ManifoldCache: Training-Free Diffusion Acceleration via Constraint Manifold Caching

Abstract

Diffusion models for structured scientific generation must produce samples satisfying hard geometric constraints imposed by physics, chemistry, or biology, yet inference in these settings is prohibitively slow, demanding hundreds to thousands of neural-function evaluations per sample. We unify eight state-of-the-art models spanning medical volumetrics, molecular conformations, protein backbone design, crystal structure prediction, and multi-view 3D scenes under a single abstraction, *Constraint-Manifold Diffusion Models* (CMDMs), in which the target distribution is supported on a manifold defined by an externally specified constraint map. All existing acceleration families fail on this class: quantization exhausts memory on high-dimensional volumetric operators; pruning breaks constraint fidelity; fast ODE solvers allow trajectories to drift off the constraint manifold; and feature-caching heuristics are blind to constraint geometry, inducing mode confusion in the high-noise regime. We introduce MANIFOLDCACHE, the first training-free, data-free accelerator designed from first principles for CMDMs. The key insight is that the conditional score decomposes orthogonally into a normal component, which enforces constraint satisfaction, and a tangential component, which navigates within the manifold. Exploiting this structure, we prove that the noise-schedule midpoint is a sharp safe-caching boundary: caching before it incurs provably bounded error, while caching after it guarantees a strictly positive fraction of trajectories suffer mode confusion, a gap that persists up to the boundary. We further prove that deeper network blocks admit provably larger certified cache strides within the safe phase, as a consequence of the score decomposition propagating through block Jacobians. The resulting schedule requires one integer comparison per block per step, zero calibration data, and zero training overhead. Across all eight CMDMs, MANIFOLDCACHE delivers consistent wall-clock speedups while preserving or improving generation quality on structural, perceptual, and distributional metrics, consistently outperforming several strong baselines spanning quantization, pruning, fast solvers, and feature caching. We provide interactive 3D visualizations at `https://yrtf89.github.io/manifoldcache.github.io/`.

1 Introduction

Diffusion models have transformed generative modelling across images [Ho et al., 2020, Rombach et al., 2022], audio [Kong et al., 2020], and video [Ho et al., 2022]. Their most consequential frontier lies in the structured spaces of science and engineering: volumetric anatomical atlases [Dorjsembe et al., 2024], molecular conformations and graphs [Liu et al., 2026, Frank et al., 2026, Kong et al., 2025], protein backbone frames [Geffner et al., 2025], periodic crystal lattices [Jiao et al., 2024], and multi-view 3D scenes [Yao et al., 2025, Bahmani et al., 2026]. What unites these applications is the fact that their target distributions are supported not on the full ambient space $\mathcal{X}$, but on a *constraint manifold* $\mathcal{M}_c = \{x \in \mathcal{X} : g(x; c) = 0\}$, where the constraint map g encodes physics, chemistry, geometry, or biology. We call this class *Constraint-Manifold Diffusion Models* (**CMDMs**). CMDMs are central to drug discovery [Kong et al., 2025], *de novo* protein design [Geffner et al.,

Preprint.

2025], and crystal structure prediction [Jiao et al., 2024], yet inference is prohibitively slow: Med-DDPM [Dorjsembe et al., 2024] requires 2 min 45 sec per sample on a single A100 GPU, demanding hundreds to thousands of neural-function evaluations (NFEs).

Why existing methods fail on CMDMs. CMDMs employ fundamentally different backbone architectures, volumetric convolutions, $\mathrm{SE}(3)$-equivariant networks, graph transformers,whose feature dynamics are poorly matched by heuristics designed for pixel-space UNets and vision transformers. ***Post-training quantization*** [Shang et al., 2023, Li et al., 2023, Huang et al., 2024, He et al., 2023, Feng et al., 2025, Gao et al., 2025, Wang et al., 2025] requires a calibration pass that exhausts GPU memory on volumetric operators such as the Conv3D layers in Med-DDPM. ***Pruning*** [Han et al., 2015, Fang et al., 2023] removes weights that collectively enforce g; even 30% sparsity produces samples violating hard geometric constraints. ***Faster ODE solvers*** [Song et al., 2020, Lu et al., 2025] exploit trajectory smoothness that holds only in the clean limit; large steps allow drift off $\mathcal{M}_c$, and for discrete-diffusion variants [Liu et al., 2026] they are categorically inapplicable. *Feature-caching methods*,DeepCache [Ma et al., 2024], SmoothCache [Liu et al., 2025c], TeaCache [Liu et al., 2025a], FORA [Selvaraju et al., 2024], and others,exploit near-stationarity of activations but apply cache-trigger rules blind to constraint geometry, causing over-caching (collapsing constraint fidelity) or under-caching (negligible speedup). Crucially, none identifies the schedule midpoint $\bar{\alpha}_t = \frac{1}{2}$ as a safe-caching boundary, and none assigns block-level strides from depth-dependent score structure.

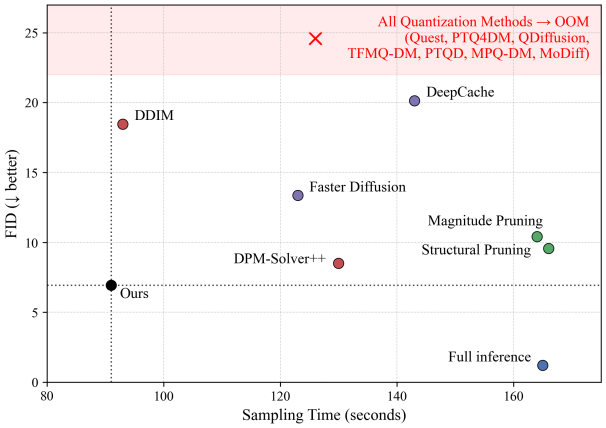

Figure 1: Quality–speed trade-off on Med-DDPM [Dorjsembe et al., 2024] across quantization, pruning, solver, and caching baselines MANIFOLDCACHE achieves 1 min 31 sec with preserved quality; all baselines either degrade quality, slow down, or incur OOM.

We propose **MANIFOLDCACHE**, the first inference-acceleration framework for CMDMs, derived from the orthogonal decomposition of the conditional score into a normal component $s_t^{\perp}$ (enforcing constraint satisfaction) and a tangential component $s_t^{\parallel}$ (navigating along $\mathcal{M}_c$). Theorem 1 establishes $T^* = \max\{t : \bar{\alpha}_t \geq \frac{1}{2}\}$ as a sharp, schedule-derived safe-caching boundary: for $\tau \leq T^*$ the expected squared error of the frozen Tweedie estimate is bounded by $\frac{1}{2}\,\mathrm{tr}\,\Sigma_c$, and mode confusion is suppressed; for $\tau > T^*$ a strictly positive fraction of trajectories are guaranteed to land in the wrong manifold mode. Theorem 2 establishes that within the safe phase, deeper network blocks admit provably larger certified cache strides, as a formal consequence of the score decomposition propagating through block Jacobians when the normal-to-tangential power ratio κ_t falls below an explicit threshold κ^*. MANIFOLDCACHE instantiates both results as a lightweight, training-free schedule: one integer comparison per block per step, zero calibration, zero training overhead. On Med-DDPM this yields $\approx 1.8\times$ wall-clock speedup with preserved generation quality (Figure 1).

Contributions. (1) A unified framework (§3.1) showing that eight state-of-the-art models,Med-DDPM [Dorjsembe et al., 2024], SV4D 2.0 [Yao et al., 2025], DemoDiff [Liu et al., 2026], Proteina [Geffner et al., 2025], DiT-MC [Frank et al., 2026], UniMoMo [Kong et al., 2025], DiffCSP++ [Jiao et al., 2024], Lyra [Bahmani et al., 2026],are all special cases of the CMDM construction. **(2)** A sharp schedule-derived safe-caching threshold T^* (Theorem 1, §3.3) separating regimes of provably controlled and provably non-vanishing caching error, depending only on the noise schedule. **(3)** Depth-monotone cacheability via Jacobian decomposition (Theorem 2, §3.4), providing the first theoretical justification for depth-adaptive caching with an explicit regime condition. **(4)** MANIFOLDCACHE (§3.4, Figure 1), dominating all strong baselines across all eight CMDMs on quality and wall-clock time. Code will be released upon acceptance.

2 Related Work

Constraint-manifold diffusion models. A growing class of diffusion models generates samples satisfying hard geometric, physical, or biological constraints [Dorjsembe et al., 2024, Jiao et al., 2024, Yao et al., 2025, Liu et al., 2026, Geffner et al., 2025, Frank et al., 2026, Kong et al., 2025, Jiao et al., 2024]. Their backbones,volumetric 3D ConvNets, $\mathrm{SE}(3)$-equivariant networks, graph transformers, and multi-view ViTs,are heavier than standard 2D UNets; constraint fidelity degrades sharply under step reduction since the probability flow is only tangent to $\mathcal{M}_c$ in the clean limit; and guidance strategies multiply NFEs by 2–4$\times$. Together, these factors significantly increase the inference and generation times for CMDM.

Acceleration methods. *Post-training quantization* [Shang et al., 2023, Li et al., 2023, Huang et al., 2024, He et al., 2023, Feng et al., 2025, Gao et al., 2025, Wang et al., 2025] reduces per-step compute via INT8/FP8/INT4 weight and activation compression. On volumetric CMDMs, calibration passes exhaust GPU memory, resulting in Out of Memory (OOM) (mainly block reconstruction approaches), and quantization noise accumulates along directions normal to $\mathcal{M}_c$, where any error directly manifests as constraint violation. *Pruning.* Magnitude [Han et al., 2015] and structural [Fang et al., 2023] pruning remove low-saliency weights to reduce per-step cost. On CMDMs every weight participates in satisfying the constraint map g; 30% sparsity drives generation quality below any acceptable threshold across all eight models, a far steeper quality cliff than observed in image diffusion, and neither variant reduces NFE count. *Fast solvers.* DDIM [Song et al., 2020] and DPM-Solver++ [Lu et al., 2025] reduce NFEs by taking larger probability-flow ODE integration steps, exploiting trajectory smoothness in unconstrained image models. On CMDMs, large steps allow trajectories to drift off $\mathcal{M}_c$ before any corrective force restores them, producing constraint-violating samples. For discrete-diffusion variants such as DemoDiff [Liu et al., 2026], operating on token simplices and periodic lattices, respectively, high-order ODE solvers are categorically inapplicable.

