BACKWARD LATENT CERTAINTY: WHEN MORE CERTAIN FORECASTS LIE FARTHER

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual world models predict how scenes evolve to support physical reasoning and planning. Forecasting in a frozen encoder's feature space lets them reuse pretrained representations without reconstructing future images. This forecast can follow the future motion of every object still in view. The view cannot keep every object inside it. After an object slips out and the image no longer captures it, the forecast continues. The predictor still reports a confidence with that continued forecast. That confidence does not identify an accurate physical position. The forecasts it treats as more certain lie farther from the true position. We call this failure *backward latent certainty*. We measure this on VideoSAUR and SAVi, encoders that world models widely use and that we keep frozen. On CLEVRER, take the objects that have left the view and keep the forecasts the model calls safest. However many are kept, their average position error is higher than if every such forecast had been kept. Averaged over every number kept, it is 19.2% higher on VideoSAUR and 23.9% higher on SAVi. Mix in the objects that are still visible, and the same rule keeps forecasts with lower position error. We split position error into the error of the predicted features and the error of reading a position from the true features. The published score ranks the first error in the right direction. The second error is the larger one after the object leaves the view, and the score does not contain it. A Gaussian example shows that encoder collapse is enough to make the more certain forecasts farther from the true position, even when the latent predictor is exactly right. We leave those position forecasts unchanged. Using the true positions, we correct that confidence. Averaged over every number kept, the forecasts the corrected confidence keeps have mean position error 19.4% lower on VideoSAUR and 19.5% lower on SAVi than the forecasts the original confidence keeps.

1 INTRODUCTION

Visual world models learn how scenes evolve so that agents can anticipate object motion and evaluate possible actions. An encoder maps visual observations into latent features, and a dynamics model predicts how those features will change. DINO-WM and V-JEPA 2-AC build predictors on frozen encoders, reusing pretrained visual representations without reconstructing future images. This design supports planning toward image goals in DINO-WM and robot pick-and-place tasks in V-JEPA 2-AC (Zhou et al., 2024a; Assran et al., 2025).

The usefulness of feature prediction depends on what its targets preserve. After an object disappears, confidence in a frozen-feature forecast does not identify an accurate physical position. The forecasts it treats as more certain lie farther from the true position. We call this failure *backward latent certainty*. Position-bearing information can contract while features remain predictable (Figure 1), so the predictor can be certain about features that poorly specify an object's location.

A confidence filter turns this distinction into a decision with a measurable cost. In a tracking pipeline, it could decide which extrapolated positions to pass onward while observations are unavailable. Favoring inaccurate positions would then leave a worse accepted set despite rejecting more forecasts.

On out-of-view CLEVRER objects, projected risk and position error rank in opposite directions. The forecasts are ordered by that risk from low to high. At every acceptance count, those kept at the low-risk end have higher mean position error than the full set. Averaged over every acceptance count, the

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

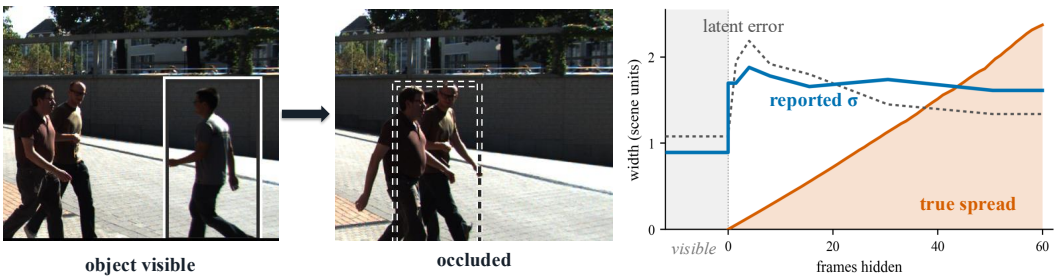

Figure 1: **Feature predictability need not imply position reliability.** The photographs illustrate an object before and after occlusion. The dashed box marks its hidden position. After the object disappears, feature uncertainty can stay low.

gap is 19.2% for VideoSAUR and 23.9% for SAVi. Pooled across visibility states, the same score still lowers accepted error. An aggregate evaluation can therefore conceal harmful selection within the unobserved population.

To locate the missing physical error, we separate decoded prediction error from readout error on the true future representation, retaining their cross term. Projected variance ranks the first component positively even where it inversely ranks total position error. The second component dominates after loss of view. A Gaussian construction shows how encoder collapse can produce this mismatch even with a correctly specified latent predictor.

With the forecasts held fixed, the accepted error depends on the score. Using the true positions, we correct that confidence. Averaged over every acceptance count, the corrected confidence lowers the error of the forecasts it keeps by 19.4% on VideoSAUR and 19.5% on SAVi.

We make three contributions.

- We find backward latent certainty in the systems we test. We measure its selection cost on records that have left view. A filter can accept fewer position forecasts and still raise their mean error.
- We locate the main physical error that latent covariance projection omits. We use matched records for that split. Projected variance still ranks decoded prediction error in the right direction. A Gaussian analysis shows how encoder collapse produces this mismatch.
- We correct the confidence with world-error supervision and leave every forecast unchanged. The corrected confidence lowers the error of the forecasts it keeps by 19.4% and 19.5%. The correction also improves ranking on autoregressive rollouts and on IntPhys.

2 Related Work

I-JEPA and V-JEPA learn representations by feature prediction (Assran et al., 2023; Bardes et al., 2024). DINO-WM and V-JEPA 2-AC forecast frozen features (Zhou et al., 2024a; Assran et al., 2025). DreamerV3 and TD-MPC2 instead update the encoder in the same training loop as the dynamics (Hafner et al., 2025; Hansen et al., 2024). Variational JEPA, Branch-JEPA, HAUWM, and UWM-JEPA model latent distributions, multiple successors, horizon-dependent uncertainty, or beliefs during rollout (Huang, 2026; Song et al., 2026; Wan et al., 2026; Radha & Goktas, 2026). Several also assess physical targets. Branch-JEPA scores decoded trajectories, and UWM-JEPA probes hidden state. Our question is whether a fixed Gaussian predictor's uncertainty misranks physical errors and whether its forecast selection can improve without retraining.

The reliability of an uncertainty estimate depends on its prediction target. Proper scoring rules, heteroscedastic losses, ensembles, and post-hoc calibration improve uncertainty estimates for a specified target (Malik et al., 2019; Seitzer et al., 2022b; Lakshminarayanan et al., 2017; Guo et al., 2017; Kuleshov et al., 2018). C^3 learns dense uncertainty for generated video and maps latent confidence into pixel-space heatmaps (Mei et al., 2025). We evaluate object position, including when it is unobserved, which differs from visible image fidelity. Object-centric models make this dis-

tinction measurable through matched object-level forecasts (Kipf et al., 2021; Elsayed et al., 2022; Zadaianchuk et al., 2023; Seitzer et al., 2022a; Wu et al., 2022; 2023). Over-compressed top-layer representations make classifier confidence harder to recalibrate post hoc (Wang & Zhang, 2024), and related work studies rate constraints and probabilistic representations (Apostolopoulou et al., 2024; Kirchhof et al., 2023a;b). Here the compressed representation is itself the prediction target, so latent uncertainty can rank its own errors positively while inversely ranking errors in a decoded physical quantity. We study how this affects the errors admitted by confidence filtering when objects leave the camera view.

Selective prediction evaluates the errors remaining after uncertain predictions are rejected (Geifman & El-Yaniv, 2019; Zhou et al., 2024b). At equal rejection budgets, scores can retain substantially different errors. Conformal risk control also uses evaluation-side outcomes to calibrate decisions (Angelopoulos et al., 2024; Farinhas et al., 2024). We study the quality of the ordering used to select predictions, which is separate from the guarantees conformal methods provide under their stated assumptions.

