JEPA-UQ: MEASURING UNCERTAINTY IN LATENT WORLD MODEL ROLLOUTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent world model rollouts need uncertainty estimates tied to future prediction error. JEPA-UQ defines a measurement protocol that specifies prediction-time information, target loss, horizon, output meaning, and statistical unit. We study one frozen visual encoder and two variance-based predictive systems on 14 nuScenes logs outside the project's recorded history. Both systems show a positive score–loss association at H1–H5, with weaker correlations at longer horizons. At H1 and H2, AUROC intervals lie above chance. Overall, score rankings carry information about continuous rollout loss across all five horizons. Interval evidence for thresholded failure events is concentrated at H1 and H2.

1 INTRODUCTION

Latent world models forecast future representations from recent observations, offering compact predictions for planning and analysis in representation space. JEPA-style visual learning has made feature prediction a practical route to reusable video representations (Bardes et al., 2024). A recursive predictor then uses each predicted latent as part of the context for later steps, creating a practical reliability question at prediction time: which rollouts are likely to incur large error when the future arrives?

Average prediction loss summarizes accuracy across scenes. A decision made before a scene unfolds, however, depends on the error that scene will incur, especially when the error is large. We therefore ask whether, before the future is revealed, we can identify rollouts that will later incur large errors. Recursive prediction can alter the relation between an uncertainty score and realized error at each horizon. A probability forecast makes a more specific claim: it assigns a numerical risk to the declared future loss and can be tested against that target.

Latent rollout evaluation makes these choices consequential. Loss magnitude depends on the encoder's coordinate system and its normalization, and each horizon defines a distinct future event. Uncertainty methods may return a scalar disagreement score, a probability, or a trajectory distribution. A useful evaluation keeps those meanings explicit: it declares the prediction-time information set, the observable future target, the loss, and the interpretation of each output.

JEPA-UQ is a measurement protocol for these choices. It treats uncertainty as a prediction-time statement about a specified future rollout loss, with separate claims for ordinal risk ranking, quantitative risk estimation, and joint trajectory distributions. The protocol sets data roles, failure events, calibration rules, metrics, and resampling units for evaluation across horizons. In this first instantiation, one frozen visual encoder and a camera-only latent predictor on nuScenes (Caesar et al., 2020) define the evaluated latent space and rollout. We first describe two stochastic predictive systems on a historically exposed held-out split. A frozen transfer test then evaluates them on 14 logs outside the project's recorded history.

The paper contributes a risk-targeted definition of latent rollout uncertainty, a horizon-resolved Level A evaluation, and a frozen case study on locally unseen logs with explicit evidence boundaries. Levels B and C specify interfaces for numerical risk and trajectory distributions. The nuScenes experiment measures normalized latent error from camera input under a single encoder. Figure 1 summarizes the prediction-time information flow and evaluation design.

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.

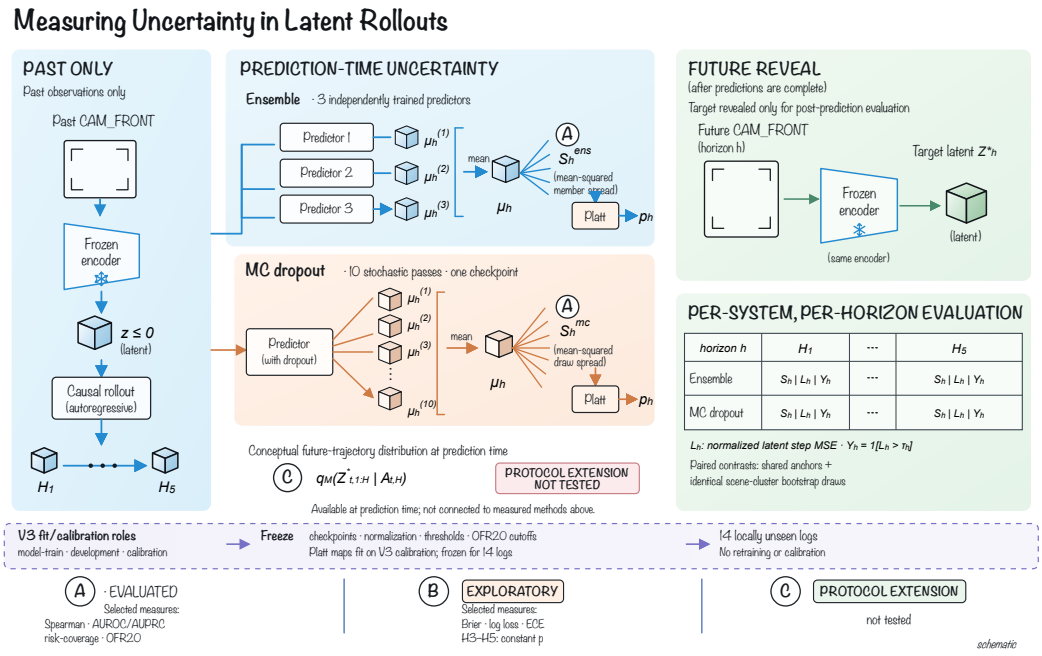

Figure 1: Overview of prediction-time information, the three uncertainty-claim levels, and the frozen transfer evaluation on 14 locally unseen logs. V3 denotes Protocol Version 3.

2 Related Work

Latent prediction and world models. V-JEPA learns video representations through feature prediction (Bardes et al., 2024). That line motivates forecasting in representation space. JEPA-UQ studies a downstream measurement question for a frozen representation and a recursive latent predictor: whether an output issued before the future provides information about the later rollout loss. Evaluation therefore ties the target to the specified encoder, predictor, horizon, and data distribution. Work on action-conditioned world models may enter the protocol through planned controls that are genuinely available when prediction begins.

World-model uncertainty and rollout evaluation. Plan2Explore uses ensemble disagreement over predicted latent states to guide self-supervised exploration (Sekar et al., 2020). PETS propagates uncertainty through sampled trajectories from probabilistic dynamics ensembles (Chua et al., 2018). Calibrated model-based reinforcement learning studies predictive uncertainty as a planning input and calibrates its probability estimates (Malik et al., 2019). Adaptive conformal prediction supplies multistep motion uncertainty sets for planning among dynamic agents (Dixit et al., 2023). V-JEPA 2 adds an action-conditioned latent world model for robotic planning (Assran et al., 2025). JEPA-UQ targets realized loss from a frozen latent rollout and reports its ranking, thresholded failure, and probability measurements by horizon. The primary camera-only case study uses one encoder and two predictive systems. An exploratory appendix analysis evaluates a learned predictive-variance score on the same test population. Broader coverage of model families and control settings remains future work.

Predictive uncertainty and selective prediction. Deep ensembles (Lakshminarayanan et al., 2017) and Monte Carlo dropout (Gal & Ghahramani, 2016) provide established ways to generate variability across predictions. Such variability is an estimator output; its relation to realized error is an empirical question for the declared task. Uncertainty behavior under dataset shift has been studied by Ovadia et al. (2019), showing the value of evaluating reliability as conditions change. Calibration work distinguishes probability values from observed frequencies (Guo et al., 2017). Selective prediction evaluates the loss retained at each coverage level (Geifman & El-Yaniv, 2019). JEPA-UQ applies these ideas to a horizon-indexed latent rollout loss and specifies the prediction-time information available to each estimator.

