JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 5, pp. 1351-1358, May. 2026

Check for updates

웨이블릿 기반 저비트 양자화를 이용한 확산 모델 경량화

서 용 석* · 임 동 혁
한국전자통신연구원 콘텐츠연구본부 책임연구원

Lightweight Diffusion Models via Wavelet-Based Low-Bit Quantization

Yongseok Seo* · **Dong-Hyuck Im**

Principal Researcher, Content Research Division, Electronics and Telecommunications Research Institute, Daejeon 34129, Korea

[요 약]

확산 모델은 최첨단 이미지 생성 성능을 제공하지만, 높은 메모리 연산 요구로 인해 운용 비용이 매우 크다. 본 연구는 SDXL을 위한 웨이블릿 기반 저비트 양자화 파이프라인을 제안하며, 가중치 중심 채널별 평활화와 U-Net 중간 특징맵에 적용하는 DWT를 결합한다. DWT는 에너지를 LL 서브밴드에 집중시켜 구조적 정보를 보존하고, 서브밴드 특성에 따라 비트폭을 달리하는 차등 양자화를 가능하게 한다. 또한 제안한 평활화는 가중치 중심의 스케일 재배치를 통해 활성값 이상치의 영향을 완화한다. RTX 6000 GPU에서 실험한 결과, 제안한 Wavelet-INT4 모델은 최대 VRAM을 약 70% 절감하고, 1024×1024 해상도에서 FP32 대비 1.24배 빠른 추론을 달성했으며, 동일 프롬프트에서 CLIP-FID 차이 5 이하로 품질도 유지했다. 이는 특징맵에 대한 웨이블릿 분석과 가중치 중심 평활화의 결합이 품질, 속도, 메모리 효율 사이의 균형을 개선하여, 소비자용 GPU 환경에서도 실용적인 저비트 확산 모델 운용 가능성을 보여줌을 의미한다.

[Abstract]

Diffusion models achieve state-of-the-art performance in image generation; however, their substantial memory and computational requirements make practical deployment highly resource-intensive. To address this limitation, this study proposes a wavelet-based low-bit quantization pipeline for Stable Diffusion XL (SDXL). The framework integrates weight-centric channel-wise smoothing with a discrete wavelet transform (DWT) applied to intermediate U-Net feature maps. The proposed approach leverages the energy compaction property of the DWT, where most structural information is concentrated in the LL subband. This enables subband-wise adaptive quantization, in which different bit-widths are assigned based on the characteristics of each frequency component. Furthermore, the weight-centric smoothing mechanism mitigates activation outliers through scale rebalancing, improving quantization stability. Experimental results on an RTX 6000 GPU show that the proposed Wavelet-INT4 model reduces peak VRAM usage by approximately 70% and achieves 1.24× faster inference than FP32 at a resolution of 1024×1024. Despite aggressive compression, the model maintains high visual fidelity, with a CLIP-FID gap of 5 or less under identical prompts. These results indicate that combining wavelet-based feature decomposition with weight-centric smoothing effectively improves the trade-off between image quality, computational efficiency, and memory usage, enabling practical deployment of low-bit diffusion models on consumer-grade hardware.

색인어 : 확산 모델, 경량화, 양자화, 저비트, 웨이블릿 변환

Keyword : Diffusion Models, Lightweight, Quantization, Low-Bit, Wavelet Transform

http://dx.doi.org/10.9728/dcs.2026.27.5.1351

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 09 March 2026; **Revised** 14 April 2026
Accepted 04 May 2026

***Corresponding Author; Yongseok Seo**

Tel: +82-42-860-1137
E-mail: yongseok@etri.re.kr

Ⅰ. 서 론

딥러닝 기반 생성 모델 중 확산 모델(Diffusion Models)은 이미지 생성 및 텍스트-이미지 합성 분야에서 뛰어난 성능으로 주목받고 있다[1]. 확산 모델은 데이터에 점진적으로 노이즈를 추가하는 과정을 학습하고, 이를 역으로 제거하며 새로운 데이터를 생성하는 방식으로, 타 생성 모델 대비 높은 학습 안정성과 우수한 일반화 성능을 제공한다. 그러나 이러한 탁월한 성능에도 불구하고, 확산 모델은 방대한 파라미터 수와 반복적인 계산 과정으로 인해 막대한 연산량과 메모리 요구량이 매우 커서, 고사양 GPU가 아닌 환경에서는 실시간 활용이 어렵다는 한계가 있다[2],[3]. 이러한 문제를 해결하기 위해 모델 경량화 연구가 활발히 진행중이며, 그 중 양자화(Quantization) 기법은 모델 파라미터를 저정밀도로 표현하여 연산을 가속하고 메모리 사용을 줄이는 핵심 방안으로 주목받고 있다[2].

모델 양자화는 뉴럴넷의 가중치나 활성값을 부동소수점 대신 낮은 비트의 정수로 표현하여 연산 효율을 높이는 기법이다[2]-[4]. 일반적으로 각 레이어별 데이터 분포에 맞추어 적절한 스케일 팩터와 제로포인트를 선택한 후, 연속적인 실수 값을 스케일링 및 반올림하여 이산적인 정수값으로 매핑함으로써 메모리 대역폭의 병목 현상을 해소한다[2]. 이러한 방식은 특히 곱셈 연산을 가속하고 전력 소모를 대폭 줄여주며, 확산 모델과 같이 다단계 변환을 거치는 구조에서는 양자화에 따른 이득이 단계별로 누적되어 전체 추론 성능을 비약적으로 향상시킬 수 있다. 다만 양자화로 인한 정밀도 손실 때문에 출력 품질 저하가 발생할 수 있어, 이를 최소화하기 위한 다양한 연구가 진행 중이다[2]-[9].

