No Single Basis Wins: A Cross-Family Study of Diffusion Feature Forecasting and the Limits of Training-Free Basis Selection

Krishi Attri
Soft Robotics and Bionics Laboratory, Seoul National University
ORCID 0009-0005-4695-6467 kattri@snu.ac.kr

Abstract

Feature-caching methods accelerate diffusion and flow-matching samplers by skipping the network on most denoising steps and *forecasting* the cached features from recent compute-step anchors. The literature has treated the forecast basis as a ladder to climb: monomials (TaylorSeer), scaled Hermite polynomials (HiCache), rational and multistep variants (FoCa, HyCa), global Chebyshev fits (Spectrum), each rung validated on a single sampler family. We went to the ladder's mathematical endpoint, the exponential solution class of the local feature-ODE, fitted it with rank-truncated Dynamic Mode Decomposition (Prony's method in its modern multivariate form), and expected it to win everywhere the polynomials do. It does not. The exponential basis wins on every flow-matching 3D generator we test (+0.13/+0.24 F-score over the deployed Hermite arm at intervals 5/6 on Hunyuan3D-2.1; geometry-lossless through interval 6 on SAM3D at 1.56×), yet on DiT-XL/2 ImageNet-256 with the literature-standard 250-step DDPM sampler the ranking inverts outright: a sign-correct TaylorSeer is near-lossless (paired-noise FID drift 2.27 at 3.81×), the corrected Hermite sweeps the interval ladder (3.54/6.46/10.74 at intervals 4/6/8), and the exponential basis drifts 5–9× more than the corrected polynomial at every interval. No basis in our study wins both families. The obvious remedy fails too. Training-free holdout selection, which is exact on synthetic regime switches (120/120 windows) and harmless to helpful on the 3D generators, serves the losing exponential arm on DiT in *both* holdout modes, at FID drift 18.11 where the corrected polynomial sits at 3.54; this refutes a prediction we pre-registered before the deciding run, that a horizon-matched holdout (which recovers the oracle 20/20 on the matched synthetic regime) would fix the cell. The mechanism deserves to be stated plainly: the richer exponential parameterization fits, and therefore backcasts, the in-window history better at any holdout distance while extrapolating forward worse, so in-window backcast fit carries no signal about forward extrapolation quality on real features. We close with an uncomfortable methodological finding. A one-character sign bug made our Hermite baseline *anti-extrapolative*, and it survived every end-to-end benchmark we and our integrations ran, because near-reuse fails safe: quality metrics cannot distinguish a forecaster from a damped cache, and hyperparameter tuning actively conceals the defect. We propose directional closed-form regression tests as a community-level remedy, and report every affected number as-released alongside corrected values.

1 Introduction

Diffusion and flow-matching models dominate generative modeling of images, video, and 3D shapes, but their iterative samplers remain expensive: 20–250 network evaluations per sample, each a full transformer forward. A productive line of training-free acceleration exploits the temporal redundancy of the sampler itself. The features computed at adjacent denoising steps are highly correlated, so one can *cache* features at sparse "anchor" steps and *skip* the network in between. Early methods reused the cache directly; TaylorSeer [1] reframed reuse as degree-zero *forecasting* and showed that higher-order polynomial extrapolation of the cached features is markedly better, establishing the cache-then-forecast paradigm. Successors refine the forecast basis: HiCache [2] stabilizes the expansion with dual-scaled Hermite polynomials; FoCa [3] treats caching as a stiff-ODE integration problem; HyCa [4] routes feature dimensions to a pool of classical multistep solvers; Spectrum [6] replaces the local Taylor window with a global Chebyshev fit. Each is presented as a better basis, each is evaluated on one sampler family, and each implicitly assumes that a single basis can win.

This paper interrogates that assumption rather than adding a rung to the ladder. We organize the study around three research questions.

RQ1: Is the forecast-basis ranking a property of caching, or of the sampler family? The ladder's mathematical endpoint is identifiable: if, across a short window of steps, the cached feature stream evolves under a slowly varying near-linear operator, the trajectory locally solves a linear feature-ODE whose exact solution class is a finite sum of damped or oscillatory *exponentials*, of which every polynomial is only a local truncation that diverges under extrapolation. Section 3 states this as a falsifiable hypothesis with its two scope conditions, and we build the strongest implementation we can, rank-truncated Dynamic Mode Decomposition (DMD) [12, 13], the noise-robust multivariate descendant of Prony's method [10], so that a refutation cannot be blamed on a weak estimator. The cross-family test (Sec. 5) then returns a split, not a winner (Fig. 1). On four flow-matching 3D generators (Hunyuan3D-2.1 and -2-mini, SAM3D, TRELLIS v1/v2) the exponential basis behaves exactly as the feature-ODE argument predicts: it degrades gracefully where the deployed polynomial arm collapses (+0.125/+0.241 F-score at intervals 5/6 on Hunyuan3D-2.1), is exactly lossless at interval 5 on Hunyuan3D-2-mini, and stays geometry-lossless through interval 6 on SAM3D at 1.56×. On DiT-XL/2 ImageNet-256 with the literature-standard 250-step DDPM sampler, the ranking inverts outright: a sign-correct TaylorSeer monomial forecast is near-lossless (paired-noise FID drift 2.27 at 3.81× speedup), the corrected Hermite sweeps the interval ladder (3.54/6.46/10.74 at intervals 4/6/8), and the exponential basis drifts 1.7–1.9× more than even a near-reuse Hermite control, 5–9× more than the corrected one, at every interval tested (Table 3). The universal reading of the feature-ODE hypothesis is falsified; the conditional reading survives. No basis in our study wins both families, and we argue no published basis result should be extrapolated across families without a cross-family benchmark.

RQ2: Can training-free selection recover the per-family optimum? Given the split, the natural remedy is a *selection mechanism* rather than a basis: backcast a held-out snapshot the schedule already paid for with both bases, serve whichever demonstrably wins on the trajectory at hand, at zero model cost. We evaluate this mechanism to its honest conclusion (Sec. 6). On synthetic regime switches it is essentially exact (120/120 windows); on the 3D generators it is harmless to helpful. On DiT it fails, twice. The default 1-step holdout serves the exponential arm (FID drift 18.08) where the corrected polynomial sits at 3.54. We reconstructed that inversion in a controlled oscillatory-with-trend regime, found a *horizon-matched* holdout that recovers the oracle arm 20/20 there (the 1-step test picks the losing arm in 12/20), pre-registered the prediction that horizon matching would fix the DiT cell, and ran the A/B: the horizon variant also serves the exponential arm and lands at FID drift 18.11, statistically identical to the 1-step holdout. The prediction is *refuted*, and the refutation is more informative than the fix would have been. On this family's features the exponential model's strictly richer parameterization fits, and therefore backcasts, the in-window history better at *any* holdout distance while extrapolating forward worse (Fig. 2), so in-window backcast fit carries no signal about forward extrapolation quality: model selection by training error, locally disguised as validation. The practical recommendation is basis-by-model-family defaults, with holdout selection as a regime-switch safety net; selection from real feature trajectories is an open problem.

