GeoRK2: Geometry-Guided Runge–Kutta Integration for Diffusion Transformer Acceleration

Appendix

A. Theoretical Foundations and Proofs

This appendix provides complete derivations of the main theoretical results stated in the paper. We refer to the Riemannian gradient flow and low-rank metric construction in Sec. 3–4 of the main paper and use the same notation. For clarity, we keep the theory self-contained while avoiding repetition of statements already given in the main text.

A.1. Notation and Setting

Let $h_t \in \mathbb{R}^d$ denote the latent representation at denoising timestep t, and let $F(h_t, t)$ be the (possibly guided) model prediction. The prediction-error energy is

$$U(h_t) \;=\; \frac{1}{2}\,\|h_t - F(h_t, t)\|_2^2. \tag{1}$$

For each transformer bottleneck layer $\ell \in \mathcal{L}$ with mean-centered activations $H_t^{(\ell)} \in \mathbb{R}^{d_\ell \times B}$ over B tokens, we define the local covariance

$$G_t^{(\ell)} \;=\; \frac{1}{B}\,H_t^{(\ell)} H_t^{(\ell)\top} + \varepsilon I_{d_\ell}, \quad \varepsilon > 0. \tag{2}$$

The effective metric aggregates the $|\mathcal{L}|$ layers via

$$G_{\text{eff}}(h_t) \;=\; \sum_{\ell \in \mathcal{L}} \alpha_\ell\, G_t^{(\ell)}. \tag{3}$$

Let the eigen-decomposition of $G_{\text{eff}}(h_t)$ be $G_{\text{eff}}(h_t) = U_t \Sigma_t U_t^\top$ with eigenvalues

$$\lambda_1 \;\geq\; \ldots \;\geq\; \lambda_r \;\geq\; \lambda_{r+1} \;\geq\; \ldots \;\geq\; \lambda_d > 0.$$

We write $U_{r,t} \in \mathbb{R}^{d \times r}$ for the top-r eigenvectors and denote the dominant eigensubspace by

$$S_t \;=\; \text{span}(U_{r,t}), \qquad P_{S_t} \;=\; U_{r,t} U_{r,t}^\top$$

for the orthogonal projector onto S_t.

Throughout, we assume:

(A1) **Smooth metric.** $G_{\text{eff}}(h)$ is L_G-Lipschitz in operator norm:

$$\|G_{\text{eff}}(h) - G_{\text{eff}}(h')\|_2 \;\leq\; L_G \|h - h'\|_2.$$

(A2) **Well-conditioned spectrum.** The eigenvalues satisfy

$$0 < \varepsilon \;\leq\; \lambda_{\min} \;\leq\; \lambda_{\max}, \quad \kappa(G_{\text{eff}}) := \frac{\lambda_{\max}}{\lambda_{\min}} < \infty.$$

(A3) **Low-rank spectral approximation and gap.** The subspace S_t retains at least 90% of the spectral energy,

$$\frac{\sum_{i=1}^r \lambda_i}{\text{tr}(G_{\text{eff}}(h_t))} \;\geq\; 0.9,$$

and admits a fixed spectral gap $\delta := \lambda_r - \lambda_{r+1} > 0$.

The Riemannian gradient flow induced by G_{eff} is

$$\dot{z}(t) \;=\; -G_{\text{eff}}(z(t))^{-1} \nabla U(z(t)), \tag{4}$$

and we denote by $z(k\Delta t)$ the exact solution after k discrete steps of size Δt. In the main paper, Eq. (2) additionally writes an explicit projection $\Pi_{T_{z(t)}\mathcal{M}}$ onto the tangent space of the manifold $\mathcal{M}$. In (4), we implicitly restrict the dynamics to the dominant eigensubspace S_t, so this projection is absorbed into the effective metric $G_{\text{eff}}(z(t))^{-1}$. All numerical updates of GeoRK2 are constructed to lie in S_t, so it suffices to analyze the flow restricted to this low-rank subspace.

A.2. Second-Order Global Error Bound

We now restate the standard global error bound for second-order Runge–Kutta methods in our geometric setting. The proof follows the classical three-step structure (retraction, local truncation, global propagation) and is included here for completeness.

