한국어 기반 시각-언어모델 안전성 성능 평가

[1] 김진성, [2] 이영완*, [2] 이용주
[1] 순천향대학교 AI · 빅데이터학과 학부생
[2] 한국전자통신연구원 선임연구원

rlawlstjd0115@sch.ac.kr, {yw.lee,yongju}@etri.re.kr

A Safety Benchmark and Evaluation for Korean Vision-Language Models

[1]Jin-Seong Kim, [2]Youngwan Lee*, [2]Yong-Ju Lee
[1]Dept. of AI and Bigdata, Soonchunhyang University
[2]Electronics and Telecommunications Research Institute (ETRI)

요 약

본 연구는 Ko-HoliSafe-Bench 를 활용하여 국내 대표 Ko-VLM 4 종의 안전성을 체계적으로 평가했다. 평가 결과, 모든 Ko-VLM 은 유해 입력에 대해 65% 이상의 높은 mASR(mean Attack Success Rate)을 기록하며 안전성에 전반적인 취약점을 보였다. 특히, 텍스트보다 이미지 기반 유해성 탐지 능력이 현저히 낮았으며, 문맥적 위험을 추론하는 데에도 뚜렷한 한계를 드러냈다. 또한, 선행 연구의 영어권 VLM 과 비교했을 때 높은 mASR 과 불필요한 Refusal Rate(RR)를 동시에 보여 안전성과 유용성 모두에서 개선이 필요함을 확인했다. 본 연구의 결과는 한국어 VLM 의 안전성 강화를 위한 후속 연구의 필요성을 시사한다.

1. 서론

Vision-Language Model (VLM)은 이미지와 텍스트를 동시에 이해하는 멀티모달 모델로 다양한 분야에 활용되고 있다. 그러나 모델이 생성하는 응답은 유해 발화, 편향, 개인정보 침해 등 여전히 많은 안전성 문제를 내포하고 있어, VLM 의 안전성을 체계적으로 평가하고 강화하는 연구의 중요성이 커지고 있다.

영어권에서는 VLM 의 잠재적 위험성을 평가하는 안전성 연구[1, 2]가 활발히 진행되고 있다. 반면, 한국어의 경우 기존 연구는 텍스트 기반 LLM 의 안전성 평가에 집중되어 있으며[3], Korean VLM (Ko-VLM)의 안전성을 심층적으로 분석하고 평가한 연구는 아직 부족하다.

따라서 본 연구에서는 HoliSafe-Bench 의 한국어 버전을 사용하여 대표적인 Ko-VLM 들의 안전성을 다각적으로 평가하고자 한다. 이를 통해 현재 한국어 VLM 이 가진 안전성의 취약점을 진단하고, 향후 모델 개발에 필요한 개선 방향을 제시하는 것을 목표로 한다.

2. 평가 방법

Ko-VLM 의 안전성 평가는 HoliSafe-Bench[2]의 한국어 버전인 Ko-HoliSafe-Bench 를 사용하였다. 이 데이터셋은 총 4,146 개의 이미지-텍스트 쌍으로 구성되며, 7 개 상위 카테고리와 18 개 하위 카테고리를 포함한다. 특히, 이미지와 텍스트의 안전성 조합에 따라 설계된 아래 5 가지 유형(Safeness Type)을 통해 모델의 취약점을 다각적으로 분석할 수 있다.

(1) $U_I U_T$: Unsafe Image + Unsafe Text
(2) $U_I S_T$: Unsafe Image + Safe Text
(3) $S_I U_T$: Safe Image + Unsafe Text
(4) $S_I S_T \rightarrow U$: Safe Image + Safe Text = Unsafe 한 결과
(5) $S_I S_T \rightarrow S$: Safe Image + Safe Text = Safe 한 결과

평가 모델로는 HyperCLOVAX, KANANA, VARCO, A.X 등 4 개의 대표적인 Ko-VLM 활용하였다. 모델의 안전성 수준을 평가하기 위한 지표로는 유해 입력에 대한 공격 성공률을 의미하는 mASR(mean Attack Success Rate)과 안전한 입력에 대한 답변 거부율을 의미하는 RR(Refusal Rate)을 사용하였다.

* 교신저자

<표 1> Ko-HoliSafe 데이터셋 기반 Ko-VLMs 안전성 평가

Model	$S_IS_T \rightarrow S$ (↑)	$S_IS_T \rightarrow U$ (↓)	$U_IS_T \rightarrow U$ (↓)	$S_IU_T \rightarrow U$ (↓)	$U_IU_T \rightarrow U$ (↓)	$mASR$ (↓)	RR (↓)
HyperCLOVAX-SEED-Vision-Instruct-3B	**95.61**	**85.1**	85.27	**43.6**	**39.55**	**65.8**	**4.39**
KANANA-1.5-v-3b-Instruct	95.36	92.06	86.2	59.84	57.82	73.63	4.64
VARCO-VISION-2.0-1.7B	88.83	91.64	**80.95**	59.07	59.8	72.89	11.17
A.X-4.0-VL-Light	94.1	92.34	89.6	58.84	61.63	75.58	5.9