3 Task Formulation

3.1 Forecasts and physical-state targets

The forecasting pipeline is

$$x_t \xrightarrow{\text{render}} o_t \xrightarrow{E} z_t, \qquad (\hat{z}_{t+1}, \hat{\Sigma}_{t+1}) = P(\mathcal{C}_t). \tag{1}$$

Here x_t is the full scene and $s_t = f(x_t) \in \mathbb{R}^2$ is the position of the scored object. We suppress its index, so $z_t \in \mathbb{R}^D$ is one object latent, computed deterministically at evaluation. The predictor uses the scene history $\mathcal{C}_t$, containing all slots in its input window, to return that object's next-latent mean and diagonal covariance. In the main experiments, a frozen linear readout $g(z) = A^\top z + b$, with $A \in \mathbb{R}^{D\times 2}$ and $b \in \mathbb{R}^2$, produces the position forecast $\hat{s}_{t+1} = g(\hat{z}_{t+1})$. Rendering and encoding depend on appearance as well as position, so z_t may lose information about s_t and is not a function of position alone.

3.2 Evaluating uncertainty

We measure squared errors in each space as

$$\varepsilon^z_{t+1} = \|\hat{z}_{t+1} - z_{t+1}\|^2, \qquad \varepsilon^s_{t+1} = \|g(\hat{z}_{t+1}) - s_{t+1}\|^2. \tag{2}$$

Suppressing the next-step subscript $t+1$, the latent and world scores are

$$\hat{v}^z = \operatorname{tr}\hat{\Sigma}, \qquad C_0 = A^\top \hat{\Sigma} A, \qquad \hat{v}_0 = \operatorname{tr} C_0 = \sum_d \hat{\sigma}^2_d \|A_{d\cdot}\|^2, \tag{3}$$

where $A_{d\cdot}$ is row d of A. Each space is evaluated with its own score and error. We reserve $\hat{\sigma}^2_d$ for a variance component and $\hat{v}$ for scalar scores. Eq. 3 applies to linear readouts, while the dense-feature and single-step IntPhys audits use scalar proxies.

Spearman correlation $\rho(\hat{v}, \varepsilon)$ measures risk ranking. We call $\rho(\hat{v}_0, \varepsilon^s) < 0$ *backward latent certainty* on the specified evaluation subset. It is a failure of risk ranking, and the magnitude of $\hat{v}$ is evaluated separately. To measure the cost of using a score, let $S_c(\hat{v})$ contain its lowest-scoring nominal fraction c of evaluation records and define selective risk

$$\mathcal{R}(c; \hat{v}) = \frac{1}{|S_c(\hat{v})|} \sum_{i \in S_c(\hat{v})} \varepsilon^s_i. \tag{4}$$

Finite-sample acceptance counts and ties follow Eq. 10. AURC averages the retained error over every acceptance count from one record to all of them, so it fixes no acceptance fraction. Its keep-all value equals keep-all MSE. We evaluate it directly, since a negative rank correlation alone need not imply an increase in mean squared error after selection. For magnitude, let $y_i = \log(\delta + \varepsilon^s_i)$, with $\delta = 10^{-6}$, and $\ell_i = \log \hat{v}_i$. We report

$$R^2_{\log} = 1 - \frac{\sum_i (y_i - \ell_i)^2}{\sum_i (y_i - \bar{y})^2}. \tag{5}$$

A negative $R^2_{\log}$ means worse log-error prediction than $\bar{y}$, not negative rank correlation. Global scaling separates an overall score offset from input-dependent gains.

3.3 EVALUATION PROTOCOLS

We distinguish an object leaving the view from inter-object occlusion. CLEVRER annotates whether each object is inside the view, so its gap is the number of raw frames since it was last in view. Single-step forecasts are evaluated on these naturally occurring out-of-view intervals, and the separate IntPhys audit uses its occlusion annotations. During autoregressive rollout, we recursively feed predicted latents back to P for $h = 1, \ldots, 10$ steps. The uncertainty head emits a local conditional covariance at each step without propagating it, so we evaluate it as a score of the resulting forecast rather than as the marginal covariance of the full rollout.

4 BACKWARD LATENT CERTAINTY

4.1 CONFIDENCE CAN FAVOR LARGER WORLD-STATE ERRORS

On CLEVRER, the projected world score reverses the position-error ranking after objects leave the camera view. Four of six out-of-view bins show backward latent certainty on VideoSAUR and all six show it on SAVi (Table 1). SAVi shows it immediately after view is lost. VideoSAUR shows it from gap 6 to 10. The sign change therefore occurs within gap bins, not only after pooling different durations. On held-out records with gap ≥ 3, pooled ρ is -0.28 for VideoSAUR and -0.53 for SAVi.

Table 1: **Projected uncertainty can reverse the position-error ranking.** Spearman $\rho(\hat{v}_0, \varepsilon^s)$ on held-out CLEVRER records at $\beta = 0.5$. Gap counts raw frames since last in view. Full video-bootstrap intervals are in Table 6.

Encoder	0	1–2	3–5	6–10	11–20	21–40	41+
VideoSAUR	$+0.38$	$+0.71$	$+0.42$	-0.27	-0.32	-0.16	-0.26
SAVi	$+0.17$	-0.42	-0.58	-0.45	-0.49	-0.50	-0.33

Using the score to select forecasts exposes a concrete cost. On the out-of-view selection population, retained error exceeds keep-all error at every acceptance count below full retention (Figure 4, Table 9). Averaged over every acceptance count, the retained error is 17.42 (CI $[16.00, 18.80]$) against keep-all 14.61 (CI $[13.58, 15.62]$) for VideoSAUR, and 11.13 (CI $[10.07, 12.21]$) against 8.99 (CI $[8.21, 9.73]$) for SAVi. Pooling visibility states reverses this aggregate conclusion. Averaged over every acceptance count, selection lowers MSE from 0.513 and 0.860 to 0.082 and 0.773 for VideoSAUR and SAVi, respectively. In-view records dominate that population, and the harmful selection occurs within out-of-view records.

Two assignment controls preserve the pooled backward latent certainty. Excluding possible slot takeover leaves negative point estimates in the affected bins (Table 8). Fixing assignments at the last in-view frame available by the final context frame also avoids future-dependent matching. Pooled gap-≥ 3 correlations remain negative at -0.24 (CI $[-0.32, -0.15]$) and -0.49 (CI $[-0.55, -0.41]$), respectively. For SAVi, last-context matching changes 873 of 1614 out-of-view assignments (54%), yet bin correlations remain between -0.51 and -0.25.

A dense DINOv2 audit extends the target comparison beyond slots. A predictor forecasts patch features, and learned queries decode positions (held-out $R^2 = 0.77$) and supply attention-weighted token variance as a risk proxy. On identical out-of-view records, latent correlations are $+0.56$ to $+0.78$, versus -0.19 to -0.43 for position, whose six intervals exclude zero (Figure 2c).

4.2 ENCODER COLLAPSE AND THE MISSING PHYSICAL ERROR

An exact decomposition identifies the physical error omitted by projected covariance. Write $z = z_{t+1}$, $\hat{z} = \hat{z}_{t+1}$, and $s = s_{t+1}$. For any readout g, defining $e_P = g(\hat{z}) - g(z)$ and $e_g = g(z) - s$ gives the exact, generally non-orthogonal decomposition

$$\varepsilon^s = \underbrace{\|e_P\|^2}_{\text{prediction}} + \underbrace{\|e_g\|^2}_{\text{readout}} + \underbrace{2e_P^\top e_g}_{\text{cross term}} . \tag{6}$$

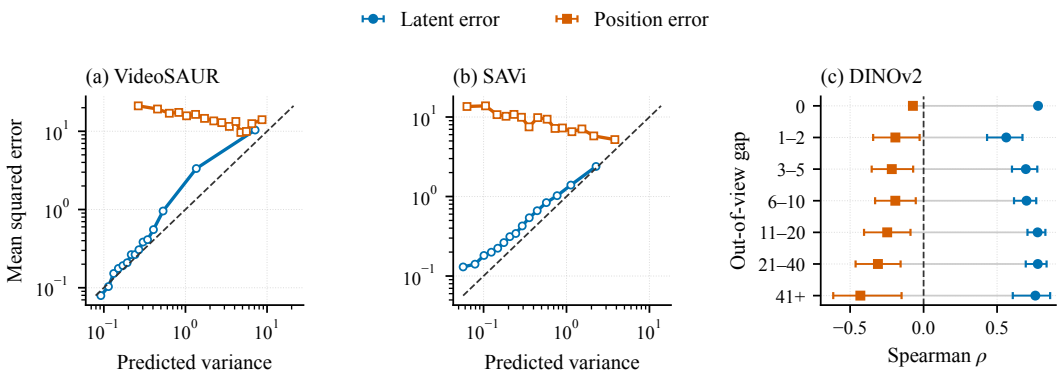

Figure 2: **The target changes what uncertainty can rank.** (a,b) Predicted variance versus binned mean squared error for VideoSAUR and SAVi. Blue uses all-frame latent targets, and orange uses out-of-view position targets. The diagonal denotes equal variance and error, and these curves use different subsets. (c) DINOv2. Both correlations within a gap bin use the same object-frame samples. Horizontal bars are video-bootstrap 95% intervals, and sample counts are in Table 10.