기존 학습 후 양자화(PTQ; Post-Training Quantization) 방식은 활성값 이상치(outlier)가 양자화 범위를 과도하게 확장하여 대부분의 채널에서 유효 비트 수를 감소시키는 문제를 갖는다. 또한 정확도 측면에서 유리한 채널별 활성값 양자화는 하드웨어 친화적인 일반적인 INT8 행렬곱셈 구현과 직접 양립하기 어려워, 정확도와 효율을 동시에 만족하는 PTQ 설계가 쉽지 않다. 최근 few-step 확산 양자화 연구는 이러한 문제가 더욱 두드러지며, 균일 비트 폭 기반 양자화가 소수의 매우 민감한 레이어에 의해 병목이 발생되고 시각적 품질뿐 아니라 텍스트-이미지 정합성까지 동시에 저하시킬 수 있음을 보고하였다. 따라서 few-step 확산용 U-Net에서는 이상치 완화, 구조 정보 보존, 그리고 레이어별 감도를 함께 고려하는 PTQ 설계가 필요하다.

본 연구에서는 Stable Diffusion 기반 모델에 저비트 양자화 기법을 적용하여 고효율의 경량화된 확산 모델을 구현하고자 한다. 구체적으로는 점진적 적대적 확산 증류를 통해 생성 단계를 단축한 SDXL-Lightning 모델[10]을 대상으로 선정하였다. 본 논문에서는 해당 모델의 핵심 구조인 U-Net[11]에 저비트 양자화를 적용함으로써, FP32 정밀도와 유사한 품질을 유지하면서도 추론 속도를 가속하고 메모리 사용량을 절감하는 방법론을 제시한다. 더 나아가 양자화 과정에 웨이블릿 기법을 도입하여 정보 보존 능력을 한층 강화하였다. 특징맵을 주파수 영역으로 분해하여 정보의 중요도에 따라 차등적인 양자화를 수행함으로써, 기존 방식보다 더 낮은 비트 정밀도에서도 이미지의 세부 디테일을 효과적으로 유지하고 메모리 절감 효과를 극대화할 수 있음을 확인하였다.

Ⅱ. 관련 연구

최근 고품질 이미지 생성을 위한 확산 모델, 특히 Stable Diffusion XL(SDXL)[12] 계열은 높은 표현력을 제공하지만, 그 대가로 막대한 연산량과 메모리 요구가 동반된다. 이에 따라 다양한 가속·경량화 기법이 제안되었으며, 양자화는 그 중에서도 하드웨어 친화적이고 학습 없이 적용 가능하다는 장점으로 널리 연구되고 있다. PTQ 기반 기법으로는 대표적으로 SmoothQuant[13]이 있다. 대규모 언어모델(LLM)의 8비트 양자화를 위해 제안된 이 기법은 활성값 분포의 이상치 문제를 완화하여 가중치와 활성값 모두 INT8 양자화를 가능하게 한 방법이다. Xiao 등은 LLM의 경우 특정 채널의 활성값이 유난히 커서 INT8로 직접 양자화하면 정확도 저하가 발생함을 지적하고, “활성값의 어려움을 가중치로 이전”하는 아이디어를 제시하였다. 구체적으로, 각 레이어에서 평활화 팩터 D를 정의해 활성값 X를 D로 나누고 가중치 W에 D를 곱하는 선형 변환을 수행한다. 이렇게 하면 X와 W의 곱 XW는 변환 전과 동일하지만, 활성값의 분포가 균일해져 양자화하기 쉬워지고 대신 가중치의 분포가 다소 넓어지게 된다. 결과적으로 두 변수 모두 INT8로 무난히 표현 가능해져, LLM에서 부동소수점 대비 성능 저하 없이 약 2배 메모리 절감과 1.5배 속도 향상을 이루었다고 보고 되었다.

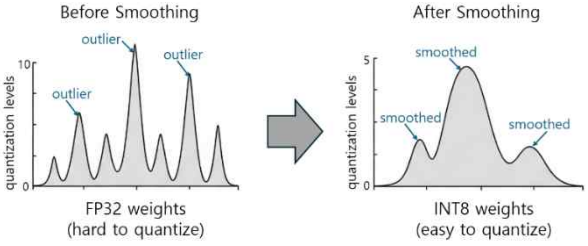


그림 1. SmoothQuant 기법의 개념
Fig. 1. SmoothQuant's intuition

SmoothQuant는 추가 학습 없이 적용 가능하고 연산 오버헤드가 거의 없는 간결한 방법으로 트랜스포머 기반 언어 모델을 중심으로 검증되었으나, 확산 모델의 경우에는 활성 분포의 시공간 변화가 커서 직접적인 적용에 한계가 있다. 특히 Stable Diffusion 모델의 경우 attention 블록과 convolution 연산이 복잡하게 얽혀있어, 일괄적인 비트폭 할당은 생성 이미지 품질 열화의 원인이 된다. 이에 MiXDQ[14]는 레이어별 민감도 분석을 통해 민감도가 높은 층은 상대적으로 높은 정

밀도로 유지하고, 나머지 층에는 다양한 비트폭을 혼합 적용하는 혼합 정밀도 양자화 기법을 제안하였다. MixDQ는 SDXL-turbo[15] 1-step 설정에서 혼합 정밀도 양자화를 통해 FP16 대비 약 2~3배의 메모리 절감과 최대 1.5배의 속도 향상을 보고하였으며, 일부 설정에서는 FID 증가를 매우 작게 유지하였다. 또한, BOS(Begin-of-Sentence) 토큰과 관련된 텍스트 임베딩 층이 특별히 민감하다는 분석을 통해 BOS-aware 양자화 기법을 제안한 바 있다. 그러나, 이 방법은 균일 양자화 대신 복잡한 단계(BOS-aware 양자화, Metric_Decoupled 민감도 분석)를 포함하고 있으며, BOS 토큰에 대한 부동 소수점 특징을 미리 계산하고 오프라인으로 저장해야 하는 추가적인 과정이 필요한 단점이 있다.

