LogiScope-VQA : Benchmarking Vision-Language Models for Logistics Hazard Identification in Industrial Scenarios

Anonymous authors
Paper under double-blind review

Abstract

Large Multimodal Models (LMMs) large-scale deployment in industrial warehouse settings specifically necessitates that models exhibit human-expert-level hazard-oriented perception, understanding, and reasoning capabilities. However, the scarcity of real industrial data, tightly coupled to commercial terms, significantly hampers further advancement. To bridge this gap, we curate **LogiScope-VQA** to investigate the practical applicability of mainstream LMMs in real-world logistics operations. LogiScope-VQA comprises 2,476 images and 2,918 videos primarily sourced from real-world logistics parks, along with 10,274 VQAs meticulously curated and validated by human annotators. Grounded in 18 core objects and 20 risk types, we devise 39 subtasks aligned with three principal themes: industrial element perception, warehouse knowledge understanding, and potential risk reasoning. Furthermore, we incorporate dynamic thinking-budget configurations and dual-dimensional risk bias analyses to elucidate the properties of LMMs. Extensive experiments unveil that even powerful proprietary models, including GPT-5.5, Gemini-3.1-Pro, and Claude-Opus-4.7, exhibit a significant gap relative to human performance. The unique challenge of jointly integrating perception, understanding, and reasoning for hazard identification poses substantial headroom for further improvement on LogiScope-VQA. We additionally reveal the pervasive security bias issue that impedes LLMs' practical deployment in real-world settings. We hope LogiScope-VQA will stimulate the community to build next generation multimodal foundation models practically deployable in authentic industrial logistics scenarios. We provide evaluation code at the anonymous repository and upload representative data samples as supplementary material. The full project will be fully open-sourced upon paper acceptance.

1 Introduction

Large Multimodal Models (LMMs) have achieved remarkable success on a broad spectrum of question-answering tasks in naturalistic settings, showcasing strong perceptual and reasoning abilities as well as extensive embedded knowledge. Nevertheless, the limited evaluation of LMMs in industrial scenarios has hindered the conversion of this potential into practically deployable systems and tools. Industrial environments pose unique challenges that are largely underexplored in current benchmarks, particularly due to safety-oriented task requirements and the relatively dense spatial distribution of targets.

Logistics warehousing is a representative industrial scenario in which continuous monitoring and timely safety risk recognition are crucial for accident prevention, regulatory compliance, and efficient warehouse operations. Previous studies have begun to develop multimodal benchmarks for industrial logistics, but their applicability to real-world assessment remains constrained by two inherent obstacles: *(i)* In terms of *evaluation depth*, industrial logistics environments naturally raise concerns regarding commercial cost, privacy, safety, and liability. Consequently, IndustryEQA instead relies on physically simulated data produced within virtual environments. Collecting data from real-world industrial environments for systematic evaluation remains a significant gap. *(ii)* In terms

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

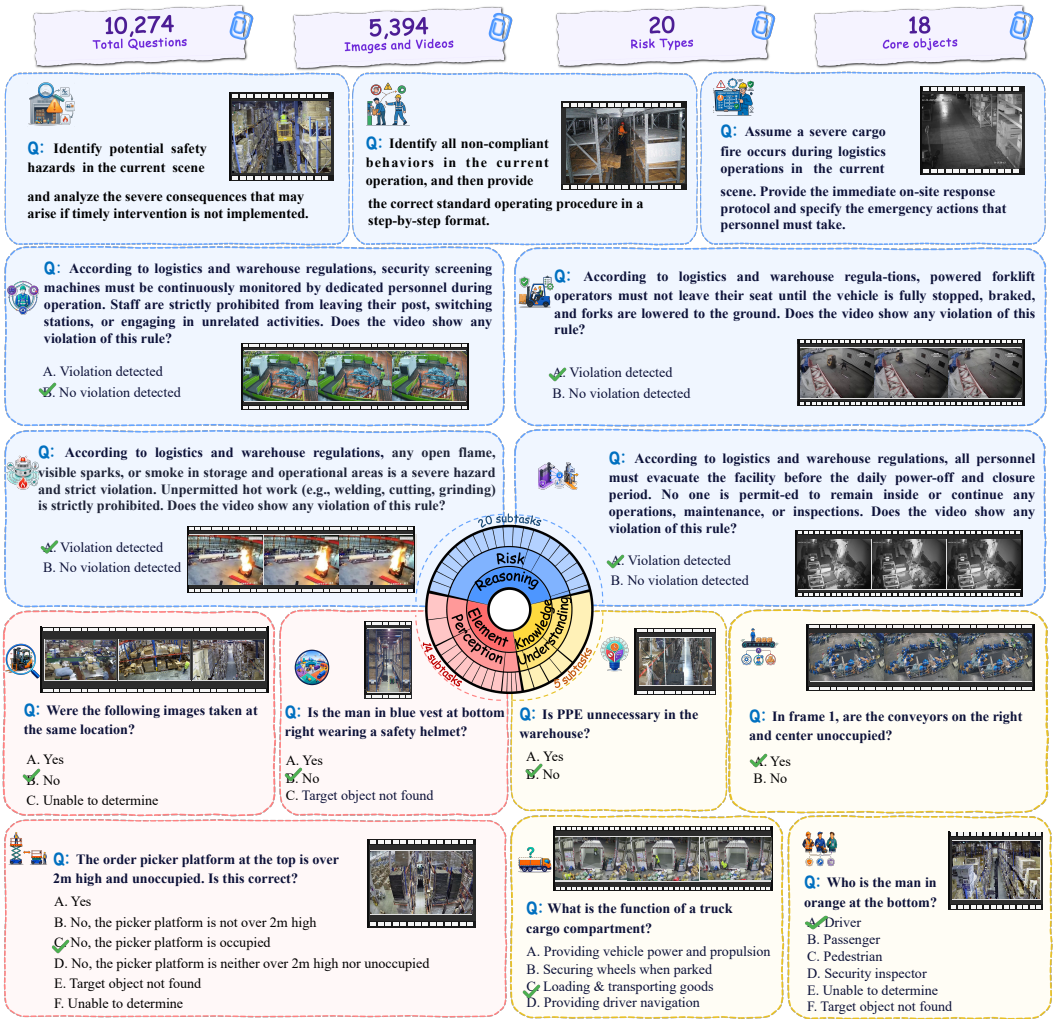


Figure 1: Overview diagram the LogiScope-VQA, evaluating LMMs in real-world industrial scenarios across three key competency dimensions: industrial element perception, warehouse knowledge understanding, and potential risk reasoning. (The full task taxonomy is provided in Appendix B.3.)

of *evaluation breadth*, practical deployment of LMMs involves diverse safety-critical dimensions, including object perception, spatial reasoning, path planning, and action recognition. However, ARMBench and iSafetyBench are limited to evaluating specific tasks under a single question format. Comprehensive multi-task evaluation of logistics safety-oriented perception and reasoning capabilities has yet to be established. Therefore, the safety-oriented assessment of mainstream LMMs in real-world industrial logistics scenarios remains an open question.

To systematically evaluate LMMs for logistics safety inspection, we introduce **LogiScope-VQA**, an industrial benchmark consisting of 2,476 images, 2,918 videos, 1,0274 VQAs, 20 risk factors, and 18 logistics objects. As illustrated in Fig. 1, we adopt a progressive curriculum to broadly assess the deployment value of LMMs in real-world industrial scenarios. For *industrial element perception*, 3.5 million surveillance clips collected from the global intelligent logistics company Cainiao capture the dynamics and heterogeneous operational scenarios within logistics parks. For *warehouse knowledge understanding*, recruited logistics experts collectively devote 52 person-days to annotating essential factors, including spatial locations, object attributes, hazardous actions, etc. For *potential risk reasoning*, we additionally incorporate open-ended problems and an LLM-as-a-Judge evaluation protocol to assess LMMs' reliability and fairness in sophisticated thinking. Coupled evaluation tasks pose fundamentally new challenges to LMMs in satisfying the demands of industrial logistics deployment.

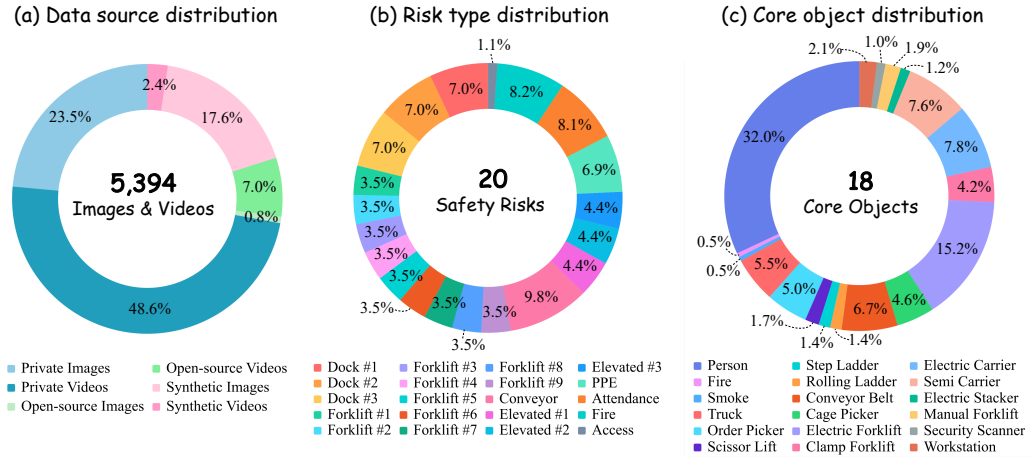


Figure 2: Core statistical distributions in LogiScope-VQA, comprehensively covering 5,394 visual samples, 20 safety risks, and 18 core objects.

By conducting a holistic evaluation of 20 mainstream open-source and advanced proprietary LMMs, we summarize the **key findings** as follows: *(a)* As exemplified by Qwen3.5-Plus, open-source LMMs closely track proprietary performance and achieve parity with GPT-5.5 and Claude-Opus-4.7. Nevertheless, LMMs still have a pronounced performance disparity relative to human experts, indicating substantial room for improvement. *(b)* Industrial scenarios pose unprecedented challenges to fine-grained visual perception. LMMs consistently perform below expectations on the fine-grained industrial element perception tasks. Open-source models attain average accuracies of 0.39 on FG-S and 0.36 on FG-C. *(c)* Tasks involving industrial element perception and warehouse knowledge understanding are predominantly driven by visual cues. In contrast, potential risk reasoning requires models proficient in sophisticated thinking to identify risk factors. *(d)* Mainstream LMMs exhibit a pervasive risk-averse bias when responding to risk-prediction queries. Proprietary commercial models tend to adopt conservative responses, whereas open-source models provide more permissive and candid reports. Only 25% of LMMs pass our fairness-criteria audit. Collectively, these findings emphasize the challenges intrinsic to the LogiScope-VQA benchmark and delineate directions for future research and model improvement.

