape. Empirical coverage fluctuates around this guarantee (Remark 3); panel (b) compares widths at matched coverage.

Horizon regularization improves interval efficiency. In E3, the horizon objective yields $\rho = 0.96$, compared with 0.61 without it. At matched coverage, width at $h = 16$ is 1.21 times the width at $h = 1$, compared with 2.43 times without the objective.

Alignment improves discrimination; calibration controls coverage. The Bridge ablations (Table 2) show that alignment raises AF-fail AUROC from 0.689 to 0.782, with approximately 6% wider intervals (normalized widths 0.94 and 1.00). Without alignment, ρ reaches 0.98. At $\lambda_a = 2.0$, BCE and the ranking surrogate train the shared output for discrimination while the other objectives fit its magnitude, introducing a trade-off with interval efficiency. The version without alignment provides narrower intervals when error-magnitude estimation is the priority. Conformal rescaling restores coverage; removing it preserves AUROC but reduces empirical coverage from 0.902 to 0.773.

Table 1: E2, action-fidelity failure discrimination on the replay-labelled subset (2,140 segments, 412 trajectories, base rate 0.31 on Bridge; 1,875, 356, 0.28 on DROID). The upper block never sees AF-fail, the lower trains on it. Oracle visual proxies are marked † because they consume the ground-truth future, which is why Proposition 2 does not cover them (Remark 2). Best per column in bold; our rows shaded.

	Bridge		DROID	
Method	AUROC ↑	AUPRC ↑	AUROC ↑	AUPRC ↑
PSNR†	0.612	0.401	0.604	0.392
SSIM†	0.634	0.423	0.621	0.410
LPIPS†	0.661	0.452	0.648	0.438
MC-dropout	0.598	0.388	0.586	0.375
S-QUBED (port)	0.643	0.431	0.631	0.418
Deep ensemble	0.657	0.447	0.644	0.432
UNISafe (port)	0.665	0.456	0.652	0.441
HAUWM (port)	0.672	0.463	0.659	0.448
C^3 (CS-BC)	0.681	0.471	0.668	0.457
C^3 (FSC)	0.694	0.485	0.679	0.469
CalibWM w/o alignment	0.689	0.479	0.674	0.463
Trained on the AF-fail *label*				
Supervised probe on AF-fail	0.751	0.556	0.733	0.535
$C^3 + \mathcal{L}_{\text{align}}$	0.762	0.570	0.744	0.549
CalibWM (ours)	**0.782**	**0.594**	**0.761**	**0.571**
Δ vs. best label-free	+0.088	+0.109	+0.082	+0.102
Δ vs. best label-using	+0.020	+0.024	+0.017	+0.022
95% clustered CI	[.009, .031]	[.011, .037]	[.006, .028]	[.009, .035]

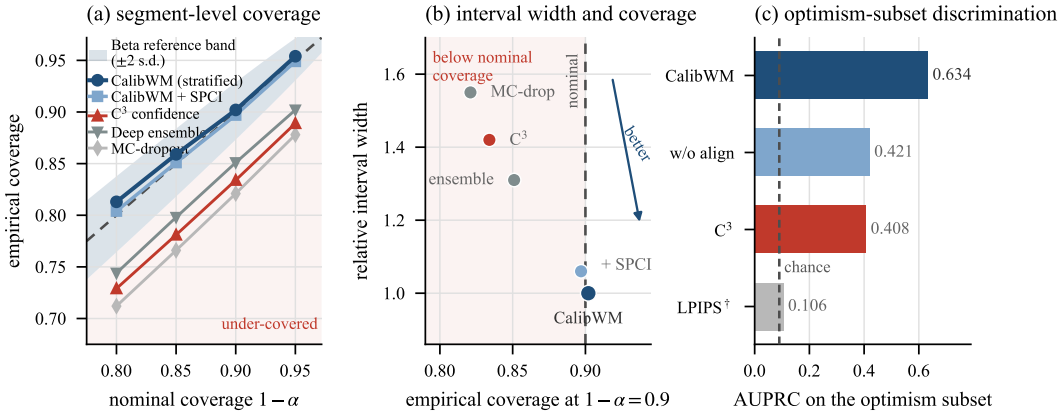

Figure 2: Coverage, interval width, and optimism-bias discrimination. (a) Segment-level coverage against the nominal level. The shaded band shows ±2 standard deviations of the $\text{Beta}(k, n_h + 1 - k)$ fluctuation around the marginal-coverage interval in Proposition 1 under the i.i.d., continuous-score reference model of Remark 3. The conformalized variants fall within this reference band; the remaining methods under-cover. (b) Interval width for methods reaching nominal coverage; the under-covered region is shaded out because width comparisons require matched coverage. (c) AUPRC on the human-annotated optimism subset, selected for high visual quality. Its failure rate is 0.09, the random-ranking reference. CalibWM reaches 0.634, compared with 0.106 for the oracle visual proxy LPIPS† and 0.421 without alignment.

Table 2: Ablations on Bridge with the replay-labelled subset. Coverage is pooled at nominal 0.9, width is relative to the full model at h=16, and OOD AUROC is averaged over the five shift axes. The right column names the effect tested by each ablation. Variants retaining the conformal layer have marginal coverage under Proposition 1's assumptions and empirical coverage near nominal here. Removing that layer reduces empirical coverage; its width is greyed because widths are comparable only at matched coverage.

Variant	AF AUROC	$\rho \to 1$	Cov.	Wid.	OOD AUROC	**Claimed role**
Full CalibWM	**0.782**	0.96	**0.902**	1.00	**0.806**	
w/o horizon, Eq. (6)	0.771	0.61	0.901	2.00	0.803	interval efficiency
w/o spatial localization	0.744	0.95	0.901	1.18	0.781	localization
w/o conformal layer	0.782	0.96	0.773	0.82	0.806	coverage only
w/o alignment, Eq. (4)	0.689	**0.98**	0.902	**0.94**	0.804	**the central claim**
epistemic teacher → dropout	0.779	0.96	0.901	1.02	0.742	E4 only

Epistemic scores respond to the evaluated shifts. E4 perturbs background, lighting, novel object, end effector, and viewpoint using real held-out assets and deterministic segment-seeded transformations. Across these five shifts, the epistemic branch distinguishes shifted inputs while the aleatoric branch remains near chance. Replacing the epistemic teacher with MC-dropout changes E2 little and reduces mean OOD AUROC from 0.806 to 0.742. These are empirical branch responses; Eq. (17) defines the variance components without guaranteeing their response to arbitrary shifts.

5 Related work

Reliability benchmarks. MiraBench evaluates physics adherence, action-following fidelity, and optimism bias in video generation (Yang et al., 2026). Its annotations distinguish visual quality from preservation of action outcomes. CalibWM uses them for label validation and the optimism diagnostic; Appendix B gives the benchmark context.

World-model uncertainty. HAUWM calibrates ensemble disagreement across horizons (Wan et al., 2026); UNISafe uses epistemic uncertainty and conformal thresholds for latent safety filtering (Seo et al., 2025); S-QUBED decomposes clip-level text-to-video uncertainty (Mei et al., 2025a). C^3 learns subpatch visual confidence from prediction-error threshold labels (Mei et al., 2025b). CalibWM combines action-fidelity supervision with spatial uncertainty and per-horizon calibration. Matched reimplementations and label-using controls separate these design choices from supervision access (Appendix C).

