WHAT LIES BENEATH THE CURVE? SCALING LAWS IN THE PRESENCE OF EXACT POSTERIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

How close are neural networks to the best they could possibly do? Standard benchmarks cannot answer this because they lack access to the true posterior $p(y|x)$. We use class-conditional normalizing flows as oracles that make exact posteriors tractable on realistic images (AFHQ, ImageNet). This enables five lines of investigation. **Scaling laws:** Prediction error decomposes into irreducible aleatoric uncertainty and reducible epistemic error; the epistemic component follows a power law in dataset size, continuing to shrink even when total loss plateaus. **Limits of learning:** The aleatoric floor is exactly measurable, and architectures differ markedly in how they approach it: ResNets exhibit clean power-law scaling while Vision Transformers stall in low-data regimes. **Soft labels:** Oracle posteriors contain learnable structure beyond class labels: training with exact posteriors outperforms hard labels and yields near-perfect calibration. **Distribution shift:** The oracle computes exact KL divergence of controlled perturbations, revealing that shift *type* matters more than shift magnitude: class imbalance barely affects accuracy at divergence values where input noise causes catastrophic degradation. **Active learning:** Exact epistemic uncertainty distinguishes genuinely informative samples from inherently ambiguous ones, improving sample efficiency. Our framework reveals that standard metrics hide ongoing learning, mask architectural differences, and cannot diagnose the nature of distribution shift.

1 INTRODUCTION

Every paper reporting test accuracy implicitly asks: how good is this model? But good compared to what? Without access to the true posterior $p(y|x)$, we cannot tell whether a model is at 50% or 99% of the theoretical maximum. When a loss curve flattens, we cannot distinguish irreducible noise (aleatoric uncertainty) from gaps in the model's knowledge (epistemic uncertainty) (Gal & Ghahramani, 2016; Guo et al., 2017). Neural Scaling laws (Kaplan et al., 2020) tell us loss decreases as $N^{-\alpha}$, but loss conflates both Hoffmann et al. (2022). When performance degrades under distribution shift, we cannot tell whether the shift was large or small (Hendrycks & Dietterich, 2019) Gal & Ghahramani (2016).

A class-conditional normalizing flow (Dinh et al., 2017; Kingma & Dhariwal, 2018) trained on real images provides a concrete reference point. The flow defines an explicit $p_\theta(x \mid y)$ from which we can both sample and evaluate exact likelihoods. We treat this not as an approximation to nature, but as the complete specification of a synthetic world in which Bayes-optimal posteriors are computable in closed form (Dinh et al., 2017). In this world, the expected cross-entropy of any classifier decomposes exactly:

$$\mathcal{L}(q_\theta) = \underbrace{\mathbb{E}_x[H(p(y|x))]}_{\text{aleatoric (irreducible)}} + \underbrace{\mathbb{E}_x[\mathrm{KL}(p(y|x)\|q_\theta(y|x))]}_{\text{epistemic (reducible)}}$$

This decomposition, intractable on any standard benchmark, is a direct consequence of Bayes-optimal risk decomposition and is the basis for all our experiments. As illustrated in Figure 1, we train state-of-the-art normalizing flows on realistic image datasets and conduct a thorough validation to confirm the generated images match real data statistics and are not memorized (Section 2.3). We find:

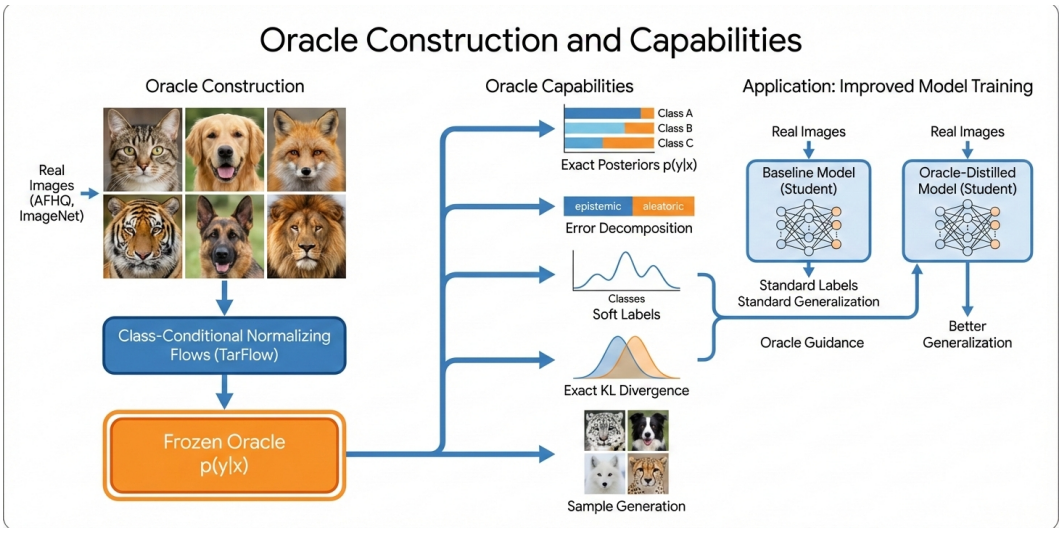

Figure 1: **Normalizing flows enable exact posterior computation on realistic images. (Left)** We train class-conditional flows on AFHQ/ImageNet, creating a frozen oracle with tractable likelihoods. **(Center)** The oracle decomposes prediction error into aleatoric (irreducible) and epistemic (reducible) components, quantities hidden in standard benchmarks. **(Right)** We apply this framework to diagnose scaling laws, quantify distribution shift with exact KL divergence, and train models with exact soft labels.

- **Scaling laws beyond the loss curve.** Epistemic error follows a power law $\mathrm{KL} \propto N^{-\alpha}$ that continues to shrink even when total loss plateaus. The exponents differ across architectures.
- **The limits of learning are exactly measurable.** The aleatoric floor is a hard limit no architecture can beat. Architectures differ markedly in how they approach it: ResNets show clean power-law scaling while Vision Transformers stall without pretraining.
- **Soft labels from exact posteriors enable better learning.** Training with $p(y|x)$ instead of argmax labels outperforms hard labels and yields near-perfect calibration Szegedy et al. (2015).
- **Distribution shift is exactly quantifiable.** The oracle computes $\mathrm{KL}[p_{\text{shifted}} \| p_{\text{baseline}}]$ exactly, revealing that shift *type* matters far more than magnitude: class imbalance barely affects accuracy at divergence values where input noise causes collapse.
- **Active learning with ground-truth uncertainty.** Exact epistemic uncertainty distinguishes genuinely informative samples from inherently ambiguous ones, improving sample efficiency over standard acquisition functions that confound aleatoric and epistemic uncertainty Gal et al. (2017).

Our oracle defines a synthetic world, not nature. But the phenomena it reveals (hidden epistemic progress, architectural scaling gaps, misleading divergence measures) are properties of the *models and metrics*, not the data source Belghazi et al. (2021).

2 The Oracle Framework

We build an oracle benchmark by training class-conditional normalizing flows that define tractable densities $p_\theta(x \mid y)$ on high-dimensional inputs Papamakarios et al. (2021). Once the flows are trained and frozen, they specify a synthetic world in which the posterior is available by Bayes' rule,

$$p(y \mid x) = \frac{p_\theta(x \mid y)\, \pi_y}{\sum_{y'} p_\theta(x \mid y')\, \pi_{y'}}. \tag{1}$$

Because likelihoods are tractable, we can compute Bayes-optimal posteriors exactly under the oracle model (up to floating-point arithmetic) via Bayes' rule (Dinh et al., 2017). This in turn makes our

loss decomposition and controlled scaling, shift, and active-learning experiments possible. Practical details on training and scaling to many classes (ImageNet) are provided in the appendix Deng et al. (2009).

2.1 Flows as Oracles

A normalizing flow defines an invertible mapping between inputs x and latent variables z (Dinh et al., 2017; Kingma & Dhariwal, 2018): the forward map f_θ sends $x \mapsto z$, while the inverse f_θ^{-1} generates samples by mapping $z \mapsto x$ for $z \sim \mathcal{N}(0, I)$. The key property is that densities are tractable Papamakarios et al. (2021):

$$\log p_\theta(x) = \log p_Z(f_\theta(x)) + \log \left|\det J_{f_\theta}(x)\right|. \tag{2}$$

We train one flow per class, yielding class-conditional densities $p_\theta(x \mid y)$, and then freeze the parameters. For any input x, we compute the oracle posterior via Bayes' rule (Eq. 1).