The curated benchmark bridges the last-mile from well-trained, simulation-validated models to practical application, providing actionable guidance for real-world readiness. For LMMs developed for large-scale industrial deployment, achieving strong performance on LogiScope-VQA is necessary to demonstrate holistic security expertise and expert-level perception and reasoning in operational environments. We hope the unique visual perception challenges, risk reasoning characteristics, and response bias tendencies introduced by LogiScope-VQA will provide useful insights for future work.

2 LogiScope-VQA

2.1 Overview of Logistics Benchmark

We introduce the LogiScope-VQA benchmark, a novel benchmark meticulously curated to evaluate the perception and reasoning capabilities of foundation LMMs for logistics safety in real-world industrial scenarios. The visual corpus is derived from the warehouse surveillance of the global intelligent logistics company Cainiao, comprising 3.5 million raw clips that, after rigorous cleaning and filtering, are distilled into 5,394 high-quality visual samples. Logistics questions are automatically generated from human-annotated object attributes and hand-crafted question templates, comprehensively covering single-choice, multiple-choice, and open-ended formats. Corresponding answers are authored, annotated, and cross-validated by recruited logistics experts through a hierarchical workflow, resulting in 10,274 high-quality VQA instances. Collectively, the benchmark's thorough data release and holistic experimental evaluation are underpinned by one year of surveillance across industrial warehouses worldwide, 52 person-days of annotation by logistics experts, and 324 million tokens consumed during LMM API inference.

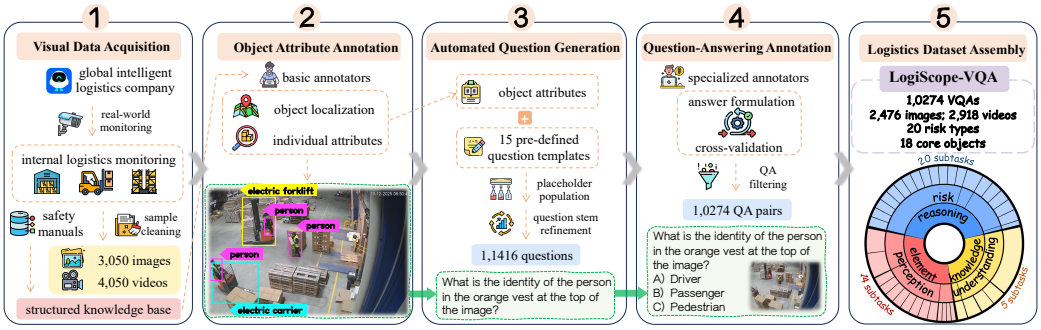


Figure 3: The curation pipeline of LogiScope-VQA, rigorously presenting the operational protocol alongside example outputs.

Grounded in diverse visual content and granular attribute annotations, our benchmark encompasses single-choice, multiple-choice, and open-ended VQA problems. As illustrated in Fig. 2, we achieve comprehensive coverage of key elements in industrial logistics scenarios, spanning 18 core objects and 20 risk types. We aim to directly challenge LMMs to perceive logistics objects and integrate vertical knowledge to uncover latent industrial hazards. Eventually, LogiScope-VQA is designed to rigorously evaluate the essential skills of LMMs, spanning industrial element perception, warehouse knowledge understanding, and potential risk reasoning.

Our benchmark proposes four key challenges for multimodal foundation models. Across diverse tasks, it requires LMMs to (i) perceive dense object distributions in industrial environments, (ii) acquire and operationalize logistics-specific domain knowledge, (iii) integrate multimodal information to perform reasoning, and (iv) anticipate and identify potential safety risks via sophisticated thinking. We aim for the curated benchmark to serve as a comprehensive LMM testbed for identifying logistics safety risks in industrial scenarios, thereby precisely evaluating the real-world deployment potential of mainstream foundation models.

2.2 Distinctive Positioning of LogiScope-VQA

Distinct from existing benchmarks, we require LMMs to perform safety risk–oriented question-answering grounded in industrial contexts, enabling an intuitive assessment of practical operational value. The most closely related prior efforts are IndustryEQA (Li et al., 2026) and iSafetyBench (Abdullah et al., 2025). However, IndustryEQA derives its industrial data from simulated virtual environments, while iSafetyBench focuses solely on hazardous action recognition, constraining the breadth of visual elements and task coverage. In contrast, our benchmark primarily sources data from real-world logistics warehouses, a nearly tenfold size over the previous dataset, and conducts a comprehensive multi-dimensional evaluation across perception, understanding, and reasoning.

2.3 Dataset Curation Pipeline

The construction of LogiScope-VQA followed a hierarchical curation pipeline that focused on data coverage and annotation quality, as illustrated in Fig. 3. In **step 1**, we aggregated one year of surveillance video data from Cainiao's global warehouse parks, amounting to 3.5 million raw clips. This massive corpus was then cleaned through video segmentation, camera-zone grouping, object detection pre-labeling and deduplication, yielding more than 7K high-quality visual samples as the initial seed data. Targeted editing is performed on approximately 900 samples with Nano Banana 2 (Gemini, 2023) to rebalance the distribution of rare risk types. (The detailed image editing prompt is presented in Appendix B.4.) These over 7K seed samples were subsequently filtered in step 4 based on VQA pair quality, retaining the final 5,394 visual samples. Concurrently, we identified core objects of interest within logistics safety and transformed their associated unstructured safety manuals into a structured schema. In **step 2**, basic annotators spent 12 person-days annotating 20 core objects and their basic attributes. Specifically, they first drew bounding boxes to identify designated objects (*e.g.*, persons, electric forklifts) on the keyframes, and then labeled the corresponding attributes (*e.g.*, frame location, gender). To ensure accuracy, a third-party audit was conducted, with

substandard annotations returned for iterative correction. In **step 3**, the labeled object attributes were populated into the placeholders of the pre-defined 15 question templates to batch-generate initial VQA pairs, followed by a refinement operation. Such operation not only resolves referential ambiguity by adding specific modifiers and descriptive constraints to object names, but also improves sentence fluency by ensuring grammatical correctness and natural phrasing. In **step 4**, four specialized annotators leveraged their domain expertise to annotate ground-truth answers and simultaneously filter out low-quality VQA pairs. This three-round cross-validation process, which took 40 person-days in total, comprised: (i) initial answer annotation and filtering of ill-formed pairs (*i.e.*, those with ambiguous referents, unnatural phrasing, or missing options), which retained 5,394 visual samples from the initial 7K seed samples; (ii) independent re-annotation by a senior expert to accept concordant cases and adjudicate discordant cases with reference to the original labels to determine preferred answers; and (iii) a stratified review that prioritized verifying cases with prior inter-round inconsistency. Consequently, approximately 90% of the data was retained as high-quality VQA pairs. In **step 5**, to ensure dataset diversity, we balanced the distribution of core objects, risk types, and hallucination rates by removing over-represented categories. Ultimately, the entire pipeline produced LogiScope-VQA, a comprehensive dataset comprising 10,274 high-quality VQA pairs. (Thorough details of the data processing pipeline are provided in Appendix B.)

3 EXPERIMENT

3.1 EVALUATION SETUPS

LMM Baselines We evaluate the difficulty of LogiScope-VQA using diverse state-of-the-art LMMs as baselines and establish robust reference points for future research. For proprietary LMMs, six leading commercial models are selected: Claude-Sonnet-4.6 (Team., 2026b), Claude-Opus-4.7 (Team., 2026a), Gemini-3.1-Pro (Gemini, 2023), GPT-5.4, GPT-5.5 (OpenAI, 2025), and Qwen3.7-Plus QwenTeam (2026b). These models represent the current upper bound of closed-source multimodal capabilities and serve as strong performance ceilings for LogiScope-VQA. For open-source LMMs, models spanning multiple architectural families and parameter scales are included: MiMo-VL-7B-RL (Team, 2025), GLM-5.2 (Z.ai, 2026), GLM-4.7 (Z.ai, 2025), Kimi-K2-Thinking (AI, 2025), Kimi-K2.6 (AI, 2026), InternVL-3.5 (InternVL, 2025), LLaVA-v1.6-7B (LLaVA, 2024), Qwen3.5-Plus (QwenTeam, 2026a), and multiple Qwen3-VL variants (8B, 32B, 235B-A22B, Plus) (Bai et al., 2025). Moreover, we incorporate the Random Choice and Frequent Guess baselines to provide chance-level and prior-biased reference points. While computed in the standard manner for multiple-choice questions, the Random Choice accuracy is treated as zero for open-ended questions, given the extremely low probability that a solution randomly sampled from the vast candidate space would be consistent with the ground-truth answer. Meanwhile, the Frequent Guess baseline uses the most frequent canonical answer within each task category as the prediction for all open-ended questions.

Human Performance To assess holistic human performance, we recruited ten college students with no expertise in industrial logistics to represent the novice level, and three senior logistics specialist with over three years of on-site warehouse safety experience to represent the expert level. Human participants completed all questions independently based solely on professional knowledge, without external references or model outputs. This setup ensures that the reported performance reflects genuine domain expertise rather than test-taking strategies or external aids, serving as a reliable upper bound for model evaluation. (The detailed employment terms for human evaluators are provided in the Appendix C.3.)

Decoding Settings For all problem instances, we use a unified prompt template that explicitly instructs the models to produce answers in a standardized format. For single-choice and multiple-choice questions, correctness is determined via direct keyword matching. For open-ended questions, we develop an LLM-as-a-Judge workflow to extract the key conclusion from each response for answer matching. Eventually, we adopt micro-averaged accuracy as the evaluation metric. All evaluated models are accessed via their official APIs or publicly available checkpoints, with thinking modes enabled throughout inference. For each task, we repeat each test three times and report the average performance to ensure reliability and reproducibility. The sampling temperature for each

model is set to its recommended or default value. (Detailed prompt templates and decoding configurations are in Appendix C.1. Thorough video evaluation protocols are described in Appendix C.2.)

