JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 10, pp. 2981-2989, Oct. 2026

Check for updates

DiT의 Attention을 Multi-Scale Retention으로 대체한 DiR 모델의 이미지 생성 성능 연구

이 중 현[1] · 최 웅[2*]
[1]강남대학교 ICT융합공학부 학사과정 [2]강남대학교 컴퓨터공학부 교수

Study on the Image-Generation Performance of Diffusion with Retention: Replacing Self-Attention in Diffusion Transformers with Multi-Scale Retention

Jung-Hyeon Lee[1] · Woong Choi[2*]
[1]Bachelor's Degree, College of ICT Convergence Engineering, Kangnam University, Yongin 16979, Korea
[2]Professor, Department of Computer Engineering, Kangnam University, Yongin 16979, Korea

[요 약]

본 논문은 Diffusion Transformer(DiT)의 셀프 어텐션(Self-Attention) 모듈을 Multi-Scale Retention(MSR)으로 대체한 DiR(Diffusion with Retention) 모델을 제안한다. 기존 DiT는 $O(N^2)$ 연산 복잡도로 인해 고해상도에서 메모리 및 연산 효율이 급격히 저하된다. 이를 해결하기 위해, 청크 내부에는 정확한 $\gamma^{|i-j|}$ 감쇠를 적용하고 청크 간에는 KV 상태를 순방향·역방향으로 누적하는 양방향 청크 단위 MSR(Bidirectional Chunkwise MSR)을 도입하여 O(N) 메모리 복잡도를 달성하였다. CelebA-HQ 256 데이터셋을 이용한 학습 실험에서, DiR-B/2 모델은 약 200K 스텝 학습 후 FID 30.06을 기록하였다. 이는 동일 조건에서 학습된 DiT-B/2의 FID 26.92 대비 약 3.14점 높은 수준이다. 반면 처리 속도 측면에서는 1024×1024·배치 16 조건에서 FlashAttention을 적용한 DiT 대비 처리 시간이 약 40% 단축되었으며, O(N) 수준의 안정적인 메모리 사용량을 유지하였다. 본 연구는 성능 우위가 아닌 구조적 대체 가능성 검증에 목적이 있으며, 결과는 MSR이 확산 기반 이미지 생성에서 실용적인 어텐션 대안이 될 수 있음을 시사한다.

[Abstract]

A diffusion transformer (DiT) suffers from $O(N^2)$ complexity in self-attention, which severely limits scalability at high resolutions. To address this, we propose diffusion with retention (DiR), which replaces self-attention with multi-scale retention (MSR). DiR introduces bidirectional chunkwise MSR: It applies an exact $\gamma^{|i-j|}$ decay within each chunk and accumulates key-value states across chunks in both forward and backward directions. This design achieves O(N) memory complexity. On CelebA-HQ 256, DiR-B/2 reaches a Fréchet inception distance (FID) of 30.06 at approximately 200,000 training steps, compared to FID of 26.92 for DiT-B/2 under identical conditions—a gap of about 3.14 points. In terms of speed, DiR-B/2 reduces processing time by approximately 40% compared with FlashAttention-enabled DiT at 1024×1024 resolution with a batch size of 16, while maintaining stable O(N) memory usage. Rather than aiming to outperform DiT, this work verifies the structural feasibility of a linear-complexity attention alternative, suggesting that MSR is a practical option for diffusion-based image generation.

색인어 : 디퓨전 트랜스포머, 멀티스케일 리텐션, 이미지 생성, 선형 복잡도, RetNet
Keyword : Diffusion Transformer, Multi-Scale Retention, Image Generation, Linear Complexity, RetNet

http://dx.doi.org/10.9728/dcs.2026.27.10.2981

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 25 June 2026; **Revised** 10 August 2026
Accepted 14 September 2026

***Corresponding Author; Woong Choi**
Tel: +82-31-280-3753
E-mail: wchoi.pine@gmail.com

Ⅰ. 서 론

이미지 생성 분야에서 Denoising Diffusion Probabilistic Model(DDPM)[1]과 Latent Diffusion Model(LDM)[2]이 주목받은 이후, Peebles 등이 제안한 Diffusion Transformer(DiT)[3]는 U-Net 기반 디노이저를 순수 Transformer 아키텍처로 대체하여 확산 모델의 스케일러빌리티와 이미지 품질을 크게 향상시켰다. 그러나 DiT는 셀프 어텐션의 $O(N^2)$ 복잡도로 인해 고해상도에서 메모리와 연산 비용이 급격히 증가한다.

구체적으로, 512×512 이미지를 패치 크기 2로 처리하면 시퀀스 길이 N=1,024가 되며, 1024×1024에서는 N=4,096으로 늘어난다. FlashAttention[4]이 I/O 최적화를 통해 실용적 해결책을 제공하지만, 이는 이차 복잡도 자체를 극복하지 못한다. 본 연구는 이 문제의 해결책으로, RetNet[5]의 Multi-Scale Retention(MSR)을 이미지 생성을 위한 확산 모델에 적용한 DiR(Diffusion with Retention)을 제안한다.

이미지 패치는 전후 모든 방향의 공간 문맥을 함께 활용할 때 유리하므로, 본 연구에서는 자기회귀 언어모델링을 위해 설계된 단방향(Causal) RetNet 원형을 양방향으로 확장하는 접근을 채택한다. 본 연구의 DiR 아키텍처는 청크 단위 양방향 KV 상태 누적을 통해 O(N) 메모리로 이를 실현한다.

