SafeBuild-Bench: A Temporal-Robust Construction Safety Benchmark with Graph-Enhanced Data Mining

Yi Cui*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
ycui785@connect.hkust-gz.edu.cn

Zilin Wang*
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
zwang374@connect.hkust-gz.edu.cn

Yijie Xu, Qianyi Cai, Huizai Yao
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
{yxu409,qcai603,hyao032}@connect.hkust-gz.edu.cn

Shuai Jiang
School of Management
Northwestern Polytechnical University
Xi'an, China
shuaijiangai@gmail.com

Bingzhuo Zhong†
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China
bingzhuoz@hkust-gz.edu.cn

Hui Xiong†
The Hong Kong University of Science and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science and Technology
Hong Kong, China
xionghui@ust.hk

Abstract

Multimodal safety systems for construction sites are often trained and evaluated on data that are redundant, long-tailed, and temporally mismatched to deployment, making training and benchmarking inefficient and sometimes misleading. We introduce SafeBuild-Bench, a temporal-robust benchmark for construction safety designed to evaluate multimodal large language models under realistic time and site shifts. SafeBuild-Bench is mined from 100K+ raw industrial frames collected across unseen sites and different dates, and contains 4K+ expert-verified samples with retained temporal metadata for robustness evaluation. It covers hierarchical tasks ranging from multiple-choice hazard detection to free-form reasoning, with the latter assessed using an LLM-as-a-judge scheme to support scalable evaluation. To build SafeBuild-Bench from massive and redundant streams, we further develop a data selection pipeline, **GEMS** (Graph-Enhanced Multimodal Selection), which combines epistemic uncertainty with dual-graph structure to identify an informational core for curation. We validate this pipeline via a data-quality proxy: on the public LLaVA-mix-665K corpus, fine-tuning LLaVA-v1.5 on only the selected 1% subset matches or surpasses the full-data baseline on robustness-oriented benchmarks, suggesting that **GEMS** can remove redundancy while preserving critical long-tail content. Benchmarking state-of-the-art multimodal LLMs on SafeBuild-Bench shows substantial performance degradation under temporal shift, highlighting a key gap for high-stakes industrial deployment. We release the SafeBuild-Bench dataset, evaluation scripts, and the **GEMS** codebase to support reliable development and comparison of safety-focused multimodal models. Dataset, code, and benchmark results are publicly available at https://github.com/safebuild/gems.

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; • **Information systems** → *Data mining*.

Keywords

Benchmark, Data-Centric AI, Temporal Robustness, Data Efficiency, Construction Safety

ACM Reference Format:
Yi Cui, Zilin Wang, Yijie Xu, Qianyi Cai, Huizai Yao, Shuai Jiang, Bingzhuo Zhong, and Hui Xiong. 2026. SafeBuild-Bench: A Temporal-Robust Construction Safety Benchmark with Graph-Enhanced Data Mining. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '26), August 2026, TBD.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/1122445.1122456

1 Introduction

Multimodal Large Language Models (MLLMs) have improved general-purpose visual understanding and instruction following [5, 8, 26, 32], but their reliability in safety-critical domains is still uncertain. Industrial safety monitoring spans multiple settings, for example, manufacturing lines, logistics, energy facilities, and construction sites, where failures can cause injuries, downtime, and compliance risks. Deploying MLLMs in these workflows requires evidence that models generalize beyond curated test sets and remain stable under real operational changes. In this work, we study construction safety monitoring as a concrete and high-impact domain where such reliability questions directly affect deployment decisions.

*Yi Cui and Zilin Wang contributed equally to this research.
†Bingzhuo Zhong and Hui Xiong are corresponding authors.

KDD '26, TBD

ACM ISBN 978-1-4503-XXXX-X/26/08
https://doi.org/10.1145/1122445.1122456

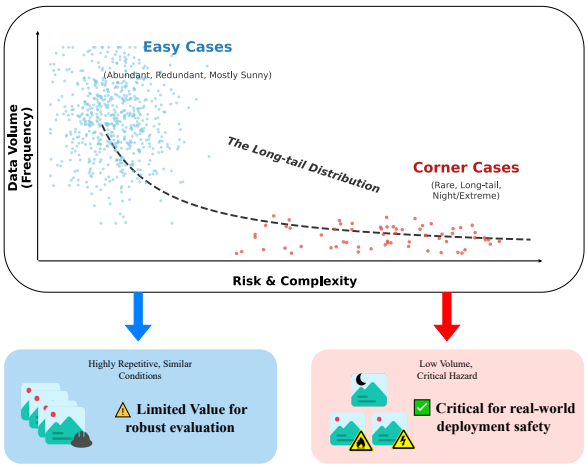

Figure 1: The dichotomy of construction safety data. (Top) Abundant data (blue) is repetitive and low-risk, whereas critical threats form a sparse, complex long tail (red). (Bottom) "Easy cases" dominate volume, while rare corner cases are often missed yet drive real-world safety incidents.

In practice, construction safety monitoring produces archives of inspection records in the form of image–text pairs. Despite their size, these archives are often data-rich but information-poor: much of the volume is repetitive and low-risk, while rare but high-impact hazards form a long tail [42]. Figure 1 illustrates this imbalance. Meanwhile, construction sites evolve over time due to weather, construction progress, camera maintenance, and lighting, which leads to a temporal distribution shift [43]. Evaluation protocols that ignore time shift can overestimate deployment performance [43]. Many existing datasets and benchmarks [2, 9, 40] adopt random splits under an i.i.d. assumption; this can inflate scores when models exploit stable backgrounds or site-specific cues rather than safety-relevant semantics. Performance then drops on footage from later dates, a phenomenon we refer to as *temporal decay* [44]. At the same time, manual benchmark curation is costly and can miss rare but high-impact failure modes that matter in real deployments.

