Text-to-ASCII: 대형 언어 모델 기반 구조적 시각 표현 생성

조성환, 정의림, 박천음*

국립한밭대학교

josseong1227@gmail.com, erjeong@hanbat.ac.kr, *parkce@hanbat.ac.kr

Structural Visual Generation with Large Language Models

Seong-Hwan Jo, Euirim Jeong, Cheoneum Park*

Hanbat National University

요 약

기존 확산(Diffusion) 기반의 텍스트-투-이미지(Text-to-Image, T2I) 모델은 구조적 수정이 필요할 때마다 전체 이미지를 재생성해야 하는 한계를 가진다. 이로 인해 반복적인 레이아웃 조정 과정에서 연산 비용과 메모리 소모가 증가하는 문제가 발생한다. 이를 해결하기 위해 본 논문은 이미지 생성 이전 단계에서 시각적 구조를 아스키 문자 기반의 텍스트 시퀀스로 표현하는 방법을 제안한다. 사전 학습된 텍스트-투-이미지 모델을 이용하여 이미지를 생성한 뒤, Canny Edge 기반 구조 추출 기법을 적용한다. 추출된 구조 정보는 Random Forest(RF) 분류기를 통해 구조 중심의 아스키 표현으로 변환된다. 생성된 아스키 표현에는 공간 기반 시퀀스 압축과 k-mers 기반 토큰화를 적용한다. 아스키 시퀀스 생성을 위해 대형 언어 모델(Large Language Model, LLM)에 LoRA(Low-Rank Adaptation) 기반 미세 조정을 적용한다. 이를 통해 입력 텍스트에 대응하는 구조적 아스키 표현을 생성하도록 모델을 학습한다.실험에서는 동일한 이미지 생성 모델을 사용하여 아스키 구조 조건의 유무에 따른 생성 결과를 비교한다. 정량 및 정성 평가 결과, 제안한 방법은 텍스트 프롬프트만을 사용하는 기존 방식 대비 구조적 일관성과 캡션 정합도가 향상됨을 확인하였다.

I. 서론

최근 확산 기반의 T2I 모델의 발전으로 텍스트 입력으로부터 고품질 이미지를 생성하는 것이 가능해졌으며, 이러한 기술은 창작, 디자인, 콘텐츠 제작 등 다양한 분야로 활용 범위를 넓혀가고 있다. 그러나 실제 활용 환경에서는 사용자가 의도한 결과를 얻기까지 프롬프트를 반복적으로 수정해야 하는 경우가 많다.

일반적인 T2I 모델은 변경이 요구될 경우 이미지 전체를 다시 생성하는 방식에 의존한다. 이로 인해 반복적인 샘플링 과정이 필요하며, 결과적으로 연산 자원과 메모리 소모가 증가하는 문제가 발생한다. 특히 초기 아이디어 구상 단계나 레이아웃을 빈번하게 조정하는 작업 환경에서는 이러한 특성이 시간적·비용적 비효율로 이어질 수 있다 [1].

이에 본 논문은 이미지 생성 이전 단계에서 시각적 구조를 텍스트 형태로 표현하고,이를 LLM이 텍스트 수준에서 생성 및 조작할 수 있도록 하는 새로운 접근 방식을 제안한다. 아스키 아트(ASCII Art)는 텍스트 문자들의 조합을 통해 시각적 형상을 표현하는 방식으로, 이미지를 텍스트 기반의 구조적 표현으로 변환함으로써 2차원 공간 정보를 LLM이 처리 가능한 순차적 문자 시퀀스로 대응시킬 수 있다.

이러한 아스키 기반 구조 표현은 시퀀스 처리에 강점을 지닌 LLM의 특성과 높은 호환성을 보이며, 고비용의 이미지 생성 과정을 수행하지 않고도 텍스트 단계에서 구조의 수정과 탐색이 가능하다는 장점을 가진다. 이는 반복적인 수정이 요구되는 환경에서 연산 효율성을 개선할 수 있는 효과적인 대안으로 작용한다.

