Journal of KIIT. Vol. 23, No. 4, pp. 21-26, Apr. 30, 2025. pISSN 1598-8619, eISSN 2093-7571 
http://dx.doi.org/10.14801/jkiit.2025.23.4.21

# 일관된 웹툰 이미지 자동 생성을 위한 텍스트 프롬프트 최적화 기법

윤지원*[1], 윤서빈*[2], 유석종**

# Consistent Webtoon Image Generation based on Text Prompt Optimization

Jee-Won Yoon*[1], Seo-Bin Yoon*[2], and Seok-Jong Yu**


## 요 약

최근 웹툰 시장이 급속도로 성장하면서 다양한 텍스트 콘텐츠를 웹툰으로 변환하고자 하는 시도가 이루어지고 있다. 그러나 텍스트의 맥락을 이해하고 이를 시각적으로 일관성 있게 표현하면서 웹툰의 특성을 고려한 자동 생성 시스템에 대한 연구는 아직 미미한 실정이다. 본 논문에서는 다양한 장르의 텍스트를 웹툰으로 자동 변환 생성하기 위한 프롬프트 최적화 시스템을 제안한다. 제안하는 시스템은 GPT-4와 DALL-E 3를 기반으로 하며, CLIP 모델을 활용한 이미지 평가와 프롬프트 개선 메커니즘을 구현하였다. 본 연구는 기존 연구와 달리 CLIP 모델 기반의 정량적 평가 방식을 도입하고, 연속된 이미지 간의 시각적 일관성을 유지하는 메커니즘을 구현하였다.

## Abstract

As the webtoon market has grown rapidly, there have been attempts to convert various text content into webtoons. However, research on automated conversion systems that understand the context of text and maintain visual consistency while considering webtoon characteristics remains limited. This paper proposes a prompt optimization system for automatic conversion of diverse text genres into webtoons. The proposed system is based on GPT-4 and DALL-E 3, implementing image evaluation and prompt enhancement mechanisms utilizing the CLIP model. This study differentiates itself from existing research by introducing a CLIP model-based quantitative evaluation method and implementing mechanisms to maintain visual consistency between sequential images.


Keywords

prompt engineering, feedback loop, webtoon generation, text-to-image generation, CLIP model

* 숙명여자대학교 소프트웨어학부 학사과정
- ORCID[1]: https://orcig.org/0009-0003-3386-2133
- ORCID[2]: https://orcid.org/0009-0000-0113-8608
** 숙명여자대학교 소프트웨어학부 교수(교신저자)
- ORCID: https://orcid.org/0000-0002-1631-4034

· Received: Jan. 08, 2025, Revised: Mar. 11, 2025, Accepted: Mar. 14, 2025
· Corresponding Author: Seok-Jong Yu
Dept. of Computer Science, Sookmyung Womens's University,
Cheongpa-ro 100, 47-gil, Cheongpa-ro, Yongsan-gu, Seoul, Korea
Tel.: +82-2-710-9831, Email: sjyu@sookmyung.ac.kr

# Ⅰ. 서 론

웹툰 산업의 성장과 함께 텍스트 콘텐츠를 웹툰으로 변환하고자 하는 시도가 이루어지고 있다. 특히 교육과 출판 분야에서는 기존 문학 작품을 새로운 형태로 재해석하기 위해 웹툰 형식의 활용이 활발히 이루어지고 있다[1]. 그러나 텍스트의 맥락을 정확히 이해하고 이를 웹툰의 특성에 맞게 시각화하는 과정은 많은 시간과 비용이 소요되며, 전문 웹툰 작가의 역량을 필요로 한다[2]. 또한 기존의 AI 기반 웹툰 생성 시스템인 Deeptoon[3]과 Lore Machine[4]은 주로 개별 장면의 시각적 완성도에 중점을 두어, 스토리의 연속장면들 간의 서사적 연결성과 맥락적 일관성을 보장하는 데 한계를 보였다. 통합 제작 파이프라인과 자동 스토리보드 생성 기능을 제공하지만, Deeptoon은 장면 전환이 부자연스럽고 수동 조작이 많이 필요한 한계가 있으며, Lore Machine의 경우 관련 논문이나 공개된 기술이 없다는 문제가 있다.