These gaps raise a practical question: *Can we build a construction-safety benchmark that (1) explicitly measures generalization to future time periods and unseen sites, (2) concentrates on long-tail hazards instead of redundant data, and (3) remains feasible to build from large-scale data streams?* To address this, we introduce **SafeBuild-Bench**, a construction safety benchmark designed for robustness evaluation under both temporal and site shift. SafeBuild-Bench is mined from 100,000+ raw image–text pairs collected across multiple sites and dates, and distilled into a compact set of expert-verified evaluation instances under strict protocols. Each instance retains temporal metadata, which enables users to stratify results by time period and site identity, rather than relying on a single fixed split. SafeBuild-Bench covers hierarchical tasks that range from hazard identification to hazard description, and we provide executable evaluation scripts together with a standardized LLM-as-a-judge scheme to score free-form outputs at scale.

Building such a benchmark from redundant streams requires scalable candidate mining before expert verification. We therefore develop an automated curation pipeline, **GEMS** (Graph-Enhanced Multimodal Selection), to extract a small set of informative and diverse candidates from large archives. **GEMS** combines epistemic uncertainty, which ranks samples where current MLLMs are unsure, with a dual graph structure that enforces coverage across scenes and reduces near duplicates. The result is an *informational core*: a compact subset that (i) preserves coverage of the major scene modes and hazard categories, (ii) concentrates hard and rare hazards that are otherwise diluted by repetition, and (iii) removes repeated or near-identical frames that contribute little new information.

We validate the general utility of **GEMS** with a proxy study on a public instruction-tuning corpus. Fine-tuning LLaVA-v1.5 on only the top 1% **GEMS**-selected subset, about 6.6K samples on LLaVA-mix-665K, matches or exceeds full-data training on robustness-oriented benchmarks. This result supports the use of **GEMS** as a practical mining tool for benchmark construction, since it removes redundancy while retaining high-value long-tail content.

We benchmark a diverse set of state-of-the-art MLLMs on SafeBuild-Bench and observe performance degradation under temporal shift, highlighting a reliability gap for safety-critical deployment. Beyond aggregate scores, SafeBuild-Bench enables analysis across task levels and shift types, providing more targeted evidence about when and how current models fail in construction safety understanding.

In summary, our contributions are as follows:

(1) We release **SafeBuild-Bench**, an expert-verified construction safety benchmark mined from 100,000+ raw image–text pairs, featuring long-tail hazards and realistic time and site shift evaluation, with hierarchical annotations and expert rationales covering hazard identification and hazard description.

(2) We provide a metadata-driven evaluation design that enables stratified temporal and site-level robustness analysis, together with executable evaluation scripts and a standardized LLM-as-a-judge scheme for scalable scoring of free-form outputs.

(3) We introduce **GEMS** as an automated selection pipeline that ranks informative candidates while enforcing diversity for expert verification, and we validate its effectiveness on a public dataset under data-budget constraints.

2 Related Work

2.1 Benchmarks for Multimodal LLMs

Early vision–language evaluation largely centered on static VQA-style datasets [15, 16, 18, 35], and domain-focused benchmarks [30, 49], which primarily report answer-level accuracy under largely i.i.d. assumptions. As multimodal LLMs (MLLMs) emerged, broader and more diagnostic benchmarks were proposed. *MME* [12] and *MM-Bench* [29] aim to cover a wider range of perception and cognition skills with standardized protocols and controlled question design. In parallel, hallucination-oriented benchmarks such as *POPE* [24] and *MMHal-Bench* [36] evaluate object-level faithfulness and response reliability, increasingly leveraging *LLM-as-a-judge* (or LLM-assisted normalization) to scale evaluation beyond exact-match metrics. Despite improved coverage and scalability, most existing benchmarks remain temporally static and do not explicitly stress-test MLLMs

