Novelty Assessment Report

Paper: Let Features Decide Their Own Solvers: Hybrid Feature Caching for Diffusion Transformers
PDF URL: https://openreview.net/pdf?id=URbsHlTK8c
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Diffusion Transformers (DiTs) offer state-of-the-art fidelity in image and video synthesis, but their iterative sampling process remains a major bottleneck due to the high cost of transformer forward passes at each timestep. To mitigate this, feature caching has emerged as a training-free acceleration technique that reuses or forecasts hidden representations. However, existing methods often apply a uniform caching strategy across all feature dimensions, ignoring their heterogeneous dynamic behaviors. Therefore, we adopt a new perspective by modeling hidden feature evolution as a mixture of ODEs across dimensions, and introduce \textbf{HyCa}, a Hybrid ODE solver inspired caching framework that applies dimension-wise caching strategies. HyCa achieves near-lossless acceleration across diverse domains and models, including 5.56$\times$ speedup on FLUX and HunyuanVideo, 6.24$\times$ speedup on Qwen-Image and Qwen-Image-Edit without retraining. \emph{Our code is in supplementary material and will be released on Github.}

Core Task Landscape

This paper addresses: **Accelerating Diffusion Transformer Inference through Feature Caching**

A total of **50 papers** were analyzed and organized into a taxonomy with **23 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Core Feature Caching Mechanisms**
- **Adaptive Caching Strategies**
- **Predictive and Forecasting-Based Caching**
- **Learning-Based Caching Optimization**
- **Architectural and Structural Enhancements**
- **Redundancy Analysis and Profiling**
- **Domain-Specific Caching Applications**
- **Hybrid and Multi-Paradigm Acceleration**
- **Universal and Cross-Architecture Caching**
- **Specialized Application Domains**

Complete Taxonomy Tree

- Accelerating Diffusion Transformer Inference through Feature Caching Survey Taxonomy
- Core Feature Caching Mechanisms
  - Uniform Temporal Caching (3 papers)
  - [1] Accelerating diffusion transformer via error-optimized cache (Junxiang Qiu, 2025) View paper
  - [10] Fora: Fast-forward caching in diffusion transformer acceleration (Selvaraju, 2024) View paper
  - [29] Frdiff: Feature reuse for universal training-free acceleration of diffusion models (Junhyuk So, 2024) View paper
  - Token-Level Selective Caching (4 papers)
  - [4] Token caching for diffusion transformer acceleration (Jing Lou, 2024) View paper
  - [9] Accelerating Diffusion Transformers with Token-wise Feature Caching (Zou Chang, 2024) View paper
  - [22] Compute Only 16 Tokens in One Timestep: Accelerating Diffusion Transformers with Cluster-Driven Feature Caching (Zhi-xin Zheng, 2025) View paper
  - [40] Rethinking Token-wise Feature Caching: Accelerating Diffusion Transformers with Dual Feature Caching (Chang Zou, 2024) View paper
  - Hierarchical and Block-Level Caching (5 papers)
  - [8] Learning-to-cache: Accelerating diffusion transformer via layer caching (Michael Bi Mi, 2024) View paper
  - [15] BWCache: Accelerating Video Diffusion Transformers through Block-Wise Caching (Tang Zhiqing, 2025) View paper
  - [18] Accelerating Diffusion Transformer-Based Text-to-Speech with Transformer Layer Caching (Sakpiboonchit, 2025) View paper
  - [27] Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers (Zhang Hui, 2025) View paper
  - [38] Accelerating Vision Diffusion Transformers with Skip Branches (Guanjie Chen, 2024) View paper
  - Dual and Multi-Stage Caching (2 papers)
  - [5] Accelerating diffusion transformers with dual feature caching (Zou Chang, 2024) View paper
  - [50] H2-Cache: A Novel Hierarchical Dual-Stage Cache for High-Performance Acceleration of Generative Diffusion Models (Sung Min-gyu, 2025) View paper
- Adaptive Caching Strategies
  - Runtime-Adaptive and Content-Aware Caching (3 papers)
  - [16] Less is Enough: Training-Free Video Diffusion Acceleration via Runtime-Adaptive Caching (Zhou Xin, 2025) View paper