Table 1: Main results on LogiScope-VQA across 10 evaluation tasks. *Abbr.*, FP-S: Fine-grained Perception of Single-instance, FP-C: Fine-grained Perception of Cross-instance, CP: Coarse-grained Perception; WC: Warehouse Commonsense, SR: Spatial Relation, OR: Operator Role; PAC: Perimeter Access Control, FM: Fire Monitoring, PSD: Personnel Safety Duty, EOC: Equipment Operation Compliance. Optimal proprietary and open-source models are highlighted in purple and yellow.

Model	Overall	Industrial Element Perception			Warehouse Knowledge Understanding			Potential Risk Reasoning			
		FP-S	FP-C	CP	WC	SR	OR	PAC	FM	PSD	EOC
Heuristics Baselines											
Random Choice	0.33	0.24	0.22	0.32	0.31	0.15	0.07	0.37	0.46	0.46	0.47
Frequent Guess	0.47	0.32	0.27	0.63	0.39	0.24	0.40	0.74	0.49	0.64	0.80
Proprietary Large Multimodal Models											
Claude-Sonnet-4.6	0.64	0.63	0.64	0.86	0.90	0.58	0.20	0.49	0.76	0.58	0.62
Claude-Opus-4.7	0.70	0.64	0.61	0.88	0.89	0.60	0.30	0.74	0.82	0.65	0.79
Gemini-3.1-Pro	0.68	0.72	0.70	0.91	0.92	0.66	0.30	0.48	0.82	0.58	0.61
GPT-5.4	0.69	0.59	0.58	0.90	0.93	0.59	0.25	0.58	0.84	0.69	0.78
GPT-5.5	0.71	0.62	0.68	0.78	0.95	0.59	0.20	0.79	0.82	0.65	0.81
Qwen3.7-Plus	0.66	0.67	0.64	0.89	0.95	0.56	0.40	0.48	0.83	0.53	0.69
Open-source Large Multimodal Models											
Kimi-K2-Thinking	0.49	0.45	0.54	0.72	0.96	0.18	0.05	0.15	0.30	0.29	0.14
Kimi-K2.6	0.62	0.61	0.60	0.89	0.93	0.45	0.20	0.57	0.85	0.59	0.66
Qwen3-VL-8B	0.50	0.48	0.44	0.89	0.87	0.09	0.05	0.42	0.73	0.57	0.65
Qwen3-VL-32B	0.55	0.55	0.54	0.83	0.92	0.27	0.15	0.46	0.74	0.60	0.59
Qwen3-VL-235B	0.64	0.64	0.59	0.89	0.91	0.55	0.25	0.43	0.73	0.60	0.67
Qwen3-VL-Plus	0.34	0.53	0.45	0.76	0.91	0.15	0.05	0.20	0.57	0.30	0.13
Qwen3.5-Plus	0.70	0.66	0.65	0.90	0.93	0.61	0.45	0.48	0.84	0.59	0.77
InternVL3.5	0.33	0.48	0.39	0.88	0.69	0.04	0.05	0.44	0.83	0.42	0.17
LLaVA-v1.6-7B	0.50	0.34	0.30	0.48	0.60	0.21	0.05	0.85	0.62	0.69	0.82
MiMo-VL-7B-RL	0.44	0.46	0.39	0.65	0.32	0.05	0.10	0.44	0.74	0.70	0.60
GLM-4.7	0.48	0.40	0.41	0.70	0.95	0.20	0.15	0.06	0.40	0.29	0.16
GLM-5.2	0.53	0.44	0.45	0.72	0.95	0.20	0.15	0.41	0.45	0.37	0.38
MiniMax-M2.1	0.43	0.47	0.49	0.54	0.93	0.19	0.15	0.50	0.43	0.40	0.60
MiniMax-M2.5	0.46	0.50	0.54	0.61	0.93	0.18	0.15	0.42	0.41	0.43	0.59
Human Performance											
Novice Student	0.73	0.80	0.81	0.85	0.76	0.58	0.75	0.75	0.80	0.70	0.67
Logistics Expert	0.95	0.97	0.97	0.96	1.00	0.99	0.95	0.93	0.95	0.95	0.91

3.2 Overall Results

3.2.1 Main Performance Analysis

Performance Across Task Dimensions As reported in Tab. 1, proprietary models maintain an advantage, yet the gap with open-source models has essentially closed at the top, where the best open-source model now ties the strongest proprietary systems. These leading models have edged past the untrained human (novice student) while remaining far below the logistics expert, confirming that our benchmark measures learnable professional expertise rather than common intuition. The key finding, however, is that fine-grained perception in logistics safety scenarios constitutes a shared, industry-wide bottleneck. All top models cluster within a narrow, low band on FG-S and FG-C, falling short of even the novice humans and trailing the experts by a wide margin. This is partly because surveillance images in these settings are typically wide-angle, low-resolution, and densely cluttered with targets and backgrounds, posing a far greater challenge than the clean, high-quality images common in standard benchmarks. In stark contrast, on knowledge-intensive tasks such as AR and EOC, models already outperform untrained humans. This contrast reveals an interesting structural complementarity between human and machine capabilities. Humans naturally excel at

Table 2: Cross-task performance comparison with and without thinking mode. Colored cells indicate significant changes.

Model	Industrial Element Perception		Warehouse Knowledge Understanding		Potential Risk Reasoning	
	non-think	think	non-think	think	non-think	think
LLaVA-v1.6-7B	0.28	0.30 ↑7.1%	0.27	0.31 ↑14.8%	0.65	0.71 ↑9.2%
Kimi-K2.6	0.66	0.66	0.71	0.57 ↓19.7%	0.68	0.74 ↑8.8%
MiniMax-M2.1	0.14	0.10 ↓28.6%	0.38	0.38	0.62	0.63
GLM-5.2	0.15	0.12 ↓20.0%	0.39	0.39	0.50	0.54 ↑8.0%
Qwen3-VL-235B	0.63	0.64	0.66	0.69	0.70	0.71
Qwen3.5-Plus	0.69	0.68 ↓1.5%	0.72	0.74	0.72	0.80 ↑11.1%
Qwen3.7-Plus	0.69	0.69	0.71	0.68 ↓4.2%	0.69	0.77 ↑11.6%

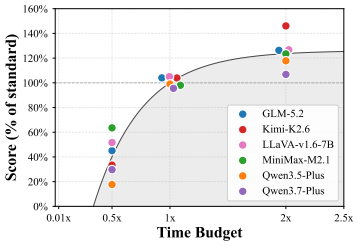


Figure 4: Impact of thinking time budget on reasoning-driven tasks.

spotting fine-grained visual details in logistics safety scenarios even without training, whereas models, despite their vast knowledge, often behave like a well-read scholar with poor eyesight.

Statistical Error Analysis We manually examined 150 randomly sampled error cases and categorized their root causes. Perceptual errors constitute the largest proportion of 48% errors, including detail misidentification (33%) and spatial misperception (15%). Reasoning errors (30%) are the next most frequent, followed by lack of knowledge (18%), and other errors (4%). This distribution corroborates that fine-grained visual perception, rather than high-level reasoning or knowledge coverage, remains the primary bottleneck for current LMMs in logistics scenarios. (Detailed error distribution is in Appendix C.5.)

Performance Evaluation of Real and Synthetic Data To increase the challenge of multimodal hazard identification, synthetic data is incorporated to systematically diversify logistics risk factors. Herein, we compare model performance across synthetic and real data. As detailed in Appendix C.4, the marginal performance disparity between these data sources indicates that the synthetic distributions exhibit high fidelity to realistic scenarios, thereby further validating the reliability of synthetic data for robust model evaluation.

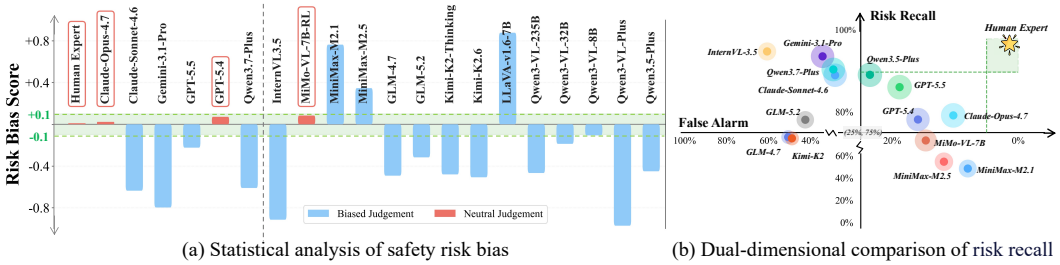
3.2.2 In-depth Thinking Analysis

We conduct an analysis of how thinking mode and time budgets affect different task dimensions. As shown in Tab. 2, red, orange and blue cells present notable increases, moderate increases and noticeable drops, respectively, indicating that enabling thinking mode yields a mixed effect rather than a uniform improvement. Across models, the overall ranking remains largely stable, but the performance shift is highly task-dependent, suggesting that thinking mode interacts differently with perception-heavy and reasoning-heavy tasks. *Industrial Element Perception* is the least responsive dimension, while *Warehouse Knowledge Understanding* shows modest gains, which may be constrained by the perceptual layer itself. However, *Potential Risk Reasoning* benefits the most, especially for stronger models such as Qwen3.5-Plus. This is because the process of observing the scene, matching safety rules and drawing inferences is reasoning-dominant, making it particularly well aligned with explicit chain-of-thought enabled by thinking mode.

Building on this observation, we further probe whether the reasoning advantage persists as the thinking time budget increases for the same *Potential Risk Reasoning* VQA samples. Figure 4 indicates a clear saturation pattern across all models. At low budgets, performance degrades severely due to output truncation, while further gains become marginal beyond the standard budget. Therefore, to balance performance and token consumption, thinking mode should be selectively enabled depending on the task's primary reliance on visual perception or reasoning. And the time budget should be set near the saturation point.

3.2.3 Safety Risk Bias Analysis

We investigate whether models' safety risk judgments exhibit systematic bias or random noise with two experiments, as shown in Fig. 5. In panel (a), we quantify global directional skew via the indicator RBS, defined as $\text{Risk Bias Score} = \tanh\left(\frac{\log((\text{FNR}+\epsilon)/(\text{FPR}+\epsilon))}{c}\right)$, where FNR and FPR denote the false negative rate and false positive rate, and $\epsilon = 10^{-8}$ and $c = 2$. Positive, negative, and



(a) Statistical analysis of safety risk bias (b) Dual-dimensional comparison of risk recall

Figure 5: Empirical analysis of safety risk biases in mainstream LMMs. Panel (a) depicts the holistic characterization of deviations in risk bias scores. Panel (b) provides the detailed comparison across the risk-recall and false-alarm dimensions.

near-zero RBS indicate optimistic, conservative, and neutral judgments, respectively. Our evaluation reveals two critical findings. **First, conservative bias dominates**, with most models exhibiting a negative RBS. This systematic over-reporting tendency likely reflects a safety-first alignment strategy that prioritizes high recall, penalizing missed detections more heavily than false alarms. **Second, a comprehension gap separates proprietary and open-source models**. While top proprietary models like GPT cluster near the unbiased neutral zone (RBS ≈ 0), open-source counterparts scatter across the entire score range. This divergence implies a cognitive divide between genuine risk comprehension and superficial pattern matching. Elite proprietary models may leverage event-level risk reasoning to accurately distinguish benign events from hazards, whereas less capable models lack such depth. Consequently, the latter likely rely on defensive over-reporting to mitigate uncertainty, manifesting as systemic conservative bias.