최근 GPT-4, DALL-E 3와 같은 생성형 AI 모델의 발전으로 텍스트 처리와 이미지 생성 기술이 비약적으로 향상되었고, 텍스트와 이미지를 함께 처리하는 멀티모달(Multimodal) 모델을 통해 텍스트-이미지 간 의미적 연관성을 강화할 수 있게 되었다[5]. 최근 연구[6]에 따르면, 생성된 이미지 평가 방법인 CLIP 모델과 확산 모델(Diffusion model)을 결합한 이미지 생성 방식은 텍스트와 이미지 간의 의미적 연관성 향상에 효과적이라고 평가되고 있으며, 특히 캐릭터의 외형의 일관성과 배경의 조화를 유지하는 데 중요한 역할을 한다.

본 연구에서는 웹툰과 같이 높은 수준의 일관성을 필요로 하는 환경에서 텍스트 이미지 자동 생성을 위한 프롬프트 최적화 연구를 제안하고자 한다. 본 시스템은 GPT-4와 DALL-E 3를 기반으로 하며, CLIP 모델을 통해 생성된 이미지와 원문 텍스트 간의 의미적 일치도를 정량적으로 평가하고 프롬프트를 개선하는 피드백 루프를 구현한다. 특히 사용자 개입을 최소화하면서, 웹툰의 연속적 스토리 전개와 시각적 일관성 유지를 위한 알고리즘을 제시한다.

본 논문의 구성은 다음과 같다. 2장에서는 텍스트-이미지 변환과 웹툰 자동 생성 관련 연구들을 살펴보고, 3장에서는 제안하는 시스템의 구조를 상세히 기술한다. 4장에서는 제안된 시스템의 성능을 평가하기 위한 실험 결과를 제시하고, 5장에서는 결론 및 향후 연구 방향을 제시한다.

# Ⅱ. 관련 연구

## 2.1 텍스트-이미지 변환 생성 모델

텍스트-이미지(Text-to-image) 생성은 컴퓨터 비전과 자연어 처리가 교차하는 중요한 연구 분야로, 최근 급속히 발전하고 있다. S. Reed et al.[7]은 GAN과 합성곱-순환 텍스트 인코더를 결합한 초기 모델을 제안했으며, H. Zhang et al.[8]은 StackGAN을 통해 단계적으로 이미지를 합성하고 정제하는 방식을 도입하였다. 최근에는 확산 모델이 텍스트-이미지 생성의 주류로 자리잡았다. A. Ramesh et al.[9]이 개발한 DALL-E 2는 CLIP 잠재 공간을 활용한 확산 모델을 통해 텍스트의 의미를 더 정교하게 반영하는 이미지 생성이 가능해졌다. OpenAI의 DALL-E3는 GPT-4와의 통합을 통해 복잡한 구도와 다중 객체 구성에서도 높은 정확도를 보인다. 정교한 이미지 생성을 가능하게 했지만, 프롬프트 간 미세한 차이에도 이미지 구성 요소와 스타일의 변화가 발생하여, 이미지 일관성을 위해 사용자의 섬세한 프롬프트 제어가 요구된다.

## 2.2 프롬프트 엔지니어링과 최적화 기법

텍스트-이미지 생성 모델에서 프롬프트 엔지니어링(Prompt engineering)은 생성된 이미지의 품질과 의도 반영을 결정짓는 핵심 요소이다. 특히 웹툰과 같은 연속적 이미지에서는 스토리의 자연스러운 전개와 시각적 일관성을 위해 체계적인 프롬프트 설계가 필수적이다. 이러한 맥락에서 자동화된 프롬프트 최적화 연구가 활발히 진행되고 있다. Y. Hao et al.[10]은 사용자의 입력을 모델에 적합한 프롬프트로 자동 변환하는 PROMPTIST 프레임워크를 제안하였다. 감독 학습과 강화 학습을 결합하여 사용자

의도와 단일 이미지 품질을 동시에 충족하는 프롬프트를 생성하였다. 하지만 이 방식은 일반적인 이미지 생성에 초점을 맞추고 있어, 웹툰과 같은 연속적 이미지 생성에서 요구되는 맥락적 연관성이나 스타일 일관성을 보장하기 어렵다. 추가적으로 연속적인 이미지 생성에서는 프롬프트가 편향되는 한계가 존재한다.