Unlike typical Bayesian methods that rely on sampling or variational approximations, this computation is exact *under the oracle model* Alemi et al. (2019) since the likelihoods and class prior are explicitly known: given the trained flows and the chosen class prior π, $p(y \mid x)$ is computed in closed form (implemented in log space with log-sum-exp normalization for numerical stability). The oracle therefore returns the true posterior for the synthetic world defined by the flow.

2.2 What Does "Truth" Mean Here?

Here, "true" means Bayes-optimal with respect to the oracle world induced by the trained flows. The oracle world is defined by (i) the per-class flow densities $p_\theta(x \mid y)$, (ii) the exact preprocessing pipeline used during flow training and sampling, and (iii) a class prior π_y. All "exact" statements are with respect to this specification: for the oracle world, we can compute the Bayes posterior $p(y \mid x)$ (Eq. 1) and therefore quantities such as the aleatoric term $\mathbb{E}[H(p(y \mid x))]$ and epistemic term $\mathbb{E}[\mathrm{KL}(p \,\|\, q_\theta)]$ exactly under the oracle model (Section 3).

Unless otherwise stated, we use a uniform prior over classes ($\pi_y = 1/K$). In the class-imbalance shift experiments, we modify the class prior π while keeping $p_\theta(x \mid y)$ fixed Chawla et al. (2002), isolating prior shift from conditional shift.

The main limitation is external validity: conclusions are only useful insofar as samples from the oracle world match the statistical and semantic properties of the real data used to fit the flows.

2.3 Oracle Validation

Because the oracle provides exact posteriors, we can compute the Bayes-optimal accuracy for the oracle world: 99.77% on AFHQ Choi et al. (2020) (Bayes error 0.23%). If classifiers trained on oracle data approach this bound, the oracle is producing learnable, well-structured samples. We trained five architectures on 50K oracle samples. ConvNeXt Liu et al. (2022) reaches 98.0%, ResNet He et al. (2015) 97.7%, Swin Liu et al. (2021) 97.7%, MobileNet Howard et al. (2019) 97.7%, and ViT-Base Dosovitskiy et al. (2021) 97.0% (Table 1). The remaining 1.8–2.8% gap closes with longer training. On ImageNet-64 Chrabaszcz et al. (2017), the Bayes error is 4.031%.

We validated distributional quality using six metrics. **FID** Heusel et al. (2018) measures feature distribution distance (28.44 on AFHQ, 13.48 on ImageNet-64), higher than diffusion models because flows trade sharpness for exact likelihoods. **Inception Score** Salimans et al. (2016) evaluates quality and diversity (6.24 ± 3.57 on AFHQ, 32.57 ± 4.47 on ImageNet-64). **Manifold coverage** confirms broad support (90% on AFHQ, 89.9% on ImageNet-64). **Feature variance match** shows semantic diversity is preserved (83–92% on AFHQ, 72–93% on ImageNet-64). **Memorization check**: only 36% of AFHQ samples (6.5% on ImageNet-64) have a training neighbor within feature distance 10, and visual inspection confirms these share pose/lighting but depict distinct individuals. Full metrics and protocol in Appendix B, Table 4.

2.4 Implementation

We use TarFlow (Zhai et al., 2025), a recent normalizing flow that achieves strong likelihood scores on AFHQ and ImageNet. Training follows standard practice: dequantization, logit preprocessing,

Table 1: **Classifiers approach the Bayes-optimal bound on oracle data.** Self-validation on AFHQ: all architectures reach 97–98% accuracy, within 2–3% of the theoretical optimum (99.8%).

Architecture	Accuracy (%)	Error (%)	Gap (%)
ConvNeXt	98.03	1.97	1.75
ResNet-18	97.73	2.27	2.05
Swin	97.68	2.32	2.10
MobileNet	97.65	2.35	2.12
ViT-Base	96.98	3.02	2.80

and maximum likelihood optimization. We train separate flows per class on AFHQ (3 classes: cat, dog, wild) and ImageNet Deng et al. (2009) (up to 1000 classes at 64×64 and 128×128 resolution). After training, we freeze the flows and generate 50,000 labeled samples with oracle posteriors. We compute posteriors in log space and normalize with log-sum-exp for numerical stability. Full architecture and systems details (including throughput and scaling across many classes) appear in the appendix.

3 Decomposing Prediction Error

Why does this decomposition matter? In practice, a flattening loss curve is ambiguous: it could mean the model has learned everything learnable, or it could mean the model is stuck while reducible error remains. Distinguishing these cases determines whether more data, more capacity, or a different task formulation is needed. Similarly, calibration methods aim to match model confidence to true probabilities, but without access to true posteriors, calibration can only be evaluated against empirical frequencies Guo et al. (2017). Access to the true posterior $p(y|x)$ enables a decomposition that resolves both questions. Consider a classifier $q_\theta(y|x)$ and measure its expected cross-entropy:

$$\mathcal{L}(q_\theta) = \mathbb{E}_x \left[-\sum_y p(y|x) \log q_\theta(y|x) \right] \tag{3}$$

This loss splits cleanly into two terms:

$$\mathcal{L}(q_\theta) = \underbrace{\mathbb{E}_x \left[H(p(y|x)) \right]}_{\text{aleatoric}} + \underbrace{\mathbb{E}_x \left[\mathrm{KL}(p(y|x) \| q_\theta(y|x)) \right]}_{\text{epistemic}} \tag{4}$$

The **aleatoric** term is the entropy of the true posterior averaged over inputs. It reflects irreducible uncertainty, or the ambiguity inherent in the data that no classifier can resolve. An image of a cat-like dog has high $H(p(y|x))$ regardless of model quality.

The **epistemic** term measures the gap between the model's beliefs and truth Gal & Ghahramani (2016). A perfect model achieves $\mathrm{KL} = 0$; any deviation indicates something learnable that the model has not yet captured.

On standard benchmarks, we observe only total loss. When it plateaus, we cannot tell whether the model has reached the aleatoric floor or whether epistemic error remains. Our oracle makes both terms measurable. This distinction matters: if epistemic error dominates, more data or capacity should help; if aleatoric error dominates, improvements require changing the task itself.

4 Experiments

Setup. We trained TarFlow Zhai et al. (2025) on AFHQ Choi et al. (2020) (3 classes: cat, dog, wild; ∼15K images) and ImageNet Deng et al. (2009) at 64×64 resolution. Oracle quality is validated in Section 2.3. We generated labeled samples with exact posteriors for all experiments. We trained classifiers (ResNet-50 He et al. (2015), ViT-Base Dosovitskiy et al. (2021), ConvNeXt Liu et al. (2022), Swin Liu et al. (2021), MobileNetV3 Howard et al. (2019)) using default hyperparameters on varying amounts of oracle data, from N=100 to N=10,000 (up to N=45,000 for ImageNet-64) samples. All results are averaged over 3 seeds with standard deviation shown.

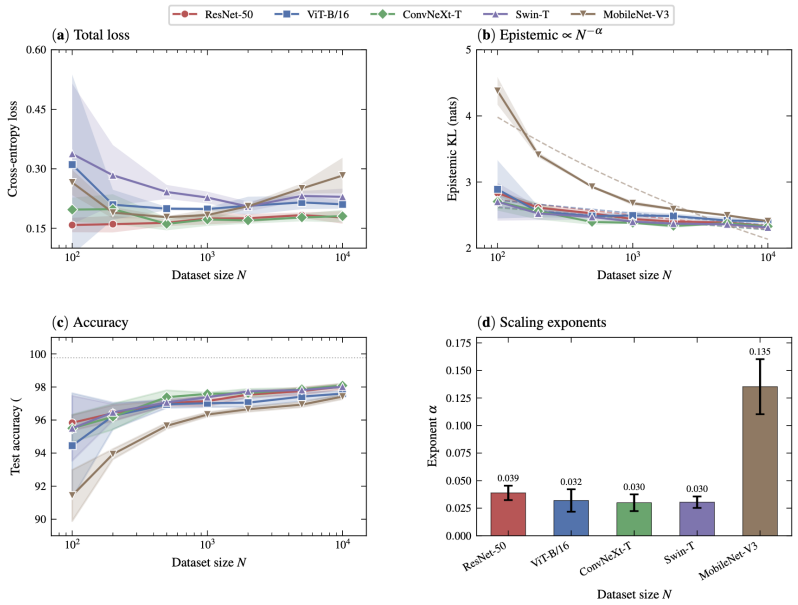

Figure 2: **Epistemic error follows a power law even when total loss plateaus. (a)** Total cross-entropy decreases with dataset size; MobileNet shows the highest loss, ResNet/ConvNeXt the lowest. **(b)** Epistemic uncertainty (KL from oracle) follows $N^{-\alpha}$; dashed lines are power-law fits. This decay continues even when total loss appears flat. **(c)** Accuracy improves from 92–96% at N=100 toward 97–98% at N=10,000. **(d)** Scaling exponents reveal architectural differences: MobileNet improves fastest (α=0.135), ViT stalls without pretraining (α=0.032).