Building on this global picture, we further conduct a paired-image probe to examine how such bias behaves under explicit risk removal. Specifically, we evaluate models on 100 surveillance samples containing genuine hazards and a corresponding set of risk-free samples, captured by the same cameras at a different time period, to measure recall and false alarm rates, respectively. As depicted in panel (b), most models maintain a risk recall above 60% but their false alarm rates vary substantially, confirming the prevalence of conservative bias. Moreover, The models are distributed across three quadrants, with those in quadrant 2 exhibiting excessive sensitivity and high false alarm rates, whereas those in quadrant 4 underperform by missing genuine threats. In contrast, models in quadrant 1 achieve the optimal trade-off between recall and false alarm rates, yet they still remain far from human expert performance. Collectively, our experiments underscore that safety risk bias is a pervasive systemic issue, which may be attributed to an interplay between internal cognitive limitations and external factors such as training data and alignment protocols.

3.2.4 Reliability of LLM-as-a-Judge

To precisely evaluate the open-ended questions in LogiScope-VQA, we adopt Claude-Opus-4.7 (temperature = 0) as the LLM-as-a-Judge evaluator. For reliability validation, we randomly sample 100 questions from the open-ended VQA set. Tab. 3 reports evaluation accuracy, F1, and Cohen's κ between each evaluator setup and the human reference, alongside the inter-annotator agreement between the two human experts as an empirical upper bound. Notably, Claude-Opus-4.7 achieves the highest consistency with human experts, closely approaching the human-human ceiling and outperforming open-source models. These results demonstrate the reliability of the LLM-as-a-Judge evaluation for our open-ended VQA questions.

Judge Setup	Acc.	F1	Cohen's κ
Phi-4-Mini	74.4	77.2	0.81
DeepSeek-V4-Pro	81.4	82.3	0.83
Claude-Opus-4.7	**87.2**	**89.6**	**0.89**
Human-Human	**92.5**	**93.3**	**0.92**

Table 3: Consistency comparison on 100 sampled open-ended VQA pairs.

4 Open Challenges and Future Directions

Evaluations on LogiScope-VQA expose two key limitations of current models. Firstly, fine-grained perception in logistics safety scenarios remains a major bottleneck, where models particularly struggle with low-resolution, wide-angle, heavily occluded and densely cluttered surveillance footage. To

address this, future efforts should aim to boost models' intrinsic visual capacity, achievable through strategies like optimizing pre-training for dense and low-resolution targets, or exploiting temporal information across video frames. Secondly, safety risk bias emerges as a significant yet overlooked issue, characterized by a prevalent unfairness, with most models heavily inclined to over-report risks. This problem deserves more attention from model developers, as industrial application models require fairness as much as accuracy. For future work, a crucial direction is to prevent such bias at the source, for example through more balanced data curation or bias-aware training objectives.

5 Related Work

5.1 Industrial Assessment of Logistics Safety

Industrial logistics safety assessment provides a compelling testbed for studying the practical applicability of LMMs. Unlike natural-scene evaluation (Liu et al., 2024; Majumdar et al., 2024), industrial environments (Mitash et al., 2023; Abdullah et al., 2025) impose the integration of sophisticated object recognition, precise spatial reasoning, and safety-critical hazard identification. However, this research area remains heavily underexplored, as industrial data source are often constrained by commercial confidentiality. IndustryEQA (Li et al., 2026) is a pioneering study on logistics safety evaluation in warehouse scenarios. However, it contains only approximately 1.3 thousand physics-simulated samples generated in virtual environments. To fully alleviate this challenge, we release LogiScope-VQA, which consists of 10,274 VQA instances annotated by logistics experts. The evaluation data is collected from 3.5 million real surveillance footage clips captured in the world's largest cross-border e-commerce warehouse. We expect the constructed benchmark to provide a direct and practical testbed for assessing the potential of LMMs in industrial applications.

5.2 Practical Application Benchmarks for LMMs

LMMs have emerged as a major frontier in AI research, offering a unified modeling framework that integrates visual perception, language modeling, and sophisticated reasoning (Radford et al., 2021; Alayrac et al., 2022; Liu et al., 2023). Through vision-language pre-training, instruction tuning, and reinforcement learning for reasoning, cutting-edge research has progressively advanced model capabilities across a broad spectrum of vertical domains. To keep pace with these advances, recent benchmarks have introduced targeted evaluations for reasoning-centric mathematical problem solving (Yue et al., 2024; Wang et al., 2026; Lu et al., 2024), knowledge-intensive medical diagnosis (Chen et al., 2024; Zhou et al., 2025; Liu et al., 2025), and visually-grounded remote sensing tasks (Wang et al., 2025b; Danish et al., 2025; Wang et al., 2025a). Logistics safety evaluation focuses on workflows, equipment states, human behaviors, and safety regulations, framing an embodied risk-oriented QA task in real industrial settings with distinctive domain-specific challenges. LogiScope-VQA is the first dedicated evaluation suite for logistics scenarios, filling a critical gap in existing research.

6 Conclusion

The proposed LogiScope-VQA represents a significant advance in assessing the capabilities of LMMs for logistics hazard identification in industrial scenarios. Concretely, we comprehensively evaluate the basic skills of industrial element perception, warehouse knowledge understanding, and potential risk reasoning. Extensive empirical evaluations reveal that LMMs fall short of expectations in industrial visual perception, and effective hazard identification hinges on coherent sophisticated reasoning. Moreover, our investigation uncovers a pervasive issue of security-related bias in current LMMs. These discrepant findings offer valuable insights for foundation model construction and practical deployment. In future work, we will further enrich the visual question-answering corpus and extend the benchmark for agentic industrial operations, continuously contributing robust resources to the multimodal evaluation research community.

AI USE STATEMENT

The conceptual innovation and motivation underlying this study are conceived independently by the human authors, without cognitive input from large language models. The manuscript is originally written by the human authors, and large language models are engaged only at the final stage to assist in polishing key academic terminology. During dataset construction, we primarily relied on proprietary real-world data, while approximately 20.5% of the dataset is generated by the large language model (*i.e.*, Nano Banana 2 (Gemini, 2023)) to systematically enrich the diversity of risk factors. For experimental evaluation, we employ large language models exclusively through the official API endpoints provided by respective vendors. All human–AI interactions are conducted in full compliance with all applicable terms of service and licensing conditions.

ETHICS STATEMENT

We are committed to responsible AI research and strict adherence to data privacy standards. The real-world surveillance footage in LogiScope-VQA was sourced directly from Cainiao's internal warehouse parks. Prior to any academic access or redistribution, the data underwent formal review and approval by our company's internal Data Governance and Legal Compliance Committee. Because the footage originates from standard operational security systems within our own facilities, where obtaining individual employee consent for secondary research use is operationally unfeasible, our ethical compliance relies on this authorized internal corporate governance coupled with strict, irreversible anonymization to protect worker privacy.

To this end, we implemented multi-layered anonymization protocols for all private data in LogiScope-VQA before release. Specifically, all privacy-sensitive content—including clearly visible human faces, employee ID badges, specific warehouse names, and commercial identifiers—has been systematically blurred or masked. This process combines automated detection algorithms with exhaustive frame-by-frame manual verification, ensuring that no individual or proprietary entity can be identified or re-identified. In contrast, since the synthetic data contains no real-world identities and the open-source data is already publicly available, both are released in their original format without additional anonymization. We confirm that the construction and release of LogiScope-VQA strictly comply with relevant ethical guidelines and data protection regulations. For detailed information regarding dataset redistribution and usage restrictions under the CC BY-NC-SA 4.0 license, please refer to Appendix A.

REPRODUCIBILITY STATEMENT

We provide evaluation code through the anonymous github repository and include a representative subset of the dataset in the supplementary material. Notably, we formally commit to publicly releasing the entire evaluation project, including complete collected and annotated dataset, comprehensive evaluation code, and a continuously updated leaderboard. We hope the open-source contributions will help break down existing data barriers in industrial safety and foster the advancement of the research community dedicated to multimodal large language model evaluation in industrial scenarios.

REFERENCES

Raiyaan Abdullah, Yogesh Singh Rawat, and Shruti Vyas. iSafetyBench: A video-language benchmark for safety in industrial environment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1433–1442, 2025.

Kimi AI. Introducing Kimi K2 thinking. *https://www.kimi.com/blog/kimi-k2-thinking*, 2025.

Kimi AI. Kimi K2.6 tech blog: Advancing open-source coding. *https://www.kimi.com/blog/kimi-k2-6*, 2026.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631, 2025*, 2025.

Pengcheng Chen, Jin Ye, Guoan Wang, Yanjun Li, Zhongying Deng, Wei Li, Tianbin Li, Haodong Duan, Ziyan Huang, Yanzhou Su, et al. GMAI-MMBench: A comprehensive multimodal evaluation benchmark towards general medical ai. *Advances in Neural Information Processing Systems*, 37:94327–94427, 2024.

Muhammad Danish et al. GEOBench-VLM: Benchmarking vision-language models for geospatial tasks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025.

Team Gemini. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805, 2023*, 2023.

Team InternVL. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Yifan Li, Yuhang Chen, Anh Dao, Lichi Li, Zhongyi Cai, Zhen Tan, Tianlong Chen, and Yu Kong. IndustryEQA: Pushing the frontiers of embodied question answering in industrial scenarios. In *Advances in Neural Information Processing Systems*, volume 38, 2026.

Bo Liu, Ke Zou, Li-Ming Zhan, Zexin Lu, Xiaoyu Dong, Yidi Chen, Chengqiang Xie, Jiannong Cao, Xiao-Ming Wu, and Huazhu Fu. GEMeX: A large-scale, groundable, and explainable medical vqa benchmark for chest x-ray diagnosis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21310–21320, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36:34892–34916, 2023.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. MMBench: Is your multi-modal model an all-around player? In *Proceedings of the European Conference on Computer Vision*, pp. 216–233. Springer, 2024.