Conformal calibration. Split conformal prediction provides finite-sample coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018; Romano et al., 2019). Adaptive and sequential methods address temporal dependence through online feedback or residual-quantile estimation (Gibbs & Candes, 2021; Xu & Xie, 2023). Our construction uses independent trajectory segments and one aggregated score per segment at each horizon. The guarantee concerns prediction-error regions; action-fidelity discrimination is trained and evaluated separately.

6 Conclusion

CalibWM aligns uncertainty with action-fidelity failure through a frozen-feature probe. Segment-level conformal calibration provides fixed-horizon prediction-error coverage under exchangeability. Matched-label comparisons show AUROC gains of 0.020 on Bridge and 0.017 on DROID. On the human-annotated optimism subset, CalibWM reaches 0.634 AUPRC versus 0.106 for oracle LPIPS. Under ordering consistency and absence of additional failure information in a generated-rollout visual proxy conditional on the probe features, the readout has AUROC no lower than that proxy. These results support action-fidelity supervision as a useful component of uncertainty estimation for generated robot rollouts, alongside horizon-aware error estimation and spatial localization.

AI USE STATEMENT

AI assistants supported manuscript organization, reference summarization, inspection of supplied code and experiment records, review and revision of mathematical derivations, assumptions, and proofs, bilingual drafting and editing, and LaTeX preparation. The authors are responsible for the final technical content and references.

ETHICS STATEMENT

CalibWM is intended to support the diagnosis of action-fidelity failures in generated robot rollouts. Our evaluation uses BridgeData V2, DROID, and existing MiraBench annotations, whose roles and label provenance are described in Appendix C.3. Deployment should validate the full robot system under its intended operating conditions and use CalibWM alongside task-specific safety checks. Its segment-level statistical coverage guarantee does not certify physical robot safety. Reuse of the datasets and annotations should respect the providers' licensing, privacy, and access requirements.

REPRODUCIBILITY STATEMENT

Section 4 specifies the datasets, trajectory-level splitting, calibration sampling, and statistical comparisons. Appendix C documents the evaluation and baseline-porting rules, and Appendix C.3 distinguishes weak, replay, and human annotations. Appendix E reports the probe architecture, loss weights, optimizer, training budget, and calibration settings. The proofs and assumptions are given in Appendices A and A.1. Each probe is trained with three seeds on a shared frozen-model cache. Reproducing the numerical results also requires the corresponding cache, dataset partition, inverse-dynamics checkpoint, and scene-reconstruction configuration.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video PreTraining (VPT): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24639–24654. Curran Associates, Inc., 2022. doi: 10.52202/068431-1789.

David Blackwell. Equivalent comparisons of experiments. *The Annals of Mathematical Statistics*, 24(2):265–272, 1953. doi: 10.1214/aoms/1177729032.

Elizabeth R. DeLong, David M. DeLong, and Daniel L. Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 44(3):837–845, 1988. doi: 10.2307/2531595.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of uncertainty in Bayesian deep learning for efficient and risk-sensitive learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1184–1193. PMLR, 2018. URL https://proceedings.mlr.press/v80/depeweg18a.html.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1050–1059. PMLR, 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Isaac Gibbs and Emmanuel Candes. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/0d441de75945e5acbc865406fc9a2559-Paper.pdf.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30, pp. 5574–5584. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/2650d6089a6d640c5e85b2b88265dc2b-Paper.pdf.

Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. DROID: A large-scale in-the-wild robot manipulation dataset. In *Proceedings of Robotics: Science and Systems*, 2024. doi: 10.15607/RSS.2024.XX.120.

Roger Koenker and Gilbert Bassett, Jr. Regression quantiles. *Econometrica*, 46(1):33–50, 1978. doi: 10.2307/1913643.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6402–6413. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018. doi: 10.1080/01621459.2017.1307116.

Zhiting Mei, Ola Shorinwa, and Anirudha Majumdar. How confident are video models? empowering video models to express their uncertainty. *arXiv preprint arXiv:2510.02571*, 2025a. URL https://arxiv.org/abs/2510.02571.

Zhiting Mei, Tenny Yin, Micah Baker, Ola Shorinwa, and Anirudha Majumdar. World models that know when they don't know: Controllable video generation with calibrated uncertainty. *arXiv preprint arXiv:2512.05927*, 2025b. URL https://arxiv.org/abs/2512.05927.

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7 (35):983–999, 2006. URL https://www.jmlr.org/papers/v7/meinshausen06a.html.

Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pp. 73–108. Springer, 1992. doi: 10.1007/978-1-4612-0919-5_6.

NVIDIA, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical AI. *arXiv preprint arXiv:2501.03575*, 2025. URL https://arxiv.org/abs/2501.03575.

Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, et al. Open X-Embodiment: Robotic learning datasets and RT-X models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6892–6903. IEEE, 2024. doi: 10.1109/ICRA57147.2024.10611477.

Harris Papadopoulos, Kostas Proedrou, Volodya Vovk, and Alex Gammerman. Inductive confidence machines for regression. In *Machine Learning: ECML 2002*, volume 2430 of *Lecture Notes in Computer Science*, pp. 345–356. Springer, 2002. doi: 10.1007/3-540-36755-1_29.

Sabela Ramos, Sertan Girgin, Léonard Hussenot, Damien Vincent, Hanna Yakubovich, Daniel Toyama, Anita Gergely, Piotr Stanczyk, Raphael Marinier, Jeremiah Harmsen, Olivier Pietquin, and Nikola Momchev. RLDS: an ecosystem to generate, share and use datasets in reinforcement learning. *arXiv preprint arXiv:2111.02767*, 2021. URL https://arxiv.org/abs/2111.02767.

Yaniv Romano, Evan Patterson, and Emmanuel J. Candès. Conformalized quantile regression. In *Advances in Neural Information Processing Systems*, volume 32, pp. 3543–3553, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/5103c3584b063c431bd1268e9b5e76fb-Abstract.html.

Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1–4):259–268, 1992. doi: 10.1016/0167-2789(92)90242-F.

Dietrich Samuel Schwarzkopf, Benjamin De Haas, and Geraint Rees. Better ways to improve standards in brain-behavior correlation analysis. *Frontiers in Human Neuroscience*, 6:200, 2012. doi: 10.3389/fnhum.2012.00200.

Junwon Seo, Kensuke Nakamura, and Andrea Bajcsy. Uncertainty-aware latent safety filters for avoiding out-of-distribution failures. In *Proceedings of The 9th Conference on Robot Learning*, volume 305 of *Proceedings of Machine Learning Research*, pp. 4442–4472. PMLR, 2025. URL https://proceedings.mlr.press/v305/seo25a.html.

Xu Sun and Weichao Xu. Fast implementation of DeLong's algorithm for comparing the areas under correlated receiver operating characteristic curves. *IEEE Signal Processing Letters*, 21 (11):1389–1393, 2014. doi: 10.1109/LSP.2014.2337313.

Erik Torgersen. *Comparison of Statistical Experiments*, volume 36 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1991. doi: 10.1017/CBO9780511666353.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. URL https://arxiv.org/abs/1812.01717.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pp. 5998–6008, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 1 edition, 2005. doi: 10.1007/b106715.

Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. BridgeData V2: A dataset for robot learning at scale. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pp. 1723–1736. PMLR, 2023. URL https://proceedings.mlr.press/v229/walke23a.html.

Shenghua Wan, Le Gan, and De-Chuan Zhan. Learning to be uncertain: Pre-training world models with horizon-calibrated uncertainty. In *International Conference on Learning Representations*, volume 2026, pp. 106107–106127, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/ad222e6e38e26e3f00d85829b9e02f43-Paper-Conference.pdf.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. doi: 10.1109/TIP.2003.819861.

Chen Xu and Yao Xie. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38707–38727. PMLR, 2023. URL https://proceedings.mlr.press/v202/xu23r.html.

Tianzhuo Yang, Zihan Shen, Zirui Mi, Zhaoyi Zhang, Jiayi Zhou, Jiaming Ji, Juntao Dai, Jiawei Chen, Boyuan Chen, and Yaodong Yang. MiraBench: Evaluating action-conditioned reliability in robotic world models. *arXiv preprint arXiv:2605.29360*, 2026. URL https://arxiv.org/abs/2605.29360.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

A PROOFS

A.1 ASSUMPTIONS AND SCOPE

Assumption 1 (Exact segment-level exchangeability)**.** *For a fixed h, the calibration and test segment scores $\{S_h^{(1)}, \ldots, S_h^{(n_h)}, S_h^{\text{test}}\}$ are jointly exchangeable. This requires two things: segments come from* mutually independent trajectories *rather than adjacent slices of one long rollout, and the score function is fixed before $\mathcal{D}_{\text{cal}}$ is seen, so f_ϕ is fit on a split disjoint from the calibration segments.*

Assumption 2 (Order-consistent readout)**.** *The readout ranks as the posterior does: r is almost surely a strictly increasing transformation of $r^\star(\Phi) = \mathbb{P}(Y = 1 \mid \Phi)$. The BCE term of Eq. (4) is strictly proper, so $r^\star$ is its unique population minimizer at $\zeta = 0$ given enough capacity and data. We assume order consistency rather than $r = r^\star$ because the ranking surrogate, the sharing of $\hat{U}$, and the certainty weight $|2p - 1|$ each bias the fitted probability without necessarily reordering it, and AUROC ignores such bias.*

The deployed readout $r = \sigma(w\, g(\hat{U}_h) + b)$, with $w > 0$, is strictly increasing in the scalar $g(\hat{U}_h)$. Order consistency therefore requires $r^\star(\Phi)$ to factor through that scalar, so $g(\hat{U}_h)$ must be sufficient for the posterior of AF-fail_h given Φ. An object max can discard information when multiple objects drive failure or when map magnitude and spatial arrangement carry separate information. In these cases, a classifier with direct access to Φ can attain a higher AUROC than r. The supervised probe in Table 1 pools $z_h(u)$ into a head without the object-max bottleneck and reaches 0.751, compared with 0.782 for r. This comparison is consistent with the assumption but does not establish it, because the supervised probe also omits the other terms of Eq. (5). A higher AUROC for that probe would motivate testing for information loss in the object-max readout.

Assumption 3 (The proxy is a garbling of Φ)**.** *The quality proxy q carries no information about Y beyond Φ: $q \perp Y \mid \Phi$, which covers both a deterministic function of Φ and a noisy observation of it.*

Remark 1 (What the guarantee says, and what it does not). The per-unit content of Proposition 1 is fixed by Ψ, not by how units co-vary. For $\Psi = \max_u$ the event $\{S_h \le \hat{Q}^{(h)}\}$ *is* containment of every unit at once, giving finite-sample simultaneous coverage per segment under no dependence assumption, at the cost of wider intervals. For the default $\Psi = \text{Quantile}_{0.95}$ it is the tolerance statement $\mathbb{P}\big(|\Omega|^{-1}\#\{u : x_h^\star(u) \in C_\alpha^{(h)}(u)\} \ge 0.95\big) \ge 1 - \alpha$, which implies neither simultaneous nor per-unit marginal coverage.

What does survive is averaging over horizons. If the test segment's horizon is drawn independently of the calibration draws, then $\mathbb{P}(\text{covered}) = \sum_h w_h\, \mathbb{P}\big(S_h^{\text{test}} \le \hat{Q}_{1-\alpha}^{(h)}\big)$ for horizon weights w_h, and both sides of Proposition 1 pass through the mixture:

$$1 - \alpha \;\le\; \mathbb{P}(\text{covered}) \;\le\; 1 - \alpha + \sum_h \frac{w_h}{n_h + 1}, \tag{10}$$

which is the same $[0.900, 0.902]$ band at $n_h = 500$. This guarantees the underlying marginal coverage for the pooled results in Tables 2 and 5; the reported empirical frequencies remain subject to sampling variation. What is *not* guaranteed is a single threshold calibrated on scores pooled across a bucket of horizons, since scores at different h are not exchangeable with one another; nor is any reading of the bound under distribution shift. And when segments are slices of one rollout, Assumption 1 fails outright and only the adaptive methods apply, with a long-run time average in place of a fixed-horizon bound.

Remark 2 (Which baselines Proposition 2 reaches). Oracle PSNR, SSIM, and LPIPS compare generated frames with $o^\star$ and can depend on information in $x^\star$ that is absent from Φ. Assumption 3 is therefore not ensured by their construction. We do not establish the conditional independence $q \perp Y \mid \Phi$ for these proxies, so Proposition 2 provides no guarantee for the rows marked † in Table 1; those comparisons are empirical. The proposition applies to proxies that satisfy Assumption 3, including deterministic functions of Φ. A proxy computed from the generated rollout alone is covered only when its information is retained in Φ or the stated conditional independence otherwise holds.

A.2 Proof of Proposition 1

Fix h and write $S_i := S_h^{(i)}$ for $i = 1, \dots, n_h$ and $S_{n_h+1} := S_h^{\text{test}}$. By Assumption 1 the vector $(S_1, \dots, S_{n_h+1})$ is exchangeable, so its law is invariant to permutation of indices. Let R denote the rank of S_{n_h+1} within the multiset $\{S_1, \dots, S_{n_h+1}\}$, with ties broken uniformly at random. Exchangeability implies R is uniform on $\{1, \dots, n_h + 1\}$, so for any integer m,

$$\mathbb{P}(R \le m) = \frac{m}{n_h + 1}. \tag{11}$$

Let $k := \lceil (1-\alpha)(n_h+1) \rceil$ and let $S_{(k)}$ be the k-th smallest of the calibration scores $\{S_1, \dots, S_{n_h}\}$, so that $\hat{Q}_{1-\alpha}^{(h)} = S_{(k)}$. The hypothesis $n_h \ge (1-\alpha)/\alpha$ gives $(1-\alpha)(n_h+1) \le n_h$ and hence $k \le n_h$, so $S_{(k)}$ exists. When that hypothesis fails, the convention $\hat{Q}_{1-\alpha}^{(h)} = +\infty$ makes the lower bound hold vacuously and the upper bound void, which is why §4 fixes a floor on the number of calibration segments per stratum. At $\alpha = 0.1$ the requirement is only $n_h \ge 9$; the far larger budget we actually use is driven by the width of the slack term below, not by finiteness of the threshold.

