FLASHDIFFUSION: FUSED TILED KERNEL SPECTRAL DECOMPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Diffusion maps, and kernel methods more generally, provide an interpretable nonlinear spectral representation basis for geometric learning. In the geometric limit, small bandwidth, these matrices tend to be high rank and thus require materializing dense Gaussian kernels requires $\mathcal{O}(N^2)$ memory. We introduce FLASHDIFFUSION, a matrix-free method that evaluates dense Gaussian kernel blocks in fused GPU tiles and couples the eigensolver to an empirical β-flow that selects the finite-sample resolution scale. A continuation over sample size and bandwidth warm-starts increasingly expensive spectral solves from coarser resolutions.

1 INTRODUCTION

Kernel nonparametric methods and their associated manifold-learning methods, e.g. Laplacian eigenmaps (Belkin & Niyogi, 2003), offer an interpretable route to nonlinear representation learning: nonlinear correlations $\mathsf{K} = \exp_{\odot}(-\beta\mathcal{H})$ are obtained by exponentiating a matrix of linear correlations $\mathcal{H} \in \mathbb{R}^{N\times N}$, e.g. pairwise squared distances. They face two coupled difficulties. First, the pairwise kernel requires $\mathcal{O}(N^2)$ arithmetic and, if materialized, $\mathcal{O}(N^2)$ memory. Second, the bandwidth β fundamentally changes the spectrum and effective rank of the kernel. Prior work has addressed the computational problem with kNN graphs Muja & Lowe (2014), low-rank approximations and fast transforms.[1] We discuss kNN separately in §A, since it sparsifies spatial interactions rather than truncating spectral rank. Low-rank methods face a distinct limitation: at the bandwidths required to resolve a substantial manifold spectrum, the Gaussian kernel is generally high rank, so aggressive compression removes spectral directions the diffusion representation is intended to recover. Bandwidth selection has meanwhile been approached through median or quantile rules, locally adaptive k-th-neighbor bandwidths (Zelnik-Manor & Perona, 2004), and the log–log kernel-sum criterion (Coifman et al., 2008); these provide useful scale estimates, but depend on the dataset's sampling and geometry and do not directly optimize finite-sample spectral resolution.

We propose that these two problems are coupled and should be treated together. With finite data, the diffusion matrix approximates an underlying continuum operator. Singer and Hein (Singer, 2006; Hein et al., 2007) showed that the kernel length scale required for consistency depends on sampling density and hence on N. In our parameterization, increasing β decreases the spatial kernel length scale, so greater sampling can support a larger β and resolve finer structure. We therefore seek an empirical resolution scale $\beta^\star(N)$, selected before the leading spectrum reaches the sampling and numerical resolution floors. Two limits clarify the picture: as $\beta \to 0$, the kernel approaches its broad-kernel, low-rank linear limit (Candanedo, 2025b); along a continuum-consistent path $N \to \infty$ with $\beta = \beta^\star(N)$, progressively finer manifold structure becomes resolvable and the kernel develops a correspondingly large effective rank (Clemente & D'Elia, 2018).

Recent work connects self-attention with diffusion maps (Candanedo, 2026), suggesting a computational strategy for the quadratic pairwise operator. FlashAttention (Dao et al., 2022; Zadouri et al., 2026) addresses the analogous $\mathcal{O}(N^2)$ memory obstruction in attention by evaluating pairwise interactions in GPU tiles rather than materializing the full matrix. FLASHDIFFUSION applies the same IO-aware principle to the Gaussian kernel: pairwise diffusion interactions are evaluated in parallel GPU tiles and used immediately as matrix-block products inside a block Lanczos eigensolver

[1] Random Fourier Features (Rahimi & Recht, 2007; 2008), Nyström features (Williams & Seeger, 2000; Long & Ferguson, 2019), and the fast Gauss transform (Greengard & Strain, 1991; Raykar & Duraiswami, 2005; Hertrich, 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

(Lanczos, 1950; Baglama et al., 2003), avoiding explicit materialization of the $\mathcal{O}(N^2)$ kernel while retaining the dense interaction at the chosen precision.

FLASHDIFFUSION remains formally $\mathcal{O}(N^2)$ in pairwise arithmetic, and each Lanczos iteration requires another quadratic kernel application. At small β the broad Gaussian kernel has rapidly decaying spectrum, so the leading eigenspace is inexpensive to resolve. Increasing β reveals additional spectral directions and makes the eigensolve progressively more demanding. We therefore bootstrap across sample size and bandwidth: starting from a subsample $n \ll N$, FLASHDIFFUSION finds $\beta^\star(n)$ and its eigenspace at cost $\mathcal{O}(n^2)$, extends the eigenspace to a larger sample by the Nyström extension (Bengio et al., 2003), and re-optimizes β. Repeating this constructs an (n, β) ladder following the empirical resolution curve $\beta^\star(n)$ toward the full dataset. For nested subsamples drawn from the same underlying distribution, increased sampling density progressively resolves finer spectral structure, allowing the full-N solve to begin from a well-adapted eigenspace. FLASHDIFFUSION therefore couples bandwidth selection, dense kernel evaluation, and eigenspace continuation while preserving the high-rank diffusion operator in the regime where its resolved manifold structure is largest.

2 FLASHDIFFUSION ALGORITHM

2.1 MATRIX-FREE DIFFUSION OPERATOR

Given N observations $\mathcal{R} \in \mathbb{R}^{N \times D}$, diffusion maps begin with the Gaussian affinity operator:

$$\mathsf{K} = \exp_\odot(-\beta\,\mathcal{H}), \tag{1}$$

where $\mathcal{H}$ contains the pairwise squared Euclidean distances. Let $q = \mathsf{K}\mathbf{1}$. The Coifman–Lafon normalization (Coifman & Lafon, 2006) defines $\mathsf{K}_\alpha = \mathrm{diag}(q)^{-\alpha}\,\mathsf{K}\,\mathrm{diag}(q)^{-\alpha}$, with degree $d = \mathsf{K}_\alpha\mathbf{1}$ and row-stochastic diffusion operator $\mathsf{P} = \mathrm{diag}(d)^{-1}\mathsf{K}_\alpha$. Although P is generally asymmetric, it is similar to the symmetric operator (with $s = q^{-\alpha} \odot d^{-1/2}$, powers taken element-wise):

$$\mathsf{M} = \mathrm{diag}(d)^{1/2}\,\mathsf{P}\,\mathrm{diag}(d)^{-1/2} = \mathrm{diag}(s)\,\mathsf{K}\,\mathrm{diag}(s). \tag{2}$$

Thus P and M share the same eigenvalues, and the diffusion spectrum can be obtained with a symmetric block Lanczos eigensolver. Note that each fresh β pays an entry fee before any Lanczos iteration can run: the normalization vectors are themselves kernel applications, $q = \mathsf{K}\mathbf{1}$ and, for $\alpha \neq 0$, $d = \mathrm{diag}(q)^{-\alpha}\,\mathsf{K}\,\mathrm{diag}(q)^{-\alpha}$, two full $\mathcal{O}(N^2)$ passes that must complete before s exists. This is a structural departure from attention, which normalizes each softmax row on-the-fly within a single pass. The resulting operator is symmetrizable, eq. 2, which attention's is not, and this purchase is what licenses the symmetric eigensolver and its guarantees. Once s is computed, each Lanczos iteration reduces to the matrix-block operation, for a block $V \in \mathbb{R}^{N \times b}$:

$$\mathsf{M}V = \mathrm{diag}(s)\,\mathsf{K}\,(\mathrm{diag}(s)V). \tag{3}$$

Normalization passes and Lanczos iterations therefore share a single computational primitive: the Gaussian application $W \mapsto \mathsf{K}W$, with $W = \mathbf{1}$, $q^{-\alpha}$, or $\mathrm{diag}(s)V$. This primitive can be evaluated directly from the observations. With mean-centered coordinates $\bar{\mathcal{R}} = \mathcal{R} - \mathbf{1}\mu^\top$, define $\mathcal{G} = \bar{\mathcal{R}}\bar{\mathcal{R}}^\top$ and $r = \mathrm{diag}(\mathcal{G})$. Translation preserves pairwise distances (and centering bounds r, taming cancellation in the distance form), hence:

$$\mathsf{K} = \exp_\odot\left(-\beta\left(r\mathbf{1}^\top + \mathbf{1}r^\top - 2\mathcal{G}\right)\right). \tag{4}$$

The kernel is thus a function of Gram blocks of the observations, and never needs to exist as a stored matrix: its tiled evaluation is the subject of §2.2.