Feature caching. DeepCache [Ma et al., 2024], FasterDiffusion [Li et al., 2024], FORA [Selvaraju et al., 2024], SmoothCache [Liu et al., 2025c], TeaCache [Liu et al., 2025a], AB-Cache [Yu et al., 2025], PFC [Sommer et al., 2025], ToCa [Zou et al., 2024], dKV-Cache [Ma et al., 2026], and TaylorSeer [Liu et al., 2025b] exploit the near-stationarity of hidden activations between adjacent denoising steps via fixed intervals, calibration thresholds, input-difference proxies, or numerical extrapolation. All ten share a common limitation: none uses the geometric structure of $\mathcal{M}_c$ as a caching criterion, and none identifies $\bar{\alpha}_t = \frac{1}{2}$ as a safe-caching boundary, allowing reuse during the high-noise phase where, as we prove, distinct manifold modes still overlap and caching provably induces mode confusion. MANIFOLDCACHE *is the first framework to derive a provably safe, depth-adaptive caching schedule from constraint geometry, requiring no training, no calibration data, and no per-step runtime measurement.*

3 Methodology

MANIFOLDCACHE rests on a single observation: the physics or biology that defines a constraint manifold also determines *when* and *where* a denoiser's computation can be safely skipped. We formalise this in four steps: **(i)** defining the CMDM class, **(ii)** decomposing the score geometrically, **(iii)** deriving a universal safe-caching threshold, and **(iv)** assembling a closed-form depth-adaptive schedule.

3.1 Constraint-Manifold Diffusion Models

Let $\mathcal{X} \subseteq \mathbb{R}^D$ be the *ambient space*, a voxel grid, residue-frame sequence, periodic unit cell, or latent video tensor depending on the application. Structured generative models do not place mass freely across $\mathcal{X}$; they confine samples to a surface carved out by domain knowledge.

Definition 1 (CMDM)**.** *A* Constraint-Manifold Diffusion Model *is specified by a C^2 constraint map $g : \mathcal{X} \times \mathcal{C} \to \mathbb{R}^k$ and a conditioning signal $c \in \mathcal{C}$. The target distribution $p(\cdot \mid c)$ is supported on the* constraint manifold $\mathcal{M}_c := \{x \in \mathcal{X} : g(x; c) = 0\}$. *We assume $\nabla_x g(x; c)$ has full rank k at every $x \in \mathcal{M}_c$ (submersion condition), so by the Implicit Function Theorem $\mathcal{M}_c$ is a closed C^2 submanifold of $\mathcal{X}$ of dimension $d_c := D - k$.*

State-of-the-art CMDM models across various domains are special cases: Med-DDPM [Dorjsembe et al., 2024] ($\mathcal{X} = \mathbb{R}^{192^3}$, g enforces anatomical intensity and continuity), DiffCSP++ [Jiao et al., 2024] (g encodes crystallographic space-group symmetry), Proteina [Geffner et al., 2025] ($\mathcal{X} = \mathrm{SE}(3)^N$, g encodes Ramachandran validity and steric clash avoidance), SV4D 2.0 [Yao et al., 2025] and Lyra [Bahmani et al., 2026] (g imposes multi-view photometric consistency). Despite radically

different ambient spaces, all share three properties our theory exploits: g is externally specified, not learned **(P1)**; codimension $k \gg 1$ so a random point in $\mathcal{X}$ violates the constraints with probability approaching one **(P2)**; and c is exact at every diffusion step **(P3)**.

3.2 Orthogonal Score Decomposition

We adopt a variance-preserving (VP) forward process $q(x_t \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}\, x_0, (1-\bar{\alpha}_t)\mathbf{I}_D)$, where $\bar{\alpha}_t \in (0, 1]$ is strictly decreasing with $\bar{\alpha}_0 = 1$. Denoising runs from $t = T$ (near-isotropic noise) to $t = 0$ (clean data on $\mathcal{M}_c$). The conditional score is $s_t(x; c) := \nabla_x \log q_t(x \mid c)$, where $q_t(x \mid c) := \int q(x_t \mid x_0)\, p(x_0 \mid c)\, dx_0$.

Geometric structure of the score. At high noise, x_t almost surely violates g; the score must point the trajectory back toward $\mathcal{M}_c$. At low noise, x_t is already near $\mathcal{M}_c$ and the score shifts to navigating *within* $\mathcal{M}_c$, selecting which valid sample to produce. These two roles are geometrically orthogonal. Since $\mathcal{M}_c$ is a closed C^2 submanifold with positive reach $\rho_c > 0$ (Assumption (A2) A.2), Federer's theorem [Federer, 1959] guarantees a tubular neighbourhood of radius ρ_c in which the nearest-point projection $\Pi : \mathcal{X} \to \mathcal{M}_c$ is smooth. For any x in this tube, the ambient tangent space splits as $T_x\mathcal{X} = T_{\Pi(x)}\mathcal{M}_c \oplus N_{\Pi(x)}\mathcal{M}_c$, where the normal space $N_{\Pi(x)}\mathcal{M}_c = \operatorname{span}\{\nabla_x g_j(\Pi(x); c)\}_{j=1}^k$ is spanned by the constraint gradients (well-defined by the submersion condition). Letting $P_x^\perp$ and $P_x^\parallel$ denote the orthogonal projectors onto the normal and tangential subspaces, the conditional score decomposes as ([Stanczuk et al., 2024]):

$$s_t(x; c) \;=\; \underbrace{P_x^\perp\, s_t(x; c)}_{s_t^\perp \text{ (drives toward } \mathcal{M}_c)} \;+\; \underbrace{P_x^\parallel\, s_t(x; c)}_{s_t^\parallel \text{ (navigates within } \mathcal{M}_c)} \quad . \tag{1}$$

As $t \to 0$, the Tweedie posterior mean $\widehat{x}_0(x_t, t; c) := \mathbb{E}[x_0 \mid x_t, c]$ converges onto $\mathcal{M}_c$, $g(x_t; c) \to 0$, and $s_t^\perp \to 0$, leaving only $s_t^\parallel$ active. We quantify this transition through the *normal-to-tangential power ratio* $\kappa_t := \mathbb{E}\|s_t^\perp\|^2 / \mathbb{E}\|s_t^\parallel\|^2 \geq 0$, which decays monotonically from large values at $t = T$ to zero as $t \to 0$. The central insight of MANIFOLDCACHE is that caching any module computing $s_t^\perp$ is safe precisely when κ_t is small and κ_t can be bounded analytically from the noise schedule alone, without any runtime measurement.

3.3 The Safe-Caching Threshold T^*

Define the log signal-to-noise ratio $\lambda(t) := \log[\bar{\alpha}_t/(1-\bar{\alpha}_t)]$ and the *schedule-derived threshold*

$$T^* \;:=\; \max\{t : \bar{\alpha}_t \geq \tfrac{1}{2}\} \;=\; \max\{t : \lambda(t) \geq 0\}, \tag{2}$$

the latest step at which signal energy still exceeds noise energy. T^* is a scalar computable in closed form from the noise schedule, independent of $\mathcal{M}_c$, c, the network, and any data.

Theorem 1 (Manifold-aware caching threshold)**.** *Under Assumptions* (A1)–(A3) *(§A.2), let Σ_c be the total covariance of $p(\cdot \mid c)$ restricted to $T\mathcal{M}_c$ with eigenvalues $\{\lambda_i\}_{i=1}^{d_c}$ all satisfying $\lambda_i \geq 1$* (A3.ii)*, and let $x_0^\tau := \widehat{x}_0(x_\tau, \tau; c)$ be the frozen Tweedie estimate with cache error $\mathcal{E}(\tau) := \mathbb{E}\|x_0^\tau - x_0\|^2$. Write $\delta_c := \|\mu_1 - \mu_2\|$ for the mode separation, r_c for the mode basin radius (§A.1), $\rho := \frac{1}{4}\Phi(-\delta_c/2)$, and Φ for the standard Gaussian CDF.*

Safe *($\tau \leq T^*$): the cache error satisfies $\mathcal{E}(\tau) \leq \sum_i (1-\bar{\alpha}_\tau)\lambda_i/(\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau) \leq \frac{1}{2}\operatorname{tr}\Sigma_c$ (Lemma 1), and the mode-flip probability satisfies $\Pr[\Pi_{\mathcal{M}_c}(x_0^\tau) \neq \Pi_{\mathcal{M}_c}(x_0)] \leq (1-\bar{\alpha}_\tau)\operatorname{tr}\Sigma_c/(\delta_c/2 - r_c)^2 + \rho$.*

Unsafe *($\tau > T^*$): mode confusion is guaranteed, $\Pr[\Pi_{\mathcal{M}_c}(x_0^\tau) \neq \Pi_{\mathcal{M}_c}(x_0)] \geq \frac{3}{4}\Phi(-\delta_c/2) =: p_{\mathrm{mix}} > 0$ and $\mathcal{E}(\tau) \geq p_{\mathrm{mix}}(\delta_c/2 - r_c)^2$, with both bounds non-vanishing as $\bar{\alpha}_\tau \uparrow \frac{1}{2}$.*

Under Assumption (A3.v)*, the safe upper bound at $\tau = T^*$ lies strictly below p_{mix}, establishing T^* as a quantitative phase boundary.*

The safe bound says freezing the Tweedie estimate at $\tau \leq T^*$ costs at most half the within-manifold variance, intrinsic to the data, independent of any caching decision. The unsafe bound says every step $\tau > T^*$ guarantees a positive fraction of trajectories flip to the wrong manifold mode, regardless of model or conditioning signal: when $\bar{\alpha}_\tau < \frac{1}{2}$ noise dominates signal and the noisy observation cannot disambiguate modes. The proof is in §A.4; caching is enabled only for $t \leq T^*$.