    - [23] Adaptive caching for faster video generation with diffusion transformers (Kahatapitiya, 2025) View paper
    - [36] Model Reveals What to Cache: Profiling-Based Feature Reuse for Video Diffusion Models (Ma Xuran, 2025) View paper
    - Frequency and Spectral-Aware Caching (2 papers)
    - [14] Towards Stabilized and Efficient Diffusion Transformers through Long-Skip-Connections with Spectral Constraints (Chen, 2025) View paper
    - [25] Freqca: Accelerating diffusion models via frequency-aware caching (Liu Jiaâ□□Cheng, 2025) View paper
    - Magnitude and Confidence-Based Caching (2 papers)
    - [28] Forecasting When to Forecast: Accelerating Diffusion Models with Confidence-Gated Taylor (Jiang, 2025) View paper
    - [32] MagCache: Fast Video Generation with Magnitude-Aware Cache (Ma Zehong, 2025) View paper
- Predictive and Forecasting-Based Caching
    - Speculative and Forecast-Verify Frameworks (1 papers)
    - [3] Speca: Accelerating diffusion transformers with speculative feature caching (Liu Jiaâ□□Cheng, 2025) View paper
    - Taylor Expansion and Polynomial-Based Forecasting (2 papers)
    - [26] From reusing to forecasting: Accelerating diffusion models with taylorseers (Liu Jiaâ□□Cheng, 2025) View paper
    - [39] HiCache: Training-free Acceleration of Diffusion Models via Hermite Polynomial-based Feature Caching (Feng Liang, 2025) View paper
- Learning-Based Caching Optimization
    - Calibration and Error Compensation (2 papers)
    - [7] Accelerating Diffusion Transformer via Increment-Calibrated Caching with Channel-Aware Singular Value Decomposition (Chen Zhiyuan, 2025) View paper
    - [13] Exposure Bias Reduction for Enhancing Diffusion Transformer Feature Caching (Z Zou, 2025) View paper
    - Gradient-Based and Differentiable Caching (1 papers)
    - [12] Accelerating Diffusion Transformer via Gradient-Optimized Cache (Liu Lin, 2025) View paper
    - Learned Approximation and Compression (3 papers)
    - [6] Fastcache: Fast caching for diffusion transformer through learnable linear approximation (Liu, 2025) View paper
    - [21] HarmoniCa: Harmonizing Training and Inference for Better Feature Cache in Diffusion Transformer Acceleration (Yushi Huang, 2024) View paper
- Architectural and Structural Enhancements
    - Skip Connections and Long-Range Preservation (1 papers)
    - [41] DDT: Decoupled Diffusion Transformer (Wang, 2025) View paper
    - Decoupled and Modular Architectures (1 papers)
    - [42] ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer (Hu Jinyi, 2024) View paper
- Redundancy Analysis and Profiling (2 papers)
    - [17] Unveiling redundancy in diffusion transformers (dits): A systematic study (Sun Xi-bo, 2024) View paper
    - [35] OmniCache: A Trajectory-Oriented Global Perspective on Training-Free Cache Reuse for Diffusion Transformer Models (Chu, 2025) View paper
- Domain-Specific Caching Applications
    - Video Generation Caching (3 papers)
    - [24] Fast and memory-efficient video diffusion using streamlined inference (Yifan Gong, 2024) View paper
    - [30] Fastercache: Training-free video diffusion model acceleration with high quality (Lv, 2024) View paper
    - [31] Foresight: Adaptive Layer Reuse for Accelerated and High-Quality Text-to-Video Generation (Adnan Muhammad, 2025) View paper
    - Text-to-Speech and Language Model Caching (3 papers)
    - [2] Accelerating diffusion language model inference via efficient kv caching and guided diffusion (Z Hu, 2025) View paper
    - [34] dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching (Liu Zhi-Yuan, 2025) View paper
    - Conditional and Controllable Generation Caching (2 papers)
    - [19] Ominicontrol2: Efficient conditioning for diffusion transformers (Tan Zhen-xiong, 2025) View paper
    - [20] Fulldit2: Efficient in-context conditioning for video diffusion transformers (He Xuan-hua, 2025) View paper
- Hybrid and Multi-Paradigm Acceleration
    - Caching with Parallelization (1 papers)
    - [11] Pipefusion: Patch-level pipeline parallelism for diffusion transformers inference (Fang Jiarui, 2024) View paper
    - Caching with Pruning and Compression (1 papers)
    - [43] Token Pruning for Caching Better: 9 Times Acceleration on Stable Diffusion for Free (Xiao Bang, 2024) View paper
    - Caching with ODE Solvers and Sampling Optimization ★ (3 papers)
    - [0] Let Features Decide Their Own Solvers: Hybrid Feature Caching for Diffusion Transformers (Anon et al., 2026) View paper
    - [44] LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers (Xuan Shen, 2024) View paper
    - [46] AB-Cache: Training-Free Acceleration of Diffusion Models via Adams-Bashforth Cached Feature Reuse (Zou Zhen, 2025) View paper
- Universal and Cross-Architecture Caching (1 papers)
    - [37] Smoothcache: A universal inference acceleration technique for diffusion transformers (Joseph Liu, 2025) View paper
- Specialized Application Domains (3 papers)
    - [45] DiTVR: Zero-Shot Diffusion Transformer for Video Restoration (Gao, 2025) View paper
    - [47] Multivariate Diffusion Transformer with Decoupled Attention for High-Fidelity Mask-Text Collaborative Facial Generation (Yushe Cao, 2025) View paper
    - [48] CELAT-DiffNet: channel-enhanced local attention transformer for underwater image enhancement based on diffusion models (Jiale Wang, 2025) View paper