### 2.3 AI 생성 이미지 평가 방법

J. Wang et al.의 연구[11]에 따르면, 이미지 생성 AI의 등장으로 생성된 이미지의 품질을 평가하는 데 집중하고 있으며, 이미지의 추상적 인식을 평가하기 위해 A. Radford et al.[12]이 텍스트와 이미지 간 의미적 연관성을 강화하는 대규모 대조 학습 방식의 CLIP 모델을 제안하였다. 이는 미세 조정 기법이나 추가 데이터가 필요한 Florence나 BLIP 등과 같은 모델에 비해 제로샷(Zero-Shot) 전이 학습 능력을 통해 새로운 데이터에 대해서도 높은 성능을 보이며, 텍스트와 이미지의 간 연관성을 파악하여 프롬프트의 각 요소가 최종 이미지 결과에 미치는 영향을 분석한다. Z. Wang et al.[13]이 제안한 PromptCharm은 CLIP 모델의 품질 평가 방식을 통해 멀티 모달 프롬프팅과 이미지 정제를 결합한 상호작용 시스템을 구현하였다. 또한 Frans 등이 제안한 CLIPDraw의 연구[14]에서는 '행복', '자아' 등의 추상적 프롬프트에 대해서도 시각화가 가능함을 보였다. CLIP 모델 기반 최적화의 한계점으로는 세부사항의 정교한 제어가 어렵다는 점을 강조하였다.

## III. 웹툰 이미지 자동 생성을 위한 프롬프트 최적화

본 연구에서는 GPT-4, DALL-E 3와 CLIP 모델을 기반으로 스토리의 맥락을 정확히 포착하고, 단계별 프롬프트의 자동 생성을 통해 이미지의 연속성을 유지하는 시스템을 제안한다.

### 3.1 시스템 구조 및 데이터 흐름

그림 1은 본 논문에서 제안하는 시스템의 단계 구조를 보여준다. 초기 텍스트를 입력받고, DALL-E 3를 통해 첫 번째 이미지가 생성된다. 이후, CLIP 모델로 생성 이미지와 입력 텍스트 간의 유사도를 평가하며, 유사도가 목표 수준에 미달할 경우 피드백 루프를 통해 프롬프트를 다시 생성한다.

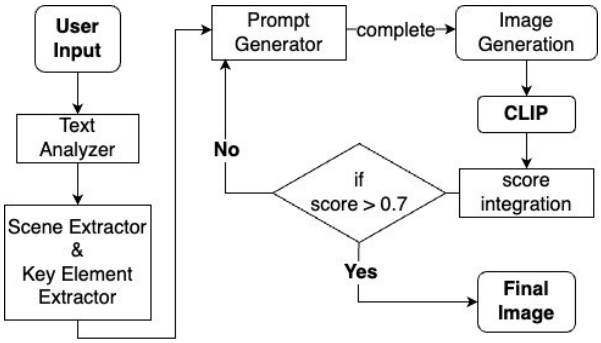



그림 1. 시스템 구조도 및 데이터 흐름
Fig. 1. System structure and data flow

### 3.2 텍스트 분석 및 처리

본 시스템은 일반 텍스트와 교육/과학 텍스트를 구분하여 처리한다. 입력 텍스트 분석 단계에서 GPT-4를 활용하여 시각화에 적합한 형태로 변환한다. 텍스트 유형별로 이미지 생성을 위한 최적화된 프롬프트 생성 과정은 그림 2의 알고리즘과 같다.

(코드1~3) 일반 서사 텍스트의 경우, 시각적 서사가 자연스럽게 전개되도록 장면 구성 방식을 결정한다. 기본 스타일 속성(선의 특성, 디테일 수준, 색상 활용 방식 등)을 정의하여 프롬프트에 반영하고, 장면을 자동으로 분할한다. (코드4~5) 사용자가 선택한 분위기 설정은 색상 팔레트 및 조명 효과와 연계되어 프롬프트의 시각적 요소로 변환된다. (코드6~10) 구도(Composition) 설정은 장면의 공간적 구성 최적화에 요구된다. 구도는 <배경과 인물 강조, 클로즈업 샷, 대화형 구도, 풍경 중심> 중에 선택되며, 캐릭터 배치, 시점, 공간 활용에 대한 변수로 변환되어 처리된다. (코드11~13) 교육/과학 텍스트의 경우, 서사적 흐름보다는 주요 개념을 추출하고 시각화 유형 <설명하기, 비교하기, 과정 보여주기 등>을 결정한다. 각 과정마다 부정 프롬프트(Negative prompt)를 덧붙여 의미전달에 불필요한 요소를 제거한다.