*교신 저자

이러한 관점에서 본 논문은 텍스트 프롬프트로부터 시각적 정합성을 갖춘 아스키 구조를 생성하는 *Text-to-ASCII* 과제를 제안한다. 이를 위해 사전 학습된 T2I 모델이 생성한 이미지를 구조적 아스키 시퀀스로 변환하여 학습 데이터셋을 구축하고, 정보 밀도를 고려한 시퀀스 압축 기법을 적용한다. 이후 LoRA 기반 미세 조정을 통해 입력 텍스트의 의미를 반영한 아스키 구조 생성 성능을 분석하고, 생성된 아스키 구조를 조건(Condition)으로 활용하여 T2I 모델이 의도된 구조를 갖는 이미지를 생성할 수 있는지를 검증한다.

II. 제안 방법

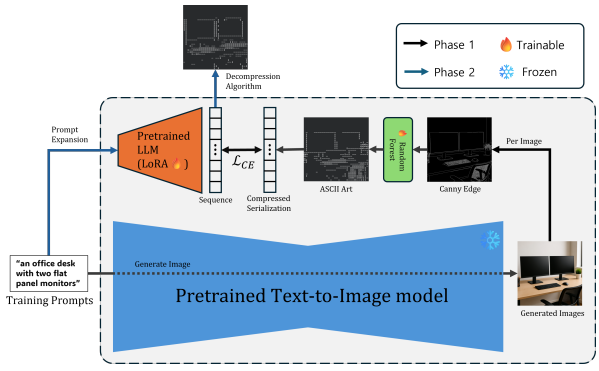

그림 1: 제안하는 Text-to-ASCII 프레임워크

본 논문에서는 텍스트 입력으로부터 구조적 시각 표현을 생성하는 Text-to-ASCII 프레임워크를 제안한다. 전체 구조는 [그림 1]과 같으며, 제안하는 방법은 T2I 모델을 활용한 아스키 아트 생성 단

계와 생성된 아스키 시퀀스를 학습하기 위한 LoRA 기반 파인튜닝 단계로 구성된다.

II.I. 구조 기반 아스키 이미지 생성

본 논문에서는 사전 학습된 T2I 모델을 이용하여 텍스트 프롬프트로부터 이미지를 생성하고, 이를 구조 중심의 아스키 표현으로 변환한다. 구체적으로, Canny Edge 검출을 통해 이미지의 주요 윤곽선과 형태 정보를 추출한 뒤, 이를 일정 크기의 그리드로 분할하고, 각 이미지 패치에 대해 학습된 RF 분류기를 적용하여 구조 중심의 아스키 이미지를 생성한다 [2].

이때, 생성된 아스키 이미지를 그대로 1차원 시퀀스로 변환할 경우 공백 문자가 다량 포함되어 시퀀스 길이가 불필요하게 증가한다. 이를 완화하기 위해 본 논문에서는 구조 정보가 존재하는 영역만을 선택적으로 표현하고, 위치 정보를 함께 인코딩하는 공간 기반 시퀀스 압축 방식을 적용한다.

II.II. LoRA 기반 파라미터 효율적 미세조정

본 논문에서는 아스키 아트 생성을 위해 LoRA를 활용한 파라미터 효율적 미세조정을 수행한다 [3]. 아스키 아트는 제한된 문자 집합으로 구성되며 연속적인 문자 패턴이 반복되는 특성을 가지므로, 이를 효과적으로 모델링하기 위해 k-mers 기법을 적용한다. 이를 위해 연속된 문자를 하나의 토큰 단위로 묶어 토큰화하고 이를 토크나이저의 신규 토큰으로 추가하여 학습을 수행한다 [4].

학습 과정에서LLM은 교차 엔트로피 손실을 통해 텍스트 캡션에 대응하는 압축 아스키 시퀀스 예측을 학습한다.

III. 실험 환경 및 결과

III.I. 실험 환경

제안 모델의 학습을 위해 사전 학습된 T2I 모델인 QWEN-image 기반의 DiffSynth-Studio Qwen-Image-Self-Generated-Dataset을 활용한다. 학습 데이터는 총 10,000장으로 구성되며, 검증 단계에서는 분포 내(in-distribution) 평가를 수행하기 위해 학습 도메인과 유사한 프롬프트 1,000개를 별도로 준비한다.

그 뒤에 이미지는 아스키 아트로 변환한다. 이때 아스키 아트에 사용하는 문자 집합은 `0`, `=`, `|`, 그리고 공백 문자로 제한한다. 구체적으로, 입력 이미지의 전체 에지 맵을 32×32 그리드로 분할한 뒤, 각 패치별로 학습된 RF 분류기를 통해 가장 적합한 아스키 문자를 예측하여 할당하고 압축하여 아스키 아트 시퀀스를 생성한다.

