Overconfidence and the Ordinality Trap in Inverse Dynamics Models for Physical AI

Rajit Rajpal
School of Mathematics
University of Edinburgh
s2592586@ed.ac.uk

Matteo Peterlongo Acquaviva
School of Informatics
University of Edinburgh

Abstract

Inverse dynamics models (IDMs) infer the action responsible for an observed transition, and are now load-bearing infrastructure in physical AI: they generate supervision for VLA policies, translate predicted futures into motor commands, and judge whether generated video is physically executable. The latter two are threshold decisions, yet IDM confidence is rarely evaluated. We audit three IDMs: OpenAI's released VPT IDM, a UniPi-style IDM we train on LIBERO, and VERA's released transformer trunk on identical data. Taking as given that calibration and discrimination are distinct quantities, we report three findings that do not follow from that. First, because modern IDMs tokenise continuous actions, practitioners score them with categorical ECE; on ordinal action bins the calibrator that exact-bin ECE selects moves probability mass *away* from the true action, improving exact-bin ECE by up to $15.8\times$ while degrading tolerant ECE on every ordinal head we tested. Bin distance is action error. Second, confidence quality depends on output-head geometry rather than the underlying model: binary heads calibrate well and rank their own errors well, while ordinal heads sharing the same forward pass do neither. Third, across two kinds of distribution shift, temperature scaling is the sole correction that never degrades ranking, and per-dimension recalibration justifies its complexity only on ordinal heads evaluated in-distribution. Consequently, confidence must be reported per dimension and calibrators selected by the downstream decision rather than by ECE. This result was produced by an autonomous agent, and its measurement-integrity work was decisive: twenty claims were withdrawn or corrected during the study, four of them scoring artefacts that each inverted a conclusion.

1 Why audit an IDM

An IDM answers $(o_t, o_{t+1}) \mapsto a_t$: what action explains this transition? The model is small and the question narrow, which is why it is treated as plumbing. Three deployed roles make that untenable. IDMs *manufacture supervision*; VPT pseudo-labelled roughly 70k hours of video with one [Baker et al., 2022]; they *execute imagination*, inverting a generated visual plan into motor commands [Du et al., 2023], and they now *judge generators*, supplying a reward signal for physically executable rollouts [Wang et al., 2026]. In all three the IDM is trusted either to be right or to say when it is not, and the latter two are threshold decisions. Nothing in a cross-entropy or Huber objective obliges it to widen its predictive distribution off-distribution.

We take as given that calibration and discrimination are distinct quantities and that monotone recalibration cannot change a ranking. The three findings below do not follow from that: an ordinality trap specific to tokenised action spaces (§3), a dependence of confidence quality on head geometry rather than on the model (§4), and one repair rule that survives two kinds of distribution shift (§5).

Preprint.

2 Targets, data and metrics

We audit three single-step IDMs; chunking is out of scope. **VPT** [Baker et al., 2022] is the only released IDM checkpoint: a non-causal transformer over 128 frames with 20 binary button heads and two 11-way *ordinal* heads over mu-law camera deltas, and a temperature of 4.0 baked into the released weights, so an "uncalibrated baseline" that ignores that field describes an already-tempered model. **LIBERO IDM** is ours, since no released direct-regression continuous IDM exists: the canonical recipe on `libero_spatial` [Liu et al., 2023] in two output geometries, 6-dim regression and the same features discretised into 21 ordinal bins, with a Bernoulli gripper head in both. **VERA** [Li et al., 2026b] contributes its released ViT trunk on identical data and splits, with our heads substituted because as published it regresses a ± 1 gripper under a squared loss and has nothing calibratable. Splits are contractor-disjoint for VPT and by-demo for LIBERO, with `libero_{goal,object,10}` as a task-shift ladder (Appendix C).

Because no-ops dominate (93.9% of button entries), every figure is restricted to non-null actions. Three further controls matter and are detailed in the appendix: action–video alignment was verified by an offset scan; window position within a recording matters, with confidence *flat* while accuracy climbs $0.6218 \rightarrow 0.9401$ over the first $\sim$24 windows of a 9.x recording (Appendix F); and three of six perturbation families move accuracy by under 0.01 and are reported as inert probes rather than assigned a decoupling ratio (Appendices G and H). We report ECE_1/ECE_2 over 15 bins, Brier and NLL for calibration; error-detection AUROC and selective risk for discrimination; and for ordinal heads the Ranked Probability Score $\mathrm{RPS} = \frac{1}{n}\sum_i \sum_k (F_{\hat{p}}(k) - F_y(k))^2$, expected bin error $\mathbb{E}_{k\sim\hat{p}}|k-y|$ in units of bins, and a tolerance-1 ECE counting a prediction correct within one bin (Appendix E; extended related work in Appendix D).

3 The ordinality trap

Once a model emits categorical tokens under cross-entropy, the default evaluation suite follows, and every member of it treats the bins as unordered. For an action space that is false in a costly way: an adjacent-bin prediction is a near-correct movement and a distant one is a collision, and exact-bin ECE scores them identically. Scored as unordered classes VPT's camera head reads as severely miscalibrated (ECE 0.204); at one-bin tolerance the same predictions read 0.028, with within-one-bin accuracy 0.989. The consequence is not conservative bookkeeping but active misdirection. Table 1 evaluates three ordinal heads across three models under exact-bin ECE and three ordinality-aware scores. **The columns disagree in the same direction every time.** Temperature scaling, the remedy exact-bin ECE selects, improves it by $2.3\times$ to $15.8\times$ while *degrading* tolerant ECE on all three heads and increasing expected bin error by 0.41 and 0.54 bins where defined. It buys a well-calibrated top-class probability by flattening the distribution until the mass no longer sits beside the true bin. On all three heads the *uncalibrated* model is best by tolerant ECE. This is not the textbook point that ECE and ranking differ. Temperature is monotone in confidence and so provably cannot change an ordering; we confirm that in §4. What it changes is *where the mass sits*, which RPS, tolerant ECE and expected bin error measure and neither exact-bin ECE nor AUROC can see. One complication we report rather than suppress: the ordinal scores disagree with each other, since temperature *improves* RPS on both LIBERO heads ($0.8937 \rightarrow 0.7779$; $1.2137 \rightarrow 0.9257$) while leaving it flat on VPT's camera. So the claim is narrower and firmer than "use tolerant ECE": **exact-bin ECE alone cannot decide whether a calibrator helps an ordinal head**.