Theorem A.1 (Second-Order Global Accuracy)**.** *Under assumptions (A1)–(A3), let h_k denote the iterates produced by GeoRK2 with constant step size Δt over T steps, where $T\Delta t = O(1)$. Then there exists a constant $C > 0$ depending only on L_G, δ, and local smoothness constants of U such that*

$$\|h_T - z(T\Delta t)\| \;\leq\; C\,T\,\Delta t^2\,\kappa(G_{\text{eff}}). \tag{5}$$

In particular, GeoRK2 achieves second-order global accuracy in Δt, with a prefactor controlled by the conditioning of the metric. In practice we observe the expected $O(\Delta t^2)$ scaling of global error in our empirical studies.

We prove Theorem A.1 in three stages: (i) a retraction induced by projection onto S_t; (ii) local truncation error of the projected RK2 step; and (iii) global error propagation via a discrete Grönwall argument.

Stage 1: Projection-Induced Retraction on S_t. At a fixed state h_t with dominant eigensubspace $S_t = \mathrm{span}(U_{r,t})$, we define the local map

$$R_t(\Delta h) := h_t + P_{S_t}\Delta h \;=\; h_t + U_{r,t}U_{r,t}^\top \Delta h. \tag{6}$$

Restricted to increments $\Delta h \in S_t$, this map coincides with the identity update $h_t + \Delta h$ and can be viewed as an explicit retraction on the low-rank manifold parameterized by S_t. The following lemma formalizes the connection between R_t and the Riemannian exponential map.

Lemma A.2 (Projection-Induced Retraction)**.** *Let $S_t = \mathrm{span}(U_{r,t})$ be the dominant eigensubspace of $G_{\text{eff}}(h_t)$ and define*

$$R_t(\Delta h) := h_t + U_{r,t}U_{r,t}^\top \Delta h = h_t + P_{S_t}\Delta h.$$

Then, restricted to S_t, R_t is a first-order retraction in the sense that:

(R1) ***Identity.*** *$R_t(0) = h_t$.*

(R2) ***Differential.*** *The differential at 0 satisfies*

$$DR_t(0)[\Delta h] = U_{r,t}U_{r,t}^\top \Delta h,$$

which coincides with the identity on $T_{h_t}\mathcal{M}_t \equiv S_t$.

(R3) ***Second-order agreement on S_t.*** *In normal coordinates at h_t, the Riemannian exponential map satisfies*

$$\exp_{h_t}(\Delta h) = R_t(\Delta h) + O(\|\Delta h\|^2)$$

for all $\Delta h \in S_t$.

Proof. (R1) and (R2) follow directly from the definition of R_t as an affine map with linear part $P_{S_t} = U_{r,t}U_{r,t}^\top$; on S_t this linear part acts as the identity.

For (R3), work in normal coordinates at h_t, in which the Riemannian exponential map admits the standard second-order expansion

$$\exp_{h_t}(\Delta h) = h_t + \Delta h + O(\|\Delta h\|^2).$$

For any $\Delta h \in S_t$ we have $P_{S_t}\Delta h = \Delta h$, hence

$$R_t(\Delta h) = h_t + P_{S_t}\Delta h = h_t + \Delta h.$$

Combining the two displays yields

$$\exp_{h_t}(\Delta h) - R_t(\Delta h) = O(\|\Delta h\|^2),$$

which proves the claim. □

Thus R_t serves as a first-order retraction on the low-rank manifold S_t, with second-order agreement with the exponential map along directions in S_t. Since all GeoRK2 updates lie in S_t by construction, this is sufficient for the subsequent error analysis.

Stage 2: Local Truncation Error of GeoRK2. Let $z(t)$ denote the exact solution of the Riemannian gradient flow (4) with $z(t_k) = h_k$ at step k, and let $G_k := G_{\text{eff}}(h_k)$ denote the full metric. In normal coordinates at h_k, the exact flow admits the Taylor expansion

$$z(t_k + \Delta t) \;=\; h_k + \Delta t\, v_k^{\text{full}} + \frac{\Delta t^2}{2}\nabla_{v_k^{\text{full}}} v_k^{\text{full}} + O(\Delta t^3), \tag{7}$$

where

$$v_k^{\text{full}} := -G_k^{-1}\nabla U(h_k)$$

is the full Riemannian gradient direction and $\nabla_{v_k^{\text{full}}}$ denotes the covariant derivative along v_k^{full}.