RQ3 (methodology): What does it take for a caching benchmark to detect a directionally wrong forecaster? While building the polynomial baseline we introduced a one-character porting bug that evaluated the Hermite basis at $-k$ instead of $+k$: every odd-order term flipped sign and the shipped forecast extrapolated *backwards* by exactly the amount it should have moved forward (Fig. 3). The bug survived every end-to-end benchmark we and our integrations ran, inverted one benchmark conclusion, and was actively concealed by hyperparameter tuning. We treat this not as an erratum but as a measurement about evaluation practice (Sec. 7): end-to-end quality ladders are structurally blind to anti-extrapolation because near-reuse fails safe, and the remedy, directional closed-form regression tests, is a one-line check any caching paper can adopt. All affected numbers are reported as-released alongside corrected values throughout.

Contributions. All four are knowledge claims; the library that produced them is an artifact, not a contribution (Appendix A).

- **Refutation of basis universality** (Sec. 5): the first cross-family evidence that the forecast-basis ranking is a property of the sampler family and its feature dynamics, not of caching per se. The exponential (feature-ODE) basis wins on flow-matching 3D generation and loses to a sign-correct polynomial everywhere on DiT-class denoising; published single-family basis rankings should be presumed family-conditional.
- **A conditional validation of the feature-ODE hypothesis** (Secs. 3, 5): stated falsifiably with two scope conditions (frozen-propagator locality; pole-estimation error vs. polynomial truncation error at the served horizons), the hypothesis predicts the 3D-family win and its own DiT-family failure. The scope conditions, not the class argument, carry the predictive content.
- **A negative result on training-free basis selection, with its mechanism** (Sec. 6): per-window holdout selection is exact on synthetic regime switches yet serves the losing arm on DiT in both holdout modes, refuting our pre-registered horizon-matching prediction. Characterization: in-window backcast fit does not predict forward extrapolation quality on real features, because the holdout rewards exactly the in-sample flexibility that hurts at serve time. Any future selector must demonstrate a signal that correlates with *forward* error.
- **A methodological finding about caching benchmarks** (Sec. 7): end-to-end quality metrics cannot distinguish a forecaster from a damped cache, so directionally wrong baselines survive entire benchmark suites and invert conclusions, and tuning conceals them. We propose directional closed-form regression tests as a reporting standard for forecast-basis papers.

2 Related work

Cache-then-forecast. TaylorSeer [1] maintains backward finite differences $\Delta^i \mathbf{F}_t$ at compute steps and serves the Taylor polynomial $\hat{\mathbf{F}}_{t+k} = \sum_i \frac{\Delta^i \mathbf{F}_t}{i!} k^i$ on skipped steps. HiCache [2] keeps the same differences but evaluates them against dual-scaled physicists' Hermite polynomials $\tilde{H}_n(x) = \sigma^n H_n(\sigma x)$, $\sigma \in (0, 1)$, which bounds the high-order terms and empirically stabilizes the forecast; it is the strongest published polynomial basis and our primary baseline. FoCa [3] integrates the feature-ODE with a BDF2 predictor and Heun corrector, a multistep *polynomial* integrator with better stability constants but the same function class. HyCa [4] observes that different feature dimensions prefer different solvers and routes dimensions to a pool of classical multistep schemes; the pool is again polynomial. SpeCa [5] adds a cheap verifier that rolls back bad forecasts, orthogonal to the basis and composable with ours. The cache-*decision* family (TeaCache [7], MagCache [8], EasyCache [9]) adaptively chooses which steps to skip; any of them can drive any forecaster discussed here. The gap this paper addresses: none of these works evaluates its basis across sampler families, none can therefore distinguish a property of its basis from a property of its benchmark family, and none reports a directional test that would catch an anti-extrapolative implementation (Sec. 7). Our results indicate the rankings they report are family-conditional.

Spectrum: the closest method. Spectrum [6] fits a *global* Chebyshev approximation to each feature trajectory over the full sampling horizon and serves the fitted polynomial on skipped steps, with an attractive non-compounding error bound. The contrast with the exponential forecaster is the classical approximation-theory split: Chebyshev polynomials are near-minimax for a bounded function on a closed interval but still diverge outside the fitted window, while a sum of exponentials with $|\lambda_j| \leq 1$ is bounded for all forward horizons. Our domain-split finding sharpens what a fair head-to-head must report: not a single operating point but the error-versus-interval curve *per sampler family*, at matched anchor budgets and wall-clock, including noise sensitivity and non-autonomous schedules. Until that head-to-head is run we claim no ordering against Spectrum in either direction.

Classical spectral estimation. Prony's 1795 method [10] recovers $\{a_j, \lambda_j\}$ in $f_t = \sum_j a_j \lambda_j^t$ from $2r$ uniform samples by rooting an annihilating polynomial; it is exact on the class but noise-fragile. The Matrix-Pencil method [11] replaces root-finding with a generalized eigenproblem and is substantially more robust. DMD [12, 13] generalizes both to vector-valued snapshots: it estimates the best-fit linear propagator

in a rank-truncated subspace and reads the poles off its eigenvalues. We use exact DMD with spectrum-based rank truncation. None of this machinery is new; its application to diffusion feature caching, and the finding that its validity is family-dependent, are.

3 Hypotheses

We state the two substantive hypotheses before any experiment, each in falsifiable form, together with the observation that would refute it. RQ3 is a methodological question rather than a hypothesis; its case-study evidence is presented in Sec. 7.

3.1 H1: the feature-ODE hypothesis

Let $\mathbf{F}_t$ denote a cached feature stream (in our experiments, the classifier-free-guidance combined velocity) indexed by sampler step t. Within a short window of steps, write the evolution of the stream as $\dot{\mathbf{F}} = M(t)\,\mathbf{F}$ with $M(t)$ slowly varying. Freezing M over the window gives the exact solution

$$\mathbf{F}_t \;=\; \sum_{j=1}^{r} a_j\, e^{\mu_j t}, \qquad a_j \in \mathbb{C}^d,\ \mu_j \in \mathbb{C}, \tag{1}$$

a finite sum of exponentials, damped when $\operatorname{Re}\mu_j < 0$ and oscillatory when $\operatorname{Im}\mu_j \neq 0$. Sampled at unit spacing this is $\mathbf{F}_t = \sum_j a_j \lambda_j^t$ with poles $\lambda_j = e^{\mu_j}$.

Proposition 1 (Polynomial divergence on the class)**.** *Let p_m be any degree-m polynomial agreeing with a trajectory of class (1) (with some $|\lambda_j| \neq 1$ or $\operatorname{Im}\mu_j \neq 0$) on $m{+}1$ anchors. Then the forecast error $\|p_m(t+k) - \mathbf{F}_{t+k}\|$ grows without bound in the horizon k, with leading order k^{m+1} times the first neglected derivative.*

This is the Lagrange remainder applied to (1); Hermite scaling, Chebyshev re-basing, BDF multistep weights, and rational Padé constructions change the constants but not the conclusion. By contrast an exponential fit with the correct poles is exact for all k, and with estimated poles the error grows only through the pole-estimation error while remaining bounded whenever $|\hat{\lambda}_j| \leq 1$.