When $\hat{z}$ and $\hat{\Sigma}$ are the true conditional mean and covariance, a linear g yields

$$\mathbb{E}[\|e_P\|^2 \mid \mathcal{C}_t] = \mathrm{tr}(A^\top \hat{\Sigma} A) = \hat{v}_0.$$

Projected variance thus models decoded prediction error. A distribution over z alone supplies neither the readout residual nor the cross term. Linearity is needed for this projection, not Eq. 6. For diagonal $\hat{\Sigma}$, the identity also requires conditionally uncorrelated latent coordinates.

The prediction component remains positively ranked in every gap bin for both encoders, with all video-bootstrap intervals above zero (Table 7). For example, VideoSAUR at gap 11 to 20 has $\rho(\hat{v}_0, \|e_P\|^2) = 0.69$ while $\rho(\hat{v}_0, \varepsilon^s) = -0.32$. Its mean prediction, readout, and cross terms are 0.66, 12.93, and -1.10, summing to position MSE 12.48.

The true future encoding is also harder to decode than the forecast under the same fixed readout. In every out-of-view bin, the mean readout error exceeds total forecast MSE, and $f_g = \mathbb{E}\|e_g\|^2 / \mathbb{E}\varepsilon^s$ ranges from 1.02 to 1.80, versus 0.32 and 0.86 in view for VideoSAUR and SAVi, respectively. This comparison exposes degraded position fidelity of the encoded target, consistent with collapse.

Information loss alone does not establish backward latent certainty. Under independence within a gap bin, nonconstant score and error ranks have zero population correlation. The bins with backward latent certainty instead have intervals below zero (Table 6), linking lower variance systematically to larger error.

A Gaussian construction establishes encoder collapse as a sufficient mechanism. Suppose a hidden two-dimensional position has covariance $u(\tau)I_2$, while its encoded position signal contracts by $a(\tau)$ amid independent nuisance variation of covariance $\nu^2 I_2$. With a fixed identity readout and a perfectly specified latent predictor,

$$\hat{v}_0(\tau) = 2\{a(\tau)^2 u(\tau) + \nu^2\}, \qquad \mathbb{E}[\varepsilon^s \mid \tau] = 2u(\tau). \tag{7}$$

Here τ denotes time without observation. If u grows while $a^2 u$ decreases, predicted risk falls as expected position error rises. Within a gap, $u_2 > u_1$ but $a_2^2 u_2 < a_1^2 u_1$ also reverses expected-risk order at fixed ν. More directly, the positional mutual information I_τ satisfies $\hat{v}_0 = 2\nu^2 e^{I_\tau}$, while the Bayes decoding error from the true future feature is $B_\tau = 2u(\tau)e^{-I_\tau}$. Lower latent variance can therefore accompany less position information and a larger irreducible decoding error. The readout term $\mathbb{E}\|e_g\|^2$ equals B_τ plus a nonnegative readout excess. Appendix C derives these relations. Gap-bin means follow the qualitative trend. From the shortest out-of-view bin to the longest, variance falls from 3.63 to 2.54 on VideoSAUR and from 1.48 to 0.54 on SAVi, while MSE rises from 4.33 to 24.73 and from 4.42 to 15.54, respectively (Table 7).

5 RISK SCORING WITH FIXED FORECASTS

Residual physical uncertainty is not identifiable from latent trajectories alone (Appendix C, Proposition). We therefore supply position labels offline while freezing E, P, g, and forecasts. Figure 3 separates fitting from runtime scoring.

5.1 RESIDUAL RISK SCORING

For the linear readouts used in the correction experiments, define

$$\log \hat{v}_R = \log \hat{v}_0 + R_E(\hat{z}, h), \qquad \hat{v}_0 = \mathrm{tr}(A^\top \hat{\Sigma} A). \tag{8}$$

The corrected score is positive and leaves $g(\hat{z})$ unchanged. A constant residual preserves the ranking, while conditioning on $\hat{z}$ and rollout step h allows forecasts to be reordered. Keeping $\log \hat{v}_0$ as an offset lets Table 3 attribute ranking changes to global, clock, or latent terms. An ablation fits the same world-error target without the offset.

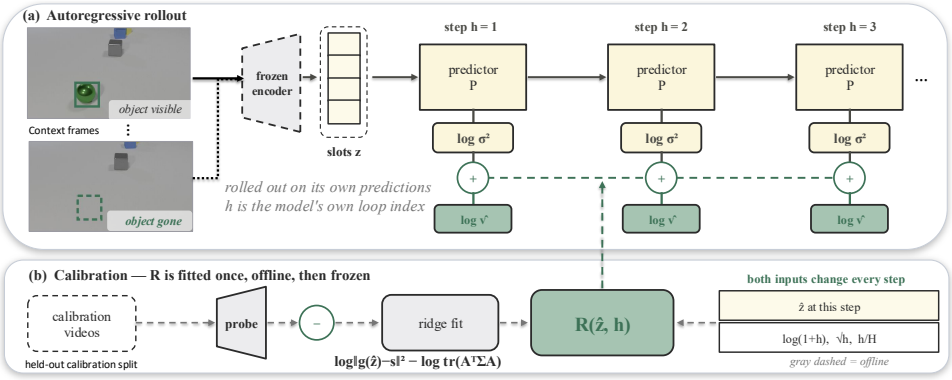

Figure 3: **Offline calibration, stepwise correction.** (a) At rollout step h, a frozen residual model corrects the risk score without changing the forecast. (b) State annotations supervise the offline ridge fit using the forecast $\hat{z}$ in Eq. 9. The diagram's R, Σ, $\log \sigma^2$, and $\log \hat{v}$ correspond to R_E, $\hat{\Sigma}$, $\log \hat{v}_0$, and $\log \hat{v}_R$. The diagram's H denotes the horizon $h_{\max}$, and h/H abbreviates $\min(h, h_{\max})/h_{\max}$. Single-step calibration uses a separate model without clock features.

5.2 OFFLINE FIT AND RUNTIME INPUTS

On calibration record i, the target is

$$r_i = \log(\delta + \varepsilon_i^s) - \log \hat{v}_{0,i}, \qquad \delta = 10^{-6}. \tag{9}$$

We fit ridge regression with $\lambda = 1$ to features $[\hat{z}, \log(1+h), \sqrt{h}, \min(h, h_{\max})/h_{\max}]$, where $h_{\max} = 10$ is the fit and test rollout horizon. Single-step scoring uses a distinct ridge on $\hat{z}$ alone, without visibility or gap features.

The residual uses $\hat{z}$ and, for rollout, h. It targets empirical log error.