이에 본 논문의 주요 기여는 다음 세 가지로 요약된다. 첫째, 거리 기반 양방향 감쇠 행렬을 새롭게 정의하여, RetNet의 단방향 MSR을 이미지 패치 간 무방향성 공간 관계를 반영하도록 확장하였다. 둘째, 청크 내부의 정확한 양방향 감쇠 연산과 청크 간 순방향·역방향 상태 누적(수식 4-10)을 결합한 양방향 청크 단위 스캔 알고리즘을 설계함으로써, 전역 문맥을 유지하면서도 O(N) 메모리 복잡도를 달성하였다. 이는 RetNet의 단방향 청크 단위 스캔을 그대로 차용한 것이 아니라, 양방향 정보를 청크 내부에서는 근사 오차 없이, 청크 간에는 극히 낮은 오차(평균절대오차 0.01 이하)로 처리하도록 새롭게 유도된 결과이다. 셋째, DiT 파이프라인에 제안 구조를 통합한 DiR 모델을 구현하고, CelebA-HQ 256 데이터셋 학습 실험을 통해 셀프 어텐션 없이도 안정적으로 학습되고 합리적인 수준의 이미지를 생성할 수 있음을 확인하였다.

Ⅱ. 관련 연구

2-1 Diffusion Transformer(DiT)

DiT[3]는 DDPM[1]의 U-Net 디노이저를 Vision Transformer(ViT)[6] 구조로 대체한 모델이다. 입력 이미지를 잠재 공간의 패치로 분할하고 Transformer 블록을 통해 디노이징하며, adaLN-Zero 기법으로 타임스텝과 클래스 조건을 각 레이어에 주입한다. 모델 규모와 이미지 품질 간의 강한 스케일링 법칙을 입증하였으나, 고해상도에서의 이차 복잡도 문제가 여전히 과제로 남아 있다.

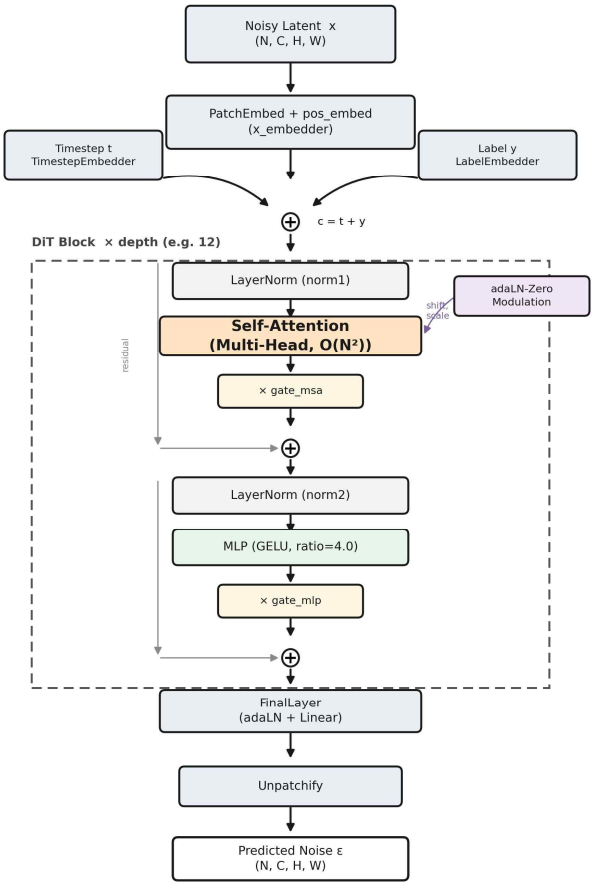


그림 1. DiT 블록 구조(Self-Attention 기반)
Fig. 1. DiT block architecture (self-attention based)

2-2 Retentive Network(RetNet)

RetNet은 병렬·청크·순환의 세 가지 동등한 연산 형태를 지원하는 MSR을 핵심 연산으로 사용한다[5]. 각 헤드 h에서 감쇠 계수 $\gamma h = 1 - 2^{(-5-h)}$를 적용하며, 멀티헤드 구조로 다양한 스케일의 거리 의존성을 동시에 학습한다. 청크 단위 연산은 O(N) 메모리로 전역 문맥을 처리할 수 있어 긴 시퀀스에 실용적이다. 단, RetNet은 언어 모델링을 위한 단방향 구조로 설계되었으므로 이미지 생성에 그대로 적용하기 위해서는 양방향 확장이 필수적이다.

2-3 선형 복잡도 기반 이미지 생성 모델

셀프 어텐션의 이차 복잡도를 완화하기 위한 시도는 RetNet 이전부터 다양하게 이루어졌다. Katharopoulos 등이 제안한 Linear Transformer는 어텐션을 커널 특징 맵의 선형 내적으로 재정의하고 행렬 곱의 결합 법칙을 활용하여 $O(N^2)$ 연산을 $O(N)$으로 낮추었으며, 이를 통해 셀프 어텐션과 순환 신경망(RNN) 사이의 수학적 동등성을 보였다[7]. 다만 해당 연구는 자기회귀(Autoregressive) 시퀀스 모델링을 가정한 단방향 구조이므로, 양방향 공간 문맥이 필요한 이

미지 패치 처리에는 그대로 적용하기 어렵다는 한계가 있다.

Mamba 기반 확산 모델인 DiS와 ZigMa는 Gu와 Dao가 제안한 선택적 상태 공간 모델(Selective State Space Model, SSM)[8]을 이미지 생성에 적용하여 선형 복잡도와 경쟁력 있는 성능을 동시에 달성하였다[9],[10]. 그러나 SSM 계열 모델 역시 기본적으로 단방향 순환 구조이므로, 이미지 패치가 요구하는 양방향 공간 문맥 처리를 위해 별도의 설계가 필요하다. DiR은 RetNet의 감쇠 기반 근사를 청크 단위 양방향 방식으로 확장함으로써, Linear Transformer 및 Mamba 계열과 구별되는 접근법을 취한다.

