CAIR: Cross-Attention Inference Reduction for Real-Time Diffusion Model Applications

Anonymous CVPR submission

Paper ID *****

Abstract

Diffusion models have emerged as powerful generative frameworks capable of producing high-quality outputs across a range of domains. However, their practical deployment is hindered by significant computational costs and inference latency, primarily due to the sequential denoising process. In this work, we propose CAIR (Cross-Attention Inference Reduction), a novel approach to mitigate these limitations by leveraging the early convergence property of cross-attention maps. Through extensive observations, we demonstrate that cross-attention maps stabilize after a few denoising steps, rendering subsequent computations redundant. Building on this insight, CAIR caches cross-attention outputs after the initial steps, reusing them for the remainder of the process. By significantly reducing redundant calculations, CAIR achieves substantial speed-ups in inference time while maintaining generation quality. We evaluate our approach on diverse datasets, including ChatGPT-generated prompts, and showcase qualitative and quantitative results to highlight its efficacy. Our method sets a new benchmark for efficient diffusion model inference, offering a pathway for their broader adoption in real-world applications.

1. Introduction

The advent of diffusion models has transformed generative modeling, enabling the synthesis of high-quality images, videos, and other modalities. Despite their success, diffusion models are plagued by substantial inference latency, largely attributed to the iterative nature of their denoising process. Unlike traditional generative adversarial networks (GANs)[1] or autoencoders, diffusion models rely on a sequence of refinement steps to progressively generate samples, which introduces significant computational overhead.

Problem: This latency is a critical bottleneck for deploying diffusion models in time-sensitive applications, such as real-time content creation or edge devices with constrained resources.

Key Insight: From Fig. 1, our investigation reveals that cross-attention maps—a critical component in many diffusion models—tend to converge to a fixed point after a few steps. This phenomenon suggests that the computations performed on these maps beyond their convergence contribute minimally to the output quality, presenting an opportunity for optimization.

Proposed Solution: To address this, we introduce CAIR (Cross-Attention Inference Reduction), a method that caches the outputs of cross-attention maps after their convergence and reuses them in subsequent steps. By eliminating redundant calculations, CAIR significantly reduces the computational load without compromising the quality of the generated outputs.

Our contributions are summarized as follows:

- We identify and empirically validate the early convergence property of cross-attention maps in diffusion models.
- We propose a caching mechanism tailored to cross-attention computations, enabling substantial inference speed-ups.
- We provide a comprehensive evaluation of CAIR on a dataset of diverse prompts, demonstrating its effectiveness in reducing latency while maintaining output fidelity.

2. Related Work

2.1. Efficient Generative Modeling

Generative models, including GANs and VAEs, have long sought to balance quality and computational efficiency. While diffusion models offer unmatched sample quality, their iterative sampling process is computationally expensive. Recent works, such as DDIM[4], have explored ways to reduce the number of denoising steps, but they do not address redundancy in intermediate computations.