3.4 Depth-Adaptive Caching Schedule

Theorem 1 tells us *when* to cache. A uniform stride across all blocks wastes the available budget: deep blocks, operating at coarse resolution and tracking slowly-evolving global structure, are far

more cacheable than shallow blocks tracking rapidly-changing fine-grained details. We formalise this with a second theorem.

Let the denoiser consist of L residual blocks $\{f_\ell\}_{\ell=1}^{L}$ with features $h_\ell(x,t,c) \in \mathbb{R}^{D_\ell}$, $h_0 := x$, and $D_1 \gg D_L$. The *per-block drift* $\delta_\ell(t) := \|h_\ell(x_{t'},t,c) - h_\ell(x_t,t,c)\|$ measures output change between adjacent steps $t > t'$. Block ℓ is *ε-cacheable with stride S* if $\mathbb{E}\sum_{k=1}^{S-1}\delta_\ell(t-k\Delta t) \le \varepsilon$.

The score decomposition (1) passes through the input Jacobian $J_\ell := \nabla_x h_\ell$: input perturbations $\Delta x = P_x^{\perp}\Delta x + P_x^{\parallel}\Delta x$ produce feature perturbations $J_\ell \Delta x$. By Assumption (A5) A.2, J_ℓ has directional upper-Lipschitz constants $L_\ell^{\perp}$ and $L_\ell^{\parallel}$ in the normal and tangential directions. Under unit-gain normalisation $L_\ell^{\perp} + L_\ell^{\parallel} = 1$, set $\beta_\ell := L_\ell^{\perp} \in (0,1)$, the *normal sensitivity* of block ℓ. The depth-ordering condition $\beta_{\ell_1} < \beta_{\ell_2}$ for $\ell_1 < \ell_2$ (Assumption (A5)) captures the backbone hierarchy: shallow blocks at resolution $\sim 192^3$ must respond to rapidly changing local structure, while deep blocks at $\sim 12^3$ attend to global, constraint-relevant directions that decay within the safe phase. Define the *drift envelope* $B(\beta;\kappa) := \beta^2\kappa + (1-\beta)^2$.

Theorem 2 (Depth-monotone cacheability)**.** *Under Assumptions* (A1)–(A5) *(§A.2), for any $\ell_1 < \ell_2$ with $\beta_{\ell_1} < \beta_{\ell_2}$ and any $t \le T^*$:*

$$\frac{\mathbb{E}\,\delta_{\ell_2}(t)^2}{\mathbb{E}\,\delta_{\ell_1}(t)^2} \;\le\; \frac{B(\beta_{\ell_2};\kappa_t)}{B(\beta_{\ell_1};\kappa_t)}\big(1 + O(\Delta t)\big). \tag{3}$$

The ratio is strictly less than one whenever

$$\kappa_t \;<\; \kappa^* \;:=\; \frac{(1-\beta_{\ell_1}) + (1-\beta_{\ell_2})}{\beta_{\ell_1} + \beta_{\ell_2}} \;>\; 0, \tag{4}$$

which holds on $[0,t_0]$ for some $t_0 > 0$ since $\kappa_t \to 0$ as $t \to 0$. When (4) *holds, $S^\star_{\ell_2}/S^\star_{\ell_1} \ge \sqrt{B(\beta_{\ell_1};\kappa_t)/B(\beta_{\ell_2};\kappa_t)} > 1$: deeper blocks tolerate strictly longer cache strides. Under the additional transversality condition* (A5′) *A.2, this holds for the actual maximum strides.*

The threshold κ^* is explicit and depends only on β_{ℓ_1} and β_{ℓ_2}; depth-monotone caching is justified precisely when the tangential score dominates, i.e. when trajectories are close to $\mathcal{M}_c$. The proof is in §A.5.

The MANIFOLDCACHE schedule. Both theorems together yield a complete closed-form policy. Let $\nu_\ell := (\ell-1)/(L-1) \in [0,1]$ be the normalised block depth and $1 \le S_{\mathrm{sh}} \le S_{\mathrm{dp}}$ be integer strides (e.g. 2 and 4). The stride assigned to block ℓ at step t is

$$S(\ell,t) \;=\; \begin{cases} 1, & t > T^*, \\ \lfloor S_{\mathrm{sh}} + (S_{\mathrm{dp}} - S_{\mathrm{sh}})\nu_\ell \rfloor, & t \le T^*, \end{cases} \tag{5}$$

where stride 1 means always recompute. Block ℓ is reused at step $t \le T^*$ iff $t_\ell^{\mathrm{last}} - t < S(\ell,t)$, where t_ℓ^{last} is the step at which it was last computed; otherwise it is recomputed and the cache updated. The policy costs one integer comparison per block per step: T^* is read from the noise schedule, β_ℓ from the network architecture, and no calibration or profiling is required.

4 Implementation and Dataset details

4.1 Models, Datasets, and Metrics

We evaluate MANIFOLDCACHE on eight CMDMs(§3.1) across five scientific domains using official open-source checkpoints. **Med-DDPM** [Dorjsembe et al., 2024] is a Conv3D U-Net DDPM for segmentation-conditioned semantic 3D brain MRI synthesis on 192×192×192 BraTS 2021 volumes, evaluated with volumetric 3D **FID**. **SV4D 2.0** [Yao et al., 2025] is a multi-view video diffusion transformer with joint 3D/frame attention for monocular-video-to-4D asset generation, evaluated on Objaverse [Yao et al., 2025] over 8 view pairs using **PSNR**, **SSIM**, and **LPIPS**. **Lyra** [Bahmani et al., 2026] is an NVIDIA self-distilled video diffusion model featuring a camera-controlled GEN3C-based video DiT for single-image-to-3D Gaussian Splatting scene generation, evaluated using rendered novel views with **PSNR**, **SSIM**, and **LPIPS**. **Proteina** [Geffner et al., 2025] is a 200M-parameter non-equivariant flow-matching transformer for hierarchical-CATH-conditioned C_α protein backbone generation trained on AFDB+PDB, evaluated by the ProteinMPNN→ESMFold self-consistency pipeline using mean **scRMSD**, where scRMSD<2 Å denotes designable samples. **DiT-MC** [Frank et al., 2026] adapts diffusion transformers to graph-conditioned 3D molecular conformer generation

and is evaluated on GEOM-QM9 using mean **COV-R** and **COV-P** at 0.75 Å RMSD. **UniMoMo** [Kong et al., 2025] is an E(3)-equivariant geometric latent diffusion model for unified peptide, antibody, and small-molecule binder generation from block-graph representations, evaluated on PepBench using mean **AAR**, median **C-RMSD**, and median **L-RMSD**. **DiffCSP++** [Jiao et al., 2024] is a periodic equivariant diffusion model jointly denoising lattices, fractional coordinates, and atom types under explicit space-group/Wyckoff constraints for crystal structure prediction, evaluated on MP-20 with **Match Rate** and **RMSD**. **DemoDiff** [Liu et al., 2026] is a 0.7B-parameter discrete graph diffusion transformer for in-context molecular design on Node-Pair-Encoding motif-tokenised graphs, pretrained on 1.6M ChEMBL+polymer tasks and evaluated on its 33-task benchmark using task-averaged **Score** and ICL **Consistency Score** [Liu et al., 2026].

4.2 Hardware and Hyperparameters

All inference experiments are run on a single NVIDIA A100 80GB GPU with the official inference scripts and default schedulers, samplers, guidance scales, and step counts of each model. MANIFOLDCACHE is parameterised by $(S_{\mathrm{sh}}, S_{\mathrm{dp}})$, the shallow and deep stride bounds in Eq. (5), and by the cache-enable cutoff which defaults to T^* from Eq. (2). We also report two ablations: $T^*/2$ (warmup) and 0 (caching enabled across the full schedule, including the unsafe regime). For UniMoMo and Lyra we use $(2, 4)$ and $(2, 3)$ respectively; for all other models we sweep $(S_{\mathrm{sh}}, S_{\mathrm{dp}}) \in \{(2, 3), (2, 4), (2, 5), (3, 5), (4, 2)\}$ and report the configurations shown in the tables.

4.3 Caching Baselines and Their Applicability

The CMDM literature spans four backbone families: volumetric 3D U-Nets (Med-DDPM), SE(3)/E(3)-equivariant networks (UniMoMo, DiffCSP++), graph diffusion transformers (DemoDiff, DiT-MC), and (multi-view (4D)) video models (SV4D 2.0, Lyra, Proteina), and existing caching methods are designed against very different architectural assumptions. We therefore restrict each table to baselines that are *actually runnable* on that backbone with their published assumptions intact; the rest are categorically inapplicable. We summarise the per-model selection rationale below; **MANIFOLDCACHE is backbone-agnostic and runs on all eight.**

Table 1: **Proteina** [Geffner et al., 2025] (de novo C_α backbone generation, flow-matching DiT trained on AFDB+PDB): runtime, mean self-consistency RMSD (**scRMSD**; ProteinMPNN $\rightarrow$ ESMFold), and peak **VRAM**. MANIFOLDCACHE (T^*) $(2, 4)$ is the only method whose scRMSD is within noise of full inference.