Ⅲ. 제안 방법: DiR

3-1 전체 아키텍처

DiR은 DiT의 전체 파이프라인(PatchEmbed → Transformer Blocks → FinalLayer → Unpatchify)을 유지하되, 각 블록의 셀프 어텐션을 Multi-Scale Retention으로 대체한다. adaLN-Zero 모듈은 DiT와 동일하게 유지하여 타임스텝 임베딩 t와 클래스 레이블 임베딩 y의 합 $c=t+y$를 각 DiRBlock에 주입한다. DiRBlock의 순전파는 다음 수식 (1)과 같다.

$$x \leftarrow x + gate(msr) \cdot MSR(AdaLN(x, shift_{msr}, scale_{msr}))$$
$$x \leftarrow x + gate(mlp) \cdot MLP(AdaLN(x, shift_{mlp}, scale_{mlp})) \quad (1)$$

여기서 shift, scale, gate 6개의 파라미터는 adaLN_modulation 레이어로부터 조건 벡터 c를 통해 생성된다.

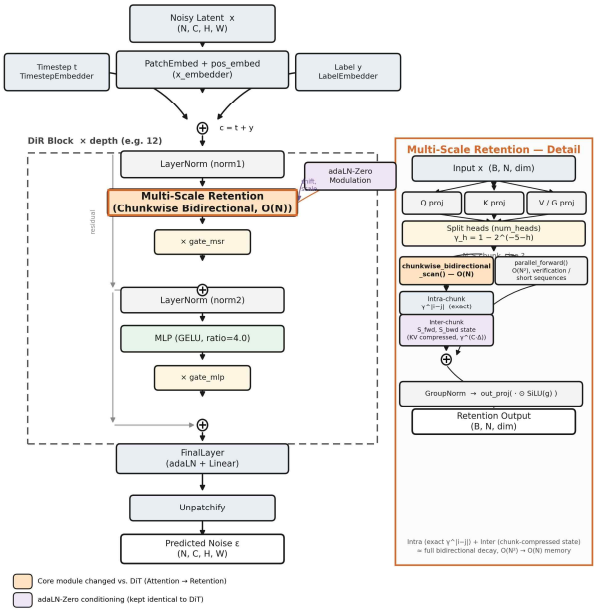


그림 2. DiR 블록 구조(Multi-Scale Retention 기반)
Fig. 2. DiR block architecture (multi-scale retention based)

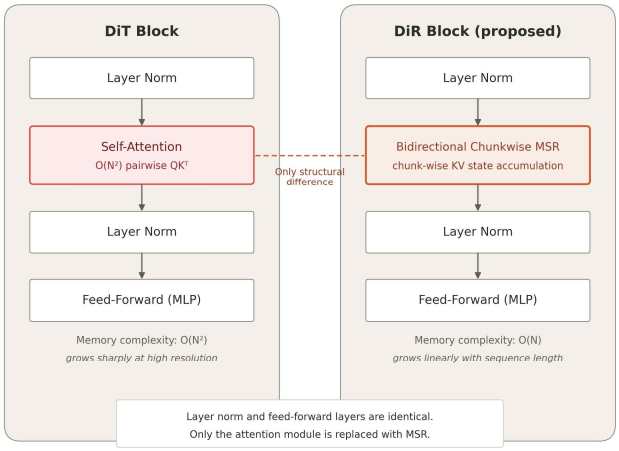


그림 3. DiT 블록과 DiR 블록 구조 비교
Fig. 3. Comparison of DiT block and DiR block structures

3-2 양방향 Multi-Scale Retention

표준 RetNet의 MSR은 단방향 인과(Causal) 감쇠 마스크를 사용한다. 이미지 패치는 방향성이 없는 2D 공간 구조를 가지므로, 본 연구는 위치 간 거리에 기반한 양방향 감쇠 행렬을 정의한다.

$$D[i,j] = \gamma^{|i-j|}, \gamma \in (0,1) \quad (2)$$

헤드 h의 감쇠 계수는 $\gamma h = 1 - 2^{(-5-h)}$이며, 멀티헤드 구조로 다양한 거리 스케일을 동시에 인코딩한다. 병렬 모드(Parallel Forward)에서의 출력은 다음 수식 (3)과 같다.

$$O = (Q \cdot K^T \odot D) \cdot V \cdot D^{-\frac{1}{2}} \quad (3)$$

이 병렬 경로는 $O(N^2)$ 연산으로 정확하며, 소규모 시퀀스 검증 및 ablation 실험에 사용된다.

3-3 청크 단위 양방향 스캔(O(N) 경로)

고해상도 이미지 생성 환경에서 선형 연산 및 메모리 복잡도를 달성하기 위해, 본 연구에서는 입력 패치 시퀀스를 고정된 청크 크기 C의 블록 단위로 분할하여 청크 내부(Intra-Chunk) 연산과 청크 간(Inter-Chunk) 연산으로 두 단계로 분리하여 수행한다. 먼저, 동일 청크 내에 존재하는 패치 쌍 간의 조밀한 기하학적 관계를 포착하기 위해 청크 내부 연산을 처리한다. 해당 구역 내의 모든 토큰에 대해서는 수식 (4) 및 수식 (5)와 같이 정확한 양방향 거리 기반 감쇠를 적용한다.

$$O(intra)[n] = (Q(n) \cdot K(n)^T \odot D(intra)) \cdot V(n) \quad (4)$$

$$D(intra)[i,j]=\gamma^{|i-j|}(i,j\in chunkn) \quad (5)$$

여기서 $Q(n),K(n),V(n)$은 n번째 청크에 해당하는 투영 행렬이며, $\odot$은 요소별 곱셈(Hadamard product)을 나타낸다. 이를 통해 국소적 영역 내에서는 마스킹에 의한 정보 손실 없이 정확한 기하학적 전방향 참조가 가능해진다. 다음으로, 청크 범위를 벗어나는 장거리 공간 의존성을 전역적으로 파악하면서도 메모리 오버헤드를 최소화하기 위해 양방향 상태 누적 기반의 청크 간 연산을 수행한다. 이를 위해 각 청크의 키(Key)와 밸류(Value) 정보는 수식 (6)과 같이 고정된 크기의 상태 행렬(State Matrix)로 압축된다.