2.2 XGEMM: TILED KERNEL OPERATOR

FlashAttention reduces memory traffic by constructing pairwise interactions in fast on-chip memory rather than materializing the full quadratic matrix in HBM (Dao et al., 2022); FLASHDIFFUSION applies the same IO-aware principle to the Gaussian kernel application in eq. 3. Unlike the attention operator, the diffusion operator is scaled on both sides, $\mathrm{diag}(s)\,\mathsf{K}\,\mathrm{diag}(s)$, and s couples the Lanczos block to global functions of the kernel (q and d): no online construction exists, and the normalization must complete before any operator application. The costs are nonetheless contained: the q and d

passes are themselves kernel applications (§2.1) and reuse the identical tile schedule with V replaced by $\mathbf{1}$ and $q^{-\alpha}$; and within an operator pass the right scaling is the precomputed block $\mathrm{diag}(s)V$ while the left scaling is a row rescale fused into the accumulation. FLASHDIFFUSION thus inherits FlashAttention's tiling while shedding its online-rescaling machinery, the hard part of the attention kernel, at the price of one or two normalization passes per new β.

For row and column tiles I and J, the observations $\bar{\mathcal{R}}_I$ and $\bar{\mathcal{R}}_J$, their norms r_I, r_J, and the Lanczos block V_J are loaded from HBM into on-chip memory. The tile is evaluated as, with $\mathcal{G}_{IJ} = \bar{\mathcal{R}}_I \bar{\mathcal{R}}_J^\top$,

$$\mathsf{K}_{IJ} = \exp_\odot\left(-\beta\left(r_I \mathbf{1}^\top + \mathbf{1} r_J^\top - 2\mathcal{G}_{IJ}\right)\right), \tag{5}$$

followed immediately by:

$$Y_I \leftarrow Y_I + \mathsf{K}_{IJ} V_J. \tag{6}$$

The intermediate matrices $\mathcal{G}_{IJ}$ and K_{IJ} exist only within the working tile and are discarded after accumulation. Iterating over J therefore evaluates the exact dense kernel application without ever storing the $N \times N$ affinity matrix. This computation maps directly onto GPU streaming multiprocessors: independent output tiles are distributed across thread blocks, the local Gram product runs on the tensor cores, and the distance transform, exponential, and accumulation are fused into the same GPU kernel as its epilogue. The implementation targets NVIDIA Ampere through Blackwell architectures (NVIDIA, 2020; 2024) using CUTLASS/CUTE primitives (Thakkar et al., 2023). The resulting XGEMM, a GEMM[2] primitive that applies the square-distance transform and elementwise exponential before a second on-chip product, retains the exact $\mathcal{O}(N^2)$ pairwise arithmetic while persisting only the observations.

2.3 β-FUNCTION FLOW

The matrix-free operator provides the spectrum at a fixed bandwidth β; FLASHDIFFUSION additionally determines the bandwidth at which the requested spectral basis is best resolved. The algorithm treats $\log\beta$ as a one-dimensional continuation coordinate and advances through a sequence of spectral probes. Each probe constructs the normalized diffusion operator at the current β, solves for the Perron mode together with the requested nontrivial eigen-pairs using block implicitly restarted Lanczos, and returns the quantities required to determine the next step and evaluate the bandwidth-selection criterion (Sorensen, 1992; Baglama et al., 2003). For centered observations, the initial bandwidth is chosen directly from the data. Since $\mathbb{E}_{ij}\|x_i - x_j\|^2 = 2\mathbb{E}_i\|x_i\|^2$, FLASHDIFFUSION initializes:

$$\beta_0 = \min\left(0.1, \frac{1}{2N^{-1}\sum_{i=1}^N \|\mathcal{R}_i\|^2}\right). \tag{7}$$

Here $\mathcal{R}$ is the centered data matrix. The cap at 0.1 places the initial probe conservatively on the broad-kernel side when the variance-based estimate would otherwise begin at a narrow kernel. The initial scale is therefore determined in $\mathcal{O}(ND)$ work before any quadratic kernel application is required.

The potential-based continuation advances on a dyadic grid, $\beta \mapsto 2\beta$. The dyadic structure is required by the effective-sample estimator below, which compares the off-diagonal kernel masses at neighboring probes β and 2β. Elementwise, the Gaussian kernel satisfies $\mathsf{K}_{ij}(2\beta) = \mathsf{K}_{ij}(\beta)^2$, so the two probes correspond exactly to adjacent scales of the same kernel family. The quantity $S_{\mathrm{off}}(2\beta)$ is accumulated during the next probe's own kernel evaluation, so evaluating the effective-sample statistic requires no kernel pass beyond those already present in the natural progression of the sweep. Neighboring probes also have slowly varying eigen-spaces, so the eigenvectors from the previous probe provide the warm start for the next and the solver follows a gradually rotating spectral subspace rather than restarting from an uninformed initialization. A non-dyadic continuation with adaptive ratios $r_{\mathrm{fast}} = 1.5$, $r_{\mathrm{slow}} = 1.2$, and $r_{\mathrm{crawl}} = 1.05$, selected from Lanczos cycle count and Perron-gap proximity, is retained for API compatibility but does not use the empirical potential selector described below.

[2]General Matrix Multiply.

Bandwidth selection uses quantities already produced by normalization and the eigensolver. The first term measures the spectral heat action of the retained nontrivial modes,

$$\mathcal{V}_{\mathrm{bias}}(\beta) = \mathrm{median}_j \left(-\log \left(\frac{\lambda_j}{\lambda_0} \right) \right). \tag{8}$$

As $\beta \to 0$, the Gaussian affinity approaches the rank-one matrix $\mathbf{1}\mathbf{1}^\top$ and the nontrivial diffusion spectrum collapses, hence $\mathcal{V}_{\mathrm{bias}} \to \infty$. In the local heat-kernel regime, $\mathcal{V}_{\mathrm{bias}}(\beta) \sim \beta^{-1}$, recording the departure of the requested eigenspace from the broad-kernel rank-one limit. Finite sampling contributes a second scale through an effective sample size measured along the sweep. Let $S_{\mathrm{off}}(\beta) = \sum_{i \neq j} e^{-\beta \|x_i - x_j\|^2}$. For neighboring dyadic probes, FLASHDIFFUSION defines:

$$n_{\mathrm{eff}}(\beta) = \frac{S_{\mathrm{off}}(\beta)^2}{N S_{\mathrm{off}}(2\beta)}. \tag{9}$$

For k equally weighted interactions per row this reduces to k, while for nonuniform weights it gives the corresponding effective interaction count. In the heat-kernel regime it scales as $n_{\mathrm{eff}} \sim N\beta^{-d/2}$. The corresponding relative statistical uncertainty is:

$$\mathcal{V}_{\mathrm{stat}}(\beta) = \frac{1}{\sqrt{n_{\mathrm{eff}}(\beta)\mathcal{V}_{\mathrm{bias}}(\beta)}}. \tag{10}$$

Because both $S_{\mathrm{off}}(\beta)$ and $S_{\mathrm{off}}(2\beta)$ are produced by consecutive probes of the dyadic sweep, the statistic requires no auxiliary kernel evaluation. At large β, the diffusion operator approaches the identity and the leading spectral gaps eventually become comparable to the arithmetic precision of the kernel evaluation and eigensolver. Define the normalized Perron gap as $g_P(\beta) = (\lambda_0 - \lambda_1)/|\lambda_0|$. FLASHDIFFUSION incorporates a smooth machine-precision barrier,

$$\mathcal{V}_{\mathrm{mach}}(\beta) = \left(\frac{\kappa_{\mathrm{P}} \eta}{g_P(\beta)} \right)^p, \tag{11}$$

with exponent $p = 6$ and safety factor $\kappa_{\mathrm{P}} = 10$, together with the hard admissibility condition:

$$g_P(\beta) > \kappa_{\mathrm{P}} \eta. \tag{12}$$

Here η is the numerical precision scale of the kernel dtype, with $\eta = 8 \times 10^{-3}$ for bf16, 10^{-3} for tf32, 10^{-7} for `fp32`, and 2.22×10^{-16} for fp64.[3] The smooth term shapes the approach to the precision boundary from within the admissible region, while the hard condition prevents selection once the Perron and first nontrivial eigenvalues can no longer be reliably separated. The sweep therefore follows a measured trajectory between two spectral limits. Small β produces a broad kernel with a nearly rank-one spectrum, increasing β progressively resolves additional nontrivial modes, and sufficiently large β approaches the identity precision boundary. FLASHDIFFUSION combines the heat-action, finite-sample, and machine-precision terms into a dimensionless empirical potential on $\log \beta$,

$$\mathcal{V}(\beta) = \mathcal{V}_{\mathrm{bias}}(\beta) + \mathcal{V}_{\mathrm{stat}}(\beta) + \mathcal{V}_{\mathrm{mach}}(\beta). \tag{13}$$

The ideal bandwidth $\beta^\star$ is the minimizer of this finite-sample resolution potential within the admissible spectral regime. The discrete sweep returns an achieved bandwidth $\beta^\times \leq \beta^\star$, determined by the probes that can be resolved safely at the available sampling and arithmetic precision. A minimum is accepted once three consecutive probes bracket a stationary point in $\log \beta$ for which $\mathcal{V}_{\mathrm{bias}}$ is falling and either $\mathcal{V}_{\mathrm{stat}}$ or $\mathcal{V}_{\mathrm{mach}}$ is rising. This direction of the flow identifies the bias-variance well associated with the resolution boundary. A quadratic fit in $\log \beta$ over the bracketing triplet estimates the vertex, which is then resolved by a final probe. If the Perron wall is reached before the bracket is completed, the last safe probe defines $\beta^\times$. The selector therefore depends on empirical spectral and sampling quantities rather than a prescribed intrinsic dimension or asymptotic bandwidth exponent.

