Journal of Audit and Accountability
Volume 1 · Number 1 · 2026

Four Hidden Features, a Different Task: Calibration After a Missing Block Moves

Logan Franklin[1] · Jacob Knight[2] · Gavin Bell[3]

[1] Canada · [2] Germany · [3] United States

Original publication: November 7, 2026 | Article 674

Abstract

Keeping the number of hidden features fixed did not preserve prediction-interval coverage. A radius calibrated for a four-feature corner block covered all four corner targets in 92.52% of held images, but covered the central block in only 25.52%. Calibrating on central-block errors instead gave 91.60% center coverage and substantially wider intervals. The images, fitted basis and evaluation cohort stayed fixed. What changed was the hidden-region task and its calibration evidence. The result illustrates why a nominal coverage label should travel with its score definition and calibration conditions, rather than with the missing-feature count alone.

Spatial-mask transfer in a public digit-feature resource. Matched center calibration is a diagnostic requiring center-region truth in separate calibration images; it is not free information at deployment.

Four Hidden Features, a Different Task: Calibration After a Missing Block Moves

> Spatial-mask transfer in a public digit-feature resource. Matched center calibration is a diagnostic requiring center-region truth in separate calibration images; it is not free information at deployment.

Abstract

Keeping the number of hidden features fixed did not preserve prediction-interval coverage. A radius calibrated for a four-feature corner block covered all four corner targets in 92.52% of held images, but covered the central block in only 25.52%. Calibrating on central-block errors instead gave 91.60% center coverage and substantially wider intervals. The images, fitted basis and evaluation cohort stayed fixed. What changed was the hidden-region task and its calibration evidence. The result illustrates why a nominal coverage label should travel with its score definition and calibration conditions, rather than with the missing-feature count alone.

What must remain comparable when an interval is reused?

A restoration interval can be calibrated using complete examples in which part of an image is deliberately hidden. It is tempting to reuse that interval when a later image has the same number of missing values. Yet the spatial position of those values changes what must be inferred and which visible measurements remain available. A count-matched missing block can therefore define a different prediction problem.

We ask whether a joint interval calibrated at an image corner keeps its empirical coverage when a four-feature block moves to the center. We compare that transferred radius with a center-matched radius estimated from the same separate calibration images. The center prediction rule and held center targets are identical in that paired comparison, so only the calibration radius changes.

Meng et al., 2025 develop zero-shot inpainting for oracle-bone inscription rubbings with a visual Mamba block. Our predictor is a fitted linear completion baseline and is not that model. The connection concerns uncertainty in missing visual content: a statement about reconstruction coverage needs to identify which region, prediction rule and calibration observations support it. Neither a small block nor a nominal percentage alone supplies that information.

Conformal-prediction introductions and regression theory distinguish a score-based calibration guarantee from an unrestricted statement about arbitrary future tasks (Shafer and Vovk, 2008; Lei et al., 2018). Our experiment measures a specific transfer that changes the score-generating task. It does not refute a guarantee whose required comparability assumptions hold; it demonstrates why those assumptions should be checked when the hidden region changes.

Complete source features before the masks are applied

The Optdigits resource provides 3,823 official training images and 1,797 official test images represented by sixty-four block-count features on an eight-by-eight grid (Alpaydin and Kaynak, 1998). Values are divided by sixteen to lie between zero and one. The task reconstructs those numerical block counts. We do not treat them as high-resolution image intensities or claim that four recovered numbers establish character readability.

All training feature vectors are used to fit one mean and a rank-sixteen principal-component basis. Singular-value decomposition supplies the basis (Golub and Reinsch, 1970). At prediction time, visible coordinates are fitted in that basis with latent ridge parameter 0.01, and the corresponding hidden

coordinates are reconstructed. Regularization provides a declared stabilization of this fixed low-rank solve (Hoerl and Kennard, 1970).

For a given mask, its sixty visible coordinates and four hidden coordinates define the projection matrices used in the latent solve. Moving the mask therefore changes the numerical affine operator even though the global mean, basis rank and penalty are unchanged. We do not claim that corner-to-center transfer varies only a label while leaving the entire point predictor unchanged.

