K-Prism: A Korean Diagnostic Benchmark for Multi-faceted Hallucination in Vision-Language Models

Min Hyuk Kim, Dong-hyok Shin, Jimi Lee, Jin Yea Jang
Korea Electronics Technology Institute, Republic of Korea
{alsgur0720,sdh876,mirine520,jinyea.jang}@keti.re.kr

Abstract

Vision-language models (VLMs) often produce hallucinated outputs that are not supported by visual or textual evidence. While existing benchmarks have advanced hallucination evaluation, they primarily target object-level or task-specific errors and often assess different failure types on separate samples. This design limits its direct comparison of hallucination patterns under shared evidence, and most prior evaluations remain English-centric. We introduce K-Prism, a Korean diagnostic benchmark for multi-faceted hallucination analysis in VLMs. K-Prism contains 1,002 multiple-choice questions across text and image tracks, organized into shared-evidence topic groups. Within each group, questions are designed to probe three diagnostic dimensions: factual, reasoning, and misleading. K-Prism uses a controlled dual-mode protocol comparing intact inputs with visually perturbed inputs to examine hallucination under reduced visual evidence. Experiments with closed- and open-source VLMs show dimension-specific degradation rather than a uniform accuracy drop. These findings demonstrate that hallucination in Korean VLMs cannot be captured by a single aggregate score and requires fine-grained diagnostic evaluation across evidence conditions and failure dimensions. Our dataset is publicly available at https://huggingface.co/spaces/KETI-NLP/K-Prism_Benchmark.

1 Introduction

A central challenge in evaluating hallucination in vision-language models (VLMs) is not merely detecting unsupported outputs, but identifying which functional failure leads to unsupported prediction (Bai et al., 2024). Existing benchmarks provide valuable diagnostics for observable hallucination patterns, including object-level inconsistencies (Li et al., 2023), multimodal conflicts (Guan et al., 2024), and domain-specific failures (Khanal et al., 2025). However, these benchmarks often evaluate different error types using separate samples, making it difficult to compare multiple facets of hallucination under the same evidential context. Hallucination evaluation also remains predominantly English-centric. Recent multilingual VLMs and visual question answering (VQA) benchmarks show that visual-language understanding varies across languages (Atuhurra et al., 2025; Tang et al., 2025), and multilingual hallucination studies further motivate language-specific evaluation beyond English (Abdaljalil et al., 2026).

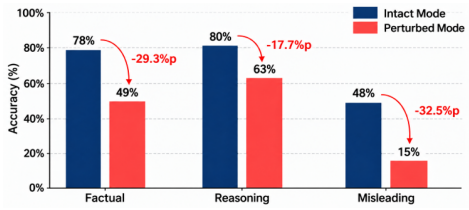


Figure 1: Accuracy comparison between intact mode and perturbed mode across factual, reasoning, and misleading dimensions.

To address these limitations, we introduce K-Prism, a Korean diagnostic benchmark for multi-faceted hallucination in VLMs. K-Prism consists of multiple questions within a shared evidential context and compares three diagnostic dimensions: factual, reasoning, and misleading. These dimensions are motivated by prior findings that hallucination in VLMs can arise from insufficient grounding, reasoning errors, and over-reliance on language priors or misleading cues (Bai et al., 2024; Guan et al., 2024; Xuan et al., 2025). The factual dimension evaluates whether a model recovers evidence-grounded information; the reasoning dimension evaluates whether a model derives a valid answer from the evidence; and the misleading dimension evaluates whether a model resists unsupported assumptions introduced by the question or distractors.

To further analyze evidence sensitivity, we evaluate each item under a dual-mode protocol: intact mode uses the original input, whereas perturbed

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 632–644
November 6-10, 2026 ©2026 Association for Computational Linguistics

Benchmark	Target Language	Diagnostic Focus	Shared Context across Dimensions	Noise Perturbation
POPE (Li et al., 2023)	English	Object Hallucination	×	×
AMBER (Wang et al., 2023)	English	Existence / Attribute / Relation	×	×
HallusionBench (Guan et al., 2024)	English	Visual Illusion & Language Hallucination	×	×
MMMU-Pro (Yue et al., 2025)	English	Robust Multimodal Reasoning	×	×
MTVQA (Tang et al., 2025)	Multilingual	Text-centric VQA	×	×
Khanal et al. (Khanal et al., 2025)	English	Domain-specific Medical Hallucination	×	×
KRETA (Hwang et al., 2025)	Korean	Text-rich VQA & Reasoning	×	×
K-Prism (Ours)	**Korean**	**Multi-faceted Functional Failures**	✓	✓

Table 1: Comparison of K-Prism with representative VLMs hallucination and multimodal evaluation benchmarks.

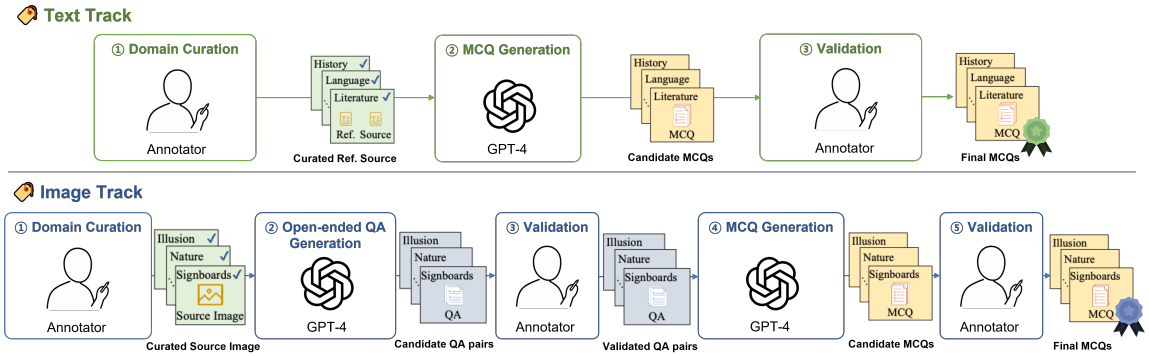


Figure 2: K-Prism data construction pipeline for the text-track and image-track.

mode removes image-track visual evidence with Gaussian noise or adds a non-informative visual distractor to text-track items. As an illustrative summary of our evaluation results, Figure 1 shows that factual, reasoning, and misleading dimensions exhibit distinct degradation patterns under perturbation, rather than a uniform accuracy drop. This observation supports the need to report hallucination behavior beyond a single aggregate score, consistent with recent diagnostic benchmarks that emphasize decomposed or multi-dimensional evaluation (Wang et al., 2023; Guan et al., 2024).

