

BRACE: Taming Sharp Irregularities via Barycentric Rational Forecasting for Fast Diffusion Transformers Inference

Jinlong Yang*

2023141461083@stu.scu.edu.cn
Sichuan University
Chengdu, China

Lizilin

2023141461067@stu.scu.edu.cn
Sichuan University
Chengdu, China

Jinke Wu*

w_rmsl@stu.scu.edu.cn
Sichuan University
Chengdu, China

Yao Zhou✉

yaozhou@scu.edu.cn
School of Artificial Intelligence, Sichuan University
Chengdu, China

Abstract

Diffusion Transformers (DiTs) have demonstrated exceptional performance in high-fidelity image and video generation. To alleviate their massive computational overhead, temporal feature caching has been proposed to bypass redundant computations. However, existing cache-then-forecast methods driven by derivative-based polynomials often cause severe quality degradation under high acceleration due to unstable long-step predictions. To address this bottleneck, we propose Barycentric Rational Forecasting with Chebyshev Enhancement (BRACE). Motivated by the observation that DiT feature trajectories are globally smooth yet frequently exhibit sharp irregularities and local non-smoothness, BRACE shifts the paradigm from derivative-driven polynomial extrapolation to feature-driven rational forecasting. Specifically, it maintains a local sliding window to cache sparse historical features and leverages adapted Chebyshev weights to formulate a barycentric rational function, directly aggregating these raw features to ensure numerical stability. Extensive experiments demonstrate that BRACE achieves state-of-the-art quality-efficiency trade-offs across various DiT architectures with negligible computational overhead. Further details are available on our project page.

CCS Concepts

• Computing methodologies → Computer vision.

Keywords

Diffusion Transformer, Acceleration, Feature Cache, Barycentric interpolation

ACM Reference Format:

Jinlong Yang, Jinke Wu, Lizilin, and Yao Zhou. 2026. BRACE: Taming Sharp Irregularities via Barycentric Rational Forecasting for Fast Diffusion Transformers Inference. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 10 pages. <https://doi.org/10.1145/3767308.3835649>

*Jinlong Yang and Jinke Wu contributed equally to this research.



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3835649>

1 Introduction

Diffusion Models (DMs) [12] have become the dominant generative modeling paradigm across diverse modalities including images [4, 5, 17], videos [15, 41], and audio [16, 20], driven by their exceptional high-fidelity content synthesis capability. The recent architectural shift to Diffusion Transformers (DiTs) [31] has further amplified this synthesis performance, coming at the cost of prohibitive computational overhead. Specifically, the iterative sampling process requires sequential forward passes through large-scale models, resulting in high latency that severely limits real-time interactive deployment and high-throughput generation.

To alleviate this bottleneck, feature caching has emerged as a prominent training-free acceleration strategy. Built on the strong temporal consistency of intermediate hidden representations, traditional caching approaches directly reuse features from preceding timesteps, but this 'cache-and-reuse' paradigm inherently degrades generation quality under long skip intervals as feature similarity diminishes with extended temporal distances [30, 35]. Recently, this research direction has evolved from static feature reuse to proactive forecasting: representative frameworks such as TaylorSeer [21] and HiCache [8] construct derivative-driven polynomial paradigms via Taylor expansions and dual-scale Hermite-inspired extrapolation, respectively. These methods achieve promising acceleration results by approximating derivatives with finite differences of sparse historical features for future state prediction.

While innovative, this derivative-driven polynomial paradigm fundamentally relies on the accuracy of instantaneous derivatives estimated from discrete historical data, and suffers from the inherent instability of polynomial extrapolation over large skip intervals. To clarify the core challenges of DiT acceleration, we empirically analyze DiT feature evolution trajectories, and reveal an inherent geometric property of DiTs: their feature trajectories, while globally smooth, frequently exhibit sharp irregularities and local non-smoothness. This dual geometric nature is visualized in Figure 1(a), and aligns with phase transition phenomena in diffusion generative processes [24], where the latent manifold undergoes rapid structural shifts across specific denoising intervals.

Numerically, these rapid transitions mimic stiff system behavior, with feature states changing drastically within extremely short intervals. This localized high curvature exposes two fundamental limitations of existing paradigms: first, finite-difference gradient estimates become highly unreliable at sharp inflection points; second,

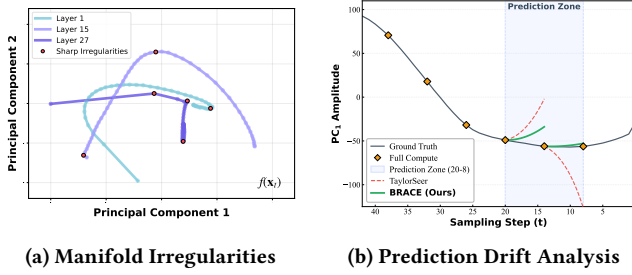


Figure 1: Feature geometric complexity. (a) Trajectories exhibit sharp irregularities across all layers. (b) At Layer 15, TaylorSeer diverges at inflection points, while BRACE robustly tracks the actual manifold.

rigid polynomials lack the structural adaptability to model abrupt shifts, leading to severe extrapolation divergence outside the local window [9]. When extrapolating across these high-frequency transition regions, derivative-centric polynomials inevitably overshoot or oscillate, accumulating severe truncation errors. As verified in Figure 1(b), derivative-driven polynomial methods like TaylorSeer [21] fail to capture these abrupt transitions due to the above deficiencies, resulting in severe prediction drift and loss of structural integrity.

In classical numerical analysis, it is well-established that rational extrapolation consistently outperforms its polynomial counterparts in modeling rapid transitions [39], and barycentric rational interpolation uniquely enables explicit detection and elimination of unattainable points and spurious interior poles [2]. This stabilizing property in interpolation heuristically inspires us to extend this approach to causal prediction, enabling the accurate capture of complex DiT feature dynamics and the effective taming of trajectory irregularities. Driven by this motivation, BRACE (Barycentric Rational Forecasting with Chebyshev Enhancement) is introduced as a novel forecasting paradigm for structurally robust diffusion acceleration and superior generation quality.

BRACE shifts the forecasting paradigm from derivative-driven polynomial extrapolation to feature-driven rational forecasting. Specifically, rather than estimating unstable local derivatives, it directly aggregates raw historical features cached in a sparse local sliding window. To proactively suppress long-step oscillations, adapted Chebyshev weights are introduced and seamlessly unified within the barycentric rational framework to enable stable causal extrapolation. This highly expressive rational predictor replaces rigid polynomials to smoothly accommodate high-curvature dynamics, bridging the gap between the global smoothness and local geometric complexity of DiT feature trajectories. This structural robustness is critical for long-step acceleration: as step size increases, derivative-driven polynomials oscillate or diverge, while BRACE maintains numerical stability and high fidelity even under aggressive skip intervals, successfully capturing localized sharp transitions without overshooting, as shown in Figure 1(b).