Each numerical prediction is clipped to the known feature range before either calibration or evaluation. This operation is common to both masks and is fixed in the design. It prevents intervals from being centered at impossible block-count values. No clipping threshold is selected from the coverage results, and the hidden held values do not enter the prediction calculation.

The two predetermined masks have zero-based indices [0, 1, 8, 9] for the corner and [27, 28, 35, 36] for the center. Both conceal a contiguous two-by-two feature block and leave sixty coordinates visible. They are selected before executing the new comparison. Their spatial location is not assumed to be semantically interchangeable merely because their size and shape match.

Separating fitting, calibration and evaluation

The official training partition supplies the basis and mean only. We permute the 1,797 official test-row indices with seed 20261108, assign the first five hundred to calibration and the remaining 1,297 to evaluation. Digit labels are not used to make this division, choose the basis or tune an interval. Calibration and evaluated images are disjoint.

Each complete calibration image supports two artificial hiding exercises, one at each mask. That access is possible because this public resource contains the complete feature vectors. In a deployed restoration problem, center-region truth might not be available for calibration. Our matched-center condition therefore represents a stated information requirement, not an interval adjustment that can be obtained without appropriate complete examples.

The source documentation describes different training and test writers, but individual writer identifiers are unavailable in these files. Randomly allocating rows within the finite test cohort does not establish transport to new writers or independent image acquisition. We preserve that distinction when interpreting the image-level working intervals and any connection to future restoration tasks.

Validation design should follow the intended transfer boundary (Roberts et al., 2017), while leakage analysis motivates keeping evaluated outcomes outside fitting and calibration choices (Kapoor and Narayanan, 2023). Here the separation is local and explicit. It does not certify broad domain validity, and it does not make a threshold calibrated for one hidden coordinate set appropriate for another by default.

Joint calibration over four coordinates

For each calibration image and mask, the nonconformity score is the largest absolute error across its four hidden coordinates. Using the maximum makes the desired event simultaneous coverage of those four targets within one image. It is not four unrelated marginal pixel events and not simultaneous coverage of all 1,297 evaluated images.

At nominal level 0.9 with five hundred calibration images, we select the 451st smallest calibration score. The same order-statistic rule is used for the corner and center tasks. A symmetric interval around each predicted coordinate uses that mask's radius, then intersects the result with the known zero-to-one feature

range. Because the true values lie in that range, the intersection does not discard an otherwise covered true value.

The rank calculation belongs to an exchangeability-based prediction framework, not a universal task-transfer mechanism. Random assignment within a fixed finite cohort supports a particular calibration experiment, while future data require the relevant distributional assumptions. Changing the hidden target coordinates and their prediction operator changes the score definition. Similar image counts do not establish that the resulting corner and center scores are interchangeable.

Extensions for covariate shift have their own weighted-comparability conditions (Tibshirani et al., 2019). We estimate no density ratio and apply none of those weighted procedures. Moving the target block also need not be a pure covariate-distribution shift with an unchanged conditional prediction problem. Citing a general shift-aware method cannot by itself justify the transferred radius used in this stress comparison.

The calibrated corner radius is 0.252649; the center radius is 0.742357. The values are measured in normalized feature-count units. They are not confidence probabilities attached to the reconstructed digit class. The substantially larger center radius reflects the observed calibration-error distribution for that task under the fixed predictor, rather than evidence that its point estimates became better after calibration.