The dual-mode comparison examines how hallucination patterns vary with multimodal grounding, language priors, and visual interference, following prior work on controlled multimodal evaluation and robustness analysis (Guan et al., 2024; Yue et al., 2025). Together with the shared-evidence design, this protocol enables K-Prism to distinguish whether model errors stem from failures in evidence-grounded factual recognition, valid inference, or resistance to unsupported assumptions under intact and perturbed inputs. As summarized in Table 1, K-Prism addresses three gaps in existing literature: it provides a Korean VLM hallucination benchmark, organizes factual, reasoning, and misleading questions under shared evidential contexts, and introduces a controlled perturbation protocol for comparing intact and visually corrupted inputs.

2 Dataset

K-Prism consists of two tracks: a text track and an image track. Both tracks use a topic-group structure, where each topic or source is paired with three multiple-choice questions (MCQs) targeting factual, reasoning, and misleading failure dimensions. Within each topic group, the three questions share the same source evidence while varying the required diagnostic dimension, allowing dimension-wise comparison without changing the underlying topic. Factual questions test explicitly stated information, reasoning questions require inference from the same source evidence, and misleading questions probe whether models rely on unsupported assumptions. In the image track, misleading questions are further designed to assess whether models accept plausible but unsupported visual attributes, states, or premises presented in the answer options. The text track contains 166 Korean text-only topic groups, yielding 498 questions, while the image track contains 168 Korean image-text topic groups, yielding 504 questions.

Data Construction. Figure 2 presents the GPT-assisted data construction pipeline of K-Prism. The overall process combines automatic candidate generation with human domain curation and validation. Annotators with native-level proficiency in Korean and sufficient domain knowledge first curate domains and source materials, and GPT-4o-mini (Hurst et al., 2024) is then used to generate

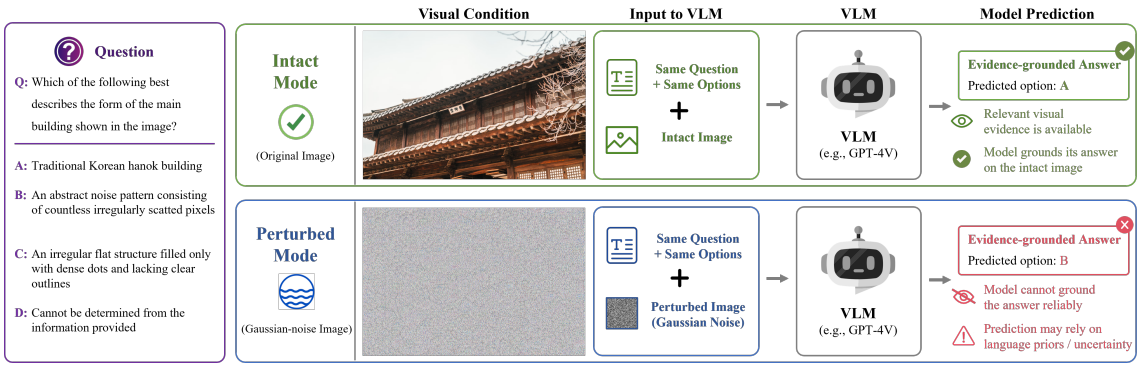


Figure 3: Example of the dual-mode evaluation protocol for the image track. Example question and answer options are English-translated from the original Korean item.

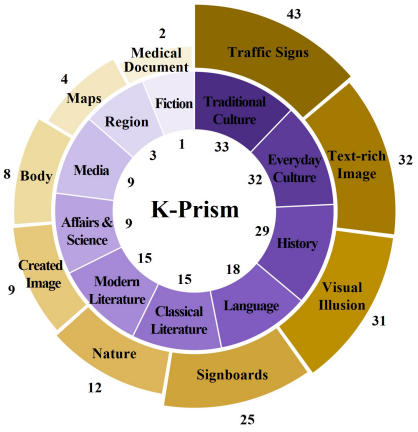


Figure 4: Domain composition of K-Prism: the text-track (inner ring) and image-track (outer ring).

candidate evaluation items. The generated items are not directly included in the benchmark; instead, annotators validate them to ensure factual correctness, linguistic naturalness, domain relevance, and consistency with the intended evaluation objective.

In the text-track, annotators curate Korean-specific domains and reference sources. GPT-4o-mini then generates candidate MCQs from the curated references. Annotators review the generated MCQs and retain only those that are grounded in the reference source, have a clearly defined answer, and are suitable for evaluating Korean-language understanding and hallucination-related failure patterns. For the image track, annotators curated visually grounded domains and source images, such as visual illusions and signboards, using everyday captures or manually created images. All sources and generated items were screened for personally identifiable information and offensive content. GPT-4o-mini first generates open-ended QA pairs from the source images, which are validated by annotators for visual grounding and answerability. The validated QA pairs are then converted into MCQs using GPT-4o-mini, followed by an additional validation step. This two-stage construction process helps ensure that the final image-track MCQs are both visually grounded and suitable for controlled hallucination evaluation. Appendix B provides detailed annotation and validation criteria for both tracks.

Dual-mode Evaluation. K-Prism evaluates each item under intact and perturbed modes. In the image track, intact mode uses the original image-text input, while perturbed mode replaces the image with same-resolution Gaussian noise and keeps the textual question and answer options unchanged. This setting serves as visual evidence ablation, testing whether models avoid unsupported visual claims when reliable visual evidence is unavailable. In the text track, where answer-relevant evidence remains textual, perturbed mode adds a fixed 224×224 noise image as a non-informative cross-modal distractor while leaving the textual input unchanged. Figure 3 illustrates the image-track protocol with an English-translated example using Qwen3-VL (Bai et al., 2025). Appendix F provides details on mode-specific gold labels and the evaluation prompt, Appendix D describes the Gaussian noise perturbation, and Appendix E presents additional examples from both text and image tracks.