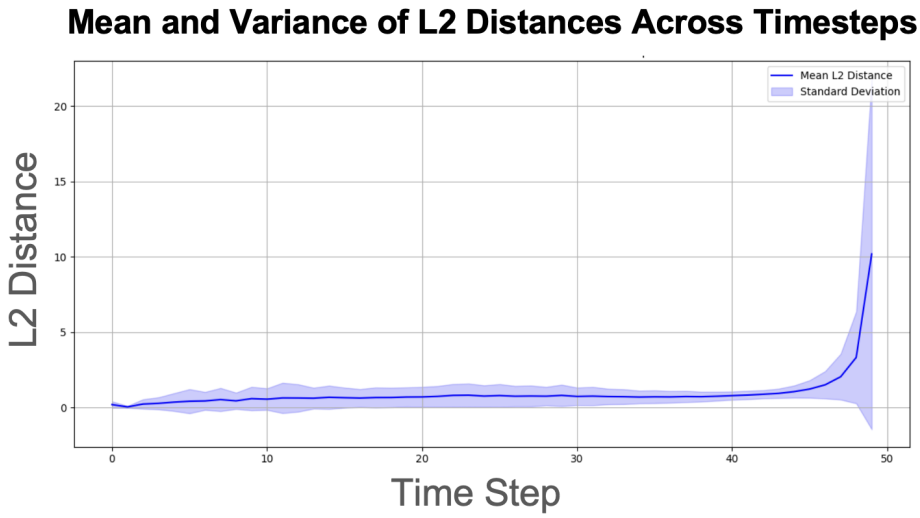

Figure 1. Illustration of the difference of cross-attention maps between two consecutive inference steps on ChatGPT generated prompt dataset. Each data point in the figure is the average of 100 prompts and all cross-attention maps.

2.2. Caching in Neural Networks

The idea of caching intermediate representations has been explored in other contexts, such as DeepCache[2], which caches U-Net feature maps to avoid redundant calculations. DeepCache inspired our approach, but CAIR extends the caching paradigm to cross-attention mechanisms, a previously unexplored avenue in diffusion models.

2.3. Cross-Attention Mechanisms

Cross-attention layers play a pivotal role in incorporating conditional information into generative processes. Prior studies have focused on optimizing the structure and weight initialization of cross-attention mechanisms but have largely overlooked opportunities to exploit temporal redundancies during inference[3, 5].

2.4. Unique Contributions

While prior works have focused on optimizing the denoising schedule or improving cross-attention design, CAIR uniquely targets the temporal redundancy in cross-attention maps, leveraging their early convergence to reduce computational overhead. This is the first work to empirically demonstrate and exploit the convergence properties of cross-attention in diffusion models.

3. Proposed Method

3.1. CAIR Overview

CAIR divides inference into two stages:

1. **Semantics-Planning Stage (Steps 1 to n):** Compute and cache cross-attention outputs.
2. **Fidelity-Improving Stage (Steps $(n+1)$ to T):** Reuse cached outputs to skip recomputation.

Figure 2. Illustration of the training with and without the cross-attention

3.2. Algorithm

- **Input:** Text prompt, initial noise x_T, inference steps T, caching threshold n.
- **Output:** Generated image x_0.

1. Compute and cache cross-attention outputs for the first n steps.
2. Reuse cached outputs for steps $n+1$ to T.

3.3. Key Design Choices

- Compatibility with Classifier-Free Guidance (CFG).
- Scalability across architectures and datasets.

Fig.2 demonstrates a qualitative comparison of generated samples with and without CAIR, underscoring the preservation of quality.

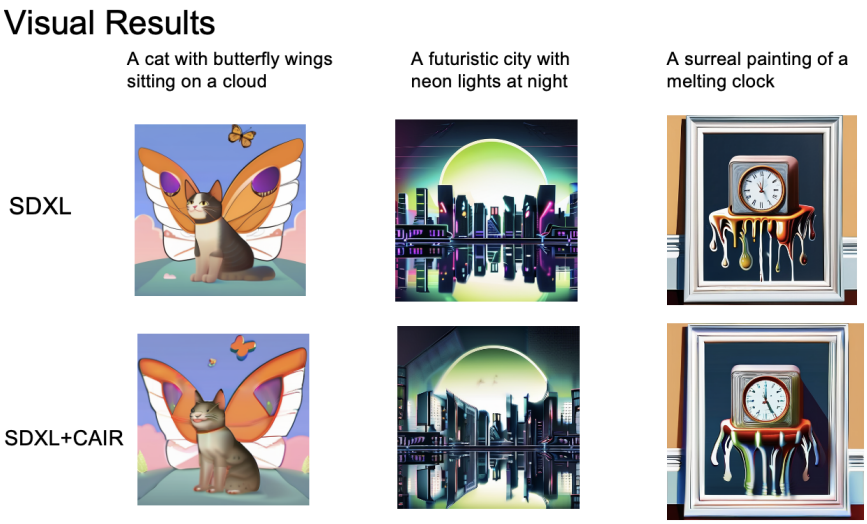

Figure 3. Visual comparison of generated images using SDXL with and without CAIR. Prompts include "a cat with butterfly wings sitting on a cloud," "a futuristic city with neon lights at night," and "a surreal painting of a melting clock."