4.1 Scaling Laws Beyond the Loss Curve

We trained classifiers on varying amounts of oracle data and decomposed prediction error into aleatoric and epistemic components Kaplan et al. (2020) (Section 3). For each model, we computed total cross-entropy loss (what standard benchmarks measure), the aleatoric component $\mathbb{E}[H(p(y|x))]$, and the epistemic component $\mathbb{E}[\mathrm{KL}(p(y|x)\|q_\theta(y|x))]$.

Figure 2 shows the result. Total cross-entropy loss decreases with dataset size on a log-log scale, consistent with Kaplan et al. (2020). MobileNet exhibits the highest loss throughout, ResNet and ConvNeXt the lowest. But the total loss curve obscures what is actually happening.

The epistemic component behaves differently. On a log-log plot, it traces a nearly perfect straight line: $\mathrm{KL} \propto N^{-\alpha}$ (dashed lines in Figure 2b) Hoffmann et al. (2022). This decay *continues even when total loss appears to plateau*, because the aleatoric floor dominates the total at large N. Standard metrics miss this ongoing learning because they conflate the two error sources.

The scaling exponents α differ across architectures (Figure 2d): ResNet-50 (α=0.039 $\pm$ 0.007), ViT-Base (0.032 $\pm$ 0.010), ConvNeXt (0.030 $\pm$ 0.008), Swin (0.030 $\pm$ 0.005), and MobileNetV3 (0.135 $\pm$ 0.025). These quantify how efficiently each architecture converts additional data into reduced uncertainty; MobileNet's much higher α indicates faster epistemic decay rate, yet it starts from a higher error floor, suggesting a less efficient initial representation. Calibration error (ECE) also decreases with N, but at a different rate than epistemic error, confirming that calibration and posterior approximation quality are distinct phenomena Guo et al. (2017).

The aleatoric floor. On AFHQ, the aleatoric uncertainty is exactly 0.012 nats (Bayes error 0.23%; see Section 2.3), with mutual information $I(X;Y) = 1.577$ bits (normalized MI = 99.5%). On ImageNet-64, the aleatoric floor is 0.10 nats (Bayes error 4.031%), higher due to the greater number of classes and inherent inter-class ambiguity.

ImageNet replication. Figure 6 shows scaling laws on ImageNet-64 up to 45K samples, a substantially more challenging dataset than AFHQ due to its fine-grained structure and 1000-way label space. Despite this higher intrinsic entropy, the total cross-entropy loss decreases rapidly with dataset size, at a faster absolute rate than in AFHQ, reinforcing the classical finding that larger datasets continue to yield measurable gains even when the absolute loss remains dominated by intrinsic class complexity.

The aleatoric component remains effectively constant at approximately 0.10 nats across all dataset sizes, confirming that all models converge to the same irreducible uncertainty induced by the oracle conditional distribution. As in AFHQ, this floor is architecture-independent and invariant to dataset size, reflecting the Bayes error of the oracle world rather than model capacity. The higher value relative to AFHQ (0.012 nats) is consistent with ImageNet-64's greater inter-class overlap and multimodality.

Epistemic uncertainty decays with dataset size and corroborates the power-law behaviour observed on AFHQ. ImageNet-64, however, exhibits a brief plateau in the mid-range (2,000-10,000 samples), reflecting a representation-transition regime in which models do not immediately convert additional data into reduced epistemic uncertainty. Beyond this regime, all architectures resume clear power-law decay, reinforcing that epistemic uncertainty, alongside total loss, is a distinct and reliable indicator of continued learning at scale.

The resulting exponents show both similarities and divergences relative to AFHQ. ResNet-50 exhibits the highest scaling rate (α=0.019 ± 0.006), closely followed by ViT-Base (0.018 ± 0.004), mirroring their behaviour on AFHQ. ConvNeXt achieves only a weak decay rate (0.003 ± 0.003, lower than ResNet and ViT-Base as in AFHQ), indicating limited epistemic improvement per additional sample, likely due to architectural biases that mismatch the ImageNet-64 feature geometry. Swin-T displays the strongest positive exponent (0.027 ± 0.01), consistent with hierarchical transformers benefiting disproportionately from larger, more diverse datasets.

MobileNet-V3 is an exception: its epistemic curve exhibits a pronounced non-monotonicity. We attribute this to MobileNet's compressed architecture and depthwise-separable convolutions, which yield an inefficient initial representation incapable of faithfully capturing ImageNet-64's high inter-class diversity. At small dataset sizes, the model overfits and becomes overconfident, producing an artificially low epistemic floor. As the dataset expands into the mid-scale regime (≈2k–10k samples), this brittle representation is forced to reorganize, triggering a spike in loss, calibration error, and epistemic uncertainty. Once past this regime, MobileNet resumes clean power-law decay, confirming that the underlying epistemic scaling law still holds once an adequate representation is established.

Together with the monotonic rise in test accuracy, these trends show that accuracy, calibration, and epistemic uncertainty respond differently to data scaling, with calibration showing partial signatures of the behaviour made explicit by epistemic uncertainty.

Architectures differ in how they approach optimality. Table 2 shows that ResNet-18 exhibits clean power-law scaling: epistemic error drops from 0.16 to 0.026 as data increases from N=100 to N=40,000 (a 6× reduction). ViT-Base behaves differently: it starts competitive at N=100 but then stalls. By N=40,000, ViT has improved by less than 1.2×. Both architectures achieve similar total loss at large N; the difference is in *how* they approach optimality. ViTs, lacking pretraining, fail to extract geometric structure from small samples Dosovitskiy et al. (2021). This is consistent with Dosovitskiy et al. (2021), but our framework quantifies the gap in information-theoretic terms.

4.2 Soft Labels Enable Better Learning

Our oracle can generate exact posterior distributions $p(y|x)$ as training labels, not just the argmax class. Training with these soft labels outperforms hard-label training at 4 out of 5 dataset sizes, with accuracy gains up to ∼1%. Models trained on soft labels also achieve near-perfect calibration (ECE = 0.018) Guo et al. (2017), confirming the posteriors encode learnable structure beyond class labels Szegedy et al. (2015); Hinton et al. (2015). Full results in Appendix D.

Table 2: **ResNets reduce epistemic error $6\times$; ViTs stall without pretraining.** Epistemic KL (nats) vs. dataset size on AFHQ. ResNet exhibits clean power-law scaling from 0.16 to 0.026; ViT improves by only $1.2\times$ over the same range.