한편, 딥러닝 모델에서 웨이블릿 변환을 활용하여 다중 해상도 정보를 통합하려는 시도도 존재한다. Phung 등은 확산 모델의 속도를 높이기 위해 웨이블릿 도메인에서 저주파와 고주파 성분을 분리 처리하는 Wavelet Diffusion 방식을 제안하였고[16], 적은 단계로도 고품질 이미지를 얻어 추론 속도를 GAN 수준으로 향상시켰다. 본 논문에서 주목하는 것은 웨이블릿 변환을 통한 양자화 효율 향상으로, Sun 등이 제안한 MWQ(Multiscale Wavelet Quantization) 방법[17]이 이에 해당한다. MWQ는 이미지의 주파수 도메인 특성에 착안하여, 모델의 가중치나 활성값을 Haar 웨이블릿 등으로 다중 밴드 분해 후 각 밴드를 따로 양자화하는 기법이다. 기존 양자화가 개별 값들의 크기만 고려하는 데 반해, 웨이블릿 양자화는 공간적 인접성과 다중 스케일 특성을 활용함으로써 정보 손실을 줄인다. 예를 들어 웨이블릿 변환으로 얻은 저주파 성분은 원본의 윤곽, 큰 형태를 담고 있어 분포가 원본과 유사한 반면 고주파 성분들은 에지나 질감 등 국소 세부 정보를 담고 있어 값 분포가 다르다. MWQ는 이러한 각 성분의 분포에 맞춤형 클리핑 범위와 양자화 스텝을 적용하여, 통합했을 때 전통적 양자화보다 더 많은 재현 상태를 제공함을 보였다. 그 결과 ResNet 계열 모델의 양자화 실험에서, 기존 방식 대비 분류 정확도 향상 및 검출/분할 성능 개선을 달성하였다. 이러한 선행 연구들은 웨이블릿을 통한 다중 해상도 표현이 양자화에 유리할 가능성을 시사하며, 본 논문에서는 이를 Stable Diffusion 모델 경량화에 활용하고자 한다.

SmoothQuant는 활성값 이상치가 양자화의 핵심 병목이라는 점을 지적하고, 활성값 쪽의 양자화 난이도를 가중치 쪽으로 이동시키는 평활화 전략을 통해 정확도와 하드웨어 효율의 균형을 맞추고자 하였다. 특히 채널별 활성값 양자화는 정확도 보존에는 효과적이지만 하드웨어 친화적인 INT8 행렬곱셈과는 잘 맞지 않는다는 점을 명확히 보여주었다. 한편 MixDQ는 few-step 텍스트-이미지 확산에서 레이어 민감도가 롱테일 특성을 보이며, 양자화가 이미지 화질뿐 아니라 텍스트-이미지 정합성에도 동시에 영향을 준다고 분석하였다. 또한 텍스트 임베딩의 이상치를 다루는 BOS-aware 양자화와 메트릭 분리 혼합 정밀도 할당 기법을 제안하여 few-step 설정의 난점을 해결하고자 하였다. 본 연구는 이러한 선행 분석을 바탕으로, SDXL-Lightning 4-step U-Net에서 변형된 평활화와 웨이블릿 기반 스킵 연결 양자화를 결합한 실용적 저비트 PTQ 프레임워크를 제안한다.

2-1 기존 PTQ 방식의 한계와 Few-Step 확산에서의 양자화

기존 PTQ가 고성능 생성 모델에 직접 적용되기 어려운 첫 번째 이유는 활성값 분포의 불균형과 이상치 때문이다. SmoothQuant는 일부 채널의 큰 활성값이 전체 양자화 범위를 지배하면서 대부분의 값이 매우 낮은 유효 비트 수로 표현되고, 그 결과 큰 양자화 오류가 발생한다고 설명하였다. 또한 활성값 양자화에서 정확도를 유지하려면 채널별 수준의 세밀한 처리가 필요하지만, 이러한 방식은 하드웨어 가속 행렬곱셈 커널과 직접 결합되기 어렵다. 결국 기존 PTQ는 정확도 보존과 하드웨어 효율 사이에서 구조적 trade-off를 가진다.

두 번째 이유는 few-step 확산이 다단계 확산보다 양자화 잡음에 더 민감하다는 점이다. MixDQ는 반복 디노이징 단계가 적은 few-step 설정에서는 양자화 오류가 후속 단계에서 충분히 상쇄되지 못하며, 민감도 역시 모든 레이어에 균등하게 분포하는 것이 아니라 일부 매우 민감한 레이어에 집중된다고 분석하였다. 더 나아가 양자화의 영향은 단순한 블러나 아티팩트와 같은 시각적 품질 저하에 그치지 않고, 프롬프트의 핵심 구성 요소가 누락되는 방식의 텍스트-이미지 정합성 저하로도 이어질 수 있다. 따라서 few-step 확산 PTQ는 생성 이미지 화질과 정합성을 동시에 고려하는 문제로 다루어져야 한다.

Ⅲ. 제안 방법: 웨이블릿 기반 저비트 양자화 기법

본 장에서는 Stable Diffusion 모델의 경량화를 위하여 제안하는 웨이블릿 기반 스킵 연결 및 저비트 양자화를 소개한다. 본 방법은 기존의 PTQ 방식에서 흔히 발생하는 양자화 오류 및 품질 저하 문제를 완화하기 위하여, 수정된 가중치 중심 평활화 기법과 웨이블릿 기반 저비트 양자화를 U-Net 파이프라인에 통합한다.