The event $\{S_{n_h+1} \le S_{(k)}\}$ contains the event $\{R \le k\}$: if the test score's rank among all $n_h + 1$ values is at most k, then at most $k-1$ calibration scores are strictly below it, so it cannot exceed the k-th smallest calibration score. Therefore by Eq. (11),

$$\mathbb{P}\big(S_{n_h+1} \le S_{(k)}\big) \;\ge\; \frac{k}{n_h+1} \;=\; \frac{\lceil (1-\alpha)(n_h+1) \rceil}{n_h+1} \;\ge\; 1-\alpha, \tag{12}$$

where the last inequality uses $\lceil z \rceil \ge z$ with $z = (1-\alpha)(n_h+1)$.

For the upper bound, assume the $n_h + 1$ scores are almost surely distinct, so that R is exactly uniform without randomization. Note that a continuous marginal law for S_h is *not* enough here, since exchangeability permits $S_1 = \cdots = S_{n_h+1}$ almost surely, as when all scores are a common continuous random variable, and joint distinctness is what rules this out. Under it, $\{S_{n_h+1} \le S_{(k)}\} = \{R \le k\}$ almost surely, and

$$\mathbb{P}\big(S_{n_h+1} \le S_{(k)}\big) = \frac{\lceil (1-\alpha)(n_h+1) \rceil}{n_h+1} < \frac{(1-\alpha)(n_h+1)+1}{n_h+1} = 1 - \alpha + \frac{1}{n_h+1}, \tag{13}$$

using $\lceil z \rceil < z + 1$. This is the claimed two-sided bound. □

Remark 3 (The slack is not the width of the expected empirical band). Proposition 1 bounds coverage averaged over the calibration draw and test point. Its slack $1/(n_h+1)$ therefore bounds a marginal probability, not the deviation of a measured coverage frequency. For the sampling reference, fix the fitted score function and assume that calibration and test scores are i.i.d. with a continuous common distribution. Across calibration draws, the coverage probability conditional on the calibration set then follows $\mathrm{Beta}(k, n_h + 1 - k)$ (Vovk et al., 2005; Angelopoulos & Bates, 2023). Its standard deviation is approximately $\sqrt{\alpha(1-\alpha)/(n_h+2)}$, or 0.013 at $n_h = 500$ and $\alpha = 0.1$, compared with the 0.002 marginal slack. An independent finite test set adds binomial sampling variation, with standard deviation approximately $\sqrt{\alpha(1-\alpha)/n_{\text{test}}}$ when conditional coverage is near $1-\alpha$. Thus 0.887 empirical coverage at nominal 0.9 can be consistent with sampling variation under this reference model. The coverage figures use the Beta fluctuation as a reference band; it is not an additional guarantee under exchangeability alone and does not include finite-test-set variation. The marginal upper bound in Proposition 1 still requires almost surely distinct scores. At $n_h = 500$ we compare single-stratum deviations with the 0.013 calibration standard deviation; reducing this below 0.01 would require approximately 900 calibration segments under the i.i.d. reference model.

A.3 Proof of Proposition 2

We proceed in two steps.

Step 1: the posterior is an AUROC-optimal ranking statistic. Let $Y \in \{0, 1\}$ and let Φ be the feature context. Consider any score $s(\Phi)$ measurable with respect to $\sigma(\Phi)$. Thresholding s at level c defines a test of $H_0 : Y = 0$ against $H_1 : Y = 1$ with true positive rate $\mathrm{TPR}_s(c) = \mathbb{P}(s(\Phi) > c \mid Y = 1)$ and false positive rate $\mathrm{FPR}_s(c) = \mathbb{P}(s(\Phi) > c \mid Y = 0)$. Let p_1 and p_0 be the conditional densities, with respect to a common dominating measure, of Φ given $Y = 1$ and $Y = 0$. By the

Neyman–Pearson lemma (Neyman & Pearson, 1992), at every level β the most powerful test of size β is a threshold test on the likelihood ratio $\Lambda(\Phi) = p_1(\Phi)/p_0(\Phi)$, randomized on the boundary when Λ has an atom there. For a score, we use the threshold ROC with independent boundary randomization to interpolate ties. Its area is the usual AUROC, which assigns half weight to tied positive–negative pairs. This curve need not be concave and is distinct from the envelope of all tests attainable from an experiment. Neyman–Pearson optimality implies that the threshold ROC of Λ lies pointwise above that of every $\sigma(\Phi)$-measurable score. By Bayes' rule,

$$r^\star(\Phi) = \mathbb{P}(Y = 1 \mid \Phi) = \frac{\pi_1 p_1(\Phi)}{\pi_1 p_1(\Phi) + \pi_0 p_0(\Phi)} = \frac{\pi_1 \Lambda(\Phi)}{\pi_1 \Lambda(\Phi) + \pi_0}, \tag{14}$$

which is a strictly increasing function of $\Lambda(\Phi)$ for fixed priors $\pi_0, \pi_1 > 0$. A strictly increasing transformation preserves the induced ordering and hence the entire ROC curve, so $r^\star$ generates the same pointwise dominant ROC as Λ. Since AUROC is the integral of the ROC curve, $\mathrm{AUROC}(r^\star \to Y) \geq \mathrm{AUROC}(s \to Y)$ for every $\sigma(\Phi)$-measurable s. A strictly increasing map leaves the ROC curve unchanged, so under Assumption 2 the deployed readout r has the same ROC as $r^\star$ and inherits this optimality.

Step 2: a visual proxy is a garbling and is therefore dominated. If q is a deterministic function of Φ, it is a $\sigma(\Phi)$-measurable score and Step 1 applies directly, giving $\mathrm{AUROC}(q) \leq \mathrm{AUROC}(r^\star)$. In general, Assumption 3 says the conditional law of q given Φ does not depend on Y, so the experiment $\mathcal{E}_q = \{\mathcal{L}(q \mid Y = y)\}_{y\in\{0,1\}}$ is obtained from $\mathcal{E}_\Phi = \{\mathcal{L}(\Phi \mid Y = y)\}_{y\in\{0,1\}}$ by applying a Markov kernel $\mathcal{K}(\cdot \mid \Phi)$, which is exactly Blackwell's notion of one experiment being a *garbling* of another (Blackwell, 1953); see Torgersen (1991) for the modern treatment. For dichotomies, the comparison theorem is equivalent to pointwise dominance of the attainable ROC. We use only its elementary direction, that a garbling cannot increase the attainable power, and not the converse. Writing $\mathrm{TPR}_{\mathcal{E}}(\beta)$ for the largest true positive rate any test based on experiment $\mathcal{E}$ achieves at false positive rate β,

$$\mathrm{TPR}_q(\beta) \;\leq\; \mathrm{TPR}_{\mathcal{E}_q}(\beta) \;\leq\; \mathrm{TPR}_{\mathcal{E}_\Phi}(\beta) \;=\; \mathrm{TPR}_{r^\star}(\beta) \qquad \text{for every } \beta \in [0,1]. \tag{15}$$

The first inequality holds because thresholding the single statistic q is one particular test available under $\mathcal{E}_q$, the second is Blackwell dominance, and the equality is Step 1, which identified $r^\star$ as attaining the envelope of $\mathcal{E}_\Phi$. Integrating over $\beta \in [0,1]$,

$$\mathrm{AUROC}(q \to Y) = \int_0^1 \mathrm{TPR}_q(\beta)\, d\beta \;\leq\; \int_0^1 \mathrm{TPR}_{r^\star}(\beta)\, d\beta = \mathrm{AUROC}(r^\star \to Y). \tag{16}$$

Combining with Assumption 2 gives $\mathrm{AUROC}(r) \geq \mathrm{AUROC}(q)$.