Dataset size N	ResNet-18	ViT-B-16
100	0.160	0.131
1,000	0.090	0.141
5,000	0.058	0.099
10,000	0.033	0.117
40,000	0.026	0.117

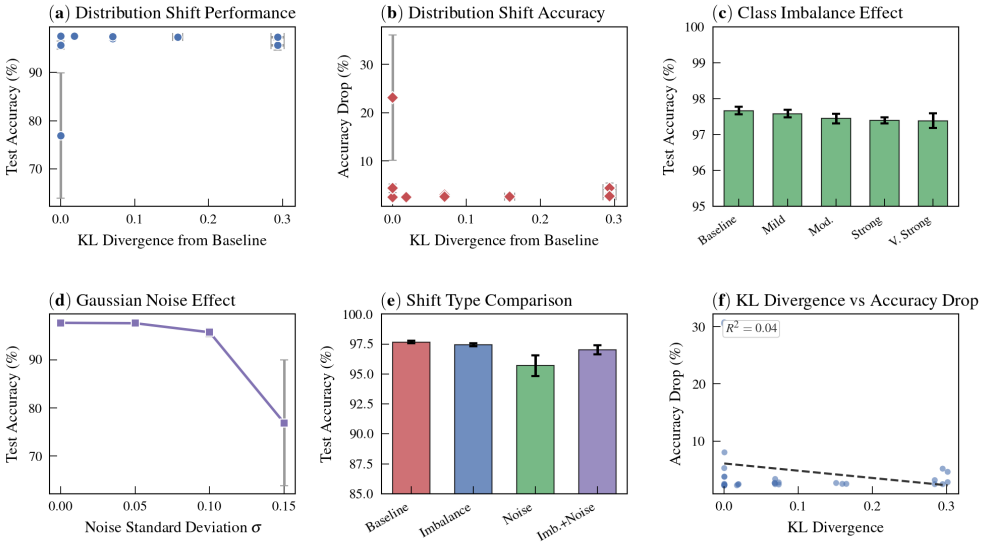

Figure 3: **What shifts matters more than how much: KL magnitude alone poorly predicts performance (R^2=0.04). (a)** Test accuracy vs. exact KL divergence: class imbalance (blue) maintains $\sim$97% accuracy; Gaussian noise (low points) collapses to $\sim$77% at comparable KL. **(b)** Accuracy drop vs. KL: imbalance (red diamonds) clusters near zero regardless of divergence magnitude. **(c)** Prior shift alone barely affects accuracy (97.4–97.7% across all imbalance levels). **(d)** Covariate shift causes exponential degradation (σ=0.15 $\rightarrow$ 77%). **(e)** Aggregated comparison confirms noise dominates. **(f)** Linear regression: $R^2 = 0.04$ shows aggregate KL is uninformative.

4.3 Distribution Shift: Exact Quantification via Oracle

On standard benchmarks, distribution shift is observed but never measured: we see accuracy drop but cannot quantify how much the distribution actually changed. Our oracle computes $\mathrm{KL}[p_{\text{shifted}} \| p_{\text{baseline}}]$ exactly by evaluating the true log-likelihood $\log p(x|y)$ under both distributions. This enables controlled experiments that are impossible on real benchmarks: we introduce perturbations of known type and magnitude, measure the exact divergence, and observe how models degrade.

We introduce two types of controlled perturbation to the oracle distribution:

- **Class imbalance:** Shifting the class prior from uniform to skewed ratios (e.g., 40/35/25 up to 70/20/10), which increases KL divergence through the marginal $p(y)$ Chawla et al. (2002).
- **Gaussian noise:** Adding noise with standard deviation $\sigma \in \{0.05, 0.10, 0.15\}$ to generated images, which increases KL divergence through the conditional $p(x|y)$.

The results (Figure 3) show a clear asymmetry. Class imbalance produces KL divergences from 0.018 to 0.293, yet accuracy remains flat at 97.4–97.7%. Gaussian noise at σ=0.15 produces a

comparable KL divergence but causes accuracy to collapse to $\sim$77%, a 20-percentage-point drop. The same KL magnitude leads to dramatically different outcomes depending on *where* the shift occurs.

Class imbalance shifts the marginal $p(y)$ but leaves the class-conditional $p(x|y)$ intact, so learned features remain discriminative. Gaussian noise corrupts $p(x|y)$ directly, destroying the features that classifiers depend on Hendrycks & Dietterich (2019). A linear fit of KL divergence against accuracy drop yields $R^2 = 0.042$ (Figure 3f): aggregate divergence does not predict robustness. What shifted matters more than how much.

4.4 Active Learning with Ground-Truth Uncertainty

Active learning aims to select the most informative training samples Sener & Savarese (2018). On standard benchmarks, informativeness must be estimated via heuristics (e.g., entropy of model predictions) Lewis & Gale (1994); Gal et al. (2017). Our oracle provides exact epistemic uncertainty, enabling a critical distinction: separating samples that are *epistemically* uncertain (informative) from those that are *aleatorically* uncertain (inherently ambiguous). On standard benchmarks, a high-entropy sample could be either; the oracle tells us which.

We compare three acquisition strategies on ImageNet-64: Random, Max Entropy, and Max Epistemic (selecting by exact epistemic uncertainty). Max Epistemic yields a 1.6$\times$ larger accuracy gain than Max Entropy under the same budget and needs 47.8% fewer labels to match mid-regime performance, consistently beating both baselines. This supports that predictive entropy is a weak proxy for information gain under substantial aleatoric noise. Full results and learning curves are in Appendix (Figure 9).

5 Discussion

Implications for practice. Validation loss can be misleading: epistemic error may keep shrinking even after loss plateaus, so early stopping can leave accuracy untapped. Prefer tracking epistemic proxies (e.g., ensemble disagreement), noting that in practice these must be estimated Lakshminarayanan et al. (2017); Gal & Ghahramani (2016).

Limitations. Our oracle is a synthetic world, not nature: transfer to real images depends on how well the flow captures their statistics. Key caveats include (i) a domain gap between flow samples and photographs, (ii) limited scale (AFHQ has 3 classes; full ImageNet runs are ongoing), and (iii) possible sensitivity to hyperparameters. We view this as complementary to standard benchmarks: it offers ground truth in a controlled setting, not claims about nature's Bayes error.

6 Conclusion

We introduce a framework that treats normalizing flows as *oracles* rather than models, enabling exact computation of quantities that standard benchmarks can only estimate. Knowing the data-generating process lets us separate what classifiers learn from what remains fundamentally uncertain.

Empirically, **epistemic error follows a clean power law** ($\mathrm{KL} \propto N^{-\alpha}$) even when total loss plateaus: models keep improving in ways standard metrics miss. This decay is consistent across architectures and scales to ImageNet (1000 classes). The decomposition also exposes clear architectural gaps: ResNets reduce epistemic error by $\sim 6\times$ over the same data range where ViTs improve by only $\sim 1.2\times$, highlighting the value of convolutional inductive bias in low-data regimes.

Under distribution shift, **what shifts matters more than how much**: class imbalance has little effect at KL levels where input noise causes $\sim$20-point accuracy drops. This suggests robustness evaluations should characterize shift *type*, not just magnitude.

Our oracle framework complements real-world benchmarks by providing ground truth in a controlled setting, informing architecture choice, training decisions, and robustness evaluation beyond aggregate loss curves.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep Variational Information Bottleneck, October 2019. URL http://arxiv.org/abs/1612.00410. arXiv:1612.00410 [cs].

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and R. Devon Hjelm. MINE: Mutual Information Neural Estimation, August 2021. URL http://arxiv.org/abs/1801.04062. arXiv:1801.04062 [cs].

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16:321–357, June 2002. ISSN 1076-9757. doi: 10.1613/jair.953. URL http://arxiv.org/abs/1106.1813. arXiv:1106.1813 [cs].

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains, April 2020. URL http://arxiv.org/abs/1912.01865. arXiv:1912.01865 [cs].

Patryk Chrabaszcz, Ilya Loshchilov, and Frank Hutter. A Downsampled Variant of ImageNet as an Alternative to the CIFAR datasets, August 2017. URL http://arxiv.org/abs/1707.08819. arXiv:1707.08819 [cs].

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848. URL https://ieeexplore.ieee.org/document/5206848. ISSN: 1063-6919.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using Real NVP, February 2017. URL http://arxiv.org/abs/1605.08803. arXiv:1605.08803 [cs].

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. URL http://arxiv.org/abs/2010.11929. arXiv:2010.11929 [cs].

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*, pp. 1050–1059. PMLR, June 2016. URL https://proceedings.mlr.press/v48/gal16.html.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep Bayesian Active Learning with Image Data, March 2017. URL http://arxiv.org/abs/1703.02910. arXiv:1703.02910 [cs].

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks, August 2017. URL http://arxiv.org/abs/1706.04599. arXiv:1706.04599 [cs].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Dan Hendrycks and Thomas Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations, March 2019. URL http://arxiv.org/abs/1903.12261. arXiv:1903.12261 [cs].

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium, January 2018. URL http://arxiv.org/abs/1706.08500. arXiv:1706.08500 [cs].

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat].

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training Compute-Optimal Large Language Models, March 2022. URL http://arxiv.org/abs/2203.15556. arXiv:2203.15556 [cs].

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, Quoc V. Le, and Hartwig Adam. Searching for MobileNetV3, November 2019. URL http://arxiv.org/abs/1905.02244. arXiv:1905.02244 [cs].