Hypothesis 1 (Exponential-basis superiority)**.** *On cached diffusion feature streams, an exponential forecast fitted on the live snapshot window achieves lower forecast error at the served horizons than any polynomial forecast at matched window, and therefore lower end-to-end quality drift at matched skip schedule.*

H1 is falsifiable precisely because the class argument is conditional, on two scope conditions that any test must respect and that any refutation will implicate. First, the proposition is a statement about the *class*, conditional on the frozen-M hypothesis; whether a given sampler family's measured feature stream is in the regime where the polynomial divergence dominates *at the horizons actually served* is an empirical question. Second, with estimated poles the exponential forecast inherits pole-estimation error that compounds geometrically in the horizon; on trajectories that a low-order polynomial already captures at the served horizons, that estimation error can exceed the truncation error the polynomial pays. A family on which H1 fails through the second condition would exhibit a specific signature: polynomial drift small in absolute terms, exponential drift worse than *near-reuse*. Section 5.3 reports exactly this signature on DiT, and Sec. 5.4 renders the verdict: H1 holds on the flow-matching 3D family and is refuted as a universal claim.

3.2 H2: the holdout-selection hypothesis

Hypothesis 2 (Training-free selection)**.** *The winning basis for an upcoming skip window can be identified at zero model cost by backcasting a held-out snapshot, ground truth the schedule already paid for, with both bases and serving whichever reproduces it with smaller relative error.*

H2 is attractive because it requires no prior knowledge of the family and adapts within a run. It carries an identifiable failure mode, stated here because the experiments turn on it: a holdout drawn from the same short history the model was fit on measures *in-window* fit, and a basis with a richer parameterization can win

the backcast while losing the forward extrapolation. Whether that failure mode is realized on real feature streams is the empirical content of RQ2. We additionally pre-registered a specific prediction before running the deciding experiment (Sec. 6): **P1 (pre-registered)**: *if* the 1-step holdout's DiT failure is a horizon mismatch (ranking at backcast distance 1 inverting against the served distance $N-1$), *then* a horizon-matched holdout, which selects 20/20 on the matched synthetic regime, will recover the polynomial arm on DiT. Section 6 resolves P1: *refuted*.

4 Experimental apparatus

This section specifies the instruments needed to test H1 and H2 fairly: the schedule (held fixed), the strongest exponential estimator we could build (so a refutation of H1 indicts the class, not the estimator), its validity conditions (so DMD is only ever served where its assumptions hold), and the two holdout selectors. Implementation and cost engineering are deferred to Appendix A; nothing in the paper's claims rests on them beyond the matched-wall-clock accounting cited in Table 3.

4.1 The cache-then-forecast skeleton, held fixed

We adopt the HiCache/TaylorSeer schedule unchanged. Sampling runs T steps; the first E (`first_enhance`) steps are always computed; thereafter one step in every N (the *interval*) is computed and the remaining $N-1$ are skipped. At a compute step the network runs and the CFG-combined velocity $\mathbf{F}_t \in \mathbb{R}^d$ is recorded; at a skipped step a forecast $\hat{\mathbf{F}}_t$ is served to the integrator in its place. The only experimental manipulation is the forecast formula and, for RQ2, which formula gets served; everything else is controlled.

4.2 The exponential estimator: rank-truncated DMD

At each compute step we append $\mathbf{F}_t$ to a snapshot buffer of the last $n+1 \le 6$ compute-step velocities (a deliberately *short* window: Sec. 3.1 froze $M(t)$ only locally, and a long window would average over drifting dynamics). Stack the buffer into

$$X = [\mathbf{F}_0, \dots, \mathbf{F}_{n-1}] \in \mathbb{R}^{d\times n}, \qquad X' = [\mathbf{F}_1, \dots, \mathbf{F}_n] \in \mathbb{R}^{d\times n},$$

where indices are in buffer order and $d \gg n$. The exact-DMD estimate of the one-spacing propagator A ($X' \approx AX$) proceeds:

$$\begin{aligned} &X = U\Sigma V^H \quad \text{(economy SVD)}, \qquad r = \#\{\, i :\ \sigma_i > 10^{-4}\sigma_1 \,\} \quad \text{(rank truncation)},\\ &\tilde{A} = U_r^H X' V_r (\Sigma_r + \rho I)^{-1} \in \mathbb{C}^{r\times r}, \qquad \tilde{A} = W\Lambda W^{-1} \quad \text{(eigendecomposition)},\\ &\Phi = X' V_r (\Sigma_r + \rho I)^{-1} W \in \mathbb{C}^{d\times r}, \qquad b = \Phi^{+}\mathbf{F}_n \quad \text{(amplitudes at the newest snapshot)}, \end{aligned} \tag{2}$$

with a small ridge $\rho = 10^{-8}$, computed in double precision on the flattened stream. The forecast at horizon k snapshot-spacings past the newest snapshot is

$$\hat{\mathbf{F}}_{n+k} \;=\; \mathrm{Re}\big(\Phi\,(\Lambda^k\, b)\big)\,, \tag{3}$$

where Λ^k is the elementwise principal power, valid for *fractional* k. Fractional horizons matter because the buffer is spaced N sampler-steps apart while skipped steps lie between anchors: the horizon in spacing units is $k = (t - t_{\text{last}})/N$, typically a proper fraction. For the exponential class the fractional power is exact: $\Lambda^{1/N}$ *is* the one-step propagator. The fit (2) is performed once per compute step and its eigendecomposition reused across the window (Appendix A quantifies the cost). Degenerate fits (SVD/eig failure, non-finite forecast) fall back to last-value reuse; rank truncation is the noise defense, discarding directions whose singular values sit below $10^{-4}\sigma_1$. (Throughout, "DMD" abbreviates Dynamic Mode Decomposition, never Distribution Matching Distillation.)

4.3 Validity conditions: the identifiability floor and the uniform-spacing rule

These two rules ensure the estimator is only consulted where its assumptions hold, so the experiments measure the class, not estimator misuse.

Four snapshots, not three. The trajectory is real-valued, so complex poles come in conjugate pairs: one oscillatory mode $re^{\pm i\omega}$ contributes the two real functions $r^t \cos \omega t$ and $r^t \sin \omega t$, which are *two* real degrees of freedom. Rank $r = 3$ (one conjugate pair plus one real mode, the generic minimal oscillatory case) requires three snapshot *pairs*, hence **four snapshots**. With only two pairs the pencil is rank-deficient for this class and the estimated poles alias: empirically the forecast error jumps from $\sim 5 \times 10^{-9}$ (three pairs) to $\sim 2 \times 10^{-1}$ (two pairs) on clean synthetic trajectories. Below the floor we do not serve DMD at all.

Uniform-spacing tail. The propagator estimated by (2) advances exactly one snapshot-spacing per application, so the buffer must be uniformly spaced. Real schedules are not: the warm-up boundary changes the compute cadence, and adaptive deciders change it dynamically. We therefore walk back from the newest snapshot and keep the longest suffix with constant spacing, fitting only on it. If the uniform tail is shorter than the floor, we fall back to the Hermite forecast (the polynomial path is always maintained in parallel at negligible cost; it shares the finite differences the schedule already keeps). DMD thus acts only where it is valid, and warm-up is automatically covered by the published polynomial behavior.

4.4 The selection instruments: two holdout designs

H2 requires a selector; we test two designs that differ only in where the held-out target sits relative to the served horizon.

1-step holdout. At a compute step with uniform tail of length ≥ 5, hold out the newest snapshot; fit DMD on the remaining tail and back-cast the held-out snapshot at horizon 1; do the same with a degree-2 Newton-forward polynomial (the polynomial analogue at matched window); serve, for the whole upcoming skip window, whichever basis reproduced the held-out snapshot with smaller relative error. The selection is cached per compute step, so the selector costs one extra small SVD per *compute* step, amortized over all skips. The yardstick is a *forward* $(+k)$ polynomial, so the selection logic is independent of the Hermite sign convention of Sec. 7.