Equality condition. If q is almost surely a strictly *increasing* transformation of $r^\star$ it induces the same ordering, hence the same ROC curve and the same AUROC. Monotonicity alone is not enough, since a strictly decreasing transformation reverses the ordering and gives $1 - \mathrm{AUROC}(r^\star)$, which is why the direction has to be fixed. Conversely, because the dominance above is pointwise and the ROC curves are bounded, the two integrals can agree only if $\mathrm{TPR}_q(\beta) = \mathrm{TPR}_{r^\star}(\beta)$ for almost every β, so q must already be an ROC-optimal statistic for Φ. That does not force q to be an increasing transformation of $r^\star$ in degenerate cases: if $r^\star$ is almost surely constant, both curves are the diagonal and any q attains equality. Away from that degeneracy, ROC coincidence means q orders concordantly with $r^\star$ on every pair the posterior separates. □

Remark 4 (Mutual information does not order AUROC). The data processing inequality gives $I(q; Y) \leq I(\Phi; Y)$, but this does not imply an AUROC ordering. Mutual information and AUROC are different functionals of the joint distribution: two scores can have equal mutual information with Y and different AUROC, and the score with more mutual information can have lower AUROC. Blackwell comparison gives the required ordering by comparing the attainable ROC at every operating point.

B Related work

MiraBench evaluates 12 model configurations using over 16,000 structured human annotation decisions (Yang et al., 2026). Within the same model family, scaling from 2B to 14B reduces failure preservation from 48.7 to 12.8. Failure preservation measures how often the model depicts failed outcomes as failures. These measurements complement the post-training comparison in §1.

Reliability benchmarks for world models. A recent line of work asks whether action-conditioned video models are usable as simulators at all, not whether their frames look real. MiraBench (Yang et al., 2026) organizes this into physics adherence, action-following fidelity, and optimism-bias detection, and finds that visual quality scores fail to predict task completion under the commanded action, that scale does not fix this, and that post-training which improves task completion can simultaneously destroy a model's willingness to depict failure. Such benchmarks supply the diagnosis and the human-annotated labels that make an action-fidelity calibration target feasible, but not a calibrated uncertainty signal; CalibWM consumes their labels and works on the solution side.

Uncertainty for world models. HAUWM (Wan et al., 2026) uses a horizon-calibrated objective to make ensemble disagreement grow with the prediction horizon. It is evaluated by control return on latent control models, without spatial localization or calibration metrics. We transfer its horizon-calibration idea to a spatiotemporal uncertainty probe. UNISafe (Seo et al., 2025) combines epistemic uncertainty with a conformally calibrated threshold for latent safety filtering in a generative world model. Its scalar gate detects departures from the training distribution. CalibWM instead calibrates an action-fidelity readout and retains spatial and per-horizon resolution. S-QUBED (Mei et al., 2025a) decomposes text-to-video uncertainty at the clip level and validates it through rank correlation with semantic similarity. C^3 (Mei et al., 2025b) uses strictly proper scoring rules for subpatch-level confidence in action-conditioned video models. Its target is $\mathbf{1}[|\hat{x}-x^\star| \le \varepsilon]$, evaluated against error magnitude; our target is $\mathrm{AF\text{-}fail}_h$, evaluated against action-fidelity failure. UNISafe already provides a finite-sample conformal guarantee. Here, the horizon objective and per-horizon calibration address uncertainty and coverage at each rollout horizon.

Conformal prediction and uncertainty quantification. Split conformal prediction gives distribution-free finite-sample coverage under exchangeability (Vovk et al., 2005; Papadopoulos et al., 2002; Lei et al., 2018; Romano et al., 2019; Angelopoulos & Bates, 2023), which autoregressive rollouts violate along time. The standard responses are adaptive conformal inference, driving long-run time-average coverage to nominal by online feedback (Gibbs & Candes, 2021), and sequential predictive conformal inference, re-estimating the residual quantile with a quantile random forest (Xu & Xie, 2023; Meinshausen, 2006). Our contribution is not a new conformal method but a choice of calibration unit, one scalar per segment per horizon, under which exchangeability is plausibly satisfied rather than violated by treating correlated pixels as independent. Elsewhere we build on standard machinery: deep ensembles (Lakshminarayanan et al., 2017) and MC-dropout (Gal & Ghahramani, 2016), the aleatoric and epistemic distinction of Kendall & Gal (2017) and Depeweg et al. (2018), proper scoring (Gneiting & Raftery, 2007) with quantile regression (Koenker & Bassett, 1978), and low-rank adapters (Hu et al., 2022).

C Protocol details

C.1 Experiments and their falsification criteria

C.2 Porting rules for the transferred baselines

Three baselines were designed for settings other than a spatiotemporal uncertainty map over an action-conditioned rollout: HAUWM (Wan et al., 2026) for latent control models evaluated by return on DeepMind Control, MetaWorld, and RoboDesk; UNISafe (Seo et al., 2025) for latent safety filtering; and S-QUBED (Mei et al., 2025a) for clip-level text-to-video. To make the comparison informative and not rhetorical, each port keeps everything fixed except the component under test. All three read the same cached features, use a probe of the same depth and width as ours, train for the same number of steps with the same optimizer, and emit the same per-segment, per-horizon, per-position contract. What differs is only the set of capabilities the source method has. None of the three supervises against action fidelity, so no port receives $\mathcal{L}_{\mathrm{align}}$; on top of that, the S-QUBED port pools over space before its output heads, since it is a clip-level method and cannot localize, and the UNISafe port omits the horizon-monotone term. The HAUWM port is the most capable of the three, keeping horizon calibration and spatial resolution.

Each ablation removes one component from the full model; each port omits $\mathcal{L}_{\mathrm{align}}$ and any additional components absent from its source design. On Bridge, all three ports fall below the no-

Table 3: The four experiments and the optimism diagnostic, with their evaluation criteria. The E1 Beta band is a sampling diagnostic under the i.i.d., continuous-score reference model of Remark 3, not a deterministic constraint on empirical coverage. Here ρ is the horizon slope ratio and $\hat{U}^{\mathrm{epi}}$ the epistemic branch.