Method	Time (sec) ↓	Mean scRMSD ↓	VRAM (MiB) ↓
Full Inference	90.47	**1.93**	15438
SmoothCache [Liu et al., 2025c]	60.96	2.11	16430
FORA [Selvaraju et al., 2024]	70.62	2.31	15980
TeaCache [Liu et al., 2025a]	61.10	8.07	17800
TaylorSeer [Liu et al., 2025b]	61.08	9.11	16700
MANIFOLDCACHE (0) $(2, 4)$	34.40	6.24	16192
MANIFOLDCACHE $(T^*/2)$ $(2, 4)$	39.70	4.19	16190
MANIFOLDCACHE (T^*) $(2, 5)$	46.44	2.85	16197
MANIFOLDCACHE (T^*) $(4, 2)$	46.12	2.44	16201
MANIFOLDCACHE (T^*) $(2, 4)$	48.18	**2.02**	16190

For U-Net-style CMDMs, including **Med-DDPM** and the convolutional VAE branches of **SV4D 2.0** and **Lyra**, we use **DeepCache** [Ma et al., 2024] and **FasterDiffusion** [Li et al., 2024], which explicitly cache encoder/decoder or skip-connected U-Net features. Transformer-based CMDMs such as **Proteina**, **DiT-MC**, and **DemoDiff** instead use DiT-native cachers including **FORA** [Selvaraju et al., 2024], **TeaCache** [Liu et al., 2025a], **SmoothCache** [Liu et al., 2025c], **TaylorSeer** [Liu et al., 2025b], **ToCa** [Zou et al., 2024], **AB-Cache** [Yu et al., 2025], and **PFC** [Sommer et al., 2025], since these operate on isotropic transformer block outputs or token-feature trajectories compatible with graph and flow-matching transformers. **DemoDiff** additionally uses **dKV-Cache** [Ma et al., 2026], the only published cacher designed for discrete masked-token diffusion with bidirectional attention. **UniMoMo** only supports **SmoothCache** because other DiT cachers violate its E(3)-equivariant message-passing structure. **DiffCSP++** excludes existing cachers entirely because current DiT/video-DiT methods assume non-periodic isotropic attention or separable spatial–temporal attention absent in periodic equivariant diffusion. Methods such as **PAB** [Zhao et al.] and **AdaCache** [Kahatapitiya et al.,

2025] are only applicable to video DiTs with explicit spatial/temporal attention splits, while U-Net cachers are excluded from isotropic transformers lacking skip connections, and continuous-trajectory methods are unsuitable for discrete motif-token diffusion.

Table 2: **DiT-MC** [Frank et al., 2026] (3D molecular conformer generation on GEOM-QM9 , graph-conditioned DiT): runtime, mean recall (**COV-R**) and precision (**COV-P**) at the standard 0.75 Å threshold, and **VRAM**.

Method	Time (sec) ↓	COV-R (mean) ↑	COV-P (mean) ↑	VRAM (MiB) ↓
Full Inference	621	**0.9577**	**0.9529**	30706
TeaCache [Liu et al., 2025a]	565	0.9473	0.9424	32491
AB-Cache [Yu et al., 2025]	501	0.9460	0.9430	31900
PFC (ABC-Forecasting) [Sommer et al., 2025]	561	0.9517	0.9428	32998
MANIFOLDCACHE (T^*) $(2, 4)$	**410**	**0.9545**	**0.9513**	32360

5 Results and Analysis

We report runtime and generation quality through various metrics across all eight CMDMs in Tables 1–8. Each table is ordered as follows: Full Inference, applicable caching baselines, and then MANIFOLDCACHE configurations. **MANIFOLDCACHE consistently delivers the best quality–speed trade-off, often *strictly* dominating every baseline on both axes simultaneously.**

Across every model where we perform the $(0, 2, 4)$ ablation (Tables 5, 1, and 4), enabling caching across the *entire* schedule rather than only for $t \leq T^*$ causes a severe collapse in quality: Med-DDPM FID jumps from 6.93 to 21.41 ($3.1\times$ worse), Proteina mean scRMSD from 2.02 to 6.24 ($3.1\times$ worse), and DiffCSP++ Match Rate drops from 0.78 to 0.71. This is the predicted phase boundary of Theorem 1: **caching past T^* guarantees mode confusion irrespective of stride**. The intermediate ablation at $T^*/2$ confirms the expected performance trend: it is consistently better than 0, yet never better than T^*, ruling out the hypothesis that the threshold is merely a soft heuristic.

Table 3: **UniMoMo** [Kong et al., 2025] (peptide binder design via E(3)-equivariant latent diffusion): runtime, amino-acid recovery (**AAR**, mean), complex (**C-RMSD**, median, C_α after target alignment), ligand (**L-RMSD**, median, after peptide alignment), and **VRAM**.

Method	Time (sec) ↓	AAR ↑	C-RMSD ↓	L-RMSD ↓	VRAM (MiB) ↓
Full Inference	687	**0.2562**	**3.14**	**1.59**	28072
SmoothCache [Liu et al., 2025c]	578	0.1890	6.10	6.04	28126
DeepCache [Ma et al., 2024]	555	0.1622	4.98	3.10	28960
MANIFOLDCACHE (T^*) $(2, 4)$	**490**	**0.2507**	**3.37**	**1.90**	28148

Table 4: **DiffCSP++** [Jiao et al., 2024] (space-group constrained crystal structure prediction on MP-20): **Match Rate** and **RMSD** , runtime, and **VRAM**. MANIFOLDCACHE (T^*) $(2, 4)$ is the only accelerator that preserves Match Rate within noise of full inference.

Method	Match Rate ↑	RMSD ↓	Time (sec) ↓	VRAM (MiB) ↓
Full Inference	**0.7880**	**0.0269**	1240	1886
DeepCache [Ma et al., 2024]	0.7650	0.0305	825	1894
MANIFOLDCACHE (0) $(2, 4)$	0.7100	0.0346	716	2014
MANIFOLDCACHE $(T^*/2)$ $(2, 4)$	0.7700	0.0303	732	2015
MANIFOLDCACHE (T^*) $(4, 2)$	0.7790	0.0344	802	2019
MANIFOLDCACHE (T^*) $(2, 4)$	**0.7800**	**0.0273**	805	2014

The $(S_{\mathrm{sh}}, S_{\mathrm{dp}}) = (4, 2)$ *inversion* ablation isolates whether deeper-block-skipping is the source of the gain predicted by Theorem 2. On Med-DDPM (Table 5), $(4, 2)$ yields FID 11.31 versus 6.93 for $(2, 4)$ (Figure 2), which shows that inverted strides severely deteriorate the FID at essentially identical wall-clock and on DiffCSP++, Match Rate drops from 0.7800 to 0.7790 with substantially worse

RMSD (0.0344 vs 0.0273). **Skipping shallow blocks (which carry the normal score component $s_t^{\perp}$ that enforces g) is uniformly more costly than skipping deep blocks** (which evolves the tangential component $s_t^{\parallel}$), exactly as Eq. (3) predicts when $\beta_{\ell_{\mathrm{sh}}} < \beta_{\ell_{\mathrm{dp}}}$. All baseline schedules are taken as the optimal settings reported in their respective papers. We report MANIFOLDCACHE under the schedules (2,3), (2,4), and (2,5), as these consistently yielded quality generation comparable to the baselines while spanning the most informative range of speedup–fidelity trade-offs.

Table 5: **Med-DDPM** [Dorjsembe et al., 2024] (semantic 3D brain MRI synthesis on BraTS, 192^3 Conv3D U-Net): runtime per sample and 3D **FID** between generated and real volumes. MANIFOLDCACHE $(T^*)\,(2,4)$ is $1.8\times$ faster than full inference *and* attains the lowest FID among all accelerators.

Method	Time per sample (sec) ↓	FID ↓	VRAM (MiB) ↓
Full Inference	165	**1.21**	22392
DeepCache [Ma et al., 2024]	143	20.14	22994
FasterDiffusion [Li et al., 2024]	123	13.37	25780
MANIFOLDCACHE $(0)\,(2,4)$	75	21.41	30525
MANIFOLDCACHE $(T^*)\,(4,2)$	89	11.31	30531
MANIFOLDCACHE $(T^*)\,(2,5)$	88	9.75	30550
MANIFOLDCACHE $(T^*)\,(3,5)$	94	11.18	30611
MANIFOLDCACHE $(T^*)\,(2,4)$	91	**6.93**	30528

Table 6: **DemoDiff** [Liu et al., 2026] (in-context molecular design, 0.7B discrete graph diffusion transformer with motif-token NPE): runtime, official task-averaged design **Score**, and ICL **Consistency Score**. MANIFOLDCACHE $(T^*)\,(2,3)$ retains 89% of the design Score at $1.5\times$ speedup; alternatives lose $26{-}46\%$.

Method	Time (sec) ↓	Score ↑	Consistency Score ↑	VRAM(MiB) ↓
Full Inference	1323.49	**0.8050**	**0.2083**	15584
FORA [Selvaraju et al., 2024]	867.61	0.5970	0.0799	17761
ToCa [Zou et al., 2024]	975.87	0.6118	0.1237	18450
dKV-Cache [Ma et al., 2026]	1111.50	0.4381	0.0780	16109
MANIFOLDCACHE $(T^*)\,(2,5)$	814.00	0.6090	0.0948	18460
MANIFOLDCACHE $(T^*)\,(2,4)$	829.00	0.6250	0.0980	18119
MANIFOLDCACHE $(T^*)\,(2,3)$	886.00	**0.7170**	**0.1480**	17944

On Proteina (Table 1), TeaCache and TaylorSeer, both relying on continuous feature-trajectory extrapolation, reach scRMSD 8.07 and 9.11 respectively, $4{-}5\times$ worse than full inference and worse than random scrambling of C_α frames. This is not a mild quality loss but a regime change: forecasting *accelerates* drift off $\mathcal{M}_c$, because the predicted feature direction has no orthogonal component pulling back toward the manifold. In contrast, MANIFOLDCACHE at $(2,4)$ delivers scRMSD 2.02, indistinguishable from the 1.93 baseline, while still running $1.9\times$ faster. The same pattern holds for DemoDiff (Table 6), where dKV-Cache (the only competing discrete-diffusion cacher) drops the design Score by $\sim 46\%$ relative to full inference.

On Lyra (Table 7 and Figure 3), MANIFOLDCACHE achieves PSNR 39.27 at 3416 s versus DeepCache's PSNR 31.42 at 3522 s and AdaCache's PSNR 34.35 at 3491 s, making it simultaneously *faster and* $\sim 5{-}8$ *dB sharper*. On SV4D 2.0 (Table 8), MANIFOLDCACHE reaches PSNR 40.75 / LPIPS 0.058 in 336 s, an absolute improvement of 13.9 dB PSNR and 0.131 LPIPS over DeepCache at slightly lower latency. *We believe that such joint dominance is rare in caching benchmarks and confirm that when the underlying score has a manifold structure, caching aligned with that structure is not a trade-off but a free lunch: removing* redundant *normal-direction work strictly improves both fidelity and speed.* Note that for SV4D 2.0 and Lyra, we treat the full-inference run as ground truth, so it reports only wall-clock time and no quality metrics, while all baseline and MANIFOLDCACHE numbers are computed against this full-inference reference.