In summary, our core contributions are as follows:

- **The BRACE Framework:** We propose BRACE, a training-free rational forecasting framework for efficient DiT inference. By synergizing a sparse historical cache with adapted

Chebyshev weights within a barycentric rational framework, BRACE ensures superior numerical stability and tames abrupt transitions where traditional derivative-driven methods fail.

- **Geometric Analysis of DiT Dynamics:** We reveal the dual geometric nature of DiT evolution: globally smooth trajectories frequently punctuated by sharp irregularities, and empirically validate that rational bases inherently outperform polynomials in fitting DiT feature trajectories, providing structural motivation for our approach.
- **State-of-the-Art Acceleration:** Extensive evaluations across diverse architectures and modalities—including DiT-XL/2, FLUX.1-dev, and HunyuanVideo—demonstrate that BRACE consistently outperforms existing baselines, maintaining exceptional generation fidelity even under aggressive skip intervals.

2 Related Work

2.1 Computation Reduction Strategies

Previous efforts to accelerate diffusion models primarily focus on either reducing the number of sampling steps or compressing the denoising network itself. Step-reduction methods accelerate the generative process through efficient ODE solvers like DDIM [37] and DPM-Solver [26, 27, 47], knowledge distillation [28, 33, 34], or few-step paradigms such as Rectified Flow [23] and Consistency Models [38]. However, these approaches often degrade generation quality or require expensive retraining [29]. Alternatively, network compression techniques like pruning [7, 44], quantization [19, 36], and token merging [3, 45] reduce FLOPs per step but entail a strict efficiency–expressivity trade-off, risking generation fidelity under aggressive compression. Consequently, developing a plug-and-play acceleration paradigm that dynamically exploits redundancy without compromising the original synthesis quality remains a critical open challenge across diverse model architectures.

2.2 Feature Caching

As a prominent training-free alternative, feature caching skips redundant intermediate computations. Existing methods generally fall into two paradigms: cache-then-reuse and cache-then-forecast. Reuse-based approaches typically choose to directly copy cached representations, spanning various spatial granularities from global spatial-temporal feature reuse (e.g., DeepCache [30], FORA [35], PAB [48]) to fine-grained token-level caching (e.g., ToCa [50], TokenCache [25], DaTo [45]). However, regardless of the spatial granularity, the temporal representation mismatch inevitably grows with larger skip intervals, fundamentally degrading generation quality.

Forecasting-based approaches, pioneered by TaylorSeer [21], actively predict feature evolution to bridge this temporal gap. Recent advancements seek to refine this paradigm: HiCache [8] improves the standard Taylor expansion by introducing scaled Hermite coefficients, while FOCA [49] and SpeCA [22] integrate verify-and-calibrate mechanisms into the forecasting framework to mitigate extrapolation drift. Nevertheless, because their underlying prediction modules inherently rely on volatile derivative estimates and rigid polynomial bases, they remain brittle near sharp transitions.

Our method instead performs highly expressive barycentric rational extrapolation directly from raw historical features, ensuring robust trajectory alignment even under aggressive skip intervals.

3 Method

3.1 Preliminaries

Diffusion models generate data by iteratively reversing a predefined forward noising process via a neural network $\epsilon_\theta(\mathbf{x}_t, t)$. Starting from Gaussian noise, the latent variable $\mathbf{x}_t$ is progressively refined over T timesteps:

$$\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \alpha_t}} \epsilon_\theta(\mathbf{x}_t, t) \right) + \sigma_t \epsilon. \quad (1)$$

Requiring T sequential model evaluations, the sampling process incurs massive computational overhead.

To mitigate this, *forecast-based caching* approximates future intermediate feature maps $\mathbf{F}^l$ at layer l . Specifically, using a recursive discrete difference operator Δ , the m -th order Taylor prediction for a skip interval s is expressed as:

$$\mathbf{F}_{\text{pred}}(\mathbf{x}_{t-s}^l) = \mathbf{F}(\mathbf{x}_t^l) + \sum_{i=1}^m \frac{\Delta^i \mathbf{F}(\mathbf{x}_t^l)}{i!} (-s)^i. \quad (2)$$

According to Taylor’s theorem [40], the accuracy of this polynomial extrapolation is governed by the Lagrange remainder:

$$\mathcal{R}_m = \frac{\mathbf{F}^{(m+1)}(\xi)}{(m+1)!} (-s)^{m+1}, \quad \xi \in [t-s, t]. \quad (3)$$

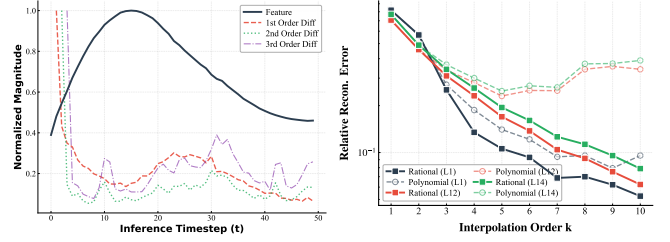
As shown in Eq. 3, the Taylor forecasting error $\mathcal{R}_m$ is theoretically bounded by the high-order derivative $\|\mathbf{F}^{(m+1)}(\xi)\|$ and the skip interval s^{m+1} . This formulation exposes a critical vulnerability: if intermediate features exhibit non-smooth dynamics, the resulting spikes in true derivatives will be severely amplified by large skip steps, risking catastrophic extrapolation divergence.

3.2 Feature Trajectory Insight

Coupling this theoretical vulnerability with the actual dual nature of feature trajectories—globally smooth yet punctuated by sharp transitions—it becomes imperative to identify an optimal forecasting tool. To determine the most robust mathematical framework for such dynamics, we conduct two parallel empirical analyses focusing on the numerical stability of finite differences and intrinsic basis expressiveness, as depicted in Fig. 2.

As shown in Fig. 2(a), although raw feature trajectories appear globally smooth, their high-order finite differences produce erratic spikes at localized transitions. This empirically confirms the vulnerability identified in Eq. (3): localized non-smoothness triggers an explosion in $\|\mathbf{F}^{(m+1)}\|$. Consequently, traditional derivative-driven extrapolation relying on these unstable estimates suffers from severe error amplification over long steps.

Beyond numerical stability, rational bases theoretically surpass polynomials in complex function approximation [39]. Fig. 2(b) empirically validates their inherent optimality for feature dynamics: as the interpolation order increases, the polynomial baseline reveals a fundamental geometric mismatch, severely drifting at sharp trajectory turns instead of fitting them.



