한국어 기반 Text-to-Video 프롬프트 엔지니어링 시스템 연구 : CogVideoX를 중점으로

임영기[1], 최지희[1], 엄태선[1], 이우성[1], 구영현*

㈜피씨엔[1], *세종대학교

yglim@pcninc.co.kr, jhchoi@pcninc.co.kr, tseom@pcninc.co.kr, wooslee@pcninc.co.kr, *yhgu@sejong.ac.kr

A Study on the Text-to-Video Prompt Engineering System Based on Korean : Focusing on CogVideoX

Younggi Lim, Jihui Choi, Taeseon Eom, Wooseong Lee, Yeonghyeon Gu*

PCN R&S Headquarters., *Sejong Univ.

요 약

대다수의 Text-to-Video(T2V) AI 모델은 영어 사용을 전제로 연구되고 있기 때문에, T2V 모델 비전문가 혹은 국내 사용자들은 텍스트 프롬프트 작성에 많은 어려움을 겪는다. 이를 해결하기 위해 본 논문은 한국어 기반 Text-to-Video 프롬프트 엔지니어링 시스템(Video Prompt Engineering based on Korean, VPE-Ko)을 제안한다. VPE-Ko는 한국어 대화를 통해 사용자로부터 영상 생성에 필요한 핵심 정보를 수집하여 T2V 모델에 최적화된 프롬프트를 생성하는 시스템이다. 본 연구는 CogVideoX를 중점으로 진행되었다. VPE-Ko는 T2V 프롬프트의 5가지 핵심 요소를 도출해 이를 활용하여 VPE-Ko 시스템 프롬프트 및 멀티-턴 챗봇 시스템을 구축한다. 본 연구는 Qwen2 언어 모델을 활용해 수행됐으며, 한국어 멀티-턴 데이터셋을 생성 및 사용한다. 실험 결과 VPE-Ko는 G-Eval 4.9, 프롬프트 만족도 점수 4.5, 영상 만족도 점수 4.6을 기록하며 우수한 성능을 보였다. 더욱 자세한 내용 GitHub에서 확인할 수 있다 : https://github.com/wndaasa/VPE-Ko/

Ⅰ. 서 론

최근 Text-to-Video(T2V) 생성형 AI 모델의 발전이 활발히 이루어지며 텍스트 프롬프트 엔지니어링의 중요성이 강조되고 있다. 하지만 대부분의 T2V 모델은 기본적으로 영어 사용을 전제하고 있으며, 프롬프트 가이드 또한 영어 작성을 권장하고 있다. 이는 한국어가 익숙한 국내 사용자들, 특히 T2V 모델 비전문가들에게 큰 어려움으로 작동한다.

프롬프트 엔지니어링을 위한 연구는 다방면으로 진행되어 왔다. Sarkar 외[1]는 소규모 LLM으로도 효율적인 프롬프트 작성이 가능함을 보여준다. Ein-Dor 외[2]는 LLM을 활용해 단계별 논의 수행을 통해 효과적인 프롬프트를 제작하는 프레임워크를 제안한다. Zhang 외[3]는 효과적인 대화형 AI 개발을 위해서는 전반적인 기술 발전이 필요함을 강조한다. 손민준 외[4]는 데이터셋에 따라 적합한 프롬프트 엔지니어링 기법이 달라질 수 있음을 분석했다. 이처럼 프롬프트 엔지니어링을 위한 AI 시스템 연구는 많으나, 한국어 혹은 T2V 모델에 특화된 프롬프트 엔지니어링 시스템 연구는 많이 부족한 상황이다.

이에 본 논문은 한국어 기반 T2V 프롬프트 엔지니어링 시스템(T2V Prompt Engineering system based on Korean, VPE-Ko)을 제안한다. 본 논문의 제안점은 두 가지이다. 첫 번째는 T2V 프롬프트의 핵심 요소를 도출해 설계한 시스템 프롬프트이다. 두 번째는 프롬프트 엔지니어링을 위한 한국어 멀티-턴 챗봇 구현이다. VPE-Ko의 유효성은 사용자 작성 프롬프트와 VPE-Ko 기반 프롬프트를 한 쌍으로 묶어 G-Eval과 사용자 설문조사의 2가지 방법으로 측정했다.

Ⅱ. 방법론

본 연구는 한국어 기반 T2V 프롬프트 엔지니어링 시스템인 VPE-Ko를 제안한다. VPE-Ko는 T2V 모델을 처음 사용하는 한국어 사용자가 전문적인 프롬프트 지식 없이도 고품질 프롬프트를 작성하도록 지원한다.