Horizon-matched holdout. The 1-step test is low-variance but myopic: it ranks the bases at distance one when the window will be served at distance up to $N-1$, and rankings can invert between those distances. The horizon-matched variant backcasts at the actual skip distance of the window, $h \approx (N-1)/\text{spacing}$ gaps, against the *served* damped-Hermite arm: for $h \geq 4$ the DMD fit uses the newest h snapshots and backcasts the snapshot h gaps older (a fresh fit, extrapolating backwards over the served distance); for $h < 4$ it degrades to a forward prefix backcast at distance h. The single far-out target is higher-variance than the 1-step test; Sec. 6 reports where each design wins on the controlled suite and what both do on DiT.

5 RQ1: Does the basis ranking transfer across families?

Throughout, "Hermite" is our reimplementation of HiCache's dual-scaled Hermite forecast, "DMD" is the exponential forecaster of Sec. 4.2 on the *same* schedule, and speedups are measured solo (uncontended GPU). Wall-clock at a given interval is set mainly by the skip schedule (both formulas are small next to a forward pass), so quality at matched interval is the primary comparison. Cells affected by the Hermite sign bug are explicitly labeled as-released versus corrected (Sec. 7); DMD and TaylorSeer$(+k)$ cells are sign-independent. Figure 1 previews the section's verdict.

5.1 The mechanism in isolation: H1 on its own class

Before any model enters the loop, we verify that the estimator realizes the class advantage the proposition promises; otherwise the cross-family test would confound class with implementation. Synthetic trajectories are drawn from the solution class (1) (two damped/oscillatory modes, 64 channels, 20 seeds); each method sees the same 8-anchor window and forecasts H steps past it (H is the reach of interval $H+1$). Table 1 reports mean relative ℓ_2 error; all rows postdate the sign fix.

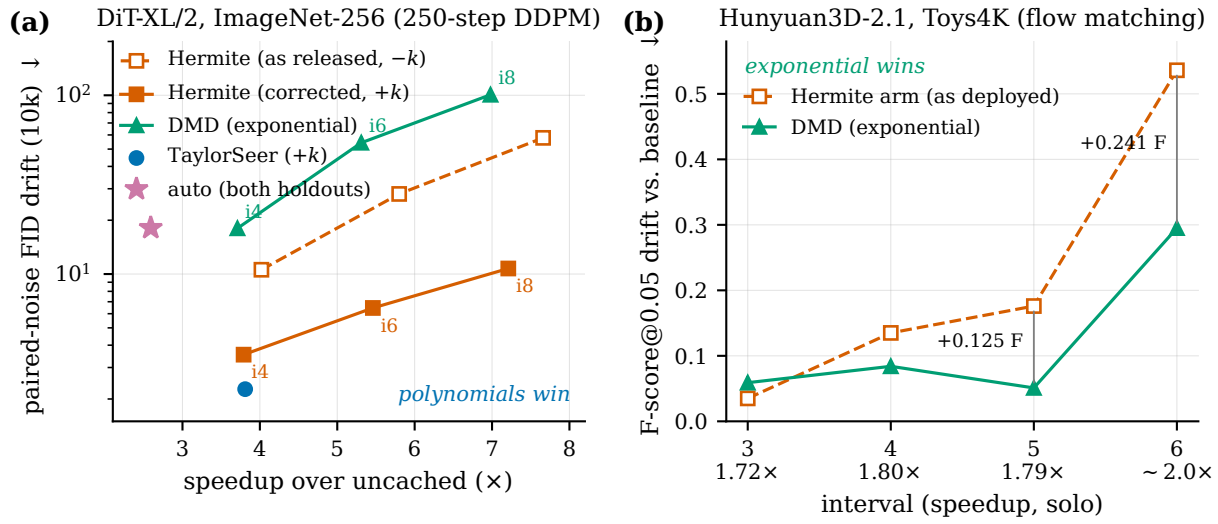

Figure 1: The domain split in one image: no single forecast basis wins. (a) DiT-XL/2 ImageNet-256, 250-step DDPM: paired-noise FID drift (10k, log scale) versus measured solo speedup, intervals 4/6/8 annotated. Every polynomial arm dominates the exponential (DMD) arm, the corrected ($+k$) Hermite dominates the as-released near-reuse variant at matched interval, and the sign-correct TaylorSeer at interval 4 is the best cell overall (drift 2.27 at 3.81×). The holdout selector (`auto`, star) serves the losing exponential arm in both modes and pays the selector overhead on top. **(b)** Hunyuan3D-2.1 on Toys4K: F-score@0.05 drift against the uncached baseline versus skip interval (solo speedup beneath each tick). The ranking inverts: the deployed Hermite arm collapses beyond interval 4 while DMD degrades gracefully, leading by +0.125/+0.241 F-score at intervals 5/6. Sources: Table 3 and Sec. 5.2.

On the class, the exponential forecast is exact and flat in H; the monomial diverges as H^{m+1}; the rational basis improves the constants but still diverges; the damped Hermite trades small-H accuracy for bounded growth. Under noise, DMD's rank truncation rejects the noise subspace and degrades to ∼0.3 where the monomial amplifies to 15. Two stress scenarios probe validity limits. Under *drifting* (non-autonomous) dynamics the exponential basis remains the most accurate (rel. error 0.50 at H=8 vs. 13.7 monomial, 2.5 rational, 3.2 Hermite). Under an abrupt *regime switch* inside the cached window, the designed failure mode of any whole-window fit, the forced exponential fit degrades (9.4 at H=8); Sec. 6 shows the holdout selector contains exactly this failure. Full six-scenario tables ship with the repository (`benchmarks/MICROBENCH_RESULTS.md`).

Instrument ablations (same suite): histories of 5–6 outperform both shorter (under-determined, floor) and longer (averaging over drifting dynamics) buffers; removing rank truncation is harmless on clean trajectories and harmful under noise, as predicted; serving DMD from 3 snapshots degrades catastrophically (aliasing, Sec. 4.3) and the Hermite fallback removes the failure entirely; replacing the fractional power Λ^k, $k = (t - t_{\text{last}})/N$, with nearest-integer powers visibly degrades mid-window steps; and $\sigma = 0.5$ is optimal for the corrected ($+k$) Hermite in all five scenarios, so the tuned default does not encode the sign bug (Sec. 7).

This establishes only that the mechanism works *on the class*. Whether a given model's feature stream is in the class at the served horizons is what the next two subsections measure, with opposite outcomes.