Team LLaVA. LLaVA: v1.6. *https://ollama.com/library/llava:v1.6*, 2024.

Pan Lu, Liang Xu, Swaroop Mishra, Lichen Xia, Yicheng Nie, Chen Zhu, Minjie Zhang, Tom Goldstein, William Yang Wang, and Hannaneh Hajishirzi. MathVista: Evaluating mathematical reasoning of foundation models in visual contexts. In *The Twelfth International Conference on Learning Representations*, 2024.

Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, Karmesh Yadav, Qiyang Li, Ben Newman, Mohit Sharma, Vincent Berges, Shiqi Zhang, Pulkit Agrawal, Yonatan Bisk, Dhruv Batra, Mrinal Kalakrishnan, Franziska Meier, Chris Paxton, Sasha Sax, and Aravind Rajeswaran. OpenEQA: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Chaitanya Mitash, Fan Wang, Shiyang Lu, Vikedo Terhuja, Tyler Garaas, Felipe Polido, and Manikantan Nambi. ARMBench: An object-centric benchmark dataset for robotic manipulation. *arXiv preprint arXiv:2303.16382*, 2023.

OpenAI. OpenAI GPT-5 system sard. *arXiv preprint arXiv:2601.03267*, 2025.

QwenTeam. Qwen3.5: Towards native multimodal agents. *https://qwen.ai/blog?id=qwen3.5*, 2026a.

QwenTeam. Qwen3.7: Multimodal agent intelligence. *https://qwen.ai/blog?id=qwen3.7-plus*, 2026b.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Anthropic. Claude Team. Introducing claude opus 4.7. *https://www.anthropic.com/news/claude-opus-4-7*, 2026a.

Anthropic. Claude Team. Introducing claude sonnet 4.6. *https://www.anthropic.com/news/claude-sonnet-4-6*, 2026b.

Xiaomi LLM-Core Team. MiMo-VL technical report. *arXiv preprint arXiv:2506.03569*, 2025.

Fengxiang Wang, Mingshuo Chen, Xuming He, Yi-Fan Zhang, Yueying Li, Feng Liu, Zijie Guo, Zhenghao Hu, Jiong Wang, Jingyi Xu, et al. OmniEarth-Bench: Towards holistic evaluation of earth's six spheres and cross-spheres interactions with multimodal observational earth data. *arXiv preprint arXiv:2505.23522*, 2025a.

Fengxiang Wang et al. XLRS-Bench: Could your multimodal llms understand extremely large ultra-high-resolution remote sensing imagery? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025b.

Xiaohan Wang, Mingze Yin, Yilin Zhao, Gang Liu, and Dian Li. LiveK12Bench: Have large multimodal models truly conquered high school-level examinations? *arXiv preprint arXiv:2605.26781*, 2026.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 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*, pp. 9556–9567, 2024.

Z.ai. GLM-4.7: Advancing the coding capability. *https://z.ai/blog/glm-4.7*, 2025.

Z.ai. GLM-5.2: Built for long-horizon tasks. *https://z.ai/blog/glm-5.2*, 2026.

Hanjing Zhou, Mingze Yin, Wei Wu, Mingyang Li, Kun Fu, Jintai Chen, Jian Wu, and Zheng Wang. ProtCLIP: Function-informed protein multi-modal learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 22937–22945, 2025.

A Data Usage and Licensing

LogiScope-VQA is released under the **Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0)**. It is strictly intended for academic and non-commercial research purposes only. Users are prohibited from using this data for any commercial activities, including but not limited to training commercial models, selling derived products, or integrating it into profit-generating services. Furthermore, any derivative works based on this dataset must be distributed under the same license. By accessing this dataset, users agree to adhere to these terms.

B Curation Details

B.1 Hierarchical Curation Pipeline

Visual Data Acquisition The raw surveillance corpus comprises 3.5 million video clips collected from Cainiao's global warehouse parks between April 15, 2025 and April 10, 2026. The data spans diverse operational scenarios, including high rack areas, loading docks, storage zones, light-duty picking shelves, office areas, etc. Notably, we declare that **no surveillance data from the United States was included in this collection**. To transform this massive raw corpus into a usable foundation, we implemented a rigorous cleaning process. Continuous videos were first segmented into clips or keyframes via fixed-interval sampling or risk-triggered alarm retrieval. These segments were then grouped by camera zones to align with specific operational areas. Subsequently, a pre-trained object detection model identified core objects, allowing us to filter out irrelevant background frames and balance object category distributions. The process concluded with deduplication, yielding over 7,000 high-quality, unique visual samples that serve as the foundational seed data for benchmark construction. With this visual foundation in place, a structured knowledge base from unstructured

safety manuals became essential. Specifically, logistics specifications are parsed into hierarchical JSON schemas based on semantic tuples (**`Object`**,**`Attribute`**,**`Value`**). In this formulation, **`Object`** denotes one of the 18 core entities critical to warehouse safety monitoring, such as person, electric carrier and security scanner. **`Attribute`** captures either the intrinsic state of an object (*e.g.*, posture, PPE compliance) or its interaction relationship with other entities (*e.g.*, human-truck proximity). **`Value`** consists of predefined enumerations derived from domain regulations, strictly constraining the attribute space. For instance, a valid tuple might be `(person, attrs_hoi_rack, climbing)`, which explicitly encodes the specific interaction of a person climbing a rack. Some examples of the JSON schema are shown in Fig. 6.

Object Attribute Annotation In Object Attribute Annotation, visual grounding was established via lightweight, cost-effective manual labeling. Specifically, annotators followed a "box-then-attribute" protocol: they first drew bounding boxes to localize core objects, and then marked basic attributes (e.g., frame location, occlusion status). For video samples, this process was applied exclusively to the single keyframe exhibiting the most salient risk characteristics, thereby maximizing annotation efficiency. Furthermore, to enrich semantic details, the "person" object received extended tags such as gender and upper-garment color. Crucially, all labels were constrained to predefined values aligned with the logistics knowledge base. To ensure final precision, a third-party quality audit was conducted, involving iterative sampling and correction cycles to rectify any remaining inaccuracies.

Automated Question Generation This step transformed object attribute annotations into natural language queries through two sequential operations. First, we leveraged a library of 15 predefined question templates (Fig. 7) to generate raw question stems. Specifically, twelve multiple-choice templates were designed to assess cognitive understanding across dimensions such as holistic scene interpretation, single-object recognition, and multi-object interaction. The remaining three open-ended templates evaluated capabilities in safety risk perception, consequence prediction, hazard rectification, and emergency response planning. Based on these templates, we performed deterministic population by mapping extracted core objects and attribute values into the corresponding slots. Second, we applied rule-driven linguistic refinement to enhance naturalness and precision. This refinement operation served two purposes. It synthesized multi-attribute constraints into natural pre-nominal modifiers to disambiguate and precisely localize target objects (*e.g.*, refining "person" with attributes to "the woman in a blue upper garment in frame 3"). Meanwhile, it made minimal grammatical adjustments (*e.g.*, quantifiers, prepositions or verbs) only when necessary, thereby preserving the structural integrity of the majority of queries.

Question-Answering Annotation We performed Question-Answering Annotation to assign ground-truth answers and simultaneously filter out low-quality QA pairs. This process involved four specialized annotators and totaled 40 person-days of effort at a cost of $1,000. In the first round, two specialists divided the dataset (containing over 7,000 visual samples) equally and independently annotated their respective halves, producing initial labels for each VQA pair. During this round, they explicitly discarded VQA pairs with ambiguous referents, unnatural phrasing, or missing correct options, retaining only well-formed and verifiable pairs (corresponding to 5,394 final visual samples) for subsequent verification. In the second round, one senior expert independently re-annotated every VQA pair and compared the result with the first-round label. If the labels matched, the pair was tentatively accepted; if they disagreed, the expert re-evaluated the pair to adjudicate the discrepancy, either upholding the initial label or correcting it. In the third round, VQA pairs with consistent labels from the previous two rounds underwent a rapid check, while significant attention was devoted to verifying those pairs with prior disagreements. This intensive cross-validation ensured the precision of the final ground-truth answers.

Dataset Assembly As is common in real-world scenarios, logistics safety surveillance data also exhibited a long-tail distribution, where the occurrences of core objects and safety hazards were highly uneven. To achieve a trade-off between diversity and balance, we rebalanced the VQA pairs across multiple dimensions, such as hallucination proportions, core object types and hazard categories, ultimately yielding LogiScope-VQA.



Figure 6: JSON schema examples for *person*, *truck* and *electric forklift*.

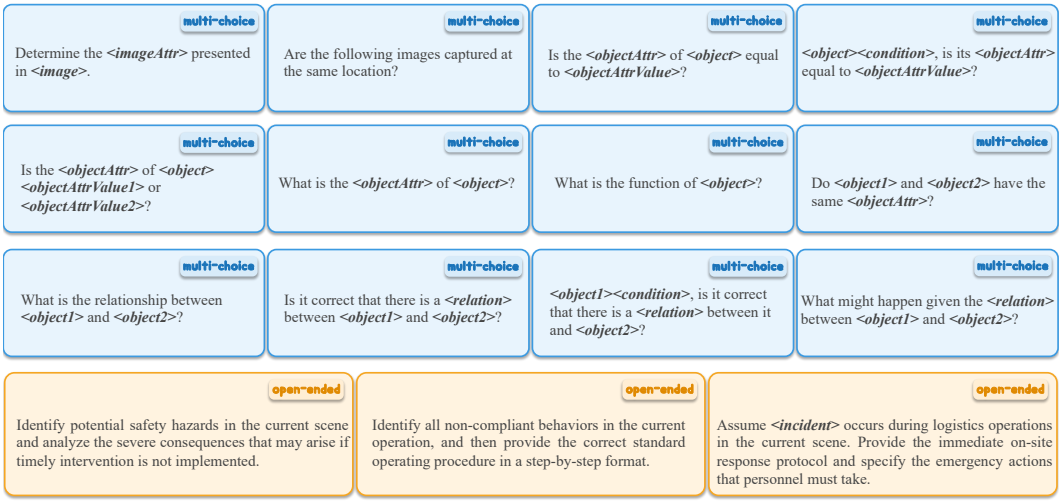


Figure 7: Question templates for generating question stems.