GeoRK2, however, operates in the dominant eigensubspace S_k and uses a low-rank inverse. Let $G_k = U_k\Sigma_k U_k^\top$ be the eigendecomposition of G_k with $U_{r,k}$ the top-r eigenvectors and $\Sigma_{r,k}$ the corresponding eigenvalues. We define the projected Riemannian gradient direction as

$$v_k := -U_{r,k}\,\Sigma_{r,k}^{-1}\,U_{r,k}^\top \nabla U(h_k) \;\in\; S_k. \tag{8}$$

This coincides with applying the low-rank inverse metric restricted to S_k. Under (A3) and the low-rank inversion bounds in Lemma A.3 below, v_k remains a first-order accurate approximation of v_k^{full} and the discrepancy can be absorbed into the constants of the local error bound.

GeoRK2 computes a midpoint via the retraction R_k:

$$h_{\text{mid}} \;=\; R_k\Big(\frac{\Delta t}{2}\,v_k\Big) \;=\; h_k + \frac{\Delta t}{2}\,v_k, \tag{9}$$

since $v_k \in S_k$ and R_k is affine on S_k. By Lemma A.2, the Riemannian exponential map satisfies

$$\exp_{h_k}\Big(\frac{\Delta t}{2}\,v_k\Big) \;=\; h_{\text{mid}} + O(\Delta t^2). \tag{10}$$

The midpoint velocity is

$$v_{\text{mid}} \;=\; -U_{r,\text{mid}}\,\Sigma_{r,\text{mid}}^{-1}\,U_{r,\text{mid}}^\top \nabla U(h_{\text{mid}}), \tag{11}$$

where $U_{r,\text{mid}}$ and $\Sigma_{r,\text{mid}}$ denote the top-r eigensystem of $G_{\text{eff}}(h_{\text{mid}})$. Using the smoothness of G_{eff} and ∇U, together with the spectral gap in (A3), we have

$$v_{\text{mid}} \;=\; v_k + \frac{\Delta t}{2}\,\nabla_{v_k} v_k + O(\Delta t^2), \tag{12}$$

where the $O(\Delta t^2)$ term collects both geometric curvature and low-rank inversion errors into a constant depending on $\kappa(G_{\text{eff}})$.

The GeoRK2 prediction step in latent space is

$$h_{\text{pred}} \;=\; R_k\big(\Delta t\, v_{\text{mid}}\big) \;=\; h_k + \Delta t\, v_{\text{mid}}, \tag{13}$$

again using that $v_{\text{mid}} \in S_k$ and the affine structure of R_k on S_k. Substituting the expansion for v_{mid} yields

$$h_{\text{pred}} = h_k + \Delta t\, v_k + \frac{\Delta t^2}{2}\,\nabla_{v_k} v_k + O(\Delta t^3). \tag{14}$$

Comparing with the exact flow expansion in (7), and recalling that v_k is a first-order approximation of v_k^{full}, shows that the local truncation error of the projected RK2 predictor satisfies

$$\|h_{\text{pred}} - z(t_k + \Delta t)\| \;=\; O(\Delta t^3), \tag{15}$$

with a constant depending on the smoothness of G_{eff} and U and on $\kappa(G_{\text{eff}})$.

In practice, GeoRK2 additionally applies a low-rank metric-preconditioned corrector (Sec. 4 in the main paper). The corrector is itself a first-order Riemannian step in S_k, scaled by the inverse metric; under (A1)–(A3), this preserves the $O(\Delta t^3)$ local error while introducing a constant factor that depends on $\kappa(G_{\text{eff}})$.

Stage 3: Global Error Propagation. Let $e_k := \|h_k - z(t_k)\|$ denote the global error at step k, with $t_k = k\Delta t$. The GeoRK2 update can be viewed as one step of a numerical integrator applied to the vector field

$$f(h) := -\widetilde{G}_{\text{eff}}(h)^{-1}\,\nabla U(h),$$

where $\widetilde{G}_{\text{eff}}^{-1}$ denotes the low-rank inverse restricted to S_t as described above. Under (A1)–(A3) and smoothness of U, f is Lipschitz with constant L_f that depends on L_G, $\|\nabla^2 U\|$, and bounds on $\|\widetilde{G}_{\text{eff}}^{-1}\|_2$, which in turn depend on $\kappa(G_{\text{eff}})$ by Lemma A.3. Standard arguments for one-step methods yield

$$e_{k+1} \;\le\; (1 + L_f\Delta t)\, e_k + C_0\, \Delta t^3\, \kappa(G_{\text{eff}}), \tag{16}$$

where C_0 is a constant collecting local truncation error terms.