ID	Question	Primary measurement	Falsified if
E1	Is coverage correct, in distribution and under shift?	Empirical against nominal coverage on the α grid; interval width	Coverage falls below nominal by more than two standard deviations of the Beta fluctuation of Remark 3, or nominal coverage is attained only at a width no better than the baselines achieve
E2	**Does $\hat{U}$ predict action-fidelity failure?**	AUROC and AUPRC of r against C^3 and visual proxies, replay-labelled subset	A 95% CI of ΔAUROC against the *label-free* C^3 head strictly below 0 challenges the empirical ordering and the applicability of the proposition's assumptions; separately, the practical claim needs the CI against the strongest *label-using* control to exclude 0 from above
E3	Does uncertainty grow correctly with horizon?	ρ, per-horizon coverage, Spearman against h	Temporal calibration is significantly worse than the horizon-aware baseline
E4	Does $\hat{U}^{\mathrm{epi}}$ separate shift, and $\hat{U}^{\mathrm{ale}}$ not?	Per-axis OOD AUROC of both branches over five shift axes	$\hat{U}^{\mathrm{epi}}$ and $\hat{U}^{\mathrm{ale}}$ are indistinguishable under shift
D	Are imagined successes that really fail caught?	AUPRC on the optimism-bias subset, read against the chance floor set by its base rate	No AUPRC gain over the oracle visual proxies

alignment ablation at 0.689 AUROC: HAUWM reaches 0.672, UNISafe 0.665, and S-QUBED 0.643. The HAUWM port omits alignment and the attribution split; UNISafe additionally omits the horizon term, and S-QUBED additionally lacks spatial localization. The 0.017 gap between the no-alignment ablation and the HAUWM port is consistent with an indirect effect of attribution distillation through the shared trunk (§3.1), which also supplies the quantile head. These comparisons use matched architectures and training budgets. They evaluate reimplementations of the source ideas within our architecture, rather than the authors' released systems. The label-using controls in Table 1 provide the complementary comparison with access to the same action-fidelity supervision.

C.3 Label provenance

The replay route works as follows. For the subset of Bridge and DROID trajectories whose scenes we reconstruct, we initialize the simulator at the recorded state, execute the recorded action sequence $a^{\star}_{t:t+H}$, and record whether the task's success predicate fires. The onset of failure is the first horizon at which the simulated end-effector pose leaves a tolerance tube around the recorded trajectory, and failure is propagated forward from it, so $\mathrm{AF\text{-}fail}_h$ is monotone in h within a segment. Reconstruction succeeds for 412 of the Bridge trajectories and 356 of the DROID trajectories, yielding the segment counts quoted in Table 1. Neither of the two limitations that follow is removable. The verdict is a simulator's, so the sim-to-real gap in contact dynamics enters the label; and reconstruction selects for scenes a simulator can express, which biases the subset towards rigid-body manipulation and away from deformable objects.

MiraBench's 906 released segments include an optimism-bias subset of 410 annotated for whether a real failure was imagined as a success, and an action-following subset of 210. These annotations describe videos produced by MiraBench's own twelve generator configurations, not by our C^3 backbone, so a MiraBench label cannot be attached to a segment in our cache. We therefore use them in two restricted ways: to fit and validate τ_a against human judgement on the videos they do describe, and as a held-out benchmark on which every method, ours included, is run over MiraBench's videos. The optimism numbers in §4.2 are of the second kind.

The evaluation unit is a (segment, horizon) window. The optimism diagnostic retains windows with high visual quality to test failures that are not apparent from visual corruption. Most retained windows are action-faithful; the failure rate is 0.09, compared with 0.31 on the Bridge replay subset. The positive cases depict successful-looking outcomes for actions that failed. The 0.09 rate provides the random-ranking AUPRC reference; the performance of visual proxies is measured empirically in Figure 2(c).

Weak labels are produced by an IDM applied to generated frames, while the probe reads backbone features from those same frames. Evaluation against these labels therefore partly measures agreement with the IDM's errors. Primary conclusions use replay labels, which avoid this particular dependence. Cohen's κ quantifies agreement between the two routes and informs interpretation of weak-label results where replay is unavailable.

The IDM's reliability can change under distribution shift on generated frames.

C.4 What prior work publishes

Table 4 summarizes the metrics reported by the closest prior work. The compared methods do not report action-fidelity failure discrimination: C^3 evaluates latent error magnitude without AUROC or AUPRC, horizon-aware and safety-filter methods report control return and safety rates on non-video benchmarks, and the clip-level method reports rank correlation with text-video similarity. We therefore evaluate the methods on a shared cache using a common action-fidelity target.

C^3 reports Bridge ECE and MCE only as bar charts, so those cells are left untranscribed. Its MCC head has a positive headline correlation despite the expected negative sign for confidence; the -0.130 shown here comes from its re-analysis with bin upper bounds capped at 0.2. Our comparisons do not depend on that published value. The real-robot OOD result is not assigned to a specific head because the source does not identify the variant.

Table 4: Published numbers for the closest prior work, under the metrics that paper actually uses. Shepherd's π is the robust rank correlation between a method's scalar output and the latent prediction error. C^3 emits a *confidence*, so its correlation is negative by construction, whereas CalibWM outputs an *uncertainty* and should correlate positively: magnitudes are comparable, signs are not. All entries are from Mei et al. (2025b), which reports no baselines of its own. Here "n.r." marks a quantity the source does not report.

Method	**Data**	**Shep. π**	**ECE**	**MCE**	**AF-fail AUROC**
C^3 (FSC)	Bridge	-0.373	n.r.	n.r.	n.r.
C^3 (CS-BC)	Bridge	-0.172	n.r.	n.r.	n.r.
C^3 (MCC)	Bridge	-0.130	n.r.	n.r.	n.r.
C^3 (CS-BC)	DROID	-0.149	7.28×10^{-2}	1.74×10^{-1}	n.r.
C^3	real-robot OOD	n.r.	9.98×10^{-2}	1.71×10^{-1}	n.r.

Because that table has no ensemble row to compare against, one clarification about our own deep-ensemble baseline belongs here. We run a genuine parameter ensemble, K members drawn from the approximate posterior, and its Shepherd's π of 0.42 in Table 5 is the strongest association with latent error of any method we ran that never sees $\mathrm{AF\text{-}fail}$. That is not evidence against E4. E4 predicts that $\hat{U}^{\mathrm{ale}}$ tracks error *magnitude* well while failing to separate distribution *shift*, and the two claims are about different targets; E4 is falsified only if $\hat{U}^{\mathrm{ale}}$ also separates shift. The distinction matters because varying the generation seed at fixed parameters, which is sometimes described as an ensemble, is $\hat{U}^{\mathrm{ale}}$ under Eq. (17) and not a parameter ensemble at all.

C.5 E2 as an ordering

Figure 3(a) shows the Bridge AUROC ordering from Table 1, with trajectory-clustered intervals. The oracle visual proxies rank below CalibWM despite access to the ground-truth future; these comparisons are empirical and lie outside Proposition 2. The three label-using methods are closer together, with a $+0.020$ gain over the strongest control. Panel (b) illustrates ROC shapes constructed from the

AUROC values under an equal-variance binormal model; it does not report empirical performance at individual thresholds.