The ablation without the offset predicts $\log(\delta + \varepsilon^s)$ from the same features. A readout-residual baseline predicts $\log(\delta + \|g(z) - s\|^2)$ from $\hat{z}$ using training-set position annotations and forms $\hat{v}_{\mathrm{RO}} = \hat{v}_0 + \exp(q(\hat{z}))$. Log fitting and the omitted cross term keep it from estimating total MSE without bias.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

The evaluation covers CLEVRER (Yi et al., 2019) and IntPhys (Riochet et al., 2018). On CLEVRER the out-of-view selection sets are 1267 VideoSAUR records and 1387 SAVi records. We freeze

VideoSAUR and SAVi encoders and train heteroscedastic masked slot transformers with β-NLL at $\beta = 0.5$. For rollout scoring, VideoSAUR predictors are also trained at $\beta \in \{0, 1\}$. The predictors are not conditioned on actions. A linear position readout is fitted on in-view training positions and frozen. Position labels supervise the readout and the risk regressors, and match each forecast to its object. Tables use the fixed predictor seed 42, with video-bootstrap intervals. Appendix B reports rollout correlations for predictors trained at seeds 43 and 44.

The selection audit uses a fixed 1500-video cohort with two-fold video cross-fitting, so each record is scored by a regressor fitted on the other 750 videos and all scores select from the same 1267 VideoSAUR and 1387 SAVi out-of-view records. We use 2000 paired video-bootstrap resamples for selection differences, refitting the cross-fitted scores in each replicate. Gap and rollout intervals condition on their fitted scoring models.

6.2 Reducing the error admitted by the filter

World-error supervision reduces the original selector's error on the same out-of-view population (Table 2). It keeps less error than the published score at every acceptance count below full retention (Table 9). Residual scoring lowers AURC from 17.42 to 14.03 and from 11.13 to 8.96, reductions of 19.4% and 19.5% with every forecast fixed, with paired difference intervals $[-4.27, -2.45]$ and $[-2.78, -1.55]$ (Figure 4).

Table 2: **Risk scores on identical out-of-view forecasts.** AURC point estimates, where smaller is better. Keep-all AURC equals keep-all MSE, 14.61 and 8.99 for VideoSAUR and SAVi, respectively. Log-affine and world-error fits use video cross-fitting. Readout heads are fit on the training positions.

Risk score	VideoSAUR AURC↓	SAVi AURC↓
Projected variance $\hat{v}_0$	17.42	11.13
Log-affine score	17.43	11.13
Readout residual (in-view train)	17.27	10.58
Readout residual (all train)	16.79	10.52
World-error residual $\hat{v}_R$	14.03	8.96
without offset (ablation)	12.90	8.72

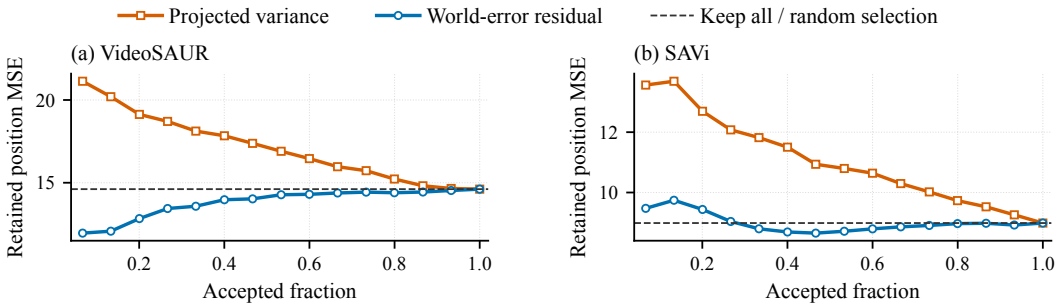

Figure 4: **Selection on the identical out-of-view population.** Each point accepts the same number of forecasts under the two scores. The dashed line is keep-all MSE, also the expected MSE of uniform random selection at any fixed acceptance count. Curves show point estimates.

6.3 Comparing risk scores

The ablation that fits the same target without the offset reaches AURC 12.90 and 8.72 for VideoSAUR and SAVi, respectively. World-error supervision therefore lowers AURC below every baseline in Table 2 with or without the offset. Removing the offset lowers AURC further only for VideoSAUR, with a paired difference of -1.13 (CI $[-1.84, -0.48]$). The two are comparable on SAVi, with CI $[-0.54, 0.12]$, and on autoregressive rollouts (Table 3). Keeping the offset makes the gain decomposable, and a zero residual returns the published ranking. In view, the corrected score also lowers AURC below the published score (Table 4).

Within gap bins, the correction raises the rank correlation in 10 of 12 out-of-view bins, with paired intervals above zero (Table 6). These include all six SAVi bins. VideoSAUR gap 6 to 10 rises from -0.27 to $+0.39$ (corrected CI $[0.25, 0.52]$), and gap 11 to 20 rises by 0.30 (change CI $[0.17, 0.42]$).

6.4 Scoring longer autoregressive forecasts

Rollout tests scoring when the uncorrected ranking is already positive on CLEVRER. Table 3 separates global, clock, and latent terms. On VideoSAUR, global scaling raises $R^2_{\log}$ from 0.089 to 0.325 without changing ranking. The combined correction reaches 0.547 and raises ρ from 0.491 to 0.731, and SAVi changes from 0.174 to 0.474.

Table 3: **Scoring fixed autoregressive forecasts.** CLEVRER at $\beta = 0.5$, where higher is better. All rows use the exact projected trace. Table 5 gives record counts.

	VideoSAUR		SAVi	
Risk score	$\rho \uparrow$	$R^2_{\log} \uparrow$	$\rho \uparrow$	$R^2_{\log} \uparrow$
Original $\hat{v}_0$	0.491	0.089	0.174	-0.244
Global scaling	0.491	0.325	0.174	-0.133
Clock $R_E(h)$	0.679	0.457	0.327	-0.012
Latent $R_E(\hat{z})$	0.627	0.459	0.367	0.112
Combined $R_E(\hat{z}, h)$	0.731	0.547	0.474	0.210
without offset (ablation)	0.729	0.550	0.492	0.239

AURC falls from 0.355 to 0.198 on VideoSAUR and from 1.381 to 0.966 on SAVi. For VideoSAUR, combined and clock-only AURC have 95% intervals $[0.188, 0.209]$ and $[0.271, 0.310]$, and SAVi gives $[0.918, 1.014]$ and $[1.115, 1.232]$. These separate intervals support a benefit from latent features in rollout, beyond the clock-only scores of 0.290 and 1.175, respectively. VideoSAUR's loss sweep preserves the combined correction's advantage over the original score at all three β values (Table 11).

6.5 Reusing the scoring intervention

Across VideoSAUR predictors sharing an encoder but trained at different β, off-diagonal transfer correlations range from 0.659 to 0.754 (Table 12).

We transfer the VideoSAUR encoder and $\beta = 0.5$ predictor used on CLEVRER to IntPhys (Riochet et al., 2018), fitting a linear readout and correction locally. On held-out IntPhys rollouts, ρ rises from -0.035 to 0.480 and $R^2_{\log}$ from -0.502 to 0.216. AURC falls from 0.186 to 0.106, below the clock-only 0.142. The backbone transfers, while the readout and risk model use target-dataset supervision.

7 Discussion

7.1 Why world-error supervision lowers the accepted error

The correction lowers accepted error because it supplies the error term that the published score omits. Projected variance models decoded prediction error and ranks it positively in every gap bin (Table 7). The omitted term is the readout of the true future features. Keeping $\log \hat{v}_0$ as an offset retains the correctly ranked term, and the residual fit to world error carries the omitted one.

The supervision target decides how much of the omitted error the score recovers (Table 2). Readout-residual heads add the readout term without the cross term, and their AURC stays close to the published score. World-error supervision lowers it, and the ablation without the offset shows that the gain comes from this target.

The omitted error grows with time out of view, as encoder collapse predicts. Mean projected variance falls across gap bins while position MSE rises (Table 7), and the Gaussian construction in Section 4.2 yields this pattern with an exactly specified latent predictor.

Table 4: **Scores and the information they use.** AURC point estimates on the out-of-view selection population and on in-view records of the same cohort, where smaller is better.

