DiScoFormer: Plug-In Density and Score Estimation with Transformers

Vasily Ilin[1] Peter Sushko[2]

[1]Department of Mathematics, University of Washington [2]Allen Institute for Artificial Intelligence

The Problem

Estimating a density f and its *score* $\nabla \log f$ from i.i.d. samples underpins generative modeling, Bayesian inference, and kinetic theory.

- **KDE:** distribution-agnostic but cursed by dimension.
- **Neural score matching:** accurate, but *retrained per target*.

One network. Any distribution. Density *and* score.

DiScoFormer: A Sequence-to-Operator Model

Learn two permutation- and affine-equivariant operators on i.i.d. samples:

$$T(X)_i \approx f(X_i), \qquad S(X)_i \approx \nabla \log f(X_i).$$

Permutation equivariance — Transformer w/o positional encodings.
Affine equivariance — whitening layer + rotation augmentation.
Whitening centers, decorrelates, and rescales each cloud to a canonical coordinate system before attention acts on the sample geometry.

Theory: Attention *is* Kernel Smoothing

Prop. 3.2 (arbitrary inputs)

For any PSD B, softmax attention is a *reweighted Gaussian kernel*:

$$\frac{\exp(x_i^\top B\, x_j)}{\sum_k \exp(x_i^\top B\, x_k)} = \frac{w_j\, e^{-\frac{1}{2}\|x_i - x_j\|_B^2}}{\sum_k w_k\, e^{-\frac{1}{2}\|x_i - x_k\|_B^2}}$$

with $w_j = e^{\frac{1}{2}\|x_j\|_B^2}$.

Prop. 3.3 (constructive)

A two-layer Transformer ($d_{\text{model}} \geq d{+}1$, one head) *exactly* outputs the KDE score.

KDE is a special case of DiScoFormer.

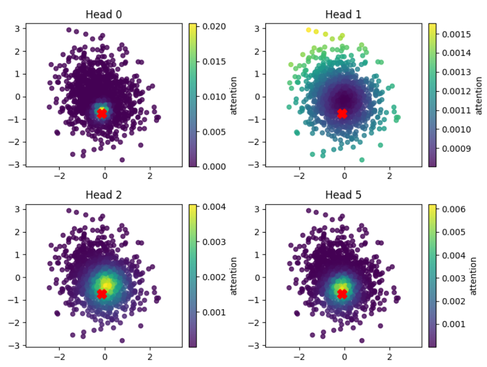


Fig. 1. Emergent head specialization: multi-scale, anisotropic kernels.

Score Estimation

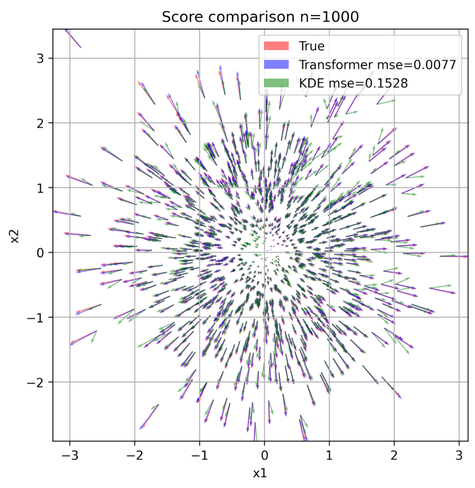


Fig. 2. 2D Gaussian: transformer score (blue) is more accurate than Silverman KDE (green).