Narrative

Core task: accelerating diffusion transformer inference through feature caching. The field has organized itself around several complementary strategies for reducing computational overhead in diffusion models. Core Feature Caching Mechanisms establish foundational techniques such as token-level reuse (Token Caching[4], KV Caching Diffusion[2]) and dual-stream approaches (Dual Feature Caching[5]), while Adaptive Caching Strategies introduce runtime flexibility through methods like Runtime-Adaptive Caching[16] and cluster-driven selection (Cluster-Driven Caching[22]). Predictive and Forecasting-Based Caching leverages Taylor expansions and confidence gating (TaylorSeers[26], Confidence-Gated Taylor[28]) to anticipate future features, whereas Learning-Based Caching

Optimization trains policies or networks to decide what and when to cache (Learning-to-Cache[8]). Architectural and Structural Enhancements modify model designs directly (Long-Skip-Connections[14], Decoupled Diffusion Transformer[41]), and Redundancy Analysis and Profiling systematically identify reusable computations (Unveiling Redundancy[17], Profiling-Based Reuse[36]). Domain-Specific Caching Applications tailor strategies to video generation (Adaptive Caching Video[23]) or text-to-speech (Text-to-Speech Caching[18]), while Hybrid and Multi-Paradigm Acceleration combines caching with ODE solvers or sampling optimizations, and Universal and Cross-Architecture Caching aims for broad applicability across model families (OmniCache[35]).

Recent work has explored trade-offs between caching granularity, error accumulation, and computational savings. Fine-grained token-wise methods (Token-wise Feature Caching[9], Rethinking Token-wise Caching[40]) offer precise control but may introduce overhead, whereas block-level or layer-skipping approaches (BlockDance[27], Skip Branches[38]) achieve coarser speedups with simpler logic. Hybrid Feature Caching[0] sits within the Hybrid and Multi-Paradigm Acceleration branch, combining caching with ODE solver refinements to balance quality and speed—a direction also pursued by LazyDiT[44] and AB-Cache[46], which similarly integrate sampling optimizations. Compared to purely adaptive schemes like Runtime-Adaptive Caching[16] or purely predictive methods like TaylorSeers[26], Hybrid Feature Caching[0] emphasizes synergy between multiple acceleration paradigms, aiming to mitigate the exposure bias and error drift that can arise when caching decisions are made in isolation from the underlying numerical solver.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers

Authors: Xuan Shen, Zhao Song, Yufa Zhou, Bo Chen, Yanyu Li, et al. (15 authors total) | **Year/Venue**: 2024 • AAAI Conference on Artificial Intelligence | **URL**: View paper

Abstract

Diffusion Transformers have emerged as the preeminent models for a wide array of generative tasks, demonstrating superior performance and efficacy across various applications. The promising results come at the cost of slow inference, as each denoising step requires running the whole transformer model with a large amount of parameters. In this paper, we show that performing the full computation of the model at each diffusion step is unnecessary, as some computations can be skipped by lazily reusi...

Relationship Analysis

Both papers belong to the 'Caching with ODE Solvers and Sampling Optimization' category, integrating feature caching with advanced numerical methods to accelerate diffusion transformer inference. They overlap in their core approach of reusing cached features to skip redundant computations during the denoising process. However, the original paper (HyCa) models feature evolution as a mixture of ODEs with dimension-wise hybrid solvers selected through clustering, while LazyDiT employs a lazy learning framework with trainable layers that dynamically decide which computations to skip based on similarity between consecutive steps.

---

2. AB-Cache: Training-Free Acceleration of Diffusion Models via Adams-Bashforth Cached Feature Reuse

Authors: Zou Zhen, ZHANG Chengwei, Huang, Jie, Zhao Feng, et al. (9 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Diffusion models have demonstrated remarkable success in generative tasks, yet their iterative denoising process results in slow inference, limiting their practicality. While existing acceleration methods exploit the well-known U-shaped similarity pattern between adjacent steps through caching mechanisms, they lack theoretical foundation and rely on simplistic computation reuse, often leading to performance degradation. In this work, we provide a theoretical understanding by analyzing the denois...

Relationship Analysis

Both papers belong to the same taxonomy category of integrating caching with advanced ODE solvers for accelerating diffusion transformer inference. They overlap in applying numerical ODE methods (Adams-Bashforth in AB-Cache, hybrid solvers including Adams-Bashforth/Adams-Moulton in HyCa) to predict cached features during diffusion sampling. The key difference is that HyCa introduces dimension-wise clustering to assign different solvers to feature groups based on their dynamic behaviors, while AB-Cache applies a uniform k-th order Adams-Bashforth linear combination across all dimensions without heterogeneous treatment.

Contributions Analysis

Overall novelty summary. The paper proposes HyCa, a hybrid caching framework that models hidden feature evolution as a mixture of ODEs and applies dimension-wise caching strategies. It resides in the 'Caching with ODE Solvers and Sampling Optimization' leaf, which contains only three papers total, including this work and two siblings (AB-Cache and LazyDiT). This represents a relatively sparse research direction within the broader taxonomy of 50 papers across 23 leaf nodes, suggesting the integration of ODE-inspired solvers with feature caching remains an emerging area rather than a saturated one.