Domain. Following multi-disciplinary benchmarks such as MMMU (Yue et al., 2024), K-Prism spans diverse knowledge and reasoning contexts, as shown in Fig. 4. For Korean-specific coverage, we refer to CLIcK (Kim et al., 2024) and KULTURE Bench (Wang et al., 2024), which emphasize Korean language, literature and socially grounded

Model	Text Track						Image Track					
	Intact Mode			Perturbed Mode			Intact Mode			Perturbed Mode		
	Factual	Reasoning	Misleading	Factual	Reasoning	Misleading	Factual	Reasoning	Misleading	Factual	Reasoning	Misleading
Closed-source												
GPT-4o-mini	85.7	100	2.3	83.9 (↓ 1.8)	100 (-)	0.6 (↓ 1.7)	95.1	78.3	81.3	22.8 (↓ 72.3)	27.7 (↓ 50.6)	4.8 (↓ 76.5)
Gemini-3.1-pro	67.2	70.8	25.6	77.3 (↑ 10.1)	87.5 (↑ 16.7)	0.0 (↓ 25.6)	57.2	52.4	68.0	7.2 (↓ 50.0)	21.0 (↓ 31.4)	52.4 (↓ 15.6)
Claude-4.7-opus	93.4	99.4	24.4	94.6 (↑ 1.2)	99.4 (-)	13.1 (↓ 11.3)	65.0	60.2	73.4	2.4 (↓ 62.6)	6.6 (↓ 53.6)	15.0 (↓ 58.4)
Open-source												
Qwen3-VL-4B-Thinking	83.3	96.4	20.8	72.6 (↓ 10.7)	89.8 (↓ 6.6)	1.7 (↓ 19.1)	89.8	86.1	89.8	11.4 (↓ 78.4)	22.3 (↓ 63.8)	27.7 (↓ 62.1)
Qwen3-VL-8B-Instruct	84.5	98.2	29.7	86.3 (↑ 1.8)	95.8 (↓ 2.4)	0.0 (↓ 29.7)	89.8	72.9	77.7	0.6 (↓ 89.2)	33.1 (↓ 39.8)	7.2 (↓ 70.5)
Qwen3-VL-8B-Thinking	79.1	96.4	29.7	76.7 (↓ 2.4)	92.2 (↓ 4.2)	0.0 (↓ 29.7)	81.9	77.7	78.3	18.7 (↓ 63.2)	27.1 (↓ 50.6)	24.1 (↓ 54.2)
Qwen3-VL-30B-A3B-Thinking	83.3	99.4	22.0	75.0 (↓ 8.3)	89.2 (↓ 10.2)	0.0 (↓ 22.0)	90.4	70.5	83.1	34.3 (↓ 56.1)	43.4 (↓ 27.1)	13.2 (↓ 69.9)
Qwen3-VL-32B-Thinking	88.6	99.4	29.7	79.7 (↓ 8.9)	92.2 (↓ 7.2)	0.6 (↓ 29.1)	71.7	60.8	77.1	31.9 (↓ 39.8)	28.3 (↓ 32.5)	38.0 (↓ 39.1)
HyperCLOVAX-SEED-Vision-Instruct-3B	80.9	92.2	0.0	77.9 (↓ 3.0)	91.0 (↓ 1.2)	0.0 (-)	71.1	59.0	76.5	31.9 (↓ 39.2)	78.9 (↑ 19.9)	29.5 (↓ 47.0)
LLaVA-OneVision-1.5-4B-Instruct	80.9	95.8	19.6	79.1 (↓ 1.8)	96.4 (↑ 0.6)	0.6 (↓ 19.0)	77.7	82.5	92.8	7.8 (↓ 69.9)	36.8 (↓ 45.7)	26.5 (↓ 66.3)
Deepseek-VL-7B-chat	29.1	14.2	5.8	28.5 (↓ 0.6)	2.9 (↓ 11.3)	2.3 (↓ 3.5)	50.0	46.3	6.0	0.0 (↓ 50.0)	1.2 (↓ 45.1)	1.2 (↓ 4.8)
A.X-4.0-VL-Light	86.3	97.6	0.6	86.3 (-)	98.8 (↑ 1.2)	0.0 (↓ 0.6)	81.3	83.1	87.3	0.0 (↓ 81.3)	13.9 (↓ 69.2)	5.4 (↓ 81.9)
Intern-VL3-14B	89.8	97.0	16.6	84.5 (↓ 5.3)	94.0 (↓ 3.0)	0.0 (↓ 16.6)	76.5	75.3	82.5	59.6 (↓ 16.9)	53.6 (↓ 21.7)	29.5 (↓ 53.0)
Intern-VL3-38B	82.7	97.0	32.7	76.7 (↓ 6.0)	89.8 (↓ 7.2)	5.9 (↓ 26.8)	74.1	74.1	83.1	29.5 (↓ 44.6)	86.7 (↑ 12.6)	69.8 (↓ 13.3)
Intern-VL3-38B-Instruct	82.7	97.0	31.5	76.7 (↓ 6.0)	91.0 (↓ 6.0)	7.1 (↓ 24.4)	75.3	77.1	80.7	50.0 (↓ 25.3)	85.5 (↑ 8.4)	74.7 (↓ 6.0)

Table 2: Evaluation results on K-Prism across text and image tracks. Each model is evaluated under Intact and Perturbed Modes for three failure dimensions: factual, reasoning, and misleading. All values indicate accuracy (%).

knowledge. While K-Prism adopts broad hallucination dimensions for comparability with prior work, these dimensions are operationalized through Korean-specific linguistic, cultural, historical, regional, and visual evidence. For visually grounded domains, we follow Korean text-rich VQA benchmarks such as KRETA (Hwang et al., 2025). The text track covers Traditional Culture, Everyday Culture, History, Language, Classical Literature, Modern Literature, Affairs & Science, Media, Region, and Fiction; the image track covers Traffic Signs, Text-rich Image, Visual Illusion, Signboards, Nature, Created Image, Body, Maps, and Medical Document.

3 Experiments

Evaluation Setup. We evaluate all models using an evaluation pipeline adapted from VLMEvalKit (Duan et al., 2024), an open-source toolkit for evaluating VLMs across proprietary API-based models and open-source models. In particular, we follow its parsing and scoring procedures for multiple-choice evaluation, where model responses are mapped to option choices and compared with the ground-truth answers. We adopt the multiple-choice system prompt used in MMMUPro (Yue et al., 2025) to ensure a consistent answer format across models. Each model is evaluated on both the text track and image track of K-Prism under two evaluation modes: intact mode and perturbed mode. For each track and mode, we report accuracy over three failure dimensions.