Benchmark design. WILDS organizes evaluation around naturally occurring distribution shifts and standardized data access (Koh et al., 2021). RobustBench fixes evaluation conditions for adversarial robustness (Croce et al., 2021). OpenOOD v1.5 documents protocol choices across out-of-distribution detection settings (Zhang et al., 2024). These benchmarks make evaluation conditions explicit and state the scope of their evidence. We apply the same evaluation discipline to future losses from causal latent rollouts, and report results by horizon, metric, and system. The first instance covers one encoder and one visual data setting.

3 Problem formulation and uncertainty semantics

Consider a prediction issued at time t for a maximum horizon H. Let $\mathcal{A}_{t,h}$ contain information legitimately available when the prediction for horizon h is issued. It may include observations and states through t, past actions, and a declared future control plan already known at t. A camera-only instance can restrict this set to historical visual latents. To enforce causality, we stipulate that predictions may not use future observations, future target latents, realized losses, or statistics fitted from evaluation futures.

In a given evaluation, the frozen system M contains a target encoder, rollout predictor, and uncertainty estimator. The predictor produces a trajectory $\widehat{Z}_{t,1:H} = M_{\mathrm{pred}}(\mathcal{A}_{t,H})$. At horizon h, a future observation passes through the frozen target encoder to yield z^*_{t+h}. The realized step loss is

$$L_{t,h} = \ell_h(\hat{z}_{t+h}, z^*_{t+h}).$$

After the future occurs, the target becomes observable. For a cumulative view, we define $L^{\mathrm{cum}}_{t,h} = \sum_{k=1}^{h} w_{h,k} L_{t,k}$ with nonnegative weights fixed before evaluation. The present instance sets $w_{h,k} = 1/h$. Step loss and cumulative mean loss correspond to different events, each with its own threshold.

Under the declared evaluation population, the ideal prediction-time target is the conditional distribution of future loss,

$$F^*_{M,h}(e \mid \mathcal{A}_{t,h}) = \Pr(L_{t,h} \le e \mid \mathcal{A}_{t,h}, M).$$

Its functionals include expected loss $r^*_{M,h} = \mathbb{E}[L_{t,h} \mid \mathcal{A}_{t,h}, M]$ and failure probability $p^*_{M,h}(\tau_h) = \Pr(L_{t,h} > \tau_h \mid \mathcal{A}_{t,h}, M)$ for a prespecified threshold τ_h. These quantities serve as conceptual risk targets. One scene provides one realized future and one realized loss. Assessment of a risk estimator consequently uses repeated observations across held-out scenes from the stated population.

For each horizon, an uncertainty output has the form

$$U_{t,h} = G_h(\mathcal{A}_{t,h}; M, \mathcal{D}_{\mathrm{train}}, \mathcal{D}_{\mathrm{cal}}, \xi),$$

where ξ denotes permitted prediction-time randomness. Each output is specified by four elements: admissible information, future loss target, horizon, and semantic claim. This definition distinguishes the measured object from the algorithm used to produce a score.

3.1 Three claim levels

At **Level A**, a scalar $S_{t,h} \in \mathbb{R}$ states that larger values indicate greater risk for the declared loss. Its numerical scale represents an order relation, preserved by any strictly increasing transformation. Empirical tests ask whether high scores precede large losses, identify threshold failures, and select cases with lower retained loss. These questions map to Spearman correlation for loss ordering, AUROC and AUPRC for failure ranking, and selective risk for retained loss.

At **Level B**, an output estimates a numerical risk such as $r^*_{M,h}$ or $p^*_{M,h}(\tau_h)$. Probability predictions receive proper scores and reliability diagnostics on evaluation data after their mapping is fitted on calibration data. A probability-valued output makes a numerical claim whose calibration is established through evaluation.

At **Level C**, a system specifies a joint predictive distribution $q_M(Z^*_{t,1:H} \mid \mathcal{A}_{t,H})$ over the future target-latent trajectory. Joint structure matters for cumulative losses and dependencies between rollout steps. Proper distributional scoring or coverage checks belong to this level. The empirical spread of ensemble members supplies a Level A score in the primary comparison. The appendix also reports an exploratory learned predictive-variance score at Level A. The three levels describe increasingly specific output meanings. This case study tests Level A directly. Level B probabilities provide a transfer diagnostic whose calibration remains unvalidated; the fitted forecasts become constant at H3–H5. Level C remains a protocol interface.

Across these levels, the primary construct is total predictive reliability. Aleatoric and epistemic variation may contribute to future loss. Recursive rollout carries prediction errors and model mismatch through successive steps. JEPA-UQ evaluates whether the final score tracks realized risk, leaving source decomposition to dedicated analyses.

4 JEPA-UQ MEASUREMENT PROTOCOL

4.1 ELIGIBILITY AND INFORMATION FLOW

For evaluation, each system accepts a fixed historical context and issues a causal latent rollout and a horizon-specific uncertainty output before target futures are revealed. The target encoder and loss evaluator are frozen across systems in a comparison. During recursive prediction, later steps consume the model's evolving predicted context. Teacher-forced prediction can be reported as an additional diagnostic. The protocol records the origin of any action or state variable included in the prediction-time information set.

In **Track J**, each uncertainty method belongs to a complete predictive system and may have its own point predictions and realized losses. A paired comparison then concerns those complete systems. A fixed-predictor track would hold point predictions and failure labels shared while comparing uncertainty estimators. The present experiment uses Track J.

For each scene, anchor, horizon, and system, the evaluation file retains one record. Each record identifies the data role, scene and log, elapsed horizon time, prediction loss, uncertainty score, failure label, optional probability, and validity status. Shared anchor identities make paired comparisons possible. Every excluded or invalid output requires an explicit count and reason.

4.2 DATA ROLES AND TARGETS

Four roles govern the first instantiation. Model-train data fit predictor parameters and channel normalization. Development data select training choices and checkpoints. Calibration data determine failure thresholds, low-score cutoffs, and score-to-probability mappings. Evaluation data provide the realized futures for final measurement. Protocol version 3 (V3) separates these roles by log token.

For a target latent with N tokens and D channels, channel statistics μ_d and σ_d come from model-train target latents. The standardized latent is $\tilde{z}_{n,d} = (z_{n,d} - \mu_d)/\max(\sigma_d, \epsilon)$. The V3 implementation fixes $\epsilon = 10^{-6}$. Its primary realized loss is

$$L_{i,h} = \frac{1}{ND}\|\tilde{\hat{z}}_{i,h} - \tilde{z}^*_{i,h}\|_F^2, \qquad L^{\mathrm{cum}}_{i,h} = \frac{1}{h}\sum_{k=1}^{h} L_{i,k}.$$

Both losses are tied to this encoder and normalization. Comparisons across unrelated latent spaces require a target defined and validated for that comparison.

4.3 FAILURE EVENTS AND SCORE CALIBRATION

For each h, a fixed reference predictor supplies calibration errors. Using these errors, we fix the step failure threshold as $\tau_h = Q_{0.90}(\{L^{\mathrm{ref}}_{i,h} : i \in \mathcal{D}_{\mathrm{cal}}\})$. The observed event is $Y_{i,h} = \mathbf{1}[L_{i,h} > \tau_h]$. All compared systems use the same threshold. Another 0.90 quantile of cumulative reference errors

defines cumulative failure. This shared construction gives a stable event definition while allowing system-specific failure prevalence as point predictions change.