The taxonomy reveals that most caching research clusters around core mechanisms (uniform temporal, token-level selective, hierarchical block-level) and adaptive strategies (runtime-adaptive, frequency-aware, magnitude-based). HyCa's parent branch, 'Hybrid and Multi-Paradigm Acceleration,' also includes leaves for caching with parallelization and caching with pruning, indicating the field is exploring synergies between caching and complementary acceleration techniques. The scope note for HyCa's leaf explicitly excludes 'pure caching without solver integration,' positioning this work at the intersection of numerical methods and feature reuse—a boundary less explored than standalone caching or standalone solver optimization.

Among 30 candidates examined, the contribution-level analysis shows mixed novelty signals. 'Heterogeneous Feature Dynamics' (10 candidates, 0 refutable) and 'State-of-the-Art Acceleration Performance' (10 candidates, 0 refutable) appear to have no clear prior work overlap within the limited search scope. However, 'HyCa: Hybrid Feature Caching Framework' (10 candidates, 1 refutable) encounters at least one candidate that provides overlapping prior work, suggesting the core framework design may share conceptual or technical elements with existing methods. The scale of this search—30 papers total—means these findings reflect top semantic matches rather than exhaustive coverage.

Given the sparse population of the ODE-solver-caching leaf and the absence of refutation for two of three contributions, the work appears to occupy a relatively novel niche within the examined scope. The single refutable candidate for the framework contribution indicates some prior overlap exists, but the limited search scale and the emerging nature of this hybrid paradigm suggest the paper may still offer substantive advances. A broader literature review would be needed to confirm whether the dimension-wise ODE mixture modeling and the specific solver integration represent genuine departures from existing hybrid acceleration methods.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Heterogeneous Feature Dynamics in Diffusion Transformers

Description: The authors demonstrate that hidden feature dimensions in Diffusion Transformers evolve according to distinct temporal patterns rather than a single unified process. Through clustering analysis, they reveal that these dynamics are consistent across prompts, timesteps, and resolutions, motivating the need for dimension-specific solvers.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Forecast then calibrate: Feature caching as ode for efficient diffusion transformers

URL: View paper

Brief Assessment

Forecast Then Calibrate[69] focuses on modeling feature evolution as a single feature-ODE and applying predictor-corrector methods, rather than analyzing heterogeneous dynamics across feature dimensions with distinct temporal patterns and clustering.

2. Dynamic diffusion transformer

URL: View paper

Brief Assessment

Dynamic Diffusion Transformer[64] focuses on computational redundancy across timesteps and spatial regions in diffusion transformers, proposing dynamic width and token strategies. It does not demonstrate or analyze heterogeneous feature dynamics with distinct temporal patterns across dimensions through clustering analysis as claimed in the original contribution.

3. Diffusion Transformers for Tabular Data Time Series Generation

URL: View paper

Brief Assessment

Tabular Time Series[71] focuses on tabular data time series generation using diffusion transformers for heterogeneous tabular data (numerical and categorical fields), not on analyzing hidden feature dimension dynamics within diffusion transformers themselves. The candidate addresses a completely different problem domain.

4. Transformer-Based spatiotemporal graph diffusion convolution network for traffic flow forecasting

URL: View paper

Brief Assessment

Traffic Flow Forecasting[73] focuses on spatial-temporal traffic prediction using graph neural networks and transformers for transportation systems, not on analyzing heterogeneous feature dynamics within diffusion transformers for generative modeling.

5. Precipitation Nowcasting Using Diffusion Transformer With Causal Attention

URL: View paper

Brief Assessment

Precipitation Nowcasting[68] focuses on weather forecasting using diffusion transformers for precipitation prediction, not on analyzing heterogeneous feature dynamics or temporal patterns in diffusion model hidden representations. The paper does not address feature dimension clustering or dimension-specific solvers.

6. Diffusion models for intelligent transportation systems: A survey

URL: View paper

Brief Assessment

Transportation Systems Survey[70] focuses on diffusion models for intelligent transportation systems (traffic forecasting, autonomous driving, etc.), not on analyzing internal feature dynamics within diffusion transformer architectures. The survey does not examine heterogeneous temporal patterns of hidden feature dimensions.

7. Emergent Temporal Correspondences from Video Diffusion Transformers

URL: View paper

Brief Assessment

Emergent Temporal Correspondences[65] analyzes temporal correspondence in video diffusion transformers through cross-frame attention mechanisms, focusing on how different layers and timesteps contribute to temporal matching. This differs from the original paper's focus on heterogeneous feature dynamics across dimensions within individual layers, where different feature dimensions evolve according to distinct temporal patterns requiring dimension-specific solvers.