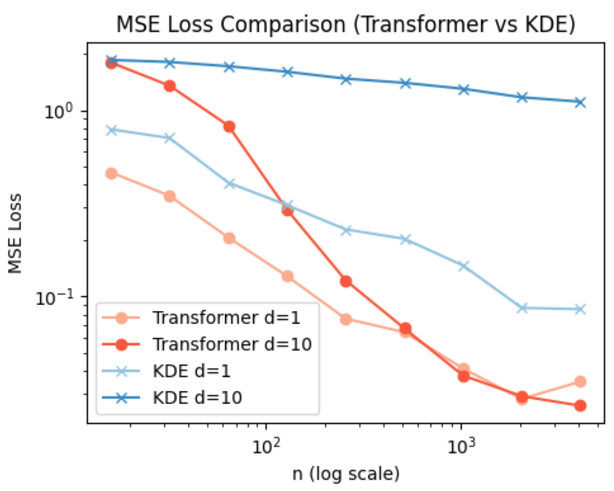


Fig. 3. Score MSE on GMMs in $d{=}1, 10$: DiScoFormer outperforms KDE across sample sizes.

Density Estimation

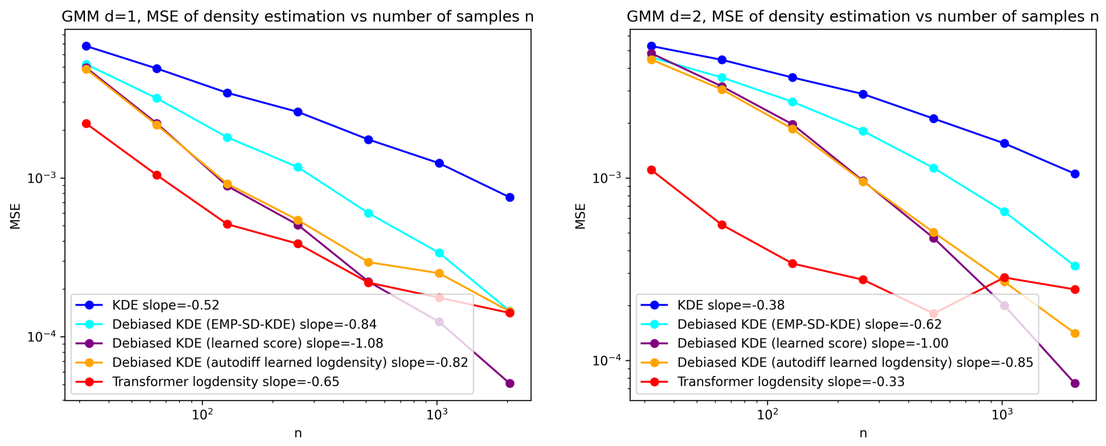
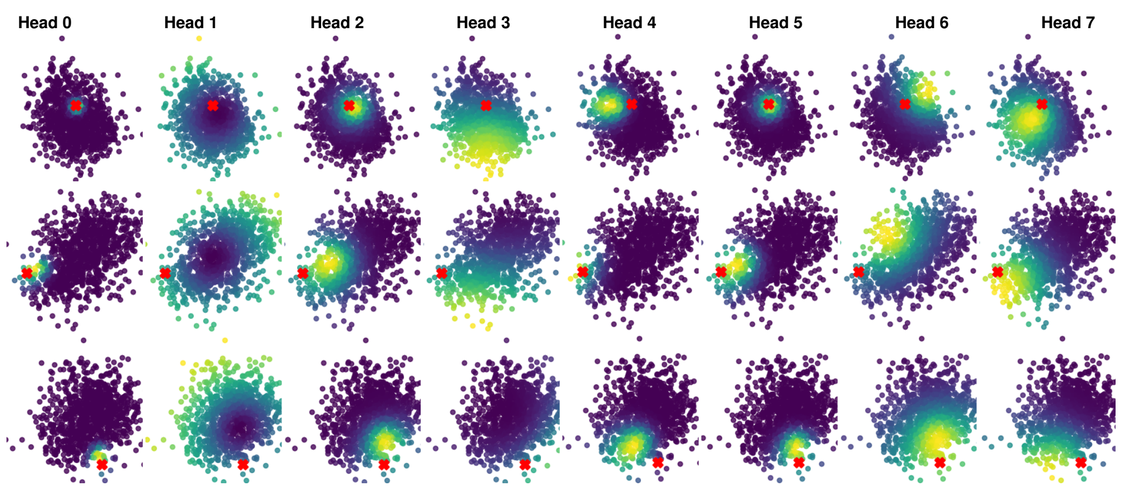


Fig. 4. Density MSE, $d{=}1$ (left), $d{=}2$ (right). SD-KDE with our learned score and DiScoFormer scale best; model trained only at $n{=}2048$.