4 Confidence quality tracks head geometry

The two heads of a single VPT forward pass behave completely differently in distribution: the button head is 0.9249 accurate with ECE 0.0470 and AUROC 0.8289, the camera head 0.5551 accurate with ECE 0.3960 and AUROC 0.5598. LIBERO agrees; its Bernoulli gripper reaches ECE 0.0040 at AUROC 0.9811 and stays usable across the task-shift ladder (0.8047 at `libero_10`) while the same model's continuous dimensions fall to $R^2 = -0.2714$, worse than the mean, with conformal coverage collapsing from 0.875 to 0.493 at nominal 0.90 (Table 6, appendix). Per-dimension and reliability breakdowns, and the LIBERO analogue of Figure 1, are in Appendix B. Table 2 sweeps six post-hoc corrections on VPT's button head. Every one reduces ECE, the best by $9\times$; AUROC spans 0.8146 to 0.9053 against an uncalibrated 0.8993, so **the ECE-optimal method is never the**

Table 1: Three ordinal heads, three models, in distribution. Exact-bin ECE says temperature helps enormously; tolerant ECE and expected bin error say it hurts. Only full-distribution methods are compared, since confidence-scalar maps leave the ordinal scores undefined.

head	method	ECE_1	$\mathrm{ECE}_{\pm 1}$	RPS	bin err.
VPT camera (11 bins)	uncalibrated	0.4272	**0.0064**	0.2664	
	temperature	**0.1875**	0.2583	0.2588	
LIBERO conv (21 bins)	uncalibrated	0.3596	**0.0644**	0.8937	**1.0733**
	temperature	**0.0434**	0.3692	**0.7779**	1.4828
LIBERO VERA ViT (21 bins)	uncalibrated	0.5503	**0.2209**	1.2137	**1.2896**
	temperature	**0.0348**	0.3550	**0.9257**	1.8318

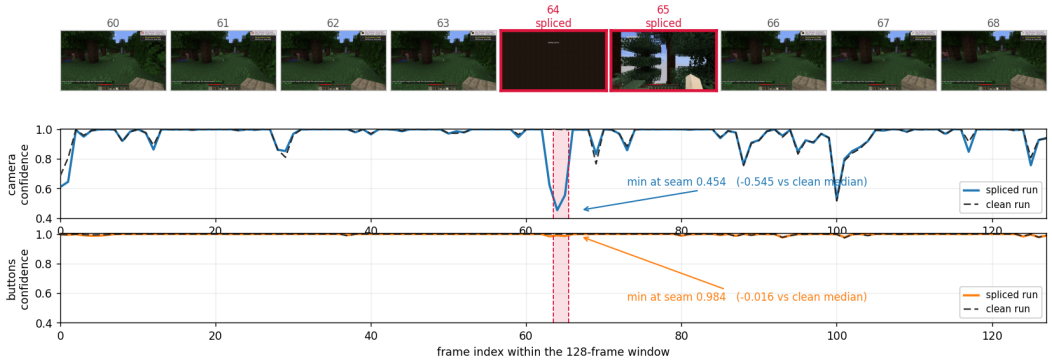

Figure 1: VPT over one 128-frame window, with two successor frames spliced in from another contractor's recording, so that no action can explain the transitions at the seam. Camera confidence collapses there ($0.999 \to 0.454$) while button confidence passes straight through ($1.000 \to 0.984$): the ordinal head reacts, the binary head does not. The example is the *median* of 24 scanned windows by camera seam drop, chosen by order statistic rather than by eye.

AUROC-optimal one and both Platt variants *lose* roughly 0.08 of ranking against doing nothing. A practitioner selecting on ECE, which is standard, ships a model worse at knowing when it is wrong. Nor is confidence uniform *within* a head: across the 20 button dimensions AUROC when pressed spans 0.632 to 0.955, so a gate on one pooled confidence averages over a spread as wide as the gap between heads.

A probe needing no labels. Splicing successor frames from elsewhere creates transitions no action can explain, so confidence there is a hallucination by construction. On LIBERO both heads separate these from clean data (gripper AUROC 0.9695, pose 0.8600) while falling to chance on interior frames, where motion is coherent but the scene is wrong: confidence responds to physical impossibility, not unfamiliarity. Temperature and Platt scaling leave the binary head's detection *exactly* unchanged at 0.9695, as monotonicity requires (Table 3, appendix).

5 What survives shift, and what to do

Transferring each map unchanged to two *different* kinds of shift (synthetic occlusion and a natural recorder-version change), the orderings disagree, and that disagreement is the result. Under occlusion a kNN feature-density temperature wins on both axes at once (0.0294 ECE at 0.8257 AUROC, against 0.0996 and 0.7894 uncalibrated), the only method in our study to improve calibration and ranking together; on the natural shift that advantage evaporates and plain temperature is best. What holds on *both* is narrower: **temperature never degrades ranking and is competitive everywhere**, while both Platt variants degrade it under both. Two conditions sufficed to eliminate two successive apparent winners, which is why we distrust a recommendation resting on two (Table 4, appendix). Action-wise scaling [Zollo and Zemel, 2025] is conditional rather than blanket: it wins on ordinal

Table 2: Post-hoc corrections on VPT's button head, held-out contractors, non-null. "n/a" = undefined for confidence-scalar maps; † changes predictions, so its AUROC is not comparable.

method	ECE_1	ECE_2	NLL	AUROC
uncalibrated (raw)	0.0491	0.0599	0.0448	0.8993
as shipped (T=4)	0.0281	0.0391	0.0162	0.9038
temperature	**0.0057**	**0.0077**	0.0155	0.9038
last-layer Laplace	0.0186	0.0233	**0.0145**	0.8939
kNN density	0.0074	0.0159	0.0189	**0.9053**
Platt, global	0.0330	0.0485	n/a	0.8180
Platt, action-wise	0.0276	0.0445	n/a	0.8146
per-dim Platt†	0.0083	0.0093	0.0153	0.9021

heads in distribution (both LIBERO trunks, all four metrics) and fails on binary heads, where it is the worst method in Table 2. Its advantage does not transfer, being worse on ECE at every shifted rung (Table 5, appendix).