(a) Numerical Instability of Dif- (b) Rational vs. Polynomial Ex-
ferences pressiveness

Figure 2: DiT feature dynamics and basis expressiveness. (a) Magnitude of finite differences. Localized non-smoothness leads to an explosion in $\|\mathbf{F}^{(m+1)}\|$ as the order increases. (b) Interpolation error comparison. Unlike drifting polynomials, rational bases accommodate local singularities under identical historical constraints.

These empirical insights thus motivate a paradigm shift from derivative-driven polynomial estimation to feature-driven rational extrapolation. Rather than relying on high-order derivatives, our approach forecasts future states by directly aggregating stable historical features. Crucially, by replacing rigid polynomials with rational bases, the self-normalizing denominator inherently “absorbs” localized non-smoothness, thereby effectively preventing the error explosion characteristic of polynomial drift. Ultimately, this combination of derivative-free aggregation and rational structural robustness establishes a fundamentally more stable foundation for long-step DiT acceleration, ensuring high generative fidelity even across sharp feature transitions.

3.3 The BRACE Framework

Driven by the numerical insights from our empirical analysis, Barycentric Rational Forecasting with Chebyshev Enhancement (BRACE) is proposed, a training-free framework where the forecasting paradigm is shifted from derivative-driven polynomials to feature-driven rational aggregation. As illustrated in Figure 3, the inference process alternates at a fixed interval k between full computation and rapid forecasting, a procedure formally detailed in Algorithm 1.

During full computation steps, BRACE couples exact intermediate features $\mathbf{F}_\tau$ and their timesteps into state tuples $\mathbf{S} \triangleq (\mathbf{F}_\tau, \tau)$, pushing them into a FIFO Local Sliding Window. This cache $\mathcal{H}$ preserves the historical context on the feature manifold.

Conversely, when a skip interval is triggered, BRACE activates its rational prediction mechanism. Specifically, a Domain Mapping operation first normalizes the cached timestamps into a canonical interval to ensure numerical stability. Subsequently, these mapped states are then synthesized into the predicted feature $\mathbf{F}_t$ via the Barycentric Rational Extrapolation module, which leverages Adapted Chebyshev Weights to robustly accommodate the potential non-smoothness identified in our earlier analysis.

3.3.1 Local Sliding Window (FIFO Queue). Serving as the foundation for state caching, BRACE employs a fixed-capacity FIFO queue to manage the local historical context. Formally, let $\mathbf{S}_j \triangleq (\mathbf{F}_{\tau_j}, \tau_j)$ denote a state tuple. At any given inference step, the active sliding

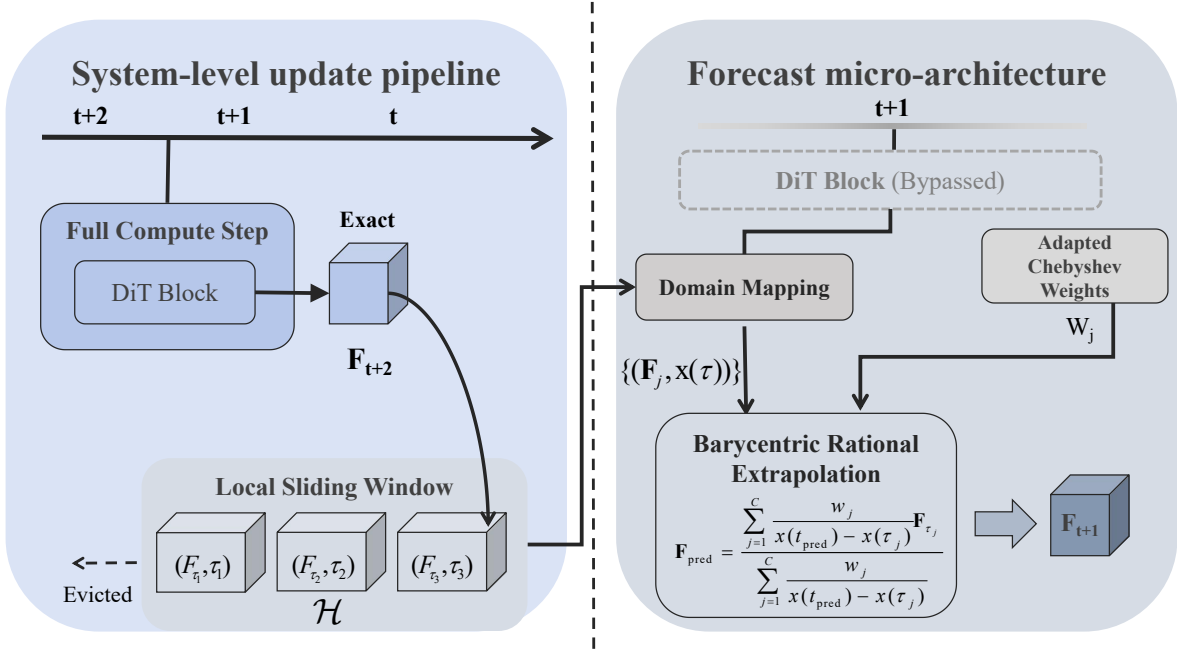


Figure 3: Overview of the BRACE framework. Left: The system-level pipeline manages a FIFO Sliding Window to maintain a sparse historical cache $\mathcal{H}$ of exact DiT features. Right: The forecast micro-architecture bypasses neural network computation by synthesizing future features through Domain Mapping and Barycentric Rational Extrapolation, guided by Adapted Chebyshev Weights with negligible overhead while maintaining high fidelity.

Algorithm 1 BRACE Sequential Inference (per layer l)

Require: Total timesteps T , capacity C , interval k
Internal State: Local Sliding Window $\mathcal{H} \leftarrow \emptyset$

```

1: for  $t = T$  down to 1 do  $\triangleright$  Discrete reverse diffusion trajectory
2:   if  $t \pmod k == 0$  then  $\triangleright$  Full Compute Step
3:      $F_t^l \leftarrow \text{ForwardBlock}_l(\text{input})$ 
4:      $\mathcal{H} \leftarrow \text{Update}(\mathcal{H}, (F_t^l, t), C)$   $\triangleright$  FIFO strategy
5:      $\tilde{F}_t^l \leftarrow F_t^l$ 
6:   else  $\triangleright$  Forecast micro-architecture
7:      $x(t) \leftarrow \text{DomainMapping}(t, \mathcal{H})$   $\triangleright$  Yields  $|x(t)| > 1$ 
8:      $F_{\text{pred}}^l \leftarrow \text{BarycentricRationalExtrap}(\mathcal{H}, x(t), W_j)$ 
9:      $\tilde{F}_t^l \leftarrow F_{\text{pred}}^l$   $\triangleright$  Bypass DiT Block
10:  end if
11: end for

```

window of capacity C is defined as an ordered sequence:

$$\mathcal{H} = (S_1, S_2, \dots, S_C). \quad (4)$$

Upon a full computation step that yields a new state S_{new} , the FIFO mechanism updates the cache via a shift-and-append operation:

$$\mathcal{H} \leftarrow (S_2, S_3, \dots, S_C, S_{\text{new}}), \quad (5)$$

where the oldest state S_1 is evicted.