8. A survey on diffusion models for time series and spatio-temporal data

URL: View paper

Brief Assessment

Time Series Survey[66] discusses heterogeneous temporal modalities in time series contexts, not feature dynamics within diffusion transformer architectures. The candidate focuses on time series and spatio-temporal data applications, while the original paper analyzes hidden feature evolution patterns within diffusion transformers during image/video generation.

9. Post-Training Quantization for Diffusion Transformer via Hierarchical Timestep Grouping

URL: View paper

Brief Assessment

Hierarchical Timestep Grouping[67] focuses on quantization challenges in diffusion transformers, specifically addressing channel-dependent activation outliers and timestep-dependent outlier distributions for post-training quantization. This is a different technical problem from modeling heterogeneous feature dynamics for acceleration purposes.

10. Spatio-Temporal Probabilistic Forecasting of Wind Speed Using Transformer-Based Diffusion Models

URL: View paper

Brief Assessment

Wind Speed Forecasting[72] focuses on spatio-temporal wind speed prediction using diffusion models for probabilistic forecasting, not on analyzing heterogeneous feature dynamics within diffusion transformer architectures during image/video generation.

Contribution 2: HyCa: Hybrid Feature Caching Framework

Description: HyCa is a training-free acceleration framework that models hidden feature evolution as a mixture of ODEs. It clusters feature dimensions by their temporal behaviors and assigns the optimal ODE solver to each cluster through a one-time offline optimization, enabling efficient and adaptive feature prediction during inference.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. MPQ-DM: Mixed Precision Quantization for Extremely Low Bit Diffusion Models

URL: View paper

Brief Assessment

MPQ-DM[57] focuses on mixed-precision quantization for diffusion models to reduce bit-width and computational cost, not on hybrid ODE solvers or feature caching strategies for acceleration during inference.

---

2. Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model

URL: View paper

Brief Assessment

Timestep Embedding Cache[54] focuses on video diffusion models and uses timestep embedding-modulated noisy inputs as caching indicators, whereas HyCa models feature evolution as a mixture of ODEs with dimension-wise clustering and hybrid solver assignment. The technical approaches and problem formulations differ fundamentally.

---

3. Frdiff: Feature reuse for universal training-free acceleration of diffusion models

URL: View paper

Prior Art Analysis

FrDiff[29] demonstrates prior work on feature caching for diffusion model acceleration that predates the ORIGINAL paper's HyCa framework. Both papers address the same core problem: exploiting temporal redundancy in diffusion models through feature reuse/caching to reduce computational costs. FrDiff[29] already proposes reusing intermediate feature maps across timesteps based on temporal similarity analysis, introduces a score mixing technique to balance coarse and fine-grained generation, and provides an automated tuning method (Auto-FR) to optimize caching intervals. The ORIGINAL paper's claim of being first to model feature evolution as a mixture of ODEs and assign dimension-wise solvers represents an incremental refinement rather than a fundamentally novel contribution, as the core insight of temporal redundancy exploitation through feature caching was already established by FrDiff[29].

Evidence

Evidence 1 - **Rationale**: FrDiff[29] explicitly proposes feature reuse as a method to exploit temporal coherence, demonstrating prior work on the same core mechanism that HyCa claims as novel. - **Original**: training-free feature caching has emerged as a promising solution. it exploits the temporal coherence of hidden representations by reusing features, thereby reducing redundant computation. - **Candidate**: by reusing these intermediate feature maps with higher temporal similarity, we can significantly reduce computation overhead while maintaining output quality. building on this insight, we propose a new optimization potential namedfeature reuse (fr).

Evidence 2 - **Rationale**: Both papers conduct temporal similarity analysis to identify which features can be reused/cached, showing FrDiff[29] already performed this type of analysis. - **Original**: we analyze how each feature dimension changes over timesteps and group them into clusters based on their dynamics. as shown in fig. 2(a), some dimensions fluctuate sharply with oscillatory patterns, indicating stiffness or multimodal behavior, while others evolve smoothly and predictably - **Candidate**: according to our extensive analysis, specific modules within diffusion models show considerable similarity in their feature maps across adjacent frames. by reusing these intermediate feature maps with higher temporal similarity, we can significantly reduce computation overhead while maintaining outp...

Evidence 3 - **Rationale**: FrDiff[29] already implements selective feature reuse based on temporal patterns, which is conceptually similar to HyCa's dimension-wise solver assignment based on clustering. - **Original**: we introducehyca, a hybrid caching framework that models hidden feature evolution as a mixture of odes and applies suitable solvers for every dimension. hyca begins with unsupervised clustering, grouping dimensions with similar dynamic behaviors - **Candidate**: for non-keyframe timesteps t′ /∈ k, the computation of thes(·) is replaced by the saved memory from the nearest early timestept ∈ k, as follows: yt′ i = fi(mt i, t′) + xt′ i . (5) hence, by skipping the operationss(·), a significant amount of computation can be saved

Evidence 4 - **Rationale**: FrDiff[29]'s Auto-FR provides automated optimization of feature reuse policies, similar to HyCa's one-time solver selection, demonstrating prior work on automated tuning for feature caching. - **Original**: hyca assigns the most suitable solver to each cluster. normally, identifying the best solver would require running inference on a large set of images and comparing quantitative metrics. however, surprisingly, we found that cluster assignments are highly stable across resolutions, timesteps, and even... - **Candidate**: in this approach, we apply a timestep-wise learnable parameterαt = sigmoid(θt) with a hard gating mechanism update. it's important to note that the network parameters are frozen and training-free; only the gating parameters are updated. the reuse policy is trained to maximize quality while minimizin...