The one choice we recommend without qualification. At fixed architecture the regression head reaches $R^2 = 0.8057$ and the discretised head 0.7959: discretisation costs 0.0098 R^2 and buys a per-dimension predictive distribution a point estimator structurally lacks, since a Huber-trained head admits no calibration without inventing a noise model or falling back on conformal intervals. If an IDM's confidence will be consumed downstream, discretise.

6 Conclusion

Every post-hoc method we tried makes an IDM's confidence *read* honestly while leaving it nearly useless for deciding which predictions to believe; and on ordinal action spaces the standard remedy is worse still, moving probability mass away from the physically correct action while the metric that selected it improves. Three things follow: choose the correction by the decision, since a gate needs ranking and temperature is the only method here that never damaged it; never score tokenised continuous actions with exact-bin ECE alone; and prefer the discretised head when the confidence will be consumed, at a cost of 0.0098 R^2. We cannot claim a universal collapse under shift, three of six perturbation families moved accuracy by under 0.01, and of the three that bit only additive noise inverted error detection while occlusion did the opposite. Reaching that scope required withdrawing twenty claims of our own, collected in Appendix A. We acknowledge that three IDMs is a small sample; diagnosing more is the obvious next step.

Use of AI

The experiments and writing in this paper were carried out with an autonomous coding agent, Claude Code (Opus 5), working under close human direction. The agent wrote the research code, ran the sweeps on a shared GPU cluster, and produced the tables and figures, while the authors set the scope, chose the models and metrics, and checked the results. Several early conclusions were wrong and were corrected after human review; the substantive corrections are recorded in Appendix A. The authors take full responsibility for the content.

References

Pulkit Agrawal, Ashvin Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. In *Advances in Neural Information Processing Systems*, 2016. arXiv:1606.07419.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

Bowen Baker, Ilge Akkaya, Peter Zhokhov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video PreTraining (VPT): Learning to act by watching unlabeled online videos. In *Advances in Neural Information Processing Systems*, 2022.

Yilun Du, Sherry Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems*, 2023.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2016.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, 2017.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Kerui Li, Zhe Jing, Xiaofeng Wang, Zheng Zhu, Yukun Zhou, Guan Huang, Dongze Li, Qingkai Yang, and Huaibo Huang. StableIDM: Stabilizing inverse dynamics model against manipulator truncation via spatio-temporal refinement. *arXiv preprint arXiv:2604.17887*, 2026a.

Sizhe Lester Li, Evan Kim, Xingjian Bai, Tong Zhao, Tao Pang, Max Simchowitz, and Vincent Sitzmann. Turning video models into generalist robot policies. *arXiv preprint arXiv:2605.27817*, 2026b. VERA; project page `https://vera.csail.mit.edu`.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. LIBERO: Benchmarking knowledge transfer for lifelong robot learning. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*, 2023.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International Conference on Machine Learning*, 2017. arXiv:1705.05363.

Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, 2022.

Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *International Conference on Learning Representations*, 2025. Oral.

Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. In *International Joint Conference on Artificial Intelligence*, 2018.

Ruixiang Wang, Qingming Liu, Yueci Deng, Guiliang Liu, Zhen Liu, and Kui Jia. EVA: Aligning video world models with executable robot actions via inverse dynamics rewards. *arXiv preprint arXiv:2603.17808*, 2026.

Thomas P. Zollo and Richard Zemel. Confidence calibration in vision-language-action models. *arXiv preprint arXiv:2507.17383*, 2025.

The material below was in the body of earlier drafts and is retained for completeness and reproducibility. None of it is required to follow Parts 1–3.

A Retraction log

Twenty claims made during this project were withdrawn or corrected. We report the count because it is the honest denominator for everything in the main text: a reader should know how many times the study's own conclusions moved before settling. The substantive ones, most consequential first:

1. **The VPT button head is badly calibrated in distribution.** Reported as accuracy 0.715, ECE 0.257, AUROC 0.541, "barely above chance". Withdrawn: those frames were the first four windows of each recording, which are loading screens and spawns. On representative frames the head is 0.9249 accurate with ECE 0.0470 and AUROC 0.8289. The correction reverses the sign of the conclusion for this head, and was caught only because two of our own pipelines disagreed on a quantity both claimed to compute identically. What survives is sharper: confidence is *flat* across the two regimes (0.9723 vs 0.9718) despite a 0.29 accuracy gap.
2. **Below-chance error detection on clean data.** Reported from two recordings (accuracy 0.37, AUROC 0.47); at 52 recordings the values became 0.715 and 0.541, and then see above. Retracted twice.
3. **Choosing temperature by NLL costs 0.133 AUROC.** Measured on the single demonstration recording shipped with VPT. At scale the cost is 0.0253 on the camera head and exactly 0 on the button head, where AUROC is invariant for all $T \geq 3$ as a binary head must be. The direction survived; the magnitude was overstated fivefold.
4. **Action-wise Platt scaling is the repair that works.** Reported as lifting button AUROC to 0.8348 under occlusion against global Platt's 0.5639, and as "the only prediction-preserving method improving both axes". Withdrawn on two counts. The occlusion cache it was measured on held only windows 0–3, and on the rebuilt full cache the figures are 0.7693 against 0.7002; both *below* the uncalibrated baseline of 0.7894. In distribution on the button head it is the worst method we tested, 0.085 below doing nothing. What survives is the narrower published claim: action-wise beats global, on ordinal heads, in distribution.
5. **kNN density tempering is the repair that works.** Its replacement, withdrawn within a day: it wins on both axes under severe occlusion (0.0294 / 0.8257) and loses to plain temperature on a natural recorder-version shift. Two successive "winners" each failed a second shift condition, which is why the final claim is stated over two shifts rather than one.
6. **Session starts and play are two regimes.** The window-position effect is a *gradient*, not a step: 9.x button AUROC climbs $0.4906 \rightarrow 0.9177$ over roughly 24 windows and plateaus only near window 16, so our `-min-window 4` filter was itself insufficient (windows 4–7 sit at 0.5393). It is also version-specific, absent in 6.x, so not a general property of "the first frames of a video".
7. **The splice asymmetry between VPT's two heads reflects head geometry.** On VPT the ordinal head detected an impossible transition at AUROC 0.9584 and the binary head managed 0.6829, and we read the gap as a property of output geometry. Running the same probe on LIBERO put its two heads within 0.005 of each other (0.8651 and 0.8600, with a matched self-donor control at 0.8172 and 0.7904). The asymmetry is a property of VPT. What replicates is the boundary/interior dissociation, both models detect physical impossibility, neither detects unfamiliarity, which is the claim we now make.
8. **Temperature scaling cannot change ranking.** True only for binary heads. For $K > 2$ the maximum probability depends on the whole logit vector, so samples can swap order; and the corrected version is worse for temperature scaling, not better.
9. **Window-edge frames will be worse calibrated.** They are better. Refuted, mechanism unknown.
10. **LIBERO's R^2 plateau is an aleatoric ceiling.** Withdrawn: the privileged-state probe recovers 0.941, so it was architectural.