		Out of view		In view	
Score	Run-time input	VideoSAUR	SAVi	VideoSAUR	SAVi
Keep all	None	14.61	8.99	0.346	0.750
Projected variance $\hat{v}_0$	Forecast	17.42	11.13	0.054	0.666
World-error residual $\hat{v}_R$	Forecast	14.03	8.96	0.036	0.429
without offset (ablation)	Forecast	12.90	8.72	0.035	0.454
Fixed reversal	Direction set on the result	12.95	6.89	1.105	0.935
Time out of view	Target-time visibility	9.66	6.46	n.a.	n.a.

Two further rules in Table 4 confirm this account. Reversing the published score lowers out-of-view AURC, as expected when its ranking is backward there, but raises in-view AURC above keep-all on VideoSAUR and SAVi. Ordering forecasts by time out of view also lowers out-of-view AURC, as expected when the omitted error grows with that time. Both rules use information the predictor does not have at run time. The reversal direction is set after seeing the out-of-view result, and the time comes from the visibility annotation at the target frame. The corrected score uses only the forecast and lowers AURC both in view and out of view.

These results show that the correction works through the readout error missing from the published score, that recovering it requires supervision on total world error, and that this error follows the collapse of position information over time out of view.

7.2 What the correction requires and where it holds

The correction is fit offline on calibration videos. At run time the residual takes the forecast $\hat{z}$ and, during rollout, the step h.

Its gain is stable across predictor training. A residual fit on one VideoSAUR predictor keeps rollout correlations of 0.659 to 0.754 when applied to predictors trained at other β (Table 12). VideoSAUR's loss sweep keeps the advantage over the published score at all three β values (Table 11).

The gain extends beyond single-step forecasts on CLEVRER. Inside gap bins it raises the rank correlation in 10 of 12 out-of-view bins, including all six on SAVi (Table 6). On autoregressive forecasts the combined residual raises ρ from 0.491 to 0.731 on VideoSAUR and from 0.174 to 0.474 on SAVi. The same encoder and predictor, with the readout and residual fit on IntPhys labels, raise rollout ρ from -0.035 to 0.480.

These results show that the correction is fit offline, leaves every forecast unchanged, holds across training losses, and carries over to within-gap ranking, autoregressive rollout, and IntPhys.

8 Conclusion

A latent world model forecasts in a feature space, and its confidence is measured in that space. The physical state a task reads from those features lives in a different space. When the two spaces diverge, the forecast grows less accurate in position while the latent confidence stays high, and the forecasts it treats as more certain lie farther from the true position.

Widely used slot and dense encoders all show this failure on CLEVRER once objects leave the view. Averaged over every acceptance count, keeping the forecasts the confidence calls safest raises position error by 19.2% and 23.9% relative to keeping all of them. The confidence still ranks the error of the predicted features. The larger error comes from reading position out of the true features, and encoder collapse is enough to produce it. We keep every forecast fixed and correct the confidence with a residual fit to world-position error. The corrected confidence lowers that average error by 19.4% and 19.5%, and it also improves ranking on autoregressive rollouts and on IntPhys.

The audit already spans slot and dense encoders. Future work will extend the audit and the correction to encoders trained jointly with the dynamics and to more complex scenes.

AI USE STATEMENT

We have not used generative AI tools for any task with required disclosure. Additionally, we used generative AI tools to polish wording. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International conference on learning representations*, volume 2024, pp. 55198–55218, 2024.

Ifigeneia Apostolopoulou, Benjamin Eysenbach, Frank Nielsen, and Artur Dubrawski. A rate-distortion view of uncertainty quantification. *arXiv preprint arXiv:2406.10775*, 2024.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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.

Gamaleldin Elsayed, Aravindh Mahendran, Sjoerd Van Steenkiste, Klaus Greff, Michael C Mozer, and Thomas Kipf. Savi++: Towards end-to-end object-centric learning from real-world videos. *Advances in Neural Information Processing Systems*, 35:28940–28954, 2022.

António Farinhas, Chrysoula Zerva, Dennis Ulmer, and Andre Martins. Non-exchangeable conformal risk control. In *International Conference on Learning Representations*, volume 2024, pp. 50952–50966, 2024.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International conference on machine learning*, pp. 2151–2159. PMLR, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pp. 1321–1330. PMLR, 2017.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

Nick Hansen, Hao Su, and Xiaolong Wang. Td-mpc2: Scalable, robust world models for continuous control. In *International Conference on Learning Representations*, volume 2024, pp. 47376–47405, 2024.

Yongchao Huang. VJEPA: Variational joint embedding predictive architectures as probabilistic world models. *arXiv preprint arXiv:2601.14354*, 2026.

Thomas Kipf, Gamaleldin F Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional object-centric learning from video. *arXiv preprint arXiv:2111.12594*, 2021.

Michael Kirchhof, Enkelejda Kasneci, and Seong Joon Oh. Probabilistic contrastive learning recovers the correct aleatoric uncertainty of ambiguous inputs. In *International Conference on Machine Learning*, pp. 17085–17104. PMLR, 2023a.

Michael Kirchhof, Bálint Mucsányi, Seong Joon Oh, and Enkelejda Kasneci. URL: A representation learning benchmark for transferable uncertainty estimates. *Advances in Neural Information Processing Systems*, 36:13956–13980, 2023b.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *International conference on machine learning*, pp. 2796–2804. PMLR, 2018.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *International Conference on Machine Learning*, pp. 4314–4323. PMLR, 2019.

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*, 2025.

Santosh Kumar Radha and Oktay Goktas. Uwm-jepa: Predictive world models that imagine in belief space. *arXiv preprint arXiv:2605.25313*, 2026.

Ronan Riochet, Mario Ynocente Castro, Mathieu Bernard, Adam Lerer, Rob Fergus, Véronique Izard, and Emmanuel Dupoux. IntPhys: A framework and benchmark for visual intuitive physics reasoning. *arXiv preprint arXiv:1803.07616*, 2018.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, et al. Bridging the gap to real-world object-centric learning. *arXiv preprint arXiv:2209.14860*, 2022a.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. *arXiv preprint arXiv:2203.09168*, 2022b.

Zhi Song, Ximing Xing, Zhenchao Tang, Hanbo Huang, Jiehui Huang, Weilong Yan, Tianxu Lv, Minghao Yang, Zhongzheng Niu, Bing He, Lusheng Wang, and Jianhua Yao. Branch-jepa: Finite-support predictive distributions for jepa world models, 2026.

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.

Deng-Bao Wang and Min-Ling Zhang. Calibration bottleneck: Over-compressed representations are less calibratable. In *Forty-first International Conference on Machine Learning*, 2024.

Ziyi Wu, Nikita Dvornik, Klaus Greff, Thomas Kipf, and Animesh Garg. Slotformer: Unsupervised visual dynamics simulation with object-centric models. *arXiv preprint arXiv:2210.05861*, 2022.

Ziyi Wu, Jingyu Hu, Wuyue Lu, Igor Gilitschenski, and Animesh Garg. Slotdiffusion: Object-centric generative modeling with diffusion models. *Advances in Neural Information Processing Systems*, 36:50932–50958, 2023.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. CLEVRER: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019.

Andrii Zadaianchuk, Maximilian Seitzer, and Georg Martius. Object-centric learning for real-world videos by predicting temporal feature similarities. *Advances in neural information processing systems*, 36:61514–61545, 2023.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. DINO-WM: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024a.

Han Zhou, Jordy Van Landeghem, Teodora Popordanoska, and Matthew B Blaschko. A novel characterization of the population area under the risk coverage curve (aurc) and rates of finite sample estimators. *arXiv preprint arXiv:2410.15361*, 2024b.

A EVALUATION DETAILS AND SELECTIVE PREDICTION

The CLEVRER VideoSAUR and SAVi slot features are used without modifying the features or encoder weights. Predictor training uses the supplied training split.

