Next-Scale 인버전 기반 레퍼런스-가이드 T2I 스타일 전이 연구

오민석, 순동현, 이상철†
대구경북과학기술원

harrymark0@dgist.ac.kr, dhsoon@dgist.ac.kr, sangchul.lee@dgist.ac.kr

Next-Scale Inversion Based Reference-Guided T2I Style Transfer

Oh Min Seok, Soon Dong Hyeon, Sang Chul Lee†
DGIST

요 약

텍스트-투-이미지(Text-to-Image, T2I) 생성에서 스타일 전이(Style Transfer)는 레퍼런스 이미지가 지닌 색감, 조명 분위기 등 시각적 양식을 유지하면서, 텍스트 프롬프트가 요구하는 새로운 객체나 구성을 생성하는 문제로 볼 수 있다. 기존 확산모델 기반 접근은 강력하나 다단계 샘플링으로 인해 추론 비용이 높거나, 스타일 개인화(Personalization)를 위해 추가 학습이 요구되는 경우가 많다. 본 연구는 비트 단위 시각 오토리그레시브(Visual AutoRegressive, VAR) 모델인 *Infinity*를 베이스라인으로 채택하고, 다음 스케일(next-scale) 오토리그레시브 샘플링에서의 노이즈 인버전을 이용해 학습 없이(training-free) 레퍼런스 스타일을 주입하는 T2I 스타일 전이 파이프라인을 제안한다. 정성·정량 평가를 통해 제안한 모델이 프롬프트 정합성을 유지하면서 레퍼런스 스타일 유사도를 향상시키며, 오토리그레시브 모델의 확산모델 대비 빠른 추론의 장점을 유지함을 보인다.

Ⅰ. 서 론

T2I 스타일 전이는 텍스트가 지정한 객체와 구성을 생성하는 내용 충실도와, 레퍼런스 이미지의 시각적 양식을 반영하는 스타일 일관성을 동시에 요구한다. 최근 확산모델은 고품질 생성의 표준이 되었지만, 다단계 샘플링은 속도, 비용 측면에서 부담이 크다. 반면 오토리그레시브 계열은 토큰을 순차 예측하며 빠른 추론 성능을 보였으며, 특히 [1] 는 비트 단위 토큰화 및 self-correction 으로 고해상도 생성 성능을 달성했다. 한편, 확산모델 편집 에서 널리 쓰이는 노이즈 인버전 개념은 최근 VAR 계열에도 확장되어, next-scale 오토리그레시브 샘플링에서의 노이즈를 복원해 소스 이미지 재구성과 텍스트 기반 편집을 가능케 하는 기법이 제안되었다 [2]. 그러나 이러한 next-scale 인버전은 주로 소스 이미지 보존형 편집에 초점이 맞춰져 있으며, 레퍼런스 스타일 이미지를 인버전해 T2I 생성 과정에 스타일을 주입하는 스타일 전이로의 체계적 적용은 상대적으로 덜 탐구되었다. 본 연구의 목표는 스타일 레퍼런스 이미지 한 장만으로 추가 학습 없이 T2I 생성에서 스타일을 반영하도록 만드는 것이다.

Ⅱ. 본론

2.1 문제 정의 및 사전 배경 지식

본 연구는 레퍼런스-가이드 T2I 스타일 전이를 메인 태스크로 정의한다. 스타일 레퍼런스 이미지 I_s와 텍스트 프롬프트 p_t가 주어졌을 때, 출력 I_{out}은 p_t가 요구하는 새로운 객체·구성을 생성하면서도, I_s의 시각적 양식을 최대한 유지해야 한다. 또한 본 연구는 LoRA 나 파인튜닝과 같은 추가 학습 없이 사전학습된 모델만으로 이를 달성하는 것을 목표로 한다. 기존 next-scale 오토리그레시브 인버전 계열은 주로 소스 이미지 보존형 편집을 목표로 하여, 동일한 이미지의 구조와 배경을 유지한 채 일부 속성만 변경하는 설정에 초점을 맞춘다. 반면 본 연구는 인버전의 목적을 소스 복원이 아니라 레퍼런스 스타일 추출로 재정의한다. 즉, 레퍼런스 스타일 이미지에서 복원된 역노이즈가 그 스타일을 선택하게 만든 샘플링 요인을 담는다고 보고, 이를 타깃 프롬프트 기반 생성에 주입함으로써 스타일 전이를 수행한다. 베이스라인인 Infinity 는 멀티스케일 토큰 예측을 수행한다. 스케일 l 에서의 토큰을 $r^{(l)} \in \{0,1\}^{H_l W_l \times D}$ 로 두면, 모델은 텍스트 조건 c 하에서 각 위치의 비트 분포를 로짓 $\mathrm{l}^{(l)}$ 로 출력한다. 비트 단위로 정리하면 $V = 2D$ 이고, 로짓은 $\mathrm{l}^{(l)} \in R^{B\times(H_l W_l \cdot D)\times 2}$. 샘플링은 Gumbel-max 트릭으로 표현할 수 있다.

$$r^{(l)} = \arg\max_{b\in\{0,1\}} \left(\frac{1}{\tau_l}\mathrm{l}_b^{(l)} + g_b^{(l)}\right),$$

$$g_b^{(l)} \sim \mathrm{Gumbel}(0,1),$$

여기서 τ_l 은 스케일별 온도(temperature)이다.