5.2 Flow-matching 3D generators: H1 holds

Hunyuan3D-2.1 (flow-matching DiT, flat velocities). Image-to-3D on the Toys4K set, F-score@0.05 of the accelerated mesh against the *uncached* baseline geometry after Go-ICP alignment (one rotationally degenerate object, a perfect sphere, is excluded; remaining cells reproduce to ±0.01 across independent runs). Table 2 and Fig. 1b: at interval 3 the Hermite arm is slightly ahead (0.876 vs 0.852) but collapses beyond its ceiling while DMD degrades gracefully: +0.125 at interval 5, +0.241 at interval 6. On the deployed Hunyuan3D-2-mini, DMD at interval 5 is exactly lossless (0.794 vs baseline 0.794). The Hermite arm in these tables is each fork's forecaster *as deployed*, which inherited the as-released ($-k$) convention (Sec. 7); these

Table 1: Forecast error H steps past an 8-anchor window on trajectories from the feature-ODE solution class (mean rel. ℓ_2, 20 seeds). Top: clean. Bottom: 1% Gaussian snapshot noise. All rows use the sign-correct ($+k$) implementations.

	basis	H=1	H=2	H=3	H=4	H=6	H=8
clean	TaylorSeer (monomial)	1.5e-2	8.0e-2	2.6e-1	6.2e-1	2.3e0	6.5e0
	Padé / FoCa (rational)	4.9e-2	1.1e-1	1.7e-1	2.4e-1	5.3e-1	1.2e0
	HiCache (Hermite, $+k$)	1.5e-1	3.0e-1	4.6e-1	6.3e-1	1.0e0	1.7e0
	DMD (exponential)	**4.7e-9**	**1.4e-8**	**3.0e-8**	**5.3e-8**	**1.2e-7**	**2.2e-7**
1% noise	TaylorSeer (monomial)	9.8e-2	3.7e-1	9.0e-1	1.9e0	6.0e0	1.5e1
	Padé / FoCa (rational)	7.4e-2	4.6e-1	1.2e0	1.7e0	2.5e0	3.3e0
	HiCache (Hermite, $+k$)	1.5e-1	3.0e-1	4.7e-1	6.4e-1	1.1e0	1.8e0
	DMD (exponential)	**2.3e-2**	**4.5e-2**	**7.5e-2**	**1.2e-1**	**2.1e-1**	**3.0e-1**

Table 2: Hunyuan3D-2.1, Toys4K F-score@0.05 vs. uncached baseline (0.911 = self-score); identical schedules, only the basis differs. The Hermite column is the fork's deployed (as-released) forecaster; see text. Speedup is wall-clock, solo.

interval	Hermite arm (as released)	DMD (ours)	speedup
3	**0.876**	0.852	1.72×
4	0.776	**0.827**	1.80×
5	0.735	**0.860**	1.79×
6	0.375	**0.616**	∼2.0×

cells stand as measured. The baseline-relative DMD claims (graceful degradation, exact losslessness) do not involve the Hermite arm at all, and the DiT control of Sec. 5.3 shows the as-released variant is a competitive near-reuse baseline rather than a strawman.

SAM3D (structured PyTree velocities). The forecaster applies leaf-wise to structured latents. Against the vanilla output: the Hermite arm at its lossless interval 3 gives 1.44×; DMD remains geometry-lossless (F_1=1.000, Chamfer drift 0.013) through interval 6 at **1.56**×. The extra speed comes entirely from the wider lossless interval the basis affords on this model family.

TRELLIS v1/v2 (sparse-structure stage). Swapping only the forecast basis inside an aggressively cached pipeline (carved-hybrid schedule, 2.8×): Hermite arm 0.825, DMD 0.829 F-score@0.05 (n=31; vanilla 0.839) at matched speed. The edge is small but consistent at the deployed interval, widens at higher intervals, and reproduces on the 4B-parameter v2 model (+0.03–0.04 F-score at intervals 3–4).

5.3 DiT-XL/2, ImageNet 256: H1 fails through its second scope condition

The literature-standard image benchmark inverts the 3D ranking. Protocol (`benchmarks/dit_imagenet/`): official DiT-XL/2 checkpoint, 250-step DDPM, cfg 1.5, batch 64; *paired-noise* FID at N=10k. The per-step ancestral noise is re-seeded identically per batch across cells, so FID of a cached cell against the uncached baseline measures pure cache-induced drift with no RNG floor (a lossless cache reads ≈ 0; an interval-1 control that runs the full cache machinery with zero forecast steps reads -0.00). Table 3 reports the complete ladder: every as-released cell, every corrected re-run, and the holdout A/B of Sec. 6; Fig. 1a plots it.

Three observations resolve H1 on this family. **(1) Polynomial forecasting is near-lossless on this workload.** The sign-correct TaylorSeer monomial at interval 4 drifts 2.27 FID from the paired baseline at 3.81× speedup, with absolute FID 8.95 versus the baseline's 8.89: forecasting on DiT-class denoising is, at this interval, essentially free, and the basis ladder has little room to improve it. **(2) The corrected Hermite confirms it across the interval sweep.** The sign-fixed ($+k$) Hermite drifts 3.54/6.46/10.74 at intervals 4/6/8, a 3.0–5.4× improvement over the as-released near-reuse variant (10.57/28.06/57.79): the corrected polynomial at interval 8 (7.2× speedup) drifts *less* than the as-released one at interval 4. **(3)**

Table 3: DiT-XL/2 ImageNet 256×256 (250-step DDPM, cfg 1.5, paired noise, N=10k); the complete ladder. The primary metric is the *drift* FID of each cached cell against the uncached baseline under identical per-step noise (a lossless cache reads ≈ 0; the interval-1 validity control reads −0.00). "As released" rows were measured with the sign-bugged anti-extrapolative Hermite (Sec. 7), which behaves as damped near-reuse; "corrected" rows are the sign-fixed ($+k$) re-runs. FID values are from the 10k runs; latencies marked * were re-timed post-eigencache at n=512 (Sec. A) because their FID runs predate the per-window eigendecomposition cache. The absolute column uses a 10k pytorch_fid reference and is not comparable to published ADM-protocol FIDs; it is consistent across rows.

method	int.	ms/img	speedup	FID drift ↓	FID vs. ref ↓
baseline (uncached)	—	1791	1.00×	0.00	8.89
TaylorSeer (monomial, $+k$)	4	470	3.81×	**2.27**	**8.95**
HiCache Hermite (corrected, $+k$)	4	472	3.79×	3.54	9.60
HiCache Hermite (as released, $-k$)	4	445	4.02×	10.57	15.09
DMD (exponential)	4	483*	3.71×	18.02	21.47
`auto`, 1-step holdout (as released)	4	691*	2.59×	18.08	21.54
`auto`, 1-step holdout (corrected)	4	691*	2.59×	18.11	21.57
`auto`, horizon holdout (corrected)	4	692*	2.59×	18.11	21.57
HiCache Hermite (corrected, $+k$)	6	328	5.46×	**6.46**	**11.61**
HiCache Hermite (as released, $-k$)	6	309	5.80×	28.06	31.06
DMD (exponential)	6	337*	5.31×	54.24	55.57
HiCache Hermite (corrected, $+k$)	8	248	7.21×	**10.74**	**15.10**
HiCache Hermite (as released, $-k$)	8	234	7.66×	57.79	59.73
DMD (exponential)	8	256	6.98×	100.65	100.99

The exponential basis loses at every interval. DMD drifts 18.02/54.24/100.65 at intervals 4/6/8: 1.7–1.9× worse than even the as-released near-reuse control, and 5–9× worse than the corrected Hermite. The exponential fit is not merely suboptimal here; it underperforms approximately doing nothing. This is precisely the refutation signature predicted in Sec. 3.1: the parsimonious reading is that the 250-step DDPM stream is locally so smooth that low-order polynomial truncation error is negligible at the served fractional horizons, while DMD pays geometric pole-estimation error.

Latency provenance. FID values in Table 3 are from the 10k-image runs; the `dmd` and `auto` latencies were re-timed post-optimization on the same GPU at n=512 (Appendix A), since their original 10k cells predate the per-window eigendecomposition cache. All other latencies are from the runs that produced the FID values.