---

4. Blended latent diffusion

URL: View paper

Brief Assessment

Blended Latent Diffusion[53] focuses on local text-driven image editing using latent space blending, not on hybrid ODE solvers for feature caching across diffusion transformer architectures. The technical approaches are fundamentally different.

---

5. Approximate caching for efficiently serving {Text-to-Image} diffusion models

URL: View paper

Brief Assessment

Approximate Caching[55] focuses on caching intermediate noise states in diffusion models for text-to-image generation, not on hybrid ODE solvers for feature dimension clustering. The candidate addresses a different technical problem in the diffusion model pipeline.

---

6. Token Pruning for Caching Better: 9 Times Acceleration on Stable Diffusion for Free

URL: View paper

Brief Assessment

Token Pruning Caching[43] focuses on token-level pruning based on dynamics to extend feature dynamics across timesteps, while HyCa models feature evolution as a mixture of ODEs with dimension-wise solver assignment. The candidate addresses token pruning rather than hybrid ODE solver strategies for feature dimensions.

---

7. Sortblock: Similarity-Aware Feature Reuse for Diffusion Model

URL: View paper

Brief Assessment

SortBlock[56] focuses on block-wise feature reuse based on similarity ranking across timesteps, not on modeling feature evolution as a mixture of ODEs with cluster-specific solver assignment as in HyCa.

---

8. Deepcache: Accelerating diffusion models for free

URL: View paper

Brief Assessment

DeepCache[51] focuses on uniform feature caching by reusing high-level features across consecutive steps in U-Net architectures, without clustering dimensions by temporal behaviors or assigning different ODE solvers to different feature groups. HyCa's core novelty lies in modeling feature evolution as a mixture of ODEs with dimension-wise clustering and adaptive solver assignment, which is fundamentally different from DeepCache's uniform caching approach.

---

9. Cachequant: Comprehensively accelerated diffusion models

URL: View paper

Brief Assessment

CacheQuant[52] focuses on jointly optimizing model caching and quantization for diffusion models, whereas HyCa addresses hybrid ODE solver assignment for feature dimensions based on their temporal dynamics. The candidate does not demonstrate prior work on dimension-wise solver assignment or mixture of ODEs modeling.

---

10. Model Reveals What to Cache: Profiling-Based Feature Reuse for Video Diffusion Models

URL: View paper

Brief Assessment

Profiling-Based Reuse[36] focuses on video diffusion models with foreground/background segmentation-based caching, while HyCa addresses general diffusion transformers using ODE-based dimension-wise solver assignment. The technical approaches are fundamentally different.

Contribution 3: State-of-the-Art Acceleration Performance Across Diverse Tasks

Description: The authors demonstrate that HyCa achieves near-lossless acceleration across multiple domains and models, including 5.56× speedup on FLUX and HunyuanVideo, and 6.24× speedup on Qwen-Image and Qwen-Image-Edit, without requiring retraining. The method is also compatible with distillation techniques, reaching up to 24.4× speedup.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Speca: Accelerating diffusion transformers with speculative feature caching

URL: View paper

Brief Assessment

Speca[3] focuses on speculative feature caching with a 'forecast-then-verify' mechanism for diffusion transformers, achieving different acceleration ratios (6.34× on FLUX, 6.1× on HunyuanVideo) than HyCa's reported performance (5.56× on FLUX and HunyuanVideo, 6.24× on Qwen-Image). The technical approaches differ fundamentally: Speca[3] uses speculative sampling with verification mechanisms, while the original paper employs hybrid ODE solvers with dimension-wise caching strategies.

---

2. Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning

URL: View paper

Brief Assessment

Similarity-Based Pruning[59] focuses on token pruning for diffusion models, achieving acceleration through reducing input data size. HyCa addresses a different technical problem—feature caching across timesteps using hybrid ODE solvers—rather than token reduction, making direct comparison of novelty claims inappropriate.

---

3. Forecasting When to Forecast: Accelerating Diffusion Models with Confidence-Gated Taylor

URL: View paper

Brief Assessment

Confidence-Gated Taylor[28] focuses on module-level vs. block-level prediction strategies and dynamic caching mechanisms for taylor-based acceleration, rather than demonstrating state-of-the-art performance across the same diverse set of tasks (FLUX, HunyuanVideo, Qwen-Image, Qwen-Image-Edit) with comparable speedup metrics and distillation compatibility as claimed in the original paper.