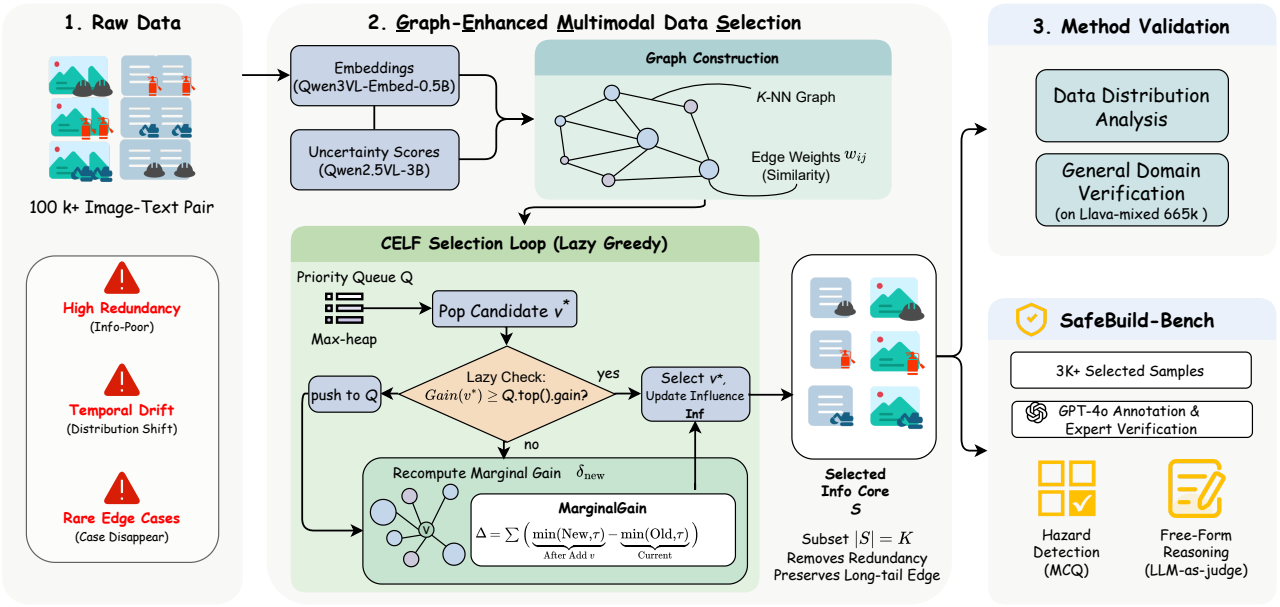

Figure 2: Overview of the GEMS framework. The pipeline involves: (1) Graph Construction: Modeling the data manifold using multimodal embeddings and uncertainty scores; (2) Iterative Selection: A CELF-based lazy greedy algorithm maximizes marginal gain Δ, balancing uncertainty propagation with redundancy saturation (τ); and (3) Validation: The method is verified on general dataset to ensure efficacy before constructing the domain-specific SafeBuild-Bench.

under realistic *time* and *site* shifts. SafeBuild-Bench is designed to complement these efforts by focusing on temporal robustness and distribution shift in a high-stakes industrial domain.

2.2 Data Selection for LLMs and MLLMs

Data curation and selection have become essential for improving training efficiency and robustness in large-scale instruction tuning. Prior work spans influence-based and gradient-driven selection, uncertainty-/difficulty-aware sampling, and geometric coverage or coreset-style methods that reduce redundancy while preserving distributional diversity [11, 20, 28, 41]. Graph-based formulations have recently provided a way to model inter-sample dependence. In particular, the *uncertainty-aware influence maximization* paradigm selects examples that are simultaneously informative (high uncertainty) and representative (strong influence on neighbors), enabling submodular objectives with greedy optimization and approximation guarantees [17]. For multimodal instruction data, recent methods [6, 45] explicitly balance informativeness, uniqueness, and representativeness to curate compact yet effective subsets for MLLM fine-tuning. Our pipeline, **GEMS**, follows this line but targets large, redundant industrial streams: it integrates epistemic uncertainty with a dual-graph structure to identify an informational core suitable for both efficient training and robust benchmarking.

2.3 Hazard Understanding in Construction Sites

Construction site monitoring has traditionally relied on standard computer vision tasks such as object detection for Personal Protective Equipment compliance and related hazard events [21]. While effective for closed-set categories, these supervised approaches typically require dense annotations, for example, bounding boxes, and scale poorly when new hazard types or safety rules must be added. More recently, vision–language models have been explored for open-vocabulary or zero-shot safety assessment in construction settings [2, 33, 34, 40]. However, existing benchmarks often lack a public, expert-verified evaluation set with explicit protocols for time and site shift, and they rarely emphasize long-tail, high-impact hazards within large and redundant inspection streams. **SafeBuild-Bench** complements this line by providing an expert-verified public benchmark mined from real inspection archives, with hierarchical tasks covering hazard identification and hazard description, and with retained temporal metadata to support robustness analysis across time periods and sites.

3 Construction Pipeline for SafeBuild-Bench

Creating a robust industrial benchmark requires moving beyond random sampling. We propose a systematic construction pipeline, as illustrated in Figure 2, comprising four stages: Raw Data Acquisition, Graph-Enhanced Mining (**GEMS**), Method Validation on General Domain dataset, and Expert-in-the-Loop Annotation, separated in the following subsections.

3.1 Raw Data Acquisition and Preprocessing

In collaboration with multiple large-scale construction sites, we curated an extensive dataset of image-text pairs over a five-month period from July to November 2025. These data were collected during the safety inspection phases of the construction process.

Multiple safety experts and field personnel were involved in the acquisition and subsequent verification of the collected samples. The raw dataset comprises over **100,000** image-text pairs sampled from more than 50 construction locations. This collection encompasses diverse scenes (e.g., scaffolding, foundation pits, tower cranes), various facilities (e.g., tower cranes, scaffolding, distribution boxes), and a range of environmental conditions (e.g., nighttime, rainy, and foggy weather). To ensure strict adherence to privacy regulations, an automated anonymization pipeline was implemented. All identifiable faces and license plates were detected and obscured using a specialized obfuscation model. Furthermore, a random subset of the data was manually verified to ensure the complete absence of Personally Identifiable Information (PII).

3.2 The GEMS Selection Engine

We propose **GEMS** (Graph-Enhanced Multimodal Selection), a data-centric framework distilling the "informational core" from redundant industrial streams. As shown in Figure 2, **GEMS** comprises three stages: (1) Multimodal Representation Learning, (2) Epistemic Uncertainty Quantification, and (3) Graph-Theoretic Submodular Selection.