Main Results. We evaluate various VLMs, including GPT-4o (Hurst et al., 2024), Gemini-3.1 (Google DeepMind, 2026), Claude-4.7 (Anthropic, 2026), the Qwen3-VL series (Bai et al., 2025), HyperCLOVAX (NAVER, 2025), LLaVA-OneVision (Li et al., 2024), Deepseek (Lu et al., 2024), A.X-4.0 (Lab, 2025), and the Intern-VL3 series (Zhu et al., 2025). Table 2 shows that K-Prism reveals substantial variation across tracks, evaluation modes, and failure dimensions. Most models achieve high accuracy on text-track factual and reasoning questions, but show much lower performance on misleading questions. In the image track, the performance gap between intact and perturbed modes is considerably larger, indicating that visual perturbation exposes differences in model robustness and uncertainty-aware behavior. Several models exhibit strong text-track performance but sharply degrade under image perturbation, while some larger models, particularly in the Intern-VL3 family, maintain relatively high performance in perturbed mode. Aggregated across all models and tracks, performance drops by 29.35%p for factual questions, 17.70%p for reasoning questions, and 32.59%p for misleading questions under perturbation. To assess the uncertainty of these changes, paired item-level bootstrap analysis yields an average standard deviation of 3.42 percentage points across all combinations of models, tracks, and diagnostic dimensions. For HyperCLOVAX and LLaVA-OneVision, whose text-track accuracies change only slightly, the corresponding standard deviations range from 0.60 to 1.03 percentage points, indicating low perturbation sensitivity. These non-uniform drops indicate that perturbed-mode accuracy reflects dimension-specific evidence sensitivity rather than a single notion of visual robustness. In particular, the large degradation on misleading items suggests that models often over-commit to plausible but unsupported premises when reliable evidence is weakened or absent. These results support K-Prism's decomposed evaluation across

modalities, evidence conditions, and failure dimensions. Appendix A provides a detailed analysis of the experiments, while Appendix C presents the Korean-to-English translation ablation and Appendix D reports additional experiments with occluded and black-image inputs.

4 Conclusion

We presented K-Prism, a Korean benchmark for multi-faceted hallucination analysis in vision-language models. K-Prism evaluates factual errors, reasoning failures, and susceptibility to misleading information under two controlled modes, intact mode and perturbed mode, across text-track and image-track datasets. Built through a GPT-assisted pipeline with human curation and validation, K-Prism covers Korean-specific textual and visually grounded image domains, comprising 166 and 168 topic groups, respectively. Experiments on closed- and open-source VLMs show substantial variation across tracks, modes, and failure dimensions, highlighting the need for decomposed and controlled evaluation of hallucination in VLMs.

Limitations

K-Prism focuses on diagnosing hallucination behaviors rather than proposing a mitigation method. The benchmark decomposes hallucination into factual errors, reasoning failures, and susceptibility to misleading information, and evaluates model behavior under intact and perturbed modes. However, it does not directly introduce training or inference-time strategies to reduce these failures. Future work can use K-Prism to develop mitigation strategies tailored to each failure dimension. For factual errors, retrieved evidence or external knowledge can be used to verify whether the selected answer is supported by reliable sources. For reasoning failures, models can be encouraged to generate intermediate reasoning steps and check their consistency before producing the final answer. For misleading cases, models can be trained to detect insufficient or corrupted visual evidence and abstain from making visually grounded claims when the image does not provide reliable information.

Additionally, K-Prism adopts a multiple-choice evaluation format for scalable and consistent comparison across models. Although this design enables reliable automatic scoring, it may not fully capture hallucination behaviors in open-ended generation, such as unsupported rationales, fabricated details, or partially correct but misleading explanations. In future work, we plan to extend K-Prism to free-form generation by converting the existing items into open-ended prompts and evaluating generated claims as supported, contradicted, or unverifiable based on reference answers and evidence annotations, with human or automatic scoring. This extension would enable a more comprehensive evaluation of hallucination behaviors beyond multiple-choice prediction.

K-Prism prioritizes controlled diagnostic comparison over large-scale coverage. By organizing factual, reasoning, and misleading questions within shared-evidence topic groups, the benchmark reduces confounds arising from differences in topic or image difficulty. However, its current scale of 1,002 questions limits the breadth of coverage. Future work will expand the benchmark across additional topics and domains while preserving the shared-evidence design.

Use of AI Assistants

We used AI assistants, including GPT-based assistants, during dataset construction and model evaluation. In dataset construction, AI assistants were used to generate initial multiple-choice question candidates, refine question wording, and produce distractor options based on curated source materials and annotation guidelines. All generated items were manually reviewed, revised, or discarded by human annotators before inclusion in K-Prism. Selected visual-illusion, created-image, and body sources were GPT-4o-mini-assisted and manually screened.

We also used closed-source vision-language models through their official interfaces or APIs as evaluation targets in our experiments. These models were queried only to obtain responses to the benchmark questions under the specified intact and perturbed evaluation modes. The authors retained full responsibility for the dataset design, annotation validation, experimental setup, result analysis, and final manuscript content.

Acknowledgments

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI)

References

Samir Abdaljalil, Parichit Sharma, Erchin Serpedin, and Hasan Kurban. 2026. Halluverse-M^3: A multitask multilingual benchmark for hallucination in llms. *arXiv preprint arXiv:2602.06920*.

Anthropic. 2026. Claude 4.7 claude system card. Technical report, Anthropic.

Jesse Atuhurra, Iqra Ali, Tomoya Iwakura, Hidetaka Kamigaito, and Tatsuya Hiraoka. 2025. Vlures: Benchmarking vlm visual and linguistic understanding in low-resource languages. *arXiv preprint arXiv:2510.12845*.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Google DeepMind. 2026. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/. Accessed: 2026-05-19.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Taebaek Hwang, Minseo Kim, Gisang Lee, Seonuk Kim, and Hyunjun Eun. 2025. Kreta: A benchmark for korean reading and reasoning in text-rich vqa attuned to diverse visual contexts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33421—-33432.

Bidur Khanal, Sandesh Pokhrel, Sanjay Bhandari, Ramesh Rana, Nikesh Shrestha, Ram B Gurung, Cristian Linte, Angus Watson, Yash R Shrestha, and Binod Bhattarai. 2025. Hallucination-aware multimodal benchmark for gastrointestinal image analysis with large vision-language models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 235–245. Springer.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346.

SKT AI Model Lab. 2025. A.x 4.0 vl light.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, and 1 others. 2024. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*.

NAVER. 2025. HyperCLOVAX-SEED-Vision-Instruct-3B. Accessed: 2026-09-18.

Philip Wootaek Shin, Ajay Narayanan Sridhar, Lakshmi Sivani Devarapalli, Rui Zhang, Jack Sampson, and Vijaykrishnan Narayanan. 2026. When relations break: Analyzing relation hallucination in vision-language model under rotation and noise. In *Proceedings of the 4th Workshop on Advances in Language and Vision Research (ALVR)*, pages 180–185.