Risk Name	Definition
Warehouse Access Control Presence After Night Lockdown	*In accordance with the warehouse safety regulation "Warehouse Access Control":* During the daily power-off and lockdown period, all personnel must finish their work in advance and evacuate the warehouse area in an orderly manner; no one may remain behind or continue warehousing, equipment maintenance, inspection, or debugging activities for any reason.
Key Post Attendance Unattended Security Scanner Post	*In accordance with the warehouse safety regulation "Key Post Attendance":* While the security scanner is operating, a dedicated staff member must remain on duty at all times to monitor the scanned goods and equipment status in real time; the attendant must not leave the post without authorization, swap posts, or engage in any activity unrelated to monitoring.
Personal Protective Equipment No Helmet in High-Rack Area	*In accordance with the warehouse safety regulation "Personal Protective Equipment":* All personnel entering the high-rack storage area—whether for operations, patrols, or temporary passage—must properly wear a standards-compliant safety helmet with the chin strap fastened; this area poses risks of falling goods and beam impacts, and anyone without protective equipment is strictly forbidden to enter.
Elevated Work #1 Non-Designated Climbing Tool	*In accordance with the warehouse safety regulation "Elevated Work":* All elevated work must use designated, inspection-approved tools such as aerial work platforms, inventory cages, order pickers, A-frame ladders, or mobile ladders; stepping on racks, cargo boxes, forklift forks, or other non-purpose-built structures for work at height is strictly prohibited.
Elevated Work #2 Multiple Riders on Climbing Device	*In accordance with the warehouse safety regulation "Elevated Work":* Dedicated elevating devices such as lifting cages and order pickers are limited to a single occupant; carrying two or more persons simultaneously is strictly prohibited, as overloading significantly reduces equipment stability and increases the risk of tipping and falls.
Elevated Work #3 Elevated Work Without Protection	*In accordance with the warehouse safety regulation "Elevated Work":* When working at height on an A-frame ladder or mobile ladder, a safety helmet must be worn properly; when using mobile platforms such as aerial work platforms, inventory cages, or order pickers, a protective helmet is required and a safety harness must be secured to a reliable anchor point at all times.
Fire Monitoring Open Flame or Smoke	*In accordance with the warehouse safety regulation "Fire Monitoring":* In warehouse and operation areas, any open flame, visible firelight, flying sparks, or smoke emission constitutes a serious safety hazard and is treated as a violation upon discovery. Unpermitted hot work—such as welding, cutting, or grinding that produces high-temperature sparks—is strictly prohibited.
Dock Loading & Unloading #1 Personnel in Truck Danger Zone	*In accordance with the warehouse safety regulation "Dock Loading & Unloading":* While a truck is starting, reversing, or moving, no person may stay in or cross the danger zone within 3 meters in front of or behind the vehicle or 1 meter on either side. This zone covers the vehicle's blind spots and dynamic impact area and poses severe collision and crushing risks. Before moving, the driver must fully observe the surroundings with mirrors, the reversing camera, and other aids, and proceed slowly only after confirming that no one is in or approaching the blind spots.
Dock Loading & Unloading #2 Personnel in Moving Truck Bed	*In accordance with the warehouse safety regulation "Dock Loading & Unloading":* While a truck is starting, moving, or repositioning, no person may remain inside or on top of the cargo bed; after loading and unloading are completed, all personnel must evacuate to a safe area, and the driver may move the vehicle only after confirming the cargo bed is empty.
Dock Loading & Unloading #3 Dock Climbing or Jumping	*In accordance with the warehouse safety regulation "Dock Loading & Unloading":* Personnel must get on and off the dock only via designated passages such as dedicated stairs, ramps, or dock levelers; climbing dock guardrails, crossing edge structures, or jumping from the dock down to the ground or vehicles is strictly prohibited. Such behavior can easily lead to missteps, slips, and falls, causing serious personal injury.
Forklift Driving Compliance #1 Overheight Load on Forklift	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* When powered forklifts (electric forklifts, electric pallet trucks, semi-electric pallet trucks, pallet stackers, or clamp trucks) travel with a load, the stacked cargo height must not exceed 1.8 meters or the upper limit of the mast, so as to prevent rollover from an excessively high center of gravity or collisions caused by blocked vision; cargo stability must be checked before loading to ensure safe operation.
Forklift Driving Compliance #2 Loaded Forklift Not Reversing	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* When powered forklifts (electric forklifts, electric pallet trucks, semi-electric pallet trucks, pallet stackers, or clamp trucks) travel with a load, the forks must face backward with the cargo held tight against the mast, and the truck must travel in reverse to ensure clear vision and a stable center of gravity; driving forward with a load is strictly prohibited.
Forklift Driving Compliance #3 Insufficient Forklift Safety Distance	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* During powered forklift operations (electric forklifts, electric pallet trucks, semi-electric pallet trucks, pallet stackers, or clamp trucks), all personnel other than the operator must keep a sufficient safety distance: where boundary lights are installed, stepping into the warning light zone is strictly prohibited; where no lights are installed, keep clear of the areas within 2 meters in front of and behind the forklift and 1 meter on its sides.
Forklift Driving Compliance #4 Forklift Pushing Cargo on Ground	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* Using powered forklifts (electric forklifts, electric pallet trucks, semi-electric pallet trucks, pallet stackers, or clamp trucks) to push cargo or obstacles along the ground is strictly prohibited, as such operation can easily cause equipment loss of control, cargo tipping, or personnel being run over; to move heavy objects, use dedicated handling tools or place the cargo securely on the forks before transport.
Forklift Driving Compliance #5 Extra Riders on Forklift	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* Powered forklifts (electric forklifts, electric pallet trucks, semi-electric pallet trucks, pallet stackers, or clamp trucks) may only be operated by a single licensed driver, and carrying any other person is strictly prohibited; the forks, fork carriage, or any part of the truck body must never be used as a platform for carrying people.
Forklift Driving Compliance #6 Forklift Forks Not Lowered	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* While powered forklifts (electric forklifts, pallet stackers, or clamp trucks) are traveling with or without a load, the forks must be lowered to a safe height of about 15 cm above the ground; traveling with raised forks is strictly prohibited.
Forklift Driving Compliance #7 Operator Leaning Out of Cabin	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* While a powered forklift (electric forklift, clamp truck, etc.) is in motion, the operator must stay inside the cab at all times and must not extend the head, hands, torso, or any other body part out of the cab. Such behavior can easily cause scraping collisions with racks, door frames, pipes, or other equipment, resulting in serious personal injury.
Forklift Driving Compliance #8 Dismounting a Moving Forklift	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* Before a powered forklift (electric forklift, electric pallet truck, semi-electric pallet truck, pallet stacker, or clamp truck) has come to a complete stop with the brake applied, the operator is strictly prohibited from leaving the driver's or operator's seat; for a temporary absence, the operator must first stop the truck, apply the handbrake, and lower the forks to the ground, confirming the equipment is in a stationary and safe state before leaving the seat.
Forklift Driving Compliance #9 Overloaded Inventory Cage	*In accordance with the warehouse safety regulation "Forklift Driving Compliance":* When an inventory cage is mounted on the forklift forks, the cage is limited to a single occupant; carrying more than one person is strictly prohibited, as overloading significantly reduces equipment stability and increases the risk of tipping and falls.
Conveyor Operation Compliance Standing on Running Conveyor	*In accordance with the warehouse safety regulation "Conveyor Operation Compliance":* While the conveyor is running, standing, sitting, lying, stepping across, or treading on its surface is strictly prohibited to prevent entanglement in mechanical parts and serious injury; for maintenance or cleaning, the conveyor must first be stopped and powered off, lockout-tagout procedures must be performed, and personnel may enter the work area only after confirming the equipment is completely stationary.

Figure 8: Definitions of 20 risk in potential risk reasoning.

B.2 Detailed Risk Definitions

We provide precise definitions for 20 specific risks in Fig. 8, which are involved in the potential risk reasoning tasks.

B.3 Full Task Taxonomy

Fig. 9 presents the full task taxonomy.

B.4 Prompt for Targeted Image Editing

As illustrated in Fig. 10, our prompt template for targeted image editing employs multiple placeholders, accompanied by a concrete example demonstrating its usage. The diversity of resulting prompts stems from the rich set of allowable values for each placeholder, which we comprehensively enumerate in Fig. 11.

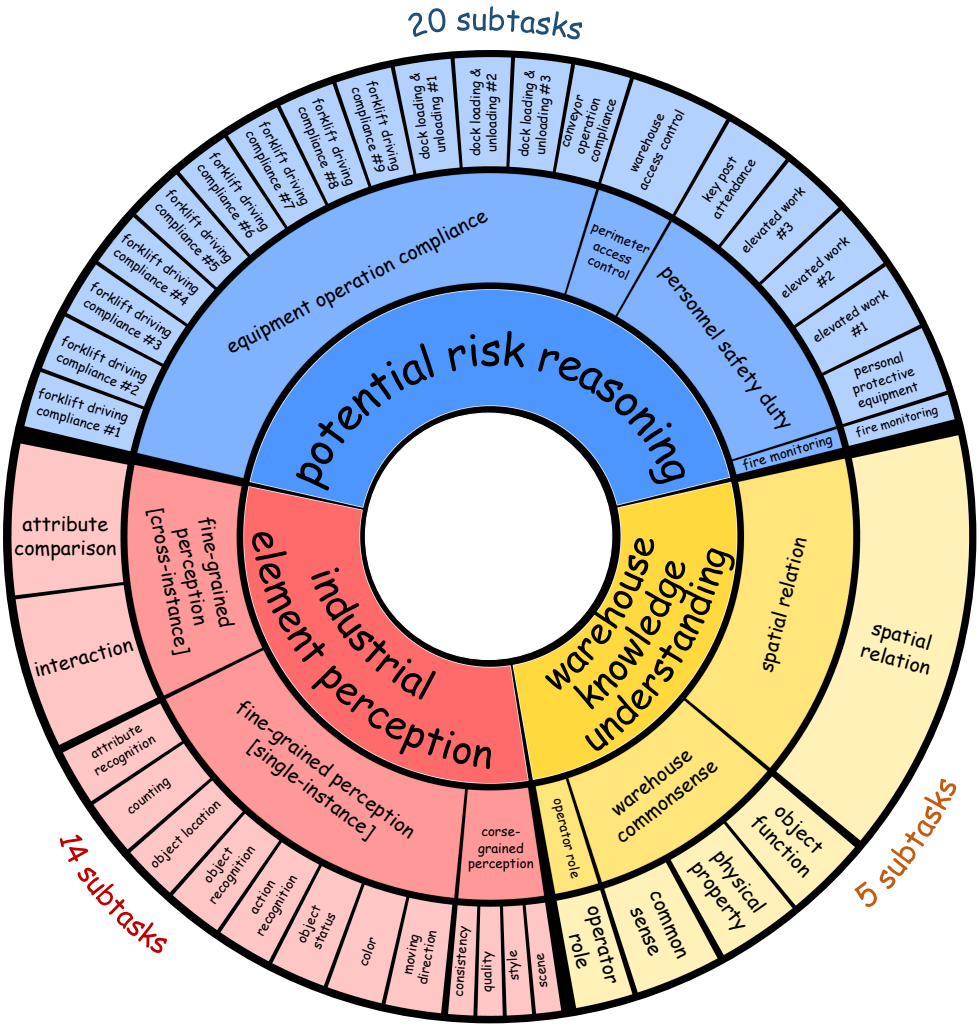


Figure 9: Full task taxonomy of LogiScope-VQA.