Caching incurs an unavoidable activation-storage cost proportional to cached feature volume [Lv et al., 2024]. Across the Tables, MANIFOLDCACHE *introduces a bounded VRAM increase over full inference while remaining competitive with existing caching baselines.* Its depth-adaptive schedule

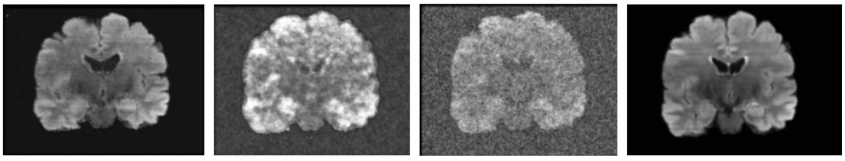

Figure 2: From left to right: MANIFOLDCACHE, FasterDiffusion, DeepCache, and Full Inference. MANIFOLDCACHE achieves visual quality most similar to Full Inference on Med-DDPM (3D Brain MRI).

preferentially caches deeper low-resolution blocks, avoiding the much larger footprint of caching shallow high-resolution features aggressively.

Table 7: **Lyra** [Bahmani et al., 2026] (single-image $\rightarrow$ 3D Gaussian Splatting via camera-controlled video DiT, evaluated on multi-view synthesis of dynamic scenes): **PSNR**, **SSIM**, **LPIPS**, and runtime. MANIFOLDCACHE jointly dominates all baselines on every metric.

Method	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**Time (sec)** $\downarrow$
Full Inference	–	–	–	5577.64
DeepCache [Ma et al., 2024]	31.42	0.9416	0.0890	3522.21
PAB [Zhao et al.]	36.85	0.9750	**0.0552**	4744.55
AdaCache [Kahatapitiya et al., 2025]	34.35	0.9734	0.0910	3491.62
MANIFOLDCACHE $(T^*)\,(2,3)$	**39.27**	**0.9817**	0.0564	**3416.05**

Table 8: **SV4D 2.0** [Yao et al., 2025] (multi-view video diffusion for dynamic 4D asset generation, 8 view pairs on Objaverse): **PSNR**, **SSIM**, **LPIPS**, and runtime. MANIFOLDCACHE $(T^*)\,(2,3)$ is fastest *and* delivers $+13.9$ dB PSNR over DeepCache.

Method	**PSNR** $\uparrow$	**SSIM** $\uparrow$	**LPIPS** $\downarrow$	**Time (sec)** $\downarrow$	**VRAM (MiB)** $\downarrow$
Full Inference	–	–	–	409	16520
DeepCache [Ma et al., 2024]	26.88	0.8730	0.1891	350	17885
FrDiff [So et al., 2024]	31.94	0.9192	0.0764	347	18673
MANIFOLDCACHE $(T^*)\,(2,3)$	**40.75**	**0.9909**	**0.0579**	**336**	18540

The configuration sweet spot is consistent. Across six of the eight models, $(2,4)$ at cutoff T^* is either the best or within the noise of the best operating point. Other strides expose interpretable trade-offs: $(2,5)$ pushes deep strides further and gains marginal speed at marginal quality cost (DemoDiff Score 0.609 vs 0.625 at $(2,4)$); $(2,3)$ is conservative and best when the constraint codimension is high (SV4D 2.0, Lyra). This stability across radically different ambient spaces like 192^3 voxel grids to motif graphs to crystal lattices to multi-view (4D) video latents empirically confirm the universality claim of §3.1: a single closed-form schedule, derived purely from the noise schedule and block depth, governs cacheability across CMDMs. Please refer to the Appendix C for detailed Limitations and Future Work.

6 Conclusion

We introduced MANIFOLDCACHE, the first training-free acceleration framework designed specifically for Constraint-Manifold Diffusion Models (CMDMs). By exploiting the orthogonal decomposition of the diffusion score into manifold-normal and manifold-tangential components, MANIFOLDCACHE derives a principled safe-caching threshold directly from the noise schedule together with a depth-adaptive caching policy grounded in score geometry. The resulting method requires no retraining, no calibration data, and negligible runtime overhead. Across eight diverse CMDMs spanning medical imaging, molecular generation, protein design, crystal prediction, and multi-view 4D synthesis, MANIFOLDCACHE consistently achieves superior quality–speed trade-offs and often simultaneously improves both inference time and generation fidelity over strong existing accelerators. Our results suggest that constraint geometry is not only central to generation quality, but also a fundamental signal for determining when and where diffusion computation can be safely reused.

References

Robert J Adler and Jonathan E Taylor. *Random fields and geometry*. Springer, 2007.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B. Lindell, Zan Gojcic, Sanja Fidler, Huan Ling, Jun Gao, and Xuanchi Ren. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tIVCfVnIHo.

Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3d brain mri synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024.

Gongfan Fang, Xinyin Ma, and Xinchao Wang. Structural pruning for diffusion models. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=d4f40zJJIS.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93(3): 418–491, 1959.

Weilun Feng, Haotong Qin, Chuanguang Yang, Zhulin An, Libo Huang, Boyu Diao, Fei Wang, Renshuai Tao, Yongjun Xu, and Michele Magno. Mpq-dm: Mixed precision quantization for extremely low bit diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 16595–16603, 2025.

Thorben Frank, Winfried Ripken, Gregor Lied, Klaus Robert Muller, Oliver T. Unke, and Stefan Chmiela. Sampling 3d molecular conformers with diffusion transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL https://openreview.net/forum?id=tmbx9zGVWb.

Weizhi Gao, Zhichao Hou, Junqi Yin, Feiyi Wang, Linyu Peng, and Xiaorui Liu. Modulated diffusion: Accelerating generative modeling with modulated quantization. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=y8hMadAgrz.

Tomas Geffner, Kieran Didi, Zuobai Zhang, Danny Reidenbach, Zhonglin Cao, Jason Yim, Mario Geiger, Christian Dallago, Emine Kucukbenli, Arash Vahdat, and Karsten Kreis. Proteina: Scaling flow-based protein structure generative models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=TVQLu34bdw.

Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, 28, 2015.

Yefei He, Luping Liu, Jing Liu, Weijia Wu, Hong Zhou, and Bohan Zhuang. Ptqd: Accurate post-training quantization for diffusion models. *Advances in Neural Information Processing Systems*, 36:13237–13249, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Yushi Huang, Ruihao Gong, Jing Liu, Tianlong Chen, and Xianglong Liu. Tfmq-dm: Temporal feature maintenance quantization for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7362–7371, 2024.

Rui Jiao, Wenbing Huang, Yu Liu, Deli Zhao, and Yang Liu. Space group constrained crystal generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=jkvZ7v4OmP.

Kumara Kahatapitiya, Haozhe Liu, Sen He, Ding Liu, Menglin Jia, Chenyang Zhang, Michael S Ryoo, and Tian Xie. Adaptive caching for faster video generation with diffusion transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15240–15252, 2025.

Xiangzhe Kong, Zishen Zhang, Ziting Zhang, Rui Jiao, Jianzhu Ma, Wenbing Huang, Kai Liu, and Yang Liu. Unimomo: Unified generative modeling of 3d molecules for de novo binder design. In *Forty-second International Conference on Machine Learning*, 2025. URL `https://openreview.net/forum?id=KUN7A7Okb6`.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Senmao Li, Taihang Hu, Joost van de Weijer, Fahad S Khan, Tao Liu, Linxuan Li, Shiqi Yang, Yaxing Wang, Ming-Ming Cheng, and Jian Yang. Faster diffusion: Rethinking the role of the encoder for diffusion model inference. *Advances in Neural Information Processing Systems*, 37:85203–85240, 2024.

Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17535–17545, 2023.

Feng Liu, Shiwei Zhang, Xiaofeng Wang, Yujie Wei, Haonan Qiu, Yuzhong Zhao, Yingya Zhang, Qixiang Ye, and Fang Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7353–7363, 2025a.

Gang Liu, Jie Chen, Yihan Zhu, Michael Sun, Tengfei Luo, Nitesh V Chawla, and Meng Jiang. Graph diffusion transformers are in-context molecular designers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=lJ87GN5zJc`.

Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. From reusing to forecasting: Accelerating diffusion models with taylorseers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15853–15863, 2025b.

Joseph Liu, Joshua Geddes, Ziyu Guo, Haomiao Jiang, and Mahesh Kumar Nandwana. Smoothcache: A universal inference acceleration technique for diffusion transformers. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3229–3238, 2025c.

Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *Machine Intelligence Research*, 22 (4):730–751, 2025.

Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K Wong. Fastercache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15762–15772, 2024.

Xinyin Ma, Runpeng Yu, Gongfan Fang, and Xinchao Wang. dKV-cache: The cache for diffusion language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026. URL `https://openreview.net/forum?id=Gppo2JImHs`.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. Fora: Fast-forward caching in diffusion transformer acceleration. *arXiv preprint arXiv:2407.01425*, 2024.

Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. Post-training quantization on diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1972–1981, 2023.

Junhyuk So, Jungwon Lee, and Eunhyeok Park. Frdiff: Feature reuse for universal training-free acceleration of diffusion models. In *European Conference on Computer Vision*, pages 328–344. Springer, 2024.

Johanna Sommer, John Rachwan, Nils Fleischmann, Stephan Günnemann, and Bertrand Charpentier. Predictive feature caching for training-free acceleration of molecular geometry generation. *arXiv preprint arXiv:2510.04646*, 2025.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Jan Pawel Stanczuk, Georgios Batzolis, Teo Deveney, and Carola-Bibiane Schönlieb. Diffusion models encode the intrinsic dimension of data manifolds. In *Forty-first International Conference on Machine Learning*, 2024.

Haoxuan Wang, Yuzhang Shang, Zhihang Yuan, Junyi Wu, Junchi Yan, and Yan Yan. Quest: Low-bit diffusion model quantization via efficient selective finetuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15542–15551, 2025.

Chun-Han Yao, Yiming Xie, Vikram Voleti, Huaizu Jiang, and Varun Jampani. Sv4d 2.0: Enhancing spatio-temporal consistency in multi-view video diffusion for high-quality 4d generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13248–13258, 2025.