For a Level A score, a horizon-specific Platt map provides an exploratory Level B failure probability. Its slope is constrained to be nonnegative, and score standardization uses calibration data. The V3 calibration also sets $\kappa_{m,h} = Q_{0.20}(\{S_{i,h}^{(m)} : i \in \mathcal{D}_{\text{cal}}\})$ for method m. The overconfident-failure rate at this 20th-percentile cutoff (OFR20) is the failure fraction among evaluation cases with $S_{i,h}^{(m)} \leq \kappa_{m,h}$. The observed retained fraction accompanies OFR20 because a frozen cutoff can select a different share of evaluation cases.

4.4 Metrics and statistical units

Each horizon receives the mean, median, and 90th percentile of the point predictor's step loss. Level A results include score–loss Spearman correlation, failure AUROC and AUPRC, and risk–coverage. Event prevalence accompanies AUPRC because it differs across the two predictive systems. The curve records retained mean loss as high-score cases are removed. Its area is normalized as $(\text{AURC} - \text{AURC}_{\text{oracle}})/(\text{AURC}_{\text{random}} - \text{AURC}_{\text{oracle}})$, and risk at 80% coverage gives a fixed point. OFR20 uses the frozen calibration cutoff and is accompanied by the fraction actually retained. Exploratory Level B probabilities receive Brier score, log loss, equal-mass ten-bin ECE, and reliability diagrams. These outputs appear by horizon in the results and appendix.

The positive-count reporting floor of 10 entered during a quality revision after the original freeze. Under this later rule, step-failure AUROC, AUPRC, normalized AURC, and risk at 80% coverage are N/E for a system and horizon with fewer than 10 positive windows. The revision preserved the per-sample predictions, losses, scores, thresholds, and calibration artifacts. Its timing and scope are recorded in Appendix A.

Overlapping anchors from one scene share visual context. The primary V3 bootstrap resamples scenes together with all their windows, using identical draws for paired system contrasts. A three-log bootstrap supplies a sensitivity view whose intervals have limited precision because they rest on only three independent log groups.

Table 1: Output meanings and the measurements available in the current instance.

Output claim	Observable question	Applicable measures	This study
Level A ranking	Does a score order future loss or failures?	Spearman, AUROC, AUPRC, risk–coverage	Ensemble and MC dropout
Level B numerical risk	Does estimated risk match observed outcomes?	Brier, log loss, reliability	Exploratory transfer; H3–H5 constant
Level C trajectory law	Does a joint law represent future target uncertainty?	Proper scores, trajectory coverage	Protocol extension

5 nuScenes visual latent rollout instance

The camera-only instance uses nuScenes `CAM_FRONT` clips, a frozen LeVJEPA-VideoMix-Large encoder, and a causal predictor that rolls forward across five latent steps. Each step represents approximately 2.13 seconds. The ensemble uses the mean prediction from three independently trained predictors and scores their member variance. MC dropout uses ten stochastic passes from one frozen checkpoint and scores the variance across predictions. These are distinct predictive systems with system-specific losses and failure labels; cross-system comparisons therefore evaluate complete systems. Appendix A gives the architecture, checkpoint selection, role manifests, and implementation details.

The study contains two evaluation sets, summarized in Table 2. The 3-log V3 holdout had prior project exposure and provides protocol context. The 14-log test falls outside the project's recorded training, development, calibration, held-out, and earlier experimental logs. The sets share no log tokens. We report their results separately and do not compare their point estimates.

The locally unseen test reuses the V3 checkpoints, normalization, failure thresholds, probability mappings, and OFR20 cutoffs. It contains 304 aligned windows from 157 scenes. No model training or calibration uses these logs. Primary intervals use 1,000 scene-cluster bootstrap draws; 1,000 log-cluster draws provide a sensitivity analysis. Full alignment, provenance, and audit details appear in Appendix A.

6 RESULTS

Table 2 distinguishes the two evaluation populations. The log sets are disjoint.

Table 2: Evaluation populations and their roles in the study.

Evaluation set	Logs	Scenes	Windows	Exposure and role
V3 historical holdout	3	38	76	Prior project exposure; protocol context
Locally unseen logs	14	157	304	Outside recorded project history; primary transfer test

6.1 FIXED LOG-DISJOINT V3 EVALUATION WITH HISTORICAL EXPOSURE

The historically exposed V3 holdout supplies protocol context. Its step-loss and Level A estimates appear in Appendix Table eftab:v3-preliminary. The sample contains 38 scenes, 76 windows, and three logs. We report its estimates separately from the primary locally unseen evaluation.

In auditing this workflow, we were surprised to find that the V3 OFR20 evaluator had read the calibration cutoffs without applying them. Earlier estimates used each test cell's own 20th-percentile cutoff. Before the locally unseen evaluation, we re-evaluated the saved V3 predictions with the frozen calibration cutoffs. This changed OFR20, retained fraction, recorded cutoff, and their intervals; predictions, losses, and other metrics remained fixed. Appendix efapp:ofr-repair records the cell-level changes.

6.2 LOCALLY UNSEEN LOG EVALUATION

The locally unseen test contains 304 valid windows from 157 aligned scenes in 14 logs. Table 3 reports both uncertainty systems at all five horizons. Sample-file hashes and acceptance checks are documented in Appendix A.

Table 3: Locally unseen log case study of the reused V3 protocol; encoder pretraining exposure is unknown. Values in brackets are 95% intervals from 1,000 scene-cluster bootstrap draws. Failures count windows above the frozen step threshold among 304 windows per system and horizon. N/E follows the later positive-count reporting floor of 10. Ens. denotes the three-seed ensemble and MC denotes MC dropout. V3 and new-log point estimates come from different test populations.

Horizon	System	Failures	Spearman [95% CI]	AUROC [95% CI]
H1	Ens.	29	0.6415 [0.5500, 0.7187]	0.7822 [0.6975, 0.8604]
	MC	43	0.5717 [0.4754, 0.6573]	0.7312 [0.6478, 0.8101]
H2	Ens.	28	0.5991 [0.5024, 0.6831]	0.7506 [0.6649, 0.8411]
	MC	40	0.5389 [0.4286, 0.6378]	0.7403 [0.6506, 0.8260]
H3	Ens.	6	0.5048 [0.3939, 0.6040]	N/E
	MC	10	0.4967 [0.3869, 0.6103]	0.4163 [0.2139, 0.6540]
H4	Ens.	7	0.3644 [0.2328, 0.4943]	N/E
	MC	12	0.3713 [0.2321, 0.4983]	0.4472 [0.2514, 0.6253]
H5	Ens.	14	0.2972 [0.1679, 0.4174]	0.5534 [0.3548, 0.7257]
	MC	20	0.2958 [0.1651, 0.4305]	0.4785 [0.3475, 0.6100]

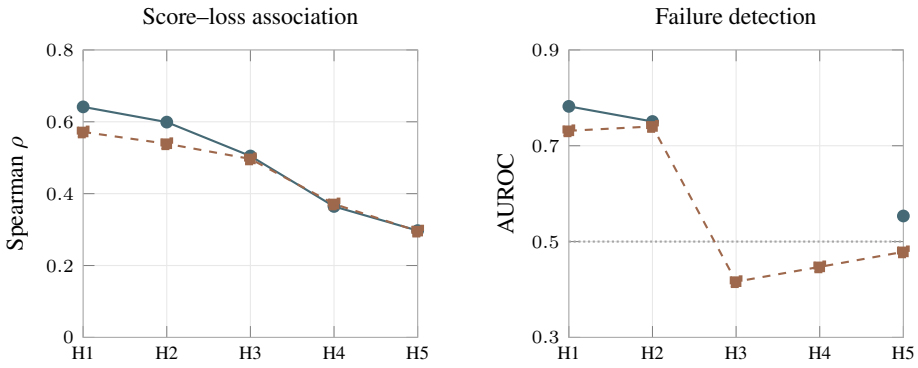

Figure 2: Locally unseen log test of the reused V3 protocol; encoder pretraining exposure is unknown. Blue circles show the ensemble, and brown squares show MC dropout. The dotted line marks chance AUROC at 0.5. Ensemble H3 and H4 have no AUROC markers because their failure counts fall below the reporting floor. Table 3 gives scene-cluster intervals and failure counts.