**Algorithm 1** Optimized Text-to-Scene Generation

$k \leftarrow \text{ComputeSceneCount}(T, \text{type}, C)$
$S \leftarrow \text{ExtractScenes}(T, \text{type})$
$S \leftarrow \text{EnhanceScenes}(S)$
$E \leftarrow \begin{cases} \text{ExtractConcepts}(T), & \text{if type} = \text{"educational"} \\ \text{ExtractStoryElements}(T), & \text{otherwise} \end{cases}$
$E' \leftarrow \text{SelectTopK}(E, k)$
**for each** $s \in S$ **do**
  $p \leftarrow \text{ApplyComposition}(s, C)$
  $p \leftarrow \text{ApplyStyle}(p, C)$
  $p \leftarrow \text{ApplyMood}(p, C)$
**end for**
$D \leftarrow \text{EnhanceScenes}(S)$
$P^* \leftarrow \text{CLIPAnalyzer.enhance}(D, \text{negative\_prompt})$
**return** $P^*$

그림 2. 프롬프트 생성을 위한 알고리즘 의사코드
Fig. 2. Pseudocode for prompt generation by text type

### 3.3 이미지-텍스트 유사도 평가 및 프롬프트 최적화

본 시스템은 웹툰의 연속된 장면에서 시각적 일관성을 유지하면서 텍스트의 의미를 정확히 반영하기 위하여 CLIP 모델을 활용한 이미지-텍스트 유사도 평가와 GPT-4 기반의 프롬프트 최적화를 수행한다. CLIP 점수는 생성 이미지와 원문 텍스트 및 이전 이미지와의 유사도를 0~1(높음) 사이의 값으로 나타낸다. 유사도 평가는 개별 이미지와 프롬프트 간의 의미적 일치도(70%)와 연속 장면 간 시각적 일관성(30%)을 기준으로 수행되었으며, 이 비율은 다수의 실험결과에서 도출하였다. 시스템은 CLIP 모델의 임베딩($f_I$)을 통해 현재 이미지($I_t$)와 이전 이미지($I_{t-i}$) 간의 코사인 유사도($\cos$)를 계산한다. 이때 최종 일관성 점수(*Consistency*)는 직전 장면과의 일관성에 더 높은 중요도를 부여하여 시간적 거리에 따른 가중치($w_i$)를 적용한다.

$$Consistency(I_t) = \frac{1}{n}\sum_{i=1}^{n} w_i \cdot \cos(f_I(I_t), f_I(I_{t-i})) \quad (1)$$

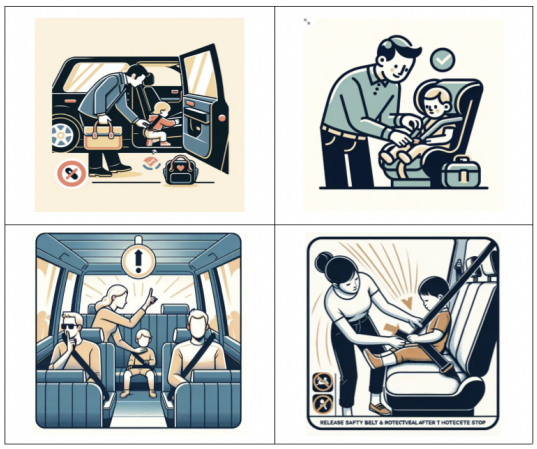
현재 유사도 점수가 목표 임계값(0.7)에 미달하는 경우, 시스템은 프롬프트 최적화를 위해 피드백 루프를 최대 3회까지 수행한다. 피드백 과정에서 CLIP 모델이 누락된 시각적 요소들을 식별하여 프롬프트에 반영하고, 이전 장면들과의 스타일 일관성을 추가하여 유사도를 평가한다. 개선 과정에서, 추상적인 표현은 구체적인 단어로 치환되며 이전 장면의 스타일 정보를 참조하여 현재 장면 프롬프트에 반영한다.