This finite-capacity design strategically confines the forecasting horizon to a sparse local window, effectively decoupling temporally decayed features while ensuring the trajectory satisfies local smoothness. As a result, by maintaining a strictly local cache, BRACE structurally bypasses complex global nonlinearities that could otherwise destabilize the rational predictor.

3.3.2 Canonical Domain Mapping. To strictly align the diffusion timesteps with the theoretical foundations of the adapted weights, a Canonical Domain Mapping is prepended to the extrapolation process on the active cache $\mathcal{H}$. Specifically, it transforms each raw state tuple $S_j = (F_{\tau_j}, \tau_j)$ into a canonical state tuple $(F_{\tau_j}, x(\tau_j))$ via an affine mapping. Mathematically, this transformation is essential because the numerical stability of the rational formulation is optimally preserved within the standard interval $[-1, 1]$. To project the raw timesteps into this standard domain, the mapping boundaries are determined dynamically by the active cache:

$$x(\tau) = 2 \cdot \frac{\tau - \tau_{\min}}{\tau_{\max} - \tau_{\min}} - 1, \quad (6)$$

where $x(\tau_j) \in [-1, 1]$ for all cached states. By binding features directly to their canonical coordinates, this mapping ensures scale-invariance across different diffusion schedulers, while providing well-defined theoretical error bounds and preventing numerical instabilities during finite-precision inference.

3.3.3 Barycentric Rational Extrapolation. Following the domain mapping stage, the mapped cache $\mathcal{H} = \{(F_{\tau_j}, x(\tau_j))\}_{j=1}^C$ is aggregated to synthesize the future feature F_{pred} at the target timestep

t_{pred} . Specifically, we employ the second barycentric form. By assigning fixed weights independent of node intervals, it is fundamentally transformed into a stable rational function.

$$\mathbf{F}_{\text{pred}} = \frac{\sum_{j=1}^C \frac{w_j}{x(t_{\text{pred}}) - x(\tau_j)} \mathbf{F}_{\tau_j}}{\sum_{j=1}^C \frac{w_j}{x(t_{\text{pred}}) - x(\tau_j)}}, \quad (7)$$

where w_j denotes the adapted Chebyshev weights configured for numerical robustness and stable long-step extrapolation.

Leveraging this rational structure, the prediction is decomposed into a feature aggregation numerator and a self-normalizing denominator. This formulation provides two critical advantages: it inherently bounds predicted features to prevent polynomial divergence via self-normalization, and intrinsically absorbs localized non-smoothness by aggregating historical states rather than relying on unstable derivatives. Consequently, this design secures robust, high-fidelity long-step extrapolation.

3.3.4 Adapted Chebyshev Weights. The stability and approximation accuracy of Eq. (7) are determined by the rational weights w_j . In this context, it is well known that classical approximation theory shows that Chebyshev–Gauss–Lobatto nodes offer superior numerical stability and possess well-defined analytical weights [2]. Inspired by the numerical stability of classical Chebyshev weights, we formulate an adapted weighting scheme tailored to the nonlinear manifolds of DiT features:

$$w_j = (-1)^{j+1} \cdot \delta_j, \quad \delta_j = \begin{cases} \frac{1}{2} & j = 1, \\ \gamma & j = C, \\ 1 & \text{otherwise.} \end{cases} \quad (8)$$

This formulation incorporates two critical structural designs. First, the alternating sign $(-1)^{j+1}$ serves as an inherent stabilizer to prevent numerical collapse in the extrapolation regime. Second, a boundary sensitivity coefficient γ replaces the standard Lobatto terminal weight to flexibly accommodate varying model dynamics.

While applying these weights typically requires non-uniform Chebyshev sampling ($x_k = -\cos(\frac{k-1}{C-1}\pi)$), BRACE intentionally restricts the cache to an ultra-sparse capacity ($C \leq 3$) to prevent long-range feature corruption. Crucially, in this low-order regime, these optimal nodes are mathematically identical to equidistant points. This equivalence enables theoretically optimal sampling via a simple fixed interval, thereby eliminating complex dynamic adjustments and circumventing the numerical instability that traditionally plagues barycentric extrapolation [42].

3.4 Error Propagation and Stability Analysis

To rigorously establish the numerical stability of BRACE under large skip intervals, we analyze its pointwise extrapolation error $\mathbf{E}(x) = \mathbf{F}(x) - \mathbf{F}_{\text{extrap}}(x)$. Assume that the feature trajectory $\mathbf{F}$ is Lipschitz continuous on the inference domain, such that $\|\mathbf{F}(x) - \mathbf{F}(x_i)\| \leq L|x - x_i|$, where L is the Lipschitz constant (equivalent to the supremum of the first-order derivative $\|\mathbf{F}'\|_\infty$). Substituting this Lipschitz condition into the barycentric error identity yields a robust upper bound on the error norm.

$$\|\mathbf{E}(x)\| = \left\| \frac{\sum_{i=1}^C \frac{w_i}{x - x_i} (\mathbf{F}(x) - \mathbf{F}(x_i))}{\sum_{j=1}^C \frac{w_j}{x - x_j}} \right\| \leq \frac{\sum_{i=1}^C |w_i| \cdot L}{|D(x)|} \quad (9)$$

where $D(x) = \sum_{j=1}^C \frac{w_j}{x - x_j}$ denotes the barycentric denominator.

For extrapolation at target $x = x_C + s$ ($s > 0$), BRACE’s uniformly spaced cached nodes yield $x_j = x_C - (C - j)k$ with a constant interval k , which inherently bounds the extrapolation range to $s \leq k$. Substituting these reformulates $D(x)$ concerning s .

$$D(s) = \sum_{i=1}^C \frac{w_i}{s + (C - i)k} = \frac{1}{k} \sum_{i=1}^C \frac{w_i}{\sigma + C - i} \quad (10)$$

where $\sigma = s/k \in (0, 1]$ represents the relative skip ratio. Crucially, the summation term $\sum_{i=1}^C w_i / (\sigma + C - i)$ is strictly bounded below by a non-vanishing structural constant $\lambda > 0$ for all $\sigma \in (0, 1]$. A rigorous formal proof of this property is provided in the Supplemental Material. Consequently, by incorporating the factor $1/k$ from the definition of $D(s)$, it holds that $|D(s)| \geq \lambda/k$, thereby directly establishing the localized error bound:

$$\|\mathbf{E}(x_C + s)\| \leq \frac{\sum_{i=1}^C |w_i| \cdot L}{\lambda} \cdot k \quad (11)$$

This suggests that the extrapolation error in BRACE is theoretically governed by the interval k and the first-order derivative (Lipschitz constant L). This advantage is highly pronounced compared to conventional m -th order Taylor extrapolation (Eq. 2): for a prediction offset s , Taylor methods incur a truncation error of $\mathcal{O}(s^{m+1} \|\mathbf{F}^{(m+1)}\|_\infty)$, leading to potential polynomial divergence and unbounded amplification of high-order derivatives during sharp transitions. By bypassing these numerical hazards and theoretically bounding the error via k and L , BRACE ensures predictable stability even under aggressive skipping.