11. `libero_goal` **will be the mildest shift** because the model has no language input. Refuted: R^2 falls $0.78 \rightarrow 0.26$. Goal-agnostic architecture does not imply goal-invariant performance.

B Supplementary tables and figures

Referenced from the main text but held here for space. Nothing in this subsection duplicates a float in the body; the earlier draft's full text has been removed rather than reproduced, since keeping it defined every body label twice.

Splice detection on LIBERO under each correction.

Table 3: Splice detection on LIBERO under each correction, cross-donor. Maps fitted on clean labelled frames, applied to the unlabelled splice conditions. The binary head's column is constant by construction: temperature and Platt-on-confidence are monotone in confidence, and AUROC depends only on ordering. Scored on the log-odds margin, since 73.5% of clean gripper confidences are exactly 1.0.

condition	correction	gripper AUROC	pose AUROC
boundary (impossible)	uncalibrated	**0.9695**	0.8600
	temperature	**0.9695**	0.8158
	Platt (conf)	**0.9695**	**0.8701**
interior (coherent, wrong scene)	uncalibrated	0.3344	0.3137
	temperature	0.3344	0.3627
	Platt (conf)	0.3344	0.3242

Transfer of every fitted map to two kinds of distribution shift.

Table 4: Transfer, button head, non-null: every map fitted once in distribution and applied unchanged. Occlusion is synthetic and severe; the recorder-version change is natural and mild (it makes the model *easier*, so absolute numbers are higher). Best per column in bold. The two columns do not agree on a winner, which is why we report both.

	occlusion (severity 3)			natural (9.x $\rightarrow$ 6.x)		
method	ECE_1	AUROC	selR@80	ECE_1	AUROC	selR@80
uncalibrated	0.0996	0.7894	0.0556	0.0352	0.9473	0.0026
as shipped (T=4)	0.0765	0.8081	0.0556	0.0146	**0.9498**	0.0026
temperature	0.0443	0.8081	0.0556	0.0078	**0.9498**	0.0026
last-layer Laplace	0.0645	0.7214	0.0643	**0.0072**	0.9402	0.0038
kNN density	**0.0294**	**0.8257**	**0.0530**	0.0148	0.9489	0.0027
Platt, global	0.0808	0.7002	0.0656	0.0204	0.8985	0.0070
Platt, action-wise	0.0755	0.7693	0.0648	0.0148	0.8966	0.0084

Action-wise versus global Platt across two architecture families.

Table 5: Action-wise versus global Platt on 21-bin ordinal pose heads across a real task-shift ladder, for *two architecture families* on identical data, splits and evaluation. The conditional structure replicates: action-wise wins on every metric in distribution and loses on calibration at every shifted rung, in both a 1.17M convolutional trunk and a 12.35M ViT trunk. Bold marks the better map within each block and rung.

set	map	ECE_1	ECE_2	AUROC	selR@80
convolutional trunk (1.17M; fitted $a \in [0.221, 0.399]$, $T = 6.71$)					
in-dist.	global	0.0231	0.0331	0.7122	0.5210
	action-wise	**0.0160**	**0.0203**	**0.7187**	**0.5158**
`libero_goal`	global	**0.1534**	**0.1549**	0.6910	0.7412
	action-wise	0.1584	0.1625	**0.6948**	**0.7383**
`libero_object`	global	**0.1496**	**0.1524**	**0.6788**	**0.7348**
	action-wise	0.1570	0.1643	0.6736	0.7354
`libero_10`	global	**0.2219**	**0.2446**	0.6329	0.8348
	action-wise	0.2270	0.2516	**0.6350**	**0.8343**
VERA ViT trunk (12.35M; fitted $a \in [0.085, 0.176]$, $T = 18.96$)					
in-dist.	global	0.0228	0.0281	0.6625	0.3635
	action-wise	**0.0202**	**0.0248**	**0.6836**	0.3635
`libero_goal`	global	**0.1321**	**0.1377**	0.6305	0.7757
	action-wise	0.1348	0.1446	**0.6361**	**0.7713**
`libero_object`	global	**0.1279**	**0.1329**	**0.6226**	**0.7864**
	action-wise	0.1350	0.1461	0.6218	0.7833
`libero_10`	global	**0.1616**	**0.1680**	0.6116	0.8301
	action-wise	0.1667	0.1748	**0.6405**	**0.8241**

The LIBERO task-shift ladder on the frozen checkpoint.

Table 6: LIBERO IDM across a real shift ladder, on the frozen checkpoint. Maps and conformal quantiles are fitted once, in distribution, and transferred unchanged. Note the in-distribution column is the *held-out test* split ($R^2 = 0.7762$); the 0.8057 quoted elsewhere is best *validation* R^2, which is what the architecture grid is selected on. We report both rather than collapsing them.

	in-dist.	goal	object	`libero_10`
mean R^2	0.7762	0.3462	0.0242	**-0.2714**
gripper ECE	0.0040	0.0515	0.0799	0.1189
gripper AUROC	0.9811	0.9100	0.8839	0.8047
coverage @ nominal 0.90	0.8750	**0.6028**	0.6478	**0.4930**

The LIBERO analogue of Figure 1.