Three evaluated interval conditions

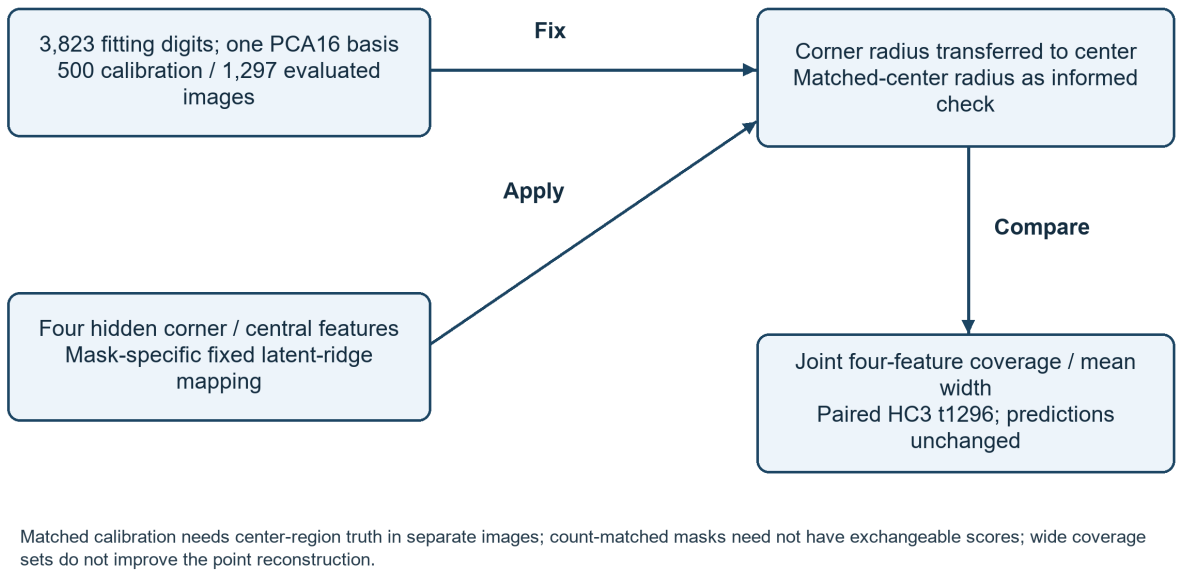

Figure 1. *A common training-only basis supplies two mask-specific completion rules. Corner and center error radii are estimated on the same separate calibration images; transferred and matched radii then cover identical held center predictions and targets.*

We evaluate the corner task with the corner radius, the center task with the corner radius, and the center task with its matched radius. Coverage requires all four held targets to lie in their coordinate intervals. Mean width first averages the four interval widths within image and then averages images, keeping the image as the reporting unit.

Model 1 summarizes the matched-minus-transferred center-coverage difference. Model 2 summarizes the corresponding width difference. Both regressions have an intercept and 1,297 paired image rows, use HC3 covariance and t1296 intervals, and contain no other controls. The two models condition on one fitted basis, one calibration split and the resulting fixed radii.

Heteroskedasticity-consistent covariance states a working image-level variance calculation (White, 1980). It does not propagate the variability of selecting another five hundred calibration rows or capture all writer dependence. The shared estimated radius is particularly important: a tight conditional interval for a width difference does not establish equally precise uncertainty about deployment performance under a new calibration draw.

Table 1. *Matched-center minus corner-calibrated center outcomes, N=1,297, intercept-only OLS, HC3 standard errors and t1296 intervals. M1 is a joint-coverage proportion change; M2 is mean width in normalized feature units. R-squared is zero to numerical rounding.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.660756 (0.013157)	[0.634945, 0.686566]	1297
M1: fit	R-squared = 0.0000	-	-	1297
M2	Mean outcome/contrast	0.491731 (0.000038)	[0.491658, 0.491805]	1297
M2: fit	R-squared = 0.0000	-	-	1297

Model labels: M1 = Matched-center minus corner-calibrated center coverage; M2 = Matched-center minus corner-calibrated center interval width.

The corner radius covers all four corner targets in 1,200 of the 1,297 evaluated images, or 0.925212. Applying that same radius to center predictions covers 331 images, or 0.255204. The nominal label has not changed, yet the evaluated coverage has. This directly identifies the mask-transfer issue under the common image cohort.

The matched center radius covers 1,188 images, or 0.915960. Compared with the transferred condition, the increase is 0.660756, standard error 0.013157 and interval [0.634945, 0.686566]. This is about 66.08 percentage points. It is not a relative percentage improvement in point-prediction accuracy, and neither condition changes the held center point predictions (Stevens, 1946).

Mean center interval width increases from 0.402628 under the transferred radius to 0.894359 under matched calibration. The difference is 0.491731, standard error approximately 0.0000376 and interval [0.491658, 0.491805]. The narrow conditional interval reflects the very similar width increases induced by two fixed global radii. It does not include uncertainty about how those radii would vary with a new calibration sample.

Because the matched center radius is larger in this draw, every transferred center interval is contained in its matched counterpart. Coverage therefore cannot decrease when those two fixed sets are compared. The empirical quantity is how many additional images become covered and how much width is spent, not whether a larger interval can contain more targets. A positive paired gain is not evidence that exchangeability has been restored or that the nominal rate is certified for a new mask.