제안 방법의 각 구성요소는 서로 다른 실패 요인에 대응하도록 설계되었다. 수정된 가중치 중심 평활화는 활성값과 특징 사이의 불균형과 이상치로 인한 양자화 범위 왜곡을 완화하고, 웨이블릿 기반 스킵 연결 양자화는 스킵 경로에서 전달되는 고주파 구조 정보의 손실을 줄이며, 하이브리드 저비트 할당은 경로 및 레이어별 민감도 차이를 반영하여 메모리 절감과 품질 유지 사이의 균형을 맞춘다. 이러한 문제 분해는 활성값 이상치 분석과 레이어 민감도 불균형 분석에 기반한 최근 PTQ 연구의 관찰과도 일치한다.

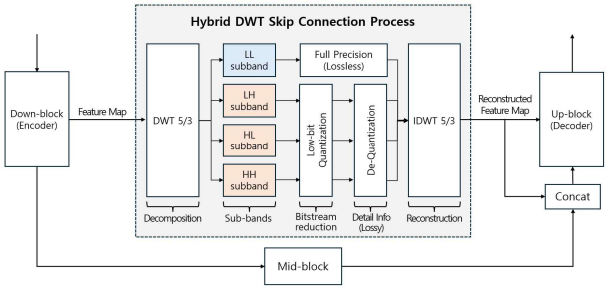


그림 2. 하이브리드 웨이블릿 기반 스킵 연결 구조
Fig. 2. Hybrid wavelet-based skip connection architecture

3-1 웨이블릿 기반 스킵 연결(Wavelet-Based Skip Connection)

고해상도 이미지 생성 과정에서 U-Net 구조의 스킵 연결은 인코더의 세부 특징을 디코더로 전달하는 필수적인 역할을 수행하지만, 해상도가 높아짐에 따라 기하급수적으로 증가하는 메모리 점유 및 대역폭 병목 문제를 야기한다. 본 연구에서는 이를 해결하기 위해 특징맵을 공간적 해상도가 아닌 주파수 성분별로 분해하고, 정보의 중요도에 따라 비트 정밀도를 차등 할당하는 하이브리드 웨이블릿 기반 양자화 기법을 제안한다. 이는 복원 품질 저하를 최소화하면서도 메모리 효율을 극대화하는 것을 목적으로 하며, 인코더의 각 다운블록에서 추출된 중간 특징맵에 대해 DWT를 수행한다. 본 방법론에서는 연산 효율성과 정수 기반의 무손실 복원 가능성을 고려하여 LeGall-5/3 필터를 채택하였다. DWT를 통해 4개의 서브밴드로 분해되며, 저주파 성분인 LL 밴드는 이미지의 거시적인 구조와 형태 정보를 담고 있으며, 고주파 성분인 LH, HL, HH 밴드는 세부적인 엣지 및 텍스처 정보를 포함한다. 분해된 각 서브밴드의 시각적 중요도에 따라 정보 손실과 압축률 사이의 최적의 균형을 찾기 위해 다음과 같은 하이브리드 양자화 전략을 적용한다. 이미지 생성의 핵심 뼈대가 되는 LL 밴드는 정보 손실을 최소화하기 위해 FP16(Full Precision) 또는 INT8 양자화를 선택적으로 적용하고, 상대적으로 정보 밀도가 낮고 시각적 민감도가 떨어지는 고주파 밴드들에 대해서는 4비트 양자화를 적용한다. 특히, 두 개의 4비트 데이터를 하나의 8비트 단위로 패킹함으로써 물리적인 메모리 사용량을 원본 대비 획기적으로 절감한다. 디코더의 각 업블록에서는 스킵 연결을 통해 전달된 패키지 데이터를 호출한다. 먼저 저장된 고주파 데이터를 역양자화한 후, 높은 정밀도로 보존된 LL 밴드와 결합하여 웨이블릿 역변환을 수행한다. 이를 통해 원본과 동일한 해상도를 가진 특징맵을 재구성한다. 최종적으로 재구성된 특징맵은 디코더 내에서 업샘플링된 특징맵과 채널 방향으로 결합되어 잔차 정보로 활용됨으로써 최종 고해상도 이미지 생성에 기여한다.

3-2 가중치 평활화(Weight Smoothing)

Stable Diffusion 모델 내 U-Net의 가중치 분포는 매우 비균일하며, 특히 특정 채널의 값 범위가 비정상적으로 커지는 경향이 있어 양자화 시 심각한 정밀도 저하를 유발한다. 본 연구에서는 이를 해결하기 위해, SmoothQuant[13]의 평활화 매커니즘을 변형하여 가중치 중심의 정규화 방식을 제안한다. SmoothQuant의 핵심 원리는 양자화 전 가중치와 활성값의 분포를 사전에 평활화(smoothing)하여, 양자화 과정에서 발생하는 클리핑 손실과 양자화 오차를 동시에 최소화하는 것이다. 이를 위해 [13]에서는 활성값의 이상치를 억제하기 위해 채널별 스케일 조절 방식을 도입하였으며, 활성값을 스케일링 인자로 나누는 동시에 가중치에 같은 값을 곱함으로써 수학적 동등성을 유지한다. 실제 구현에서는 해당 스케일을 이전 레이어의 파라미터에 오프라인으로 흡수하여, 추론 시 별도의 활성값 스케일링 연산을 추가하지 않았다. 이를 통해 전체적인 데이터 분포를 균일하게 재조정하여 양자화 효율을 극대화한다. 아래에서 X는 레이어 입력인 활성값, W는 가중치 텐서, s는 채널별 스케일링 인자이며, 다음 식과 같이 활성값 및 가중치를 조정한다.