The continuation provides several geometric diagnostics without an additional kernel pass. In the local-manifold regime, the log-derivative of the off-diagonal kernel mass across a dyadic pair yields the empirical intrinsic dimension:

$$\widehat{d}(\beta) = -2 \frac{\Delta \log S_{\mathrm{off}}}{\Delta \log \beta}. \tag{14}$$

[3] The defaults reported here use `fp32` throughout, with $\eta = 10^{-7}$. The precision cascade described below is enabled only when multiple dtypes are requested.

The corresponding Singer-type scaling exponent can then be reported diagnostically. These quantities characterize the observed flow and permit comparison with continuum asymptotics, while the selected bandwidth is determined from the measured finite-sample spectrum rather than from a prescribed value of $\widehat{d}$ or a hard-coded scaling law.

When multiple kernel dtypes are supplied, FLASHDIFFUSION executes the continuation as a precision cascade. A sweep is first performed with the earliest requested dtype, and its selected state initializes the next precision level. The final bandwidth and eigenspace are taken from the last dtype in the cascade. This separates the precision used to explore the β trajectory from the precision required to resolve its final spectral boundary while preserving the same normalization, matrix-free kernel application, and block Lanczos machinery at every stage (Bradbury et al., 2018; Paszke et al., 2019). The default configuration uses a single dtype, `fp32`. The complete algorithm therefore consists of three nested operations: XGEMM supplies matrix-free Gaussian applications, block implicitly restarted Lanczos converts those applications into finite spectral probes, and the β flow uses successive probes to approach the finite-sample resolution optimum. The output is the achieved bandwidth $\beta^\times$, its nontrivial diffusion eigenvalues, and the corresponding symmetric eigenvectors of the normalized operator, from which the Markov right eigenvectors are obtained by diagonal rescaling.

3 EXPERIMENT: T^6 MANIFOLD

We benchmark FLASHDIFFUSION on a manifold whose continuum spectrum is known in closed form. The flat six-torus $T^6 = (S^1)^6$ with unequal radii:

$$R = (1.00,\ 1.08,\ 1.17,\ 1.27,\ 1.38,\ 1.50) \tag{15}$$

each circle embedded as $(R_a \cos\theta_a,\ R_a \sin\theta_a)$ with independent uniform angles $\theta_a \sim \mathrm{Uniform}[0, 2\pi)$, is a curvature-free product manifold in $\mathbb{R}^{12}$, zero-padded to $D = 32$ and sample-centered.[4] The continuum Laplace–Beltrami eigenvalues are $\sum_a m_a^2/R_a^2$ with integer mode numbers m_a. The unequal radii break the inter-circle degeneracies, and because $R_{\max}/R_{\min} = 1.50 < 2$ the entire first-harmonic ladder lies strictly below the first second-harmonic level: the largest first harmonic is $1/R_{\min}^2 = 1.00$, while the smallest second harmonic is $4/R_{\max}^2 \approx 1.778$, leaving a gap of ≈ 0.78 between the first-harmonic band and the rest of the spectrum. Each circle contributes a degenerate doublet $\{\cos(m\theta_a), \sin(m\theta_a)\}$ per harmonic m, so the first-harmonic band consists of six doublets; the leading doublet, that of the largest-radius circle ($s = 5$, $R_s = 1.50$), provides an exact target for the recovered modes.

We compute the leading $k = 3$ nontrivial modes at $\alpha = 1/2$ in FP32 on NVIDIA RTX PRO 6000 (Blackwell SM120) GPUs. Throughout, $\beta^\star(N)$ denotes the ideal resolution scale of §2.3: the largest bandwidth at which the requested leading spectrum remains distinguishable from the sampling and numerical resolution floors. Section 2.3 shows that $\beta^\star$ exists but is not computable from finite data; the in-loop heat-action selector returns an achievable approximation $\beta^\times(N) \le \beta^\star(N)$, subject to the Perron ceiling $\beta^\times$ that enforces the machine-precision bandwidth floor. The selector runs as it would on an unlabeled dataset: no test-side bandwidth grid, quantile rule, locally adaptive rule, or ground-truth angle enters $\beta^\times$, and the ground-truth angles enter only the post-hoc validation metric described below. The third computed mode ($k = 3$) is retained only as a Lanczos buffer, letting the solver certify that the multiplicity-two target doublet has closed before a possible fourth mode.

The sample count varies over the power-of-two sequence $N = 2^p$, $p = 10, \cdots, 22$, i.e. $1,024$ through $4{,}194{,}304$ observations. Timing variance is largest for small problems while the per-solve cost grows quadratically, so the number of repetitions decays geometrically from 64 at the smallest N to 1 at the largest, following $n_{\mathrm{rep}}(N) = \mathrm{clip}\left(\lceil 64^{(22-\log_2 N)/12} \rceil,\ 1,\ 64\right)$, giving $n_{\mathrm{rep}}(2^{10}) = 64$, $n_{\mathrm{rep}}(2^{17}) = 6$, and $n_{\mathrm{rep}}(2^{22}) = 1$. We report mean wall-clock time with one standard deviation over repeats.

[4] Only the first twelve ambient coordinates are non-zero; the padding fixes the ambient dimension for the kernel without adding structure. Samples are centered by subtracting the sample mean, which preserves the anisotropic radii and matches the centered input expected by FLASHDIFFUSION. All angles are drawn from a single seed (42) per N so that the validation diagnostic is reproducible; neither the seed nor the angles enter the solve, only the post-hoc score of Figure 1c.

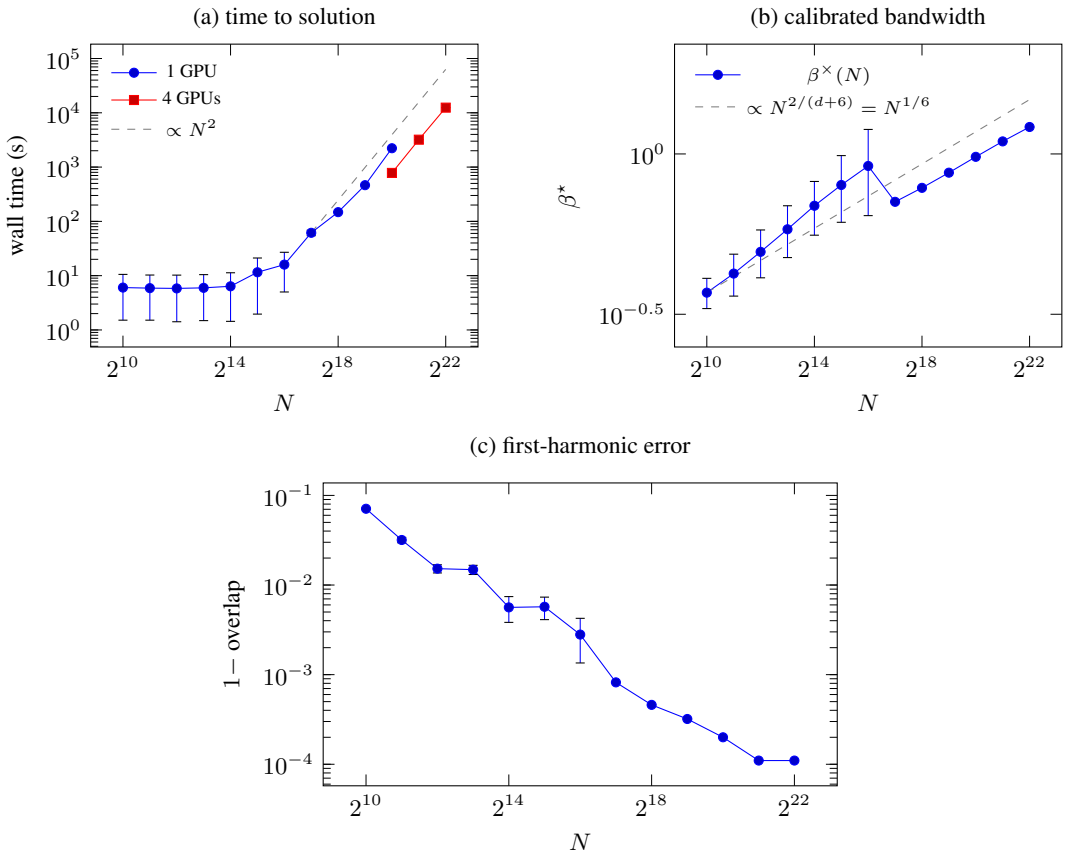


Figure 1: FlashDiffusion on uniform samples of the flat torus T^6 (RTX PRO 6000). **(a)** Wall time versus sample count; the dashed guide is exact $\mathcal{O}(N^2)$ scaling, and the flat region below $N \approx 2^{14}$ is the fixed launch/solver overhead. Four-GPU points continue the curve to $N = 2^{22} \approx 4.2$M. **(b)** The bandwidth emitted by the (n, β) ladder, against the Singer exponent $2/(d+6) = 1/6$ for $d = 6$ (dashed, anchored at the first point). **(c)** Recovery of the first harmonic band: projection error of the leading nontrivial eigenspace onto the exact T^6 harmonics, decaying with sample count toward the resolution floor. Error bars are one standard deviation over repeated runs where multiple runs were performed.