Attention Heads

Attention heads specialize into local, far-field, and directional kernels.



High-Dimensional Scaling in $d{=}100$

Score and density estimation on random 2-component diagonal-covariance GMMs in $d{=}100$ ($n{=}2048$ context, 256 queries).

Method	Score MSE	Rel. MSE	Density MSE
Best KDE (Oracle-h)	1.090	54.6%	781
DiScoFormer	$\mathbf{0.167}$	$\mathbf{8.4\%}$	$\mathbf{20.8}$

6.5× better score, 37× better density than the best KDE. DiScoFormer explains 91.6% of the score signal vs. 45.4% — a regime where kernel methods structurally fail.

Sample-Size Scaling

Trained on $n{=}2048$; evaluated at test-time sample sizes up to $n{=}2^{17}{=}131{,}072$ ($64\times$ training). KDE throws OOM past $n{\approx}16$k; DiScoFormer keeps improving in *both* accuracy and wall-clock speed.

	$d{=}2$		$d{=}10$	
n	**Ours**	KDE	**Ours**	KDE
256	$\mathbf{14.7\%}$	43.8%	$\mathbf{7.5\%}$	65.6%
1,024	$\mathbf{8.8\%}$	33.6%	$\mathbf{4.7\%}$	61.3%
4,096	$\mathbf{7.2\%}$	24.6%	$\mathbf{3.3\%}$	57.1%
16,384	$\mathbf{6.8\%}$	17.2%	$\mathbf{2.8\%}$	52.9%
65,536	$\mathbf{5.4\%}$	OOM	$\mathbf{2.8\%}$	OOM
131,072	$\mathbf{5.4\%}$	OOM	$\mathbf{2.7\%}$	OOM

Score rel. MSE on GMMs — accuracy keeps decreasing with n far beyond training.

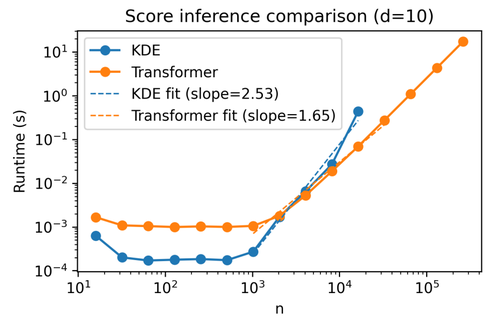


Fig. 5. Wall-clock runtime in $d{=}10$ on an L40S GPU. Both are $O(n^2)$ asymptotically, but attention kernels beat naive KDE past $n{=}2048$; KDE OOMs at $n{=}2^{15}$.

Ablation: Whitening is Essential for OOD

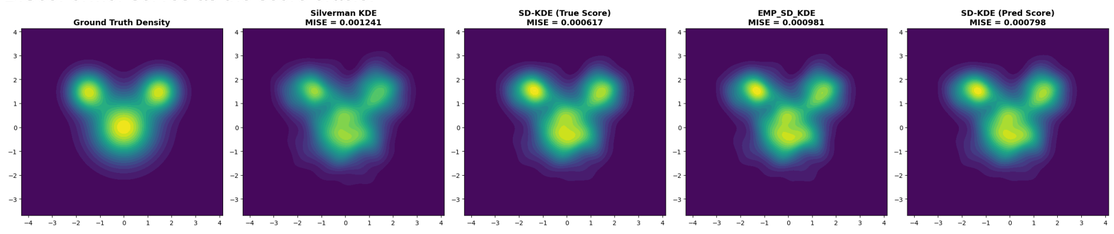
		Score Rel. MSE	Density Rel. MSE
ID	Whitening	$\mathbf{16.9\%}$	$\mathbf{1.1\%}$
	No whitening	18.6%	1.2%
OOD	Whitening	$\mathbf{25.8\%}$	$\mathbf{0.9\%}$
	No whitening	1487%	12.2%

Without whitening, the model is **15× worse than predicting zero** on OOD scales. Whitening supplies scale-equivariance and robust generalization.