4 Experiments

4.1 Experimental Setup

BRACE is evaluated across three representative generative tasks: (i) class-conditional image generation on ImageNet-1K [6] using DiT-XL/2 [31]; (ii) text-to-image generation on DrawBench [32] via the rectified flow-based FLUX.1-dev [17, 18, 23]; and (iii) text-to-video generation with HunyuanVideo [15] on the HunyuanLarge architecture. To provide a holistic assessment, all experiments are conducted under official inference protocols and default sampling parameters, with caching methods compared at identical acceleration ratios for fair visualization. Quantitative performance is assessed using FID-50K [11], sFID, and Inception Score for DiT-XL/2; ImageReward [43], CLIP Score [10], CycleReward [1], PSNR, SSIM [13], and LPIPS [46] for FLUX.1-dev; and the 16-dimensional VBench suite [14] for HunyuanVideo. Finally, computational efficiency is measured via relative FLOPs and latency against baselines.

4.2 Class-Conditional Image Generation

Quantitative Analysis. As demonstrated in Table 1, BRACE consistently outperforms all baselines in both generative fidelity and efficiency across various skip intervals, achieving the lowest FID and sFID in every configuration. This advantage stems from our second-form barycentric framework, which inherently tames the trajectory irregularities that plague conventional methods. Specifically, while derivative-driven polynomial approaches [8, 21] suffer from extrapolation instability—often leading to structural vulnerabilities and ghosting artifacts—BRACE maintains superior geometric integrity.

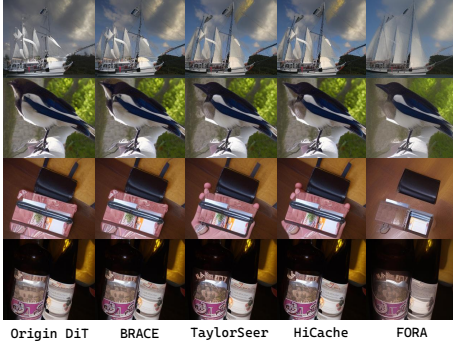


Figure 4: Qualitative comparison of different acceleration methods on DiT-XL/2

Table 1: Quantitative comparison of class-conditional image generation on ImageNet with DiT-XL/2

Method	Lat. (s)↓	FLOPs (T)↓	Spd.↑	FID↓	sFID↓	IS↑
DDIM-50 steps [37]	1.18	23.74	1.00×	2.29	4.29	237.31
DDIM-25 steps [37]	0.62	11.87	2.00×	2.91	4.57	232.22
DDIM-15 steps [37]	0.42	7.12	3.33×	5.09	6.11	205.54
FORA($N=4$) [35]	0.45	6.66	3.56×	4.73	8.45	214.20
Taylor ($N=4$, $O=4$) [21]	0.61	6.66	3.56×	2.59	5.18	232.90
HiCache ($N=4$, $O=4$) [8]	0.57	6.66	3.56×	2.51	5.20	232.72
BRACE($N=4$, $C=3$, $\gamma=0.5$)	0.46	6.66	3.56×	2.46	4.90	233.49
DDIM-12 steps [37]	0.33	5.70	4.17×	7.94	8.07	183.88
FORA($N=5$) [35]	0.37	5.24	4.53×	5.81	9.61	199.96
Taylor ($N=5$, $O=4$) [21]	0.55	5.24	4.53×	2.73	5.34	228.75
HiCache ($N=5$, $O=4$) [8]	0.50	5.24	4.53×	2.67	5.48	229.44
BRACE($N=5$, $C=3$, $\gamma=0.5$)	0.42	5.24	4.53×	2.59	5.03	228.97
DDIM-10 steps [37]	0.30	4.75	5.00×	12.17	11.24	159.20
FORA($N=6$) [35]	0.34	4.76	4.98×	9.22	14.86	169.02
Taylor ($N=6$, $O=4$) [21]	0.50	4.76	4.98×	3.17	6.36	222.00
HiCache ($N=6$, $O=4$) [8]	0.46	4.76	4.98×	3.09	6.30	220.55
BRACE($N=6$, $C=3$, $\gamma=0.5$)	0.39	4.76	4.98×	3.06	5.75	218.20

By bypassing volatile high-order derivative estimations, BRACE strikes a favorable balance between mathematical expressiveness and numerical robustness, providing a more stable alternative for feature extrapolation in accelerated diffusion inference.

Qualitative Comparison. Figure 4 compares different methods on DiT-XL/2 under matched acceleration settings. FORA [35] produces over-smoothing and color shifts, whereas TaylorSeer and HiCache [8, 21] exhibit texture and structural distortions, particularly in Row 3. In contrast, BRACE better preserves clean backgrounds, sharp details, and semantic structure.

4.3 Text-to-Image Generation

Quantitative Analysis. Table 2 reports the results on FLUX.1-dev [17], where BRACE achieves a new state-of-the-art in both semantic alignment and structural fidelity. Across all skip intervals, BRACE consistently yields the highest ImageReward [43], CLIP Score [10], CycleReward [1]; Furthermore, BRACE achieves the lowest LPIPS and highest SSIM in nearly all settings, effectively

confirming that our rational extrapolation effectively preserves intricate fine-grained geometric structures and minimizes perceptual distortion compared to derivative-driven baselines.

Qualitative Comparison. Figure 5 compares BRACE with baseline methods on FLUX.1-dev [17] at an extreme 5.55× speedup. Under this challenging setting, TaylorSeer [21] exhibits semantic and text-structure distortions, HiCache [8] causes over-smoothing and blurred text, and FORA [35] produces noticeable typographical errors. In column 6, the baselines corrupt the double “f” by merging or deforming the characters, whereas BRACE preserves sharp, legible typography consistent with the full-step baseline.

4.4 Text-to-Video Generation

Quantitative Analysis. Table 3 extends our evaluation to text-to-video generation, where BRACE establishes a new state-of-the-art in preserving spatiotemporal fidelity under extreme acceleration. Specifically, at a high skip interval ($N = 6$), BRACE achieves the highest VBench score of 80.14, significantly outperforming both direct reuse (FORA [35]: 78.8) and derivative-driven forecasting methods like TaylorSeer [21] (79.87) and HiCache [8] (79.93). Notably, it successfully recovers the performance level closest to the unaccelerated 50-step baseline, effectively minimizing the quality gap typically observed in high-speed video inference.