## IV. 실험 및 성능 평가

제안 시스템의 성능 평가를 위해 다양한 장르의 텍스트로 이미지 생성 실험을 수행하였다. 실험 데이터는 역사 텍스트, 동화, 묘사 중심 소설, 시, 교육/과학 텍스트 등 총 10종의 데이터셋을 포함하였다. 각 데이터셋은 평균 1500단어 내외로 구성되며, 서로 다른 스타일 설정과 구도를 적용하였다. 각 데이터셋당 3회의 실험 통해 평균 CLIP 점수를 산출했다.

### 4.1 이미지 생성 결과

그림 3은 (a)안전 프로토콜 교육 텍스트[15]와 (b) 서사적 변화가 포함된 우화 텍스트[16]를 사용하여 생성한 이미지 결과이다.



(a) 안전 프로토콜
(a) Safety protocol for child transportation


(b) 토끼와 거북이
(b) Rabbit and turtle

그림 3. 텍스트 데이터셋별 웹툰 이미지 생성 결과
Fig. 3. Webtoon images generated by text dataset

그림 3(a)는 '단계 보여주기' 방식으로, (b)는 '만화(Cartoon)' 스타일로 각각 4컷 이미지를 생성하였다. (a)는 단계별 지침이 명확한 이미지 장면으로 변환되어, 충분한 의미 전달이 가능하였다(CLIP점수=1.0). (b)는 등장인물의 감정과 장면의 분위기가 적절히 표현되어 서사적 연결성이 잘 유지되었다(0.97). 텍스트별로 안정적인 성능을 보여 주었으나, DALL-E 3의 내재된 무작위성으로 인해 동일한 프롬프트를 사용하더라도 캐릭터의 외형이 완전히 일치하지 않는 한계가 관찰되었다. 이러한 스타일 편차는 장면 개수가 증가할수록 누적되는 경향을 보였다.

### 4.2 정량적 성능 평가

생성된 이미지 품질을 정량적으로 평가하기 위해 CLIP 유사도 점수와 이미지 생성 소요 시간을 측정하였다. 각 장면 생성마다 전체 파이프라인 수행 시간을 측정한 결과, 4컷의 서사 텍스트와 교육 텍스트 생성 시, 각각 평균 243초와 73.5초가 소요되었다. 또한, 텍스트 유형별 성능 분석 결과, 교육/과학 텍스트는 거의 모든 유형에서 1.0의 CLIP점수를 기록하였으나, 비교 설명(Compare) 유형은 0.65의 비교적 낮은 점수를 보였다.

표 1. 텍스트 유형별 CLIP 점수 결과
Table 1. Comparison of CLIP scores by text type

| Text type | CLIP score |
|---|---|
| Science (Show process) | 1.00 |
| Science (Explain) | 1.00 |
| Fairy tale (Minimalist) | 1.00 |
| Descriptive novel | 1.00 |
| Background novel | 0.89 |
| Fairy tale (Webtoon) | 0.88 |
| Poetic text | 0.86 |
| Science (Compare) | 0.65 |
| Historical | 0.62 |

서사 텍스트의 경우, 구체적 묘사가 풍부한 텍스트에서 더 높은 유사도를 보였다. 묘사 중심 소설(Descriptive novel)과 동화(Fairy tale-Min)는 1.0인 반면, 역사 텍스트는 0.62로 측정되었다. 이는 범용 데이터 위주로 학습한 GPT 모델이 역사적, 문화적 맥락이 중요한 텍스트에서는 성능이 제한적이라는 것을 시사한다고 볼 수 있다.

## Ⅴ. 결론 및 향후 과제

본 연구에서는 일관된 웹툰 이미지 자동 변환을 위한 프롬프트 최적화를 위해 GPT-4와 DALL-E 3를 결합한 생성 파이프라인과 피드백 과정을 구현하였다. 실험 결과, 생성된 웹툰 이미지가 평균 0.89점의 CLIP 유사도 점수를 기록하여, 제안한 최적화 방식이 이미지의 일관성 및 스토리의 연속성 유지에 효과적임을 확인하였다. 다만 5컷 이상의 이미지를 생성할 경우 스타일 편차가 누적된다는 점과 CLIP점수가 높더라도 실제 웹툰의 품질이 충분히 보장되지 않는 한계가 있었다. 향후 연구로, 이미지 일관성 개선을 위해 다른 확산 모델 또는 이미지 특징 추출 모델과의 병합 연구를 제안할 수 있다.