Jingqun Tang, Qi Liu, Yongjie Ye, Jinghui Lu, Shu Wei, An-Lan Wang, Chunhui Lin, Hao Feng, Zhen Zhao, Yanjie Wang, and 1 others. 2025. Mtvqa: Benchmarking multilingual text-centric visual question answering. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7748–7763.

Jiawei Wang, Yushen Zuo, Yuanjun Chai, Zhendong Liu, Yicheng Fu, Yichun Feng, and Kin-Man Lam. 2025. Safeguarding vision-language models: Mitigating

vulnerabilities to gaussian noise in perturbation-based attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2773–2782.

Junyang Wang, Yuyang Wang, Guohai Xu, Jingang Zhang, Yucheng Gu, Haiyang Jia, Ming Yan, Wei Ji, and Jitao Qiao. 2023. Amber: An llm-free multi-dimensional benchmark for object hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xiaonan Wang, Jinyoung Yeo, Joon-Ho Lim, and Hansaem Kim. 2024. Kulture bench: A benchmark for assessing language model in korean cultural context. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 914–927.

Weihao Xuan, Qingcheng Zeng, Heli Qi, Junjue Wang, and Naoto Yokoya. 2025. Seeing is believing, but how much? a comprehensive analysis of verbalized calibration in vision-language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1408–1450.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, and 1 others. 2025. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Additional Analysis

Model-wise Analysis. Table 2 shows that strong performance in the text-track does not necessarily translate to robustness in the image-track. Many models, including proprietary ones, perform well under intact inputs but degrade substantially when visual evidence is perturbed, particularly on misleading questions. Open-source models show larger variation, with some relying heavily on textual priors or fragile visual grounding. The InternVL3 family is comparatively more robust in the perturbed setting, suggesting better resistance to corrupted visual evidence and over-confident hallucination.

Track-wise analysis. The text-track and image-track exhibit clearly different difficulty profiles. In the text-track, factual and reasoning questions tend to be easier because the required evidence is primarily contained in the textual input or can be inferred from it. However, misleading questions show substantially lower accuracy, demonstrating that models can still be misled by questions that encourage unsupported assumptions even without informative visual evidence. This indicates that hallucination is not limited to image-grounding failures, but can also arise from language-driven priors and question-level bias.

The image-track reveals a stronger gap between intact and perturbed modes. When the original image is provided, many models can exploit visual evidence to answer factual and reasoning questions correctly. However, when the image is replaced with non-informative perturbation, performance drops more sharply. This trend shows that the image-track is more effective at exposing whether a model can distinguish valid visual evidence from corrupted or absent evidence. In particular, the large degradation on misleading questions suggests that models often fail to abstain or select uncertainty-aware answers when the visual input no longer supports the question.

Domain-wise Analysis. Figure 5 compares the domain-wise performance of closed-source models across intact and perturbed modes. Overall, the results show that visual perturbation leads to a clear performance drop in many visually grounded domains, such as Text-rich Image, Traffic Signs, Created Image, Body, Signboards, Medical Document, Nature, and Map. This indicates that these domains require reliable visual evidence, and replacing the original image with non-informative perturbation substantially weakens the models' ability to select the correct answer.

However, the effect of perturbation is not uniform across domains. In several domains, including Affairs & Science, Media, Fiction, and Classical Literature, the perturbed-mode accuracy is comparable to or higher than the intact-mode accuracy. This increase should not be interpreted as improved visual understanding under perturbation. Rather, it suggests that these domains contain questions that can often be solved using textual cues, world knowledge, or strong language priors, even when

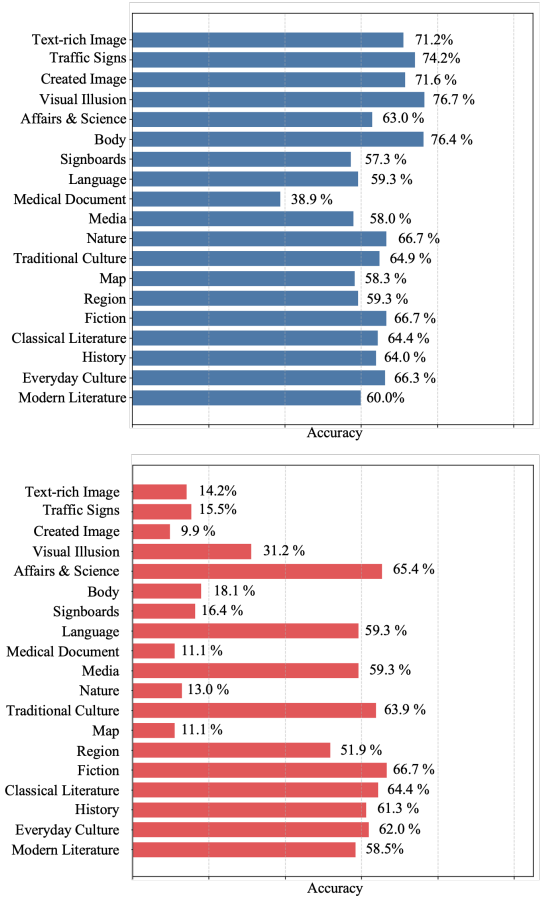


Figure 5: Domain-wise accuracy comparison of closed-source VLMs under intact and perturbed modes. The upper panel shows intact mode performance, while the lower panel shows perturbed mode performance.

the visual input is corrupted. In such cases, the original image may provide limited additional evidence, or may even introduce distracting visual signals, while the perturbed input encourages the model to rely more heavily on its language-based prior. Therefore, the increased accuracy under perturbation reflects reduced dependence on visual grounding rather than genuine robustness to visual corruption.

These results support the need for domain-wise and condition-wise analysis in hallucination evaluation. A single aggregate score may obscure the fact that some domains are highly sensitive to visual perturbation, whereas others are dominated by textual or parametric knowledge. The contrast between visually dependent domains and knowledge-oriented domains shows that hallucination behavior is not homogeneous across inputs, but varies according to the type of evidence required by each domain.

Comparison across failure dimensions. The three failure dimensions show different degradation patterns under perturbation. Factual questions exhibit a large drop, indicating that even visually grounded recognition or retrieval-style questions become unreliable when the visual evidence is corrupted. Reasoning questions show a comparatively smaller drop, suggesting that some models may rely more on textual reasoning or prior knowledge when solving these questions. Misleading questions show the largest degradation, indicating that models are especially vulnerable when the question itself encourages an unsupported interpretation. These differences confirm that hallucination should not be treated as a single homogeneous behavior, since each dimension captures a different source of model failure.