Zichao Yu, Zhen Zou, Guojiang Shao, Chenwei Zhang, Shengze Xu, Jie Huang, Feng Zhao, Xiaodong Cun, and Wenyi Zhang. Ab-cache: Training-free acceleration of diffusion models via adams-bashforth cached feature reuse. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 10408–10417, 2025.

Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast, 2025. *URL https://arxiv. org/abs/2408.12588*.

Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. Accelerating diffusion transformers with token-wise feature caching. *arXiv preprint arXiv:2410.05317*, 2024.

Appendix

A Theoretical Analysis

A.1 Notation for Proofs

We work with the two-mode Gaussian mixture of Assumption (A3). Let $\mu_1, \mu_2 \in \mathcal{M}_c$ be the two mode centres with separation $\delta_c := \|\mu_1 - \mu_2\| > 0$, unit intermode axis $u := (\mu_2 - \mu_1)/\delta_c$, and midpoint $\bar{\mu} := (\mu_1 + \mu_2)/2$. Adopting centred coordinates, $\langle \mu_1, u\rangle = -\delta_c/2$ and $\langle \mu_2, u\rangle = \delta_c/2$. The within-mode covariance is $\Sigma_c^{\text{in}} := P_u^{\perp} + \sigma_u^2 P_u$, where $P_u := uu^{\top}$ and $P_u^{\perp} := \mathbf{I}_{d_c} - P_u$, so Σ_c^{in} has eigenvalue σ_u^2 along u and eigenvalue 1 in the remaining $d_c - 1$ tangential directions. The total marginal covariance of the equal-weight mixture is $\Sigma_c := P_u^{\perp} + \Lambda_c P_u$ with $\Lambda_c := \sigma_u^2 + \delta_c^2/4$, so $\operatorname{tr}\Sigma_c = (d_c - 1) + \Lambda_c$ and the eigenvalues of Σ_c are 1 (multiplicity $d_c - 1$) and Λ_c (multiplicity 1).

We use the following quantities throughout the proofs. $V_\tau^u := (1-\bar{\alpha}_\tau)+\bar{\alpha}_\tau\sigma_u^2$ is the marginal variance of $x_{\tau,u} := \langle x_\tau, u\rangle$ given either mode, arising from $x_\tau = \sqrt{\bar{\alpha}_\tau}x_0 + \sqrt{1-\bar{\alpha}_\tau}\varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$. $\rho := \frac{1}{4}\Phi(-\delta_c/2) > 0$ is a tail threshold, where Φ denotes the standard Gaussian CDF. $r_c := \sqrt{(d_c-1)+\Lambda_c} + \sqrt{2\Lambda_c \log(1/\rho)}$ is the mode basin radius, chosen so that $\Pr[\|x_0 - \mu_k\| \geq r_c] \leq \rho$ (derived below via the Borell–TIS inequality [Adler and Taylor, 2007]). $M_t := \widehat{x}_0(x_t, t; c) = \mathbb{E}[x_0 \mid x_t, c]$ is the Tweedie posterior mean. $\pi_k(x_\tau) := \Pr[\text{mode } k \mid x_\tau]$ is the posterior mode probability. $\Pi_{\mathcal{M}_c}(x) \in \{\mu_1, \mu_2\}$ denotes nearest-mode projection.

A.2 Standing Assumptions

(A1) Network regularity. Each block f_ℓ is C^2 along the sampling trajectory with $\|J_\ell\|_{\text{op}} \leq K_\ell$ and $\|\nabla^2 h_\ell\|_{\text{op}} \leq M_\ell$ a.s., where $J_\ell := \nabla_x h_\ell$.

(A2) Manifold regularity. $\mathcal{M}_c$ is a closed C^2 submanifold of $\mathcal{X}$ with positive reach $\rho_c > 0$ (guaranteed by Definition 1 when g is C^2 with the submersion condition holding).

(A3) Two-mode Gaussian mixture. $p(x_0 \mid c) = \frac{1}{2}\mathcal{N}(\mu_1, \Sigma_c^{\text{in}}) + \frac{1}{2}\mathcal{N}(\mu_2, \Sigma_c^{\text{in}})$ (notation as in §A.1), with the following conditions: (i) $0 < \sigma_u \leq 1$; (ii) $\Lambda_c > 1$ (equivalently $\delta_c > 2\sqrt{1-\sigma_u^2}$), ensuring all eigenvalues of Σ_c satisfy $\lambda_i \geq 1$; (iii) $r_c < \delta_c/2$, ensuring distinct modal supports; (iv) $\delta_c > 0$; (v) *(sharp threshold)* $\frac{(d_c-1)+\Lambda_c}{2(\delta_c/2-r_c)^2} < \frac{1}{2}\Phi(-\delta_c/2)$, ensuring a quantitative gap between the safe and unsafe mode-flip bounds at $\bar{\alpha}_\tau = \frac{1}{2}$.

(A4) ODE remainder. $\mathbb{E}\|R(\Delta t)\|^4 \leq C_R \Delta t^4$ for the Euler remainder $R(\Delta t)$ of the probability-flow ODE.

(A5) Directional Lipschitz constants and depth ordering. For each block ℓ and $t \leq T^*$, there exist $L_\ell^{\perp}, L_\ell^{\parallel} > 0$ with $\|J_\ell P_x^{\perp} v\| \leq L_\ell^{\perp}\|v\|$ for $v \in N_{\Pi(x)}\mathcal{M}_c$ and $\|J_\ell P_x^{\parallel} w\| \leq L_\ell^{\parallel}\|w\|$ for $w \in T_{\Pi(x)}\mathcal{M}_c$. Under unit-gain normalisation $L_\ell^{\perp} + L_\ell^{\parallel} = 1$, set $\beta_\ell := L_\ell^{\perp}$ and $L_\ell^{\parallel} = 1 - \beta_\ell$. The depth-ordering condition $\beta_{\ell_1} < \beta_{\ell_2}$ for $\ell_1 < \ell_2$ holds throughout the safe phase.

(A5′) Transversality (for Theorem 2, two-sided bound only). There exist $\theta_* \in (0, \pi/2]$, $\zeta \in (0, \cos\theta_*]$, and $\gamma_* > 0$ such that for all non-zero $J_\ell v$, $J_\ell w$ with $v \in N_{\Pi(x)}\mathcal{M}_c$, $w \in T_{\Pi(x)}\mathcal{M}_c$: $\langle J_\ell v, J_\ell w\rangle/(\|J_\ell v\|\|J_\ell w\|) \geq -\cos\theta_* + \zeta$, so $\alpha := \cos\theta_* - \zeta \geq 0$. Additionally: $\|J_\ell v\| \geq \underline{L}_\ell^{\perp}\|v\|$ and $\|J_\ell w\| \geq \underline{L}_\ell^{\parallel}\|w\|$, with $\gamma_* := \min(\underline{L}_\ell^{\perp}/L_\ell^{\perp}, \underline{L}_\ell^{\parallel}/L_\ell^{\parallel}) > 0$. The constraint $\zeta \leq \cos\theta_*$ (i.e., $\alpha \geq 0$) is necessary for the AM-GM step in the lower bound (Step 6).

(A6) Temporal regularity. $\mathbb{E}\,\delta_\ell(t_1)^2 \leq \eta\,\mathbb{E}\,\delta_\ell(t_2)^2$ for all t_1, t_2 within any cache stride window, for some $\eta \geq 1$.

A.3 Lemma 1: Exact Posterior Variance

Lemma 1 (MMSE upper bound)**.** *For any prior $p(x_0 \mid c)$ with mean zero and covariance Σ_c having eigenvalues $\{\lambda_i\}_{i=1}^{d_c}$, and the VP forward kernel $q(x_\tau \mid x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_\tau}x_0, (1-\bar{\alpha}_\tau)\mathbf{I})$, the Tweedie estimate $M_\tau = \mathbb{E}[x_0 \mid x_\tau]$ satisfies*

$$\mathbb{E}\|M_\tau - x_0\|^2 \;\leq\; \sum_{i=1}^{d_c} \frac{(1-\bar{\alpha}_\tau)\lambda_i}{\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau}.$$

Proof. The nonlinear MMSE estimator $M_\tau = \mathbb{E}[x_0 \mid x_\tau]$ achieves lower mean-squared error than any linear estimator. Work in the eigenbasis of Σ_c. The linear MMSE estimator of $x_0^{(i)}$ from $x_\tau^{(i)} =$

$\sqrt{\bar{\alpha}_\tau} x_0^{(i)} + \sqrt{1-\bar{\alpha}_\tau}\varepsilon^{(i)}$ has MSE $\mathrm{Cov}(x_0^{(i)}) - \mathrm{Cov}(x_0^{(i)}, x_\tau^{(i)})^2/\mathrm{Var}(x_\tau^{(i)}) = \lambda_i - \bar{\alpha}_\tau\lambda_i^2/(\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau) = (1-\bar{\alpha}_\tau)\lambda_i/(\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau)$, using $\mathrm{Cov}(x_0^{(i)}, x_\tau^{(i)}) = \sqrt{\bar{\alpha}_\tau}\lambda_i$ and $\mathrm{Var}(x_\tau^{(i)}) = \bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau$. Since $\mathbb{E}\|M_\tau - x_0\|^2 \le \mathbb{E}\|\hat{x}_{0,\mathrm{lin}} - x_0\|^2 = \sum_i (1-\bar{\alpha}_\tau)\lambda_i/(\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau)$, the result follows. □

A.4 Proof of Theorem 1

Setup. Freezing the sampler at cache time τ yields output $x_0^\tau = M_\tau := \widehat{x}_0(x_\tau, \tau; c)$. Since $M_t = \mathbb{E}[x_0 \mid x_t, c]$ is a reverse-time martingale (by the tower property: $\mathbb{E}[M_s \mid x_t] = \mathbb{E}[x_0 \mid x_t] = M_t$ for $s < t$) with $M_0 = x_0$ a.s., the cache error is $\mathcal{E}(\tau) = \mathbb{E}\|M_\tau - x_0\|^2$.