Qualitative Comparison. Figure 6 shows that TaylorSeer [21] and FORA [35] lose entities, facial details, or coherent motion in the bird-cat and swimming scenes. In the running-horse scene, TaylorSeer and HiCache [8] produce structural artifacts such as extra legs. Across these cases, the competing methods struggle to preserve both frame-level details and cross-frame dynamics. In contrast, BRACE better preserves semantic content, anatomical structure, and temporal consistency.

4.5 Ablation Study

To evaluate the impact of window capacity and boundary sensitivity on the performance of BRACE, we conduct comprehensive ablations on FLUX.1-dev and DiT-XL/2. Regarding window capacity (Table 4), $C = 3$ provides the optimal approximation of feature curvature. Further increasing C to 4 leads to performance degradation primarily attributed to the accumulation of long-range extrapolation noise, while reducing C to 2 also weakens the feature fitting capability and overall model performance.

We further analyze the boundary sensitivity γ (Table 5). Reconstruction metrics (PSNR, SSIM) improve monotonically as $\gamma \rightarrow 1.0$. However, perceptual quality (*ImageReward*) peaks at $\gamma = 0.7$ and subsequently declines; thus, $\gamma = 0.7$ is established as the default. A more systematic exploration of the underlying causal factors driving this divergence between reconstruction and perceptual metrics is therefore deferred to future work.

Beyond parameter tuning, we investigate different extrapolation schemes. As shown in Table 5, while the *First Barycentric Form* provides a reasonable baseline, it becomes susceptible to numerical drift under aggressive extrapolation. In contrast, our *Second Barycentric Form* ensures robust structural stability. Within this rational framework, we benchmark our adapted Chebyshev weights against three classical schemes: Uniform, Berrut, and Floater-Hormann [9].

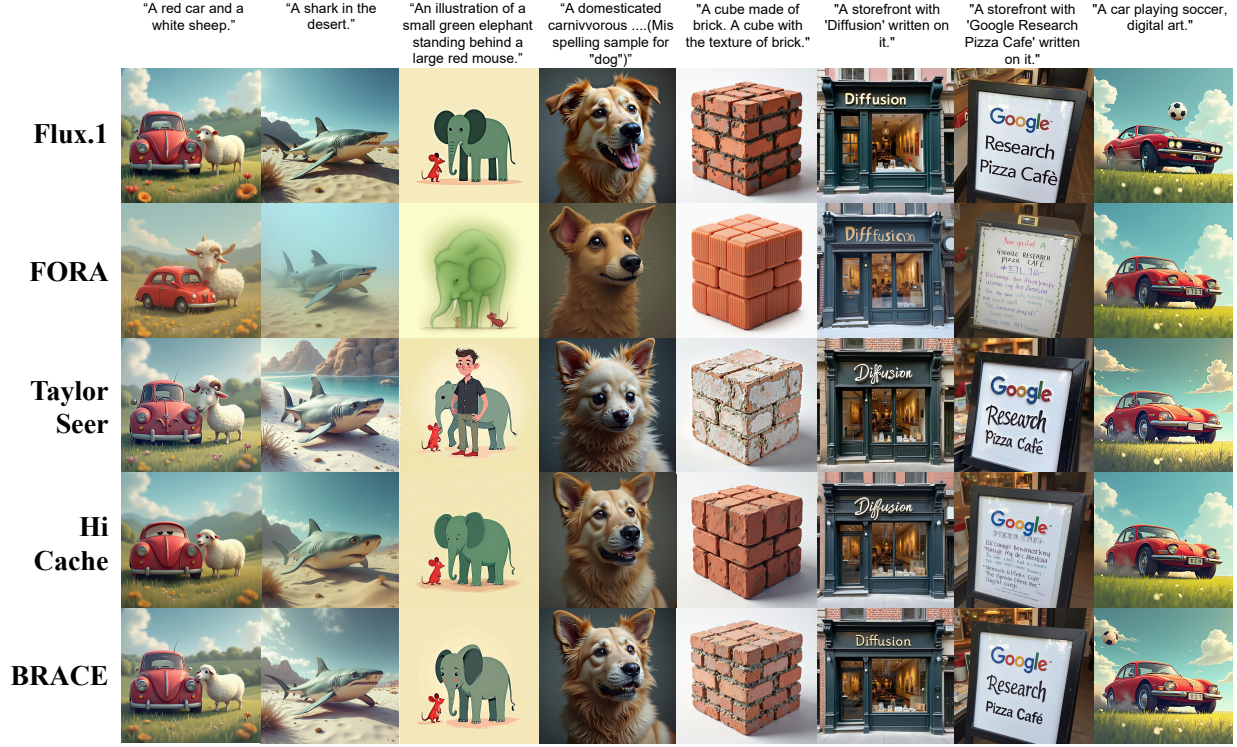


Figure 5: Visualization examples for different acceleration methods on Flux.

Table 2: Quantitative comparison of text-to-image generation on FLUX

Method	Latency(s)↓	FLOPs(T)↓	Speed↑	Img. Reward↑	CLIP Score↑	Cycle Reward↑	PSNR↑	SSIM↑	LPIPS↓
FLUX.1 [dev] - 50 steps [†]	17.18	3719.50	1.00×	0.9898	27.4611	0.8800	∞	1.0000	0.0000
FLUX.1 [dev] - 25 steps	8.80	1859.75	2.00×	0.9375	27.3388	0.8466	19.2599	0.7749	0.3583
FLUX.1 [dev] - 20 steps	7.12	1487.80	2.62×	0.9387	27.2505	0.8466	18.1371	0.7484	0.4006
FORA ($N=5$) [35]	5.01	893.54	4.16×	0.8253	27.2377	0.8353	15.9899	0.6677	0.5206
TaylorSeer ($N=5, O=1$) [21]	5.14	893.54	4.16×	0.9919	27.5143	0.8540	18.4497	0.7424	0.4146
HiCache ($N=5, O=1$) [8]	5.12	893.54	4.16×	0.9350	27.4034	0.8346	19.2595	0.7503	0.3911
BRACE[‡] ($N=5, C=2, \gamma=0.7$)	5.04	893.54	4.16×	1.0021	27.5886	0.8596	19.3671	0.7598	0.3770
FORA ($N=6$) [35]	4.40	744.81	4.99×	0.7836	27.1541	0.8087	15.7298	0.6663	0.5253
TaylorSeer ($N=6, O=1$) [21]	4.57	744.81	4.99×	0.9850	27.5199	0.8480	17.5396	0.7078	0.4657
HiCache ($N=6, O=1$) [8]	4.55	744.81	4.99×	0.9427	27.6898	0.8539	18.5159	0.7155	0.4503
HiCache ($N=6, O=2$) [8]	5.03	744.81	4.99×	0.9892	27.6586	0.8428	18.5044	0.7268	0.4316
BRACE[‡] ($N=6, C=2, \gamma=0.7$)	4.43	744.81	4.99×	1.0025	27.7246	0.8709	18.6275	0.7285	0.4246
FORA ($N=7$) [35]	4.10	670.44	5.55×	0.7289	27.0135	0.7962	15.5290	0.6548	0.5437
TaylorSeer ($N=7, O=1$) [21]	4.31	670.44	5.55×	0.9420	27.4039	0.7970	16.8563	0.6844	0.5062
HiCache ($N=7, O=1$) [8]	4.20	670.44	5.55×	0.9191	27.6949	0.8494	18.2778	0.7032	0.4708
HiCache ($N=7, O=2$) [8]	4.72	670.44	5.55×	0.9881	27.5651	0.8384	17.9911	0.7098	0.4592
BRACE[‡] ($N=7, C=2, \gamma=0.7$)	4.12	670.44	5.55×	0.9884	27.7372	0.8626	18.1925	0.7137	0.4491