$$KV(n)=K(n)^T\cdot V(n)\in\mathbb{R}^{(D\times D)} \quad (6)$$

압축된 $D\times D$크기의 청크별 상태 정보는 수식 (7) 및 수식 (8)에 정의된 바와 같이 순방향(Forward) 및 역방향(Backward)으로 각각 스캔되며, 청크 간 거리 감쇠 인자인 γ^C를 통해 공간적 거리에 비례하여 순차적으로 누적된다.

$$S(fwd)[n]=\gamma^C\cdot(S(fwd)[n-1]+KV(n-1)),S(fwd)[0]=0 \quad (7)$$

$$S(bwd)[n]=\gamma^C\cdot(S(bwd)[n+1]+KV(n+1)),S(bwd)[nc-1]=0 \quad (8)$$

여기서 n_c는 이미지 전체 시퀀스에서 도출된 총 청크의 개수를 의미한다. 최종적인 청크 간 출력항은 순방향과 역방향 상태 정보를 결합한 뒤, 쿼리 행렬 $Q(n)$과 헤드 차원 수(D)의 스케일링 인자를 적용하여 수식 (9)와 같이 산출된다.

$$O(inter)[n]=Q(n)\cdot(S(fwd)[n]+S(bwd)[n])\cdot D^{-\frac{1}{2}} \quad (9)$$

최종적인 Retention 레이어의 출력 $O[n]$은 수식 (10)과 같이 청크 내부 및 청크 간 연산 결과의 선형 합산으로 완성된다.

$$O[n]=O(intra)[n]+O(inter)[n] \quad (10)$$

결과적으로 각 청크의 전역 공간 정보가 청크 당 하나의 $D\times D$ 상태 행렬로 압축되므로, 모델의 전체 메모리 복잡도는 기존 DiT의 $O(N^2)$에서 $O(N\cdot(C+D^2/C))$ 수준으로 획기적으로 감소한다. 이는 C가 고정된 값일 때 시퀀스 길이 N에 대해 선형 복잡도 $O(N)$이 됨을 의미한다. 본 연구의 이미지 생성 실험에서는 수렴 최적화를 고려하여 청크 크기를 $C=256$으로 설정하여 기본 연산을 수행하였다.

청크 간 연산에서 각 청크의 Key·Value를 하나의 상태 행렬로 압축할 수 있는 이유는, 표준 Self-Attention의 Softmax(QKᵀ)V 연산을 커널화된 선형 어텐션 형태로 재해석하면 어텐션 출력이 순서에 무관하게 KᵀV의 누적합으로 표현 가능하기 때문이다. 즉, 원래 Attention이 모든 (query, key) 쌍을 개별적으로 비교하는 것과 달리, 본 연구의 상태 누적 방식은 청크 단위로 key-value 쌍의 가중합을 미리 축적해 두고, query 시점에 이 축적된 상태만을 참조한다. 이 과정에서 정보 손실이 발생하는 지점은 두 가지이다. 첫째, 청크 경계를 넘는 원소 쌍에 대해서는 정확한 |i−j| 기반 감쇠 대신 청크 단위로 이산화된 감쇠(discretized decay)가 적용되므로, 청크 내부 원소 간 정밀한 상대 거리 정보가 근사된다. 둘째, 상태 행렬의 크기가 시퀀스 길이와 무관하게 고정되어 있어(청크당 1개), 매우 긴 시퀀스에서는 여러 청크의 정보가 같은 크기의 상태로 압축되며 이 과정에서 정보 압축에 따른 손실이 누적될 수 있다. 이러한 근사가 실제 정확도에 미치는 영향은 본 절 말미에 제시한 오차 분석(청크 내부 오차 0, 청크 간 최대 절대오차 0.04 미만)을 통해 정량적으로 확인하였다.

```
Input: Q, K, V ∈ R^(N×d), chunk size C
1: Split Q, K, V into chunks of size C → {Q_m, K_m, V_m}, m = 1..N/C
2: for each chunk m:
3:     Compute intra-chunk output (Eq. 4–5) with exact γ^|i−j| decay
4:     Compute chunk-level KV state S_m (Eq. 6)
5: Forward pass:  S_m^fwd = γ^C · S_(m-1)^fwd + S_m   (Eq. 7)
6: Backward pass: S_m^bwd = γ^C · S_(m+1)^bwd + S_m   (Eq. 8)
7: for each chunk m:
8:     Combine intra-chunk output with S_m^fwd, S_m^bwd (Eq. 9)
9: Output = concat over all chunks + final projection (Eq. 10)
```

그림 4. Bidirectional Chunkwise MSR 연산 절차
Fig. 4. Procedure of Bidirectional Chunkwise MSR

앞선 과정을 통해 산출된 최종 Retention 연산 결과는 GroupNorm[11] 정규화를 거친 후, 비선형 활성화 함수인 SiLU 기반의 게이트 벡터와 요소별 곱셈(Hadamard product)을 통해 최종 출력으로 변환된다. 이를 수식으로 나타내면 수식 (11)과 같다.

$$Output=W(out)\cdot(GroupNorm(Retention(x))\odot SiLU(W(g)\cdot x)) \quad (11)$$

여기서 W_{out}과 W_g는 각각 출력 및 게이트 투영을 위한 가중치 행렬이다. 한편, 제안된 청크 단위 양방향 감쇠 근사 기법의 수학적 타당성을 검증하기 위해 단일 헤드($D=1$, $N=12$, $C=4$, $\gamma=1-2^{-5}$) 환경에서 사전 오차 분석을 수행하였다. 그 결과, 청크 내부 원소의 산출 오차는 0으로 나타나 마스킹 손실 없이 정확한 감쇠가 적용됨을 확인하였다. 또한, 상태 누적이 적용된 청크 간 원소의 경우에도 최대 절대 오차가 0.04 미만, 평균 절대 오차가 0.01 이하로 측정되어 본 연구에서 제안한 $O(N)$ 근사 기법이 이미지 생성 도메인에서 실용적이고 높은 수치적 신뢰성을 확보하고 있음을 입증하였다.