Part (i): Safe regime $\bar{\alpha}_\tau \ge \frac{1}{2}$. By Lemma 1 (MMSE $\le$ linear MMSE, valid for any prior with covariance Σ_c): $\mathcal{E}(\tau) \le \sum_i (1-\bar{\alpha}_\tau)\lambda_i/(\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau)$. Since $\lambda_i \ge 1$ (A3.ii), we have $\bar{\alpha}_\tau\lambda_i + 1 - \bar{\alpha}_\tau \ge 1$, so each term is bounded by $(1-\bar{\alpha}_\tau)\lambda_i$, giving

$$\mathcal{E}(\tau) \;\le\; (1-\bar{\alpha}_\tau)\,\mathrm{tr}\,\Sigma_c \;\le\; \tfrac{1}{2}\,\mathrm{tr}\,\Sigma_c, \tag{6}$$

where the last step uses $1 - \bar{\alpha}_\tau \le \frac{1}{2}$. For the mode-flip probability, two events must fail jointly. The *basin event* $\mathcal{A} = \{\|x_0 - \mu_k\| \le r_c\}$: writing $x_0 - \mu_k = \Sigma_c^{\mathrm{in},1/2} z$ with $z \sim \mathcal{N}(0, \mathbf{I})$, the function $f(z) := \|\Sigma_c^{\mathrm{in},1/2} z\|$ is $\sqrt{\Lambda_c}$-Lipschitz (the largest eigenvalue of $\Sigma_c^{\mathrm{in},1/2}$ is $\sqrt{\Lambda_c}$), with $\mathbb{E}f \le \sqrt{\mathrm{tr}\,\Sigma_c^{\mathrm{in}}} \le \sqrt{(d_c - 1) + \Lambda_c}$. The Borell–TIS inequality [Adler and Taylor, 2007] gives $\Pr[\|x_0 - \mu_k\| \ge \mathbb{E}f + t] \le e^{-t^2/(2\Lambda_c)}$. Setting $e^{-t^2/(2\Lambda_c)} = \rho$ gives $t = \sqrt{2\Lambda_c \log(1/\rho)}$, so $\Pr[\mathcal{A}^c] \le \rho$. The *accuracy event* $\mathcal{B} = \{\|M_\tau - x_0\| < \delta_c/2 - r_c\}$: by Markov and (6), $\Pr[\mathcal{B}^c] \le (1-\bar{\alpha}_\tau)\,\mathrm{tr}\,\Sigma_c/(\delta_c/2 - r_c)^2$, where $\delta_c/2 - r_c > 0$ by (A3.iii). On $\mathcal{A} \cap \mathcal{B}$, the triangle inequality gives $\|M_\tau - \mu_k\| < (\delta_c/2 - r_c) + r_c = \delta_c/2$, so no mode flip occurs. A union bound yields the stated safe mode-flip bound.

Part (ii): Unsafe regime $\bar{\alpha}_\tau < \frac{1}{2}$. Fix mode $k = 1$ (by symmetry the bound applies to both modes). In centred coordinates ($\mu_1 = -\frac{\delta_c}{2}u$, $\mu_2 = \frac{\delta_c}{2}u$), write $x_{\tau,u} := \langle x_\tau, u\rangle$. The marginal distribution of $x_{\tau,u}$ given mode k is $\mathcal{N}(\sqrt{\bar{\alpha}_\tau}\langle \mu_k, u\rangle, V_\tau^u)$.

Step 1: Log-posterior ratio. Since the two mode-conditional marginals of $x_{\tau,u}$ are Gaussian with the same variance V_τ^u and means differing by $\sqrt{\bar{\alpha}_\tau}\delta_c$:

$$\log\frac{\pi_2(x_\tau)}{\pi_1(x_\tau)} \;=\; \log\frac{p(x_{\tau,u} \mid \text{mode } 2)}{p(x_{\tau,u} \mid \text{mode } 1)} \;=\; \frac{\sqrt{\bar{\alpha}_\tau}\delta_c}{V_\tau^u}\, x_{\tau,u},$$

so $\pi_2 \ge \frac{1}{2}$ if and only if $x_{\tau,u} \ge 0$.

Step 2: Mode-flip probability. Conditional on mode 1: $x_{\tau,u} \sim \mathcal{N}(-\sqrt{\bar{\alpha}_\tau}\delta_c/2,\, V_\tau^u)$. Since $\bar{\alpha}_\tau > 0$ and $\delta_c > 0$:

$$\Pr[x_{\tau,u} \ge 0 \mid \text{mode } 1] \;=\; \Phi\Big(\frac{\sqrt{\bar{\alpha}_\tau}\delta_c/2}{\sqrt{V_\tau^u}}\Big) \;>\; \Phi(0) = \tfrac{1}{2} \;>\; \Phi(-\delta_c/2),$$

so a majority of mode-1 trajectories already have $\pi_2 \ge \frac{1}{2}$ at cache time $\tau > T^*$.

Step 3: Tweedie flip (Lemma 2).

Lemma 2 (Tweedie mode flip)**.** *Whenever $x_{\tau,u} \ge 0$, the Tweedie estimate satisfies $\langle M_\tau, u\rangle \ge 0$, i.e., $\Pi_{\mathcal{M}_c}(M_\tau) = \mu_2$.*

Proof. The Tweedie formula gives $M_\tau = \pi_1 m_1 + \pi_2 m_2$, where $m_k := \mathbb{E}[x_0 \mid x_\tau, \text{mode } k]$. By the Gaussian regression formula ($\mathrm{Cov}(x_{0,u}, x_{\tau,u}) = \sqrt{\bar{\alpha}_\tau}\sigma_u^2$, $\mathrm{Var}(x_{\tau,u}) = V_\tau^u$):

$$\langle m_k, u\rangle \;=\; \langle \mu_k, u\rangle + \frac{\sqrt{\bar{\alpha}_\tau}\sigma_u^2}{V_\tau^u}(x_{\tau,u} - \sqrt{\bar{\alpha}_\tau}\langle \mu_k, u\rangle) \;=\; \frac{\sqrt{\bar{\alpha}_\tau}\sigma_u^2\, x_{\tau,u} + (1-\bar{\alpha}_\tau)\langle \mu_k, u\rangle}{V_\tau^u}.$$

Hence $\langle m_2 - m_1, u\rangle = (1-\bar{\alpha}_\tau)\delta_c/V_\tau^u > 0$, and $\langle m_1, u\rangle = [\sqrt{\bar{\alpha}_\tau}\sigma_u^2\, x_{\tau,u} - (1-\bar{\alpha}_\tau)\delta_c/2]/V_\tau^u$. On $\{x_{\tau,u} \ge 0\}$, we have $\pi_2 \ge \frac{1}{2}$, so:

$$\begin{aligned}\langle M_\tau, u\rangle &= \langle m_1, u\rangle + \pi_2\langle m_2 - m_1, u\rangle \\ &\ge \frac{\sqrt{\bar{\alpha}_\tau}\sigma_u^2\, x_{\tau,u} - (1-\bar{\alpha}_\tau)\delta_c/2}{V_\tau^u} + \frac{(1-\bar{\alpha}_\tau)\delta_c}{2V_\tau^u} \;=\; \frac{\sqrt{\bar{\alpha}_\tau}\sigma_u^2\, x_{\tau,u}}{V_\tau^u} \;\ge\; 0,\end{aligned}$$

where the final inequality uses $x_{\tau,u} \ge 0$. □

By Steps 2–3, a mode flip occurs whenever $x_{\tau,u} \geq 0$, which happens with probability $\Phi(\sqrt{\bar{\alpha}_\tau}\delta_c/(2\sqrt{V_\tau^u})) \geq \Phi(-\delta_c/2)$ under mode 1. Let $\mathcal{A}_1 = \{\|x_0 - \mu_1\| \leq r_c\}$ and $\mathcal{F}_1 = \{\Pi_{\mathcal{M}_c}(M_\tau) = \mu_2\}$. By a union bound: $\Pr[\mathcal{A}_1 \cap \mathcal{F}_1 \mid \text{mode } 1] \geq \Phi(-\delta_c/2) - \rho = \frac{3}{4}\Phi(-\delta_c/2) =: p_{\text{mix}}$. By symmetry (equal priors, equal covariances), $\Pr[\Pi_{\mathcal{M}_c}(x_0^\tau) \neq \Pi_{\mathcal{M}_c}(x_0)] \geq p_{\text{mix}} > 0$. On $\mathcal{A}_1 \cap \mathcal{F}_1$: $\langle M_\tau, u\rangle \geq 0$ and $\langle x_0, u\rangle \leq -\delta_c/2 + r_c$, so $\|M_\tau - x_0\| \geq \delta_c/2 - r_c > 0$ by (A3.iii), giving $\mathcal{E}(\tau) \geq p_{\text{mix}}(\delta_c/2 - r_c)^2$.

