

Peer Review

Review of: "Influence Functions for Scalable Data Attribution in Diffusion Models"

Evgenii Arsentev¹

1. Independent researcher

1. What the work claims

The paper builds a principled influence-function framework for data attribution in diffusion models. Its central technical move is to separate two questions that earlier work had blurred: which generalised Gauss-Newton (GGN) matrix to use as the Hessian approximation for the diffusion objective, and how to make that matrix tractable. On the first question, the authors set out two splits of the loss (§2.2.1, §3.1.1): an “MC-Fisher” GGN that linearises only the network (Eq. 9, estimated with sampled auxiliary targets as in Eq. 10, derivation in Appendix B), and an “empirical Fisher” that linearises the composition of network and loss (Eq. 11). They argue that the second is what TRAK and D-TRAK implicitly use, and that the first is better motivated because the diffusion loss is nonlinear in the network output. On the second question, they apply K-FAC, choose the “expand” variant for weight-shared layers on the grounds that the diffusion problem is an expand-type setting (§3.1.2), and add int8 quantisation of query gradients with query batching (§3.2, Appendix C).

Empirically, on DDPMs trained on CIFAR-2 and CIFAR-10, the paper reports that K-FAC Influence beats TRAK in every LDS setting and beats D-TRAK when the loss measurement is used (Figures 2a and 2b, text of §4), surfaces more influential points under retraining with 2% and 10% of the data removed (Figure 3), and is far less sensitive to damping than the TRAK family (Figures 7 to 10). The second half of the paper is more unusual and, to me, more interesting: §4.1 and Appendix E document three ways influence functions misbehave here. Substituting a different per-timestep loss can predict a target timestep better than the correct one (Observation 1, Figure 11). Removal of data almost always raises loss, while the

predictions are roughly symmetric (Observation 2, Figures 15 and 16). And the ELBO on generated samples is nearly invariant to which 50% of the data a model saw (Observation 3, Figure 17).

2. Whether the conclusion holds on the data

I want to begin with what this paper does that most attribution papers do not. It reports its own failure modes as findings rather than burying them, and it recasts prior methods as points in one design space, so that the comparison K-FAC Influence vs. TRAK isolates two concrete choices (GGN type, and K-FAC vs. random projection) instead of comparing whole pipelines. The baselines are also treated generously. TRAK and D-TRAK get their best damping from a sweep on the very LDS benchmark they are scored on (Appendix G.3), a projection dimension of 32768, and float32 after the authors found float16 degraded them (Figure 20). Beating a tuned baseline with an untuned method is the right way round.

The LDS protocol is described with enough detail to judge its strength (Appendix G.5): $M = 100$ subsampled datasets, each containing 50% of the data, $K = 5$ seeds per subset, so $100 \times 5 = 500$ retrained models per benchmark, scored on 200 generated samples, with 5000 Monte Carlo samples to estimate each “true” measurement. That is a serious benchmark for 32×32 images. Standard errors are reported over generated samples (Figure 2 caption, Figure 3 caption).

Where I think the claims run slightly ahead of the evidence:

Scope. All experiments are on CIFAR-2 and CIFAR-10 at 32×32 with a pixel-space DDPM U-Net (Appendix G.1, G.2). The introduction motivates the work with commercial text-to-image systems and copyright litigation, and the title says “scalable”. I could not find any experiment on a latent diffusion model, a conditional model, or a dataset beyond 50000 images. The method plausibly scales, since K-FAC is per-layer and the paper covers convolutional layers (§3.1.2) and the linear layers inside attention (G.2, G.3), but scalability is argued, not measured. No wall-clock time, memory footprint, or GPU-hours appear for K-FAC Influence versus TRAK. I would add a cost table, even for CIFAR-10 alone.

Significance of method differences. As rendered on Qeios, the numerical LDS values inside Figures 2a, 2b, and 3 are not legible in the text, so I could not check the margins myself. The captions say standard errors are shown. With 200 query samples, I would ask for a paired test across samples (the same 200 samples are scored by every method), because the paired difference is what supports “K-FAC Influence outperforms TRAK in all settings”. Unpaired standard errors can hide or exaggerate a small gap.

The Hessian ablation is confounded with damping. The §4 paragraph on Figure 4 concludes that the results “seem to imply that a better Hessian approximation directly results in better downstream data

attribution performance”, and in the very next sentence states plainly that approximation quality is not evaluated directly and that damping was not swept for all variants. Given that Figures 8 and 10 show the empirical-Fisher methods are exactly the ones that are damping-sensitive, the expand-vs-reduce and MC-vs-empirical gaps in Figure 4 could partly reflect damping rather than curvature quality. That sentence would be safer as a hypothesis.

“without the need for method-specific hyperparameter tuning” (abstract). The evidence is good (Figures 7 and 9: small damping works if enough MC samples are drawn), but K-FAC Influence still has two knobs, damping and the number of MC samples, set to 250 for all CIFAR runs (G.3). The claim is fairer stated as “insensitive to damping over the range tested, given 250 MC samples”.