The CLEVRER slot predictors use three observations, seven 128-dimensional slots, six transformer layers, sixteen attention heads, and feedforward width 2048. They are trained for 40 epochs with learning rate 5×10^{-4} and batch size 256. Training uses the β-NLL loss (Seitzer et al., 2022b), in which each latent dimension contributes $\lfloor \hat{\sigma}_d^{2\beta} \rfloor \{\frac{1}{2} \log \hat{\sigma}_d^2 + (\hat{z}_d - z_d)^2/(2\hat{\sigma}_d^2)\}$, where $\lfloor \cdot \rfloor$ stops gradients, and $\beta = 0$ recovers the Gaussian NLL. Predictor seeds are 42, 43, and 44, and seed 0 controls bootstrap resampling. The encoders and readouts remain fixed across seeds. Both evaluation protocols sample input frames at interval 5. Single-step and rollout windows have strides 7 and 17, and horizons 1 and 10. The gap always counts raw frames, before this temporal subsampling.

CLEVRER annotates at every frame whether each object lies inside the view. An eligible target must have an assigned slot and must previously have been in view. For each encoder, training alternates between per-video Hungarian assignments and a shared linear ridge readout using only matched in-view positions in 10,000 training videos. The fit uses $\lambda = 1$, a penalized intercept, four random assignment restarts, and at most fifteen iterations per restart, retaining the lowest-cost fit. Evaluation freezes that readout and computes one Hungarian assignment per video from its in-view position errors. That assignment stays fixed while the object is out of view, so slots are never rematched to hidden positions frame by frame.

We retain the lowest 75% of videos by alignment cost. The gap audit applies this rule to all validation videos and splits the retained videos by ID into calibration and evaluation halves. Its evaluation half contains 1875 videos for each encoder, with 1404 and 1614 out-of-view object-frames and 133,679 and 129,705 in-view frames for VideoSAUR and SAVi, respectively. It supplies the within-gap ranking and decomposition results in Tables 1, 6 to 8.

The selection audit (Table 2, Figure 4, and Table 9) starts from the first 2000 alignment videos and applies the 75% threshold within that cohort, leaving 1500 videos. All its records with gap ≥ 3 form the selection population of 1267 VideoSAUR and 1387 SAVi records, contributed by 263 and 280 videos, respectively. We hold both cohort thresholds fixed across score comparisons. Thus these counts are not sums of the held-out gap-table records. The selection cohort shares no videos with the gap-evaluation half. Its alignment filter precedes scoring, and both the original selection cost and all corrective scores use this same cohort. Each world-error regressor fits all eligible frames in the opposite 750-video fold, and neither visibility nor gap enters these heads. Table 5 summarizes all evaluation populations.

Table 5: **Evaluation populations.** Paired counts list VideoSAUR first and SAVi second unless noted. Slot-based CLEVRER populations keep the 75% of videos with the lowest alignment cost within their source set, and the gap-audit and selection cohorts share no videos.

Population	Videos and split	Records	Used in
Gap audit	Held-out half of the filtered validation set, 1875 videos	133,679 and 129,705 in view, 1404 and 1614 out of view	Tables 1, 6 to 8
Selection	First 2000 videos filtered to 1500, two folds of 750	1267 and 1387 at gap ≥ 3 from 263 and 280 videos, 108,254 and 105,478 in view	Tables 2, 9, Fig. 4
Rollout	Selection cohort, 750 fit and 750 evaluation videos	135,445 and 132,626 evaluation records	Tables 3, 11, 12
Dense DINOv2	Separate 2000 videos, 1000 train, 1000 evaluation	172,832 object-frames	Fig. 2c and Table 10
IntPhys rollout	1834 eligible videos, 917 fit, 917 evaluation	127,980 evaluation records (VideoSAUR)	Section 6.5

The readout-residual baselines use training scenes that are disjoint from these validation cohorts. The in-view version uses only the position annotations available for readout training. The all-training version also uses annotated out-of-view positions. Ridge fits use a sum-of-squared-errors objective, $\lambda = 1$, no feature standardization, and a penalized intercept. Video-bootstrap selection intervals resample the 1500 eligible videos and refit all cross-fitted heads with fold identities fixed. Every

evaluation video's labels remain excluded from its own fitted score. Upstream encoders, predictors, readouts, and training-fitted readout-risk heads remain fixed.

The matching sensitivity analysis replaces the whole-video assignment with one computed at the last in-view raw frame available at or before the final context frame. It never uses a post-context observation to choose the match. Every out-of-view gap-audit record has an eligible alternative assignment. The takeover control removes records whose matched slot decodes within 0.5 position units of another currently in-view object. Tables report the remaining counts because this control changes the evaluated population.

We evaluate the same acceptance budget under every score. For N records and nominal coverage $c_k = k/15$, let

$$m_k = \lceil (N-1)c_k \rceil \quad (k < 15), \qquad m_{15} = N, \qquad \mathcal{R}_k = \frac{1}{m_k} \sum_{i=1}^{m_k} \varepsilon^s_{(i)}. \tag{10}$$

Records are ordered by the score being tested. A tie at the boundary contributes its group-mean error with fractional weight, so all scores accept exactly m_k records even in bootstrap replicates. AURC is the mean retained error over every integer acceptance count from 1 to N, taking the expected order within tied groups. Intervals are percentile intervals from 2000 video resamples.

The DINOv2 audit uses 2000 CLEVRER videos, each frame encoded as 256 patch tokens of dimension 384 and sampled every third frame. The first 1000 videos train the predictor and query readout, and the other 1000 evaluate them. The predictor has four transformer layers, eight heads, feedforward width 1024, history 3, $\beta = 0.5$, 12 epochs, learning rate 3×10^{-4}, and batch size 32. The readout has ten learned queries and trains for eight epochs on in-view positions. A fixed query-object assignment in each video supplies correspondences. Point estimates and intervals come from one run with seed 0. Table 10 gives all counts and correlations.

The IntPhys rollout audit uses IntPhys slots and object alignments. It fits the linear readout on IntPhys and splits the 1834 eligible videos into 917 calibration and 917 evaluation videos. A separate single-step audit trains its predictor on IntPhys slots and maps features to position with a multilayer network. Its score is the mean latent variance, $D^{-1} \sum_d \hat{\sigma}^2_d$. At gaps 1 to 2, 3 to 5, 6 to 10, and 11 to 20, its world-error correlations are -0.11, -0.11, -0.10, and -0.09, with video-level intervals excluding zero, and latent correlations range from $+0.68$ to $+0.77$. This audit measures the rank correlation of that score with position error and with latent error.

B Complete results and controls

Tables 6 to 8 report complete within-gap rankings, error components, and assignment controls with video-bootstrap intervals.

The decomposition also holds on the separate selection population. Pooled prediction-error correlations are 0.679 $[0.643, 0.713]$ and 0.522 $[0.461, 0.576]$, and readout-error correlations are -0.224 $[-0.297, -0.149]$ and -0.372 $[-0.442, -0.292]$. The readout fractions are 1.076 $[1.058, 1.094]$ and 1.028 $[1.014, 1.041]$. These video-bootstrap intervals reproduce the component distinction in the population used to measure selection costs.

Table 6: **Complete single-step ranking results.** n counts object-frames. ρ is $\rho(\hat{v}_0, \varepsilon^s)$ and $\Delta\rho$ is the change under the corrected score, with video-level 95% intervals. $\Delta\rho$ uses paired resampling. Green marks a change whose interval lies above zero.