2.2 전체 파이프라인

그림 1 은 제안하는 방법의 전체 파이프라인을 보여준다. 본 연구의 핵심은 next-scale 인버전을 이미지 편집을 위한 복원 도구가 아니라, 레퍼런스 스타일을 추출해 주입하는 스타일 캐리어 생성기로 활용한다는 점이다. 파이프라인은 크게 스타일 레퍼런스 인버전 단계와 타깃 T2I 생성 단계로 구성된다. 먼저 인버전 단계에서는 레퍼런스 스타일 이미지I_s 를 토크나이저로 인코딩하여 스케일별 레퍼런스 토큰$\widehat{r_s^{(l)}}$ 를 얻는다. 이후 teacher forcing 으로 $\widehat{r_s^{(l)}}$ 에 대한 로짓 $\mathrm{l}^{(l)}(c_s)$ 를 계산하고, $\widehat{r_s^{(l)}}$ 를 선택하게 만드는 역 Gumbel 노이즈 $g_{inv}^{(l)}$를 구성한다. 이때 teacher forcing 은 정답 토큰을 이미 알고 있는 상황에서 로짓을 안정적으로 얻기 위한 절차이다.
다음으로 생성 단계에서는 타깃 프롬프트 p_t로부터 조건 c_t 를 구성하고 스케일별 로짓 $\mathrm{l}^{(l)}(c_t)$ 를 계산한다. 이후 $g_{inv}^{(l)}$를 스케일별 가중치로 혼합하여 샘플링에 주입하고,

오토리그레시브 샘플링 및 디코딩을 통해 최종 이미지를 얻는다.

2.3 레퍼런스 인버전 기반 T2I 스타일 전이 생성

레퍼런스 스타일 이미지 I_s를 토크나이저로 인코딩하여 스케일별 토큰 $r_s^{(l)}$를 얻는다. 인버전의 목표는 특정 조건 c_s하에서 다음을 만족하는 역 노이즈 $g_{inv}^{(l)}$를 구성하는 것이다.

$$\widehat{r_s^{(l)}} = \arg\max\left(\frac{1}{\tau_l} \mathrm{l}^{(l)}(c_s) + g_{inv}^{(l)}\right).$$

기존 편집 관점에서 g_{inv} 가 원본을 재현하기 위한 샘플링 요인으로 해석되지만, 본 연구는 이를 확장하여 g_{inv}를 레퍼런스의 스타일 성분을 담는 스타일 캐리어로 해석한다. 따라서 g_{inv} 를 타깃 생성 단계에 주입하면 레퍼런스 스타일을 따르는 방향으로 샘플링이 유도될 수 있다. 모델은 타깃 프롬프트 p_t 로부터 조건 c_t 를 얻고, 스케일 l 의 로짓 $\mathbf{l}^{(l)}(c_t)$ 를 계산한다. 이후 레퍼런스 역노이즈 $g_{inv}^{(l)}$ 를 새 Gumbel 노이즈 $g_{new}^{(l)}$ 와 혼합한다:

$$g_{edit}^{(l)} = (1-\alpha_l) g_{new}^{(l)} + \alpha_l g_{inv}^{(l)},$$

$$r_{out}^{(l)} = \arg\max\left(\frac{1}{\tau_l} \mathrm{l}^{(l)}(c_t) + g_{edit}^{(l)}\right).$$

여기서 $\alpha_l \in [0,1]$ 는 스케일별 스타일 주입 강도이다. 중요한 점은 α_l 을 모든 스케일에서 일괄적으로 크게 두는 것이 아니라, 스타일 요소가 상대적으로 강하게 나타나는 스케일 구간에 더 큰 가중치를 부여하도록 설계하는 것이다. 예를 들어 낮은 스케일에서는 레퍼런스의 전역 스타일을 우선 반영하고, 중간·상위 스케일에서 프롬프트 기반의 객체 및 구조가 반영되도록 α_l을 스케줄링할 수 있다.