While these provide only basic stability, our adapted weights ultimately achieve superior alignment with the nonlinear evolution of DiT features by effectively capturing the asymmetric temporal dynamics of the generative trajectory.

5 Conclusion

We presented **BRACE**, a training-free acceleration method for diffusion transformers that revisits feature caching through the lens of sequence-based forecasting. Motivated by the inherent structural stability of barycentric rational interpolation, we abandon



Figure 6: Qualitative comparison on Text-to-Video Generation.

Table 3: Quantitative comparison of text-to-video generation on HunyuanVideo

Method	Lat. (s)↓	Spd.↑	FLOPs (T)↓	Spd.↑	VBench (%)↑
Original (50 steps)	190.2	1.00×	29773	1.00×	80.68
DDIM (22% steps) [37]	42.7	4.55×	6550	4.55×	78.73
FORA (N=6) [35]	37.1	5.13×	5359	5.56×	78.86
TaylorSeer (N=6, O=1) [21]	37.8	5.03×	5359	5.56×	79.87
HiCache (N=6, O=1) [8]	37.8	5.03×	5359	5.56×	79.93
BRACE (N=6, C=2, $\gamma=0.4$)	37.9	5.03×	5359	5.56×	80.14

Table 4: Performance on DiT-XL/2 using different window capacity

Method	Speed↑	FID↓	sFID↓	IS↑
DDIM-50 [37]	1.00×	2.29	4.29	237.31
BRACE (C = 2)	3.56×	2.53	5.15	231.45
BRACE (C = 3)	3.56×	2.46	4.90	233.49
BRACE (C = 4)	3.56×	2.58	5.32	234.19
HiCache (N = 4) [8]	3.56×	2.51	5.20	232.72
TaylorSeer (N = 4) [21]	3.56×	2.59	5.18	232.90

the reliance on conventional finite-difference derivative estimation and polynomial extrapolation—which frequently exhibit numerical instability around sharp irregularities. Instead, BRACE directly extrapolates cached hidden states using this barycentric rational form, ensuring predictable stability even under long skip

Table 5: Performance on FLUX using different boundary sensitivity

Method	Speed↑	Reward↑	PSNR↑	SSIM↑	LPIPS↓
BRACE ($\gamma = 0.4$)	4.16×	0.9500	16.5315	0.6850	0.5150
BRACE ($\gamma = 0.5$)	4.16×	1.0004	18.4343	0.7421	0.4132
BRACE ($\gamma = 0.6$)	4.16×	0.9991	19.1061	0.7565	0.3840
BRACE ($\gamma = 0.65$)	4.16×	0.9972	19.2665	0.7587	0.3791
BRACE ($\gamma = 0.7$)	4.16×	1.0021	19.3671	0.7598	0.3770
BRACE ($\gamma = 1.0$)	4.16×	0.9995	19.5934	0.7598	0.3744
BRACE (Uniform)	4.16×	0.8859	19.0233	0.7325	0.4236
BRACE (Berrut)	4.16×	0.9938	18.4343	0.7421	0.4131
BRACE (Floater-Hormann)	4.16×	0.9259	19.0242	0.7322	0.4133
BRACE (First Form)	4.16×	0.9830	18.4116	0.7408	0.4245
HiCache (N=5) [8]	4.16×	0.9350	19.2595	0.7503	0.3911
TaylorSeer (N=5) [21]	4.16×	0.9919	18.4497	0.7424	0.4146
FORA (N=5) [35]	4.16×	0.8253	15.9899	0.6677	0.5206

intervals. Across DiT [31] class-conditional generation on ImageNet, FLUX.1 [17] text-to-image generation, and large-scale text-to-video benchmarks with HunyuanVideo [15], BRACE consistently achieves a better quality–efficiency trade-off than reuse-based caching and derivative-driven polynomial forecasting. By flexibly adapting to the diverse feature dynamics of different DiT architectures, it delivers substantial speedups while maintaining strong fidelity. We hope this work serves to expand the feature forecasting paradigm from a distinct perspective.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 62376172.