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling Laws for Neural Language Models, January 2020. URL http://arxiv.org/abs/2001.08361. arXiv:2001.08361 [cs].

Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative Flow with Invertible 1x1 Convolutions, July 2018. URL http://arxiv.org/abs/1807.03039. arXiv:1807.03039 [stat].

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. BatchBALD: Efficient and Diverse Batch Acquisition for Deep Bayesian Active Learning, October 2019. URL http://arxiv.org/abs/1906.08158. arXiv:1906.08158 [cs].

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles, November 2017. URL http://arxiv.org/abs/1612.01474. arXiv:1612.01474 [stat].

David D. Lewis and William A. Gale. A Sequential Algorithm for Training Text Classifiers, July 1994. URL http://arxiv.org/abs/cmp-lg/9407020. arXiv:cmp-lg/9407020.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows, August 2021. URL http://arxiv.org/abs/2103.14030. arXiv:2103.14030 [cs].

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A ConvNet for the 2020s, March 2022. URL http://arxiv.org/abs/2201.03545. arXiv:2201.03545 [cs].

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing Flows for Probabilistic Modeling and Inference, April 2021. URL http://arxiv.org/abs/1912.02762. arXiv:1912.02762 [stat].

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs, June 2016. URL http://arxiv.org/abs/1606.03498. arXiv:1606.03498 [cs].

Ozan Sener and Silvio Savarese. Active Learning for Convolutional Neural Networks: A Core-Set Approach, June 2018. URL http://arxiv.org/abs/1708.00489. arXiv:1708.00489 [stat].

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision, December 2015. URL http://arxiv.org/abs/1512.00567. arXiv:1512.00567 [cs].

Shuangfei Zhai, Ruixiang Zhang, Preetum Nakkiran, David Berthelot, Jiatao Gu, Huangjie Zheng, Tianrong Chen, Miguel Angel Bautista, Navdeep Jaitly, and Josh Susskind. Normalizing Flows are Capable Generative Models, June 2025. URL http://arxiv.org/abs/2412.06329. arXiv:2412.06329 [cs].

Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling Vision Transformers, June 2022. URL http://arxiv.org/abs/2106.04560. arXiv:2106.04560 [cs].

A Related Work

Neural scaling laws. Kaplan et al. (2020) established that loss decreases as a power law in data, compute, and parameters. Subsequent work refined these laws for vision (Zhai et al., 2022) and explored their limits. All such studies measure total loss, conflating aleatoric and epistemic error. Our decomposition reveals that epistemic error alone follows a power law, even when total loss appears flat.

Uncertainty estimation. Bayesian deep learning and ensemble methods aim to quantify epistemic uncertainty, but without ground truth, evaluation is indirect (Alemi et al., 2019). Calibration metrics compare to empirical frequencies, not true posteriors (Guo et al., 2017). Our framework provides the missing ground truth, enabling evaluation of both calibration and posterior quality.

Knowledge distillation and soft labels. Hinton et al. (2015) showed that training on a teacher's soft predictions transfers "dark knowledge" beyond hard labels. Our soft-label experiments connect to this literature but with a key difference: our soft labels are *exact* posteriors, not approximations from a teacher model. This provides an upper bound on what soft-label training can achieve.

Distribution shift. Benchmarks like ImageNet-C (Hendrycks & Dietterich, 2019) evaluate robustness to distribution shift, but the magnitude of shift is unknown. Our oracle enables controlled experiments with exact divergence, revealing that shift type matters more than magnitude.

Active learning. Uncertainty-based acquisition functions Gal et al. (2017); Kirsch et al. (2019) select samples by estimated informativeness. Our oracle separates epistemic from aleatoric uncertainty exactly, providing an upper bound on what uncertainty-based active learning can achieve.

Normalizing flows. Flows have matured from RealNVP (Dinh et al., 2017) and Glow (Kingma & Dhariwal, 2018) to TarFlow (Zhai et al., 2025), which achieves strong likelihoods on ImageNet. We use flows differently: not as models to evaluate, but as oracles that define evaluation itself.

Synthetic benchmarks. Researchers have long used toy distributions (Gaussians, moons, spirals) with known posteriors. These lack realism. Our approach combines realistic images from AFHQ and ImageNet with the exact posterior access of synthetic benchmarks.

Dataset	FID↓	IS↑	Coverage↑	Mem. rate	NN dist. (med)	Post. stab. KL↓
AFHQ-256	28.44	6.24 ± 3.57	0.900	0.363	11.01	1.51×10^{-9}

Table 3: **Oracle passes all validation checks: high coverage, low memorization, stable posteriors.** FID measures distributional quality; coverage confirms broad manifold support (90%); memorization rate shows most samples are novel; posterior stability KL (10^{-9}) confirms robustness to small perturbations.

Table 4: **Oracle samples match real data statistics across multiple metrics.** FID, Inception Score, and manifold coverage confirm distributional quality; low memorization rates verify sample novelty. Dashes indicate metrics not yet computed.

Metric	AFHQ	ImageNet-64
FID ↓	28.44	13.48
Inception Score ↑	6.24 ± 3.57	32.57 ± 4.47
Manifold coverage	90%	89.9%
Feature variance match	83–92%	72–93%
Texture variance ratio	56–69%	—
Feature variance ratio match	—	71.6–93.3%
Memorization rate	36%	6.5%

B Oracle Validation Details

We validated the oracle along five axes on both AFHQ and ImageNet-64. This appendix provides full details for the summary in Section 2. We report quantitative checks verifying our Oracle provides (i) high-quality samples, (ii) broad support over the data manifold, and (iii) stable posteriors for uncertainty decomposition.

B.1 Distribution Quality (FID)

We measured Fréchet Inception Distance (FID) (Heusel et al., 2018), which compares the mean and covariance of Inception-v3 features between real and generated images. Our AFHQ oracle achieves FID 28.44 and ImageNet-64 achieves FID 13.48, comparable to state-of-the-art generative models on these datasets.

B.2 Manifold Coverage

A generative model might produce high-quality samples that only capture part of the true distribution (e.g., generating realistic dogs but missing rare breeds). We computed manifold coverage by embedding both real and generated samples in a pretrained feature space and measuring what fraction of real samples have a nearby generated sample within a distance threshold. In our implementation, distances are computed in ResNet feature space using k-nearest neighbors and an adaptive threshold (set to the 90th percentile of real-to-synthetic distances). Our oracle achieves 90% coverage on AFHQ and 89.9% on ImageNet, suggesting it captures broad distributional structure rather than collapsing to a few high-density modes.

B.3 Diversity (Inception Score)

We evaluated diversity using Inception Score (IS), which measures the KL divergence between the conditional label distribution $p(y|x)$ and the marginal $p(y)$ averaged over generated samples. Higher scores indicate both confident classifications and diverse outputs. We achieve 6.24 ± 3.57 on AFHQ and 32.57 ± 4.47 on ImageNet. Per-class diversity ratios (variance ratio and distance ratio between synthetic and real) range from 0.75 to 0.96 on AFHQ, confirming reasonable diversity that is slightly below real data. Across the three classes, the synthetic-to-real feature variance ratios are $\{0.746, 0.854, 0.921\}$ and the average pairwise distance ratios are $\{0.891, 0.944, 0.965\}$, indicating slightly reduced diversity relative to real data, but without severe collapse

B.4 Semantic Feature Alignment

Pixel-level metrics do not capture whether images have the right semantic content. We extracted features from a pretrained ResNet, which encode high-level structure like shape and pose, and compared statistics between real and generated images. Feature variance matches at 83–92% across classes on AFHQ, with Class 0 (cat) performing best (92%) and Classes 1–2 (dog, wild) showing larger distributional separation but still acceptable overlap. ImageNet-64 results show 72–93% feature variance match across classes.

B.5 Memorization Check

If the flow memorizes training images, our benchmark would be meaningless. We computed nearest-neighbor distances in Inception feature space between each generated image and the training set. 36% of generated samples on AFHQ have a training neighbor within distance 10 in feature space; 6.53% on ImageNet. This thresholded "near-neighbor" rate is a conservative proxy for potential memorization rather than direct evidence of exact duplication. Visual inspection confirms that "close" pairs share high-level attributes (pose, lighting) but depict different individuals. The nearest-neighbor distance distribution is well-separated from zero, with the memorization threshold clearly above the bulk of the distribution; (See Figure 4).