Table 4: Additional Level A results on the locally unseen logs. AUPRC is shown with its system-specific failure prevalence in parentheses. Normalized AURC and risk at 80% coverage use retained mean step loss. OFR20 gives the failure rate at the frozen calibration cutoff, followed by the selected count out of 304. N/E follows the later positive-count reporting floor of 10. Appendix B.2 gives the scene-cluster intervals.

Horizon	System	AUPRC (prev.)	nAURC	Risk@80	OFR20 (ret.)
H1	Ens.	0.3026 (0.0954)	0.2308	0.4445	0.0000 (18/304)
H1	MC	0.3432 (0.1414)	0.2845	0.4691	0.0417 (24/304)
H2	Ens.	0.2237 (0.0921)	0.3187	0.5720	0.0909 (22/304)
H2	MC	0.3344 (0.1316)	0.3580	0.6173	0.0800 (25/304)
H3	Ens.	N/E (0.0197)	N/E	N/E	0.0400 (25/304)
H3	MC	0.0322 (0.0329)	0.4209	0.6815	0.1364 (22/304)
H4	Ens.	N/E (0.0230)	N/E	N/E	0.1053 (19/304)
H4	MC	0.0636 (0.0395)	0.5288	0.7299	0.1364 (22/304)
H5	Ens.	0.0680 (0.0461)	0.6042	0.7234	0.0455 (22/304)
H5	MC	0.0826 (0.0658)	0.6157	0.7691	0.1429 (21/304)

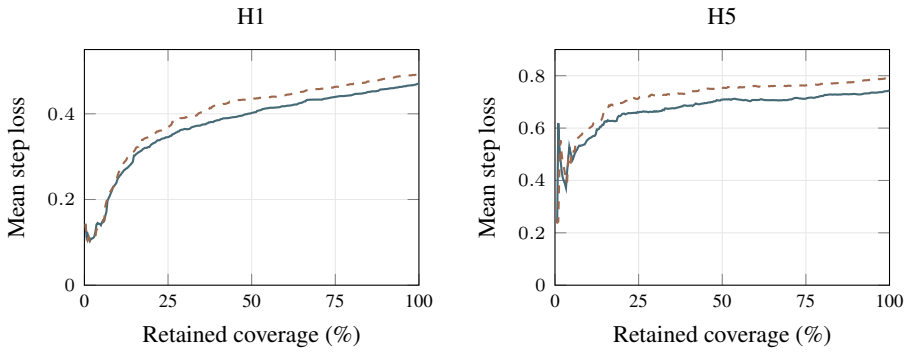

Figure 3: Retained mean step loss as high-score windows are removed on the locally unseen logs. Solid blue denotes the ensemble; dashed brown denotes MC dropout. H1 and H5 have reportable ranking cells for both systems. The two systems have distinct point predictions and loss targets, so each curve describes selectivity within its own system. Appendix B.5 shows the remaining reportable horizons.

For both systems, score–loss Spearman correlation is positive at every horizon, and all intervals exclude zero. This is the first cross-horizon signal we observe. The estimates decrease monotonically, from 0.6415 to 0.2972 for the ensemble and from 0.5717 to 0.2958 for MC dropout. H1 and H2 failure AUROC intervals lie above 0.5 for both systems. At H3–H5, the four reportable AUROC intervals include 0.5; the MC-dropout point estimates at all three horizons are below 0.5. Ensemble H3 and H4 have 6 and 7 failures, so their failure-ranking values are N/E under the floor of 10 positives. MC dropout has 10–20 failures and the ensemble has 14 at H5. These event counts limit the resolution of the later-horizon failure-detection test.

Table 4 and Figure 4 report the remaining Level A outcomes and their scene-cluster intervals. At H1, AUPRC is 0.3026 for the ensemble at 9.5% failure prevalence, and 0.3432 for MC dropout at 14.1%. Their 95% intervals are [0.1880, 0.4909] and [0.2336, 0.4725]. At H5, normalized AURC is 0.6042 [0.4185, 0.8625] for the ensemble and 0.6157 [0.4373, 0.8517] for MC dropout. The system-specific prevalences define different binary tasks, so AUPRC values have separate baselines.

Risk@80 closes 52.0% of the ensemble's random-to-oracle gap at H1, 55.3% at H2, and 24.9% at H5. MC dropout closes 45.2%, 36.9%, and 25.6% at the same horizons. These fractions use each system's own loss vector. Their permutation intervals quantify variation across random rankings. At H5 both fractions are near 25%. The paired Risk@80 contrasts use each predictor's own loss vector, so their raw-unit differences can reflect both selection and overall loss scale. The gap-closed fractions express each system's retained loss relative to its own random-to-oracle range. Normalized AURC integrates that relative performance across coverage; Risk@80 measures raw retained loss at one operating point. A corrected difference at 80% coverage can coexist with an inconclusive area comparison because these summaries use different scales and cover different parts of the risk–coverage relation. The normalized AURC comparisons provide no corrected evidence of a cross-system advantage.

The frozen OFR20 cutoff selects 18–25 windows for the ensemble and 21–25 for MC dropout out of 304. The exact counts accompany the rates in Table 4; the cutoff selects 19 windows in ensemble H4, equal to 6.25% of that cell. Figure 3 shows retained-loss curves at two reportable horizons.

Mean step loss increases from 0.4701 to 0.7435 for the ensemble and from 0.4923 to 0.7896 for MC dropout across H1–H5. Table 7 reports the corresponding medians and 90th percentiles. In the audited family of 35 paired comparisons, Holm retains the five mean-loss and five Risk@80 contrasts. No Spearman, AUROC, AUPRC, normalized AURC, or OFR20 comparison survives. Mean loss compares the complete predictive systems and their point predictions. The primary cross-system analysis uses method-specific failure labels. A same-label companion check finds label-robust differences only for H2 AUROC and H2 Risk@80; H4 is decisive under MC-dropout labels alone. This companion check tests label sensitivity and adds no second corrected family.

The frozen OFR20 cutoffs retain 0.0592–0.0822 of the new windows across the two UQ systems. At H1, the ensemble OFR20 rate is 0.0000 among 0.0592 retained. MC dropout has 0.0417 among 0.0789 retained. These are frozen-cutoff transfer diagnostics at their observed coverage. The Platt mappings yield constant failure probabilities at H3–H5 because their fitted slopes are zero. Their Brier and log loss values describe those constants; calibration resolution is unavailable at these horizons. H1–H2 probabilities vary across windows. Calibration used 86 windows from three logs, with approximately 7–9 positive events per system and horizon. These data leave Level B calibration unvalidated. Appendix B.6 gives every numerical score and the two nondegenerate reliability diagrams. Appendix B.3 describes two exploratory training-free scores, token dispersion and trajectory residual, with their definitions and multiplicity-adjusted comparisons. Appendix B.4 reports an exploratory learned predictive-variance score on this test set. H1 and H2 meet the recorded Spearman criterion; Holm-adjusted H1 comparisons favor ensemble and MC dropout on ranking and nAURC.