Multimodal Representation Learning. To capture industrial scene semantics, we map image-text pairs into a unified dense vector space. Let $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$ denote the raw pool. We employ the Qwen3-VL-Embedding [23] encoder to extract fused embeddings. Unlike unimodal approaches that process vision and language separately, we utilize the model's native capability to obtain a joint representation:

$$\mathbf{z}_i = \mathcal{F}_\theta(x_i, t_i) \in \mathbb{R}^d, \tag{1}$$

where $\mathcal{F}_\theta$ represents the encoder with frozen parameters. These high-dimensional embeddings ($\mathbf{z}_i$) serve as the geometric basis for our topological analysis.

Epistemic Uncertainty Quantification. **GEMS** hypothesizes that learning value correlates with model confusion. We use a proxy model (e.g., Qwen2.5-VL-3B-Instruct [5]) to quantify the **Epistemic Uncertainty** for each sample. Specifically, prompting the model to analyze the scene (e.g., *Identify safety hazards"*), let y be the generated response sequence of length L. We define the uncertainty score u_i as the length-normalized negative log-likelihood (NLL):

$$u_i = -\frac{1}{L}\sum_{t=1}^{L} \log P_\phi(y_t \mid y_{<t}, x_i). \tag{2}$$

A high u_i implies low generation probability, typically signaling ambiguous scenes, rare anomalies, or "unknown unknowns" (e.g., occluded hazards). We choose the NLL form for numerical stability.

Graph Construction and Optimization. We first construct a sparse k-Nearest Neighbor (k-NN) graph $G = (V, E)$ using the cosine similarity of embeddings $\mathbf{z}_i$, where weight w_{ij} represents semantic proximity. For efficiency ($N > 100k$), we employ GPU-accelerated FAISS [10]. Our objective maximizes the total information covered by a subset $S \subseteq V$, where the "influence" received by node j is defined as $\mathrm{Inf}_j(S) = \sum_{i \in S} w_{ij} \cdot u_i$. To enforce diversity, we apply a saturation threshold τ to model *diminishing returns*,

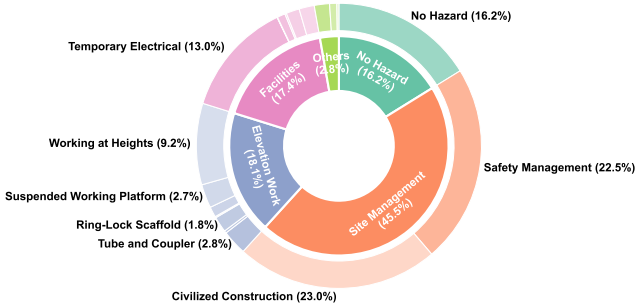

Figure 3: Category distribution of samples in SafeBuild-Bench. The benchmark covers 19 fine-grained hazard categories, which are further grouped into five higher-level safety domains: Site Management, Elevation Work, Facilities, No Hazard, and Others.

yielding the global utility function:

$$F(S) = \sum_{j \in V} \min\left(\mathrm{Inf}_j(S), \tau\right). \tag{3}$$

This monotone and submodular function guarantees a $(1 - 1/e)$-approximation via greedy strategy. To solve this efficiently, we employ the Cost-Effective Lazy Forward (CELF) algorithm [22] (Algorithm 1). By maintaining a priority queue and exploiting submodularity to perform "lazy checks" on marginal gains, CELF [22] avoids $O(K \cdot N)$ complexity, making the selection process scalable.

Overall, **GEMS** synergizes epistemic uncertainty with graph-theoretic selection to distill a compact yet informative subset from massive, redundant industrial streams. This subset effectively captures long-tail hazards and diverse scenarios, making it ideal for constructing robust evaluation benchmarks.

3.3 SafeBuild-Bench Construction

Building upon the curated construction safety data and the validated **GEMS** selection engine, we introduce SafeBuild-Bench, a benchmark for evaluating vision-language models on realistic construction safety understanding tasks.Every instance in SafeBuild-Bench retains its original collection timestamp identifier as metadata, so that users can freely partition the evaluation set along temporal dimensions to measure robustness under different deployment shifts. The complete benchmark, together with the metadata schema and executable evaluation scripts, is publicly released to facilitate reproducible and stratified robustness analysis.

Sample Selection. From the raw data pool, we select over 3,000 samples to form SafeBuild-Bench using the **GEMS** selection engine. The selection process prioritizes safety relevance, visual clarity, and coverage of diverse hazard types. Each selected sample is reviewed to ensure that the depicted scenario supports an unambiguous safety assessment and task formulation. As shown in Figure 3, the selected samples span 19 fine-grained hazard categories, organized into five higher-level safety domains, ensuring broad and structured coverage of construction safety scenarios.

Annotation and Expert Verification. All hazard labels and reference descriptions in SafeBuild-Bench are defined by construction

safety experts. For each selected sample, experts assign the primary hazard category and provide a concise reference description following established construction safety inspection standards. Ambiguous cases are discussed and resolved through expert review to ensure consistency across categories and descriptions.

To reflect realistic inspection ambiguity, hazard categories are organized into expert-defined confusion groups. Each group consists of semantically similar hazard types that are frequently misidentified in practice. These confusion groups are used to construct challenging evaluation instances by pairing the correct hazard category with visually and semantically similar alternatives.

SafeBuild-Bench consists of two complementary tasks that evaluate both categorical recognition and descriptive understanding of construction safety hazards.