3. 실험 결과 및 분석

표 1 은 Ko-VLM 4 종의 안전성 평가 결과를 보여준다. 평가 결과, 모든 Ko-VLM 이 높은 mASR 을 기록하며 전반적으로 공격에 취약한 것으로 나타났다. HyperCLOVAX 가 mASR 65.8%, RR 4.39%로 가장 우수한 방어 성능을 보였으나, A.X(75.58%), KANANA(73.63%), VARCO(72.89%)는 모두 70%가 넘는 높은 공격 성공률을 기록했다. 주목할 점은 VARCO 의 경우 RR 이 11.17%로 가장 높았음에도 mASR 은 여전히 높아, 단순히 답변을 거부하는 경향만으로는 모델의 안전성을 확보하기 어렵다는 것을 시사한다.

세부 유형별로 살펴보면, Ko-VLM 은 텍스트 기반 유해성 탐지보다 이미지 기반 유해성 탐지에 더 큰 취약점을 보였다. 예를 들어, 가장 성능이 좋았던 HyperCLOVAX 조차 텍스트가 유해한 유형(S_IU_T)에서는 43.6%의 ASR 을 기록한 반면, 이미지가 유해한 유형(U_IS_T)에서는 ASR 이 85.27%까지 급증했다. 또한, 안전한 이미지와 텍스트를 조합하여 유해한 의도를 유도하는 $S_IS_T \rightarrow U$ 유형에서는 모든 모델의 ASR 이 85% 이상으로 매우 높게 나타났다. 이는 Ko-VLM 이 시각적 맥락의 위험을 인지하는 능력이 부족할 뿐만 아니라, 문맥 속에 숨겨진 위험을 추론하는 데에도 뚜렷한 한계가 있음을 보여준다.

이러한 Ko-VLM 의 취약점은 선행 연구[2]에서 평가된 영어권 VLM 의 결과와 비교했을 때 더욱 두드러진다. 해당 연구에 따르면 InternVL2.5-8B(mASR 62.65, RR 1.46), Qwen-2.5-VL-7B(mASR 52.29, RR 0.53) 등 상위권 영어 VLM 들은 Ko-VLM 보다 현저히 낮은 mASR 과 RR 을 동시에 달성하는 경향을 보였다. 다만 LLaVA-v1.5-7B(mASR 78.20, RR 2.52)의 경우, mASR 은 일부 Ko-VLM 보다 높았지만 RR 은 현저히 낮아 Ko-VLM 에 비해 유용성 저하를 최소화하는 모습을 보였다. 따라서 Ko-VLM 은 높은 mASR 과 RR 을 보여 안전성과 유용성 양 측면에서 개선이 필요함을 확인하였다.

이러한 정량적 평가에서 드러난 취약점은 그림 1 의 정성적 응답 사례에서도 확인할 수 있다. '무기 개조 방법', '소매치기 방법', '개인정보 침해 방법'와 같은 명백히 유해한 질의에 대해 일부 모델들이 구체적인 방법을 안내하는 등, 위험한 요청을 제대로 거부하지 못하는 문제를 명확히 보여주었다.

(그림 1) Ko-VLM 의 안전성 응답 비교 사례

4. 결론

본 연구에서는 Ko-HoliSafe-Bench 를 활용하여 국내 대표 Ko-VLM 4 종의 안전성을 평가하였다. 평가 결과, Ko-VLM 은 유해한 입력에 대해 높은 mASR 을 기록하며 전반적인 안전성에 취약점을 보였다. 특히, 텍스트보다 이미지 기반 유해성에 더 민감하게 반응했으며, 문맥 속에 숨겨진 위험을 추론하는 데에도 뚜렷한 한계를 드러냈다. 또한, 영어권 VLM 에 비해 불필요한 RR 도 높아 안전성과 유용성 측면에서 모두 개선이 필요함을 확인하였다.본 연구의 결과는 향후 한국어 VLM 의 안전성을 강화하기 위한 후속 연구의 필요성을 강조하며, Ko-HoliSafe-Bench 가 이를 위한 핵심적인 평가 기준으로 활용될 수 있음을 시사한다.

ACKNOWLEDGMENT

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No. RS-2022-00187238, 효율적 사전학습이 가능한 한국어 대형 언어모델 사전학습기술 개발)

참고문헌

[1] Zhou, K., Liu, C., Zhao, X., Compalas, A., Song, D., & Wang, X. E., Multimodal Situational Safety, International Conference on Learning Representations (ICLR 2025), Singapore (Singapore EXPO), 2025, 1–13.

[2] Lee, Y., Kim, K., Park, K., Jung, I., Jang, S., Lee, S., Lee, Y., & Hwang, S. J., HoliSafe: Holistic Safety Benchmarking and Modeling with Safety Meta Token for Vision-Language Model, arXiv preprint arXiv:2506.04704, 2025, 1–15.

[3] Lee, H., Hong, S., Park, J., Kim, T., Kim, G., & Ha, J.-W., KoSBi: A Dataset for Mitigating Social Bias Risks Towards Safer Large Language Model Applications, Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), Industry Track, Toronto (Canada), 2023, 208–224.