LLM-DiT Deep Fusion 모델의 다양성 향상을 위한 코사인 유사도 기반 제어 방안 연구

신기동

대전대신고등학교(Daejeon Daeshin High.) A.C.T.(KE)

ABSTRACT: 최근 텍스트-이미지 생성 분야에서 대형 언어 모델(LLM)과 확산 트랜스포머(DiT)를 깊게 결합하는 'Deep Fusion' 아키텍처가 주목받고 있습니다. 이 접근법은 기존의 얕은 조건화 방식보다 풍부한 텍스트-이미지 정렬을 가능하게 하지만, 동일 조건에서 생성된 결과물 간의 다양성 부족 문제는 여전히 중요한 과제로 남아있습니다. 본 연구는 선행 연구(CVPR 2025)에서 규명된 Deep Fusion 의 구조적 이점을 바탕으로, '코사인 유사도'를 핵심 제어 신호로 활용하여 샘플 간 유사성을 정량화하고 제어하는 방법을 탐구합니다. 훈련 중 다양성 손실 적용, 텍스트-이미지 정렬 강화, 후처리 샘플 선택 등 다각적인 접근을 통해, 생성 모델의 텍스트 조건 일치도를 유지하면서 동시에 결과물의 다양성을 실질적으로 개선하는 것을 목표로 합니다. 본 보고서는 이러한 목표 달성을 위한 구체적인 개념 정의, 탐구 내용, 실험 설계 및 단계별 실행 계획을 기술합니다.

I. 서론

최근 텍스트 -> 이미지 생성에서 확산 모델과 대형 언어모델을 결합하는 연구가 활발히 진행되며, 특히 LLM 과 Dit(Diffusion Transformer)을 깊게 결합하는 설계가 주목받고 있다. 기존의 얕은(한 번에 인코딩된 텍스트 표현만 주입하는) 조건화 방식은 복잡한 문장 의미, 구성, 맥락을 확산 과정에 충분히 전달하지 못하는데, LLM 의 풍부한 내부 표현을 DiT 의 여러 레이어에 계층적으로 주입하면 텍스트-이미지 정렬형, 복합 명령 수행력, 구성적 표현 능력등을 실질적으로 끌어오릴 수 있어서 관심을 받고 있다.

또한 선행 연구(CVPR 2025)는 LLM 내부 표현을 DiT 와 계층적으로 공유하는 'deep fusion' 접근을 체계적으로 비교, 분석하고 이 아키텍처가 텍스트-이미지 생성에서 갖는 가능성을 규명했다는 점에서 이 연구는 단순 교체로는 성능 향상이 어렵고, LLM 의 학습 목표와 확산 모델의 조건화 방식 사이에 오차가 있으므로 세심한 설계가 필요하다는 점을 인지하였다.

한편, 생성모델의 또 다른 중요한 문제는 동일한 조건에서 생성한 결과들이 과도하게 닮아 다양성을 잃는다는 점이다. 따라서 이번 연구를 통해 LLM 과 DiT 의 깊은 결합이 제공하는 풍부한 다중 모델을 활용하여 코사인 유사도를 핵심 신호로 하여 샘플간 유사성을 정량화, 제어함으로써 텍스트 조건 하에서의 다양성과 조건 일치를 동시에 개선하는 방법을 탐구하고자 한다.

II. 이론적 배경

2.1. 확산 모델 (Diffusion Models / DiT)

확산모델은 점진적 노이즈 추가, 제거 과정을 통해 데이터 분포를 모델링하는 생성 프레임 워크로, 최근의 DiT 는 이미지 토큰을 Transformer 로 처리해 고품질 이미지를 생성한다. 하지만, DiT 와 LLM 을 레이어 단위로 결합하는 deep fusion 설계를 중심으로, 텍스트-이미지 간의 풍부한 상호작용을 실험적으로 탐구했다.

2.2. 대형언어 모델(LLM)과 Deep Fusion

단순히 LLM 을 텍스트 인코더로 치환하는 것은 기대만큼의 성능 향상을 주지 못한다는 관찰이 있다. 대신 LLM 내부를 DiT 의 각 레이어와 깊게 공유하면 LLM 의 문맥적 표현 능력을 더 자연스럽게 확산 과정에 주입할 수 있다는 아이디어가 제안되었다. 논문은 특히 'layer-wise shared self-attention' 같은 기법으로 두 스트림을 결합하는 설계와 훈련 레시피를 체계적으로 비교-제시한다.