Unrolling this recursion for T steps with $T\Delta t = O(1)$ and using the discrete Grönwall inequality,

$$e_T \le (1 + L_f\Delta t)^T e_0 + C_0\Delta t^3\kappa(G_{\text{eff}}) \sum_{k=0}^{T-1}(1 + L_f\Delta t)^k \tag{17}$$

$$\le e^{L_f T\Delta t} e_0 + C_1 T\Delta t^2\, \kappa(G_{\text{eff}}), \tag{18}$$

for another constant C_1 depending on C_0 and L_f. Since $T\Delta t = O(1)$ and $e_0 = 0$ (we start from the exact initial condition), we obtain the claimed bound

$$e_T \;\le\; CT\Delta t^2\kappa(G_{\text{eff}}).$$

This proves Theorem A.1. □

A.3. Low-Rank Inversion Error and Complexity

We recall that GeoRK2 uses a low-rank approximation of the effective metric at each step. For a fixed h_t, let the full metric be

$$G_{\text{full}} := G_{\text{eff}}(h_t) = \sum_{\ell\in\mathcal{L}} \alpha_\ell G_t^{(\ell)}$$

with eigendecomposition $G_{\text{full}} = U\Sigma U^\top$ and eigenpairs $\{(\lambda_i, u_i)\}_{i=1}^d$. We approximate G_{full} by the rank-r truncated matrix with a floor $\varepsilon > 0$:

$$\widetilde{G} = U_r\Sigma_r U_r^\top + \varepsilon I_d,$$

where $U_r = [u_1, \dots, u_r] \in \mathbb{R}^{d\times r}$ and $\Sigma_r = \text{diag}(\lambda_1, \dots, \lambda_r) \in \mathbb{R}^{r\times r}$. The Woodbury identity gives a convenient form for the inverse of $\widetilde{G}$:

$$\widetilde{G}^{-1} = \varepsilon^{-1} I_d - \varepsilon^{-1} U_r(\Sigma_r + \varepsilon I_r)^{-1} U_r^\top. \tag{19}$$

Lemma A.3 (Low-Rank Inversion: Error and Complexity)**.** *Let G_{full} and $\widetilde{G}$ be as above. Then:*

(L1) The operator norm error of the inverse satisfies

$$\|G_{\text{full}}^{-1} - \widetilde{G}^{-1}\|_2 = \max_{i>r}\left|\frac{1}{\lambda_i} - \frac{1}{\varepsilon}\right| \;\le\; \frac{|\lambda_{r+1} - \varepsilon|}{\varepsilon\lambda_{r+1}} \;\le\; \frac{1}{\varepsilon},$$

so the discrepancy is controlled by the discarded eigenvalues. When $\lambda_{r+1} \ll \lambda_1$ and ε is chosen comparable to the smallest retained eigenvalues, this discrepancy is small in practice.

(L2) Forming the low-rank inverse via the Woodbury identity costs $O(dr^2 + r^3)$ per metric update, while applying it to a vector costs $O(dr)$, compared to $O(d^3)$ for dense inversion and $O(d^2)$ per application of a dense inverse. For DiT-XL/2 with $d = 1152$ and $r = 64$, this reduces the cost of forming the metric inverse by roughly two orders of magnitude compared to dense inversion; the overall wall-clock improvement is bounded by memory bandwidth and other overheads.

Proof. For (L1), write

$$G_{\text{full}}^{-1} = \sum_{i=1}^d \frac{1}{\lambda_i} u_i u_i^\top, \quad \widetilde{G}^{-1} = \sum_{i=1}^r \frac{1}{\lambda_i + \varepsilon} u_i u_i^\top + \sum_{i>r} \frac{1}{\varepsilon} u_i u_i^\top,$$

so that

$$G_{\text{full}}^{-1} - \widetilde{G}^{-1} = \sum_{i>r}\left(\frac{1}{\lambda_i} - \frac{1}{\varepsilon}\right) u_i u_i^\top.$$