5.4 Verdict on RQ1: an empirical domain split

Across Secs. 5.2 and 5.3 the evidence is symmetric and family-shaped: the exponential basis wins on every flow-matching 3D generator tested and loses on the DiT-class denoiser at every interval tested; the polynomial basis shows the mirror image (Fig. 1). H1 read as a universal claim about diffusion feature streams is **falsified**; read conditionally, its two scope conditions carry the predictive content, and the DiT failure arrives through the second (pole-estimation vs. truncation error) exactly as stated. We know of no published cross-family comparison of forecast bases, and these results indicate one is necessary before any basis claim generalizes. We deliberately do not over-theorize the mechanism: the two families differ simultaneously in sampler (flow matching vs. 250-step DDPM), step count, modality, and guidance regime, and isolating the causal axis requires interventions (e.g. step-count sweeps on one model) we have not yet run.

6 RQ2: Can training-free holdout selection recover the optimum?

6.1 Where selection works: synthetic regimes and the 3D family

On the controlled suite the 1-step holdout behaves exactly as H2 hopes. On clean, noisy, and drifting trajectories it selects the exponential basis in 120/120 windows and matches it exactly; under an abrupt

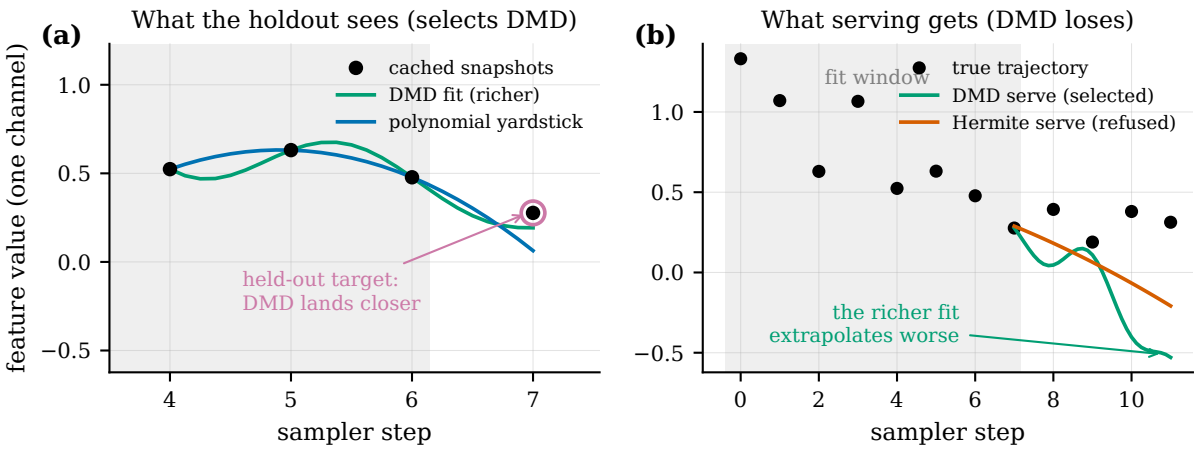

Figure 2: Why holdout selection fails on smooth real features: backcast fit does not predict forward extrapolation. One channel of an oscillatory-with-trend trajectory from the controlled suite, in the exact microbench geometry (8 cached anchors, served horizon H=4); 12 of 20 seeds show this inversion, matching the published 12/20 wrong picks. **(a)** The 1-step holdout's view: both bases are fit with the newest snapshot held out; the richer DMD fit reproduces the held-out target better, so the selector picks DMD. **(b)** The served window: the same richer fit extrapolates *forward* worse than the damped Hermite arm it displaced. The holdout rewards in-sample flexibility; forward error is governed by pole-estimation error the in-window fit cannot see.

regime switch inside the cached window it detects the misfit in 120/120 windows, serves the Hermite arm, and contains the error at 1.6 at H=8 (68× below the undamped monomial's 106, versus 9.4 for the forced exponential fit). On the flow-matching 3D generators it is harmless to helpful. Selection is demonstrably the right tool for *regime switches*; the open question was whether it also resolves the *domain split*.

6.2 The DiT failure, the pre-registered fix, and its refutation

On DiT the 1-step holdout served the exponential arm: FID drift 18.08 as released (18.11 corrected), tracking DMD's 18.02, where the corrected polynomial sits at 3.54 (Table 3). The failure initially looked like a ranking *inversion*: the basis that wins at backcast distance 1 loses at the served distance. We reconstructed that regime synthetically (oscillatory-with-trend trajectories, Fig. 2; full tables in the repository) and measured both holdout modes at the matched distance H=4: the horizon-matched holdout picks the winning arm in **20/20** windows and its error equals the oracle Hermite row exactly (0.584), where the 1-step holdout picks the losing exponential arm in 12/20 (0.877); the forced exponential fit sits at 1.09. Horizon matching is not a free win even there: the single far-out backcast is higher-variance, and on the same suite it regresses where the 1-step test is fine (under 1% snapshot noise at H=4/8: 0.238/0.522 vs 0.116/0.302; after a regime switch at H=6/8: 3.07/7.97 vs 0.92/1.56). Per a decision rule fixed before the DiT run (adopt as default only if no regression), the default stayed 1-step.

Prediction P1 (Sec. 3.2) was then put to the pre-registered DiT A/B, and **refuted**: the horizon-matched cell reads FID drift 18.11, identical to the corrected 1-step cell (18.11). On DiT's real features the horizon holdout *also* serves the DMD arm; the microbench misprediction-regime win (20/20) did not transfer.

6.3 Mechanism: backcast fit does not predict forward extrapolation

The characterization that fits all of the evidence is sharper than a distance mismatch. DMD's parameterization is strictly richer than the degree-2 polynomial yardstick at matched window (complex poles and amplitudes versus three real coefficients), so it fits, and therefore backcasts, the snapshot history better at *any* holdout distance inside the window, while extrapolating forward worse, because forward error is dominated by pole-estimation error that the in-window fit cannot see (Fig. 2). A holdout drawn from the same short history the model was fit on rewards exactly the in-sample flexibility that hurts at serve time; it is model-selection