2.3. 코사인 유사도 (Cosine Similarity)

코사인 유사도는 두 특징 벡터의 방향(각도) 일치도를 측정하는 비표로, 같은 방향일수록 값이 커지며 최대 1, 최소 -1 이 된다. 생성 모델 맥락에서는 동일 조건에서 생성한 샘플들의 특징을 비교하여 샘플들이 얼마나 닮았는가를 정량화하는데 유용하다. 또한, 코사인을 평균/상위퍼센테일로 집계하면 해당 조건에서의 중복성을 모니터링 및 제어할 수 있다.

III. 연구 내용 및 방법

3.1. 연구 목표

본 탐구의 최종 목표는 LLM 과 DiT 를 깊게 결합한 생성 시스템에서 코사인 유사도를 도구로 활용하여, 같은 지시문으로 생성된 이미지들이 프롬프트와 의미적으로

일치하면서도(품질) 서로 다양하도록(다양성) 만드는 것이다. 이를 위해 코사인 유사도를 적절히 조정하여 품질과 다양성 간의 균형점(trade-off)을 찾고, 사람이 인지하기에 두 측면을 모두 만족시키는 최적의 방안을 도출하고자 한다.

(코사인 유사도 계산: 한 조건으로 모델이 동일 미니배치에서 여러 샘플을 생성 또는 후보들을 예측할 때, 각 샘플들의 시각적 특징을 추출하고 정규화 후 코사인 행렬을 계산한다.)

3.2. 코사인 유사도 적용 방안

코사인 신호를 생성 모델에 적용할 수 있는 주요 포인트는 다음과 같다.

1. 훈련 중 다양성 제어 (Training-time diversity regularization) 한 조건(prompt)으로 모델이 동일 미니배치에서 여러 샘플을 생성할 때, 각 샘플의 시각적 특징(예: CLIP image-encoder 임베딩)을 추출하여 코사인 행렬을 계산한다. 평균(또는 상위 k%) 코사인이 높으면 벌점을 주는 손실항을 원래의 denoising 손실에 더한다. 이로써 생성 과정에서 샘플 간 각도를 벌리는 방향으로 파라미터가 업데이트된다.

2. 조건 정렬 강화 (Text–image alignment via cosine) 텍스트 임베딩(LLM 또는 CLIP text)과 생성 이미지 임베딩(CLIP image)의 코사인을 최대화하도록 보조 손실을 적용하여 프롬프트-일치도를 높일 수 있다. Deep fusion 아키텍처의 경우 LLM 의 내부 표현과 생성 이미지 표현 간의 일관성을 코사인으로 직접 측정·강화할 수 있다.

3. 후처리 샘플 선택 (Post-hoc diverse sampling) 대량(예: M 개)의 후보 샘플을 생성하고, CLIP 임베딩 기준으로 pairwise 코사인을 계산한 뒤, greedy/최적화 알고리즘으로 서로 닮지 않은 k 개의 샘플을 골라내 사용자에게 제시한다. (훈련 부담 없이 다양성 보장 가능)

4. 잠재(노이즈) 공간 규제 동일 프롬프트에서 사용하는 노이즈/시드 벡터들의 잠재 임베딩도 코사인으로 제어(입력 단계에서 다양화)하면, 간접적으로 출력 다양성에 기여할 수 있다.

3.3. 실험 설계

본 연구의 가설을 검증하기 위한 실험 설계는 다음과 같다.

1.베이스라인 (Baseline)

Deep fusion(논문에서 제시된 레시피) 또는 Shallow fusion 모델

기존 LDM/DiT 기반 텍스트-이미지 모델

2.실험군 (Experimental Groups)

(1) Training-time cosine diversity loss 추가

(2) Text–image cosine alignment loss 추가

(3) Post-hoc selection 만 적용

(4) 위 방식들의 조합

3.측정 지표 (Metrics)

Diversity metrics: Intra-prompt mean cosine (낮을수록 다양성↑), LPIPS 분산, Recall (generative recall)

Quality metrics: FID, CLIP-score (text–image similarity), Human perceptual rating(주관적 평가)

Trade-off 관찰: Mean cosine 과 FID/CLIP-score 동시 모니터링

4.하이퍼파라미터 (Hyperparameters)

다양성 손실 가중치 λ: 시작값 0.01~0.05

배치당 샘플 수 (프롬프트별): 최소 4~8 이상 권장 (코사인 평균 안정화 위해)

Top-k percent (후처리): 5~20% 실험

5.계산/효율화 모든 쌍 계산은 $O(N^2)$이므로 큰 배치는 쌍 샘플링·마스크·메모리뱅크로 최적화한다. Deep fusion 의 구조적 비용(LLM 파이프라인) 때문에 fine-tune 단계에서만 다양성 항을 켜는 접근이 현실적일 수 있다.