Corner intervals using the corner radius have mean width 0.286125. The same radius can produce another mean width on center predictions because intervals are intersected with the feature bounds and their centers differ. We therefore report realized widths rather than replace them with twice the nominal radius. The clipping convention is part of the uncertainty set's numerical definition.

Why near-nominal coverage does not identify content fidelity

A matched interval can recover a desired coverage level by becoming wide. The center intervals here occupy much of the allowed feature range. That may be appropriate for uncertainty reporting, but it should not be described as a better completed image. The point predictor and its missing-content errors remain the same between the two center-radius conditions.

Prediction-set quality involves both the coverage objective and the size of the resulting set (Gneiting and Raftery, 2007). Neither an empirical percentage close to ninety nor a large gain in coverage settles the value of a reconstruction for another task. Agreement and measurement frameworks encourage retaining the distinct quantities rather than using one as a surrogate for the other (Bland and Altman, 1986; Lin, 1989; Carroll et al., 2006).

Statistical reporting should also distinguish an observed rate from a certified universal guarantee (Wasserstein and Lazar, 2016). The finite-sample rank formula supplies a particular guarantee under its conditions; the observed rate in one held batch can differ from the nominal value. Our matched-center result is reported as an executed empirical outcome, not as proof that every future image, writer or mask will satisfy ninety-percent coverage.

Distribution-shift work emphasizes that reliability can change when the evaluation boundary moves (Ovadia et al., 2019). Here the images can remain from the same source while the reconstruction task changes spatially. This makes the missing-region definition part of the calibration specification. An error distribution learned where most targets are near the boundary of a symbol need not characterize its central region.

Restoration methods and the limits of this baseline

The fixed low-rank rule differs from modern image-completion systems. Context encoders, LaMa and RePaint use other priors and representations (Pathak et al., 2016; Suvorov et al., 2021; Lugmayr et al., 2022). Gated, partial and edge-connected inpainting organize missing-region information differently (Yu et al., 2018b; Liu et al., 2018; Nazeri et al., 2019). None is run, and their uncertainty calibration is not measured by our linear reconstruction result.

U-Net and denoising representations provide further image-processing contexts (Ronneberger et al., 2015; Vincent et al., 2008). Statistical imputation frameworks such as chained equations and missForest also involve their own modeling and uncertainty assumptions (Buuren and Groothuis-Oudshoorn, 2011; Stekhoven and Bühlmann, 2012). A calibration scheme would need to be paired with the actual predictor and target task in each case; it cannot inherit validity from the number of missing coordinates alone.

We do not infer a general explanation that corners are always easier than centers. These two masks, this digit representation and this training basis produce the reported error distributions. Another symbol collection or mask pattern could change that relationship. Influence analysis and representative sampling would be relevant before extending the result to a broader restoration inventory (Cook, 1977).

The hiding operation is also artificial. It suppresses four known block counts in complete source examples and leaves the other sixty available. It does not reproduce irregular erosion, bleed-through, scanning artifacts or a missing fragment of an inscription. Those mechanisms can alter both visible evidence and the target definition, so they would require their own calibration data rather than an interpretation borrowed from this coordinate-mask comparison.

The saved calibration file records both score vectors and their source-image indices, while the held table records joint coverage and interval widths in all three conditions. Dataset documentation preserves the

mask and information boundaries (Gebru et al., 2021), and provenance makes the radius calculation reconstructible (Wilkinson et al., 2016). In this experiment, keeping the hidden-feature count fixed preserved neither the prediction task nor the calibration evidence needed for its coverage claim.