$$\hat{X} = X/s,\ \hat{W} = W \cdot s \tag{1}$$

$$Y = (X \cdot diag(s)^{-1}) \cdot (diag(s)\, W) = \hat{X}\hat{W} \tag{2}$$

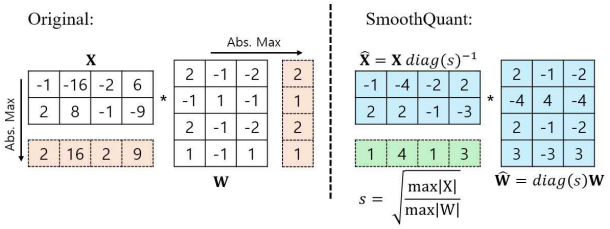


그림 3. SmoothQuant 주요 아이디어
Fig. 3. Main idea of SmoothQuant

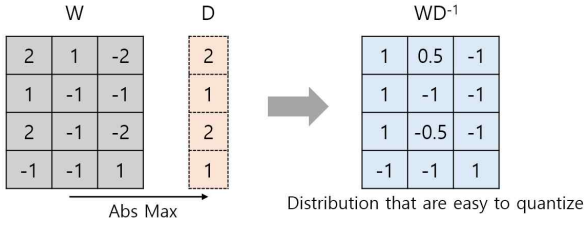


그림 4. 가중치 중심의 평활화 적용
Fig. 4. Weight-centric smoothing

그러나 기존의 SmoothQuant 기법을 Stable Diffusion 모델에 직접 적용할 경우, 생성된 이미지가 극심한 노이즈 형태를 띠는 현상이 발생한다. 확산 모델의 각 레이어를 분석한 결과, 해당 모델의 활성값 분포는 LLM 모델에 비해 훨씬 불규칙하고 비균일한 특성을 보였다. 이러한 분포 특성상 활성값의 채널별 최댓값에 의존하여 스케일링을 수행할 경우, 오히려 수치적 안정성이 저하되어 모델의 성능을 해치는 것으로 파악되었다. 따라서 본 논문에서는 활성값이 아닌 가중치의 채널

별 최댓값을 기준으로 평활화를 수행하고, 수학적 동등성을 보존하기 위한 보정 항은 구현상 이전 레이어 파라미터에 사전 흡수하는 방식으로 처리하였다. 즉 U-Net의 모든 선형 및 컨볼루션 레이어의 가중치 행렬 W에 대해 그림 4와 같이 채널별 최댓값으로 구성된 평활화 벡터 D를 산출하여 이를 나누어주고, 이때 대응하는 활성값 보정은 이전 레이어 파라미터에 사전 흡수하여 런타임 오버헤드를 추가하지 않도록 하였다. 이 과정을 통해 U-Net의 전체 가중치 분포는 양자화 친화적인 범위로 스케일링되며, 결과적으로 정수 표현 오차를 감소시켜 고품질의 양자화된 이미지 생성을 가능하게 한다.

3-3 저비트 양자화(Low-Bit Quantization)

앞서 기술한 과정을 통해 획득한 스킵 연결의 서브밴드별 성분과 평활화된 가중치 텐서에 대해 그림 5와 같은 비대칭 선형 양자화를 적용하여 저비트 데이터로 변환한다. 본 단계에서는 데이터의 최소값과 최댓값을 독립적으로 매핑하는 비대칭 방식을 채택하여, 비균일한 분포 내에서도 정보 손실을 방지한다. 이를 위한 정수 양자화 스케일 인자 s와 zero-point z는 다음과 같이 정의된다.

$$s = (q_{\max} - q_{\min}) / (r_{\max} - r_{\min}) \tag{3}$$

$$z = round(q_{\min} - r_{\min} \cdot s) \tag{4}$$

여기서 $q_{\max}, q_{\min}$은 정수 표현 범위(예: INT8일 경우 −128~127), $r_{\max}, r_{\min}$은 실수 가중치 범위이다. 최종적으로 양자화는 다음식과 같이 수행된다.

$$Q(W') = clip(round(W' \cdot s + z), q_{\min}, q_{\max}) \tag{5}$$

이때 사용되는 zero-point는 양자화 전 0 값이 양자화된 구간으로 매핑되었을때의 값을 의미하고, 스케일 인자 s와 zero-point z를 양자화 파라미터라 하며 모델의 가중치와 활성값의 분포를 통해 결정된다.