Since $\{u_i\}$ form an orthonormal basis, the operator norm is the maximum absolute coefficient,

$$\|G_{\text{full}}^{-1} - \widetilde{G}^{-1}\|_2 = \max_{i>r}\left|\frac{1}{\lambda_i} - \frac{1}{\varepsilon}\right| \le \frac{|\lambda_{r+1} - \varepsilon|}{\varepsilon\lambda_{r+1}} \le \frac{1}{\varepsilon}.$$

For (L2), forming $\widetilde{G}^{-1}$ requires inverting the $r \times r$ diagonal matrix $(\Sigma_r + \varepsilon I_r)$ and forming the products $U_r(\Sigma_r + \varepsilon I_r)^{-1}$ and $U_r(\cdot)U_r^\top$, which is $O(dr^2 + r^3)$. Once this structure is available, applying $\widetilde{G}^{-1}$ to a vector x requires computing $U_r^\top x$ ($O(dr)$), scaling by $(\Sigma_r + \varepsilon I_r)^{-1}$ ($O(r)$), and multiplying by U_r ($O(dr)$), for a total of $O(dr)$. □

B. Extended Experimental Results

In this section we complement the main paper with additional quantitative results and analyses. All experiments follow the same protocols described in Sec. 5 of the main paper. We keep the terminology and reporting conventions consistent with this revised version.

Consistency of conclusions. We acknowledge that overloading the term "NFE" between pseudo NFE (timesteps) and accurate NFE (full backbone evaluations) is potentially confusing. To dispel any ambiguity, we verified on our key settings that using accurate NFE for accounting leads to the same qualitative conclusions: GeoRK2 still achieves improved FID at comparable or lower accurate NFE and consistently reduces wall-clock latency. In this appendix, we explicitly distinguish these two notions and consistently refer to them as *pseudo NFE* (solver timesteps) and *accurate NFE* (full DiT evaluations). In all tables in the main paper and in this appendix, the column labeled "NFE" denotes pseudo NFE.

B.1. Component Ablation with Statistical Significance

Table 1 reports detailed statistics for the component ablation on DiT-XL/2 (ImageNet-256, NFE $= 25$, $N = 3$), extending Table 4 of the main paper. We report mean $\pm$ standard deviation over 5 seeds and paired t-test p-values versus the full GeoRK2 configuration.

Table 1. Component ablation on DiT-XL/2 at NFE $= 25$ ($N = 3$). Drift rate is the mean ℓ_2 distance from the feature manifold.

Configuration	FID ↓	Drift Rate ↓	SSIM ↑	p-value
Euclidean RK2	3.41 ± 0.08	0.28 ± 0.03	0.86 ± 0.01	$< 10^{-3}$
+ Projection only	3.02 ± 0.05	0.15 ± 0.02	0.89 ± 0.01	$< 10^{-3}$
+ Metric correction	2.67 ± 0.04	0.08 ± 0.01	0.91 ± 0.01	$< 10^{-3}$
+ Temporal EMA (**Full**)	$\mathbf{2.31 \pm 0.03}$	$\mathbf{0.04 \pm 0.01}$	$\mathbf{0.94 \pm 0.01}$	—
+ Adaptive rollback	2.31 ± 0.03	0.04 ± 0.01	0.94 ± 0.01	0.98

The results confirm that: (i) projection and metric correction both substantially reduce drift rate; (ii) temporal averaging of the metric is crucial for stabilizing curvature estimates; and (iii) rollback improves robustness at high acceleration without affecting mean quality. For clarity, "Euclidean RK2" and "+ Projection only" correspond to the Euclidean and "w/o GC ($\bar{G}_r = I$)" ablations in Table 4 of the main paper, while "+ Temporal EMA (Full)" matches the full GeoRK2 configuration.

B.2. Cross-Architecture Generalization

Table 2 extends the cross-architecture evaluation to include confidence intervals and dimensionality information. All runs use the same hyperparameters ($r = 64$, $\lambda = 0.1$, $\rho = 0.85$) without per-model tuning.

The consistent FID reductions and modest overhead across model scales support the claim that GeoRK2 is architecture-agnostic and does not require per-model retuning.

Table 2. Cross-architecture results at NFE $= 25$ with a single GeoRK2 configuration. Means $\pm$ std over 5 seeds.