LLM 학습에는 `Qwen2.5-Coder-7B-Instruct` 모델을 기반으로 LoRA 튜닝을 수행한다.또한 출력 형식의 안정성을 확보하기 위해, 학습 과정에서 구조화된 시스템 프롬프트를 사용한다.

III.II. 정량적 성능 평가

제안한 Text-to-ASCII 프레임워크의 성능을 정량적으로 평가하기 위해, 이미지 생성 모델인 Juggernaut XL v9을 사용하되 입력 조건에 따른 생성 결과를 비교한다. 이를 위해 텍스트 프롬프트만을 사용하는 기본 T2I 방식과, 아스키 구조 표현을 조건으로 추가한 생성 방식을 비교한다. 평가 결과는 [표 1]에 정리한다.

표 1: 아스키 구조 조건 유무에 따른 텍스트-이미지 생성 성능 비교 (BLIP-2 임베딩 기반 코사인 유사도).

Generation Condition	Sim(I–T)↑	Sim(I–I)↑
Text Prompt Only (Baseline)	62.93	36.84
Text + ASCII Structure (Ours)	**65.33**	**38.12**

III.III. 정성적 비교 결과

[그림 2]는 동일한 텍스트 캡션에 대해 입력 조건에 따른 이미지 생성 결과를 비교한 예시를 보여준다.

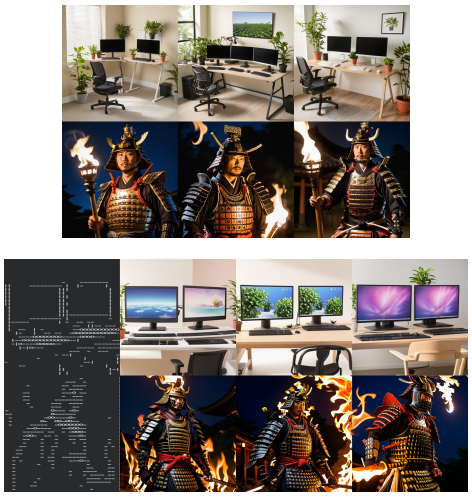

그림 2: (a) 텍스트 프롬프트만을 사용한 생성 결과. (b) 아스키 구조 표현을 조건으로 추가하여 생성한 결과.

첫 번째 이미지는 텍스트 프롬프트만을 사용한 기본 T2I 생성 결과로, 장면의 전반적인 의미는 반영되지만 객체의 배치와 구조적 일관성이 이미지마다 불안정하게 나타난다. 반면 두 번째 이미지는 제안한 파이프라인을 통해 학습된 LLM이 생성한 아스키 구조를 조건으로 사용한 결과로, 객체의 위치와 배열이 보다 안정적으로 유지되며 캡션과의 구조적 정합성이 향상된 것을 확인할 수 있다.

IV. 결론

본 논문은 프롬프트로부터 이미지의 구조적 정보를 반영한 아스키 아트를 생성하여 T2I 모델을 제어하는 Text-to-ASCII 프레임워크를 제안하였다. 실험 결과, 제안 방법은 기존 방식 대비 객체의 구조적 일관성을 안정적으로 유지하며 텍스트-이미지 간 정합성을 향상시킴을 확인하였다. 향후 아스키 표현의 해상도와 문자 집합을 확장하여 더욱 정교한 구조 제어를 수행하고, 적용하는 연구를 진행할 계획이다.

참 고 문 헌

[1] J. Huang, Z. Shen, H. Zhang, Z. Lin, and S. Cohen, "Diffusion model-based image editing: A survey," *arXiv preprint arXiv:2402.17525*, 2024.

[2] Y. Wang, H. Dong, Z. Luo, and Y. Yang, "Text-to-ascii: Generating structured ascii art from text," *arXiv preprint arXiv:2503.14375*, 2025.

[3] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representations (ICLR)*, 2022.

[4] S. Suzuki, H. Ueda, Y. Shiraishi, and K. Nakai, "Genomic language models with k-mer tokenization strategies for plant genome annotation and regulatory element strength prediction," *Plant Molecular Biology*, 2025.