한편, 청크 단위 양방향 스캔 대신 수식 (2)-(3)의 양방향 감쇠 행렬을 청크 분할 없이 전체 시퀀스에 그대로 적용하는 단순 적용 방식(Naive Full Bidirectional MSR) 역시 원리

적으로는 가능하다. 그러나 이 방식은 모든 패치 쌍의 관계를 $O(N^2)$ 복잡도로 직접 계산해야 하므로, Self-Attention을 MSR로 단순 대체하는 것만으로는 DiT가 겪는 이차 복잡도 문제 자체를 해결하지 못하며, 고해상도·대배치 조건에서는 표 2의 FlashAttention 미적용 DiT와 동일하게 메모리 사용량이 급증하여 OOM에 이를 것으로 예상된다. 즉, 단순 적용 방식은 본 연구가 목표로 하는 효율성 문제에 대한 해결책이 되지 못한다. 반면 생성 품질 측면에서 두 방식이 얼마나 근접한 결과를 내는지는, 위에서 제시한 오차 분석(청크 내부 오차 0, 청크 간 평균절대오차 0.01 이하)을 통해 사실상 수치적으로 거의 동일함이 확인되었다. 다만 이 검증은 소규모 시퀀스에서의 수치적 오차 분석에 기반한 것으로, 실제 CelebA-HQ 학습·생성 실험에서 단순 적용 방식과 제안 방식 간 FID/IS 및 처리 속도를 직접 비교하는 것은 단일 GPU(RTX 4070 Ti Super, 16GB) 환경의 메모리 제약으로 인해 수행하지 못하였다.

Ⅳ. 실험 및 결과

4-1 실험 환경 및 학습 설정

본 연구에서는 제안 모델의 이미지 생성 성능과 연산 효율성을 객관적으로 비교하기 위해, 기준 모델인 DiT-B/2와 제안 모델인 DiR-B/2를 동일한 파라미터 규모로 구성하여 실험하였다. 두 모델은 모두 레이어 깊이(Depth) 12, 은닉 차원(Hidden Dimension) 768, 어텐션 헤드(Heads) 수 12, 패치 크기 2로 동일하게 설정하여 공정한 비교 조건을 갖추었다. 제안 모델인 DiR-B/2에는 선형 복잡도 달성을 위한 청크 단위 스캔의 기준 크기(C)를 256으로 고정하여 적용하였다.

모든 학습 및 벤치마크 실험은 NVIDIA GeForce RTX 4070 Ti Super(16GB VRAM) 1대로 구성된 단일 GPU 환경에서 수행하였다. 학습 데이터셋은 약 30,000장의 인물 이미지로 구성된 CelebA-HQ 256 데이터셋[12]을 사용하였으며, 원본 CelebA[13] 이미지를 기반으로 Progressive GAN 방식으로 고해상도화된 데이터를 256×256 해상도로 크기 조정(Resize)한 후 중앙 자르기(Center Crop)와 무작위 수평 뒤집기(Random Horizontal Flip)를 적용하고 [0.5, 0.5, 0.5] 기준으로 정규화하였다.

모델 최적화에는 Loshchilov와 Hutter가 제안한 AdamW[14] 옵티마이저를 사용하였으며, 학습률(Learning Rate) 1×10^{-4}, 가중치 감쇠(Weight Decay) 0.01을 적용하였다. 배치 크기는 4로 설정하였고, 감쇠율 α=0.9999의 지수 이동 평균(EMA)을 통해 평가용 가중치의 안정성을 확보하였다. 이미지는 사전 학습된 VAE 인코더(stabilityai/sd-vae-ft-mse)를 통해 잠재 공간으로 변환한 뒤, Nichol과 Dhariwal이 개선한 디퓨전 학습 기법[15]을 기반으로 1,000 타임스텝의 디노이징 과정을 학습하도록 설계하여 약 200,000(200K) 스텝까지 학습을 진행하였다.

성능 평가는 생성된 이미지 10,000장과 실제 이미지 분포 간의 거리를 측정하는 FID(Fréchet Inception Distance)[16]와 IS(Inception Score)[17]를 지표로 사용하였으며, 연산 효율성은 추론 시간(s)과 피크 메모리(MB) 사용량을 벤치마크하여 평가하였다. 표 1은 본 실험에 사용된 GPU 환경 및 주요 학습 하이퍼파라미터를 정리한 것이다.

표 1. 실험 환경 및 학습 하이퍼파라미터
Table 1. Experimental environment and training hyperparameters

Item	Configuration
GPU	NVIDIA RTX 4070 Ti Super (16GB VRAM)
Dataset	CelebA-HQ 256 (approx. 30,000 images)
Image Resolution	256 × 256
Model Configuration	Depth 12 / Hidden 768 / Heads 12 / Patch 2
Chunk Size (DiR, C)	256
Batch Size	4
Optimizer	AdamW
Learning Rate	1×10^{-4}
EMA Decay Rate	0.9999
VAE	stabilityai/sd-vae-ft-mse
Diffusion Timesteps	DDPM 1,000 steps
Total Training Steps	200,000 (200K)
Evaluation Metrics	FID, IS (based on 10,000 generated images)

4-2 메모리 효율성 분석

DiR-B/2는 셀프 어텐션을 Multi-Scale Retention으로 대체하면서 청크 단위 양방향 스캔을 통해 어텐션 연산의 메모리 복잡도를 $O(N^2)$에서 $O(N)$으로 낮추도록 설계되었다. 이러한 구조적 이점을 정량적으로 확인하기 위해, 해상도(512×512, 1024×1024)와 배치 크기를 달리하며 DiT-B/2와 DiR-B/2의 처리 시간 및 Peak VRAM 사용량을 비교하는 3-way 벤치마크를 수행하였다. DiT-B/2는 FlashAttention 적용 여부(ON/OFF)에 따라 구분하여 측정함으로써, 본 연구의 청크 단위 양방향 스캔이 FlashAttention과 같은 별도의 커널 최적화 없이도 어느 정도의 효율성을 확보하는지 함께 확인하였다.