본 논문은 오픈소스 T2V 모델 중 프롬프트 이해도가 높으며, 1360×768 고해상도 영상 생성이 가능한 CogVideoX[5]를 중심으로 연구됐다. 우선 공식 문서를 분석해 VPE-Ko 시스템을 위한 고품질 영상 생성 핵심 요소를 도출했다. VPE-Ko 핵심 요소는 표 1에서 확인할 수 있다.

표 1 VPE-Ko 핵심 요소

핵심 요소	예시
주요 객체	건물, 사람, 동물 등
카메라 구도	전경을 담는 구도, 특정 객체에 집중한 구도 등
카메라 종류	드론, 1인칭 카메라 등
영상의 속도	느린 속도, 빠른 속도 등
조명	자연광, 네온 조명 등

VPE-Ko 학습에 사용된 데이터셋

VPE-Ko 핵심 요소에 중점을 둔 한국어 대화 데이터셋을 자체 생성하여 연구를 진행했다. 데이터셋은 두 개의 챗봇에 각각 Assistant와 User 역할을 부여한 뒤 자연스럽게 대화시켜 구축했다. Assistant는 VPE-Ko 시스템 역할로 핵심 요소를 수집하고, User는 VPE-Ko 사용자로서 핵심 요소 정보를 제공한다. 이렇게 생성된 5,000개의 대화 내용과 이를 기반으로 한 24,769건의 한국어 대화쌍 데이터셋은 Qwen2[6]의 학습에 사용됐다. 데이터셋 구축을 위한 챗봇은 ChatGPT-API를 사용했다.

VPE-Ko 시스템 프롬프트 설계

VPE-Ko 시스템 프롬프트는 멀티-턴 챗봇의 지시문으로 사용되며, 설계 방향성은 아래와 같다.

1. VPE-Ko 핵심 요소의 체계적인 수집을 위한 순차적 질문 전략 설계
2. VPE-Ko 핵심 요소 외에도 키워드 가중치, 부정 프롬프트 등 일반적인 프롬프트 가이드 포함
3. T2V 프롬프트 스타일 분석을 위한 Few-Shot 예시 프롬프트 제시

* 교신저자

4. 영어 프롬프트, 한국어 번역 프롬프트, 부정 프롬프트, 개선 사항 제안의 네 가지 표준 출력 형식 제안

VPE-Ko 멀티-턴 챗봇 시스템

본 논문이 제안하는 VPE-Ko 멀티-턴 챗봇 시스템은 사용자와의 연속적인 한국어 대화를 기반으로 고품질 영어 프롬프트를 생성한다. 멀티-턴 챗봇 구현을 위해 이전 대화 내용 전체를 다음 입력값으로 사용했다.

VPE-Ko는 사용자가 입력한 주제에 알맞은 핵심 요소를 한국어 대화를 통해 수집한다. VPE-Ko가 핵심 요소를 충분히 수집했다고 판단하면 영어 프롬프트를 포함한 네 가지 출력을 생성한다. 이때 사용자는 일상적인 구어체 대화를 통해 프롬프트 피드백을 전달하고, VPE-Ko는 이를 반영하여 더욱 개선된 프롬프트를 생성한다. VPE-Ko 멀티-턴 챗봇 시스템 구조도는 그림 1에서 확인할 수 있다.

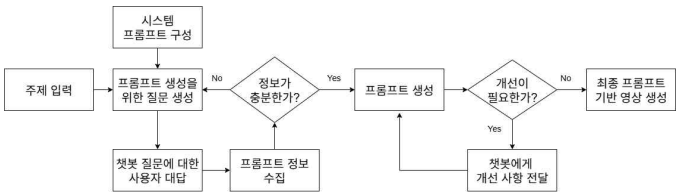

그림 1 VPE-Ko 멀티-턴 챗봇 시스템 구조도

Ⅲ. 실험

VPE-Ko 시스템의 유효성을 검증하기 위해 본 연구에서는 A-B 테스트를 진행했다. 실험 참가자는 T2V 모델 사용 경험이 없는 한국인으로 구성되었다. 우선 실험 참가자가 직관적으로 작성한 프롬프트 A와 VPE-Ko 시스템을 활용해 생성한 프롬프트 B를 준비한다. 작성된 프롬프트 A와 B는 CogVideoX 모델을 통해 각각 영상 A', B'를 생성한다.

본격적인 실험에 앞서 프롬프트 A, B와 영상 A', B'의 특징은 다음과 같다. 아무런 사전 지식 없이 작성한 프롬프트 A는 매우 짧고 정보가 단편적이다. 영상 A' 또한 CogVideoX가 이해하지 못하는 한국어 문장을 기반으로 하기 때문에 프롬프트 A와 무관한 영상이 생성됐다. 반면 프롬프트 B는 VPE-Ko의 5가지 핵심 요소가 전부 포함되어 사용자의 의도가 상세하게 반영된 영어 문장이다. 덕분에 영상 B' 또한 사용자의 생각과 의도를 그대로 반영하고 있다. 프롬프트 A, B와 생성된 영상 A', B'의 대표 예시는 표 3에서 확인할 수 있다.