Hazard Identification (MCQ). This task evaluates a model's ability to identify the primary construction safety hazard from visually and semantically similar alternatives. For each image, one correct hazard category is paired with three distractor categories sampled from the corresponding expert-defined confusion group. Models are required to select a single best answer. The task contains 2,200 instances and is evaluated using **Accuracy** and **Macro-Recall**. Accuracy measures the proportion of correctly identified hazards, while Macro-Recall computes the average recall across all hazard categories to assess robustness under class imbalance.

Hazard Description (Free-form). This task evaluates a model's ability to generate precise and informative descriptions of construction safety hazards, or to correctly confirm the absence of hazards when applicable. The task contains 1,114 instances. Performance is evaluated using two metrics: **Hazard Detection Rate**, which measures whether the model correctly identifies the presence or absence of a safety hazard, and **Description Quality**, which assesses the semantic completeness and specificity of the generated description. Description Quality is scored on a five-point scale and linearly normalized to $[0, 1]$. An automated LLM-as-a-judge protocol is used for evaluation, with expert reviewers performing random audits to ensure reliability. We provide detailed benchmark case studies and other details in the supplementary material.

4 Experiments

4.1 Experimental Setup

In our experiments, we evaluate the performance of state-of-the-art proprietary and open-source vision-language models on SafeBuild-Bench. The evaluated models cover a wide range of architectures and scales, including OpenAI's GPT-4o [19], Google's Gemini-3-Flash-Preview [14], Gemma-3-12B-IT [37], Anthropic's Claude-4.5-Sonnet [3], Moonshot's Kimi-K2.5 [38], Zhipu AI's GLM-4.5V [39], GLM-4.6V, Qwen's Qwen3-VL-Plus [4], Qwen3-VL-235B-A22B-Instruct [4], Qwen3-VL-4B-Instruct [4], Qwen3-VL-8B-Instruct [4], Qwen2.5-VL-7B-Instruct [5], OpenGVLab's InternVL3-8B [48], OpenBMB's MiniCPM-V-4.5 [46], Microsoft's LLaVA-1.5-7B-HF [27]. All models are evaluated in a zero-shot setting, without any training or fine-tuning on SafeBuild-Bench. We use a unified inference setup with identical prompt templates. The decoding temperature is fixed to 0.2 for all models, except for Kimi-K2.5, which only supports a temperature of 1.0; all other decoding parameters follow the default settings provided by each model provider. For description evaluation, we adopt a unified LLM-as-a-judge protocol using GPT-4o with fixed decoding settings. We separate the closed-source and open-source models in the result Table 1 for clarity.

Algorithm 1 GEMS Selection via CELF (Lazy Greedy)

Require: Indices V, Embeddings $\mathcal{Z}$, Uncertainty $\mathcal{U}$, Budget K, Saturation τ.
Ensure: Selected subset $S \subseteq V$ with $|S| = K$.

Init: Construct k-NN G, weights w_{ij}; $S \leftarrow \emptyset$, $\mathbf{Inf} \leftarrow \mathbf{0}_N$, $Q \leftarrow$ PQueue().
for $v \in V$ **do** // Step 3: Initial Pass
  Q.push(MarginalGain$(v, \mathbf{Inf}), v)$
end for
while $|S| < K$ **do** // Step 4: CELF Selection Loop
  $v^*, \delta_{old} \leftarrow Q$.pop(); $\delta_{new} \leftarrow$ MarginalGain$(v^*, \mathbf{Inf})$
  if $\delta_{new} \geq Q$.top().*gain* **then** // Lazy property check
    $S \leftarrow S \cup \{v^*\}$; $\mathbf{Inf} \leftarrow$ UpdateInfluence$(v^*, \mathbf{Inf})$
  else
    Q.push(δ_{new}, v^*) // Re-insert with new gain
  end if
end while
return S

function MarginalGain$(v, \mathbf{Inf})$
  $\Delta \leftarrow 0$
  for $j \in$ Neighbors(v) **do**
    $p \leftarrow w_{vj} \cdot u_v$; $\Delta \leftarrow \Delta + (\min(\mathbf{Inf}_j + p, \tau) - \min(\mathbf{Inf}_j, \tau))$
  end for
  return Δ
end function

4.2 GEMS Validation

Before presenting the SafeBuild-Bench results, we first validate the reliability of our construction pipeline. We hypothesize that if **GEMS** can effectively identify the "informational core" of a dataset, then a model trained on **GEMS**-selected data should exhibit superior robustness even with minimal samples. We evaluate our data selection method on two widely-used public instruction-tuning datasets with contrasting properties to stress-test the selection efficacy under different compression scenarios:

(1) **LLaVA-Instruct-150K** (Homogeneous Setting): This dataset serves as a test bed for *in-domain homogeneous data compression*. It contains approximately 158K samples generated by GPT-4 [1] based on the COCO [25] dataset. The data is relatively homogeneous, primarily spanning conversation, detailed description, and complex reasoning tasks grounded in daily life images.

(2) **LLaVA-Mixed-665K** (Heterogeneous Setting): In contrast, this dataset represents a *cross-task, cross-modality mixture compression* scenario. It is a heterogeneous collection (665K samples) combining LLaVA instructions with multiple academic datasets (e.g., VQAv2 [15], OCR-VQA [31], RefCOCO [7]). It also introduces task-specific response formatting prompts, adding to the distributional complexity.