측정 결과, 비교적 작은 해상도인 512×512(B=32)에서는 FlashAttention이 활성화된 DiT가 0.4203초로 가장 빠른 처리 시간을 보였으며, DiR-B/2는 0.4576초로 근소하게 뒤따르는 수준이었다. 다만 이 구간에서도 FlashAttention이 비활성화된 DiT는 처리 시간 0.5831초, Peak VRAM 6,084.5MB로 DiR-B/2(1,980.5MB) 대비 약 3배에 달하는 메모리를 사용하여, MSR의 청크 단위 연산이

표 2. 해상도/배치별 처리 시간 및 메모리 사용량 비교
Table 2. Comparison of processing time and memory usage by resolution and batch size

Resolution (Batch Size)	Model Type	Processing Time (s)	Peak VRAM (MB)
512×512 (B=32)	DiT (FlashAttention ON)	0.4203	1,573.1
	DiT (FlashAttention OFF)	0.5831	6,084.5
	DiR (Ours)	0.4576	1,980.5
1024×1024 (B=8)	DiT (FlashAttention ON)	0.7642	1,580.5
	DiT (FlashAttention OFF)	15.8169	21,068.8
	DiR (Ours)	0.4958	1,983.6
1024×1024 (B=16)	DiT (FlashAttention ON)	1.4999	2,638.7
	DiT (FlashAttention OFF)	OOM	OOM
	DiR (Ours)	0.9028	3,399.1

FlashAttention과 유사한 수준의 메모리 절감 효과를 제공함을 확인할 수 있었다.

해상도가 1024×1024로 증가하면서 격차는 더욱 뚜렷해졌다. 배치 8 조건에서 FlashAttention이 비활성화된 DiT는 처리 시간이 15.8169초로 급격히 증가하고 Peak VRAM도 21,068.8MB까지 치솟은 반면, DiR-B/2는 0.4958초, 1,983.6MB로 거의 동일한 수준을 유지하였다. 특히 주목할 점은 이 구간에서 DiR-B/2가 FlashAttention이 활성화된 DiT(0.7642초)보다도 더 빠른 처리 시간을 보였다는 것으로, 시퀀스 길이가 길어질수록 O(N) 복잡도의 이점이 본격적으로 드러나는 것으로 해석된다.

배치를 16으로 늘린 1024×1024 조건에서는 FlashAttention이 비활성화된 DiT가 16GB VRAM 한도를 초과하여 OOM(Out-Of-Memory)으로 실행 자체가 불가능하였다. 반면 DiR-B/2는 동일 조건에서 0.9028초, 3,399.1MB로 안정적으로 동작하였다. FlashAttention이 활성화된 DiT(1.4999초, 2,638.7MB)와 비교하면 DiR-B/2의 처리 시간은 약 40% 단축되었으나, Peak VRAM은 오히려 약 29% 증가하였다. 이는 DiR의 청크 단위 양방향 스캔이 FlashAttention 미적용 DiT 대비로는 처리 속도와 메모리 효율을 모두 크게 개선하지만, FlashAttention이 적용된 DiT와 비교할 경우에는 처리 속도 향상과 메모리 사용량 증가가 공존하는 trade-off 관계에 있음을 보여준다.

4-3 생성 이미지 품질 평가(FID / IS)

제안 모델의 생성 품질을 정량적으로 평가하기 위해, 각 모델에서 생성한 이미지 10,000장과 CelebA-HQ 256 학습 데이터 30,000장을 대상으로 FID[16]와 IS[17]를 측정하였다. 두 지표 모두 동일한 평가 파이프라인(torch-fidelity)과 동일한 샘플링 스텝(250 steps, DDPM) 조건 하에서 측정하였다.

표 3. DiT-B/2 vs DiR-B/2 FID / IS 비교
Table 3. Comparison of FID and IS between DiT-B/2 and DiR-B/2

Model	FID ↓	IS ↑	Training Steps
DiT-B/2 (Baseline)	26.92	2.55 ± 0.05	200K
DiR-B/2 (Proposed)	30.06	2.54 ± 0.05	200K

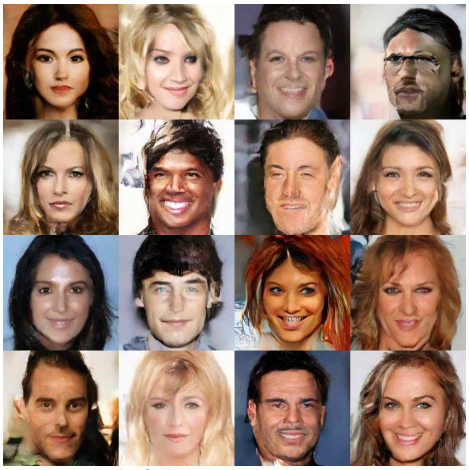


그림 5. DiT-B/2 생성 이미지(CelebA-HQ 256, 4×4 샘플)
Fig. 5. Generated samples from DiT-B/2(CelebA-HQ 256, 4×4 grid)

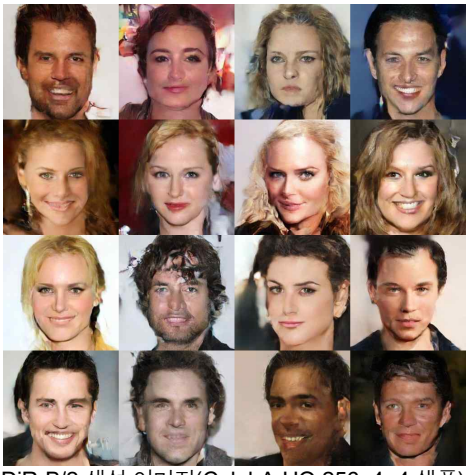


그림 6. DiR-B/2 생성 이미지(CelebA-HQ 256, 4×4 샘플)
Fig. 6. Generated samples from DiR-B/2(CelebA-HQ 256, 4×4 grid)

측정 결과, DiT-B/2는 FID 26.92, IS 2.55±0.05를 기록하였으며, DiR-B/2는 FID 30.06, IS 2.54±0.05를 기록하였다. IS는 두 모델이 거의 동일한 수준(차이 0.01 이내)으로 나타나, 생성된 이미지의 클래스 일관성·다양성 측면에서는 두 아키텍처 간 유의미한 차이가 관찰되지 않았다. 반면 FID는 DiR-B/2가 DiT-B/2 대비 약 3.14점 높게(나쁘게) 측정되어, 실제 이미지 분포와의 통계적 거리 측면에서는 현재 단계의 DiR-B/2가 DiT-B/2에 다소 못 미치는 것으로 나타났다.