Scaling within model families. Within the same model family, larger models generally show stronger overall performance, but the improvement is not uniform across tracks and failure dimensions. In particular, larger variants tend to improve factual and reasoning accuracy in the intact setting, reflecting stronger language understanding and multimodal perception. However, the gains under perturbed settings are less consistent. This suggests that increasing model size can improve general task accuracy, but does not automatically guarantee robustness to corrupted visual inputs or misleading questions.

The Intern-VL3 series shows a relatively clear size-related trend, where larger variants maintain stronger performance across both intact and perturbed modes. This indicates that scaling may help improve robustness when combined with effective multimodal training. However, the remaining performance gap between intact and perturbed modes shows that even larger models are not fully robust to visual perturbation. Similar trends can be observed in other model families, where larger models often perform better than smaller variants, but still remain vulnerable to misleading or uncertainty-sensitive questions.

These model-wise and track-wise results further support the design motivation of K-Prism. A single aggregate score would obscure important differences between models that perform similarly on clean inputs but behave very differently under perturbation. By separating evaluation across text/image tracks, intact/perturbed modes, and factual/reasoning/misleading dimensions, K-Prism enables a

more fine-grained diagnosis of hallucination behavior in VLMs.

B Annotation Guidelines

Annotators. Annotation and validation were conducted by three annotators among the paper authors, all of whom are native Korean speakers or have native-level proficiency in Korean. Each generated item was reviewed by at least two annotators. Disagreements were resolved through discussion, and items without a clear consensus were revised or removed from the benchmark. No external crowdworkers or paid participants were recruited; annotation was conducted as part of the authors' research activities.

General validation criteria. Annotators were instructed to validate all generated items before inclusion in K-Prism. They checked whether each item was grounded in the provided source evidence, had a clearly identifiable answer, and matched the intended diagnostic dimension. Items were revised or discarded if they were ambiguous, unanswerable, linguistically unnatural, factually incorrect, or inconsistent with the intended evaluation objective.

Text-track validation. For the text track, annotators reviewed each generated MCQ with respect to the curated Korean reference source. They checked whether the question was answerable from the reference, the correct option was uniquely supported, the distractors were plausible but clearly incorrect, the question and options were written in natural Korean, and the item matched the intended diagnostic dimension. Factual questions were required to test explicitly supported evidence-grounded information; reasoning questions required a valid inference beyond surface matching; and misleading questions included an unsupported assumption or misleading cue while still allowing the correct answer to be determined from the available evidence.

Image-track validation. For the image track, annotators reviewed both source images and generated QA pairs. They checked whether the image contained sufficient visual evidence, the question was visually grounded and answerable, the answer was consistent with the image, and the final options were clear and non-overlapping. Items were revised or removed if the visual evidence was unclear, multiple options could be correct, or the question could be answered solely from language priors without using the image.

Dual-mode validation. For both tracks, annotators checked whether the final MCQs were suitable for the dual-mode evaluation protocol. The intact-mode answer was required to be supported by the original input, whereas the perturbed-mode setting was designed to test whether models avoid unsupported claims when the visual evidence is corrupted or absent. Perturbed-mode questions remained linguistically identical to intact-mode questions, except that the visual input was replaced with Gaussian noise.

Inter-annotator Agreement and Adjudication. We measured inter-annotator agreement using Fleiss' κ based on the validation criteria described above. The overall agreement across the three annotators was $\kappa = 0.84$. For the misleading dimension, which involves more subtle judgments regarding unsupported assumptions and misleading cues, the agreement was $\kappa = 0.79$ for the text track and $\kappa = 0.71$ for the image track. Disagreements were resolved through discussion among the annotators, and items were revised or removed when a consensus could not be reached. After validation and adjudication, approximately 80% of the candidate items were approved for inclusion in the final benchmark.

C Translation Ablation.

To examine whether K-Prism can be reliably evaluated through translation, we translated the benchmark from Korean into English using Gemini-3.1-pro and evaluated the same models under the identical experimental setting. As shown in Table 3, the English-translated version exhibits different behavior under perturbation. Compared with the original Korean benchmark, the average perturbed accuracy decreases by 4.63 percentage points on the text track and 7.43 percentage points on the image track. These results suggest that perturbation sensitivity is not preserved under translation, indicating that evaluation on a translated English version does not fully reflect model robustness in the original Korean setting. This further motivates the need for a Korean-specific benchmark that preserves its original linguistic, cultural, regional, and visually grounded context.

D Gaussian Noise Perturbation

We use Gaussian noise as a controlled visual perturbation to degrade task-relevant visual evidence while preserving the textual question unchanged.

Model	Text Intact Avg.	Text Perturbed Avg.	Image Intact Avg.	Image Perturbed Avg.
Closed-source				
GPT-4o-mini	63.49	60.32	83.13	17.07
Gemini-3.1-pro	57.14	52.95	58.03	23.13
Claude-4.7-opus	71.83	63.69	66.06	6.63
Open-source				
Qwen3-VL-4B	64.68	49.60	80.72	12.49
Qwen3-VL-8B-Instruct	67.06	56.94	74.70	14.06
Qwen3-VL-8B-Thinking	61.71	48.02	70.68	25.10
Qwen3-VL-30B	71.83	51.79	75.30	25.39
Qwen3-VL-32B	77.98	53.57	73.49	26.91
HyperCLOVAX	56.35	50.60	82.93	21.29
LLaVA-OneVision-1.5-4B	64.48	60.12	81.33	12.85
Deepseek-VL-7B	60.71	8.75	79.32	8.03
A.X-4.0	62.30	61.51	86.75	6.02
Intern-VL3-14B	68.07	49.50	81.73	30.92
Intern-VL3-38B	60.84	47.47	73.29	42.97
Intern-VL3-38B-Instruct	79.52	48.27	74.30	46.79

Table 3: Model performance on the English-translated version of K-Prism. All values denote accuracy (%).

Model	Text Avg. (Black)	Image Avg. (Black)	Image Avg. (Occlusion)
Closed-source			
GPT-4o-mini	60.70	23.30	11.23
Gemini-3.1-pro	54.70	30.30	16.37
Claude-4.7-opus	68.87	11.23	6.43
Open-source			
Qwen3-VL-4B	63.90	31.30	2.00
Qwen3-VL-8B-Instruct	70.20	24.10	8.00
Qwen3-VL-8B-Thinking	66.10	33.50	14.30
Qwen3-VL-30B	64.80	40.20	18.10
Qwen3-VL-32B	67.90	42.30	30.10
HyperCLOVAX	67.00	56.10	37.90
LLaVA-OneVision-1.5-4B	68.00	34.40	16.50
Deepseek-VL-7B	20.80	11.20	3.20
A.X-4.0	71.60	16.50	2.00
Intern-VL3-14B	69.70	57.40	38.10
Intern-VL3-38B	68.00	71.50	49.50
Intern-VL3-38B-Instruct	69.10	79.30	59.50

Table 4: Model-wise average accuracy (%) under additional visual perturbations. Black-image replacement is applied to both the text and image tracks, while occlusion is applied only to the image track.