Table 1: Main Results on SafeBuild-Bench. We evaluate models on two core capabilities: *Hazard Identification* and *Hazard Description*. We report Global Accuracy and Macro-Recall for identification, and Hazard Detection Rate and Description Quality for description. Prices for closed-source models are reported as input/output costs per 1 million tokens (USD), based on official API pricing. The Overall score is the macro-average of the two tasks, reflecting the comprehensive safety capability. The highest score in each column is shown in bold, and the second-highest score is underlined.

Model	Cost($) / Params	Hazard Identification		Hazard Description		Overall
		Accuracy	Macro-Recall	Hazard Detection Rate	Description Quality	
Closed-Source Models						
Gemini-3-Flash-Preview [14]	0.5 / 3	**55.8**	65.8	67.6	48.4	59.4
Qwen3-VL-Plus [4]	0.1 / 1.4	40.3	53.7	**79.9**	**62.5**	59.1
Claude-4.5-Sonnet [3]	3.0 / 15.0	48.3	56.2	64.4	50.6	54.9
GPT-4o [19]	2.5 / 10.0	34.3	43.8	66.5	50.7	48.8
Open-Source Models						
Kimi-K2.5 [38]	1T	48.7	**67.7**	72.6	53.7	**60.7**
Qwen3-VL-235B-A22B-Instruct [4]	235B	40.6	51.6	62.1	51.3	51.4
Qwen3-VL-8B-Instruct [4]	8B	35.6	40.1	42.2	36.4	38.6
Qwen2.5-VL-7B-Instruct [5]	7B	36.7	36.1	57.2	42.1	43.0
Qwen3-VL-4B-Instruct [4]	4B	33.8	42.6	55.8	42.6	43.7
GLM-4.6V [39]	106B	36.8	40.2	59.6	39.3	44.0
GLM-4.5V [39]	106B	33.7	40.5	57.7	37.2	42.3
Gemma-3-12B-IT [37]	12B	27.7	40.0	55.8	40.5	41.0
InternVL3-8B [48]	8B	38.6	35.9	50.8	38.6	41.0
MiniCPM-V-4.5 [46]	8B	33.9	43.5	45.9	31.7	38.7
LLaVA-1.5-7B-HF [27]	7B	25.3	29.8	38.2	27.2	30.1

We validate **GEMS** using LLaVA-v1.5-7B [27] as the base. Comparisons include: (1) Random Selection (standard baseline); (2) DataTailor [45]; and (3) Full Data Training (ceiling). Evaluation metrics include robustness-oriented benchmarks: VizWiz [16] (real-world noise), MMMU [47] (expert reasoning), MME [12] (perception & cognition), and POPE [24] (hallucination). We group these metrics into *General Perception*, *Robustness & Reliability*, and *Knowledge & Reasoning*. The overall score is the average across all benchmarks.

Implementation Details. For **GEMS**, we set the uncertainty weight to $\lambda = 0.6$ and use Qwen2.5-VL-3B as the proxy model. We also ablate the uncertainty weight on LLaVA-Mixed-665K by testing $\lambda \in \{2, 4\}$, which increasingly up-weights model confusion in the selection objective. For all fine-tuned models, we freeze the vision tower and projector, and apply LoRA (rank 64, α 128) to all linear layers. Models are fine-tuned for 3 epochs with a learning rate of 2e-4, an effective batch size of 64, and a cosine schedule.

Results on Homogeneous Data (LLaVA-Instruct-150K). Table 2 demonstrates the efficacy of **GEMS** under a relatively homogeneous instruction distribution. First, we observe an extreme *data efficiency gain*: training with merely **1%** of the data (1.5k samples) surpasses the **Random Selection** baseline using 20% data (32k samples) on aggregate metrics (MME: 1699.8 vs. 1675.0). Notably, the 1% subset achieves the highest **POPE** score (**86.5**) among all settings, including Full Data (85.5), suggesting that filtering out redundant captions significantly mitigates hallucination. When scaling to **20%**, **GEMS** consistently breaks the performance ceiling of the full dataset, particularly on reasoning-intensive and noisy benchmarks, achieving **55.2** on VizWiz (+1.1 over Full) and **35.7** on MMMU (+2.0 over Full). Remarkably, **GEMS** achieves **99.5%** of the full-data performance using only **1%** of the samples (1.5k). At the 20% scale, it further surpasses the full dataset with a relative score of **101.1%**, suggesting that **GEMS** effectively filters out redundant or noisy data that hinders optimization.

Results on Homogeneous Data (LLaVA-Instruct-150K). Table 2(a) demonstrates the efficacy of **GEMS** under a relatively homogeneous instruction distribution. First, we observe an extreme *data efficiency gain*: training with merely **1%** of the data (1.5k samples) surpasses the **Random Selection** baseline using 20% data (32k samples) on aggregate metrics (MME: 1699.8 vs. 1675.0). Notably, the 1% subset achieves the highest **POPE** score (**86.5**) among all settings, including Full Data (85.5), suggesting that filtering out redundant captions significantly mitigates hallucination. When scaling to **20%**, **GEMS** consistently breaks the performance ceiling of the full dataset, particularly on reasoning-intensive and noisy benchmarks, achieving **55.2** on VizWiz (+1.1 over Full) and **35.7** on MMMU (+2.0 over Full). Remarkably, **GEMS** achieves **99.5%** of the full-data performance using only **1%** of the samples (1.5k). At the 20% scale, it further surpasses the full dataset with a relative score of **101.1%**, suggesting that **GEMS** effectively filters out redundant or noisy data that hinders optimization.