---

4. Learning-to-cache: Accelerating diffusion transformer via layer caching

URL: View paper

Brief Assessment

Learning-to-Cache[8] focuses on layer-level caching within diffusion transformers for image generation tasks (ImageNet), while HyCa addresses dimension-wise feature caching across multiple modalities (text-to-image, text-to-video, image editing). The candidate does not demonstrate prior work on HyCa's multi-task acceleration claims or its specific speedup achievements on FLUX, HunyuanVideo, Qwen-Image, and Qwen-Image-Edit.

---

5. Ac3d: Analyzing and improving 3d camera control in video diffusion transformers

URL: View paper

Brief Assessment

AC3D[61] focuses on 3D camera control in video diffusion transformers, not on training-free acceleration methods for diffusion transformers. The paper addresses camera motion modeling and control rather than computational acceleration techniques.

---

6. dkv-cache: The cache for diffusion language models

URL: View paper

Brief Assessment

dKV-Cache[62] focuses on accelerating diffusion language models through a KV-cache mechanism for the denoising process, not diffusion transformers for image/video generation. The technical domains and architectures are fundamentally different.

---

7. Attention-driven training-free efficiency enhancement of diffusion models

URL: View paper

Brief Assessment

Attention-Driven Efficiency[60] focuses on training-free token pruning for diffusion models using attention maps, specifically targeting single-image generation tasks. The candidate does not address multi-domain acceleration across text-to-image, text-to-video, and image editing tasks simultaneously, nor does it demonstrate compatibility with distillation techniques as claimed in the original contribution.

---

8. Accelerating Diffusion Transformers with Token-wise Feature Caching

URL: View paper

Brief Assessment

Token-wise Feature Caching[9] focuses on token-level caching strategies for diffusion transformers, achieving acceleration through selective token computation. While both methods achieve high speedups, Token-wise Feature Caching[9] operates at the token granularity rather than the dimension-wise hybrid solver approach of the original paper, representing a fundamentally different technical approach to acceleration.

---

9. -DiT: A Training-Free Acceleration Method Tailored for Diffusion Transformers

URL: View paper

Brief Assessment

DiT Training-Free[58] focuses specifically on diffusion transformers with a cache mechanism (Δ-cache) and stage-adaptive acceleration, targeting different architectural components than HyCa's hybrid ODE solver approach. The candidate does not demonstrate prior work that refutes HyCa's novelty claims about achieving near-lossless acceleration across multiple domains using dimension-wise hybrid caching strategies.

---

10. Block-wise Adaptive Caching for Accelerating Diffusion Policy

URL: View paper

Brief Assessment

Block-wise Adaptive Caching[63] focuses specifically on accelerating diffusion policy for robotic control tasks, not general diffusion transformers across image/video generation. The architectural focus (action generation vs. visual synthesis) and application domain (robotics vs. multimedia) are fundamentally different from HyCa's contributions.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Let Features Decide Their Own Solvers: Hybrid Feature Caching for Diffusion Transformers View paper
- [1] Accelerating diffusion transformer via error-optimized cache View paper
- [2] Accelerating diffusion language model inference via efficient kv caching and guided diffusion View paper
- [3] Speca: Accelerating diffusion transformers with speculative feature caching View paper
- [4] Token caching for diffusion transformer acceleration View paper
- [5] Accelerating diffusion transformers with dual feature caching View paper
- [6] Fastcache: Fast caching for diffusion transformer through learnable linear approximation View paper
- [7] Accelerating Diffusion Transformer via Increment-Calibrated Caching with Channel-Aware Singular Value Decomposition View paper
- [8] Learning-to-cache: Accelerating diffusion transformer via layer caching View paper
- [9] Accelerating Diffusion Transformers with Token-wise Feature Caching View paper
- [10] Fora: Fast-forward caching in diffusion transformer acceleration View paper
- [11] Pipefusion: Patch-level pipeline parallelism for diffusion transformers inference View paper
- [12] Accelerating Diffusion Transformer via Gradient-Optimized Cache View paper
- [13] Exposure Bias Reduction for Enhancing Diffusion Transformer Feature Caching View paper
- [14] Towards Stabilized and Efficient Diffusion Transformers through Long-Skip-Connections with Spectral Constraints View paper
- [15] BWCache: Accelerating Video Diffusion Transformers through Block-Wise Caching View paper
- [16] Less is Enough: Training-Free Video Diffusion Acceleration via Runtime-Adaptive Caching View paper
- [17] Unveiling redundancy in diffusion transformers (dits): A systematic study View paper
- [18] Accelerating Diffusion Transformer-Based Text-to-Speech with Transformer Layer Caching View paper
- [19] Ominicontrol2: Efficient conditioning for diffusion transformers View paper
- [20] Fulldit2: Efficient in-context conditioning for video diffusion transformers View paper
- [21] HarmoniCa: Harmonizing Training and Inference for Better Feature Cache in Diffusion Transformer Acceleration View paper
- [22] Compute Only 16 Tokens in One Timestep: Accelerating Diffusion Transformers with Cluster-Driven Feature Caching View paper
- [23] Adaptive caching for faster video generation with diffusion transformers View paper
- [24] Fast and memory-efficient video diffusion using streamlined inference View paper
- [25] Freqca: Accelerating diffusion models via frequency-aware caching View paper
- [26] From reusing to forecasting: Accelerating diffusion models with taylorseers View paper
- [27] Blockdance: Reuse structurally similar spatio-temporal features to accelerate diffusion transformers View paper
- [28] Forecasting When to Forecast: Accelerating Diffusion Models with Confidence-Gated Taylor View paper
- [29] Frdiff: Feature reuse for universal training-free acceleration of diffusion models View paper
- [30] Fastercache: Training-free video diffusion model acceleration with high quality View paper
- [31] Foresight: Adaptive Layer Reuse for Accelerated and High-Quality Text-to-Video Generation View paper
- [32] MagCache: Fast Video Generation with Magnitude-Aware Cache View paper
- [33] FlashDLM: Accelerating Diffusion Language Model Inference via Efficient KV Caching and Guided Diffusion View paper
- [34] dLLM-Cache: Accelerating Diffusion Large Language Models with Adaptive Caching View paper
- [35] OmniCache: A Trajectory-Oriented Global Perspective on Training-Free Cache Reuse for Diffusion Transformer Models View paper
- [36] Model Reveals What to Cache: Profiling-Based Feature Reuse for Video Diffusion Models View paper
- [37] Smoothcache: A universal inference acceleration technique for diffusion transformers View paper
- [38] Accelerating Vision Diffusion Transformers with Skip Branches View paper
- [39] HiCache: Training-free Acceleration of Diffusion Models via Hermite Polynomial-based Feature Caching View paper
- [40] Rethinking Token-wise Feature Caching: Accelerating Diffusion Transformers with Dual Feature Caching View paper
- [41] DDT: Decoupled Diffusion Transformer View paper
- [42] ACDiT: Interpolating Autoregressive Conditional Modeling and Diffusion Transformer View paper
- [43] Token Pruning for Caching Better: 9 Times Acceleration on Stable Diffusion for Free View paper
- [44] LazyDiT: Lazy Learning for the Acceleration of Diffusion Transformers View paper
- [45] DiTVR: Zero-Shot Diffusion Transformer for Video Restoration View paper