Figure 1a shows time to solution. Below $N \approx 2^{14}$ the runtime is flat at the fixed launch and solver overhead (≈ 6 s); above it the measured cost approaches the $\mathcal{O}(N^2)$ reference of the dense kernel application, eq. 3, with fitted exponent $\gamma = 0.157$ over the quadratic regime. A single GPU reaches $N = 2^{20}$ in 37 min, and four GPUs reach $N = 2^{22}$ (≈ 4.2M observations) in 3.5 h. A dense FP32 affinity matrix at $N = 2^{22}$ would occupy $N^2 \times 4$ bytes $\approx 7.0 \times 10^{13}$ bytes, or ≈ 70 TB; FLASHDIFFUSION evaluates the pairwise kernel in GPU tiles and never allocates this object. At $N = 2^{20}$ the four-GPU configuration completes in 781 s against the single-GPU 2225 s, a $2.85\times$ speedup or 71% parallel efficiency, with multi-GPU runs sharded over the row axis.

Figure 1b reports the selector's calibrated bandwidth $\beta^\times(N)$, the achievable approximation to the ideal resolution scale $\beta^\star(N)$ of §2.3, shown against the Singer reference slope $N^{2/(d+6)} = N^{1/6}$ for $d = 6$ anchored at the first point.[5] Since $\beta^\times \le \beta^\star$ pointwise, the measured curve is a lower estimate of the ideal curve; the visible gap to the Singer guide can reflect either a genuine sub-Singer scaling of $\beta^\star$ or a $\beta^\star - \beta^\times$ bias whose N-dependence is not separately known. The measured slope over the

[5]The reference exponent follows from the graph-to-manifold convergence of Singer (2006); Hein et al. (2007), where the bandwidth must shrink with sample count to keep the random walk local while the neighborhood graph remains connected; for a d-dimensional manifold sampled with a Gaussian kernel, the resulting bandwidth scales as $N^{-2/(d+6)}$, so $\beta^\star$ grows as the reciprocal.

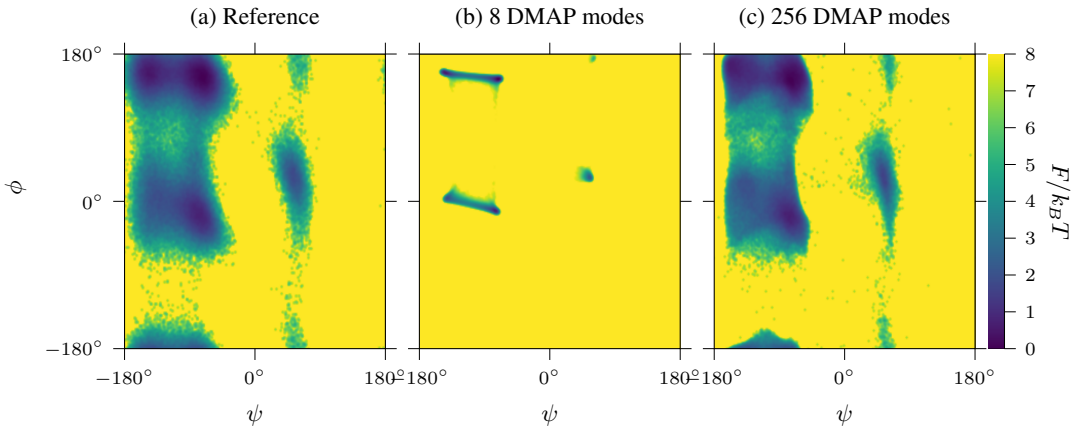


Figure 2: Ramachandran free-energy distributions for alanine dipeptide. **(a)** Reference distribution, **(b)** reconstruction using 8 diffusion-map (DMAP) modes, and **(c)** reconstruction using 256 DMAP modes. The horizontal coordinate is the dihedral angle ψ and the vertical coordinate is ϕ; both span $[-180°, 180°]$, with $(\phi, \psi) = (0, 0)$ at the center of each panel. The shared color bar reports the free energy F/k_BT from 0 to 8.

quadratic regime is 0.157, to be compared with $1/6 \approx 0.1667$; the measured curve runs somewhat shallower, $\approx 20\%$ below the Singer guide at the largest N. A visible break in the per-N error bars near $N = 2^{17}$ coincides with the drop of n_{rep} into the single-digit regime and may also reflect a change in the selector's effective safety margin as the spectrum hardens; it does not shift the central trend of $\beta^{\times}(N)$.

Figure 1c reports recovery of the known leading eigenspace. Because the continuum doublet is degenerate and may rotate arbitrarily, we score the two-dimensional span rather than individual eigenvectors: the whitened cross-covariance between the leading two nontrivial modes and $\{\cos\theta_s, \sin\theta_s\}$ yields a rotation-invariant overlap in $[0, 1]$, accumulated block-wise in `fp64` so that the diagnostic remains stable at multi-million N.[6] The subspace error: $(1 - \mathrm{overlap})$ decays from 7×10^{-2} at $N = 2^{10}$ to $\approx 10^{-4}$ at $N = 2^{22}$, confirming that the in-loop calibration tracks the resolvable geometry as sampling deepens: the leading harmonic is recovered to three significant digits by the time the sample count reaches 10^6.

4 Experiment: Alanine Dipeptide

We evaluate FlashDiffusion on the MDShare alanine-dipeptide benchmark (Nüske et al., 2017; Wehmeyer & Noé, 2018), consisting of three independent 250 ns molecular-dynamics trajectories and $N = 750{,}000$ configurations. Each configuration contains the 10 heavy atoms and is mapped to an E(3)-invariant representation using distances from every heavy atom to the first 8 heavy atoms in the dataset ordering as anchors (Candanedo, 2025a), giving $D = 80$ features. FlashDiffusion is run with $\alpha = 1/2$ in `fp32` and computes 256 nontrivial diffusion modes on eight NVIDIA RTX PRO 6000 GPUs. The eigensolve required 7315 s (2.03 h). Explicit storage of the corresponding dense `fp32` kernel would require $750{,}000^2 \times 4 \approx 2.25$ TB, which is avoided by the tiled matrix-free operator. This experiment uses the full N throughout and exercises the β continuation directly rather than the sample-size ladder.

The β flow required eight dyadic probes and selected $\beta^{\star} = 276.4$ from the empirical potential. The sweep continued beyond the selected point to characterize the precision boundary: $\beta = 552.8$

[6] Writing A for the matrix whose columns are the leading two nontrivial modes, B for the columns $(\cos\theta_s, \sin\theta_s)$, and G_{AA}, G_{BB}, G_{AB} for their centered covariance blocks, the score is $\sum_{ij} C_{ij}^2/2$ with $C = G_{AA}^{-1/2} G_{AB} G_{BB}^{-1/2}$, the squared RV coefficient between the two subspaces. It equals 1 iff the subspaces coincide and 0 iff they are orthogonal, up to floating-point round-off. The accumulation is blocked at 10^6 rows and carried in `fp64` so that the $N \times 2$ blocks are never materialized at full precision.

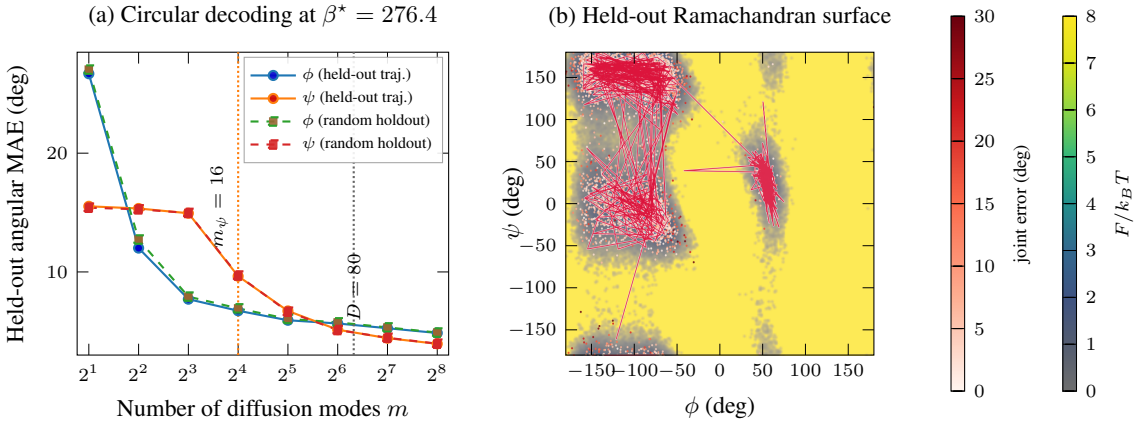


Figure 3: **Spectral resolution of alanine-dipeptide coordinates. (a)** Mean absolute wrapped error for the backbone dihedrals ϕ and ψ as the number of retained diffusion modes m increases. Solid curves evaluate the completely held-out trajectory, while dashed curves evaluate 5,000 randomly held-out frames from the decoder-training trajectories. The vertical markers indicate the ambient feature dimension $D = 80$ and the detected onset $m_\psi = 16$ of the principal ψ-resolution regime. **(b)** Held-out Ramachandran free-energy surface at $\beta^\star = 276.4$, with random-holdout frames colored by joint angular reconstruction error and a strided segment of the decoded held-out trajectory overlaid as a curve.