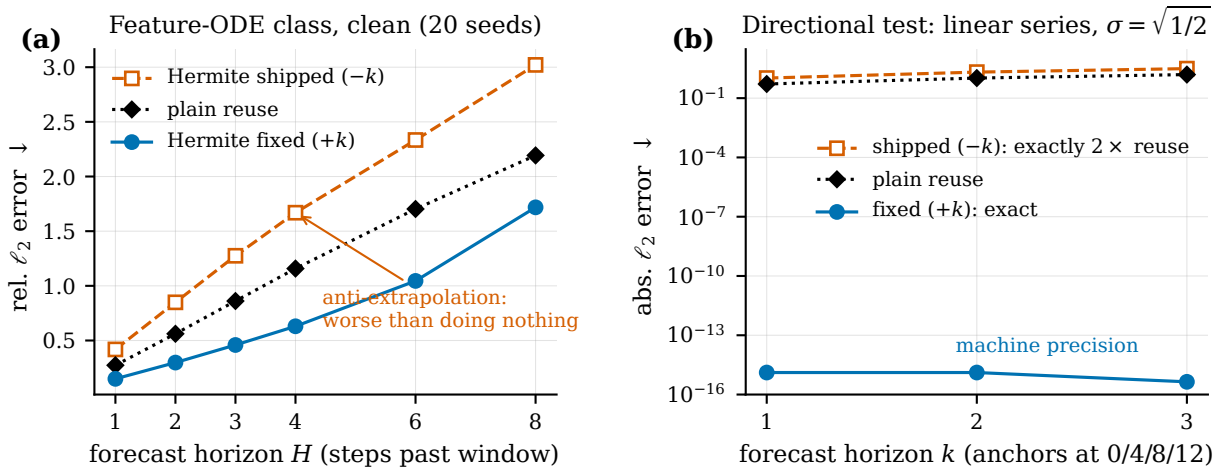

Figure 3: The anti-extrapolation defect, and the test that catches it. (a) Controlled suite (feature-ODE class, clean, 20 seeds): the shipped $(-k)$ Hermite is strictly *worse than plain reuse* at every horizon, while the one-character fix $(+k)$ beats both. Nothing about the shipped curve looks broken; it is smooth, bounded, and produces plausible quality ladders end to end. **(b)** The directional closed-form test: on a linear series with anchors at 0/4/8/12 and $\sigma = \sqrt{1/2}$, a correct first-order forecast is exact to machine precision ($\sim 10^{-15}$, note the log scale) while the sign-flipped variant errs at exactly 2× reuse. One assertion separates a forecaster from a damped cache; FID and F-score ladders cannot.

by training error, locally disguised as validation. On the synthetic suites the two bases' fit quality separates because the generating dynamics genuinely change (regime switch) or genuinely match one class, so the holdout is informative; on DiT's smooth real features both bases fit the history near-perfectly and the holdout margin carries no signal about the forward direction.

6.4 Verdict on RQ2

Holdout selection is correct on synthetic regime switches (120/120), harmless-to-helpful on the flow-matching 3D generators, and wrong on DiT in both modes. H2 is **refuted on real DiT features**, and the refutation generalizes beyond our selector: *any* selection mechanism validated only on synthetic trajectories, and any mechanism whose signal is in-window fit quality, inherits this failure mode. We therefore recommend **basis-by-model-family defaults** (polynomial for DiT-class denoising; exponential for the flow-matching 3D family), with holdout selection retained as a safety net against intra-run regime switches, and we state per-window basis selection from real feature trajectories as an open problem: a working selector needs a signal that correlates with *forward* extrapolation error, which in-window backcast fit demonstrably is not on at least one major model family.

7 RQ3: Why caching benchmarks cannot see a directionally wrong forecaster

We document this incident in full because we believe its failure mode is endemic to the feature-caching literature, not specific to our code: the evaluation methodology used by every paper in Sec. 2, end-to-end quality ladders, is structurally unable to detect it.

The instrumented defect. Our Hermite forecaster evaluated the dual-scaled Hermite basis at $x = -k$ instead of $x = +k$, where k is the number of steps *past* the newest anchor (one character, in two modules). The finite differences are forward slopes, so the forecast must evaluate forward; upstream TaylorSeer uses $+k$, but step-index conventions differ across caching codebases (some index timesteps descending, some count

"steps since refresh" with the opposite sign), and the flip happened in porting. Since $\tilde{H}_n(-x) = (-1)^n \tilde{H}_n(x)$, every odd-order term flipped sign: the shipped forecast extrapolated *backwards* by exactly the amount it should have moved forward. Anti-extrapolation, structurally similar to damped reuse.

Why every benchmark passed. Near-reuse fails safe. An anti-extrapolative forecast is wrong but bounded and smooth, so images render, meshes close, FID and F-score degrade gradually with interval, and nothing crashes. End-to-end benchmarks measure aggregate quality, not extrapolation direction, and the buggy baseline produced fully plausible quality ladders. Worse, hyperparameter tuning actively concealed the bug: for the buggy variant the optimal Hermite scale is $\sigma = 0.3$ rather than 0.5, because shrinking σ shrinks the wrong-signed odd terms toward plain reuse. The only way to tune anti-extrapolation is to mute it, so a tuning loop converges to a config that looks tuned and hides the defect. This is a general statement about the evaluation practice of the field: any of the published forecast bases could ship anti-extrapolative in some integration today, and the standard evidence would not show it.

What it inverted. On the controlled suite the shipped Hermite loses to plain last-value reuse in *every cell of every scenario* (Fig. 3a): the published strongest polynomial basis, as shipped by us, was strictly worse than doing nothing on the trajectory class. The corrected ($+k$) Hermite beats reuse in every cell of the published per-horizon suite except one (drifting dynamics at H=8, where any correct extrapolation overshoots) and is 1.45–2.47× more accurate than shipped across scenarios. The benchmark conclusion "Hermite forecasting does not beat reuse here" was therefore an artifact of one character. Noise did not vindicate the bug: under 1% snapshot noise the fixed variant still beats shipped 2.16× and reuse 1.51×; the backwards term was not useful damping.

Blast radius and re-run policy. The DMD path (forward eigenvalue powers) and the holdout selection logic (forward polynomial yardstick) are sign-independent; the selector's *output* changes only where it serves the Hermite arm. The controlled tables (Table 1 and the repository file) were regenerated after the fix. The DiT Hermite/holdout FID cells were measured as-released, re-run under the corrected code, and reported in labeled columns side by side (Table 3); the analysis of Sec. 5.3 was constructed before the re-runs landed, so that no conclusion depended on them, and the corrected cells confirmed it (the quantitative ladder improved 3.0–5.4×; every qualitative conclusion held). The 3D-generator forks deployed the as-released forecaster; those results stand as measured, with the Hermite arm labeled accordingly (Sec. 5.2).

The generalizable remedy: directional closed-form tests. The fix is not better end-to-end benchmarks but a different *kind* of test, one with a closed-form answer that is sensitive to direction. On a linear series with anchors at 0/4/8/12, a correct first-order forecast is exact at $\sigma = \sqrt{1/2}$ (error $\sim 10^{-15}$), while the sign-flipped variant has error exactly 2× reuse (Fig. 3b); this is now an assertion in our test suite, and the analogous one-line check would have caught the bug in any of the codebases we ported from or to. We propose that any paper introducing a forecast basis report such a directional test alongside its quality ladder: a basis that fails it can still produce a publishable-looking quality ladder, which is precisely the problem.

8 Threats to validity

Construct validity. (i) The DiT ladder is FID-10k under the paired-noise protocol; the FID-50k headline trio has not been run. The 10k estimator bias is identical across cells and the primary metric is differential, so the ladder ordering stands, but absolute numbers are not ADM-protocol comparable. (ii) The 3D evaluation protocols (F-score@0.05 after Go-ICP alignment against the uncached baseline; paired-noise FID drift) are self-designed; we report them in full and release the harnesses, but no external group has yet replicated them. **Internal validity.** (iii) The 3D Hermite arm is the as-released variant as deployed in the forks; corrected re-runs there have not been performed. Baseline-relative 3D claims are unaffected, and the corrected-vs-as-released DiT contrast (Table 3) bounds the likely effect. (iv) All baseline arms are our reimplementations on a shared schedule rather than the original authors' code; the sign-bug episode (Sec. 7) is direct evidence of the risk this carries, which is why every arm now passes the directional test and the as-released numbers remain in the tables. **External validity.** (v) The DiT side of the split rests on a single model, dataset,

sampler, and guidance configuration (DiT-XL/2, ImageNet-256, 250-step DDPM, cfg 1.5); the flow-matching side rests on four generators from one modality family. The domain split is established empirically at two families; its mechanism is hypothesized, not isolated, and the families differ along several confounded axes. (vi) Like all forecast caches, quality claims are schedule-relative: a different decision layer (TeaCache-style) changes anchor placement and could shift every crossover; the Spectrum head-to-head specified in Sec. 2 remains future work. **Model validity.** (vii) The frozen-propagator model is local: when dynamics shift inside a window the exponential fit can misfit; holdout selection solves exactly that synthetic failure but, per Sec. 6, fails on DiT's real features in both modes, and we know of no in-window signal that predicts forward extrapolation quality there. (viii) The double-precision SVD adds memory traffic proportional to $d\,n$; we have not profiled a fused low-precision variant.