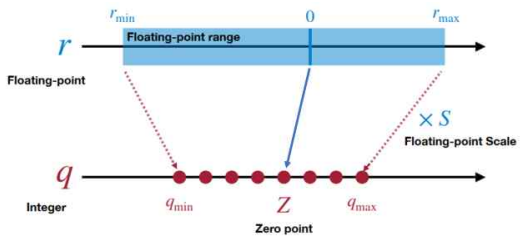


그림 5. 비대칭 선형 양자화 방법
Fig. 5. Asymmetric linear quantization method

Ⅳ. 실험 결과

본 논문에서 제안한 하이브리드 양자화 기법의 효용성을 검증하기 위해, SDXL 기본 모델에 점진적 적대적 확산 증류를 적용한 SDXL-Lightning 4-step 모델을 대상으로 실험을 수행하였다. Few-step 확산 모델은 디노이징 반복 횟수가 적기 때문에 양자화 오류가 후속 단계에서 충분히 보정되기 어렵고, 따라서 다단계 설정보다 PTQ에 더 도전적인 평가 환경으로 간주된다. 해당 모델은 단 4단계의 추론만으로 1024×1024 해상도의 고품질 이미지를 생성하도록 최적화되어 있으며, 본 실험에서는 공개된 사전학습 체크포인트를 활용하였다. 실험 환경은 NVIDIA RTX 6000 (48GB) 및 RTX 3060 (12GB) GPU에서 각각 수행되었다. 모델 추론에는 float32 기반 환경에서 진행하되, U-Net 모듈을 제안된 양자화 모델로 대체하여 성능을 비교 분석하였다. 성능 지표로는 동일한 텍스트 프롬프트에 대한 평균 이미지 생성 시간과 GPU 메모리 피크 사용량, CLIP-FID를 측정하였다. 비교 대상은 FP32 정밀도의 기존 모델, 가중치 평활화가 적용된 INT8 양자화 모델, 평활화를 제거한 naive PTQ(INT8, No Smoothing), 그리고 본 논문에서 제안한 웨이블릿 기반 하이브리드 INT4(Wavelet-INT4) 양자화 모델이다. 이를 통해 동일 비트폭 조건에서 평활화의 효과와, 더 낮은 비트폭에서의 웨이블릿 기반 압축 효과를 각각 분리하여 비교하였다. SmoothQuant가 활성값 양자화의 어려움을 비교 실험으로 보여주었고, MixDQ 역시 few-step 확산 PTQ의 난점을 베이스라인과의 직접 비교를 통한 실험으로 제시한 바 있다. MixDQ는 SDXL-turbo 1-step 설정에서 FP16의 FID 17.15에 대해 naive PTQ W8A8은 103.96, Q-Diffusion W8A8은 76.18, MixDQ W8A8은 17.03을 보고하였다. 이는 few-step 확산 PTQ에서 단순 균일 양자화가 매우 큰 성능 저하를 유발할 수 있음을 보여준다. 다만 해당 결과는 SDXL-turbo 1-step 기준이며 본 연구의 SDXL-Lightning 4-step 설정과는 동일하지 않으므로, 본 논문에서는 이를 관련 연구의 참고 결과로 제시하고, 직접 비교는 동일한 4-step 조건에서 재현한 naive PTQ, 가중치 중심 평활화 기반 INT8, Wavelet-INT4 baseline을 중심으로 수행하였다. 본 논문은 동일 설정에서의 내부 비교를 중심으로 분석하며, MixDQ 및 기타 PTQ 방법과의 수치 비교는 대상 모델과 step 설정 차이로 인해 관련 연구 논의로 한정하였다.

4-1 추론 시간 및 메모리 효율성

표 1은 GPU 환경별 모델 설정에 따른 이미지 1장당 평균 생성 시간, 최대 VRAM 사용량, 그리고 CLIP-FID를 비교한 결과를 나타낸다. FP32 모델의 경우 RTX 3060에서는 오프로딩 없이 실행이 불가능하였으며, 오프로딩을 적용하더라도 추론 시간이 크게 증가하였다. 반면 INT8, naive PTQ(INT8, No Smoothing), 그리고 Wavelet-INT4 모델은 두 GPU 환경 모두에서 오프로딩 없이 실행 가능하였다. 이는 few-step 확산 모델에서도 저비트 양자화가 메모리 병목을 완화하는 데 실질적으로 기여함을 보여준다. 먼저 동일한 8비트 설정에서 naive PTQ와 가중치 중심 평활화가 적용된 INT8 모델을 비교하면, 두 방법의 추론 시간과 VRAM 사용량은 거의 유사하

였다. RTX 6000 환경에서 naive PTQ는 1.61초, 6683MB였고, 평활화가 적용된 INT8 모델은 1.67초, 6690MB로 측정되었으며, RTX 3060에서도 큰 차이가 없는 결과를 보였다. 이는 가중치 중심 평활화가 별도의 무거운 런타임 연산을 추가하기보다, 동일한 INT8 양자화 경로 내에서 주로 품질 안정화에 기여함을 시사한다. 즉, 평활화의 효과는 효율 향상 자체보다는 동일한 메모리·연산 예산에서 양자화 오차를 더 안정적으로 제어하는 데 있다고 해석할 수 있다. 한편, Wavelet-INT4 모델은 RTX 6000에서 1.80초, 5419MB를 기록하여, FP32 대비 약 68.5%의 피크 메모리 절감과 INT8 대비 약 19.0%의 추가 메모리 절감을 달성하였다. RTX 3060에서도 5416MB 수준으로 유지되어, 6GB급 메모리 환경에서도 오프로딩 없이 실행 가능함을 확인하였다. 이는 웨이블릿 분해 후 고주파 성분을 4비트로 패킹하여 저장함으로써 스킵 연결에서 발생하는 메모리 사용량을 효과적으로 줄인 결과로 볼 수 있다. 추론 시간 측면에서는 INT8 계열이 가장 빠르게 측정되었고, Wavelet-INT4는 이보다 다소 느린 경향을 보였다. 이는 현재 PyTorch 환경에서 별도의 INT4 연산 커널을 직접 활용하기 어렵기 때문에, 4비트 표현을 위해 두 개의 값을 하나의 INT8에 패킹하고 복원하는 부가 연산이 필요하기 때문이다. 그럼에도 불구하고 Wavelet-INT4는 FP32 대비 여전히 더 낮은 메모리 사용량과 실용적인 추론 속도를 유지하였으며, 결과적으로 few-step 고해상도 확산 모델에서 품질-속도-메모리 사이의 실용적인 절충점을 제공함을 확인하였다.

표 1. SDXL-Lightning 모델(4 step, 1024²)에 대한 실험 결과
Table 1. Experimental results on SDXL-Lightning (4-step, 1024^2): time (s/img) ↓, peak VRAM (MB) ↓, and CLIP-FID ↓

Model	Offload	RTX 6000		RTX 3060		CLIP-FID
		Time	VRAM	Time	VRAM	
FP32	No	2.24	17227	Out of memory		13.87
FP32	Yes	5.54	10225	13.68	10229	14.67
naive PTQ (INT8, No Smoothing)	No	1.61	6683	4.45	6689	19.24
INT8	No	1.67	6690	4.68	6693	16.16
Wavelet-INT4	No	1.80	5419	5.58	5416	18.75
Wavelet-INT4	Yes	2.67	2599	7.70	2599	18.82