Part (iii): Sharpness at T^*. Evaluating the safe mode-flip bound at $\bar{\alpha}_\tau = \frac{1}{2}$ gives $U_{\text{safe}} := \operatorname{tr}\Sigma_c/(2(\delta_c/2 - r_c)^2) + \rho$. The unsafe lower bound is $p_{\text{mix}} = \frac{3}{4}\Phi(-\delta_c/2)$. Their gap is $p_{\text{mix}} - U_{\text{safe}} = \frac{1}{2}\Phi(-\delta_c/2) - \operatorname{tr}\Sigma_c/(2(\delta_c/2 - r_c)^2) > 0$ by (A3.v), establishing T^* as a quantitative phase boundary. □

A.5 Proof of Theorem 2

Step 1: Linearise block drift. By Taylor's theorem and (A1): $h_\ell(x_{t'}, t, c) - h_\ell(x_t, t, c) = J_\ell \Delta x + E_\ell$, with $\Delta x := x_{t'} - x_t$ and $\|E_\ell\| \leq \frac{M_\ell}{2}\|\Delta x\|^2 = O(\Delta t^2)$. Squaring, taking expectations, and bounding the cross term via Cauchy–Schwarz (which gives $O(\Delta t^3)$):

$$\mathbb{E}\,\delta_\ell(t)^2 = \mathbb{E}\|J_\ell \Delta x\|^2 + O(\Delta t^3). \tag{7}$$

Step 2: Decompose the Euler step. The probability-flow ODE advances as $x_{t'} = x_t + \Delta t\, A_t s_t(x_t; c) + R(\Delta t)$, so $\Delta x = \Delta t A_t s_t + R(\Delta t)$. Using $s_t = s_t^\perp + s_t^\parallel$ and the projectors: $P_x^\perp \Delta x = \Delta t A_t s_t^\perp + P_x^\perp R$ and $P_x^\parallel \Delta x = \Delta t A_t s_t^\parallel + P_x^\parallel R$. With $\Theta_t := A_t^2 \mathbb{E}\|s_t^\parallel\|^2 > 0$ and $\kappa_t := \mathbb{E}\|s_t^\perp\|^2/\mathbb{E}\|s_t^\parallel\|^2$, Assumption (A4) gives $\mathbb{E}\|P_x^\perp R\|^2, \mathbb{E}\|P_x^\parallel R\|^2 = O(\Delta t^4)$, so:

$$\mathbb{E}\|P_x^\perp \Delta x\|^2 = \Delta t^2 \Theta_t \kappa_t + O(\Delta t^3), \quad \mathbb{E}\|P_x^\parallel \Delta x\|^2 = \Delta t^2 \Theta_t + O(\Delta t^3). \tag{8}$$

Step 3: Upper bound on drift. By (A5) and the triangle inequality: $\|J_\ell \Delta x\| \leq L_\ell^\perp \|P_x^\perp \Delta x\| + L_\ell^\parallel \|P_x^\parallel \Delta x\|$. Using $(a+b)^2 \leq 2(a^2 + b^2)$, taking expectations, and substituting (8) into (7):

$$\mathbb{E}\,\delta_\ell(t)^2 \;\leq\; \overline{B}_\ell\, \Delta t^2 \Theta_t + O(\Delta t^3), \quad \overline{B}_\ell := 2[(L_\ell^\perp)^2 \kappa_t + (L_\ell^\parallel)^2] = 2B(\beta_\ell; \kappa_t). \tag{9}$$

Step 4: Certified strides. For ε-cacheability with stride S, it suffices that $(S-1)\mathbb{E}\delta_\ell \leq \varepsilon$. Since $\mathbb{E}\delta_\ell \leq \sqrt{\mathbb{E}\delta_\ell^2} \leq \sqrt{\overline{B}_\ell \Delta t^2 \Theta_t} + O(\Delta t^{3/2}) =: \overline{D}_\ell$ (Jensen), the certified stride $\widehat{S}_\ell := 1 + \varepsilon/\overline{D}_\ell \leq S_\ell^\star$ is a valid lower bound on the true maximum cacheable stride.

Step 5: Depth-monotone comparison. For $\ell_1 < \ell_2$ with $\beta_{\ell_1} < \beta_{\ell_2}$ (A5):

$$\frac{\widehat{S}_{\ell_2} - 1}{\widehat{S}_{\ell_1} - 1} = \frac{\overline{D}_{\ell_1}}{\overline{D}_{\ell_2}} = \sqrt{\frac{B(\beta_{\ell_1}; \kappa_t)}{B(\beta_{\ell_2}; \kappa_t)}}(1 + O(\Delta t)).$$

This ratio exceeds one iff $B(\beta_{\ell_1}; \kappa_t) > B(\beta_{\ell_2}; \kappa_t)$, i.e., $(\beta_{\ell_1}^2 - \beta_{\ell_2}^2)\kappa_t > (1 - \beta_{\ell_2})^2 - (1 - \beta_{\ell_1})^2$. Dividing both sides by $(\beta_{\ell_2}^2 - \beta_{\ell_1}^2) > 0$ and negating: $\kappa_t < \kappa^* := [(1 - \beta_{\ell_1}) + (1 - \beta_{\ell_2})]/(\beta_{\ell_1} + \beta_{\ell_2})$, which is (4). Since $\kappa_t \to 0$ as $t \to 0$, the condition holds on $[0, t_0] \cap [0, T^*]$ for some $t_0 > 0$, establishing (3) and the certified stride comparison. □ (Theorem 2a)

Step 6: Actual strides under transversality. Write $v := P_x^\perp \Delta x$ and $w := P_x^\parallel \Delta x$. From (A5′) with $\alpha := \cos\theta_* - \zeta \geq 0$: $2\langle J_\ell v, J_\ell w\rangle \geq -2\alpha\|J_\ell v\|\|J_\ell w\| \geq -\alpha(\|J_\ell v\|^2 + \|J_\ell w\|^2)$, where the last step applies AM-GM and uses $\alpha \geq 0$. Hence:

$$\|J_\ell \Delta x\|^2 = \|J_\ell v\|^2 + \|J_\ell w\|^2 + 2\langle J_\ell v, J_\ell w\rangle \;\geq\; (1 - \alpha)(\|J_\ell v\|^2 + \|J_\ell w\|^2).$$

Applying the lower-Lipschitz bounds ($\|J_\ell v\| \geq \underline{L}_\ell^\perp \|v\| \geq \gamma_* \beta_\ell \|v\|$ and similarly for w):

$$\mathbb{E}\,\delta_\ell^2(t) \;\geq\; \underline{B}_\ell\, \Delta t^2 \Theta_t - O(\Delta t^3), \quad \underline{B}_\ell := (1 - \alpha)\gamma_*^2[\beta_\ell^2 \kappa_t + (1 - \beta_\ell)^2] = \tfrac{(1-\alpha)\gamma_*^2}{2}\overline{B}_\ell. \tag{10}$$

Combining (9)–(10) with temporal regularity (A6) gives $S_\ell^\star \asymp 1 + c\varepsilon/(\overline{B}_\ell \Delta t^2 \Theta_t)^{1/2}$ with $c = \gamma_* \sqrt{(1 - \alpha)/(2\eta)} > 0$. The same comparison as Step 5 then gives $S_{\ell_2}^\star / S_{\ell_1}^\star > 1$ for $\kappa_t < \kappa^*$, completing Theorem 2b. □

Figure 3: From left to right: MANIFOLDCACHE, PAB, and AdaCache. It can be seen that MANIFOLDCACHE achieves the best visual quality on Lyra.

A.6 Discussion

Theorem 1 is proved under a two-mode Gaussian mixture (A3), which provides the minimal tractable setting for explicit mode-confusion bounds. For multi-modal distributions the pairwise argument applies to any two modes with separation $\delta_c > 2r_c$; the MMSE upper bound $\sum_i (1-\bar{\alpha}_\tau)\lambda_i/(\bar{\alpha}_\tau\lambda_i + 1-\bar{\alpha}_\tau)$ from Lemma 1 holds for any prior with covariance Σ_c. Assumption (A3.ii) ($\lambda_i \geq 1$) is a coordinate normalisation used only for the simplified bound $\frac{1}{2}\operatorname{tr}\Sigma_c$. Theorem 2 is conditional on the depth-ordering (A5) and on $\kappa_t < \kappa^*$ (4); the latter is a regime condition with explicit form, holding when trajectories are sufficiently close to $\mathcal{M}_c$. Assumption (A3.v) ensures a quantitative gap at T^* whenever modes are well-separated relative to r_c. Extensions to flow-matching models require redefining T^* via the corresponding SNR parameterisation; for stochastic samplers, injected noise adds $O(\sqrt{1-\bar{\alpha}_t})$ per cached step and is left to future work.

B Qualitative Results

Figure 3 shows that MANIFOLDCACHE achieves the best visual quality on Lyra, particularly in preserving fine details.

C Limitations and Future Work

Despite strong empirical and theoretical performance, MANIFOLDCACHE has several important limitations that motivate future work. First, our theory relies on smooth closed manifold assumptions with well-defined tangent and normal decompositions. While these assumptions cover a broad range of scientific diffusion models, some structured generation problems involve non-smooth, discrete, combinatorial, or dynamically changing constraint sets where the local geometric decomposition may not exist or may vary discontinuously across the trajectory. **Extending manifold-aware caching to hybrid discrete–continuous diffusion processes, topology-changing manifolds, or non-Euclidean latent spaces remains an open direction**. Second, the theoretical guarantees are derived under simplified regularity and mixture assumptions designed to make the analysis tractable. Real CMDMs exhibit highly multimodal, non-Gaussian, and hierarchical distributions whose geometry may evolve across denoising stages. Although the empirical results strongly support the predicted phase transition at T^*, **deriving sharper guarantees under more realistic score models, heavy-tailed distributions, or learned manifold geometries would strengthen the theoretical foundations of geometry-aware acceleration**. Third, the current schedule is analytically fixed from the noise schedule and block depth and does not adapt online to individual samples, prompts, or trajectory difficulty. Certain trajectories may approach the manifold substantially earlier or later than predicted by the global schedule, especially in highly multimodal settings. **Future work could investigate sample-adaptive or uncertainty-aware caching policies that dynamically estimate manifold proximity, score curvature, or feature drift during inference while retaining the low-overhead nature of the current framework**. Fourth, although MANIFOLDCACHE is backbone-agnostic, the present implementation focuses on inference-time feature reuse and does not jointly optimize with other acceleration mechanisms such as learned schedulers, consistency distillation, speculative denoising, low-rank adaptation, or mixed-precision execution. **Combining geometry-aware caching with solver acceleration, quantization, or token sparsification could potentially yield multiplicative speedups beyond those observed here**. Fifth, our experiments focus on single-device inference on existing pretrained checkpoints. Large-scale distributed generation systems introduce additional communication and synchronization bottlenecks that may alter the trade-off between recomputation and cache reuse. **Investigating geometry-aware caching under tensor parallelism, pipeline parallelism, or stream-**

ing generation settings is an important systems-level extension. Finally, our work studies geometry primarily as a mechanism for safe acceleration. **A broader future direction is to treat manifold structure as a unifying principle connecting controllability, robustness, interpretability, and efficiency in scientific generative models**. We believe CMDMs provide a promising foundation for studying these properties jointly, particularly in emerging multimodal settings combining physical simulation, language conditioning, molecular reasoning, and embodied 3D generation.