References

1. Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.
2. Nazeri, Kamyar; Ng, Eric; Joseph, Tony; Qureshi, Faisal Z.; Ebrahimi, Mehran. (2019). EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. arXiv:1901.00212. https://arxiv.org/abs/1901.00212.
3. Yu, Jiahui; Lin, Zhe; Yang, Jimei; Shen, Xiaohui; Lu, Xin; Huang, Thomas. (2018b). Free-Form Image Inpainting with Gated Convolution. arXiv:1806.03589. https://arxiv.org/abs/1806.03589.
4. Suvorov, Roman; Logacheva, Elizaveta; Mashikhin, Anton; Remizova, Anastasia; Ashukha, Arsenii; Silvestrov, Aleksei; Kong, Naejin; Goka, Harshith; Park, Kiwoong; Lempitsky, Victor. (2021). Resolution-robust Large Mask Inpainting with Fourier Convolutions. arXiv:2109.07161. https://arxiv.org/abs/2109.07161.
5. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
6. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.
7. Vincent, Pascal; Larochelle, Hugo; Bengio, Yoshua; Manzagol, Pierre-Antoine. (2008). Extracting and composing robust features with denoising autoencoders. Proceedings of the 25th international conference on Machine learning - ICML '08. https://doi.org/10.1145/1390156.1390294.
8. Alpaydin, E.; Kaynak, C. (1998). Optical Recognition of Handwritten Digits [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C50P49.
9. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
10. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. “U-Net: Convolutional Networks for Biomedical Image Segmentation.” *Lecture Notes in Computer Science*, 2015, pp. 234-241. https://doi.org/10.1007/978-3-319-24574-4_28.
11. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.
12. Kapoor, Sayash; Narayanan, Arvind. (2023). Leakage and the reproducibility crisis in machine-learning-based science. Patterns. https://doi.org/10.1016/j.patter.2023.100804.
13. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.
14. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.
15. Hoerl, Arthur E.; Kennard, Robert W.. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics. https://doi.org/10.1080/00401706.1970.10488634.
16. Liu, Guilin; Reda, Fitsum A.; Shih, Kevin J.; Wang, Ting-Chun; Tao, Andrew; Catanzaro, Bryan. (2018). Image Inpainting for Irregular Holes Using Partial Convolutions. arXiv:1804.07723. https://arxiv.org/abs/1804.07723.
17. Pathak, Deepak; Krahenbuhl, Philipp; Donahue, Jeff; Darrell, Trevor; Efros, Alexei A.. (2016). Context Encoders: Feature Learning by Inpainting. arXiv:1604.07379. https://arxiv.org/abs/1604.07379.
18. Stekhoven, Daniel J.; Bühlmann, Peter. (2012). MissForest-non-parametric missing value imputation for mixed-type data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr597.
19. Tibshirani, Ryan J.; Barber, Rina Foygel; Candès, Emmanuel J.; Ramdas, Aaditya. (2019). Conformal Prediction Under Covariate Shift. arXiv:1904.06019. https://arxiv.org/abs/1904.06019.
20. Meng, Zijie; Zeng, Yuanze; Chang, Xiang; Xu, Tianshuo; Chao, Fei; Cao, Xixin; Shang, Changjing; Shen, Qiang. (2025). Orpaint: a zero-shot inpainting model for oracle bone inscription rubbings with visual mamba block. Science China Information Sciences, 68(8), 189102. https://doi.org/10.1007/s11432-024-4493-4.
21. Shafer, Glenn; Vovk, Vladimir. (2008). A Tutorial on Conformal Prediction. Journal of Machine Learning Research,9(12),371-421. https://www.jmlr.org/papers/v9/shafer08a.html.

22. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

23. Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

24. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.

25. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.

26. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.

27. Lei, Jing; G’Sell, Max; Rinaldo, Alessandro; Tibshirani, Ryan J.; Wasserman, Larry. (2018). Distribution-Free Predictive Inference for Regression. Journal of the American Statistical Association. https://doi.org/10.1080/01621459.2017.1307116.

28. Lugmayr, Andreas; Danelljan, Martin; Romero, Andres; Yu, Fisher; Timofte, Radu; Van Gool, Luc. (2022). RePaint: Inpainting using Denoising Diffusion Probabilistic Models. arXiv:2201.09865. https://arxiv.org/abs/2201.09865.

29. Buuren, Stef van; Groothuis-Oudshoorn, Karin. (2011). mice: Multivariate Imputation by Chained Equations inR. Journal of Statistical Software. https://doi.org/10.18637/jss.v045.i03.

30. Golub, G. H.; Reinsch, C. (1970). Singular value decomposition and least squares solutions. Numerische Mathematik, 14(5), 403-420. https://doi.org/10.1007/BF02163027.