## References

[1] Y. K. Seol, "A Study on The Educational Utilization of Webtoons Theoretical and Practical Evidence Exploration", The Journal of the Korea Contents Association, Vol. 20, No. 6, pp. 510-520, Jun. 2020. https://doi.org/10.5392/JKCA.2020.20.06.510.

[2] S. J. Lee and J. Y. Kang, "A Study on Artificial Intelligence-Driven Paradigm Shift in the Webtoon Ecosystem", The Society of Convergence Knowledge Transactions, Vol. 11, No. 3, pp. 45-54, Sep. 2023. https://doi.org/10.22716/sckt.2023.11.3.024

[3] https://www.deeptoon.com/ [accessed: Apr. 23, 2025]

[4] https://www.loremachine.world/ [accessed: Apr. 23, 2025]

[5] K. Yu, "A Study on Webtoon Generation Using Multimodal AI", Ph.D. dissertation, Dept. of Computer Engineering, Chosun University, 2023.

[6] Z. Wang, "Enhancing Text-to-Image Generation: Integrating CLIP and Diffusion Models for Improved Visual Accuracy and Semantic Consistency", Proceedings of CONF-MLA 2024 Workshop, pp. 16-21, Nov. 2024. https://doi.org/10.54254/2755-2721/105/2024TJ0053.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis", International Conference on Machine Learning, NY, USA, pp. 1060-1069, Jun. 2016. https://doi.org/10.48550/arXiv.1605.05396.

[8] H. Zhang, et al., Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks", Proceedings of the IEEE international conference on computer vision, Venice, Italy, pp. 5907-5915, Aug. 2017. https://doi.org/10.48550/arXiv.1612.03242.

[9] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents", arXiv preprint arXiv:2204.06125, Apr. 2022. https://doi.org/10.48550/arXiv.2204.06125.

[10] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation", Advances in Neural Information Processing Systems, New Orleans, LA, USA, Nov.-Dec. 2022. https://doi.org/10.48550/arXiv.2212.09611.

[11] J. Wang, K. Chan, and C. Loy, "Exploring CLIP for Assessing the Look and Feel of Images", arXiv preprint arXiv:2207.12396, pp. 1-3, Nov. 2022. https://doi.org/10.48550/arXiv.2207.12396.

[12] A. Radford, et al., "Learning transferable visual models from natural language supervision", International conference on machine learning, pp. 8748-8763, Feb. 2021.

[13] Z. Wang, Y. Huang, D. song, L. Ma, and T. Zhang, "PromptCharm: Text-to-Image Generation through Multi-modal Prompting and Refinement", Proceedings of the CHI Conference on Human Factors in Computing Systems, Mar. 2024. https://doi.org/10.48550/arXiv.2403.04014.

[14] K. Frans, L. B. Soros, and O. Witkowski, "CLIPDraw: Exploring Text-to-Drawing Synthesis through Language-Image Encoders", arXiv preprint arXiv:2106.14843., pp. 6-7, Jun. 2021. https://doi.org/10.48550/arXiv.2106.14843.

[15] lnput Text 1, https://easylaw.go.kr/CSP/CnpClsMain.laf?popMenu=ov&csmSeq=690&ccfNo=1&cciNo=2&cnpClsNo=1 [accessed: Apr. 23, 2025]

[16] Input text 2, https://teacher.depaul.edu/Documents/TheTurtleandtheRabbitFiction3rdGrade.pdf [accessed: Apr. 23, 2025]

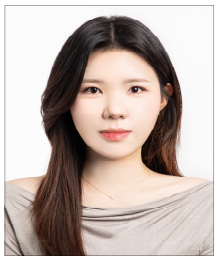
## 저자소개

윤 지 원 (Jee-Won Yoon)

2023년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 생성형 AI, 프롬프트 엔지니어링, 알고리즘

윤 서 빈 (Seo-Bin Yoon)



2023년 2월 ~ 현재 : 숙명여자대학교 소프트웨어학부 학사과정

관심분야 : 생성형 AI, 프롬프트 엔지니어링, 데이터분석

유 석 종 (Seok-Jong Yu)

1994년 2월 : 연세대학교 전산과학과(이학사)

1996년 2월 : 연세대학교 컴퓨터과학과(이학석사)

2001년 2월 : 연세대학교 컴퓨터과학과(공학박사)

2005년 3월 ~ 현재 : 숙명여자대학교 소프트웨어학부 교수

관심분야 : 데이터마이닝, 추천시스템, 정보시각화