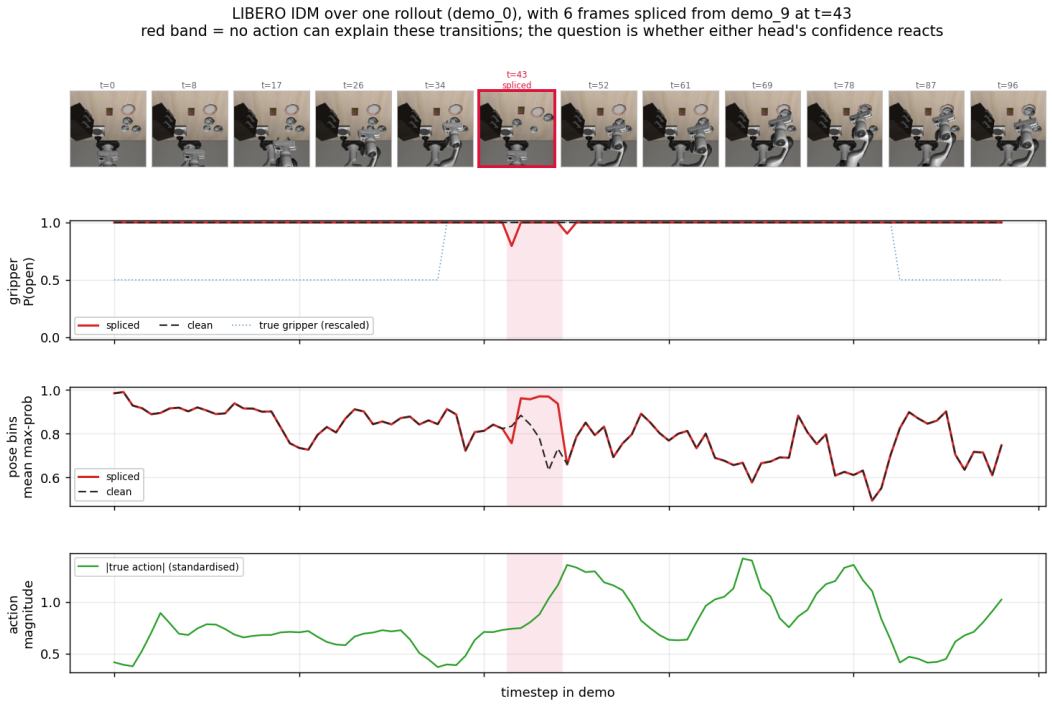

Figure 2: The LIBERO IDM over one held-out rollout, with six successor frames spliced in from a different task (red band). Frame thumbnails above the traces; both heads' confidence below. The probe needs no labels: spliced frames have no ground-truth action, so any confident prediction there is a hallucination by construction.

Per-dimension breakdown across all 20 VPT button dimensions.

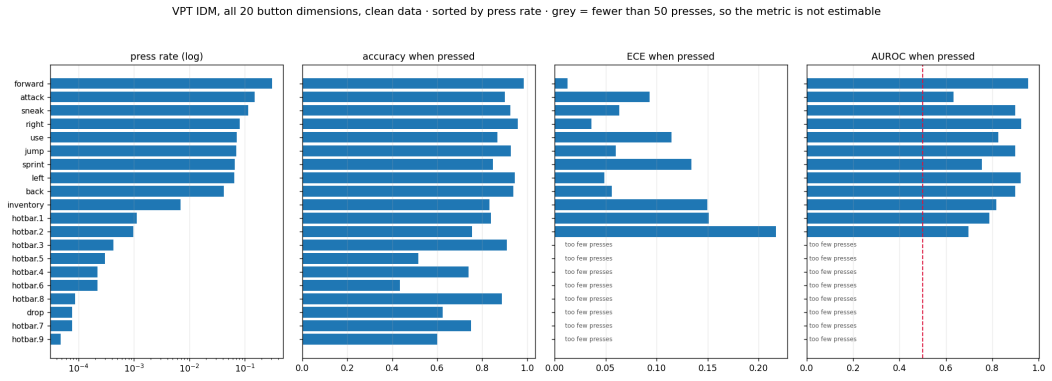

Figure 3: All 20 button dimensions, clean data, sorted by press rate. AUROC when pressed spans 0.632 to 0.955, a 0.323 spread that exceeds the gap between the two heads. Grey bars are dimensions with fewer than 50 presses, drawn as not estimable rather than as measured, which is the honest treatment of a dimension a contractor never used.

Reliability diagrams with bin populations.