3.4. 구현 시 고려사항

실무적 구현 시 다음 사항들을 고려한다.

1.특징 선택: CLIP 임베딩은 텍스트-이미지 정렬을 통합적으로 반영하므로 우선적 선택지이다. 시각적 세부/질감 차이를 잡고 싶으면 DiT 의 mid-layer 나 VGG 퍼셉추얼 특징을 병용한다.

2.학습 스케줄: 초반에는 $\lambda=0$으로 워밍업(모델 안정화) 후 λ를 점진 증가시키는 스케줄을 권장한다.

3.로깅: Epoch 별 mean_cosine, FID, CLIP-score, 샘플 격자(정성평가)를 함께 기록해 트레이드오프를 시각화한다.

4.평가 디자인: 자동 지표 외에 인간 평가(특히 다양성의 '의미성' 여부)를 반드시 포함한다. (다양성 증가는 곧바로 '유의미한 차이'를 의미하지 않을 수 있음)

IV. 결론 및 향후 계획

4.1. 결론

LLM 과 DiT 의 깊은 결합(deep fusion)이 제공하는 풍부한 표현력을 활용하여, 코사인 유사도를 핵심 제어 신호로 삼아 텍스트 조건 하의 다양성·정렬 문제를 동시에 다루는 것을 목표로 한다. CVPR 논문은 deep fusion 이 텍스트-이미지 생성에서 잠재적으로 강력한 방법임을 실험적으로 보여주었고, 본 연구는 그 구조적 이점 위에 '코사인 기반 다양성/정렬 제어'를 더해 실용적 품질·다양성 균형을 개선하고자 한다.

4.2. 기대 효과

Deep fusion 의 고차원적 텍스트 이해 능력을 보존하면서, 코사인 기반의 직접적 유사성 제어로 "같은 프롬프트에 대해 더 다양하면서도 프롬프트에 충실한" 생성 결과를 얻을 수 있다. 이는 창작 도구·콘텐츠 생산·데이터 증강 등 실무적 응용에서 사용자 선택 폭을 넓히고 반복적 유사 결과로 인한 활용 한계를 줄이는 데 기여할 것이다.

4.3. 단계별 실행 계획

1.단기 (프로토타입): 공개된 DiT + frozen LLM 기반 레시피(논문 레시피 재현)를 로컬/클라우드에서 재현한 뒤, 동일 프롬프트에서 k 샘플을 생성하여 CLIP 기반 mean_cosine 을 계산·로그한다.

2.중기 (요인 실험): Training-time diversity loss(코사인 기반)와 text–image alignment loss 를 각각·병용 적용해 FID 와 mean_cosine 변화를 비교. λ 스윕·batch size 스윕 수행.

3.장기 (확장): 사용자 주관평가(AMT 등) 및 downstream 적용(예: 콘텐츠 생성 파이프라인)에 통합하고, LLM–DiT 의 layer-wise fusion 세부 설계(어느 레이어를 결합해야 효과적인가)를 추가로 최적화한다. 또한 post-hoc selection 알고리즘을 서비스 수준에서 병행 적용해 실제 배포 시 다양성-응답성 균형을 확보한다.

참고 문헌

본 기획에서는 CVPR 2025 논문을 핵심 동기·설계 근거로 사용했으며, 논문에서 제공하는 실험 레시피와 관찰(예: deep fusion 의 layer-wise 결합, 대규모 재현 실험 등)을 기반으로 실험 설계를 제안했다. 자세한 논문 내용(아키텍처·훈련 세부)은 원문을 직접 참조하면 구현·재현에 큰 도움이 된다.

- Tang, et al. (2025). "Exploring the Deep Fusion of Large Language Models and Diffusion Transformers for Text-to-Image Synthesis". CVPR 2025. https://openaccess.thecvf.com/content/CVPR2025/papers/Tang_Exploring_the_Deep_Fusion_of_Large_Language_Models_and_Diffusion_CVPR_2025_paper.pdf