Results on Heterogeneous Mixture (LLaVA-Mixed-665K). Table 2(b) reports selection performance on the large-scale, multi-task LLaVA-665K dataset. Under this highly redundant mixture, **GEMS** improves data efficiency on robustness-oriented benchmarks. A model fine-tuned on 1% of **GEMS**-selected data (about 6.6k samples) outperforms the full-data model (665k) on in-the-wild evaluation, for example, 53.7 on VizWiz (+5.9 over full data) and 34.8 on MMMU (+2.0 over full data). Compared with the Random-8% baseline, **GEMS**-1% also yields a clear gain on MME (1805.0 vs. 1675.0). Overall, **GEMS** retains 98.3% of the full-data average while using

Table 2: GEMS validation on public instruction-tuning datasets. LLaVA-v1.5-7B is fine-tuned on subsets. MME is the aggregate of perception and cognition scores. Avg. % denotes average relative performance versus the Full Data baseline. Among data-efficient methods (excluding Full Data), the highest score in each column is shown in bold and the second-highest is underlined.

Method	Data Size	General Perception			Robustness & Reliability		Knowledge & Reasoning			Avg. %
		VQAv2	GQA	MME	VizWiz	POPE	MMMU	SciQA	TextVQA	
(a) LLaVA-Instruct-150K (Homogeneous Setting)										
Full Data	100% (158k)	73.7	60.6	1780.8	54.1	85.5	33.7	66.3	47.4	100.0
Random Selection	20% (32k)	**75.6**	60.5	1675.0	42.3	85.7	32.8	65.8	47.6	97.5
Ours: GEMS										
GEMS	1% (1.5k)	75.3	60.4	1699.8	52.1	**86.5**	34.2	**66.2**	47.7	99.5
GEMS	10% (15k)	75.4	60.3	**1768.8**	53.0	86.1	35.3	65.3	47.1	100.2
GEMS	20% (32k)	**75.6**	**60.6**	1755.2	**55.2**	85.8	**35.7**	65.8	**47.8**	**101.1**
(b) LLaVA-Mixed-665K (Heterogeneous Setting)										
Full Data	100% (665k)	79.1	63.0	1744.8	47.8	86.4	32.8	70.0	58.2	100.0
Random Selection	8% (50k)	73.7	55.0	1675.0	42.3	85.7	32.2	70.0	**53.1**	93.3
DataTailor [45]	8% (50k)	75.0	57.7	**1823.7**	46.3	82.1	33.9	**70.9**	**53.1**	96.9
Ablation: Uncertainty Weight										
GEMS w/$\lambda=2$	8% (50k)	74.9	60.4	1766.6	49.4	86.0	34.8	64.8	47.5	97.4
GEMS w/$\lambda=2$	1% (6k)	**75.1**	**60.8**	1691.8	49.3	85.9	34.2	64.7	47.8	96.9
GEMS w/$\lambda=4$	1% (6k)	72.3	58.4	1642.7	52.9	84.7	34.9	64.3	45.9	96.0
Ours: GEMS										
GEMS	1% (6k)	**75.1**	60.4	1805.0	53.7	**86.2**	34.8	65.5	47.3	**98.3**
GEMS	8% (50k)	74.8	60.1	1781.9	51.6	85.9	36.0	65.4	46.2	97.8
GEMS	15% (100k)	74.8	60.3	1728.3	**54.0**	85.3	**36.1**	66.3	46.7	98.1

only 1% of the training samples. These results suggest that selection can reduce redundancy in heterogeneous instruction mixtures while preserving the hard and informative content that matters for robustness evaluation. We use this behavior as a proxy validation signal for constructing SafeBuild-Bench: the same selection mechanism can prioritize informative and non-duplicate candidates in large industrial streams before expert verification.

4.3 Main Results

Table 1 reports the main results on SafeBuild-Bench across two core capabilities: Hazard Identification and Hazard Description. Overall, current vision-language models remain far from robust construction safety understanding. Even the best-performing models achieve Overall scores around 60, indicating substantial room for improvement across both identification and description tasks.

For **Hazard Identification**, performance varies notably across models. Gemini-3-Flash-Preview achieves the highest accuracy (55.8), while Kimi-K2.5 attains the best Macro-Recall (67.7), suggesting stronger robustness under class imbalance. In contrast, several models exhibit a large gap between Accuracy and Macro-Recall, indicating that correct predictions are often concentrated in frequent categories while rare hazards remain challenging.

For **Hazard Description**, Qwen3-VL-Plus achieves the highest Hazard Detection Rate (79.9) and Description Quality (62.5). This suggests that strong descriptive ability does not necessarily correlate with categorical identification accuracy, as Qwen3-VL-Plus lags behind top models on identification metrics. Across models, Hazard Detection Rate is generally higher than Description Quality, indicating that models can often detect the existence of hazards but struggle to provide precise and informative descriptions.

Comparing closed-source and open-source models, we observe competitive performance from large open-source systems. Kimi-K2.5 achieves the highest Overall score (60.7) among all evaluated models, outperforming several proprietary counterparts. However, smaller open-source models underperform, highlighting the importance of model scale for construction safety understanding.This trend is consistent across open-source models, where performance improvements are primarily driven by increases in model scale rather than architectural variations.

Taken together, these results reveal that construction safety remains a challenging domain for vision-language models. Strong performance on one task does not guarantee robustness on the other, underscoring the need for holistic evaluation across both hazard identification and hazard description capabilities.