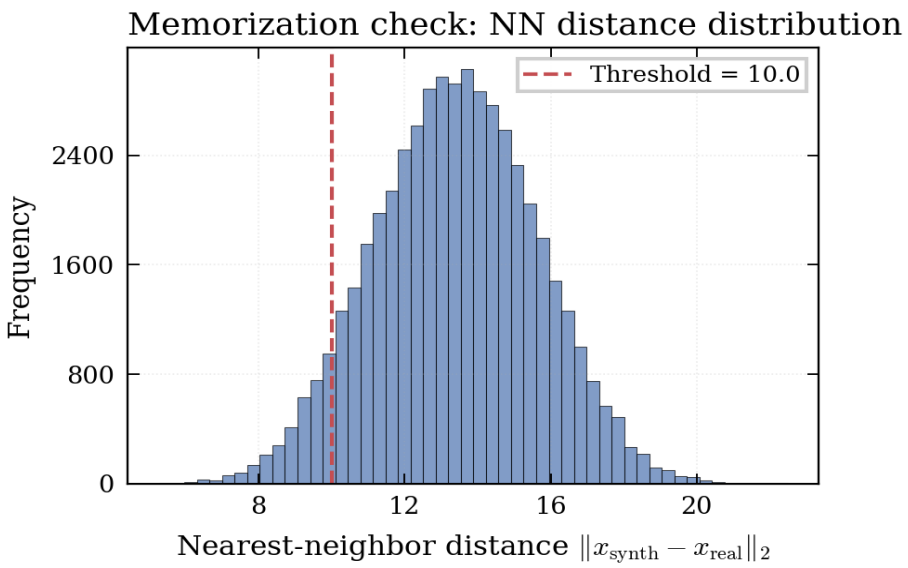

Figure 4: **Oracle samples are not memorized from training data.** Distribution of nearest-neighbor distances (ResNet feature space) between generated and training images. The dashed line marks the memorization threshold (d=10); the bulk of distances lie well above, confirming sample novelty.

B.6 Texture Analysis

Generated images are slightly smoother than real photographs: pixel-level texture variance is 56–69% of real images across all classes. This is a known property of flow-based models, which tend to soften fine-grained details while preserving global structure. For our purposes, this is acceptable because our downstream classifiers primarily rely on semantic cues (shape, pose, object parts) rather than fine-grained pixel texture. Since this smoothing effect is systematic across the oracle-generated dataset, comparative analyses across architectures remain valid.

Hyperparameter	**AFHQ** 256^2	**ImageNet** 64^2	**ImageNet** 128^2
Conditioning	Class-conditional	Class-conditional	Class-conditional
Image size	256	64	128
Channels (C)	768	768	1024
Patch size	8	2	4
# Transformer blocks	8	8	8
Flow layers / block	8	8	8
Gaussian noise std (σ)	0.07	0.05	0.15
Learning rate	1×10^{-4}	1×10^{-4}	1×10^{-4}
Batch size	256	256	768
Epochs	4000	200	320
Label dropout (p)	0.1	0.1	0.1
CFG during training	0 (disabled)	0 (disabled)	0 (disabled)

Table 5: **TarFlow hyperparameters for reproducibility.** All models use 8 Transformer blocks with 8 flow layers each. Gaussian noise refers to dequantization noise during training.

C TarFlow Architecture

We train TarFlow models from scratch on AFHQ at 256×256 and class-conditional ImageNet at 64×64 and 128×128, following the architecture introduced by Zhai et al. (2025). All configurations use an autoregressive Transformer backbone with 8 blocks and 8 flow layers per block. For all datasets, we use the same optimizer settings (learning rate 10^{-4}) and label-dropout of 0.1 for class-conditional training. Input dequantization is performed by adding Gaussian noise with dataset-specific standard deviation σ (Table 5). We compute and cache dataset statistics required for FID evaluation using the corresponding ground-truth training distribution at each resolution prior to sampling and evaluation.

Training and evaluation protocol. We train using distributed data-parallelism with dataset-dependent GPU counts (8 GPUs for AFHQ 256^2 and ImageNet 64^2; 32 GPUs for ImageNet 128^2). During training, we periodically generate samples and evaluate FID using cached ground-truth statistics computed at the matching resolution. For conditional generation at evaluation time, we additionally report classifier-free guidance (CFG) results by sampling with a nonzero guidance scale (e.g., $\gamma = 2.3$ for ImageNet 64^2), while keeping guidance disabled during training. We emphasize that all reported ImageNet and AFHQ models are trained from scratch using the above protocol.

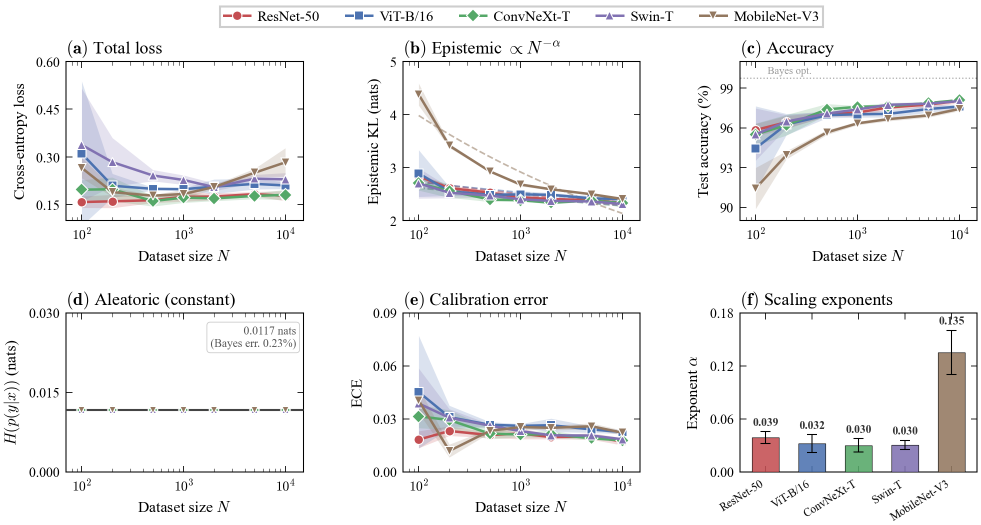

Figure 5: **Full scaling results on AFHQ.** Epistemic uncertainty follows power-law decay across all architectures, with MobileNet showing the steepest decline and ViT the shallowest.

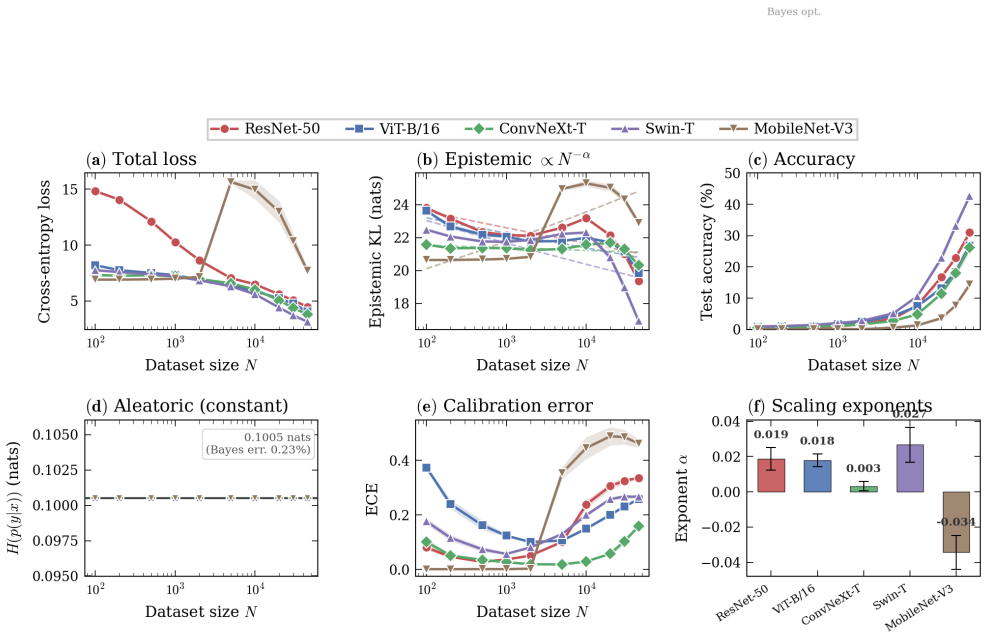

Figure 6: **Scaling laws extend to ImageNet-64 with 1000 classes.** Power-law epistemic decay holds at scale, though MobileNet shows a mid-range transition regime before resuming clean scaling.