이러한 FID 격차의 원인으로는 다음 몇 가지를 추정해 볼

수 있다. 첫째, 두 모델 모두 동일하게 200K 스텝까지만 학습되었는데, MSR은 셀프 어텐션과 다른 형태의 거리 기반 귀납 편향(inductive bias)을 가지므로 동일 스텝 수에서 DiT만큼 충분히 수렴하지 못했을 가능성이 있다. 둘째, 본 실험에서는 청크 크기 C=256을 고정하여 사용하였는데, 청크 경계에서의 양방향 상태 누적 근사(3-3절의 청크 간 연산)가 도입하는 근사 오차가 세밀한 텍스처나 고주파 디테일 표현에 영향을 미쳤을 가능성이 있다. 셋째, 단일 GPU(16GB VRAM) 환경에서 배치 크기 4로 제한된 학습 조건 자체가 두 모델 모두에 불리하게 작용했을 수 있으며, 이 조건에서의 상대적 격차가 더 큰 배치/장기 학습 환경에서도 동일하게 유지될지는 추가 검증이 필요하다. 세 가지 원인 중, 청크 간 근사오차(최대 0.04 미만)는 그 절대적 크기가 작아 FID 격차를 단독으로 설명하기는 어렵다고 판단된다. 반면 MSR의 거리 기반 감쇠(distance-based decay)는 Self-Attention과 근본적으로 다른 형태의 귀납적 편향을 가지므로, 동일한 학습 스텝(200K) 내에서 서로 다른 수렴 속도를 보였을 가능성이 상대적으로 더 유력한 설명으로 판단된다. 다만 위의 세 가지 가능성은 현재 실험 결과에 대한 사후적 해석에 해당하며, 본 연구에서는 이를 뒷받침할 별도의 통제 실험을 수행하지 못하였다. 따라서 FID 격차의 정확한 원인 규명과, 이를 완화하기 위한 아키텍처 개선은 후속 연구를 통해 검증할 필요가 있다.

V. 결 론

본 연구는 새로운 이미지 생성 방식으로서 Retention 기반 어텐션 대체 구조의 학습 가능성과 구조적 이점(O(N) 메모리)을 검증하는 데 목적이 있으며, 현재 단계에서 DiT의 생성 품질을 상회하는 것을 의도하지 않았다. 본 논문은 Diffusion Transformer의 셀프 어텐션을 Multi-Scale Retention으로 대체한 DiR 모델을 제안하였다. 양방향 감쇠 마스크와 청크 단위 양방향 스캔을 통해 이미지 도메인에서 요구되는 전방향 패치 의존성을 O(N) 메모리 복잡도로 모델링할 수 있는 구조를 설계하였으며, CelebA-HQ 데이터셋을 이용한 학습 실험을 통해 제안 구조가 안정적으로 학습 가능함을 확인하였다.

정량적 평가에서는 DiR-B/2가 DiT-B/2 대비 FID 측면에서 약 3.14점 높은(FID 30.06 vs 26.92) 결과를 보였으며, IS는 두 모델이 거의 동일한 수준(2.54 vs 2.55)으로 측정되었다. 이는 현재 단계의 DiR이 이미지 생성 품질 면에서 DiT를 완전히 대체할 만큼의 우위를 보이지는 못했음을 의미하지만, 동일한 학습 스텝과 제한된 GPU 자원이라는 동일 조건 하에서 셀프 어텐션과 구조적으로 다른 Retention 기반 메커니즘이 안정적으로 학습되고 합리적인 수준의 이미지를 생성할 수 있음을 확인하였다는 점에서 의의가 있다. 한편 해상도·배치 확장에 따른 메모리 효율성 벤치마크에서는 DiR-B/2가 FlashAttention 없이도 1024×1024·배치 16 조건까지 안정적으로 동작한 반면, FlashAttention이 비활성화된 DiT-B/2는 동일 조건에서 OOM이 발생하여, 제안 구조가 고해상도·대배치 환경에서 갖는 메모리 안정성 측면의 실질적 이점을 확인하였다. 다만 FlashAttention이 적용된 DiT-B/2와 비교하면, DiR-B/2는 처리 시간을 약 40% 단축시키는 대신 Peak VRAM은 오히려 약 29% 증가하는 것으로 나타나, 별도의 커널 최적화 없이 구조적 설계만으로 처리 속도 우위를 확보하는 대신 메모리 사용량 증가를 감수하는 trade-off가 존재함을 확인하였다.

본 연구의 한계는 다음과 같다. 첫째, 단일 GPU(16GB VRAM) 환경의 제약으로 배치 크기 4, 200K 스텝이라는 제한된 조건에서만 학습을 진행하여, 장기 학습이나 대규모 배치 조건에서도 현재의 FID 격차 및 처리 속도상의 상대적 우위가 동일하게 유지되는지 확인하지 못하였다. 둘째, Bidirectional 구조, Chunk Size, State Accumulation 등 제안 구조를 구성하는 개별 요소의 기여도를 분리 검증하는 ablation 실험을 자원 제약상 수행하지 못하였다. 셋째, DiT를 제외한 다른 선형 복잡도 계열 모델과의 직접적인 정량 비교를 동일 조건에서 수행하지 못하였다. 이러한 한계들은 본 연구의 결론을 성능 우위의 증명이 아닌 구조적 실현 가능성 검증으로 한정 짓는 근거이며, 아래의 향후 연구 과제를 통해 순차적으로 보완할 계획이다.