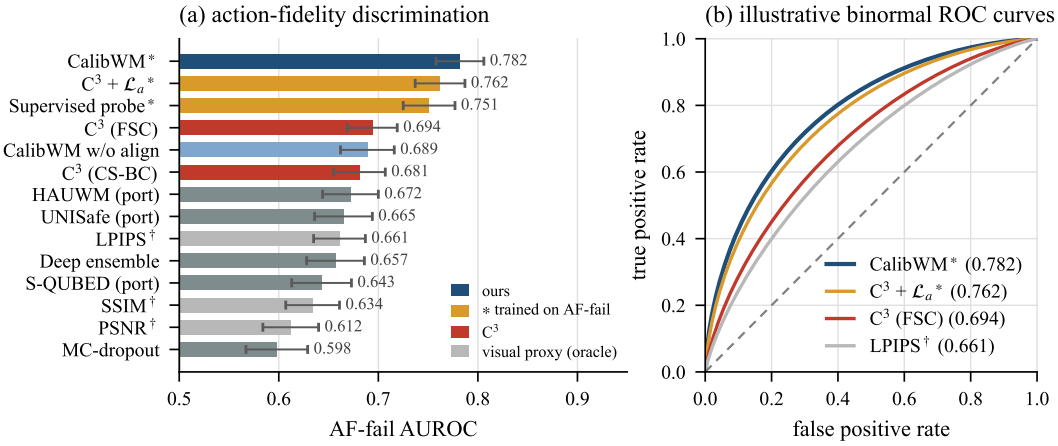

Figure 3: E2 on Bridge. (a) AF-fail AUROC with trajectory-clustered 95% intervals. Bars start at chance, and all values are annotated. Methods marked $*$ train on AF-fail; oracle visual proxies are shaded light. (b) Illustrative, non-empirical ROC curves from an equal-variance binormal model matched to the AUROC values in (a). These curves do not measure per-threshold performance. The comparison with label-free C^3 assesses the action-fidelity target on the same frozen features; the two other label-using controls assess the additional contribution of the CalibWM design.

Table 5: E1 in full, on Bridge at nominal $1 - \alpha = 0.9$. Coverage is best when closest to nominal and is marked in bold. The four greyed widths come from rows without a conformal layer, which do not target a nominal level and so are not on a comparable footing. Here π is Shepherd's π, reported against uncertainty so that larger is better; C^3 emits a confidence and its published π carries the opposite sign (Table 4), so the two are compared by magnitude. Figure 2(a,b) plots the first two columns across the full nominal grid.

Method	Cov. $\to 0.9$	Wid. $\downarrow$	ECE $\downarrow$	$\pi \uparrow$
MC-dropout	0.821	1.55	0.102	0.29
C^3 (FSC)	0.834	1.42	0.087	0.37
Deep ensemble	0.851	1.31	0.074	0.42
w/o conformal	0.773	0.88	0.061	**0.61**
w/o alignment	**0.902**	**0.94**	**0.020**	**0.61**
CalibWM + SPCI	0.897	1.06	0.024	0.58
CalibWM (strat.)	**0.902**	1.00	0.021	0.58

C.6 Full temporal calibration breakdown

Table 6 expands Figure 4(a,b) into per-horizon numbers and separates the methods that carry a per-horizon conformal layer from those that do not, which is the distinction that decides whether a coverage column is informative at all. Table 2 then removes one component at a time on the same data.

D Qualitative analysis and compute

The release includes spatial and temporal diagnostics. Horizon strips show whether $\hat{U}_h$ grows in the region of an emerging failure or across the whole frame, a distinction not captured by interval width alone. An optimism case library overlays $\hat{U}$ on segments depicting success despite real failure,

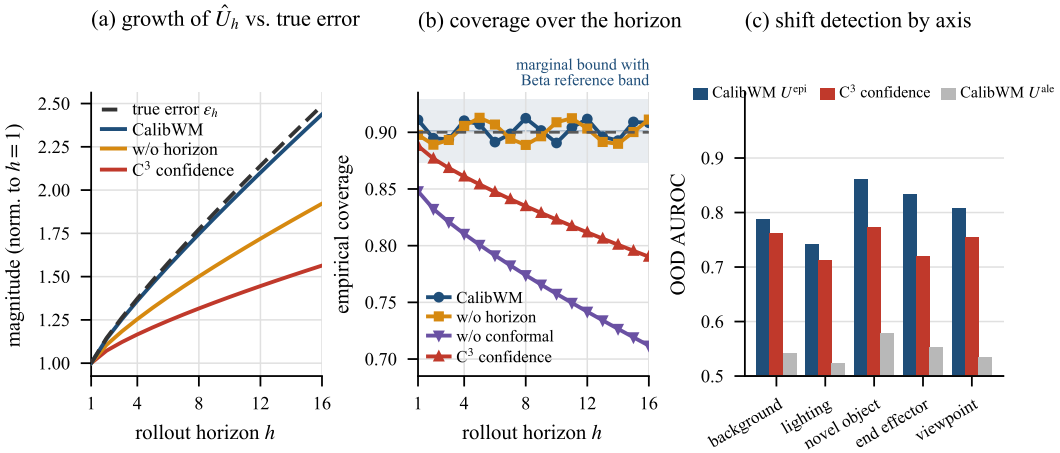

Figure 4: E3 and E4 on Bridge. (a) Growth of $\hat{U}_h$ against the true error, both normalized at h=1. (b) Per-horizon coverage. The dark strip is the $[1-\alpha,\ 1-\alpha+\frac{1}{n_h+1}]$ band Proposition 1 places *marginal* coverage in at n_h=500 under its assumptions. The pale strip adds a ± 2 standard-deviation reference for variation across calibration draws under the Beta model (Remark 3); it is not a deterministic bound on measured coverage. (c) Shift detection compares the responses of the two branches. Legend entries beginning "w/o" are CalibWM ablations.

Table 6: E3, temporal calibration on Bridge, expanding Figure 4(a,b). Coverage is at nominal 0.9 and the ideal ρ is 1; widths are relative to CalibWM at h=1. The first two rows have no per-horizon conformal layer, so their coverage is free to drift and their widths are not on a comparable footing (marked "–"). The remaining rows use per-horizon conformal calibration and attain empirical coverage near nominal. Their marginal coverage follows Proposition 1 under its assumptions; the horizon term's contribution is assessed by interval width at h=16. The two drifting rows average over h to the pooled coverage reported for the same method in Table 5 and Table 2. The last column is the Spearman correlation between $\hat{U}_h$ and ϵ_h at h=16.

Method	$\rho \to 1$	Cov. h=1	Cov. h=16	Wid. h=1	Wid. h=16	Spear. h=16
C³ (FSC)	0.37	0.888	0.790	–	–	0.31
CalibWM w/o conformal	0.96	0.848	0.712	–	–	0.52
HAUWM (port)	0.93	0.901	0.901	1.05	1.34	0.44
CalibWM w/o horizon	0.61	0.901	0.901	1.00	2.43	0.38
CalibWM (ours)	**0.96**	**0.902**	**0.901**	**1.00**	**1.21**	**0.52**

allowing inspection of uncertainty in visually clean regions with incorrect action outcomes. Matching object-level U_{obj} to the object involved in each failure gives an attribution hit rate for spatial localization.

The rollout cache dominates the compute budget. Eq. (17) needs a seed-variance per member, so a segment costs KL multi-step rollouts, 32 at the default $K=4$, $L=8$; the deployed rollout is one of them, since member 1 is the unperturbed base. Segments are rolled out in batches, and the K members share one set of base weights with only low-rank adapters swapped, so memory grows as $K \cdot \text{rank} \cdot \text{width}$ rather than as K copies of the denoiser. Everything downstream reads that cache, so probe training takes hours on one GPU, conformal calibration minutes on CPU, and no baseline ever re-runs the base model. Dropping to one seed per member would cut the cache by a factor of L, but it also removes the $\bar{\sigma}^2/L$ correction of Appendix E.1: with $L=1$ the seed noise leaks into the epistemic teacher undiminished, which is the failure E4 is built to detect.