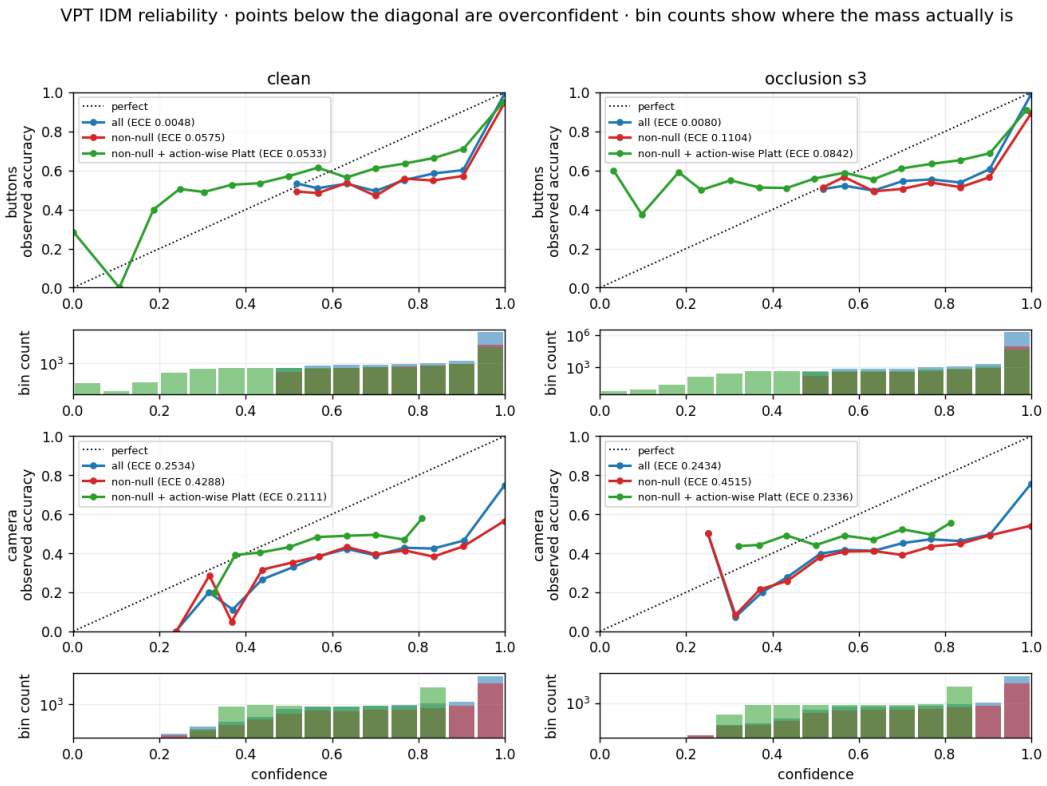

Figure 4: Reliability diagrams, clean and occluded, per head, with bin populations on a log axis beneath each. The counts are the point: a reliability curve without them hides that almost all mass sits in the top confidence bin, so a visually alarming deviation in a sparse bin can carry negligible weight.

Confidence and accuracy against perturbation severity.

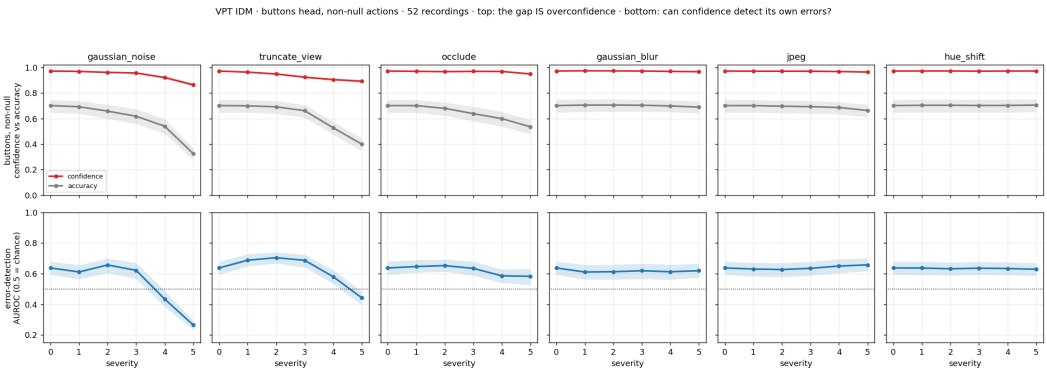

Figure 5: VPT button head, non-null actions, 52 recordings, bootstrap CIs over recordings. *Top:* confidence (red) barely moves while accuracy (grey) collapses; the widening gap is the overconfidence. *Bottom:* error-detection AUROC crosses *below* chance exactly where accuracy is worst. Note also that three of six families (blur, JPEG, hue) move nothing at all.

C Targets, data and protocol

VPT IDM (discrete, released). The only released IDM checkpoint; probing confirmed that $\{1, 2, 3, 5\}\times$ variants do not exist, so a scaling study on the released artefact is impossible. Three properties, read from source and verified against the checkpoint, shaped the protocol. Its action space is *factored*: 20 independent binary button heads (11 controls, 9 hotbar slots) and two 11-way categorical heads over mu-law-quantised camera deltas. This is not a Minecraft quirk: per-component discretisation of continuous quantities is the founding design of learned IDMs in robotics, so diagnosing VPT is diagnosing the field's standard tool. It is *non-causal* over a 128-frame window, which is legitimate (an IDM labels recorded video, and hindsight is what makes inference easier than cloning) but makes "perturb the input" ambiguous. And a temperature of 4.0 is *baked into the released checkpoint*: any study reporting "the uncalibrated baseline" without reading that field describes an already-tempered model as raw.

LIBERO IDM (continuous, ours). No released direct-regression continuous IDM exists: IDMs are disposable intermediates, so the most standard formulation appears in many papers and no published checkpoint. We therefore train the canonical recipe (siamese encoder over both frames, concatenate, MLP head) on LIBERO. Dimension 6 of the action is exactly $\{-1, +1\}$, so it receives a Bernoulli head rather than being regressed; the position dimensions span $\sim 6\times$ the range of the rotation dimensions, so targets are standardised using training-split statistics only, and we use GroupNorm, because batch statistics would make one frame's prediction depend on which others share its batch, which is fatal for a study of per-sample confidence.

VERA IDM (continuous, released code). A third arm that varies architecture family while holding the task, action space and data fixed. VERA releases four IDM algorithms; `pure_transformer_action` is a plain ViT (patch 16 here rather than 14, since the backbone requires `image_size` $\%$ `patch_size` $= 0$ and LIBERO renders at 128), embedding 384, depth 6, six heads, **12.35M parameters** against our conv net's 1.17M. Its released defaults are `input_mode="rgb_pair"`, `tokenizer_in_channels`$= 6$, `action_dim`$= 7$: a stacked frame pair in, a LIBERO action out. Two deviations we state rather than hide. We use VERA's trunk with *our* output heads, because VERA regresses the gripper (exactly ± 1 in LIBERO) under a squared loss, so as published its IDM has no calibratable output at all; holding the output geometry fixed makes this a comparison of trunks rather than of trunks and heads at once. And we train it in our harness rather than VERA's, because adopting theirs would swap the data pipeline, the splits and the evaluation simultaneously and confound the architecture comparison.

Data. For VPT: 118 contractor recordings across recorder-version indices 9.x and 6.x, giving 85 distinct contractors. Splits are *contractor-disjoint* throughout, since calibration constants fitted on one recording of a player and evaluated on another are optimistic. Two of 54 recordings sampled from 9.x are corrupt *at source*, byte-identical to the server's `Content-Length`: an 80-byte video with a 0-byte action file, and a full 172 MB video whose action file is not valid UTF-8. For LIBERO: `libero_spatial` for training (49,520/6,011/6,219 transitions under a by-demo split), with `libero_goal`, `libero_object` and `libero_10` as a shift ladder.