Prior studies have shown that visual corruptions, including noise, can affect visual grounding and multimodal reasoning, suggesting that VLMs may rely on learned language priors or superficial cues when visual evidence becomes unreliable (Wang et al., 2025; Shin et al., 2026; Favero et al., 2024). In our setting, the perturbed mode is not intended to simulate all possible real-world corruptions, but rather to provide a standardized stress test for examining whether VLMs continue to produce confident answers when reliable visual evidence is unavailable. This design allows K-Prism to measure hallucination susceptibility arising from over-reliance on language priors, answer-option biases, or misleading cues under visually unreliable conditions.

Additional Perturbation Analysis. To examine whether the observed evidence sensitivity is specific to Gaussian noise, we conduct additional perturbation experiments using black-image replacement and occlusion. Black-image replacement is applied to both the text and image tracks. For the text track, the black image serves as a non-informative visual input, whereas for the image track, it replaces the original task-relevant visual evidence. Occlusion is applied only to the image track because it requires an original source image. We randomly place multiple black boxes over the source image while limiting the total occluded area to at most 30% of the image. The textual question and answer options remain unchanged under both perturbation settings. Table 4 reports the model-wise average accuracy under black-image replacement and occlusion.

Compared with Gaussian noise, which results

in average performance drops of 7.59%p on the text track and 45.50%p on the image track, black-image replacement produces a negligible 0.33%p increase on the text track and a 36.73%p drop on the image track. Occlusion produces a 52.69%p drop on the image track. Although the magnitude of degradation varies across perturbation types, both black-image replacement and occlusion reproduce substantial performance degradation on the image track. These results indicate that the visual evidence-sensitivity observed in K-Prism is not specific to Gaussian-noise perturbation. In contrast, the negligible change under black-image replacement on the text track is consistent with the fact that task-relevant evidence in this track is primarily provided by the textual input.

E Examples

Text-track Perturbed-mode Examples Figure 6 presents representative text-track examples under the perturbed mode. In the text track, the textual question and answer options are kept unchanged, while a non-informative visual input is provided to maintain a consistent multimodal input format. Since the textual evidence remains unchanged, the gold answer is identical in both intact and perturbed modes. Blue indicates the shared gold answer, and red indicates the model prediction under the perturbed mode. The examples show that perturbation can induce different failure patterns even when the answer is available from the text. In the factual example, the model selects an uncertainty option although the correct answer is supported by the given evidence. In the reasoning example, the model chooses an anachronistic distractor, indicating a failure to reason over the historical relation required by the question. In the misleading example, the model selects a noise-based description instead of recognizing that no specific visual evidence supports the described performance scene. These examples illustrate that the text-track perturbed setting can reveal over-uncertainty, reasoning errors, and susceptibility to visually irrelevant cues.

Image-track Perturbed-mode Examples. Figure 7 shows representative image-track examples under the perturbed mode. In the image track, the original image contains the visual evidence needed to answer the question, whereas the perturbed condition removes or corrupts this evidence. We therefore distinguish the evidence-grounded answer under the intact image from the expected answer under the perturbed condition. Green indicates the gold answer under the intact mode, blue indicates the gold answer under the perturbed mode, and red indicates the model prediction under visually perturbed input. In the factual example, the intact image supports the actual geometric composition, whereas the perturbed condition makes the uncertainty option appropriate; however, the model predicts a noise-like pattern. In the reasoning example, the intact image supports the inference that the subway line is Line 7, whereas the perturbed condition no longer provides the required station-numbering evidence; nevertheless, the model selects an unsupported explanation caused by the perturbed visual condition. In the misleading example, the distractor option introduces an unsupported claim that the relevant colors or ingredients cannot be identified, even though the intact image provides clear visual evidence. The model selects this misleading option under perturbation, showing susceptibility to option-level misleading cues. Together, these cases illustrate that visual perturbation can shift model predictions away from evidence-grounded answers in different ways, including visual grounding failure, reasoning failure, and susceptibility to misleading assumptions introduced by answer options.

F Uncertainty Option Design

All 1,002 questions in K-Prism include an explicit uncertainty option among the multiple-choice answer choices. This design is consistently applied to both the text and image tracks, as illustrated in Figs. 6 and 7. The uncertainty option is intended for cases where the provided evidence is insufficient to support a definitive answer. Model responses are evaluated based on whether they select this predefined uncertainty option when appropriate, rather than through free-form expressions of uncertainty.

Perturbed-Mode Gold-Label Assignment. K-Prism uses different gold-label assignments depending on the evaluation mode. In the intact mode, the gold label corresponds to the answer supported by the original textual or visual evidence. In the perturbed mode, the controlled perturbation removes the information needed to determine the original answer. The explicitly provided uncertainty option is used as the gold label.

A model response is considered correct in the

Q: 대동법 시행 이후 조선 후기 경제 구조 변화에 대한 설명으로 가장 적절한 것은 무엇인가? *(What is the most appropriate description of the change in the economic structure of late Joseon after the implementation of the Daedong Law?)*

A: 공납이 유지되면서 지방 특산물 생산이 강화되었다. *(Local specialty production was strengthened as the tribute system was maintained.)*
B: 공납을 쌀로 통일함으로써 공인이 등장하고 상업 유통이 활성화되었다. *(As tribute payments were unified into rice, government-licensed merchants emerged and commercial distribution became more active.)*
C: 세금이 폐지되어 자급자족 경제로 회귀하였다. *(Taxes were abolished, and the economy returned to a self-sufficient system.)*
D: 주어진 정보로 알 수 없음. *(Cannot be determined from the given information.)*

(a) Factual

Q: 한글 창제 이전에 향찰과 이두 같은 차자 표기법이 사용된 배경으로 가장 적절한 것은 무엇인가? *(What is the most appropriate reason why borrowed-character writing systems such as Hyangchal and Idu were used before the creation of Hangul?)*