7 Discussion and limitations

The locally unseen test yields two findings distinguished by output meaning. Score ranks remain positively associated with realized loss through H5. For the inherited binary failure events, we observe above-chance AUROC intervals at H1 and H2; the four reportable later-horizon intervals include chance. Ensemble H3 and H4 fall below the positive-count floor and are N/E for failure ranking. At later horizons, the frozen thresholds yield 6–20 positive windows per system. Together,

these results locate a horizon-dependent boundary under the inherited events. A larger test can establish whether loss ranking and tail detection separate as horizons grow.

The ensemble has lower point-prediction loss at every horizon. Its mean prediction and failure labels differ from those of MC dropout, so we compare complete systems under Track J. Paired AUROC and normalized risk–coverage intervals span zero at H1, H2, and H5; the H3 and H4 comparisons are N/E. A fixed-predictor study could isolate estimator differences while holding point predictions constant. The inherited Platt fit produces constant probabilities at H3–H5. Calibration used 86 windows from three logs, with approximately 7–9 positive events per system and horizon. The available data do not separate sample-size effects from the nonnegative slope constraint. Level B remains unvalidated. OFR20 retains roughly 6–8% of new windows for the UQ systems. At this coverage, its transferred cutoff describes failures in a narrow subset of the test population.

The empirical target is normalized latent error under one frozen encoder with `CAM_FRONT` input. Physical-state forecasting, downstream decisions, and driving safety require their own targets and validation. All 14 new logs are disjoint from this project's recorded history. The encoder pretraining corpus is unavailable for audit, leaving potential nuScenes exposure unresolved at the representation level. An encoder with documented pretraining exclusion or a newly collected video population would test this boundary. V3's three historically exposed evaluation logs supply protocol-development context. The two test sets differ in log count, window count, and location composition, which limits any direct comparison of their point estimates. Scene-cluster intervals account for overlapping windows within scenes. The 14-log bootstrap remains a sensitivity analysis, and the V3 calibration fit rests on only 86 windows with few failures.

In the local data audit, we found no unused camera logs containing both pose and steering CAN data. Fresh confirmation of the training-free trajectory-residual score requires a new collection or dataset. The learned predictive-variance analysis adds a supervised score on the existing test population. Wider protocol claims require more encoders and uncertainty families.

8 Conclusion

JEPA-UQ ties each uncertainty claim to a future latent loss and the information available when its prediction is made. In the frozen camera-based case study, scores track realized loss across five rollout steps on 14 locally unseen logs. Failure-detection intervals lie above chance at H1 and H2; the four reportable later-horizon intervals include chance. Ensemble H3–H4 cells fall below the later reporting floor. Taken together, scores remain associated with continuous error, while detection of rare threshold events remains unresolved. Level B transfer yields constant probabilities at H3–H5, and Level C awaits empirical study. The learned score also shows a positive association with loss at H1 and H2. Holm-adjusted H1 comparisons favor the frozen stochastic scores. Tests across additional encoders and uncertainty families will establish the protocol's scope.

AI Use Statement

Generative AI tools assisted with manuscript editing. The authors remain responsible for data provenance, computations, and all claims in the submission.

References

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, et al. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025. URL https://arxiv.org/abs/2506.09985.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=QaCCuDfBk2.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pp. 11621–11631, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Caesar_nuScenes_A_Multimodal_Dataset_for_Autonomous_Driving_CVPR_2020_paper.html.

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

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/a3c65c2974270fd093ee8a9bf8ae7d0b-Abstract-round2.html.

Anushri Dixit, Lars Lindemann, Skylar X. Wei, Matthew Cleaveland, George J. Pappas, and Joel W. Burdick. Adaptive conformal prediction for motion planning among dynamic agents. In *Proceedings of the 5th Annual Learning for Dynamics and Control Conference*, volume 211 of *Proceedings of Machine Learning Research*, pp. 300–314. PMLR, 2023. URL https://proceedings.mlr.press/v211/dixit23a.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.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M. Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 5637–5664. PMLR, 2021. URL https://proceedings.mlr.press/v139/koh21a.html.

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, 2017. URL https://papers.nips.cc/paper/2017/hash/9ef2ed4b7fd2c810847ffa5fa85bce38-Abstract.html.

Ali Malik, Volodymyr Kuleshov, Jiaming Song, Danny Nemer, Harlan Seymour, and Stefano Ermon. Calibrated model-based deep reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4314–4323. PMLR, 2019. URL https://proceedings.mlr.press/v97/malik19a.html.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/8558cb408c1d76621371888657d2eb1d-Abstract.html.

Ramanan Sekar, Oleh Rybkin, Kostas Daniilidis, Pieter Abbeel, Danijar Hafner, and Deepak Pathak. Planning to explore via self-supervised world models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8583–8592. PMLR, 2020. URL https://proceedings.mlr.press/v119/sekar20a.html.

Jingyang Zhang, Jingkang Yang, Pengyun Wang, Haoqi Wang, Yueqian Lin, Haoran Zhang, Yiyou Sun, Xuefeng Du, Yixuan Li, Ziwei Liu, Yiran Chen, and Hai Li. OpenOOD v1.5: Enhanced benchmark for out-of-distribution detection. *Journal of Data-centric Machine Learning Research*, 2024. URL https://openreview.net/forum?id=cnnTnJQigs.

A Experimental details and audit record

A.1 Encoder, predictor, and role manifests

The camera-only instance uses nuScenes v1.0-trainval CAM_FRONT clips and a frozen LeVJEPA-VideoMix-Large encoder. After removing the CLS token and pooling 16 temporal slots, the encoder yields 196 spatial tokens with 1024 channels. Each input comprises three historical clips of 16 frames at a nominal 7.5 frames per second. The causal transformer predictor rolls forward five latent steps, each representing approximately 2.13 seconds. H1–H5 have nominal durations of 2.13, 4.27, 6.40, 8.53, and 10.67 seconds. Camera timestamps in the saved sample file provide actual durations. In the historical holdout, H1 spans 2.05–2.20 seconds and H5 spans approximately 10.65–10.70 seconds.

Table 5: V3 role manifests. Log tokens are disjoint across roles.

Role	Logs	Scenes	Windows
Model training	9	150	300
Development	2	24	48
Calibration	3	43	86
Historical holdout	3	38	76

The ensemble consists of three independently trained predictor runs with integer seeds 20260930, 20260931, and 20260932. Stage A checkpoint selection uses development H1 loss. Stage B trains with causal feedback across H1–H5 and selects a checkpoint by development mean H1–H5 loss. The ensemble predicts the member mean and scores mean squared member deviation. MC dropout uses ten stochastic inference passes from the seed-20260930 best Stage B checkpoint; its prediction is the draw mean and its score is mean squared draw deviation. A random score and an oracle error score serve as diagnostics. The oracle uses realized future error only for a measurement check.

Calibration uses 86 windows from three logs. Each system has approximately 7–9 positive step failures per horizon. The nonnegative Platt slopes reach zero at H3–H5, producing constant probability forecasts. The calibration sample and fitted behavior limit Level B interpretation. The held-out per-sample file contains 1,900 rows: 76 windows across five horizons for each of the two UQ systems, a reference predictor, and two diagnostics.