D Extended related work

IDMs as infrastructure. The formulation is old, Agrawal et al.'s *Learning to Poke by Poking* [Agrawal et al., 2016] recovers a discretised poke from a before/after image pair as a sum of cross-entropies, and Pathak et al. [Pathak et al., 2017] reduce it to the inverse model inside an intrinsic-curiosity module, with Torabi et al. [Torabi et al., 2018] using it for behavioural cloning from observation. What changed is the scale of the trust placed in it. VPT [Baker et al., 2022] trains an IDM on a few thousand hours of annotated play and uses it to pseudo-label roughly 70k hours; UniPi [Du et al., 2023] inverts a generated video plan step by step into actions; and a growing line of work uses an IDM's implied actions as a training signal for video generators [Wang et al., 2026]. Recent systems extend the idea in both directions: predictive IDMs that forecast and act in one trunk [Tian et al., 2025], and refinement architectures that stabilise action recovery when the manipulator leaves frame [Li et al., 2026a]. Notably several of these abandon future conditioning entirely, predicting a_t from a

causal history $o_{t-K+1:t}$ rather than from (o_t, o_{t+1}). We take the future-conditioned formulation as our scope (§2) and note that the term has broadened to cover models that are, strictly, policies.

Calibration of deep classifiers. Modern networks are systematically overconfident, and temperature scaling [Guo et al., 2017], a single scalar dividing the logits, remains the strongest simple remedy; matrix and vector scaling generalise it at the cost of changing predictions. Our contribution is not a new method but the observation that the standard selection criterion for these methods is the wrong one when the downstream use is a threshold. The distinction between calibration and *ordering* is well established in selective prediction and risk–coverage analysis [Hendrycks and Gimpel, 2017], and our results are an instance of a known theoretical point, monotone recalibration cannot change a ranking, made empirically consequential by the fact that practitioners nonetheless select on ECE.

Uncertainty in vision–language–action models. Closest to us is work on confidence calibration in VLA models [Zollo and Zemel, 2025], which diagnoses per-dimension over- and under-confidence and proposes action-wise Platt scaling as a remedy. We test that proposal rather than adopt it, on four ordinal and two binary heads across three models, and find it conditional: it beats a shared map where the dimensions are heterogeneously miscalibrated and the test distribution matches the fitting distribution, and not otherwise. Bayesian and ensemble treatments (last-layer Laplace approximations, deep ensembles [Lakshminarayanan et al., 2017], MC dropout [Gal and Ghahramani, 2016]) are the alternative family; we include Laplace and a kNN feature-density temperature [Sun et al., 2022] as input-dependent representatives, and exclude ensembles and MC dropout, the latter because the released VPT checkpoint contains no dropout at all.

Distribution-free coverage. Where a head is a point estimator there is no distribution to calibrate, and split-conformal prediction [Angelopoulos and Bates, 2021] supplies intervals with finite-sample coverage under exchangeability. We use it as the honest statement of what a Huber-trained regression head can offer, and report the coverage collapse under task shift as a measurement of how badly exchangeability fails rather than as a criticism of the method.

E Metric definitions

Two families of quantity are involved and conflating them is the mistake this paper is about. Write $\hat{p}$ for the model's confidence in its top prediction and $c \in \{0, 1\}$ for whether that prediction was correct.

Calibration: does the number mean what it says? Binning confidences into B equal-width bins,

$$\mathrm{ECE}_q \;=\; \Big(\sum_{b=1}^{B} \tfrac{n_b}{n} \,\big|\, \mathrm{acc}(b) - \mathrm{conf}(b) \big|^q \Big)^{1/q},$$

with $q = 1$ throughout and $q = 2$ reported alongside because the ℓ_2 form penalises a few badly-wrong bins more heavily than many slightly-wrong ones. We use $B = 15$. Brier score and negative log-likelihood are the two proper scoring rules; we report both because a method can improve one while worsening the other.

Discrimination: can the number be thresholded? Error-detection AUROC is the area under the ROC curve for the binary task of predicting $\neg c$ from $1 - \hat{p}$, i.e. whether *ranking* by confidence separates correct predictions from incorrect ones. It is invariant to any monotone transform of $\hat{p}$, which is exactly why temperature scaling can move ECE a great deal and AUROC not at all on a binary head. Selective risk at coverage κ is the error rate over the κ most-confident fraction, the quantity a deployed filter actually experiences, and AURC integrates it over κ.

Ordinality: is a near-miss scored as a near-miss? When bins are ordered (VPT's mu-law camera deltas, our discretised pose dimensions), cross-entropy and ECE treat an adjacent-bin prediction as a total failure, which for an action is false: bin distance is action error. We therefore also report the Ranked Probability Score,

$$\mathrm{RPS} \;=\; \tfrac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \big(F_{\hat{p}}(k) - F_y(k) \big)^2,$$

a proper scoring rule over the cumulative distribution; expected bin error $\mathbb{E}_{k\sim\hat{p}}|k-y|$, which is in units of bins; and a tolerance-1 variant of ECE in which a prediction counts as correct if it lands within one bin of the truth. §5 shows these three do not agree with exact-bin ECE about which calibrator is best, and that the disagreement is systematic across models.

Non-null conditioning. Because no-ops dominate both action spaces, every figure is reported restricted to non-null actions as well as overall. Unless stated otherwise, quoted numbers are non-null.

F Window position and the confidence gradient

Where in the recording you look. Our perturbation sweep and our calibration study were separate pipelines: the sweep re-runs the model under each corruption, while the calibration work reads cached logits. Both compute non-null accuracy with identical code. On the same 52 recordings they disagreed, 0.7148 against 0.8905 for the button head, while agreeing on the camera head (0.5465 against 0.5551). The cause is a default: the sweep took `-max-windows 4`, i.e. the *first* four 128-frame windows of each recording, while the cache used all of them. A Minecraft recording opens on a loading screen and spawn, so windows 0–3 are session starts. Splitting the cache by window position shows the effect is a *gradient*, not two regimes, and that it is confined to one recorder version (Table 7). In 9.x recordings, non-null button accuracy climbs $0.6218 \to 0.9401$ over the first $\sim$24 windows and error-detection AUROC climbs $0.4906 \to 0.9177$, reaching its plateau only around window 16, while confidence stays flat throughout, $0.9735 \to 0.9687$. In 6.x recordings there is no gradient at all: 0.9386 accuracy and 0.9276 AUROC from window 0. Two things follow. First, the model is badly wrong at session starts and equally confident there, which is a stronger statement than any two-bucket comparison. Second, dropping only windows 0–3 is insufficient: 9.x windows 4–7 still sit at AUROC 0.5393. This exactly explains the discrepancy that exposed the problem, since the sweep saw windows 4–7. Every severity rung inherited it, so the sweep as first run characterises the rising part of the curve. The camera head shows no gradient, which is why only one of the two pipelines disagreed.