- [46] AB-Cache: Training-Free Acceleration of Diffusion Models via Adams-Bashforth Cached Feature Reuse View paper
- [47] Multivariate Diffusion Transformer with Decoupled Attention for High-Fidelity Mask-Text Collaborative Facial Generation View paper
- [48] CELAT-DiffNet: channel-enhanced local attention transformer for underwater image enhancement based on diffusion models View paper
- [49] HarmoniCa: Harmonizing Training and Inference for Better Feature Caching in Diffusion Transformer Acceleration View paper
- [50] H2-Cache: A Novel Hierarchical Dual-Stage Cache for High-Performance Acceleration of Generative Diffusion Models View paper
- [51] Deepcache: Accelerating diffusion models for free View paper
- [52] Cachequant: Comprehensively accelerated diffusion models View paper
- [53] Blended latent diffusion View paper
- [54] Timestep Embedding Tells: It's Time to Cache for Video Diffusion Model View paper
- [55] Approximate caching for efficiently serving {Text-to-Image} diffusion models View paper
- [56] Sortblock: Similarity-Aware Feature Reuse for Diffusion Model View paper
- [57] MPQ-DM: Mixed Precision Quantization for Extremely Low Bit Diffusion Models View paper
- [58] -DiT: A Training-Free Acceleration Method Tailored for Diffusion Transformers View paper
- [59] Training-free and hardware-friendly acceleration for diffusion models via similarity-based token pruning View paper
- [60] Attention-driven training-free efficiency enhancement of diffusion models View paper
- [61] Ac3d: Analyzing and improving 3d camera control in video diffusion transformers View paper
- [62] dkv-cache: The cache for diffusion language models View paper
- [63] Block-wise Adaptive Caching for Accelerating Diffusion Policy View paper
- [64] Dynamic diffusion transformer View paper
- [65] Emergent Temporal Correspondences from Video Diffusion Transformers View paper
- [66] A survey on diffusion models for time series and spatio-temporal data View paper
- [67] Post-Training Quantization for Diffusion Transformer via Hierarchical Timestep Grouping View paper
- [68] Precipitation Nowcasting Using Diffusion Transformer With Causal Attention View paper
- [69] Forecast then calibrate: Feature caching as ode for efficient diffusion transformers View paper
- [70] Diffusion models for intelligent transportation systems: A survey View paper
- [71] Diffusion Transformers for Tabular Data Time Series Generation View paper
- [72] Spatio-Temporal Probabilistic Forecasting of Wind Speed Using Transformer-Based Diffusion Models View paper
- [73] Transformer-Based spatiotemporal graph diffusion convolution network for traffic flow forecasting View paper