	VideoSAUR			SAVi		
Gap	n	ρ [CI]	$\Delta\rho$ [CI]	n	ρ [CI]	$\Delta\rho$ [CI]
0	133,679	$+0.38$ $[+0.37, +0.39]$	$+0.09$ $[+0.08, +0.10]$	129,705	$+0.17$ $[+0.16, +0.18]$	$+0.38$ $[+0.37, +0.39]$
1–2	103	$+0.71$ $[+0.61, +0.79]$	$+0.18$ $[+0.10, +0.26]$	120	-0.42 $[-0.55, -0.27]$	$+0.51$ $[+0.34, +0.69]$
3–5	155	$+0.42$ $[+0.26, +0.56]$	$+0.42$ $[+0.30, +0.54]$	169	-0.58 $[-0.66, -0.48]$	$+0.61$ $[+0.47, +0.75]$
6–10	216	-0.27 $[-0.41, -0.13]$	$+0.66$ $[+0.55, +0.76]$	247	-0.45 $[-0.54, -0.34]$	$+0.50$ $[+0.38, +0.63]$
11–20	364	-0.32 $[-0.42, -0.22]$	$+0.30$ $[+0.17, +0.42]$	410	-0.49 $[-0.57, -0.40]$	$+0.55$ $[+0.43, +0.66]$
21–40	381	-0.16 $[-0.29, -0.03]$	$+0.09$ $[-0.05, +0.22]$	449	-0.50 $[-0.60, -0.38]$	$+0.47$ $[+0.30, +0.63]$
41+	185	-0.26 $[-0.43, -0.06]$	$+0.08$ $[-0.12, +0.28]$	219	-0.33 $[-0.53, -0.11]$	$+0.55$ $[+0.29, +0.78]$

Table 7: **Prediction and readout error have different rankings.** Same records as Table 6. Correlations use projected variance and include video-bootstrap 95% intervals, and $\bar{v}_0$ is mean projected variance. f_P, f_g, and $f_\times$ divide mean components by mean total MSE and sum to one. When $f_g > 1$, decoding true future features has higher mean error than decoding forecasts under the same readout.

Gap	ρ_P [CI]	ρ_g [CI]	$\bar{v}_0$	MSE	f_P	f_g	$f_\times$
VideoSAUR							
0	$+0.47$ $[+0.47, +0.48]$	$+0.28$ $[+0.27, +0.29]$	0.81	0.33	0.77	0.32	-0.09
1–2	$+0.49$ $[+0.32, +0.63]$	$+0.48$ $[+0.31, +0.63]$	3.10	3.05	0.45	1.80	-1.25
3–5	$+0.33$ $[+0.18, +0.47]$	$+0.25$ $[+0.09, +0.41]$	3.63	4.33	0.25	1.46	-0.71
6–10	$+0.61$ $[+0.51, +0.69]$	-0.23 $[-0.36, -0.09]$	3.12	8.21	0.13	1.16	-0.29
11–20	$+0.69$ $[+0.63, +0.75]$	-0.24 $[-0.34, -0.14]$	2.85	12.48	0.05	1.04	-0.09
21–40	$+0.72$ $[+0.66, +0.76]$	-0.01 $[-0.15, +0.12]$	2.94	17.28	0.02	1.05	-0.08
41+	$+0.62$ $[+0.48, +0.73]$	-0.16 $[-0.35, +0.07]$	2.54	24.73	0.02	1.04	-0.06
SAVi							
0	$+0.47$ $[+0.46, +0.47]$	$+0.12$ $[+0.11, +0.13]$	0.51	0.74	0.19	0.86	-0.04
1–2	$+0.44$ $[+0.28, +0.57]$	-0.35 $[-0.49, -0.19]$	2.00	4.39	0.06	1.19	-0.25
3–5	$+0.54$ $[+0.43, +0.63]$	-0.39 $[-0.50, -0.27]$	1.48	4.42	0.07	1.15	-0.22
6–10	$+0.46$ $[+0.34, +0.56]$	-0.38 $[-0.48, -0.27]$	1.06	6.00	0.03	1.07	-0.10
11–20	$+0.50$ $[+0.42, +0.58]$	-0.44 $[-0.53, -0.35]$	0.77	7.35	0.02	1.04	-0.06
21–40	$+0.55$ $[+0.46, +0.62]$	-0.46 $[-0.57, -0.35]$	0.72	10.36	0.01	1.02	-0.03
41+	$+0.53$ $[+0.40, +0.65]$	-0.29 $[-0.50, -0.07]$	0.54	15.54	0.01	1.02	-0.03

Table 8: **Two assignment controls.** The takeover-excluded audit reports its retained n and total-error ρ. The last-context audit uses all records in Table 6, assigning each object at its last in-view frame available by the final context frame. Changed counts the records whose assignment differs under that audit. Intervals resample videos.

Gap	Retained n	No takeover ρ [CI]	Changed/total	Last-context ρ [CI]
VideoSAUR				
1–2	91	$+0.71$ $[+0.58, +0.80]$	13/103	$+0.75$ $[+0.65, +0.82]$
3–5	136	$+0.39$ $[+0.21, +0.54]$	28/155	$+0.38$ $[+0.21, +0.53]$
6–10	160	-0.30 $[-0.45, -0.13]$	39/216	-0.19 $[-0.33, -0.04]$
11–20	280	-0.36 $[-0.46, -0.24]$	68/364	-0.30 $[-0.40, -0.19]$
21–40	279	-0.15 $[-0.30, +0.00]$	74/381	-0.08 $[-0.22, +0.07]$
41+	134	-0.26 $[-0.44, -0.06]$	23/185	-0.20 $[-0.39, +0.03]$
SAVi				
1–2	55	-0.46 $[-0.66, -0.19]$	55/120	-0.32 $[-0.48, -0.15]$
3–5	75	-0.64 $[-0.75, -0.49]$	77/169	-0.51 $[-0.61, -0.39]$
6–10	94	-0.57 $[-0.70, -0.42]$	137/247	-0.48 $[-0.57, -0.39]$
11–20	145	-0.53 $[-0.65, -0.39]$	233/410	-0.49 $[-0.56, -0.41]$
21–40	174	-0.50 $[-0.67, -0.29]$	255/449	-0.41 $[-0.52, -0.29]$
41+	89	-0.23 $[-0.55, +0.11]$	116/219	-0.25 $[-0.43, -0.05]$

The loss-sweep and transfer tables evaluate the same records with the same projected trace, eligibility rule, and video split, and both exclude objects that have not yet entered the view.

Across predictor seeds 42, 43, and 44, rollout ρ changes from 0.527 ± 0.037 to 0.746 ± 0.017 on VideoSAUR and from 0.152 ± 0.027 to 0.443 ± 0.031 on SAVi with the combined correction. These repeats use identical records and splits and vary predictor training, separately from the conditional bootstrap intervals.

The log-affine score in Table 2 regresses $\log(\delta + \varepsilon^s)$ on $\log \hat{v}_0$ and an intercept, with the same all-frame calibration folds and ridge settings as the world-error heads. Its slope is unrestricted but remains positive in both folds, at 1.174 and 1.169 for VideoSAUR and 0.361 and 0.348 for SAVi. The in-view comparison in Table 4 uses 108,254 VideoSAUR and 105,478 SAVi records from the same cohort.

Table 9: **The complete fifteen-boundary selection audit.** All scores use the identical out-of-view population and accept the same number m. Actual coverage is m/N. Corrected scores are fitted out of fold by video.

		VideoSAUR ($N = 1267$)			SAVi ($N = 1387$)		
k	$k/15$	m	Original	Corrected	m	Original	Corrected
1	0.07	85	21.13	11.96	93	13.56	9.48
2	0.13	169	20.20	12.07	185	13.69	9.74
3	0.20	254	19.13	12.83	278	12.69	9.44
4	0.27	338	18.71	13.44	370	12.08	9.04
5	0.33	422	18.12	13.58	462	11.82	8.80
6	0.40	507	17.84	13.97	555	11.50	8.69
7	0.47	591	17.38	14.03	647	10.94	8.65
8	0.53	676	16.90	14.27	740	10.80	8.72
9	0.60	760	16.46	14.30	832	10.64	8.79
10	0.67	844	15.97	14.38	924	10.30	8.86
11	0.73	929	15.73	14.43	1017	10.02	8.91
12	0.80	1013	15.23	14.40	1109	9.73	8.97
13	0.87	1098	14.82	14.44	1202	9.53	8.98
14	0.93	1182	14.65	14.53	1294	9.26	8.92
15	1.00	1267	14.61	14.61	1387	8.99	8.99

Table 10: **Dense DINOv2 audit.** Seed 0, 1000 held-out videos. Both correlations use the same attention-weighted score and object-frame samples. Intervals resample videos. Partial ρ removes the rank effect of attention entropy, a proxy for attended object extent, from both variables.