Table 7: The same cached features, button head, non-null, split by position in the recording and by recorder version. Confidence is flat down both columns while accuracy and AUROC climb steeply in 9.x, so this is a genuine calibration finding, not merely a sampling artefact: the model does not know that session starts are hard. 6.x shows no gradient, so the effect is version-specific rather than a universal property of "the first frames of a video".

	9.x recordings				6.x recordings			
windows	acc	conf	ECE	AUROC	acc	conf	ECE	AUROC
0–3	0.6218	0.9735	0.3518	0.4906	0.9386	0.9469	0.0363	0.9276
4–7	0.8742	0.9765	0.1023	0.5393	0.9511	0.9652	0.0160	0.9565
8–15	0.9161	0.9732	0.0571	0.7565	0.9707	0.9706	0.0086	0.9536
16–23	0.9357	0.9713	0.0356	0.9027	0.9671	0.9792	0.0138	0.9670
24–31	0.9401	0.9719	0.0331	0.9177	0.9603	0.9718	0.0117	0.9642
32–47	0.9286	0.9687	0.0408	0.9048	0.9576	0.9746	0.0175	0.9448

G The full perturbation ladder

Under shift, discrimination degrades but rarely inverts. Table 8 and Figure 5 give the button head across three live perturbation families, re-run with session-start windows dropped. The picture is milder than the one we first reported and it differs by family, which is the point. Gaussian noise is the only family whose AUROC crosses below chance, and it reaches 0.4571 rather than the 0.2906 we obtained before the window correction. Occlusion produces *no* inversion at all: AUROC rises monotonically to 0.7159 even as accuracy falls to 0.7077, i.e. uncertainty tracks the damage. Truncating the view peaks at 0.7866 and then decays to 0.5383. Confidence–error decoupling now reads "tracks error" on all three (1.048, 0.933, 1.189). Three of six families (blur, JPEG, hue) move nothing at all, and blur mildly *improves* the camera head; we report those as null results. The blanket claim that IDMs suffer epistemic collapse under shift is not supported by our data; the specific claim that *severe additive noise* inverts error detection is.

Table 8: VPT button head, non-null actions, 51 recordings, session-start windows dropped. Only additive noise inverts error detection; occlusion does the opposite and truncation is non-monotone. Reporting a single family, as our first draft did, would have supported whichever conclusion that family happened to give.

severity	0	1	2	3	4	5
Gaussian noise (aleatoric); the only family that inverts						
accuracy	0.8883	0.8867	0.8760	0.8446	0.7601	0.4899
confidence	0.9748	0.9725	0.9656	0.9558	0.9094	0.8243
ECE	0.0866	0.0858	0.0900	0.1115	0.1493	0.3345
AUROC	0.6245	0.6260	0.6626	0.6900	0.6102	**0.4571**
Occlusion (epistemic); no inversion; AUROC rises monotonically						
accuracy	0.8883	0.8895	0.8791	0.8539	0.8082	0.7077
ECE	0.0866	0.0847	0.0936	0.1153	0.1509	0.2190
AUROC	0.6245	0.6223	0.6480	0.6664	0.6709	**0.7159**
Truncated view (epistemic), peaks, then decays						
accuracy	0.8883	0.8898	0.8743	0.8505	0.7203	0.5248
ECE	0.0866	0.0838	0.0841	0.0850	0.1684	0.3459
AUROC	0.6245	0.7145	**0.7866**	0.7741	0.6710	0.5383

H Confidence–error decoupling by perturbation family

Most corruptions do not produce epistemic collapse. Reporting only the striking families would misrepresent the sweep. Table 9 gives the confidence–error decoupling ratio (normalised slope of reported uncertainty against normalised slope of realised error) for every family. On all three families that actually damage the model the ratio sits between 0.93 and 1.19: uncertainty does track error. The other three are inert probes; they move accuracy by less than 0.01, so the ratio is a quotient of two quantities that are both noise, and we mark them uninterpretable rather than reporting a value. An earlier version of this table gave ratios for all six and labelled one "decoupled"; that label came from dividing by an error slope that was negative, which is the effect-size guard we subsequently added.

Table 9: Confidence–error decoupling by perturbation family. Values near 1 mean uncertainty rises as fast as error does. Interpretable only where the perturbation actually hurt the model.

family	kind	CED	reading
truncate view	epistemic	1.189	tracks error
Gaussian noise	aleatoric	1.048	tracks error
occlude	epistemic	0.933	tracks error
Gaussian blur	aleatoric	inert (accuracy moves <0.01)	
JPEG	aleatoric	inert	
hue shift	epistemic	inert	

I Frame-scope versus window-scope corruption

Architectural context is not calibration failure. Corrupting a single frame inside a clean window costs at most 0.0034 non-null accuracy, against 0.3585 for whole-window corruption. Bidirectional context almost entirely repairs isolated corruption: a robustness property, and the reason window-scope and frame-scope results must be reported separately. We also predicted that window-edge frames would be worse calibrated, having one-sided context. They are better (camera ECE 0.171 at the edge, 0.215 mid-window). The mechanism is unexplained.

J Selecting temperature by NLL

Selecting temperature by NLL is nearly free, contrary to our own earlier report. An early sweep on the single demonstration recording shipped with VPT suggested that choosing T by NLL cost 0.133 AUROC on the camera head. Repeated at scale on representative frames the cost is 0.0251,

a fivefold overstatement, and on the button head it is $+0.0001$, i.e. nothing: non-null AUROC is flat at 0.9038 under every selection criterion, which is the rank-invariance a binary head must exhibit. Accuracy is invariant to machine precision throughout (0 spread), as it must be under a monotone transform of the logits. Criteria do still disagree about T (on the camera head NLL picks 20, ECE_2 and Brier pick 16, and RPS picks 12, three different answers from three defensible objectives), so the objective matters; it just matters far less than a single recording implied. We report the correction because that recording is the one a reader is most likely to reach for.