Table 6: Preliminary V3 step-loss and Level A results. Intervals are 95% scene-cluster bootstrap intervals over 38 scenes and 76 windows from three historically exposed logs. These estimates are separate from the locally unseen-log test.

System	Horizon	Mean step loss	Spearman [95% CI]	AUROC [95% CI]
Ensemble	H1	0.4593	0.8383 [0.6701, 0.9292]	0.9128 [0.8225, 0.9835]
Ensemble	H3	0.7104	0.6912 [0.4936, 0.8383]	0.7169 [0.4781, 0.9190]
Ensemble	H5	0.8113	0.5411 [0.2286, 0.7493]	0.6979 [0.4191, 0.9119]
MC dropout	H1	0.4776	0.7866 [0.5926, 0.8971]	0.8606 [0.7557, 0.9485]
MC dropout	H3	0.7487	0.6999 [0.4793, 0.8560]	0.7227 [0.4732, 0.9452]
MC dropout	H5	0.8720	0.5732 [0.2688, 0.7735]	0.7406 [0.5109, 0.9132]

A.2 Locally unseen-log alignment and result provenance

The locally unseen evaluation contains 14 nuScenes logs outside the project's recorded training, development, calibration, historical holdout, and earlier experiment logs. Complete `CAM_FRONT` data cover 158 scenes. Frozen alignment rules yield 157 aligned scenes and 304 windows. Ten scenes provide one window each; one scene has no eligible window because its 59.810 ms steering lag exceeds the inherited 50 ms CAN alignment limit. The predictor receives camera latents. CAN filters anchor eligibility and supplies no prediction-time control input. The frozen encoder processes 1,403 unique clips.

The new evaluation reuses the V3 predictor checkpoints, model-training normalization, step and cumulative failure thresholds, Platt mappings, and OFR20 cutoffs. No model training or calibration uses the new logs. The per-sample file contains 7,600 rows across 304 windows, five horizons, and five systems, including the reference predictor. Primary intervals use 1,000 scene-cluster bootstrap draws; 1,000 log-cluster draws provide a sensitivity analysis. Split, checkpoint, normalization, calibration, code, and configuration hashes are recorded in the audit manifest.

A.3 Revision timeline and integrity checks

The analysis distinguishes the original frozen evaluation artifacts from later quality revisions. A positive-count reporting floor of 10 was introduced after the original freeze. It changes which sparse failure-ranking cells are reported; it does not alter predictions, scores, losses, thresholds, or calibration mappings.

A separate audit found that the original V3 OFR20 evaluator read the calibration cutoffs but did not apply them. Its earlier output used test-set 20th-percentile cutoffs. The corrected V3 estimates apply the frozen calibration cutoffs to the saved per-sample results. OFR20, retained fraction, recorded cutoff, and their intervals changed; predictions, losses, and other metrics remained fixed. This repair preceded the locally unseen-log evaluation. Table 14 records the affected V3 cells.

The latest quality revision passed 21 pre-test and 12 post-test acceptance checks. These checks cover implementation and artifact invariants. Their timing does not make the reporting floor a pre-freeze rule. The revision manifests and check outputs provide the full audit trail.

B Additional measured outcomes

All tables in this appendix use the frozen locally unseen log result package. Its per-sample CSV digest is recorded in the companion evidence ledger. Each system and horizon has 304 valid windows. The numerical values are descriptive point estimates under the declared frozen protocol.

B.1 Step-loss distribution

Table 7: Normalized latent step-loss summaries on the locally unseen logs. Mean, median, and 90th percentile refer to the point predictor of each complete system.

Horizon	System	Mean	Median	90th percentile
H1	Ens.	0.4701	0.4649	0.6306
H1	MC	0.4923	0.4852	0.6615
H2	Ens.	0.6132	0.5907	0.8726
H2	MC	0.6467	0.6175	0.9305
H3	Ens.	0.6694	0.6352	0.9410
H3	MC	0.7087	0.6786	0.9804
H4	Ens.	0.7119	0.6852	0.9726
H4	MC	0.7538	0.7338	0.9991
H5	Ens.	0.7435	0.7138	1.0308
H5	MC	0.7896	0.7606	1.0633

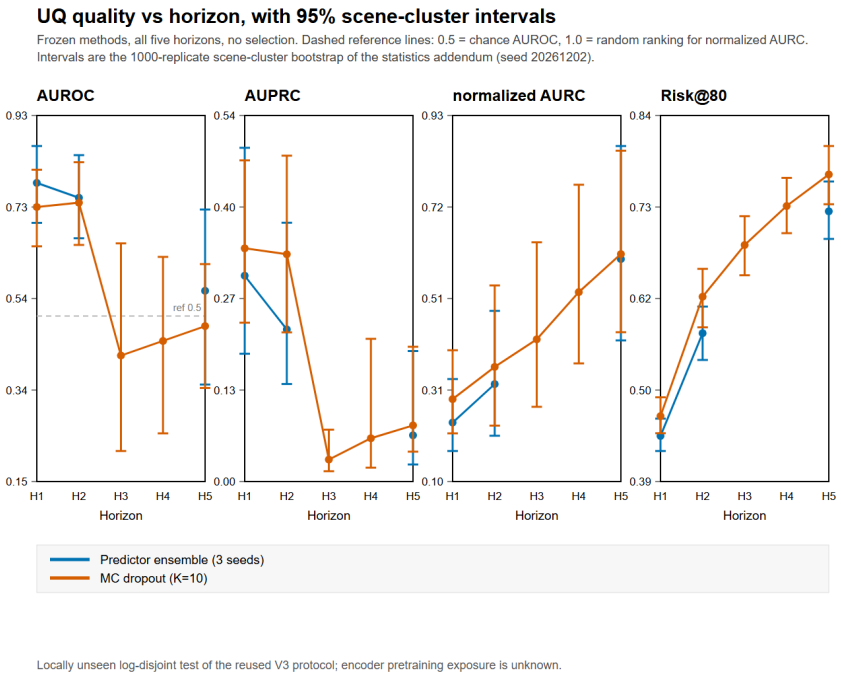

Figure 4: Scene-cluster 95% intervals for four Level A measures on the locally unseen logs. A gap marks a cell withdrawn under the positive-count floor. Risk@80 intervals describe scene-resampled evaluation estimates.

B.2 Scene-cluster intervals for additional Level A measures

Figure 4 displays the 95% scene-cluster intervals for AUROC, AUPRC, normalized AURC, and Risk@80 across the five horizons. Cells below the positive-count floor are left unplotted, including their otherwise computable intervals. AUPRC baselines vary with each system's failure prevalence, reported in Table 4.

B.3 Exploratory training-free scores

This analysis followed inspection of the original test results. The two formulas and score directions were recorded before either new score column was joined to the labels. These additions remain exploratory and leave the frozen ensemble and MC-dropout results intact. The new scores share the reference predictor's failure labels, which differ from the method-specific labels used for the frozen systems. Table 8 and Table 3 therefore do not define a common binary task for cross-table AUROC comparisons. Both scores use the reference predictor's frozen rollout from seed 20260930 and the model-train standardized latent space. They have no point predictor of their own and inherit the reference predictor's realized error and failure labels.