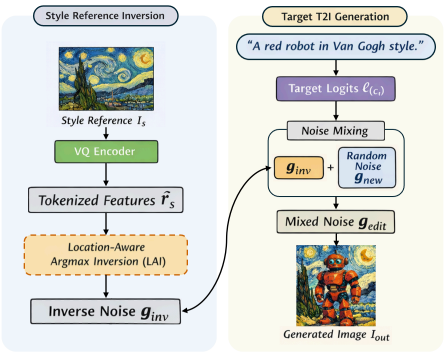

그림 1. 제안한 모델의 전체 파이프라인

2.4 정량적 비교 및 실험 결과

본 절에서는 연구에서 제안한 모델을 기존 대표 방식(State-Of-The-Art, SOTA)들과 정량 비교한다. 비교 대상은 (i) 학습 기반 개인화 계열(B-LoRA), (ii) 레퍼런스 조건부 어댑터 계열(IP-Adapter)로 구성하였다. 평가지표는 스타일 유사도 S_{style}, 텍스트 정합도 S_{text}, 추론 시간 T_{infer}로 정의한다. 스타일 유사도는 레퍼런스 이미지와 출력 이미지 의 특징을 DINO 인코더로 추출한 뒤 코사인 유사도로 계산한다. 텍스트 정합도는 CLIP 의 이미지·텍스트 임베딩을 이용하여 코사인 유사도로 측정한다. 마지막으로 추론 시간 T_{infer}는 동일한 설정 아래에서 한 장 생성 평균 시간(초)으로 측정하였다.

표 1. 기존 SOTA 모델과의 정량적 비교

Model	S_{style}(DINO)	S_{text}(CLIP)	T_{infer}(sec)	Training-free
B-LoRA	0.271	0.259	633.20	✗
IP-Adapter	0.522	0.278	10.14	✗
OURS	**0.610**	**0.285**	**1.15**	✓

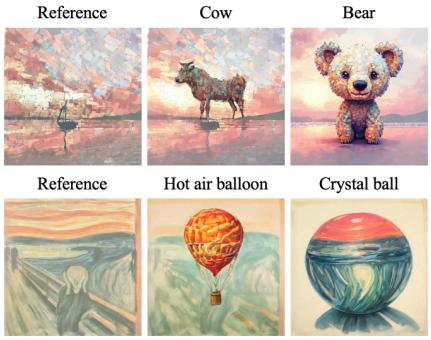

그림 2. 제안한 모델의 정성적 결과

표 1 은 정량 결과를 보여준다. 제안법은 스타일 유사도 S_{style}(DINO) 에서 높은 값을 보이며, 텍스트 정합도 S_{text}(CLIP) 또한 경쟁 방법들보다 더 높은 수준을 유지함으로써 스타일 주입으로 인해 프롬프트 반영이 무너지는 현상을 완화함을 확인하였다. 또한 추론 시간 측면에서 제안법은 오토리그레시브 기반 Infinity 의 빠른 생성 흐름 위에서 인버전-주입이 학습 없이 수행되기 때문에, 학습 기반 개인화 방법 대비 크게 빠른 경향을 보인다. 그림 2 의 정성 결과는 새 객체가 생성됨에도 레퍼런스의 스타일이 유지되는 사례를 제시하며, 이는 레퍼런스에서 복원한 $g_{inv}^{(l)}$ 를 스케일별로 주입함으로써 스타일 성분이 전달되기 때문임을 시사한다.

Ⅲ. 결론

본 연구는 Infinity 기반 next-scale 오토리그레시브 생성에서, 스타일 레퍼런스 이미지를 인버전하여 얻은 역 Gumbel 노이즈를 타깃 프롬프트 생성에 혼합함으로써, 학습 없이 스타일 전이 T2I 를 수행하는 간단한 방법을 제안하였다. 제안법은 확산모델 대비 빠른 추론 흐름을 유지하며, 레퍼런스 스타일 일관성과 텍스트 정합성을 동시에 만족하는 생성 결과를 제공한다. 향후 연구로는 공간적 가중치 $\alpha(x)$ 를 이용한 영역별 스타일 주입, 다중 레퍼런스 인버전, 인버전 노이즈의 일반화 성질 분석 및 이론적 해석을 진행할 수 있다.

ACKNOWLEDGMENT

본 연구성과는 과학기술정보통신부에서 지원하는 DGIST 기관 고유사업(26-ET-02)과 산업통상자원부의 재원으로 한국산업기술진흥원의 지원을 받아 수행된 연구임(No. RS-2024-02633871)

참 고 문 헌

[1] J. Han et al., "Infinity: Scaling Bitwise AutoRegressive Modeling for High-Resolution Image Synthesis," CVPR, 2025.

[2] Q. Dao et al.,"Discrete Noise Inversion for Next-scale Autoregressive Text-based Image Editing," arXiv,2025.