System Prompt

[BaseConstraints]

[EditInstruction]

User Prompt

[ImageDescription]

[TargetScene (action)]

[VisualDetails (orientation, clothing, safety gear)]

[ToolDefinition (tool)]

[PositionRequirement (position)]

System Prompt

Role: Expert Logistics Security Image Editor.
Task: Realistic modification for safety AI training.
Strict Constraints:
1. Do NOT change the background warehouse environment.
2. Seamlessly blend the person with lighting and shadows.
3. Ensure physical plausibility (weight, contact points).
4. High resolution, photorealistic style.

User Prompt

Editing Instruction: Detect the person and edit to simulate a high-risk safety violation.
Target Scenario: A worker is climbing barehanded up a fully loaded racking shelf, one foot on a horizontal beam, the other foot stepping upward, both hands gripping the upright column.
Visual Details: The worker is back facing the camera, wearing a blue warehouse uniform, and is wearing a yellow hard hat but NO safety harness.
Positioning: The worker is in the distance, at a high position above the ground. Height > 2 meters from the ground.

Figure 10: Prompt template and example for image generation and editing.

C Evaluation Details

C.1 Additional Decoding Settings

Fig. 12 illustrates two evaluation prompt templates. The left one is designed for solving multi-choices and open-ended questions, while the right one is utilized in the LLM-as-a-Judge workflow for a third-party LMM to assess the correctness of the model's responses (outputting "correct" or "incorrect"). In this judging workflow, the ground truth for short-answer questions typically comprises several core key points. A response is judged as "correct" if it contains all these essential elements, regardless of any additional elaboration; conversely, the omission of any single key point results in an "incorrect" classification. This zero-tolerance policy for omissions is designed to prioritize high recall in hazard detection, given the severe consequences of missed risks in industrial environments. Such a rigorous criterion serves as a stringent test, effectively validating a model's true capability and reliability in logistics safety surveillance. For improved transparency and reproducibility, Tab. 4 includes links to official documentation or model checkpoints of the evaluated LMMs, along with detailed inference configurations and the evaluation timeframe.

C.2 Video Processing and Evaluation Protocol

We detail the video processing, annotation and evaluation strategy below. For each video clip ranging from 30 seconds to 3 minutes, we uniformly sample a concise sequence of K frames ordered chronologically, where $K \in [2, 10]$ depending on the clip's specific duration and complexity. During object-attribute annotation, a single representative keyframe from these 2–10 frames is selected based on a conditional rule: if the clip contains a safety risk, the frame exhibiting the most salient risk characteristics is chosen; if no risk is present, one frame is randomly selected. For example, if a video captures the entire process of a worker climbing over a conveyor belt, the exact frame depicting the moment of crossing is selected as the keyframe. Annotators then perform object-attribute annotation exclusively on this single keyframe to identify core objects and basic attributes, bypassing the noise and cost of dense frame-by-frame labeling. Subsequently, during the question-answering annotation phase, annotators construct answers based on the entire chronologically ordered sequence of all 2-10 sampled frames (including the keyframe), rather than relying solely on the static keyframe. This ensures that the final labels accurately capture both spatial states and temporal dynamics.

Regarding how videos are consumed by the evaluated LMMs, these 2-10 sampled frames are directly fed into the models as an ordered list of images at their native resolution. This approach guarantees that every model receives equivalent temporal information spanning the entire event without arbitrary downsampling. Crucially, regardless of whether a specific VQA question can be answered using only the single keyframe (*e.g.*, no helmet in high-rack area) or inherently depends on motion and temporal events across the full sequence (*e.g.*, standing on running Conveyor), the Vision LLM always receives the complete set of 2–10 ordered frames as input. This unified input protocol prevents models from exploiting shortcuts via static image matching alone and enforces genuine temporal reasoning when evaluating dynamic industrial risks.

Element	Available Value
action	① Climbing barehanded up a loaded rack, feet on beams, hands gripping column, with no climbing aids nearby ② Standing on scissor lift guardrail edge, leaning out, one hand holding rail while other reaches for distant goods ③ Standing on electric forklift tines without cage, one hand holding mast, other searching for items on shelf ④ Standing on top A-ladder step above safety line, body overreaching laterally to grab goods from adjacent racking ⑤ Straddling racking beam with legs dangling, one hand gripping column, other retrieving item from adjacent location ⑥ standing on top of the operator cabin of an electric forklift, leaning toward the racking to inspect goods on the shelves ⑦ standing on a handcart taller than a person's height, reaching for goods on the shelf ⑧ standing on a wheeled mobile ladder platform with casters, retrieving goods from the shelf ⑨ standing alone on the platform of a single-occupancy order picker (man-up forklift) ⑩ standing side by another worker, overcrowding the platform of a single-occupancy order picker (man-up forklift) ⑪ standing inside a bright yellow, grid-walled inventory cage mounted directly onto the forks of an electric forklift ⑫ standing on a high-level order picker vehicle, inspecting goods on the racking shelves ⑬ standing side by another worker, overcrowding the cab of the same order picker vehicle ⑭ standing on top of stacked cargo boxes, reaching for goods on the shelf ⑮ using three empty pallets stacked vertically as a makeshift step ladder, standing on the top pallet and tiptoeing to retrieve an item ⑯ standing on the conveyor belt of an operating production line ⑰ climbing over an operating production line conveyor belt ⑱ a truck reversing toward the loading dock from 2-3 meters away, a worker standing directly behind the truck between the dock and the truck, only upper body visible, unaware of the reversing truck, at risk of being crushed ⑲ a truck reversing toward the loading dock from 2-3 meters away, a worker passing through the gap between the dock and the truck, only upper body visible, at risk of being crushed ⑳ a truck moving forward, a worker chasing the truck from behind to close the rear cargo door, at risk of being crushed
orientation	① back facing the camera ② side profile facing left ③ side profile facing right ④ slightly looking up ⑤ facing the camera with head down ⑥ looking down at feet ⑦ side profile facing left, leaning forward slightly ⑧ side profile facing right, leaning backward slightly ⑨ squatting position facing left ⑩ squatting position facing right ⑪ standing with hands on hips ⑫ standing with one hand on waist ⑬ body tilting slightly to the left ⑭ body tilting slightly to the right ⑮ head tilted upward looking up ⑯ head down focused on the ground
clothing	① wearing a blue warehouse uniform ② wearing a yellow reflective vest ③ wearing an orange safety jacket ④ wearing gray workwear ⑤ wearing a black hoodie with a fluorescent yellow vest with reflective strips ⑥ wearing a blue reflective vest
safety gear	① wearing a yellow hard hat but NO safety harness ② wearing a red hard hat with safety harness unhooked (hanging in air) ③ without any hard hat, head fully exposed, and NO safety harness on body ④ wearing a safety harness but NO hard hat, head fully exposed ⑤ wearing both a hard hat and a safety harness
tool	① electric forklift — a common logistics transport vehicle with two forks at the front for moving palletized goods; open-type driver cabin; metal mesh protective roof over the cabin ② handcart — a manually pushed logistics handling tool, usually with two or four wheels, similar to a large supermarket shopping cart ③ scissor lift — a logistics equipment with a lifting platform, using cross-shaped X hydraulic supports to adjust height, with guardrails on all four sides ④ mobile ladder — a wheeled movable ladder, usually straight-ladder structure, non-folding, with guardrail handholds and casters at the bottom ⑤ A-frame ladder — a foldable A-shaped ladder, without wheels ⑥ racking shelf — a steel structure for 3D storage, composed of upright columns, horizontal beams and shelves, for multi-layer stacking of pallets or turnover boxes ⑦ order picker — a dedicated industrial vehicle for order picking, with an operator platform and linked lifting device; the rider is the driver, no dedicated cabin, has a frame for safety harness, lifting column, and large motor, no forks ⑧ cargo — goods, packages or items to be transported, stored or handled in logistics, usually in cardboard or plastic boxes on pallets ⑨ empty pallet — a standard pallet without cargo, usually made of wood, plastic or metal, for subsequent loading and forklift handling ⑩ inventory cage — a cage with four-sided protective mesh mounted on forklift forks, with insertion holes at the bottom for forks, about half-person height to waist level ⑪ conveyor belt — a continuous transport system in logistics for moving goods along a production line ⑫ loading dock — the height-difference platform between the warehouse floor and the truck parking area ⑬ truck — a cargo vehicle parked at the loading dock area for loading and unloading
position	① far end, very high position, height above ground exceeds 4 meters ② near end, very high position, height above ground exceeds 4 meters ③ middle position, very high position, height above ground exceeds 4 meters ④ far end, high position, height above ground exceeds 2 meters ⑤ near end, high position, height above ground exceeds 2 meters ⑥ middle position, high position, height above ground exceeds 2 meters ⑦ far end, low position, height above ground does not exceed 2 meters ⑧ near end, low position, height above ground does not exceed 2 meters ⑨ middle position, low position, height above ground does not exceed 2 meters ⑩ conveyor belt marked area (indicated by red bounding box) ⑪ dock area below (truck parking zone)

Figure 11: Values of elements for image generation and editing.

Table 4: Inference configurations for proprietary (purple) and open-source (green) LMMs.