For predicted latent token vectors $z_h[t, :]$, token energy is $e_{h,t} = D^{-1} \sum_{c=1}^{D} z_h[t, c]^2$. The dispersion score is $S_h^{\text{disp}} = \text{sd}_{t=1,\ldots,N} \log(e_{h,t})$, with higher values declared to indicate greater risk. It measures across-token spread in log energy. The trajectory-residual score compares each prediction with a linear extrapolation from the previous two latent states:

$$z_h^{\text{lin}} = z_{h-1} + (z_{h-1} - z_{h-2}), \qquad S_h^{\text{traj}} = \frac{1}{ND} \| z_h - z_h^{\text{lin}} \|_F^2.$$

Here z_{-1} and z_0 are the last two observed history latents, and later terms use the rollout's own predictions. Higher values indicate greater departure from this linear continuation. Neither score reads future targets.

Table 8: Ranking estimates for the two exploratory scores. Each horizon has 304 windows. The scores inherit the reference predictor's failure labels; prevalence gives the AUPRC baseline. These labels differ from the method-specific labels in Table 3, so cross-table AUROC values do not compare a common binary task.

Method	Horizon	n_+ (prev.)	Spearman	AUROC [95% CI]	AUPRC
Dispersion	H1	43 (14.1%)	-0.0125	0.5387 [0.4324, 0.6499]	0.1695
Dispersion	H2	40 (13.2%)	0.1468	0.5531 [0.4422, 0.6567]	0.1845
Dispersion	H3	11 (3.6%)	0.1515	0.4514 [0.2269, 0.6865]	0.0426
Dispersion	H4	13 (4.3%)	0.1375	0.4481 [0.2459, 0.6682]	0.0504
Dispersion	H5	21 (6.9%)	0.1208	0.4430 [0.2847, 0.5856]	0.0682
Trajectory residual	H1	43 (14.1%)	0.5430	0.7056 [0.6101, 0.7821]	0.2920
Trajectory residual	H2	40 (13.2%)	0.5427	0.7359 [0.6506, 0.8258]	0.2780
Trajectory residual	H3	11 (3.6%)	0.5197	0.5324 [0.3630, 0.7161]	0.0555
Trajectory residual	H4	13 (4.3%)	0.4402	0.4655 [0.3106, 0.6339]	0.0435
Trajectory residual	H5	21 (6.9%)	0.3406	0.5038 [0.3720, 0.6461]	0.0865

Table 9: Selectivity estimates for the exploratory scores under the inherited reference-predictor loss. Lower values indicate better retained-loss performance.

Method	Horizon	Normalized AURC	Risk@80
Dispersion	H1	1.0283	0.4927
Dispersion	H2	0.9609	0.6353
Dispersion	H3	0.9776	0.6933
Dispersion	H4	0.9954	0.7442
Dispersion	H5	0.9810	0.7866
Trajectory residual	H1	0.3090	0.4712
Trajectory residual	H2	0.3676	0.6203
Trajectory residual	H3	0.3994	0.6772
Trajectory residual	H4	0.4515	0.7327
Trajectory residual	H5	0.5722	0.7716

The paired analysis compares each exploratory score with each frozen system under the corresponding method-specific labels. Holm correction is applied in two separate families: 60 ranking comparisons across Spearman, AUROC, and AUPRC, and 40 loss-unit comparisons across normalized AURC and Risk@80. No ranking comparison survives. Seventeen loss-unit comparisons survive, all favoring a frozen method. The compared scores share one reference-predictor loss vector, so these selective comparisons use a common loss target; the frozen methods' mean step loss is excluded because it repeats that shared vector. The significant loss-unit results do not establish a ranking advantage. OFR20 is also unavailable because neither score has a frozen calibration cutoff in its score units.

A companion check evaluates both scores under each frozen system's label set. Ten latent-dispersion cells are label-robust under the reported interval rule; no trajectory-residual cell is label-robust. The check finds no direction conflicts between the two label sets and is reported without a second multiplicity correction.

B.4 Exploratory learned predictive-variance score

The analysis evaluates a learned predictive-variance score for the frozen reference predictor. Five horizon-specific heteroscedastic heads were fit with three seeds per horizon to predict the reference system's normalized-latent step MSE. Training used 300 model-train windows from nine logs; 48 development windows from two logs selected checkpoints. The predictor remained frozen, and calibration and test outcomes were excluded from fitting and selection. The primary score is the arithmetic mean of the three predicted log-variance values. It is an ordinal risk estimate and is not interpreted as a calibrated probability or an aleatoric variance.

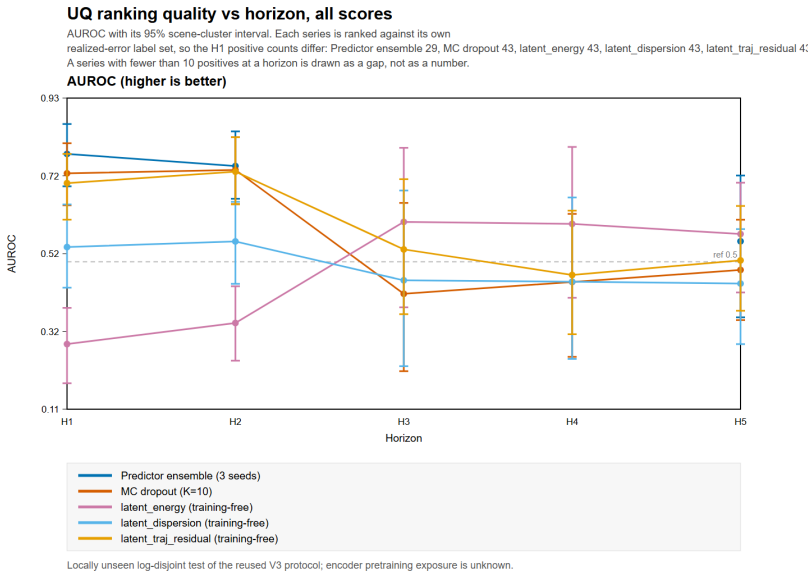

Figure 5: AUROC point estimates and scene-cluster 95% intervals for frozen and training-free scores. Every series uses its own realized-error label set, so cross-series positions do not constitute comparisons on a common binary task. Ensemble H3 and H4 values are omitted under the positive-count floor.

This analysis followed lower-clamp saturation in an earlier head run. The stabilization settings and H1–H2 decision rule were recorded before fitting. The 304-window locally unseen-log test set had already been inspected in the preceding study. These estimates therefore form an exploratory secondary analysis on the existing test population. The primary criterion required both H1 and H2 scene-cluster 95% Spearman intervals to have lower bounds above zero.

Table 10: Exploratory learned predictive-variance score against frozen reference step loss on the previously examined locally unseen-log test set. Each horizon has 304 windows. Positive counts define the AUPRC baseline. Scene-cluster intervals are primary; log-cluster intervals are a sensitivity analysis over 14 logs.

Horizon	$n_+/304$	Spearman ρ	Scene 95% CI	Log 95% CI
H1	43/304	0.3149	[0.1671, 0.4362]	[0.1722, 0.4733]
H2	40/304	0.4483	[0.3206, 0.5635]	[0.3546, 0.5380]
H3	11/304	0.3985	[0.2563, 0.5350]	[0.2940, 0.5145]
H4	13/304	0.4023	[0.2614, 0.5295]	[0.2658, 0.5175]
H5	21/304	0.3898	[0.2513, 0.5109]	[0.1967, 0.4905]

Both primary H1–H2 intervals meet the recorded criterion. The positive H3–H5 point estimates are secondary. Table 11 reports the remaining point metrics. Every reference-label cell meets the later reporting floor of 10 positives.