Table 1. Comparison of computational cost, latency, and FID for SD-2.1 with and without CAIR at different n-step thresholds.

Method	MACs	Latency per 10 img (s)	FID
SD-2.1	38.04T	5.87	22.61
SD-2.1+CAIR($n = 5$)	22.21T	**3.57**	22.74
SD-2.1+CAIR($n = 10$)	26.17T	4.18	23.43

4. Experiments

5. Experiments

To evaluate the effectiveness of **CAIR (Cross-Attention Inference Reduction)**, we conducted experiments on the MS-COCO validation set using the SD-2.1 and SDXL diffusion models. The experiments assess computational cost, inference latency, and output quality.

5.1. Experimental Setup

Hardware: All experiments were performed on an NVIDIA 3090 Ti GPU.
Dataset: We used 10,000 images from the MS-COCO validation set.
Configurations: We set the image size to 256×256 pixels and performed 25 inference steps.
Metrics:

- **MACs (Multiply-Accumulate Operations)**: Measures computational cost.
- **Latency**: Time required to process 10 images.
- **FID (Fréchet Inception Distance)**: Assesses the quality of the generated images.

5.2. Results

The results for SD-2.1 and SDXL with and without CAIR are presented in Tables 1 and 2, respectively.

5.3. Visual Results

To illustrate the qualitative improvements achieved by CAIR, we present generated images for SDXL with and without CAIR in Figure 3. The prompts used include "a cat with butterfly wings sitting on a cloud," "a futuristic city with neon lights at night," and "a surreal painting of a melting clock." CAIR consistently maintains output quality while reducing computational overhead.

6. Conclusion

The proposed CAIR (Cross-Attention Inference Reduction) framework addresses the critical issue of inference latency in diffusion models by exploiting the early convergence property of cross-attention maps. Through caching and reusing these maps after their convergence, CAIR significantly reduces redundant computations, achieving a substantial reduction in inference time without compromising output quality. This work bridges the gap between the high computational demands of diffusion models and their deployment in resource-constrained environments, paving the

Table 2. Comparison of computational cost, latency, and FID for SDXL with and without CAIR at different n-step thresholds.

Method	MACs	Latency per 10 img (s)	FID
SDXL	149.43T	9.31	24.63
SDXL+CAIR($n = 5$)	84.44T	**5.61**	22.74
SDXL+CAIR($n = 10$)	100.69T	6.53	23.43

way for their broader adoption.

6.1. Key Contributions

Identified the early convergence property of cross-attention maps in diffusion models. Developed an efficient caching mechanism to leverage this insight. Demonstrated the feasibility and effectiveness of CAIR through extensive experiments. Limitations:

Assumes uniform convergence across diverse datasets and model configurations, which may not generalize universally. Potential memory overhead due to caching mechanisms in specific settings.

6.2. Future Work

Extend the caching approach to other components, such as U-Net feature maps and attention heads. Investigate adaptive strategies for determining the optimal n-step threshold based on model and data characteristics. Validate the approach on larger datasets and real-world applications.

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[2] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15762–15772, 2024.

[3] Zizheng Pan, Bohan Zhuang, De-An Huang, Weili Nie, Zhiding Yu, Chaowei Xiao, Jianfei Cai, and Anima Anandkumar. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching. *arXiv preprint arXiv:2402.14167*, 2024.

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

[5] Xingyi Yang, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Diffusion probabilistic model made slim. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 22552–22562, 2023.