GPT-5.4 & GPT-5.5

Link: https://openai.com/index/introducing-gpt-5-5/
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 30-July 15, 2026

Claude-Sonnet-4.6 & Opus-4.7

Sonnet Access: https://www.anthropic.com/claude/sonnet
Opus Access: https://www.anthropic.com/claude/opus
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 30-July 15, 2026

Gemini-3.1-Pro

Link: https://deepmind.google/models/model-cards/gemini-3-1-pro/
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 30-July 15, 2026

Qwen3.7-Plus

Link: https://qwen.ai/blog?id=qwen3.7-plus
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 30-July 22, 2026

GLM-5.2 & GLM-4.7

Link: https://huggingface.co/zai-org/GLM-5.2
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** July 5-22, 2026

MiMo-VL-7B-RL

Link: https://huggingface.co/XiaomiMiMo/MiMo-VL-7B-RL
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** July 10-15, 2026

Kimi-K2-Thinking & Kimi-K2.6

Link: https://huggingface.co/moonshotai/Kimi-K2.6
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 25-July 22, 2026

InternVL-3.5

Link: https://huggingface.co/OpenGVLab/InternVL3_5-8B
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** July 10-20, 2026

LLaVA-v1.6-7B

Link: https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** July 10-20, 2026

Qwen Open-Source Suite

Models: *Qwen3.5-Plus, Qwen3.6-Flash, Qwen3-VL (8B, 30B-A3B, 32B, 235B-A22B, Plus)*
Qwen3 Access: https://huggingface.co/collections/Qwen/qwen3
Qwen3.5 Access: https://huggingface.co/collections/Qwen/qwen35
Qwen3.6 Access: https://huggingface.co/collections/Qwen/qwen36
Max len: $32k$ | **Batch size:** 1 | **backend:** Pytorch
Temperature: 0.2 | **Max new tokens (think/non-think):** 4096 / 256 | **Eval date:** June 25-July 22, 2026

C.3 Human Evaluation Protocol

To obtain comprehensive human performance, we recruited two distinct groups of participants with contrasting expertise levels. The first group consisted of 10 undergraduate student volunteers who served as generalist evaluators. While possessing strong general common sense and basic visual recognition capabilities, they lacked specialized knowledge in logistics operations or safety protocols. These students collectively covered the entire benchmark by answering approximately 1,000 questions each over a period of two weeks, receiving a stipend of $15 per person. The second group comprised a single professional domain expert with at least four years of hands-on experience in logistics safety and surveillance. Unlike the generalists, this expert leveraged deep domain knowledge (*e.g.*, familiarity with industrial equipment and safety protocols) to independently answer all questions in the benchmark. This expert evaluation also spanned two weeks and incurred a total cost of $700. The results labeled "Novice Student" and "Logistics Expert" in Tab. 1 correspond to the performance of the student volunteers and logistics expert, respectively.

C.4 Performance Comparison across Data Formats

Tab. 5 presents a complete performance comparison across different *data origins* and *data types*. The minimal performance gap between synthetic (*Synthetic*) and real data (*Private* and *Open-Source*) suggests that synthetic distributions faithfully mirror realistic scenarios and are reliable for model evaluation. However, a consistent accuracy discrepancy persists, with models generally achieving higher accuracy on images than on videos, likely because reasoning involving temporal dynamics and spatial transformations poses greater challenges than static visual analysis.

Table 5: Performance comparison across *data origin* and *data type*.

Model	Data Origin			Data Type	
	Private	Synthetic	Open-Source	Images	Videos
Proprietary Large Multimodal Models					
Claude-Sonnet-4.6	0.64	0.62	0.68	0.70	0.60
Claude-Opus-4.7	0.74	0.67	0.70	0.72	0.68
Gemini-3.1-Pro	0.66	0.71	0.70	0.77	0.63
GPT-5.4	0.71	0.66	0.64	0.71	0.68
GPT-5.5	0.73	0.70	0.69	0.74	0.69
Qwen3.7-Plus	0.67	0.64	0.65	0.68	0.65
Open-source Large Multimodal Models					
Kimi-K2-Thinking	0.49	0.48	0.58	0.51	0.48
Kimi-K2.6	0.64	0.57	0.67	0.65	0.60
Qwen3-VL-8B	0.54	0.44	0.45	0.50	0.50
Qwen3-VL-32B	0.58	0.50	0.55	0.56	0.55
Qwen3-VL-235B	0.64	0.64	0.67	0.68	0.62
Qwen3.5-Plus	0.72	0.67	0.69	0.73	0.69
InternVL3.5	0.33	0.32	0.37	0.37	0.31
LLaVA-v1.6-7B	0.55	0.47	0.41	0.47	0.55
MiMo-VL-7B-RL	0.48	0.34	0.36	0.41	0.46
GLM-4.7	0.48	0.44	0.48	0.49	0.46
GLM-5.2	0.54	0.52	0.51	0.51	0.53
MiniMax-M2.1	0.41	0.48	0.43	0.40	0.46
MiniMax-M2.5	0.43	0.49	0.47	0.41	0.49
Human Performance					
Novice Student	0.71	0.76	0.73	0.74	0.72
Logistics Expert	0.96	0.95	0.93	0.95	0.94

C.5 Detailed Error Distribution

We analyzed the errors of Gemini-3.1-Pro (Gemini, 2023) by manually examining 150 randomly sampled error cases to identify the root causes of mispredictions. The distribution of these errors is shown in Fig. 13.

System Prompt

You are a logistics safety VQA assistant. Answer questions strictly based on observable visual evidence. Logical reasoning is permitted, but do not fabricate details unsupported by the image.

Your response must consist of exactly two parts, in this order:

1. **Reasoning**: A thorough visual analysis and logical deduction process.

2. **Final Answer**: The format of this part depends on the question type:

- If the question provides multiple-choice options (e.g., A, B, C, D), output a JSON-formatted list containing the indices/letters of all correct answers (e.g., ["A"] or ["A", "C"]).

- If the question is open-ended (no options provided), output a concise textual response using precise safety terminology that directly addresses the query.

Do not include any additional commentary, explanations, or formatting outside these two components. Ensure the transition between Reasoning and Final Answer is clear.

User Prompt

[Question]

A) [Option A] B) [Option B] [etc.]

(a) Prompt for solving questions

System Prompt

You are an expert grader for short-answer questions. Evaluate if the student's answer covers all key points in the standard answer.

Grading Rules:

1. **Correct**: The answer includes ALL essential key points. Paraphrasing and synonymous expressions are accepted. Extra correct details are allowed and do not penalize the score.

2. **Incorrect**: The answer misses any critical key point, contains factual errors, or contradicts the standard answer.

Output Format:

- Output ONLY one word: `correct` or `incorrect`.

- No explanations or punctuation.

User Prompt

Below is the content for your evaluation:

Original Question: [question]

Standard Answer: [gold answer]

Student's Answer: [answer]

(b) Prompt for evaluating open-ended questions

Figure 12: Prompts for solving and evaluating questions.

Perceptual Error: Detail Misidentified (33%) These errors stem from the model's failure to correctly perceive fine-grained visual details in the input. For example, the model may misidentify specific object attributes, misread small text or numerical values, or confuse visually similar elements that are critical for the task. This type of error indicates a fundamental limitation in the model's visual perception rather than in its reasoning or knowledge application capabilities.

Perceptual Error: Spatial Misperception (15%) These cases involve incorrect interpretation of spatial relationships or structural layout within the visual input, such as misjudging relative positions, orientations, or spatial configurations. For example, when determining whether a person is standing on a moving conveyor belt, an individual standing in the distance may appear to have their feet positioned directly above the belt due to the viewing angle, creating a misleading impression of contact. This suggests that the model's spatial understanding of visual scenes remains incomplete, particularly in tasks requiring precise geometric reasoning from visual evidence.

Lack of Knowledge (18%) These errors occur when the model lacks the necessary domain knowledge or factual information to arrive at the correct answer. Unlike reasoning or perceptual errors, the model correctly understands the question and the visual content, but its internal knowledge base does not contain the required information. This highlights the limitations of the model's pretrained knowledge coverage in specialized or less common domains.

Reasoning Error (30%) These represent cases where the model perceives the input correctly and possesses sufficient domain knowledge, yet fails to reason through the problem to reach the correct conclusion. The model may select an incorrect solution strategy, fail to chain multiple reasoning steps coherently, or draw incorrect logical inferences from the available information. This type of error reveals gaps in the model's multi-step reasoning and logical deduction capabilities when processing complex logistics VQA tasks.

Other Errors (4%) This category includes refusal to answer (2%), where the model expresses uncertainty and declines to provide a response; answer extraction error (2%), where the model's output format could not be reliably parsed.

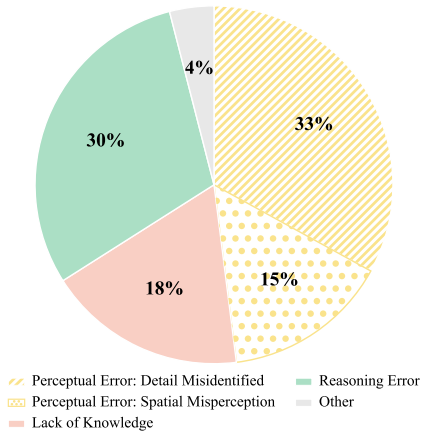


Figure 13: Error distribution of 150 manually examined error cases in Gemini-3.1-Pro.

D VISUALIZATION

To facilitate a more intuitive understanding, we provide additional visualization examples. Fig. 14 displays visual-only examples (raw images).



(a) Real video: dock operations area



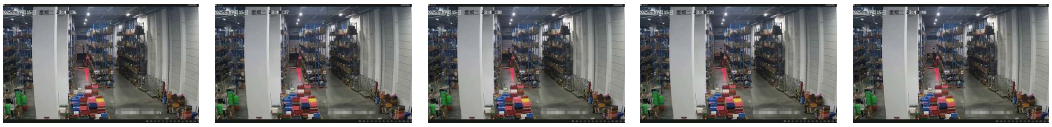
(b) Real video: storage area



(c) Real video: bulk storage area



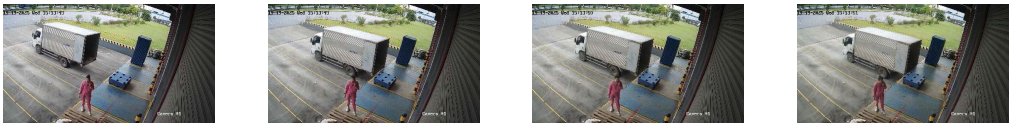
(d) Real video: nighttime scene



(e) Real video: high rack area



(f) Real video: security screening area



(g) Synthetic video: dock operations area



(h) Synthetic video: conveyor area



(i) Real images



(j) Synthetic images

Figure 14: Examples of real and synthetic visual data, covering both daytime and nighttime scenes across various zones within logistics warehouses.