remained admissible, while $\beta = 1105.7$ crossed the `fp32` Perron wall with $g_P = 3.51 \times 10^{-9}$. At $\beta^\star$, the effective interaction count is $n_{\text{eff}} \approx 4.31 \times 10^4$, approximately 5.7% of the sample count, and the measured local dimension is $\widehat{d} \approx 4.85$. Here $\widehat{d}$ is a kernel-scale diagnostic of the resolved geometry in the E(3)-invariant representation and should not be identified with the two-dimensional (ϕ, ψ) reaction coordinate alone. The selected $\beta^\star$ therefore lies comfortably inside the resolved spectral regime rather than at the `fp32` precision boundary. Following the diffusion-map autoencoder construction of Candanedo (2024), we fit a single linear decoder from the first m diffusion coordinates to the circular targets $(\cos\phi, \sin\phi, \cos\psi, \sin\psi)$. The decoder is fit by least squares with ridge regularization $10^{-8}\,\mathrm{tr}(\mathcal{G}^{(m)})/m$, where $\mathcal{G}^{(m)}$ is the centered decoder Gram matrix. Two trajectories supply the decoder-fit population, from which $5,000$ frames are drawn uniformly without replacement as an additional random holdout, while the third 250 ns trajectory is excluded entirely from decoder fitting. The diffusion eigensolve itself remains unsupervised and transductive over all configurations; only the decoder uses the split. We evaluate $m \in \{2, 4, 8, 16, 32, 64, 128, 256\}$ using the mean absolute wrapped angular error $\left|\arg\exp(i(\widehat{\theta} - \theta))\right|$. Figure 3 shows distinct resolution profiles for the two backbone coordinates. The held-out ϕ error falls from 26.69° at $m = 2$ to 7.71° at $m = 8$, while ψ remains near 15° through $m = 8$ and begins its principal decrease at $m_\psi = 16$. At $m = 256$, the completely held-out trajectory reaches 4.86° MAE for ϕ and 3.95° for ψ, compared with 4.90° and 3.96° on the random holdout. The differences are below 0.05°, so the fully held-out trajectory shows no observable degradation relative to frames drawn from the decoder-training trajectories at the reported precision. Reconstruction also continues beyond the ambient feature dimension $D = 80$: from $m = 128$ to $m = 256$, the held-out errors decrease from 5.26° to 4.86° for ϕ and from 4.42° to 3.95° for ψ. Figure 3 additionally compares the held-out Ramachandran surface with decoded ensembles at $m = 8$ and $m = 256$; the latter recovers the principal metastable regions and their geometry, while a strided held-out trajectory segment has mean joint angular error 7.1°. We use the free-energy comparison as a qualitative visualization, with the angular errors providing the quantitative reconstruction measure. The spectrum confirms that these additional modes remain part of a substantial high-rank operator. All 256 computed eigenvalues are finite, positive, and correctly ordered, with $\lambda_{\max} = 0.9985$ and $\lambda_{256} = 0.2127$. The final computed eigenvalue therefore remains approximately 21% of the leading value and lies more than six orders of magnitude above the `fp32` working precision scale used by the sweep. Together with the continued reconstruction improvement for $m > D$, this shows that the useful diffusion representation extends well beyond a low-rank approximation of the 80-dimensional input features and provides a molecular example of the high-rank Gaussian regime targeted by FLASHDIFFUSION.

5 CONCLUSION

FLASHDIFFUSION makes the pragmatic choice to retain the dense $\mathcal{O}(N^2)$ Gaussian interaction rather than replace it by a sparse or low-rank surrogate. GPU tiling removes the need to materialize the kernel, alternatives that achieve precisely this exist in the literature including KEOPS Feydy et al. (2020) and we run a comparison with our primative XGEMM in §B. However, unlike previous work we consider a synergistic connection between β-flow and tiling requirements , via (n, β) continuation, to reduce the number of expensive full-data spectral solves.

This targets the regime in which the Gaussian operator itself is high rank. Fixed-k graphs remain natural when the neighborhood graph defines the desired operator, and low-rank approximations remain effective when only a small spectral subspace is required. For data whose useful spectral rank is unknown in advance, FLASHDIFFUSION instead treats rank and spectral resolution as quantities to be measured from the finite novel sample.

The principal remaining computational limitation is the feature dimension. Each Gaussian application costs $\mathcal{O}(N^2 D)$, and therefore becomes cubic when $D \sim N$. The experiments considered here remain well inside the quadratic regime, with $D/N \approx 8 \times 10^{-6}$ on T^6 and $D/N \approx 10^{-4}$ for alanine dipeptide. Since the kernel depends on the representation through pairwise distances, feature compression provides a natural front end: PCA, Johnson–Lindenstrauss projections (Johnson & Lindenstrauss, 1984), or learned encoders can reduce D provided the induced distance distortion is controlled. This suggests a useful division of labor: the front end determines the geometry presented to the kernel, while diffusion maps perform harmonic analysis on that geometry. Jointly calibrating representation distortion and spectral resolution is a natural direction for future work.

The β flow also admits a scale interpretation related to renormalization ideas (Zinn-Justin, 2019). For a Gaussian heat kernel, the diffusion time satisfies $t \propto \beta^{-1}$, so small β corresponds to a long-time, coarse description and large β to progressively shorter-time, finer structure. FLASHDIFFUSION follows this scale dependence in the coarse-to-fine direction as sampling increases. The ideal resolution scale $\beta^\star(N)$ characterizes the finite-sample optimum, while the algorithm returns an achievable $\beta^\times$ constrained by sampling and arithmetic precision. The accompanying diagnostics n_{eff} and $\widehat{d}$ describe how the effective neighborhood size and local spectral dimension evolve along the same flow.

Empirically, FLASHDIFFUSION reaches regimes in which both sample size and spectral rank are large. On T^6, the method resolves the known first-harmonic eigenspace to approximately 10^{-4} at $N = 2^{22}$ while avoiding materialization of a dense kernel of roughly 70 TB. On alanine dipeptide, 256 diffusion modes recover the backbone dihedrals to below 5° mean absolute error on a fully held-out trajectory, while the 256th eigenvalue remains approximately 21% of the leading eigenvalue. These results illustrate the regime motivating the method: the useful spectrum can remain broad well beyond the ranks where aggressive compression would be appropriate.

Future work includes theoretical guarantees relating the empirical potential minimum to $\beta^\star$, joint calibration of feature-space compression and diffusion resolution, extensions to other radial kernel families, and applications where high-rank spectral structure is expected to be essential, e.g. in molecular dynamics, Gowers et al. (2016). FLASHDIFFUSION therefore re-frames diffusion maps as a high-rank spectral method whose geometry, resolution scale, and computational cost can be treated separately: the representation specifies the geometry, the β-flow measures how finely that geometry is resolved, and tiled matrix-free eigensolvers make the resulting dense harmonic analysis computationally accessible.

REFERENCES

James Baglama, Daniela Calvetti, and Lothar Reichel. IRBL: An implicitly restarted block-lanczos method for large-scale hermitian eigenproblems. *SIAM J. Sci. Comput.*, 24(5):1650–1677, 2003.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003. doi: 10.1162/089976603321780317.

Yoshua Bengio, Jean-François Paiement, Pascal Vincent, Olivier Delalleau, Nicolas Le Roux, and Marie Ouimet. Out-of-sample extensions for LLE, Isomap, MDS, eigenmaps, and spectral clustering. In *NeurIPS*, pp. 177–184, 2003.

Christos Boutsidis, Michael W. Mahoney, and Petros Drineas. An improved approximation algorithm for the column subset selection problem. In *SODA*, pp. 968–977, 2009. arXiv:0812.4293.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.

Jeff Calder and Nicolás García Trillos. Improved spectral convergence rates for graph laplacians on ε-graphs and k-NN graphs. *ACHA*, 60:123–175, 2022.

Julio Candanedo. Diffusion map autoencoder. *arXiv:2409.05901*, 2024.

Julio Candanedo. E(3) invariant representations of biomolecules. *bioRxiv:2025.03.26.645520*, 2025a.

Julio Candanedo. Linearized diffusion map. *arXiv:2507.14257*, 2025b.

Julio Candanedo. The diffusion attention connection. *arXiv:2604.09560*, 2026.