4-2 시각적 품질 분석

시각적 품질에 대한 정량 평가는 CLIP-FID를 중심으로 수행하였으며, 텍스트-이미지 정합성은 정성 비교를 통해 보조적으로 확인하였다. 이는 few-step 확산 모델에서 양자화 오차가 단순한 화질 저하뿐 아니라 프롬프트 콘텐츠의 누락으로도 나타날 수 있다는 최근 PTQ 연구의 관찰을 반영한 것이다. 실제로 MixDQ는 few-step 확산 양자화 평가에서 이미지 품질과 정합성을 함께 고려해야 함을 지적하며, FID와 CLIP Score를 분리하여 분석하였다. 본 연구는 동일한 문제의식을 반영하되, 현재 실험에서는 CLIP-FID와 정성 비교를 중심으로 결과를 해석하였다. 표 1의 결과를 보면, 동일한 8비트 조건에서도 가중치 중심 평활화의 유무에 따라 시각적 품질 차이가 분명하게 나타난다. 가중치 중심 평활화가 적용된 INT8 모델의 CLIP-FID는 16.16으로 FP32의 13.87 대비 2.29 증가에 그친 반면, 동일한 8비트 설정에서 평활화를 제거한 naive PTQ는 19.24로 증가하여 FP32 대비 5.37의 성능 저하를 보였다. 이는 두 방법의 추론 시간과 VRAM 사용량이 거의 유사함에도 불구하고, 평활화 전처리가 few-step 확산 모델의 8비트 양자화에서 품질 열화를 완화하는 데 중요한 역할을 한다는 점을 보여준다. 정성 비교에서도 INT8 모델은 FP32와 유사한 구조와 색감을 유지한 반면, naive PTQ는 세부 패턴의 불안정성과 색감 왜곡이 보다 뚜렷하게 나타났다. 한편 Wavelet-INT4 모델의 CLIP-FID는 18.75로, FP32 대비 4.88의 차이를 보였으나 naive PTQ보다는 더 낮은 값을 유지하였다. 이는 비트폭을 4비트 수준까지 낮추면서도 웨이블릿 기반 스킵 연결 양자화와 가중치 중심 평활화를 함께 적용함으로써, 단순 균일 양자화보다 구조적 왜곡을 더 효과적으로 억제했음을 시사한다. 특히 그림 6에서 확인할 수 있듯이, 평활화 전처리를 적용하지 않은 경우에는 색감 왜곡과 패턴 깨짐이 두드러지지만, 전처리 후에는 이러한 아티팩트가 완화된다.

그림 6. 양자화 전 평활화 전처리 효과
Fig. 6. Effect of preprocessing smoothing before quantization

또한 그림 7의 비교 결과에서도 Wavelet-INT4는 일부 세부 묘사에서 미세한 차이를 보일 뿐, 전체 장면 구성과 주요 시각 요소는 안정적으로 유지된다. 이러한 결과는 웨이블릿 분해 후 저주파 영역(LL 서브밴드)을 상대적으로 높은 정밀도로 유지하고, 고주파 성분만 공격적으로 압축하는 하이브리드 전략이 품질 유지에 기여했음을 보여준다. 즉, 주파수 분해를 통해 정보의 중요도를 사전에 반영하고, 구조적 정보가 집중된 성분의 손실을 줄임으로써 저비트 환경에서도 품질 저하를 제한할 수 있었다. 그림 8의 고해상도 생성 결과에서도 입력 프롬프트에 대응하는 전반적인 장면 정합성이 유지되는 것을 확인할 수 있다. 따라서 제안한 방법은 few-step 고해상도 확산 모델에서 이미지 품질, 추론 속도, 메모리 효율성 사이의 실용적인 절충점을 제시하며, 기존 PTQ 방식의 한계를 완화할 가능성을 시사한다.

(a) Without Quantization (FP32) (b) With Quantization (INT4) (c) With Quantization (DWT+INT4)

그림 7. 생성 이미지의 시각적 품질 비교

Fig. 7. Comparison of visual quality of generated images

(Prompt: "beautiful scenery nature glass bottle landscape, purple galaxy bottle, snowy forest in the background", "an anime illustration of a wiener schnitzel")

(a) Without Quantization (FP32)

(b) With Quantization (DWT+INT4)

그림 8. 2048x1024 크기 이미지 품질 비교

Fig. 8. Quality comparison of 2048x1024 image

(Prompt: "cherry blossom scenery, dreamy spring korean landscape, soft sunlight filtering through branches, romantic pastel color palette, enchanting garden path")

V. 결론 및 고찰

이상의 결과는 few-step 확산 PTQ의 핵심 어려움이 단순한 저비트 표현 자체에 있는 것이 아니라, 이상치에 따른 양자화 범위 왜곡, 스킵 경로의 구조 정보 손실, 그리고 레이어별 민감도 불균형이 복합적으로 작용하는 데 있음을 시사한다. 제안 방법은 이 세 문제를 함께 고려함으로써, 고해상도 few-step 확산 모델의 실용적 경량화 가능성을 보였다. 또한 본 연구에서는 고해상도 확산 모델의 추론 효율을 극대화하기 위해 웨이블릿 변환 기반의 하이브리드 저비트 양자화 기법을 제안하고 그 기술적 타당성을 검증하였다. 기존 대규모 언어 모델(LLM)에서 유효했던 SmoothQuant 기법은 확산 모델 특유의 불규칙하고 비균일한 활성값 분포로 인해 수치적 안정성을 잃고 급격한 이미지 열화를 초래하는 한계를 보였으나, 본 논문에서 제시한 가중치 중심의 평활화 전략은 이러한 분포의 불균형을 해소하며 양자화 난이도를 효과적으로 낮추는 대안이 되었다.

여기에 웨이블릿 다중 해상도 분해를 결합하여 정보의 중요도에 따라 비트 정밀도를 차등 할당함으로써, 4비트와 같은 초저비트 환경에서도 이미지의 구조적 왜곡을 최소화하고 세밀한 디테일을 보존할 수 있었다. 결과적으로 제안된 기법을 SDXL-Lightning 4-step 모델에 적용했을 때, FP32 기준 모델 대비 피크 메모리를 약 5.4GB 수준까지 낮추어, 더 낮은 메모리 환경에서의 구동 가능성을 시사하였다.