D Soft Labels: Full Results

Our oracle can generate exact posterior distributions $p(y|x)$ as training labels, not just the argmax class. We compare training with *hard labels* (one-hot from the most likely class) versus *soft labels* (the full posterior vector) across dataset sizes from N=100 to N=5,000.

Figure 7 shows that training with oracle soft labels outperforms hard-label training at 4 out of 5 dataset sizes, with accuracy gains up to $\sim$1%. The one exception is N=500, where hard labels slightly outperform.

This result validates oracle quality from a different angle: the soft posteriors contain learnable information *beyond* the class label. If the oracle were merely assigning noisy labels, soft training would not consistently outperform. Instead, the posteriors encode genuine uncertainty structure (inter-class similarities, ambiguous regions, confidence gradients) that classifiers can exploit. This form of supervision is unavailable on any standard benchmark, where labels are always hard.

Models trained on soft labels also achieve near-perfect calibration (ECE $= 0.018$), learning the full uncertainty landscape rather than just decision boundaries.

E Computing Posteriors at Scale

The bijective invertability of normalizing flow models allow us to independently calculate samples in parallel. Posteriors are obtained by iteratively cycling through smaller batches of classes per batch of image across for all possible classes. Since the results can quickly accumulate, we store the outputs of the image batch until a given size, then save the combined output as a shard for each GPU. For ImageNet64, this process has been split across 7 V100 GPUs for a total of 4,200 samples. And for ImageNet128, this process has been split across 4 H200 GPUs for a total of 16,800 samples. For the Imagenet datasets, we iteratively accumulate 50 class batches per image batch which we find to be a good balance for memory efficiency.

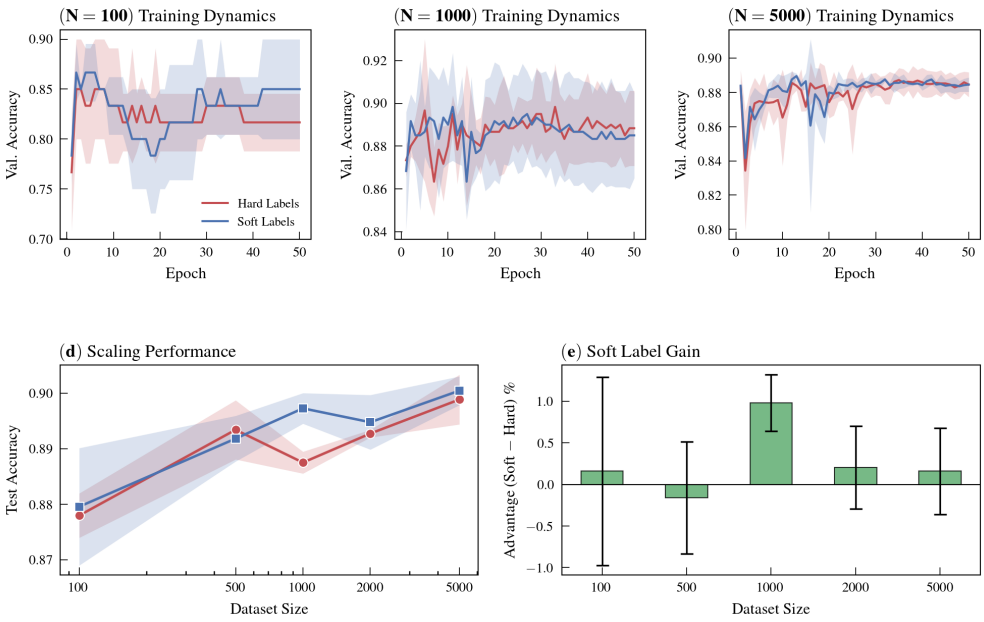

Figure 7: **Exact soft labels outperform hard labels at most dataset sizes. Top:** Training curves show soft labels (blue) tracking above hard labels (red) across three dataset sizes. **Bottom left:** Final accuracy vs. N; soft labels win at 4 of 5 sizes. **Bottom right:** Accuracy gain reaches $\sim$1%, confirming oracle posteriors encode learnable structure beyond class labels.

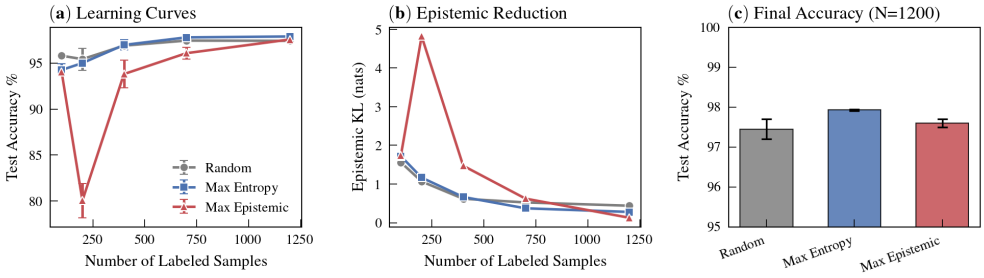

Figure 8: **Epistemic-based acquisition reduces uncertainty fastest on AFHQ. (a)** Learning curves: Max Epistemic (red) initially dips while selecting challenging samples, then recovers. **(b)** Epistemic KL reduction: Max Epistemic achieves the steepest decline, confirming it targets genuinely informative samples. **(c)** Final accuracy: all methods converge to $\sim$97–98% on this 3-class dataset.

One notable difference between the models is that OracleFlow using the AFHQ 256x256 model uses a smoothing temperature of 500 across 3 classes, compared to ImageNet 64x64 and ImageNet 128x128 which uses a smoothing temperature of 1 due to ImageNet's much higher class size of 1000.

F Distribution Shift Experimental Details

We are providing the full specification of *controlled distribution-shift* protocol used to stress-test supervised learners trained on oracle-generated data. We generate *training-time* shifts via two orthogonal knobs: (i) **label-prior shift** (class imbalance) and (ii) **covariate shift** via additive Gaussian

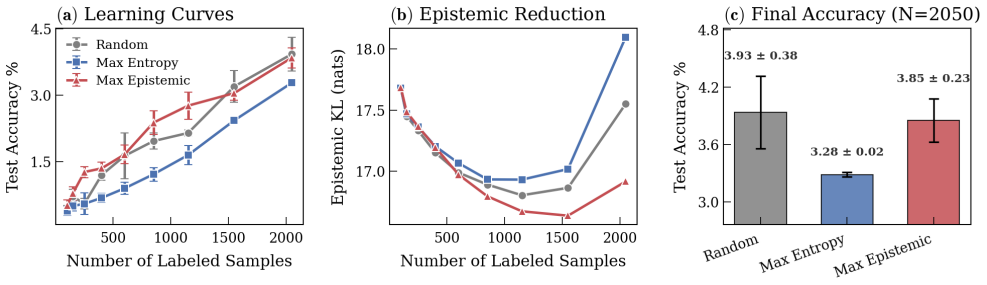

Figure 9: **Max Epistemic consistently outperforms entropy-based selection on ImageNet-64.** Selecting by exact epistemic uncertainty (orange) beats both random (gray) and max entropy (blue), confirming that predictive entropy conflates aleatoric and epistemic components while our decomposition isolates the informative signal.

noise. All models are evaluated on the *same* held-out *baseline* test set to isolate the impact of training distribution mismatch.

F.1 Base dataset and notation

Let $\mathcal{D}_0 = \{(x_i, y_i)\}_{i=1}^{N_{\text{pool}}}$ denote a fixed pool of labeled images stored in the oracle raw range $x \in [-1, 1]$ with $K = 3$ classes. We also fix a baseline test set $\mathcal{D}_0^{\text{test}}$ (size $N_{\text{test}} = 2000$) sampled once from the same source and used for *all* evaluations.

Baseline label distribution. We define the baseline label prior as uniform:

$$P_0(y = k) = \tfrac{1}{K} \qquad (K = 3). \tag{5}$$

F.2 Shift family: label-prior shift and noise perturbations

Each shifted training distribution is parameterized by $(\boldsymbol{\pi}, \sigma)$, where $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$ specifies the **target class prior** and $\sigma \geq 0$ controls **Gaussian covariate noise**.