Benjamin Charlier, Jean Feydy, Joan Alexis Glaunès, François-David Collin, and Ghislain Durif. Kernel operations on the GPU, with autodiff, without memory overflows. *J. Mach. Learn. Res.*, 22(74):1–6, 2021. arXiv:2004.11127; GitHub:getkeops/keops.

Giuseppe Clemente and Massimo D'Elia. Spectrum of the laplace-beltrami operator and the phase structure of causal dynamical triangulation. *Phys. Rev. D*, 97:124022, 2018. arXiv:1804.02294.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *ACHA*, 21(1):5–30, 2006.

Ronald R. Coifman, Yoel Shkolnisky, Fred J. Sigworth, and Amit Singer. Graph laplacian tomography from unknown random projections. *IEEE Trans. Image Process.*, 17(10):1891–1899, 2008.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS*, 2022. arXiv:2205.14135.

Petros Drineas, Michael W. Mahoney, and S. Muthukrishnan. Relative-error CUR matrix decompositions. *SIAM J. Matrix Anal. Appl.*, 30(2):844–881, 2008.

Jean Feydy, Joan Alexis Glaunès, Benjamin Charlier, and Michael M. Bronstein. Fast geometric learning with symbolic matrices. In *NeurIPS*, volume 33, 2020. GitHub:getkeops/keops.

Richard J. Gowers, Max Linke, Jonathan Barnoud, Tyler J. E. Reddy, Manuel N. Melo, Sean L. Seyler, Jan Dotson, Jan Domanski, Sébastien Buchoux, Ian M. Kenney, and Oliver Beckstein. MDAnalysis: A python package for the rapid analysis of molecular dynamics simulations. In *SciPy*, pp. 98–105, 2016.

Leslie Greengard and John Strain. The fast gauss transform. *SIAM J. Sci. Stat. Comput.*, 12(1): 79–94, 1991.

Matthias Hein, Jean-Yves Audibert, and Ulrike von Luxburg. Graph laplacians and their convergence on random neighborhood graphs. *J. Mach. Learn. Res.*, 8:1325–1368, 2007. arXiv:math/0608522.

Johannes Hertrich. Fast kernel summation in high dimensions via slicing and fourier transforms. *SIAM J. Math. Data Sci.*, 6(4):1109–1137, 2024. arXiv:2401.08260.

William B. Johnson and Joram Lindenstrauss. Extensions of lipschitz mappings into a hilbert space. In *Contemp. Math.*, volume 26, pp. 189–206, 1984.

Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *J. Res. Natl. Bur. Stand.*, 45(4):255–282, 1950.

Andrew W. Long and Andrew L. Ferguson. Landmark diffusion maps (L-dMaps): Accelerated manifold learning out-of-sample extension. *ACHA*, 47(1):190–211, 2019. arXiv:1706.09396.

Yu. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(4):824–836, 2020. arXiv:1603.09320.

James McQueen, Marina Meila, Jacob VanderPlas, and Zhongyue Zhang. megaman: Manifold learning with millions of points. *arXiv:1603.02763*, 2016.

Marius Muja and David G. Lowe. Scalable nearest neighbor algorithms for high dimensional data. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(11):2227–2240, 2014.

Feliks Nüske, Hao Wu, Jan-Hendrik Prinz, Christoph Wehmeyer, Cecilia Clementi, and Frank Noé. Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias. *J. Chem. Phys.*, 146(9):094104, 2017. doi: 10.1063/1.4976518. arXiv:1701.01665.

NVIDIA. NVIDIA A100 Tensor Core GPU Architecture, 2020. NVIDIA architecture whitepaper.

NVIDIA. NVIDIA Blackwell Architecture Technical Brief, 2024. NVIDIA technical brief.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pp. 8024–8035, 2019.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *NeurIPS*, pp. 1177–1184, 2007.

Ali Rahimi and Benjamin Recht. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *NeurIPS*, pp. 1313–1320, 2008.

Vikas C. Raykar and Ramani Duraiswami. The improved fast gauss transform with applications to machine learning. In *NIPS Workshop on Large Scale Kernel Machines*, 2005.

Amit Singer. From graph to manifold laplacian: The convergence rate. *ACHA*, 21(1):128–134, 2006.

Danny C. Sorensen. Implicit application of polynomial filters in a k-step arnoldi method. *SIAM J. Matrix Anal. Appl.*, 13(1):357–385, 1992.

Vijay Thakkar et al. CUTLASS 3.0, 2023. CUDA Templates for Linear Algebra Subroutines.

Daniel Ting, Ling Huang, and Michael I. Jordan. An analysis of the convergence of graph laplacians. In *ICML*, pp. 1079–1086, 2010. arXiv:1101.5435.

Christoph Wehmeyer and Frank Noé. Time-lagged autoencoders: Deep learning of slow collective variables for molecular kinetics. *J. Chem. Phys.*, 148(24):241703, 2018. doi: 10.1063/1.5011399. arXiv:1710.11239.

Christopher K. I. Williams and Matthias Seeger. Using the nyström method to speed up kernel machines. In *NeurIPS*, pp. 682–688, 2000.

Ted Zadouri, Markus Hoehnerbach, Jay Shah, Timmy Liu, Vijay Thakkar, and Tri Dao. Flashattention-4: Algorithm and kernel pipelining co-design for asymmetric hardware scaling. In *MLSys*, 2026. arXiv:2603.05451; GitHub:Dao-AILab/flash-attention.

Lihi Zelnik-Manor and Pietro Perona. Self-tuning spectral clustering. In *NeurIPS*, pp. 1601–1608, 2004.

Jean Zinn-Justin. *From Random Walks to Random Matrices: Selected Topics in Modern Theoretical Physics*. Oxford Graduate Texts. Oxford University Press, 2019. ISBN 9780198787754.

A SPECTRAL CONSEQUENCES OF KNN/ANN SPARSIFICATION

A common strategy for scaling diffusion maps is k-nearest-neighbor (kNN) sparsification: retain a fixed number of neighbors per sample, store the resulting weighted graph, and apply a sparse eigensolver. This strategy is representative of many large-scale spectral-embedding pipelines that combine approximate neighborhood graphs with iterative solvers, such as MEGAMAN (McQueen et al., 2016). The comparison becomes more consequential when the objective is to recover a substantial spectral basis instead of only a low-dimensional visualization. The applications considered here use on the order of $M \sim 10^2$ modes. As discussed in the main text, at the bandwidths relevant to this regime the Gaussian kernel generally has substantial effective rank, which limits the effectiveness of very low-rank approximations. The matrix-free formulation therefore preserves the dense Gaussian operator while avoiding both explicit matrix construction and a fixed sparse neighborhood representation.

The first few eigenfunctions typically vary on spatial scales comparable to the manifold diameter and can therefore be comparatively insensitive to moderate changes in local graph construction. Higher eigenfunctions vary on progressively shorter scales. Weyl counting gives a characteristic wavelength of order $(V/M)^{1/d}$ for the M-th mode on a d-dimensional manifold of volume V. Consequently, a neighborhood graph that gives a visually faithful two- or three-dimensional embedding may preserve substantially less information than is required for a larger spectral basis. We distinguish two questions below. First, how many neighbors are required for a truncated graph to reproduce the numerically relevant support of the dense Gaussian operator at a prescribed tolerance? Second, what structural changes does hard neighborhood truncation introduce when the retained neighborhood is much smaller than this support scale? The first question leads to a small-ball scaling law, while the second is addressed through retained row mass, positive semidefiniteness, and finite-sample spectral considerations.

A.1 ARITHMETIC AND NEIGHBORHOOD TRUNCATION

Consider the dense Gaussian kernel $\mathsf{K}_{ij}(\beta) = \exp\left(-\beta\|x_i - x_j\|^2\right)$. Although every entry is strictly positive in exact arithmetic, numerical computation introduces a practical relative scale below which individual contributions become increasingly insignificant compared with $O(1)$ accumulated quantities. We therefore introduce a working tolerance τ and regard interactions satisfying $\mathsf{K}_{ij}(\beta) < \tau$ as negligible at the requested arithmetic precision. We use $\tau \approx 10^{-7}$ as an `fp32`-scale working tolerance, comparable to the relative machine-epsilon scale, and $\tau \approx 10^{-15}$ as an `fp64`-scale tolerance. The resulting estimates depend only logarithmically on τ. The corresponding support radius is:

$$D_\tau = \sqrt{\frac{\log(1/\tau)}{\beta}}. \tag{16}$$

Interactions outside a ball of radius D_τ therefore have weight below the prescribed working tolerance. A kNN construction imposes a different truncation at the k-th-neighbor distance $R_k(x_i)$. This radius is row dependent and varies with the local sampling density $q(x_i)$. For the quantitative estimates below we first consider uniform sampling, $q(x) \equiv q = 1/V$, and let ω_d denote the volume of the unit d-ball. Locally, k neighbors occupy a volume of order $k/(Nq)$. For fixed β, let $\widetilde{\mathsf{K}}_{\beta,k} = \mathsf{K}_\beta \odot A_k$ denote the symmetrically masked kernel, with A_k the neighborhood mask. Reproducing all dense-kernel interactions above tolerance τ requires approximately $R_k(x_i) \gtrsim D_\tau$ for every row. Under nonuniform sampling, this condition becomes spatially heterogeneous, so a single fixed k generally corresponds to different effective support tolerances across the manifold. This distinction is also relevant to bandwidth selection. The neighborhood size required to reproduce the dense operator at tolerance τ depends on β, while the preferred bandwidth $\beta^\star$ is determined by the spectral calculation. **A fixed neighborhood graph therefore selects a truncation scale before the bandwidth of interest is known.** FlashDiffusion evaluates the Gaussian operator directly throughout the β continuation, so the same pairwise interaction model is used during bandwidth selection and eigenspace computation.