E IMPLEMENTATION DETAILS

E.1 UNCERTAINTY TEACHERS AND FINITE-SAMPLING CORRECTIONS

The two sources of randomness here have to be kept apart. A parametric source produces K members $\{\theta_m\}$ from an approximate posterior, realized as LoRA adapters (Hu et al., 2022) or MC-dropout (Gal & Ghahramani, 2016). A generative source is the sampling noise of the diffusion base at *fixed* parameters, probed with L noise seeds. With μ_m and σ_m^2 the sample mean and variance of member m over its L seeds and $\bar{\mu} = \frac{1}{K}\sum_m \mu_m$, the law of total variance, applied to the empirical distribution over the KL cached samples, gives the exact decomposition

$$\underbrace{\mathrm{Var}[\hat{x}_h(u)]}_{(U^{\mathrm{total}})^2} = \underbrace{\tfrac{1}{K}\textstyle\sum_m \sigma_m^2}_{(U^{\mathrm{ale}})^2} + \underbrace{\tfrac{1}{K}\textstyle\sum_m (\mu_m - \bar{\mu})^2}_{(U^{\mathrm{epi}})^2}. \tag{17}$$

The teachers are the square roots of these terms, so both are deviation magnitudes in the units of ϵ and the two branches share one scale. The uniform law on K sampled members stands in for the posterior over θ. Because the member means are themselves estimated from L seeds, the plug-in epistemic term is biased, with expectation $\frac{K-1}{K}$ times the posterior between-member variance, plus an additive $\frac{K-1}{K} \cdot \bar{\sigma}^2/L$ leaked from the seed noise. We undo both before caching. We use "aleatoric" for the generator's predictive spread across noise seeds at fixed parameters. E4 compares the branches empirically to assess this interpretation (Kendall & Gal, 2017; Depeweg et al., 2018).

E.2 HYPERPARAMETERS

Table 7 lists the final configuration, which is the one every reported number was produced under; where we swept a value, the range appears next to it. The loss weights were set so that the terms contribute comparable gradient magnitudes at initialization rather than tuned per dataset, and the same values are used for every ported baseline that has the corresponding term.

The probe is a 2-layer MLP with hidden width 512 and about 0.8M parameters. A 2-layer transformer with $d_{\mathrm{model}} = 384$ is nearly five times larger and changes AF-fail AUROC by less than the seed spread, so we use the MLP. The inverse dynamics window must resolve velocity: $k = 2$ is insufficient for the fastest Bridge motions, while $k = 3$ and $k = 4$ give similar results; we use $k = 3$. We fix the log-normal shape parameter ς because fitting it allows the model to widen its likelihood. Across $\varsigma \in [0.35, 0.7]$, the fitted scale changes by less than 5%, with no measurable change in coverage or AF-fail AUROC.

E.3 PROBE DEPLOYMENT AND CACHED TEACHERS

The probe's approximately 0.8M parameters are less than 1% of the 154M-parameter denoiser. Its weights are shared across tokens, with softplus outputs in the units of ϵ.

Both teachers are computed once, offline, during the rollout caching pass, with $K = 4$ parameter-perturbed members each rolled out under $L = 8$ diffusion seeds, so the cache holds $KL = 32$ rollouts per segment. Member 1 is the unperturbed base model, which lets the deployed rollout serve as one of its seeds instead of costing an extra pass. Deployment is a single forward pass with no sampling, at the inference cost of a pixel-confidence head.

E.4 SCORING PARAMETERIZATION

Since $\epsilon_h \geq 0$ is a magnitude rather than a signed residual, a zero-mean Gaussian likelihood is inappropriate. Equation (8) uses $\epsilon_h(u) + \delta \sim \mathrm{LogNormal}(\log(\hat{U}_h^{\mathrm{ale}}(u) + \delta) - \varsigma^2/2,\ \varsigma)$, so that $\hat{U}_h^{\mathrm{ale}}(u) + \delta$ is the mean of the shifted magnitude $\epsilon_h(u) + \delta$.

Retaining the terms free of ϕ gives the full NLL within the family (Gneiting & Raftery, 2007). The mean rather than the median is what makes $\hat{U}^{\mathrm{ale}}$ identifiable. Under the median parameterization this term and the distillation term in §3.2 pull one output towards values a factor $e^{\varsigma^2/2}$ apart, so the fitted scale would follow their relative weighting and not the data. For the same reason the aleatoric teacher of Eq. (17) is rescaled into a mean magnitude, by the fixed factor the log-normal model

Table 7: Final configuration. Swept ranges in parentheses.

Symbol	Meaning	Value
λ_{dist}	attribution distillation weight	0.5
λ_h	horizon calibration weight	1.0
λ_s	spatial localization weight	0.5
λ_a	action-fidelity alignment weight	1.0, warmup to 2.0
η	horizon monotonicity regularizer	0.1
η_{tv}	total variation weight	0.01
ζ	ranking term weight	0.5
β	Huber transition point	1.0 on normalized latents
q	target quantile	0.9
ς	log-normal shape parameter	0.5 (swept $[0.35, 0.7]$)
δ	conformal numerical floor	10^{-3}
L	diffusion seeds, aleatoric teacher	8
K	parameter-perturbed members, epistemic teacher	4
	rollouts cached per segment	$KL = 32$
k	inverse dynamics window length	3 (swept $\{2, 3, 4\}$)
H	rollout horizon	16
α	conformal significance	0.1
n_h	calibration segments per (domain, horizon)	500
	probe	2-layer MLP, width 512 (≈ 0.8M)
	optimizer	AdamW, lr 3×10^{-4}, wd 10^{-2}
	training	30k steps, batch 32 segments

implies, before it is used as a target. The shape ς is held fixed; its sensitivity is reported with the hyperparameters.

E.5 Units and scaling

Training and calibration use normalized errors: the cache divides ϵ by a fixed scale to make loss magnitudes comparable across data sources. The prediction $\hat{U}_h$ and offset δ use the same normalized error unit. The ratio $s_h = \epsilon_h/(\hat{U}_h + \delta)$ and its conformal quantile $\hat{Q}^{(h)}$ are dimensionless. To express the deployed region $C_\alpha^{(h)}(u)$ in raw latent units, we multiply its normalized radius by the stored scale, which is saved alongside the probe weights. A mismatch in this conversion would rescale the deployed region without appearing in metrics computed entirely in normalized units.

E.6 Guarantee labelling in reported results

Every coverage number we report is tagged with whether Proposition 1 covers it. Guaranteed under its assumptions: marginal segment-level coverage per (domain, horizon), using calibration segments from independent trajectories, and mixtures over test horizons selected independently of the calibration draws (Remark 1). Reported but not guaranteed: per-token and simultaneous per-pixel coverage under the default aggregator, coverage under distribution shift, and coverage from a single threshold calibrated on a bucket of horizons rather than on one. For the shifted regime we instead report the long-run time-average coverage of the adaptive methods, which is the guarantee those methods actually provide.