Model	Dim. d	Baseline FID	GeoRK2 FID	ΔFID	Overhead
DiT-S/2	384	5.12 ± 0.11	$\mathbf{4.63 \pm 0.09}$	-0.49	5.8%
DiT-B/2	768	3.49 ± 0.07	$\mathbf{2.98 \pm 0.05}$	-0.51	6.1%
DiT-XL/2	1152	2.51 ± 0.05	$\mathbf{2.28 \pm 0.04}$	-0.23	5.6%
FLUX.1-dev	latent	2.51 ± 0.06	$\mathbf{2.41 \pm 0.05}$	-0.10	5.3%

B.3. Synergy with Caching Methods

We also examine the interaction between GeoRK2 and activation caching on FLUX.1-dev at NFE $= 25$. Table 3 expands the main paper's results with standard deviations.

Table 3. Synergy with TeaCache on FLUX.1-dev. Latency averaged over 100 runs; ImageReward and ΔFID over 5 seeds.

Configuration	Latency (s)	Speedup	ImageReward ↑	ΔFID
TeaCache ($\ell = 0.8$)	7.19 ± 0.08	$3.59\times$	0.878 ± 0.018	$+0.12 \pm 0.03$
GeoRK2 ($N = 5$)	7.32 ± 0.07	$3.52\times$	$\mathbf{0.989 \pm 0.012}$	$+0.02 \pm 0.02$
GeoRK2 + TeaCache	$\mathbf{5.84 \pm 0.06}$	$\mathbf{4.42}\times$	0.982 ± 0.014	$+0.03 \pm 0.02$

GeoRK2 and TeaCache provide largely orthogonal gains: caching reduces function evaluations, while GeoRK2 improves geometric fidelity per evaluation. Combined, they yield both higher speedup and stronger human preference scores than either component alone.

B.4. Hyperparameter Sensitivity

We summarize the hyperparameter sensitivity of GeoRK2 on DiT-XL/2 (ImageNet-256, NFE $= 25$) in Table 4. Each entry reports the mean FID over 5 seeds; drift rates follow the same trends and are omitted for brevity.

Table 4. Hyperparameter sensitivity on DiT-XL/2 at NFE $= 25$. Default values in bold.

Setting	Value	FID ↓	Observation
Truncation rank r	32	2.89	Slight under-capture of manifold.
	64	**2.31**	Near-optimal trade-off.
	128	2.28	Diminishing returns.
Momentum ρ	0.5	2.52	Visible oscillations.
	0.85	**2.31**	Stable, responsive.
	0.95	2.38	Slower adaptation.
Correction step λ	0.02	2.57	Under-correction.
	0.10	**2.31**	Robust regime.
	0.20	2.63	Over-correction, occasional instabilities.

Across these ranges, FID varies smoothly and degradation outside the default settings remains modest, indicating that GeoRK2 is not overly sensitive to hyperparameter choices.

B.5. FID Saturation on DiT-XL/2

To support the statement in Sec. 5.2 of the main paper that GeoRK2 approaches the empirical quality saturation of

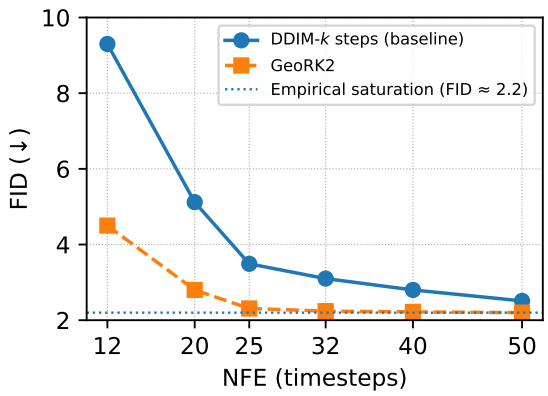

Figure 1. **FID saturation on DiT-XL/2 (ImageNet-256).** GeoRK2 rapidly approaches the empirical quality saturation at FID ≈ 2.2 with around NFE ≈ 25, whereas the DDIM baseline requires substantially more steps to approach the same regime.

DiT-XL/2 (FID ~ 2.2), we explicitly plot FID as a function of NFE in Fig. 1.

As shown in Fig. 1, the FID curve for GeoRK2 on ImageNet-256 flattens beyond NFE ≈ 25 and saturates around FID ≈ 2.2, while the baseline continues to improve more slowly. This confirms that the 25-step configuration used in our main tables already operates close to the empirical quality ceiling of the underlying DiT-XL/2 model, and that further increasing the number of steps yields only marginal gains.