실험 참가자는 프롬프트 A, B와 영상 A', B'에 대해 자신의 의도가 정확하게 반영됐는지 5점 만족도 조사를 진행한다. 또한 프롬프트 A, B는 생성형 AI 평가에 특화된 G-Eval[7] 지표를 사용하여 객관적인 유효성을 획득한다. VPE-Ko 유효성 실험은 10명의 참가자로 진행되었다.

실험 결과, G-Eval 점수는 A 3.6, B 4.9로 VPE-Ko 기반 프롬프트가 더욱 우수한 품질임을 보였다. 또한 사용자 만족도 점수에서도 A 3.6, B 4.5로 VPE-Ko가 사용자의 의도를 보다 정확하게 반영했음을 확인했다. 결정적으로 CogVideoX는 한국어 처리가 불가능해 한국어 프롬프트로 생성된 영상 A'의 만족도 점수는 2.1로 아주 낮은 반면, VPE-Ko의 영어 프롬프트를 입력하여 생성한 영상 B'의 만족도는 4.6점으로 매우 높았다. 이러한 결과는 한국어를 사용하는 T2V 비전문가에게 VPE-Ko가 매우 유용했음을 의미한다. 실험 결과는 표 2에서 확인할 수 있다.

표 2 VPE-Ko 실험 결과

종류	G-Eval	사용자 만족도	
		프롬프트	영상
A	3.6	3.6	2.1
B	**4.9**	**4.5**	**4.6**

Ⅳ. 결론

본 논문은 한국어 기반 T2V 프롬프트 엔지니어링 시스템 VPE-Ko를 제안한다. 해당 시스템은 LLM 기반 시스템 프롬프트와 멀티-턴 챗봇 구조를 활용하여 효과적인 프롬프트 엔지니어링 환경을 제공한다. 이를 위해 CogVideoX 프롬프트 핵심 요소 도출하고 2개의 LLM으로 생성한 24,769건의 한국어 대화쌍 데이터셋으로 Qwen2을 학습시켰다. VPE-Ko 실험은 사용자 만족도와 G-Eval을 활용한 A-B 테스트로 진행됐다. VPE-Ko 기반 프롬프트가 G-Eval 4.9점, 사용자 만족도 4.5점을 기록했으며, 특히 생성된 영상 만족도 4.6점을 달성하며 매우 유효한 연구임을 증명했다.

따라서 한국어 대화 기반 고품질 영어 프롬프트를 생성하는 VPE-Ko 시스템은 한국어에 익숙한 T2V 비전문가에게 큰 도움이 될 수 있으며, 더 나아가 다양한 영상 생성 AI 모델에도 적용될 수 있을 것으로 기대된다.

acknowledgment

본 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신산업진흥원의 지원을 받아 수행되었으며(고성능 컴퓨팅 지원(GPU, NPU)), 연구 수행 과정에서 귀중한 조언과 도움을 주신 이혜원 前 차장님께 깊이 감사드립니다.

참 고 문 헌

[1] Sarkar, R., et al., (2025). Conversational user-ai intervention: A study on prompt rewriting for improved llm response generation. arXiv preprint arXiv:2503.16789.

[2] Ein-Dor, et al., (2024). Conversational prompt engineering. arXiv preprint arXiv:2408.04560.

[3] Zhang, C., Dai, et al., (2025). A survey on multi-turn interaction capabilities of large language models. arXiv preprint arXiv:2501.09959.

[4] 손민준, & 이성진. (2025). 대규모 언어 모델에서 프롬프트 엔지니어링 기법에 관한 연구. 한국통신학회논문지, 50(1), 9-21.

[5] Yang, Z., et al., (2024). Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072.

[6] Team, Q. (2024). Qwen2 technical report. arXiv preprint arXiv:2407.10671.

[7] Liu, Y., et al., (2023). G-eval: NLG evaluation using gpt-4 with better human alignment. arXiv preprint arXiv:2303.16634.

표 3 VPE-Ko 대표 예시

A-B	텍스트 프롬프트 (Input)	영상 (Output)
A (A')	뜨겁게 타오르는 화산	
B (B')	A scorching volcano erupts in a breathtaking display of fiery majesty, its molten lava (flowing down:2.0) the mountain's sides like a river of liquid fire. From above, the camera captures the scene in sweeping aerial shots, emphasizing the scale and intensity of the eruption. The natural light bathes the landscape in an ethereal glow, highlighting the (crimson flames:1.8) that dance and flicker against the night sky.