향후 연구 방향으로는 청크 단위 병렬화(chunkwise parallelism)를 통한 학습 속도 최적화, 더 큰 스케일의 DiR-XL/2 모델 학습, 그리고 2차원 공간 구조에 특화된 2D 감쇠 마스크 설계를 계획하고 있다. 특히 본 실험에서 확인된 FID 격차를 줄이기 위해, 청크 크기 C에 대한 ablation, 더 긴 학습 스텝에서의 수렴 비교, 청크 간 근사 오차를 줄이는 방향의 아키텍처 개선이 우선적으로 필요할 것으로 판단된다. 또한 Dhariwal과 Nichol이 제시한 아키텍처 개선 및 가이던스 기법[18]을 참고하여, 조건부 생성 품질 향상을 위한 Classifier-Free Guidance[19] 적용도 중요한 연구 과제로 남아 있다. 아울러 단순 전체 양방향 MSR 적용 방식과 본 연구가 제안한 청크 단위 양방향 스캔 간의 실측 학습 성능(FID/IS) 및 처리 속도·메모리 사용량을 직접 비교하는 실험을, 다중 GPU 또는 그래디언트 체크포인팅 등을 활용하여 후속 연구에서 보완할 계획이다. 또한 본 연구에서 직접 재현·비교하지 못한 DiS[9], ZigMa[10] 등 선형 복잡도 계열 선행연구와도 데이터셋·해상도·학습 스텝 등 동일 조건 하에서 정량적으로 비교하는 실험을 후속 과제로 계획하고 있다.

감사의 글

이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(RS-2022-NR-069955 (2022R1A2C1092178)).

참고문헌

[1] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” in *Advances in Neural Information Processing Systems* 33, pp. 6840-6851, December 2020.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, New Orleans: LA, pp. 10684-10685, June 2022. https://doi.org/10.1109/CVPR52688.2022.01042

[3] W. Peebles and S. Xie, “Scalable Diffusion Models with Transformers,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, Paris, France, pp. 4172-4182, October 2023. https://doi.org/10.1109/ICCV51070.2023.00387

[4] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, “FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness,” in *Advances in Neural Information Processing Systems* 35, New Orleans: LA, pp. 16344-16359, November 2022.

[5] Y. Sun, L. Dong, S. Huang, S. Ma, Y. Xia, J. Xue, ... and F. Wei, “Retentive Network: A Successor to Transformer for Large Language Models,” arXiv:2307.08621, July 2023. https://doi.org/10.48550/arXiv.2307.08621

[6] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, ... and N. Houlsby, “An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in *International Conference on Learning Representations*, 2021.

[7] A. Katharopoulos, A. Vyas, N. Pappas, and F. Fleuret, “Transformers Are RNNs: Fast Autoregressive Transformers with Linear Attention,” in *Proceedings of the 37th International Conference on Machine Learning*, pp. 5156-5165, July 2020.

[8] A. Gu and T. Dao, “Mamba: Linear-Time Sequence Modeling with Selective State Spaces,” arXiv:2312.00752, December 2023. https://doi.org/10.48550/arXiv.2312.00752

[9] Z. Fei, M. Fan, C. Yu, and J. Huang, “Scalable Diffusion Models with State Space Backbone,” arXiv:2402.05608, February 2024. https://doi.org/10.48550/arXiv.2402.05608

[10] V. T. Hu, S. A. Baumann, M. Gui, O. Grebenkova, P. Ma, J. Fischer, and B. Ommer, “ZigMa: A DiT-Style Zigzag Mamba Diffusion Model,” in *Proceedings of the 18th European Conference on Computer Vision*, Milan, Italy, pp. 148-166, September 2024.

[11] Y. Wu and K. He, “Group Normalization,” in *Proceedings of the 15th European Conference on Computer Vision*, Munich, Germany, pp. 3-19, September 2018. https://doi.org/10.1007/978-3-030-01261-8_1

[12] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive Growing of GANs for Improved Quality, Stability, and Variation,” in *International Conference on Learning Representations*, Vancouver, Canada, April 2018.

[13] Z. Liu, P. Luo, X. Wang, and X. Tang, “Deep Learning Face Attributes in the Wild,” in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, pp. 3730-3738, December 2015. https://doi.org/10.1109/ICCV.2015.425

[14] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” in *Proceedings of the 7th International Conference on Learning Representations*, New Orleans: LA, May 2019.

[15] A. Q. Nichol and P. Dhariwal, “Improved Denoising Diffusion Probabilistic Models,” in *Proceedings of the 38th International Conference on Machine Learning*, pp. 8162-8171, July 2021.

[16] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium,” in *Advances in Neural Information Processing Systems* 30, Long Beach: CA, pp. 6626-6637, December 2017.

[17] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” in *Advances in Neural Information Processing Systems* 29, Barcelona, Spain, pp. 2234-2242, December 2016.

[18] P. Dhariwal and A. Nichol, “Diffusion Models Beat GANs on Image Synthesis,” in *Advances in Neural Information Processing Systems* 34, pp. 8780-8794, December 2021.

[19] J. Ho and T. Salimans, “Classifier-Free Diffusion Guidance,” arXiv:2207.12598, July 2022. https://doi.org/10.48550/arXiv.2207.12598



이중현(Jung-Hyeon Lee)

2020년～현 재: 강남대학교 ICT융합공학부 학사과정

※관심분야 : 생성형 AI, 컴퓨터 비전



최웅(Woong Choi)

2000년 : 조선대학교(공학석사)
2005년 : Tokyo Institute of Technology(공학박사-지능시스템과학)
2005년~2010년: Ritsumeikan University
2010년~2022년: National Institute of Technology, Gunma College
2022년~현 재: 강남대학교 컴퓨터공학부 교수
※관심분야 : 생성형 AI, 컴퓨터 비전