A.2 Two scales: spectral resolution and kernel support

Two distinct length scales govern the comparison. The first is the Gaussian smoothing scale $\ell_\beta \sim \beta^{-1/2}$. By Weyl counting, the number of spatial modes supported at this scale behaves as $M(\beta) \sim V\beta^{d/2}$ up to geometric constants. Resolving M modes therefore corresponds to the scaling:

$$\beta \gtrsim \left(\frac{M}{V}\right)^{2/d}. \tag{17}$$

The second is the support scale D_τ in eq. 16. Matching the dense Gaussian kernel above the working tolerance τ requires approximately

$$k^*(\beta,\tau) = \frac{N\omega_d}{V}\left(\frac{\log(1/\tau)}{\beta}\right)^{d/2}. \tag{18}$$

The quantity k^* is therefore a support-fidelity scale: it is the expected number of kernel interactions per row whose weights exceed the chosen tolerance. For the Swiss roll at $N = 10^6$, $\beta^\star = 15$, and $d = 2$, this estimate gives $k^* \approx 8 \times 10^4$ for $\tau = 10^{-7}$ and $k^* \approx 1.7 \times 10^5$ for $\tau = 10^{-15}$. The smaller `fp64`-scale tolerance includes a wider range of weak interactions and therefore increases the support count. These values are much larger than neighborhood sizes such as $k \in 32, 64, 128$ commonly used to construct sparse embedding graphs. The comparison establishes that a small fixed-k graph is a strong truncation of the finite-sample Gaussian operator at $\beta^\star$. Preservation of a particular leading eigenspace can still occur under stronger truncation because spectral accuracy also depends on the structure of the omitted interactions and on the relevant eigengaps. Along a calibrated bandwidth path, two reference exponents are useful. Balancing bias against pointwise fluctuation gives Singer's scaling $\beta^\star \sim N^{2/(d+6)}$ (Singer, 2006), while balancing against the spectral fluctuation $(N\varepsilon^{d/2})^{-1/2}$ gives $\beta^\star \sim N^{2/(d+4)}$, the corresponding rate for a spectral selection criterion (Calder & García Trillos, 2022). On the Swiss roll, the measured bandwidth selection follows $\beta^\star \sim N^{1/3}$ for $d = 2$ over the tested range. Substitution into eq. 18 gives $k^* \sim N^{4/(d+4)}$, or $k^* \sim N^{2/3}$ for $d = 2$, consistent with the measured support counts. Combining eq. 17 with eq. 18 gives a useful scaling relation. At the bandwidth associated with M resolved modes, preserving the dense-kernel support above tolerance τ requires

$$k \gtrsim \frac{N\log(1/\tau)^{d/2}}{M}. \tag{19}$$

Order-one geometric constants are omitted in eq. 19. For $d = 2$ and $N = 10^6$, this support-matching scale is approximately 10^5 neighbors for $M = 100$ and 10^4 neighbors for $M = 1000$ at the working tolerances considered here. A fixed graph with tens or hundreds of neighbors therefore corresponds to a substantially stronger truncation than retaining the numerically relevant support of the dense kernel. The same relation can be viewed at fixed k. For $d = 2$, the support count scales as β^{-1}. Using the measured Swiss-roll reference point, $k^* \approx 8 \times 10^4$ at $\beta^\star = 15$, a graph with $k = 64$ reaches a comparable support radius around:

$$\beta_{\text{support}} \approx 15\left(\frac{8\times 10^4}{64}\right) \approx 2 \times 10^4. \tag{20}$$

This value is roughly three orders of magnitude larger than the selected $\beta^\star$. A $k = 64$ graph therefore corresponds to a much shorter geometric interaction scale than the dense Gaussian operator selected by the spectral criterion.

A.3 Numerically relevant support and retained row mass

The support calculation above follows directly from the small-distance distribution of points on a smooth manifold. Define $\rho_\tau(\beta) = \Pr[\mathsf{K}_{ij}(\beta) \geq \tau]$. Equivalently, $\rho_\tau(\beta) = \Pr[d^2 \leq \log(1/\tau)/\beta]$. If F denotes the CDF of pairwise squared distances, then $\rho_\tau(\beta) = F(\log(1/\tau)/\beta)$. While the bulk of F depends on global manifold geometry, its small-distance behavior is local. For a smooth d-manifold of volume V under approximately uniform sampling, the interior small-ball law gives $\Pr[d \leq R] = (\omega_d R^d/V)(1 + \mathcal{O}(R^2))$. Hence

$$\rho_\tau(\beta) = \frac{\omega_d}{V}\left(\frac{\log(1/\tau)}{\beta}\right)^{d/2}\left(1 + \mathcal{O}(D_\tau^2)\right). \tag{21}$$

k	retained f_k	$1 - e^{-c_2 k}$
32	2.2%	2.18%
64	4.4%	4.31%
128	8.6%	8.43%
256	16.4%	16.2%
1024	50.4%	50.7%

Table 1: Gaussian kernel row mass retained under fixed-k truncation on a unit-variance Swiss roll with $N = 2 \times 10^5$ and $\beta^\star = 8.75$, averaged over 400 rows. The local $d = 2$ prediction $1 - e^{-c_2 k}$ uses $c_2 \approx 6.9 \times 10^{-4}$.

The expected support count per row is $n_\tau(\beta) = N\rho_\tau(\beta)$, and eq. 18 follows directly from this relation. Along the spectral scaling $\beta^\star \sim N^{2/(d+4)}$, the relative support density decreases with sample size while the absolute number of relevant interactions grows as $n_\tau(\beta^\star) \sim N^{4/(d+4)}$. The Gaussian operator therefore becomes increasingly localized in relative density while the number of interactions per row associated with a fixed relative tolerance continues to increase. We verified the $d = 2$ scaling on a unit-variance Swiss roll using $\tau = 10^{-7}$. The prediction is $\rho_\tau(\beta) \propto \beta^{-1}$. Across $N = 10^4$–10^8, the measured product $\rho_\tau(\beta^\star)\beta^\star$ remains approximately constant at 1.19, within 1% of $\omega_2 \log(1/\tau)/V$. Over the same range, the selected bandwidth follows $\beta^\star \sim N^{1/3}$ and the measured support count increases from approximately 3.7×10^3 to 1.7×10^6 interactions per row, consistent with $N^{2/3}$ scaling, while the relative support density decreases from 37% to 1.7%. Curvature corrections enter at relative order $\mathcal{O}(D_\tau^2)$ in the interior, while boundary effects occur within a strip of width $\mathcal{O}(D_\tau)$. The observed agreement over the tested range indicates that the leading small-ball scaling captures the dominant dependence relevant to these support estimates. Support counts depend on the chosen tolerance τ. A complementary quantity with no dependence on τ is the fraction of Gaussian row mass retained by the kNN mask. Because the Gaussian weights are nonnegative, define

$$f_k = \frac{\sum_{j \in \mathcal{N}_k(i)} \exp(-\beta d_{ij}^2)}{\sum_j \exp(-\beta d_{ij}^2)}. \tag{22}$$

Under the local uniform-density approximation,

$$f_k = \frac{\gamma(d/2, \beta R_k^2)}{\Gamma(d/2)}. \tag{23}$$

Since $R_k \sim (k/(Nq\omega_d))^{1/d}$, the argument scales as $\beta R_k^2 \propto \beta(k/N)^{2/d}$. Along $\beta^\star \sim N^{2/(d+4)}$, fixed k therefore gives $f_k \sim kN^{-4/(d+4)}$ in the small-argument regime. For $d = 2$, eq. 23 reduces locally to $f_k = 1 - e^{-c_2 k}$, where c_2 contains the sampling-density and bandwidth factors. This relation is geometric and independent of the working tolerance τ. Table 1 verifies the prediction on the Swiss roll. A fitted $c_2 \approx 6.9 \times 10^{-4}$ reproduces the measured retained fractions to within approximately 0.3% over the tested range. At this problem size, $k = 64$ retains about 4.4% of the Gaussian row mass at $\beta^\star$. Spectral error depends additionally on the organization of the omitted interactions and on the relevant spectral gaps, so the retained-mass fraction and eigenspace error need not coincide numerically. The small retained fraction nevertheless shows that fixed-k masking at this scale produces a substantial perturbation of the dense operator. This distinction is important for bandwidth selection. FlashDiffusion seeks the largest bandwidth for which a requested spectral subspace remains distinguishable from sampling and numerical fluctuations. Replacing K_β by $\widetilde{\mathsf{K}}_{\beta,k}$ changes the operator, its degree normalization, and its spectral gaps. The sparsified family therefore has its own finite-sample resolution boundary, which can occur at a different β from the corresponding boundary of the dense Gaussian family.