Applications

Application 1: Score-Debiased KDE

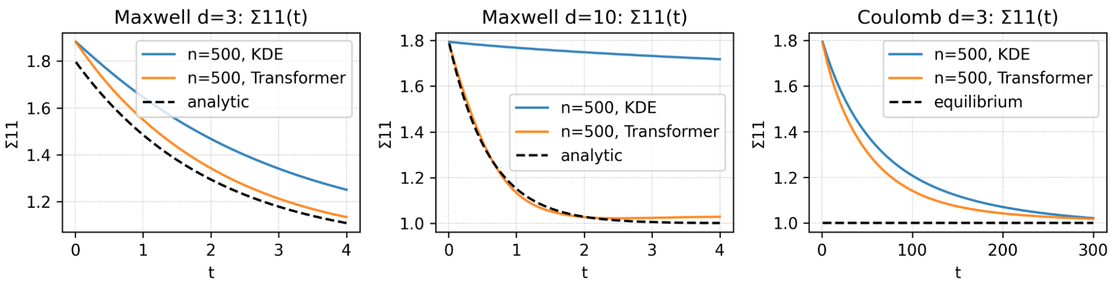
Epstein et al. 2025: given a score oracle, SD-KDE reduces KDE bias from $O(h^2)$ to $O(h^4)$ via $X \mapsto X + \frac{h^2}{2}\nabla \log f(X)$. DiScoFormer serves as the score oracle.



Application 2: Fokker–Planck Solvers (Landau)

Plug-in score for deterministic particle solvers of kinetic equations (SBTM-like) — no per-simulation retraining. DiScoFormer tracks the analytic covariance where KDE-based scores diverge.

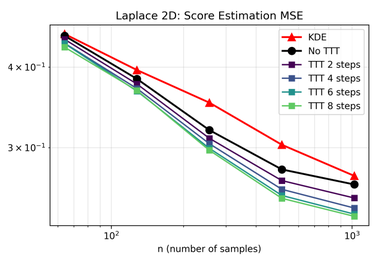
Σ11(t) for Maxwell (d=3,10) and Coulomb (d=3)



OOD Generalization: Laplace and Student-t + TTT

Trained only on GMMs; evaluated zero-shot on 2D Laplace and on 2D Student-t ($\nu{=}3$). For Student-t, test-time training adapts with the consistency loss

$$\mathcal{L}_{\text{con}} = \frac{1}{n}\sum_{i=1}^{n} \|S(X)_i - \nabla_{x_i} \log T(X)_i\|_2^2 .$$



2D Laplace

n	KDE	Transformer
512	0.381	**0.360**
1024	0.331	**0.299**
2048	0.299	**0.276**
4096	0.265	**0.260**

2D Student-t ($\nu{=}3$)

n	KDE	No TTT	TTT 4	TTT 6	TTT 8
128	0.152	0.198	0.168	0.157	**0.145**
256	0.121	0.112	0.096	0.091	**0.090**
512	0.092	0.057	0.049	**0.049**	0.051
1024	0.081	0.102	0.081	0.077	**0.077**

Takeaways

- **One model, plug-in.** Train once on GMMs; deploy zero-shot across distributions and dimensions.
- **Attention *is* KDE** — and strictly more expressive (Prop. 3.2–3.3).
- **Drop-in oracle** for SD-KDE, Fisher information, entropy, Fokker–Planck-type PDEs.