B.6. Failure Mode Quantification

At aggressive acceleration (NFE $= 8$) without rollback, we observe three characteristic failure modes over 5,000 random seeds on ImageNet-256:

- **Manifold collapse:** features collapse onto a low-dimensional region, leading to duplicated objects or missing structure.
- **Texture bleeding:** inaccurate metrics near boundaries cause textures to leak across semantic regions.
- **Semantic drift:** misaligned cross-attention leads to wrong object categories or attributes.

Table 5 reports incidence rates with and without rollback enabled.

Table 5. Failure mode incidence at NFE $= 8$ over 5,000 seeds. Rollback eliminates all observed failures.

Failure Mode	w/o Rollback	w/ Rollback
Manifold collapse	0.12% (6/5000)	0.00% (0/5000)
Texture bleeding	0.08% (4/5000)	0.00% (0/5000)
Semantic drift	0.14% (7/5000)	0.00% (0/5000)

Rollback triggers most often around timesteps 200–400, coinciding with transitions from coarse layout to fine detail where curvature peaks (Sec. C).

C. Geometric Validation

We now provide additional evidence that GeoRK2 operates on a low-rank Riemannian manifold and that its trajectories better follow the manifold geometry than flat-space baselines.

C.1. Spectral Energy Concentration

For DiT-XL/2, we compute eigenspectra of $G_t^{(\ell)}$ across layers $\ell \in \{6, 12, 18, 24\}$ and 50 timesteps, averaged over 100 random ImageNet samples. Table 6 shows that the top-64 eigen-directions consistently capture over 90% of spectral energy, with condition numbers in a moderate range.

Table 6. Spectral characteristics of $G_t^{(\ell)}$ across layers (averaged over 50 timesteps).

Layer	λ_1	λ_{64}	λ_{65}	Energy@64	κ
6	125.3	0.89	0.12	90.7%	140
12	203.7	1.24	0.15	91.3%	164
18	189.2	1.18	0.14	90.9%	160
24	156.8	0.96	0.11	90.5%	163

These statistics justify the choice of $r = 64$ as a good trade-off between geometric fidelity and computational cost, and they support the assumptions used in Lemma A.3. Empirically, we also observe that increments used by GeoRK2 lie predominantly in S_t, validating the low-rank restriction used in the theoretical analysis.

Note that Figure 3 in the main paper reports PCA variance on the raw mean-centered activations, where the top-64 principal components explain over 99% of the variance. Table 6 instead reports spectral energy with respect to the effective metric $G_{\text{eff}}(h_t)$, which reweights directions according to their contribution to the denoising dynamics. Under this metric, the top-64 eigen-directions consistently capture over 90% of the spectral energy. Our theoretical Assumption (A3) is stated in terms of this latter notion.

C.2. Trajectory Curvature and Drift

We quantify curvature κ of denoising trajectories by measuring changes in the tangent direction along projected activation paths, as described in Sec. 5.3 of the main paper. Table 7 summarizes mean and maximum curvature, drift rate, and subspace angle on DiT-XL/2 at NFE $= 25$.

The $\sim 58\%$ reduction in mean curvature and $\sim 86\%$ reduction in drift rate relative to DDIM empirically support the interpretation that GeoRK2 aligns more closely with the underlying geodesic flow, in accordance with the theoretical error bound in Theorem A.1.

Table 7. Trajectory curvature and alignment metrics (mean ± std over 500 trajectories).

Method	Mean $\kappa \downarrow$	Max $\kappa \downarrow$	Drift Rate $\downarrow$	Subspace Angle $\downarrow$
DDIM-25	0.43 ± 0.04	1.21 ± 0.11	0.28 ± 0.03	$8.7^\circ \pm 0.8^\circ$
TaylorSeer	0.31 ± 0.03	0.76 ± 0.09	0.15 ± 0.02	$6.2^\circ \pm 0.6^\circ$
GeoRK2	$\mathbf{0.18 \pm 0.02}$	$\mathbf{0.32 \pm 0.05}$	$\mathbf{0.04 \pm 0.01}$	$\mathbf{4.7^\circ \pm 0.5^\circ}$

C.3. Additional Visualizations

We further visualize activation trajectories by projecting DiT-XL/2 activations onto the top three eigenvectors of $G_t^{(\ell)}$ (Fig. 2). Flat-space samplers (DDIM, TaylorSeer) yield jagged paths that frequently drift away from high-density regions in this eigenbasis, whereas GeoRK2 produces a smoother trajectory that stays near the high-energy manifold region. These qualitative patterns are consistent with the curvature and drift statistics reported in Sec. C.