References

- [1] Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. 2025. Cycle Consistency as Reward: Learning Image-Text Alignment without Human Preferences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22934–22946.
- [2] Jean-Paul Berrut and Lloyd N Trefethen. 2004. Barycentric lagrange interpolation. *SIAM review* 46, 3 (2004), 501–517.
- [3] Daniel Bolya and Judy Hoffman. 2023. Token merging for fast stable diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4599–4603.
- [4] Qi Cai, Jingwen Chen, Yang Chen, Yehao Li, Fuchen Long, Yingwei Pan, Zhaofan Qiu, Yiheng Zhang, Fengbin Gao, Peihan Xu, Yimeng Wang, Kai Yu, Wenxuan Chen, Ziwei Feng, Zijian Gong, Jianzhuang Pan, Yi Peng, Rui Tian, Siyu Wang, Bo Zhao, Ting Yao, and Tao Mei. 2025. HiDream-II: A High-Efficient Image Generative Foundation Model with Sparse Diffusion Transformer. arXiv:2505.22705 [cs.CV] <https://arxiv.org/abs/2505.22705>
- [5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023. PixArt- α : Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426 [cs.CV] <https://arxiv.org/abs/2310.00426>
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 248–255. doi:10.1109/CVPR.2009.5206848
- [7] Gongfan Fang, Xinyin Ma, and Xinchao Wang. 2023. Structural Pruning for Diffusion Models. arXiv:2305.10924 [cs.LG] <https://arxiv.org/abs/2305.10924>
- [8] Liang Feng, Shikang Zheng, Jiacheng Liu, Yuqi Lin, Qinning Zhou, Peiliang Cai, Xinyu Wang, Junjie Chen, Chang Zou, Yue Ma, and Linfeng Zhang. 2026. Hi-Cache: A Plug-in Scaled-Hermite Upgrade for Taylor-Style Cache-then-Forecast Diffusion Acceleration. In *International Conference on Learning Representations (ICLR)*.
- [9] Michael S Floater and Kai Hormann. 2007. Barycentric rational interpolation with no poles and high rates of approximation. *Numer. Math.* 107, 2 (2007), 315–331.
- [10] Jack Hessel, Ari Holtzman, Maxwell Forbes, Roman Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. arXiv:2104.08718 [cs.CV] <https://arxiv.org/abs/2104.08718>
- [11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs.LG] <https://arxiv.org/abs/1706.08500>
- [12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf
- [13] Alain Horé and Djemel Ziou. 2010. Image Quality Metrics: PSNR vs. SSIM. In *2010 20th International Conference on Pattern Recognition*. 2366–2369. doi:10.1109/ICPR.2010.579
- [14] Ziqi Huang, Yanan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. 2024. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21807–21818.
- [15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. 2024. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603* (2024).
- [16] Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. 2021. DiffWave: A Versatile Diffusion Model for Audio Synthesis. arXiv:2009.09761 [eess.AS] <https://arxiv.org/abs/2009.09761>
- [17] Black Forest Labs. 2024. FLUX. <https://github.com/black-forest-labs/flux>.
- [18] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. arXiv:2506.15742 [cs.GR] <https://arxiv.org/abs/2506.15742>
- [19] Xiuyu Li, Yijiang Liu, Long Lian, Huanrui Yang, Zhen Dong, Daniel Kang, Shanghang Zhang, and Kurt Keutzer. 2023. Q-diffusion: Quantizing diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 17535–17545.
- [20] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. 2023. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503* (2023).
- [21] Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Junjie Chen, and Linfeng Zhang. 2025. From reusing to forecasting: Accelerating diffusion models with taylorseers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 15853–15863.
- [22] Jiacheng Liu, Chang Zou, Yuanhuiyi Lyu, Fei Ren, Shaobo Wang, Kaixin Li, and Linfeng Zhang. 2025. Specra: Accelerating diffusion transformers with speculative feature caching. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 10024–10033.
- [23] Xingchao Liu, Chengyue Gong, and Qiang Liu. 2022. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. arXiv:2209.03003 [cs.LG] <https://arxiv.org/abs/2209.03003>
- [24] Alexander Lobashev, Dmitry Guskov, Maria Larchenko, and Mikhail Tamm. 2025. Hessian geometry of latent space in generative models. *arXiv preprint arXiv:2506.10632* (2025).
- [25] Jinming Lou, Wenyang Luo, Yufan Liu, Bing Li, Xinmiao Ding, Weiming Hu, Yuming Li, and Chenguang Ma. 2024. Token caching for diffusion transformer acceleration. *arXiv preprint arXiv:2409.18523* (2024).
- [26] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems* 35 (2022), 5775–5787.
- [27] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2025. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *Machine Intelligence Research* 22, 4 (June 2025), 730–751. doi:10.1007/s11633-025-1562-4
- [28] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. 2023. Latent Consistency Models: Synthesizing High-Resolution Images with Few-Step Inference. arXiv:2310.04378 [cs.CV] <https://arxiv.org/abs/2310.04378>
- [29] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vandenberg, and Saining Xie. 2024. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*. Springer, 23–40.
- [30] Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2024. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15762–15772.
- [31] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [32] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv:2205.11487 [cs.CV] <https://arxiv.org/abs/2205.11487>
- [33] Tim Salimans and Jonathan Ho. 2022. Progressive Distillation for Fast Sampling of Diffusion Models. arXiv:2202.00512 [cs.LG] <https://arxiv.org/abs/2202.00512>
- [34] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. 2023. Adversarial Diffusion Distillation. arXiv:2311.17042 [cs.CV] <https://arxiv.org/abs/2311.17042>
- [35] Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. 2024. FORA: Fast-Forward Caching in Diffusion Transformer Acceleration. arXiv:2407.01425 [cs.CV] <https://arxiv.org/abs/2407.01425>
- [36] Yuzhang Shang, Zhihang Yuan, Bin Xie, Bingzhe Wu, and Yan Yan. 2023. Post-training Quantization on Diffusion Models. In *CVPR*.
- [37] Jiaming Song, Chenlin Meng, and Stefano Ermon. 2022. Denoising Diffusion Implicit Models. arXiv:2010.02502 [cs.LG] <https://arxiv.org/abs/2010.02502>
- [38] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. 2023. Consistency Models. arXiv:2303.01469 [cs.LG] <https://arxiv.org/abs/2303.01469>
- [39] J. STOER and R. BULIRSCH. 1966. Numerical Treatment of Ordinary Differential Equations by Extrapolation Methods. *Numer. Math.* 8 (1966), 1–13. <http://eudml.org/doc/131682>
- [40] Lloyd N. Trefethen and David Bau. 1997. *Numerical Linear Algebra*. SIAM.
- [41] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wenten Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. 2025. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314* (2025).
- [42] Marcus Webb, Lloyd N Trefethen, and Pedro Gonnet. 2012. Stability of barycentric interpolation formulas for extrapolation. *SIAM Journal on Scientific Computing*

- 34, 6 (2012), A3009–A3015.
- [43] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. ImageReward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. 15903–15935.
 - [44] Zhihang Yuan, Hanling Zhang, Lu Pu, Xuefei Ning, Linfeng Zhang, Tianchen Zhao, Shengen Yan, Guohao Dai, and Yu Wang. 2024. DiTFastAttn: Attention Compression for Diffusion Transformer Models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=51HQpkQy3t>
 - [45] Evelyn Zhang, Bang Xiao, Jiayi Tang, Qianli Ma, Chang Zou, Xuefei Ning, Xuming Hu, and Linfeng Zhang. 2024. Token Pruning for Caching Better: 9 Times Acceleration on Stable Diffusion for Free. arXiv:2501.00375 [cs.CV] <https://arxiv.org/abs/2501.00375>
 - [46] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
 - [47] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. 2023. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. *Advances in Neural Information Processing Systems* 36 (2023), 49842–49869.
 - [48] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. 2024. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588* (2024).
 - [49] Shikang Zheng, Liang Feng, Xinyu Wang, Qinming Zhou, Peiliang Cai, Chang Zou, Jiacheng Liu, Yuqi Lin, Junjie Chen, Yue Ma, and Linfeng Zhang. 2026. Forecast then Calibrate: Feature Caching as ODE for Efficient Diffusion Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*. arXiv:2508.16211 <https://arxiv.org/abs/2508.16211>
 - [50] Chang Zou, Xuyang Liu, Ting Liu, Siteng Huang, and Linfeng Zhang. 2025. Accelerating Diffusion Transformers with Token-wise Feature Caching. arXiv:2410.05317 [cs.LG] <https://arxiv.org/abs/2410.05317>