A.4 kNN TRUNCATION AND PSD

The dense Gaussian Gram matrix satisfies $\mathsf{K}_\beta \succeq 0$. For a Hadamard-truncated kernel $\widetilde{\mathsf{K}}_{\beta,k} = \mathsf{K}_\beta \odot A_k$, the Schur product theorem guarantees positive semidefiniteness when $A_k \succeq 0$. A symmetrized kNN adjacency mask generally does not satisfy this condition. Once indefiniteness is introduced,

k	$\lambda_{\min}(\mathsf{K} \odot A_k)$	neg. spectral mass	$\lambda_{\min}(A_k)$
16	-4.10	26.9%	-4.79
32	-5.84	27.8%	-7.85
64	-7.71	25.8%	-13.44
128	-7.58	19.0%	-33.07
256	-2.73	9.0%	-61.03

Table 2: Spectrum of a symmetrically kNN-truncated Gaussian kernel on the Swiss roll with $N = 3000$, $\beta^\star = 2.158$, and neighborhood masks symmetrized by union. Negative spectral mass is $\sum_{\lambda<0} |\lambda| / \sum |\lambda|$. The final column reports the smallest eigenvalue of the mask itself.

positive diagonal congruence preserves its inertia. For any positive diagonal matrix D, Sylvester's law of inertia gives

$$\text{inertia}\,(D(\mathsf{K}_\beta \odot A_k)D) = \text{inertia}\,(\mathsf{K}_\beta \odot A_k)\,. \tag{24}$$

The symmetric positive-diagonal rescalings appearing in diffusion-map normalization therefore preserve any negative eigenvalues introduced by the mask. Table 2 illustrates this effect on the Swiss roll at a problem size for which the complete spectrum can be computed directly. The table demonstrates directly that hard symmetric neighborhood masking can alter a structural property that the dense Gaussian kernel possesses exactly. Its purpose is finite-sample characterization at a size where the complete spectrum is available. PSD also has a natural diffusion interpretation. For a positive eigenvalue one may write $\lambda_m = e^{-t\nu_m}$ with $\nu_m \geq 0$ and interpret ν_m as a real decay rate. Negative eigenvalues introduced by masking correspond to complex values of ν_m under this representation. Portions of the truncated spectrum can therefore lose the real diffusion-semigroup interpretation carried by the dense Gaussian kernel.

A.5 Continuum geometry and ANN perturbations

Neighborhood-graph convergence theory shows that kNN distances estimate a density-dependent local length scale, and that the limiting differential operator depends on the graph construction, weighting, and normalization (Hein et al., 2007; Ting et al., 2010; Calder & García Trillos, 2022). Appropriately density-corrected graph constructions can converge to prescribed continuum operators. At finite N, a kNN-truncated Gaussian matrix and the dense Gaussian matrix at the same nominal bandwidth remain distinct discrete operators. This distinction is important here because FlashDiffusion selects an empirical $\beta^\star$ from finite-sample spectral information. Two graph constructions with the same continuum limit can have different finite-sample spectra and spectral gaps. Spectral convergence results such as Calder & García Trillos (2022) quantify the separation between discrete spectra and their limiting continuum objects. Approximate nearest-neighbor (ANN) methods introduce an additional perturbation. If $\widehat{A}_k$ is the returned approximate mask, then the effective kernel is:

$$\widehat{\mathsf{K}}_{\beta,k} = \mathsf{K}_\beta \odot \widehat{A}_k. \tag{25}$$

Missed or substituted edges modify the weighted degrees, the α-normalization, and ultimately the spectrum. Standard ANN recall (Malkov & Yashunin, 2020) therefore gives only a partial measure of spectral accuracy because the effect of an incorrect edge depends on its Gaussian weight and on the spectral gaps of the modes being estimated (Boutsidis et al., 2009; Drineas et al., 2008). Some ANN implementations can additionally introduce run-to-run variability, depending on their construction and execution settings. The broader issue is the additional finite-sample perturbation introduced by approximate neighborhood construction, whose accuracy must be chosen before or adapted jointly with the bandwidth search. A neighborhood budget large enough to reproduce the dense-kernel support at the as-yet unknown $\beta^\star$ must either be selected conservatively in advance or increased as the bandwidth changes. FlashDiffusion avoids this separate graph-construction approximation by evaluating the Gaussian interactions directly on GPUs during the same continuation used to determine $\beta^\star$.

A.6 SYNTHESIS

The comparison above separates several forms of approximation that affect diffusion-map calculations differently. A visually stable low-dimensional embedding can coexist with substantial changes in a larger spectral basis because the first few eigenvectors encode coarse geometry while higher eigenfunctions probe progressively shorter spatial scales. Matching the dense Gaussian kernel above a relative tolerance τ requires a neighborhood size scaling as:

$$k^*(\beta, \tau) \sim N \left(\frac{\log(1/\tau)}{\beta} \right)^{d/2}. \tag{26}$$

At a bandwidth associated with M resolved modes, this becomes:

$$k^* \sim \frac{N \log(1/\tau)^{d/2}}{M}. \tag{27}$$

These relations characterize support fidelity of the dense finite-sample Gaussian operator. Spectral fidelity can persist under stronger truncation when the omitted interactions have limited influence on the target eigenspace, making direct eigenspace comparisons an additional empirical diagnostic. The row-mass calculation provides a tolerance-independent measure of truncation strength. In the Swiss-roll experiment, $k = 64$ retains 4.4% of the Gaussian row mass at $\beta^\star$, while $k = 1024$ retains approximately half. Hard neighborhood masking can also remove the PSD structure of the Gaussian kernel and alter the deeper portion of its diffusion spectrum.

These observations complement the main-text analysis of low-rank approximations. At bandwidths relevant to resolving many diffusion modes, the Gaussian matrix can simultaneously have substantial effective rank and substantial spatial support. Low-rank compression restricts the number of retained spectral directions, while aggressive fixed-k sparsification removes a large fraction of spatial interactions. Both operations therefore modify the finite-sample Gaussian operator in distinct ways.

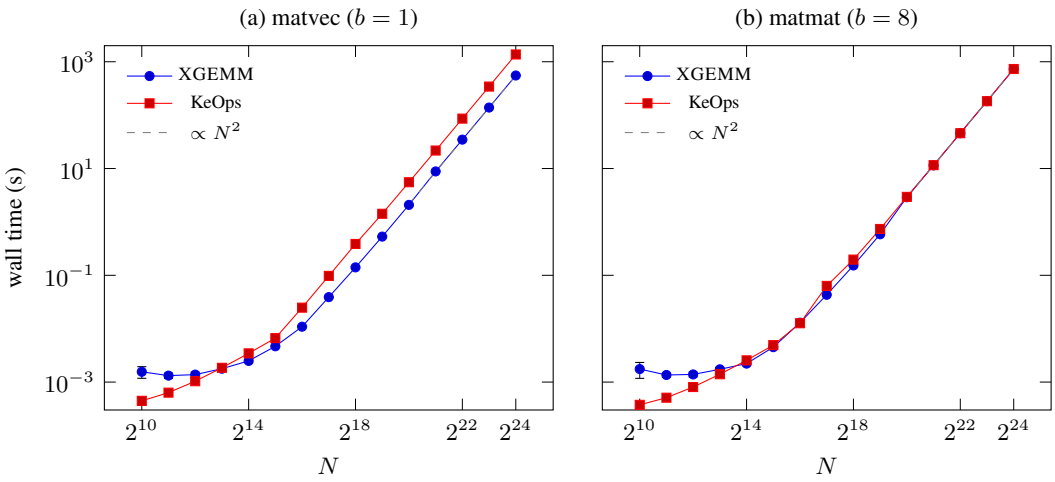


Figure 4: Raw matrix-free Gaussian-kernel scaling on the centered anisotropic flat torus T^6 ($D = 32$, $\beta = 1$) using one RTX–PRO–6000 GPU. **(a)** True matrix–vector application Kv ($b = 1$). **(b)** Eight-right-hand-side matrix–matrix application KV ($b = 8$). XGEMM–TILT and KeOps evaluate the same dense Gaussian interaction without materializing the $N \times N$ kernel matrix. Points show mean wall time and error bars show one standard deviation over repeated timed applications; the dashed guides indicate exact $\mathcal{O}(N^2)$ scaling in the large-N regime.

B KeOps Comparison

Here we demonstrate a comparison of the KeOps library[7], Feydy et al. (2020); Charlier et al. (2021), versus our matrix-vector (matvec) and matrix-matrix (matmat) primitive XGEMM. The results are shown in fig. 4, and both achieve similar results.

[7]Using `pip install pykeops==2.3` (Released: Apr 17, 2025)