(A) Label-prior (class-imbalance) shift. We sample class counts via multinomial draw:

$$(n_0, n_1, n_2) \sim \text{Multinomial}(N_{\text{train}}, \boldsymbol{\pi}), \tag{6}$$

then construct the shifted training set by sampling n_k examples from the pool restricted to class k. If n_k exceeds the available pool size for class k, we sample *with replacement* (this matches the implementation).

(B) Covariate shift via additive Gaussian noise. Given a sampled image $x \in [-1, 1]$, we generate a noised view:

$$x' = \text{clip}\,(x + \varepsilon,\, -1,\, 1), \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I). \tag{7}$$

Noise is applied *after* selecting examples according to $\boldsymbol{\pi}$ and *before* normalization.

Normalization for training. Models are trained using ImageNet-style normalization. Concretely, we map $x' \in [-1, 1]$ to $[0, 1]$ via $(x' + 1)/2$ and then apply per-channel mean/std normalization.

F.3 KL computation methodology

We report a **label-marginal KL** that quantifies the strength of the *prior shift*:

$$\text{KL}_y(P_{\text{shift}}(y) \,\|\, P_0(y)) = \sum_{k=1}^{K} \hat{\pi}_k \log \frac{\hat{\pi}_k}{1/K}, \tag{8}$$

where $\hat{\pi}_k$ is the *empirical* class frequency in the constructed shifted training set (i.e., computed from the realized multinomial sample and any resampling-with-replacement), and $\log$ is the natural logarithm (units: nats).

Table 6: **Distribution-shift training protocol.** All experiments use identical hyperparameters; only the training distribution varies.

Component	Setting
Base architecture	ResNet-50 pretrained on ImageNet-1K
Classifier head	Replace final FC with $K = 3$ outputs
Training size	$N_{\text{train}} = 5000$ (shifted)
Validation split	80/20 split of the shifted training set
Test set	Baseline test set $\mathcal{D}_0^{\text{test}}$, $N_{\text{test}} = 2000$
Optimizer	AdamW (weight decay 0.01)
Learning rate	10^{-4}
Schedule	Cosine annealing over 40 epochs
Batch size	32
Seeds	$S = 3$
Metric	Test accuracy on $\mathcal{D}_0^{\text{test}}$; error rate $= 1 -$ acc

Table 7: **Controlled shift configurations with exact KL values.** We vary class prior $\boldsymbol{\pi}$ (label shift) and noise σ (covariate shift) independently. Empirical KL matches target values closely.

Configuration	$\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$	σ	KL_y (target)	KL_y (empirical)
Balanced (Baseline)	(0.33, 0.33, 0.34)	0.00	1.00×10^{-4}	$1.27\times10^{-4} \pm 1.56\times10^{-4}$
Mild Imbalance (40/35/25)	(0.40, 0.35, 0.25)	0.00	0.017	0.018 ± 0.001
Moderate Imbalance (50/30/20)	(0.50, 0.30, 0.20)	0.00	0.069	0.070 ± 0.003
Strong Imbalance (60/25/15)	(0.60, 0.25, 0.15)	0.00	0.161	0.158 ± 0.007
Very Strong Imbalance (70/20/10)	(0.70, 0.20, 0.10)	0.00	0.296	0.293 ± 0.009
Balanced + Noise σ=0.05	(0.33, 0.33, 0.34)	0.05	1.00×10^{-4}	$1.27\times10^{-4} \pm 1.56\times10^{-4}$
Balanced + Noise σ=0.10	(0.33, 0.33, 0.34)	0.10	1.00×10^{-4}	$1.27\times10^{-4} \pm 1.56\times10^{-4}$
Balanced + Noise σ=0.15	(0.33, 0.33, 0.34)	0.15	1.00×10^{-4}	$1.27\times10^{-4} \pm 1.56\times10^{-4}$
Imbalance (50/30/20) + Noise σ=0.05	(0.50, 0.30, 0.20)	0.05	0.069	0.070 ± 0.003
Imbalance (70/20/10) + Noise σ=0.10	(0.70, 0.20, 0.10)	0.10	0.296	0.293 ± 0.009

Important implication. KL_y is *insensitive* to pure covariate shifts induced by σ when class priors remain balanced. Therefore, we **always report both** (σ, KL_y): KL_y measures label shift strength, while σ measures the covariate-noise shift strength.

F.4 Per-experiment protocol (fixed across all shift settings)

Each shift setting $(\boldsymbol{\pi}, \sigma)$ defines one training distribution. We repeat each experiment over $S = 3$ random seeds (affecting multinomial draw, sampling, and optimization).

F.5 Shift configurations and measured KL_y

Table 7 enumerates all perturbations. We report the *target* KL_y implied by $\boldsymbol{\pi}$ (using Eq. 8 with $\hat{\pi} = \pi$), and the *empirical* KL_y computed from realized label frequencies (mean $\pm$ std over seeds).

F.6 Results: performance degradation under training-time shift

Table 8 reports accuracy on the *baseline* test set $\mathcal{D}_0^{\text{test}}$ for each perturbation, averaged over three seeds. We also report ΔTestAcc (percentage points, pp) relative to the balanced baseline, and the corresponding test error rate.

Interpretation (high-level). Across the tested range, *label-prior shift alone* (up to $\mathrm{KL}_y \approx 0.29$ nats) causes only minor changes in baseline test accuracy (sub-0.3pp on average). In contrast, *covariate noise* drives substantially larger degradation, with $\sigma = 0.10$ inducing $\approx$2pp drops and $\sigma = 0.15$ leading to severe and high-variance failures. This separation is expected because KL_y captures label shift only, while σ controls covariate shift strength.

Table 8: **Noise causes 20-point accuracy drops; imbalance causes <0.3 points.** All models evaluated on the same baseline test set. Label-prior shift (up to KL=0.29) barely affects accuracy; covariate noise (σ=0.15) causes catastrophic degradation.

Configuration	Test Acc (%)	ΔTestAcc (pp)	Test Error (%)	Val Acc (%)
Balanced (Baseline)	97.67 ± 0.10	+0.00	2.33 ± 0.10	98.20 ± 0.26
Mild Imbalance (40/35/25)	97.58 ± 0.10	-0.08	2.42 ± 0.10	98.10 ± 0.44
Moderate Imbalance (50/30/20)	97.45 ± 0.13	-0.22	2.55 ± 0.13	98.40 ± 0.60
Strong Imbalance (60/25/15)	97.40 ± 0.09	-0.27	2.60 ± 0.09	98.70 ± 0.20
Very Strong Imbalance (70/20/10)	97.38 ± 0.20	-0.28	2.62 ± 0.20	98.80 ± 0.26
Balanced + Noise σ=0.05	97.58 ± 0.12	-0.08	2.42 ± 0.12	98.13 ± 0.35
Balanced + Noise σ=0.10	95.70 ± 0.87	-1.97	4.30 ± 0.87	98.00 ± 0.35
Balanced + Noise σ=0.15	76.92 ± 13.02	-20.75	23.08 ± 13.02	98.03 ± 0.42
Imbalance (50/30/20) + Noise σ=0.05	97.03 ± 0.38	-0.63	2.97 ± 0.38	98.80 ± 0.44
Imbalance (70/20/10) + Noise σ=0.10	95.65 ± 1.03	-2.02	4.35 ± 1.03	99.27 ± 0.58

Table 9: **Isolating shift axes: prior shift is benign, covariate shift is catastrophic.** Left: varying class imbalance with no noise. Right: varying noise with balanced classes. The asymmetry is stark.

Prior shift only (σ=0)				**Noise shift only (balanced priors)**		
Config	KL_y	Test Acc (%)	Δ (pp)	σ	Test Acc (%)	Δ (pp)
Balanced	≈ 0	97.67 ± 0.10	+0.00	0.05	97.58 ± 0.12	-0.08
40/35/25	0.018	97.58 ± 0.10	-0.08	0.10	95.70 ± 0.87	-1.97
50/30/20	0.070	97.45 ± 0.13	-0.22	0.15	76.92 ± 13.02	-20.75
60/25/15	0.158	97.40 ± 0.09	-0.27	—	—	—
70/20/10	0.293	97.38 ± 0.20	-0.28	—	—	—