이러한 성과는 이미지 품질과 연산 효율 사이에서 실용적인 절충점을 제시하였다는 점에서 의의가 크며, 후속 연구를 통해 확산 모델의 경량화를 한층 더 진전시켜 엣지 디바이스 환경에서도 고품질 이미지 생성을 실시간으로 구현하는 데 기여할 수 있을 것으로 기대된다.

감사의 글

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2026년도 문화체육관광 연구개발사업으로 수행되었음(연구개발과제명: 공연 콘텐츠의 고해상도(8K/16K) 서비스를 위한 AI 기반 영상확장 및 서비스 기술 개발, 연구개발과제번호: RS-2024-00395886).

참고문헌

[1] L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, ... and M.-H. Yang, "Diffusion Models: A Comprehensive Survey of Methods and Applications," arXiv:2209.00796, December 2024. https://doi.org/10.48550/arXiv.2209.00796

[2] H. Chang, H. Shen, Y. Cai, X. Ye, Z. Xu, W. Cheng, ... and H. Guo, "Effective Quantization for Diffusion Models on CPUs," arXiv:2311.16133, November 2023. https://doi.org/10.48550/arXiv.2311.16133

[3] D. Du, G. Gong, and X. Chu, "Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey," arXiv:2405.00314, May 1, 2024. https://doi.org/10.48550/arXiv.2405.00314

[4] T. Dettmers, M. Lewis, Y. Belkada, and L. Zettlemoyer, "LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale," in *Proceedings of the 36th International*

Conference on Neural Information Processing Systems (NeurIPS), New Orleans: LA, pp. 30318-30332, 2022. https://doi.org/10.48550/arXiv.2208.07339

[5] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," in *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS)*, Vancouver, Canada, pp. 6840-6851, 2020. https://doi.org/10.48550/arXiv.2006.11239

[6] Q. Zeng, C. Hu, M. Song, and J. Song, "Diffusion Model Quantization: A Review," arXiv:2505.05215, 2025. https://doi.org/10.48550/arXiv.2505.05215

[7] Y. Yang, X. Dai, J. Wang, P. Zhang, and H. Zhang. "Efficient Quantization Strategies for Latent Diffusion Models," arXiv:2312.05431, December 2023. https://doi.org/10.48550/arXiv.2312.05431

[8] X. Li, Y. Liu, L. Lian, H. Yang, Z. Dong, D. Kang, ... and K. Keutzer, "Q-Diffusion: Quantizing Diffusion Models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 17489-17499, 2023. https://doi.org/10.48550/arXiv.2302.04304

[9] L. Chen, Y. Meng, C. Tang, X. Ma, J. Jiang, X. Wang, ... and W. Zhu, "Q-DiT: Accurate Post-Training Quantization for Diffusion Transformers," arXiv:2406.17343, June 2024. https://doi.org/10.48550/arXiv.2406.17343

[10] S. Lin, A. Wang, and X. Yang, "SDXL-Lightning: Progressive Adversarial Diffusion Distillation," arXiv:2402.13929 2024. https://doi.org/10.48550/arXiv.2402.13929

[11] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in *Proceedings of the 18th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, Munich, Germany, pp. 234-241, 2015. https://doi.org/10.48550/arXiv.1505.04597

[12] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, ... and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. https://doi.org/10.48550/arXiv.2307.01952

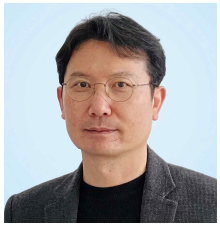
[13] G. Xiao, J. Lin, M. Seznec, H. Wu, J. Demouth, and S. Han, "SmoothQuant: Accurate and Efficient Post-Training Quantization for Large Language Models," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, Honolulu: HI, pp. 38087-38099, 2023. https://doi.org/10.48550/arXiv.2211.10438

[14] T. Zhao, X. Ning, T. Fang, E. Liu, G. Huang, Z. Lin, ... and Y. Wang, "MixDQ: Memory-Efficient Few-Step Text-to-Image Diffusion Models with Metric-Decoupled Mixed Precision Quantization," in *Proceedings of the 18th European Conference on Computer Vision (ECCV)*, Milan, Italy, pp. 285-302, 2024. https://doi.org/10.48550/arXiv.2405.17873

[15] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial Diffusion Distillation," arXiv:2311.17042, January 2024. https://doi.org/10.48550/arXiv.2311.17042

[16] H. Phung, Q. Dao, and A. Tran, "Wavelet Diffusion Models Are Fast and Scalable Image Generators," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Vancouver, Canada, pp. 10199-10208, 2023. https://doi.org/10.48550/arXiv.2211.16152

[17] Q. Sun, Y. Ren, L. Jiao, X. Li, F. Shang, and F. Liu, "MWQ: Multiscale Wavelet Quantized Neural Networks," arXiv:2103.05363, 2021. https://doi.org/10.48550/arXiv.2103.05363



서용석(Yongseok Seo)

2001년 : 영남대학교 정보통신공학과 (공학석사)
2009년 : 충남대학교 컴퓨터공학과 (공학박사)

2001년～현 재: 한국전자통신연구원 책임연구원
※관심분야 : 딥러닝, 컴퓨터 비전, 디지털 콘텐츠 보호기술

임동혁(Dong-Hyuck Im)

2006년 : 한국과학기술원 전산학과 (공학석사)
2009년 : 한국과학기술원 전산학과 (공학박사)

2009년～2012년: KT 선임연구원
2012년～현 재: 한국전자통신연구원 책임연구원
※관심분야 : 딥러닝, 컴퓨터 비전, 디지털 콘텐츠 보호기술