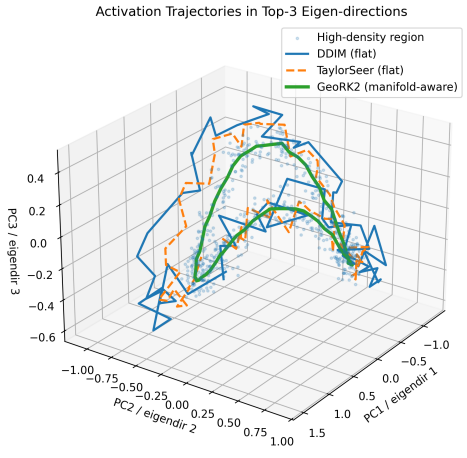

Figure 2. **Activation trajectories in the top-3 eigen-directions.** Flat-space samplers (DDIM, TaylorSeer) produce jagged, drifting paths, whereas GeoRK2 follows a smoother trajectory that stays near the high-density manifold region.

D. Runtime and Memory Profiling

Finally, we provide a fine-grained breakdown of GeoRK2's overhead relative to the base DiT-XL/2 forward pass. The numbers are representative of our PyTorch implementation and should be interpreted as indicative rather than absolute, since they depend on framework overhead and kernel fusion.

D.1. Runtime Breakdown

We profile a single DiT-XL/2 denoising step on an NVIDIA H20 96 GB GPU using PyTorch Profiler. Table 8 reports relative FLOPs and wall-clock time for each GeoRK2 component.

Using the main paper's estimate that a single DiT-XL/2 forward pass costs about 78 GFLOPs, the total GeoRK2 overhead of 0.047 in Table 8 corresponds to roughly $0.047 \times 78 \approx 3.7$ GFLOPs per denoising step, which is close to the ~ 4.2 GFLOPs figure measured in a separate profiling run and reported in the main text. Minor discrepancies arise from different profiling setups and kernel fusion.

Table 8. Runtime breakdown of GeoRK2 components on DiT-XL/2 ($d = 1152$, $r = 64$). FLOPs are reported as a fraction of a single DiT-XL/2 denoising step and amortized over 5 steps for SVD.

Component	Relative FLOPs	% of DiT step	Walltime %
Projection P_{S_t}	0.014	1.8%	1.5%
Metric inversion (Woodbury)	0.022	2.8%	2.0%
Truncated SVD (amortized)	0.011	1.4%	1.0%
Total GeoRK2 overhead	**0.047**	**6.0%**	**4.5%**

The per-step profiling yields approximately 4–6% extra walltime for a single DiT-XL/2 step, which translates into 5–8% end-to-end overhead in our standard ImageNet settings. In high-throughput regimes with large batch sizes and amortized SVD updates, the overhead can increase modestly (up to roughly 8–12%), mainly due to framework-level costs rather than pure FLOPs.

D.2. Memory Footprint

GeoRK2's additional memory comprises:

- storing $U_{r,t}$ for four bottleneck layers together with a small set of auxiliary vectors (velocities, accelerations, EMA statistics), which in total accounts for roughly 2.8MB of persistent geometry-specific state for $d = 1152$, $r = 64$;
- a rolling buffer of activations for covariance estimation (128 MB for 1024 tokens), which can often be shared with existing activation caching or checkpointing mechanisms; and
- a small number of temporary workspaces for SVD and metric updates.

On a 96 GB H20, even when the activation buffer is allocated separately, the resulting footprint remains well within the available memory budget (under 15% of total GPU memory). The 2.8 MB figure quoted in Sec. 4.2 of the main paper refers only to this persistent geometry-specific state (the stored $U_{r,t}$ and auxiliary statistics); the activation buffer is a separate, optionally shared workspace (128 MB in our implementation) and is therefore not counted in that number. In other words, the GeoRK2-specific persistent state itself is only a few megabytes on top of the base model, consistent with the figure reported in the main paper.