Gap	n	Videos	Latent ρ [CI]	Position ρ [CI]	Partial ρ
0	170,619	1000	$+0.78$ $[+0.77, +0.79]$	-0.07 $[-0.09, -0.06]$	-0.05
1–2	144	139	$+0.56$ $[+0.43, +0.67]$	-0.19 $[-0.34, -0.03]$	-0.09
3–5	216	194	$+0.70$ $[+0.60, +0.77]$	-0.22 $[-0.35, -0.07]$	-0.18
6–10	327	177	$+0.70$ $[+0.62, +0.77]$	-0.19 $[-0.33, -0.05]$	-0.14
11–20	558	166	$+0.78$ $[+0.71, +0.83]$	-0.25 $[-0.41, -0.09]$	-0.21
21–40	695	131	$+0.78$ $[+0.70, +0.84]$	-0.31 $[-0.46, -0.16]$	-0.27
41+	273	58	$+0.76$ $[+0.61, +0.86]$	-0.43 $[-0.61, -0.15]$	-0.42

Table 11: **VideoSAUR loss sweep under one protocol.** Every column uses the exact projected trace and the same 135,445 held-out rollout records. SAVi is evaluated at $\beta = 0.5$ in Table 3.

	$\rho \uparrow$			$R^2_{\log} \uparrow$		
β	Original	Latent	Combined	Original	Latent	Combined
0.0	0.499	0.626	0.714	0.057	0.486	0.555
0.5	0.491	0.627	0.731	0.089	0.459	0.547
1.0	0.490	0.670	0.778	0.099	0.456	0.568

Table 12: **Residual reuse across VideoSAUR predictors.** Each cell lists ρ and then $R^2_{\log}$. The shared sample population and projected score are identical to Table 11, so diagonal entries therefore match its combined column.

Fit β / Apply β	0.0	0.5	1.0
0.0	0.714 / 0.555	0.716 / 0.503	0.744 / 0.474
0.5	0.682 / 0.478	0.731 / 0.547	0.754 / 0.521
1.0	0.659 / 0.398	0.694 / 0.488	0.778 / 0.568

C Encoder collapse and an illustrative information limit

For Eq. 7, $\tau \geq 0$ denotes time without observation, and t remains the frame index elsewhere. Conditional on the predictor's available history and regime τ, assume $s_\tau \sim \mathcal{N}(0, u(\tau)I_2)$ and $z_\tau =$

$a(\tau)s_\tau + \eta_\tau$, where $\eta_\tau \sim \mathcal{N}(0, \nu^2 I_2)$ is conditionally independent and $\nu > 0$ is constant. All expectations below use this conditioning. Take $g(z) = z$, which preserves the position signal's scale at $a = 1$ but still incurs nuisance error $2\nu^2$. The exact latent predictor outputs $\hat{z}_\tau = 0$ and covariance $\{a(\tau)^2 u(\tau) + \nu^2\} I_2$. Thus $\hat{v}_0 = 2(a^2 u + \nu^2)$, while forecast position MSE is $2u$.

Let $u(\tau) = u_0 + \alpha\tau$ and $a(\tau) = e^{-\kappa\tau}$, with $u_0, \alpha, \kappa > 0$. Then

$$\frac{d\hat{v}_0}{d\tau} = 2e^{-2\kappa\tau}\{\alpha - 2\kappa(u_0 + \alpha\tau)\} < 0 \quad \text{for} \quad \tau > \frac{1}{2\kappa} - \frac{u_0}{\alpha}, \qquad \frac{d\mathbb{E}[\varepsilon^s \mid \tau]}{d\tau} = 2\alpha > 0.$$

The components are $e_P = -as_\tau - \eta_\tau$ and $e_g = (a-1)s_\tau + \eta_\tau$, with moments

$$\mathbb{E}\|e_P\|^2 = 2(a^2 u + \nu^2), \qquad \mathbb{E}\|e_g\|^2 = 2\{(1-a)^2 u + \nu^2\}, \qquad \mathbb{E}[2e_P^\top e_g] = 4\{a(1-a)u - \nu^2\}.$$

The cross term is negative when $\nu^2 > a(1-a)u$, and the readout term exceeds the prediction term when $a < 1/2$. These conditions can coexist at long gaps, while $\hat{v}_0$ approaches the nuisance floor $2\nu^2$.

The mutual information in nats is $I_\tau = \log(1 + a^2 u/\nu^2)$, so $\hat{v}_0 = 2\nu^2 e^{I_\tau}$. The Bayes decoder and its risk are

$$m_\tau(z) = \mathbb{E}[s_\tau \mid z_\tau = z] = \frac{au}{a^2 u + \nu^2} z,$$

$$B_\tau = \mathbb{E}\|s_\tau - m_\tau(z_\tau)\|^2 = \frac{2u\nu^2}{a^2 u + \nu^2} = 2ue^{-I_\tau}.$$

When u grows and I_τ decreases, the latent score falls while the Bayes decoding risk rises. This separates information contraction from the choice of readout. For any fixed g, conditional-mean orthogonality gives

$$\mathbb{E}\|g(z_\tau) - s_\tau\|^2 = B_\tau + \mathbb{E}\|g(z_\tau) - m_\tau(z_\tau)\|^2.$$

The first term is irreducible given the true future representation and the stated history, and the second is the readout's excess error under that same information.

Within one gap, two histories can also reverse expected-risk order if $u_2 > u_1$ but $a_2^2 u_2 < a_1^2 u_1$, with the same ν. Heterogeneity in signal loss and physical ambiguity is therefore needed to apply this argument within a gap.

Log-error fitting has a separate scale effect even under an ideal physical error model. Suppose a centered error is conditionally $\mathcal{N}(0, (v/2)I_2)$, with v its true conditional MSE, and let $Y = \log \varepsilon$ without the numerical stabilizer δ. Then $\varepsilon/v \sim \mathrm{Exp}(1)$, giving

$$\mathbb{E}[Y \mid v] = \log v - \gamma, \qquad \mathrm{Var}(Y \mid v) = \frac{\pi^2}{6}, \qquad R^2_{\log,\mathrm{oracle}} = 1 - \frac{\pi^2/6}{\mathrm{Var}(Y)}.$$

Here γ is Euler's constant and the last expression is the population score of the conditional log-mean predictor, not a bound on each finite sample. If the starting score equals v, an unpenalized population-optimal global log offset is $-\gamma$. Exponentiating the fitted log mean yields $e^{-\gamma} v \simeq 0.56v$, and multiplying it by e^γ recovers v only under these assumptions. The geometric-mean interpretation applies to $\hat{v}_R = \hat{v}_0 e^{R_E}$, and adding γ to R_E scales this score by e^γ.

For discretized variables with finite entropies, the chain rule gives

$$\mathcal{H}(s_{t+1} \mid \mathcal{C}_t) = \mathcal{H}(z_{t+1} \mid \mathcal{C}_t) + \mathcal{H}(s_{t+1} \mid z_{t+1}, \mathcal{C}_t) - \mathcal{H}(z_{t+1} \mid s_{t+1}, \mathcal{C}_t). \tag{11}$$

Here $\mathcal{H}$ denotes entropy. The last term includes appearance and other latent variation beyond position. This identity permits predictable features to coexist with unresolved physical state.

Proposition. *If two world distributions induce the same latent-trajectory law but different residual entropies $\mathcal{H}(s_{t+1} \mid z_{t+1}, \mathcal{C}_t)$, this residual is not identifiable from latent trajectories alone.*

Any latent-only estimator has the same distribution under both laws at every sample size. It cannot converge in probability to two different constants, so it cannot consistently estimate both residuals. Its expectations also coincide whenever they exist. For a construction, take two hidden continuations with distinct final positions $s_1 \neq s_2$ but identical observations throughout the interval, hence identical latent trajectories. Let one law choose both equally and the other choose only the first. Their residual entropies are one bit and zero, despite identical latent laws. This uses observation ambiguity, and distinct observations collapsed by an encoder give an analogous construction. □