9 Honest novelty statement and conclusion

Nothing in our estimator is new mathematics: Prony's method dates to 1795, the Matrix-Pencil to 1990, exact DMD to 2010, and holdout validation is elementary statistics. What we believe is new is the knowledge: (a) the first cross-family evidence that *no single forecast basis wins*, so basis rankings published on one family should not be presumed to transfer; (b) a conditional resolution of the feature-ODE hypothesis, whose scope conditions, not its class argument, predict where the exponential basis wins and where it loses; (c) a training-free per-window selection mechanism evaluated to its honest conclusion: exact on synthetic regime switches, harmless-to-helpful on the 3D family, and wrong on DiT in both holdout modes, with the characterization that in-window backcast fit does not predict forward extrapolation quality on real features, a negative result we believe the caching literature needs before more selection mechanisms are proposed on synthetic evidence; and (d) a documented demonstration that end-to-end caching benchmarks cannot detect an anti-extrapolative baseline, with the directional closed-form test that prevents it. The supporting artifact, a small dependency-free library containing the forecaster, the selector, both holdout modes, all benchmarks, and the as-released/corrected ledgers, drop-in compatible with the TaylorSeer/HiCache ecosystem, exists so that every number in this paper can be regenerated (Appendix A).

A Artifacts and reproducibility

Everything in this appendix supports reproducibility; none of it is claimed as a research contribution.

Library. The forecaster, selector, and schedule logic ship as a small dependency-free Python library (`hicache_pp`) with prepared integrations for the existing caching frameworks and model forks used in Sec. 5.2. The holdout variants of Sec. 4.4 are exposed as `backend="auto"` with `holdout="1step"` (default) or `"horizon"`; per Sec. 6 the defaults follow the basis-by-family recommendation and `auto` is positioned as a regime-switch safety net only.

Per-window eigendecomposition cache and forecast cost. Between snapshots the DMD fit inputs cannot change, so the eigendecomposition (Φ, Λ, b) of Eq. (2) is computed once per compute window and reused by every skipped step until the next anchor arrives, making the per-skip cost one $\Phi(\Lambda^k b)$ evaluation. On CPU at interval 4 (one fit amortized over three forecasts) the cached path is 2.5–3.1× cheaper per forecast across feature dimensions 8k–524k, at the $\sim (N-1)\times$ amortization ceiling; the cached and uncached paths agree to $< 10^{-12}$. On the DiT GPU the cache cuts `dmd` from 566 to 483 ms/img at interval 4 (3.17× → 3.71×) and from 434 to 337 ms/img at interval 6 (4.13× → 5.31×); these post-cache re-timings (latency only, n=512, same GPU and protocol) are the figures Table 3 carries for the affected rows, as marked. `auto` pays one extra small backcast fit per compute step on top of the served forecast (691 ms/img at interval 4), a price not worth paying on either family given Sec. 6.

Benchmark harnesses. The controlled microbenchmark (`benchmarks/forecast_microbench.py`, six scenarios, full tables in `benchmarks/MICROBENCH_RESULTS.md`), the DiT harness with the paired-noise protocol

and a resumable run queue (`benchmarks/dit_imagenet/`, results ledger `RESULTS_DIT.md` with every as-released and corrected cell), and the per-model 3D integration patches are all included. The directional closed-form regression test of Sec. 7 is an assertion in the test suite. Every figure is regenerated from the banked ledgers by `paper/figures/make_figures.py`; Fig. 2 is computed live from the microbench generators.

Availability. Code, benchmarks, ledgers, and all tables:
https://github.com/Archerkattri/hicache-plus-plus.

References

[1] J. Liu, C. Zou, Y. Lyu, J. Chen, and L. Zhang. From reusing to forecasting: Accelerating diffusion models with TaylorSeers. In *ICCV*, 2025. arXiv:2503.06923.

[2] L. Feng, S. Zheng, J. Liu, Y. Lin, Q. Zhou, P. Cai, X. Wang, J. Chen, C. Zou, Y. Ma, and L. Zhang. HiCache: A plug-in scaled-Hermite upgrade for Taylor-style cache-then-forecast diffusion acceleration. arXiv:2508.16984, 2025.

[3] S. Zheng, L. Feng, X. Wang, Q. Zhou, P. Cai, C. Zou, J. Liu, Y. Lin, J. Chen, Y. Ma, and L. Zhang. Forecast then calibrate: Feature caching as ODE for efficient diffusion transformers. In *AAAI*, 2026. arXiv:2508.16211.

[4] S. Zheng, G. Chen, Q. Zhou, Y. Lin, L. He, C. Zou, P. Cai, J. Liu, and L. Zhang. Let features decide their own solvers: Hybrid feature caching for diffusion transformers (HyCa). arXiv:2510.04188, 2025.

[5] J. Liu, C. Zou, Y. Lyu, F. Ren, S. Wang, K. Li, and L. Zhang. SpeCa: Accelerating diffusion transformers with speculative feature caching. In *ACM Multimedia*, 2025. arXiv:2509.11628.

[6] J. Han, J. Shi, P. Li, H. Ye, Q. Guo, and S. Ermon. Adaptive spectral feature forecasting for diffusion sampling acceleration (Spectrum). In *CVPR*, 2026. arXiv:2603.01623.

[7] F. Liu, S. Zhang, X. Wang, Y. Wei, H. Qiu, Y. Zhao, Y. Zhang, Q. Ye, and F. Wan. Timestep embedding tells: It's time to cache for video diffusion model. In *CVPR*, 2025. arXiv:2411.19108.

[8] Z. Ma, L. Wei, F. Wang, S. Zhang, and Q. Tian. MagCache: Fast video generation with magnitude-aware cache. In *NeurIPS*, 2025. arXiv:2506.09045.

[9] X. Zhou, D. Liang, K. Chen, T. Feng, X. Chen, H. Lin, Y. Ding, F. Tan, H. Zhao, and X. Bai. Less is enough: Training-free video diffusion acceleration via runtime-adaptive caching (EasyCache). arXiv:2507.02860, 2025.

[10] G. R. de Prony. Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques. *Journal de l'École Polytechnique*, 1(22):24–76, 1795.

[11] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. ASSP*, 38(5):814–824, 1990.

[12] P. J. Schmid. Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid Mechanics*, 656:5–28, 2010.

[13] J. H. Tu, C. W. Rowley, D. M. Luchtenburg, S. L. Brunton, and J. N. Kutz. On dynamic mode decomposition: Theory and applications. *Journal of Computational Dynamics*, 1(2):391–421, 2014.