A: 고유 문자가 없던 상황에서 한자의 음과 뜻을 빌려 한국어 어순과 표현을 기록하려는 필요가 있었기 때문이다. *(Because, in the absence of an indigenous writing system, there was a need to record Korean sounds, meanings, and expressions by borrowing Chinese characters.)*
B: 중국어를 한국어보다 우월한 언어로 만들기 위해 모든 구어 표현을 금지했기 때문이다. *(Because all native Korean expressions were prohibited in order to make Chinese superior to Korean.)*
C: 훈민정음이 이미 널리 보급되어 있었지만 이를 대체하기 위해 만들어졌기 때문이다. *(Because Hunminjeongeum was already widely distributed but was created to replace it.)*
D: 주어진 정보로 알 수 없음. *(Cannot be determined from the given information.)*

(b) Reasoning

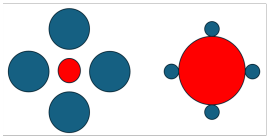
Q: 경기민요를 부르는 소리꾼의 복식이나 공연 무대의 시각적 형태는 어떻게 묘사되어 있는가? *(How is the visual appearance of the Gyeonggi folk song singer's costume or performance stage described?)*

A: 불규칙하게 흩어진 무수한 픽셀 형태의 추상적인 노이즈 패턴이 식별된다. *(An irregularly tangled abstract noise pattern composed of numerous pixels is visible.)*
B: 화려하고 깔끔한 전통 한복을 차려입은 소리꾼이 무대 중앙에서 경쾌한 손짓과 함께 노래하는 모습이 식별된다. *(A singer wearing a bright and neat traditional hanbok is visible at the center of the stage, singing with a folding fan.)*
C: 화면 전체가 동일한 단일 색상으로 덮여 있어 구체적인 형체를 전혀 파악할 수 없다. *(The entire image is a uniform single color, so no specific shape can be identified.)*
D: 어떠한 경기민요 소리꾼의 자태나 구체적인 공연 무대의 형태도 전혀 식별되지 않는다. *(No posture of a Gyeonggi folk song singer or specific form of a performance stage can be identified.)*

(c) Misleading

Figure 6: Text-track examples under the perturbed mode. Blue indicates the shared gold answer for both intact and perturbed modes, and red indicates the model prediction under the perturbed mode.

Q: 이미지에 나타난 주요 구성 요소로 가장 적절한 것은 무엇인가? *(Which of the following best describes the form of the main components shown in the image?)*



A: 왼쪽에는 작은 붉은색 원 주변에 파란색 원들이 배치되어 있고, 오른쪽에는 큰 붉은색 원 주변에 작은 파란색 원들이 배치되어 있다.*(On the left, blue circles are arranged around a small red circle, while on the right, small blue circles are arranged around a large red circle.)*
B: 불규칙하게 흩어진 무수한 픽셀 형태의 추상적인 노이즈 패턴 *(An abstract noise pattern consisting of countless irregularly scattered pixels)*
C: 명확한 외곽선 없이 촘촘한 점들로만 채워진 불규칙한 평면 구조 *(An irregular flat structure filled only with dense dots and lacking clear outlines)*
D: 이미지 내에 특정 객체나 구조를 식별할 수 있는 시각적 정보가 전혀 존재하지 않는다. *(There is absolutely no visual information in the image that would allow one to identify any specific objects or structures.)*

(a) Factual

Q: 이미지 속 지하철 노선도에 표시된 역 번호 체계를 기반으로 이 노선에 대한 설명으로 가장 적절한 것은 무엇인가? *(Based on the station numbering system shown on the subway map in the image, which of the following is the most appropriate description of this line?)*



A: 가우시안 노이즈 특유의 시각적 결함 때문에 해당 질문이 요구하는 대상을 시각적으로 재구성할 수밖에 없다. *(Due to the visual artifacts characteristic of Gaussian noise, we have no choice but to visually reconstruct the object described in the question.)*
B: 검은색 화면으로 인해 시각적 단서가 소거된 상태에서 무작위적인 속성으로 단정지어진 결과이다. *(This is the result of a conclusion drawn based on random attributes, with visual cues eliminated due to the black screen.)*
C: 첫 자리가 7로 시작하는 7호선이다. *(It's Line 7, which starts with the number 7)*
D: 이미지 내에 질문에서 지목하는 대상의 형태나 시각적 구조물이 아예 존재하지 않는다. *(The image contains absolutely no form or visual structure corresponding to the subject mentioned in the question.)*

(b) Reasoning

Q: 이미지 속 김밥에 포함된 재료의 색상으로 가장 적절한 것은 무엇인가? *(Which of the following best matches the colors of the ingredients in the gimbap shown in the image?)*



A: 실제 시각적 특징과 맥락을 종합 분석할 때, 노란색 계란, 초록색 오이, 흰색 밥 등이 함께 포함된 다양한 색상이 객관적 사실로 관찰된다. *(When conducting a comprehensive analysis of actual visual features and context, a variety of colors—including yellow eggs, green cucumbers, and white rice—are observed as objective facts.)*
B: 균일한 검은색으로만 가득 차 있어 김밥 내부 재료들이 모두 어두운 흑색 계열로만 보인다. *(Because it is filled entirely with a uniform black color, all the ingredients inside the gimbap appear to be dark shades of black.)*
C: 소금-후추 형태의 뒤섞인 회색 잡음(노이즈)으로 뒤덮여 있어 재료의 색상을 명확히 구분할 수 없는 상태이다. *(It is covered in a mottled gray pattern resembling salt and pepper, making it impossible to clearly distinguish the colors of the ingredients.)*
D: 이미지 내에 색상을 특정하거나 구분할 수 있는 뚜렷한 시각적 객체나 배경 요소가 전혀 존재하지 않는다. *(There are no distinct visual objects or background elements within the image that can be used to identify or distinguish colors.)*

(c) Misleading

Figure 7: Image-track examples under the perturbed mode. Green indicates the gold answer under the intact mode, blue indicates the gold answer under the perturbed mode, and red indicates the model prediction under visually perturbed input.

perturbed mode when it selects the designated uncertainty option from the provided answer choices. This allows the uncertainty option to be evaluated using the same multiple-choice scoring framework as the other answer choices.

Evaluation Prompt. Table 5 presents the prompt template used for multiple-choice evaluation. Following the direct multiple-choice evaluation setting of MMMU-Pro, models are instructed to select an answer from the provided options. We adapt the output format to a structured JSON response to ensure consistent parsing across different model families.

Multiple-Choice Evaluation Prompt
`Question: {question}` Passage: {passage} Options: 1. {option 1} 2. {option 2} ... Rules: - If there is one correct answer, include only one item in the answers array. - Output exactly one line of JSON. Do not include any additional text, explanation, or Markdown. Output format: {"answers":[3]} Answer JSON:

Table 5: Prompt template used for multiple-choice evaluation.