Table 11: Secondary point metrics for the learned predictive-variance score on the same 304-window test set. AUROC and AUPRC use frozen-reference failure labels; the AUPRC baseline is $n_+/304$ in Table 10. Lower nAURC and Risk@80 indicate better selectivity under the shared frozen-reference loss.

Horizon	AUROC	AUPRC	nAURC	Risk@80
H1	0.6847	0.2718	0.4653	0.4708
H2	0.7071	0.2643	0.4119	0.6170
H3	0.5110	0.1051	0.4870	0.6728
H4	0.4740	0.0440	0.4792	0.7225
H5	0.5620	0.1762	0.6366	0.7719

Paired comparisons used scene-cluster bootstrap intervals and separate Holm families for ranking (30 comparisons) and loss-unit measures (20 comparisons). The four surviving comparisons all favor a frozen score system at H1: ensemble and MC dropout have higher Spearman correlation and lower nAURC than the learned score. Table 12 gives the surviving estimates. The supported claim is limited to positive loss association for the learned score on this test population.

Table 12: Holm-surviving paired comparisons for the learned score at H1. Deltas are learned score minus comparator; negative Spearman and positive nAURC deltas favor the frozen comparator.

Measure	Comparator	Delta [scene 95% CI]	Holm p
Spearman	Ensemble	-0.3447 $[-0.4893, -0.2049]$	0.030
Spearman	MC dropout	-0.2608 $[-0.4010, -0.1258]$	0.030
nAURC	Ensemble	$+0.2493$ $[+0.1356, +0.3943]$	0.020
nAURC	MC dropout	$+0.1831$ $[+0.0795, +0.3062]$	0.038

The local role audit found no previously unused camera logs with pose and steering CAN data. This analysis adds one learned score family on the previously examined test set. A new-log confirmation of the training-free trajectory-residual score remains for future data.

B.5 Remaining risk–coverage curves

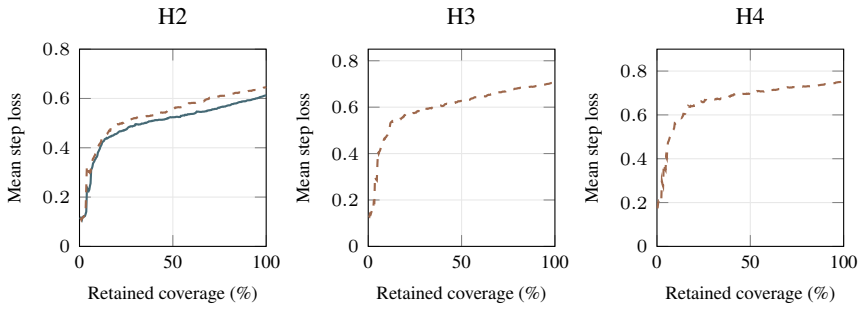

Figure 6: Retained mean step loss at H2–H4. Solid blue denotes the ensemble and dashed brown denotes MC dropout. The ensemble H3 and H4 curves are omitted under the later reporting floor for their cells. The risk–coverage values plotted here are those retained by the revised result package.

B.6 EXPLORATORY NUMERICAL RISK

Table 13: Exploratory probability scores on the locally unseen logs. Log loss is Bernoulli negative log-likelihood; ECE uses ten equal-mass bins. H3–H5 mappings issue one constant probability per system and horizon, so their numbers describe that constant output.

Horizon	System	Brier	Log loss	ECE	Mapping
H1	Ens.	0.0798	0.2813	0.0529	varying
H1	MC	0.1140	0.3763	0.0572	varying
H2	Ens.	0.0802	0.2899	0.0530	varying
H2	MC	0.1084	0.3666	0.0702	varying
H3	Ens.	0.0231	0.1327	0.0617	constant
H3	MC	0.0370	0.1812	0.0718	constant
H4	Ens.	0.0259	0.1407	0.0584	constant
H4	MC	0.0408	0.1875	0.0535	constant
H5	Ens.	0.0452	0.1965	0.0456	constant
H5	MC	0.0630	0.2518	0.0561	constant

At H3–H5 the frozen nonnegative Platt slope is zero. Each ensemble cell issues probability 0.081395. MC dropout issues 0.104651 at H3 and H5 and 0.093023 at H4. Calibration resolution is therefore unavailable at these horizons. The H1 and H2 mappings produce 304 distinct probabilities in each system and horizon. Their reliability points are shown below; the small three-log calibration set limits interpretation of any apparent fit.

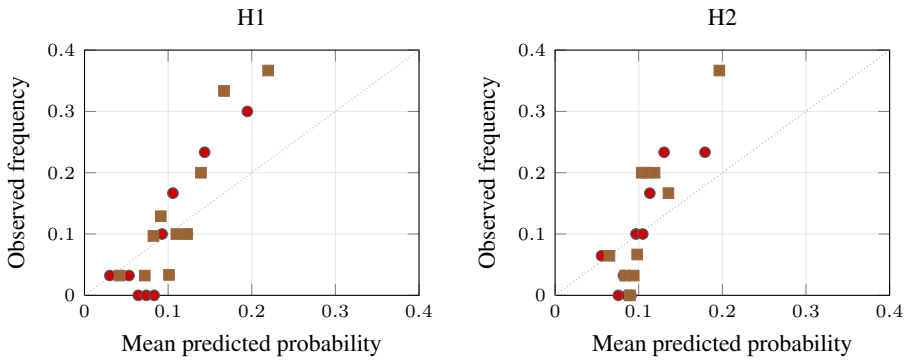

Figure 7: Ten equal-mass reliability bins for the varying H1 and H2 Platt mappings. Blue circles denote the ensemble and brown squares denote MC dropout; the dotted diagonal marks matching mean probability and observed frequency. H3–H5 have constant mappings and are listed in Table 13.

B.7 V3 FROZEN-CUTOFF REPAIR

The original V3 evaluator read the calibration OFR20 cutoffs and then used a fresh 20th percentile of the test scores within each cell. The repaired evaluator applies the calibration cutoff to the same V3 per-sample rows and to each bootstrap draw. Its repair affects OFR20, retained fraction, recorded cutoff, and their intervals. Spearman, AUROC, AUPRC, normalized AURC, probability scores, losses, failure thresholds, and the per-sample CSV retain their prior values.

Table 14: V3 OFR20 before and after frozen-cutoff repair. Parentheses contain the retained fraction. Each cell has 76 evaluation windows; the old evaluator retained 16 windows in every cell. These V3 outcomes are distinct from the locally unseen log test.

Horizon	System	Earlier OFR20 (ret.)	Repaired OFR20 (ret.)
H1	Ens.	0.0000 (0.2105)	0.0000 (0.2105)
	MC	0.0000 (0.2105)	0.0000 (0.2237)
H2	Ens.	0.0000 (0.2105)	0.0000 (0.1974)
	MC	0.0625 (0.2105)	0.0588 (0.2237)
H3	Ens.	0.0625 (0.2105)	0.0625 (0.2105)
	MC	0.1250 (0.2105)	0.1176 (0.2237)
H4	Ens.	0.0625 (0.2105)	0.0769 (0.1711)
	MC	0.1250 (0.2105)	0.1176 (0.2237)
H5	Ens.	0.1250 (0.2105)	0.1538 (0.1711)
	MC	0.1875 (0.2105)	0.2000 (0.1974)