The oracle row. §4 describes “Exact Retraining” as retraining a single model, while the LDS target averages $K = 5$ seeds (Eq. 13, G.5). Seed noise alone therefore caps how high a single-retrain oracle can score. A sentence explaining this would help readers interpret the gold markers in Figure 2.

The three observations in §4.1 are, in my reading, well supported by the figures cited, and they are honestly framed. Observation 3 in particular is a useful negative result for anyone who wants to use the ELBO as a proxy in copyright arguments.

3. Reproducibility

This is above average. The paper gives the dataset construction (CIFAR-2 as 5000 images, 2500 car and 2500 horse, following [15], G.1), the architecture and noise schedule (DDPM CIFAR-10 U-Net, G.2), sampling with 1000 steps and the “small” variance choice (G.2), training lengths (160000 steps for the full CIFAR-10 model and 80000 for subsets; 32000 and 16000 for CIFAR-2, G.7), a cosine learning-rate schedule for CIFAR-2, the TRAK projection dimension and precision (G.3), the damping convention and its factor-of-N difference from the TRAK codebase (G.4), the handling of data augmentation in both GGN variants (G.8), and the software base (the `curvlinops` package, G.3).

Things I could not find in the Qeios text:

- A link to the authors’ own code, the trained checkpoints, or the 500 retrained models per benchmark. The acknowledgements mention model weights shared by another group; it is unclear which experiments used those.
- Batch size and optimiser settings, which G.7 gives only by reference to the original DDPM procedure [13]. G.7 also gives “410 epochs” for CIFAR-10, which implies a batch size, but the number itself is not

stated.

- The random seeds for the 200 generated query samples and for the subset draws.
- Compute: hardware, time to fit K-FAC, time per query, storage of quantised gradients.
- The exact damping grid used in the sweeps (Figures 7 to 10 show it, but as rendered here the axis values are not legible).

Because this is not an LLM paper, the usual issues of model snapshots and decoding parameters do not apply. The remaining gaps are ordinary and easy to close.

4. What specifically to fix

1. **§2.1, first line.** A LaTeX margin note survived into the Qeios version: the text shows “bordercolor=orange!50” three times, then “todo:” and a note from BM about changing the notation for x to subscripts “for space/better formatting?”. Please remove it.
2. **Appendix E, paragraph headed “Observation 1. Figure 15 shows the changes in losses...”** In §4.1, Figures 15 and 16 support Observation 2, not Observation 1. The label in the appendix looks like a copy error.
3. **§4, sentence ending “throughout.8”, and footnote 8.** Marker 8 appears twice (also at the end of §3.3), and the rendered footnote 8 seems to join a note on reparameterisation with a note on validation examples. I would renumber.
4. **Figures 2a, 2b, 3.** Please add a small table with the LDS means and standard errors per method and dataset, plus a paired comparison of K-FAC Influence vs. TRAK and vs. D-TRAK across the 200 query samples. On Qeios, the in-figure numbers do not survive text extraction, and a table would also make the paper citable on exact values.
5. **§4, paragraph on Figure 4.** Either sweep damping for all four GGN/K-FAC variants, or soften “directly results in” to a hypothesis.
6. **Abstract and §1.** State that experiments are limited to unconditional pixel-space DDPMs on CIFAR-2 and CIFAR-10. If a larger model is out of reach, one sentence on expected cost scaling (K-FAC factor sizes per layer for the U-Net in G.2) would support the word “scalable”.
7. **New short subsection on cost.** Wall-clock and memory for fitting K-FAC, computing and storing int8 query gradients, and scoring all training points, next to the same numbers for TRAK at $d_{\text{proj}} = 32768$.

8. §4, “Exact Retraining”. Clarify that the oracle uses one retrained model while the LDS target averages $K = 5$ seeds, and say what ceiling that implies.
9. **Appendix G.7.** Give batch size, optimiser, and learning rate explicitly.
10. **Code availability.** A repository link, or a statement of when it will be released, belongs in the main text or Declarations.
11. **Typos I noticed**, all minor: “copmared” and “Thes appears” (§5); “funcitons” and “we can predicted” (Figure 5 caption and Appendix A); “we actually retraining a single model” (§4); “daea augmentations” and “out ‘empirical data distribution’” (G.8); “(see 20,” missing “Figure” and a closing bracket (G.3); “LDS scores on for TRAK” (Figure 20 caption); “less that” (Appendix E); “estiamtion” (Figure 8 and Figure 10 captions); and “altenrative” (opening paragraph of Appendix D).

5. Recommendation

Minor revisions. The framework is clean, the comparison to prior methods is set up fairly (and if anything, favours the baselines), and the honest catalogue of failure modes in §4.1 is a contribution in its own right. What I would change is mostly presentation: a results table with paired tests, a cost report, a softer reading of the Figure 4 ablation, a clear scope statement, and the leftover editorial artefacts. None of this touches the core result.

Declarations

Potential competing interests: No potential competing interests to declare.