4.4 Error Analysis

Hazard Identification. Figure 4 illustrates category-wise identification accuracy for representative models on the Hazard Identification task. A clear pattern emerges: model performance varies substantially across hazard categories, indicating that identification

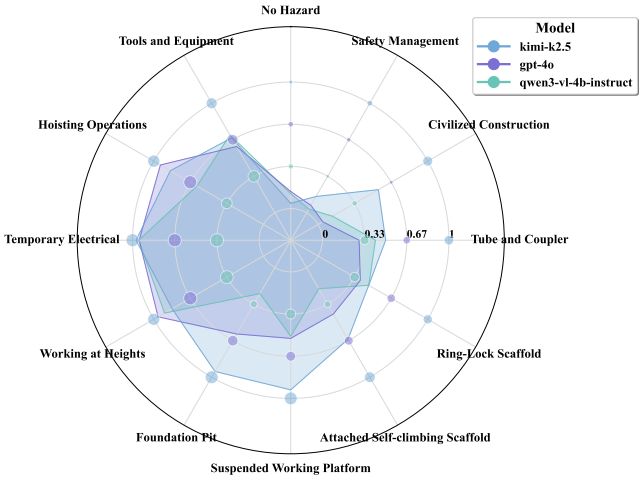

Figure 4: Category-wise identification accuracy on the *Hazard Identification* task. The radar chart compares three representative models, Kimi-K2.5, GPT-4o, and Qwen3-VL-4B-Instruct, across major hazard categories. Each axis corresponds to a hazard category, and values indicate identification accuracy within that category.

errors are strongly category-dependent rather than uniformly distributed. We also observe consistent trends across different models, suggesting shared challenges inherent to specific hazard types.

Models generally perform better on hazard categories with distinctive visual cues and well-defined objects, such as *Temporary Electrical* and *Hoisting Operations*. In contrast, categories involving subtle structural differences or contextual safety rules, such as *Tube and Coupler*, *Ring-Lock Scaffold*, and *Safety Management*, consistently exhibit lower accuracy across models. These categories often require fine-grained discrimination between visually similar configurations or rely on implicit safety norms, which remain challenging for current vision-language models.

Another notable failure mode is misclassifying *No Hazard* cases. Despite the absence of explicit safety violations, models frequently over-predict hazards, suggesting a bias toward hazard presence. This tendency indicates limited calibration in distinguishing compliant construction scenarios from genuinely unsafe ones.

Comparing models, larger models such as Kimi-K2.5 and GPT-4o show relatively more stable performance across categories, while smaller models like Qwen3-VL-4B-Instruct exhibit sharper performance drops in complex or ambiguous categories. This observation suggests that model capacity plays an important role in handling fine-grained hazard distinctions, although even large models remain far from reliable across all categories.

Hazard Description. We further analyze the Hazard Description task using GPT-4o as an automatic judge, which outputs a hazard-detection signal (HDR) and a normalized description-quality score. We define *major* description failures as cases with HDR = 0 or Final ≤ 0.5 (excluding judge errors and dataset-conflict cases). Under this criterion, Kimi-K2.5 exhibits a 27.3% major-failure rate (304/1114), while the smaller model Qwen3-VL-4B-Instruct rises to 44.4% (495/1114), indicating a capacity gap in hazard narration.

Across both models, major failures are overwhelmingly driven by breakdowns at the hazard detection stage rather than deficiencies in linguistic expression. Almost all major failures coincide with HDR = 0 (303/304 for Kimi-K2.5; 493/495 for Qwen3-VL-4B-Instruct), suggesting that once a hazard is correctly detected, models usually produce descriptions of acceptable quality.

The two models exhibit distinct failure patterns. Kimi-K2.5 often fails by describing hazards that appear plausible but do not satisfy the ground-truth safety criteria (52.0% of major failures). False alarms in genuinely safe scenes account for a further 30.9%. In contrast, Qwen3-VL-4B-Instruct is dominated by missed hazards, frequently responding with generic "no hazard" or "safe" statements when safety violations are present (70.7%). The remaining cases mainly involve mismatched hazard descriptions (21.0%).

For both models, failures are concentrated in scenarios that require contextual grounding rather than the recognition of a salient object, such as signage and housekeeping compliance, barriers and guardrails, scaffolding, and temporary electrical setups. These results indicate that hazard description remains challenging when safety violations are defined by implicit rules or subtle visual cues.

5 Discussion and Limitations

Our results highlight a practical issue in construction safety monitoring: large inspection archives are often dominated by redundancy, while rare hazards that matter for deployment form a long tail. SafeBuild-Bench is designed to make this setting measurable by retaining temporal metadata and enabling stratified analysis across time periods and sites, which better reflects real deployment shift. The proxy studies also suggest that selection can preserve robustness-relevant content under small data budgets, supporting a data-centric approach to benchmark construction. Limitations remain: SafeBuild-Bench has about 3,000 expert-verified samples, scaling will require sustained expert effort, and **GEMS** depends on proxy models for uncertainty estimation, so weak proxies may under-select certain hazard types; ensembles and complementary signals are a natural next step. Finally, extending the same recipe to other domains will require new taxonomies and expert guidelines, even if the selection mechanism remains similar.

6 Conclusion

We release SafeBuild-Bench, a construction safety benchmark for evaluating MLLMs under realistic time and site shifts. SafeBuild-Bench is mined from large-scale inspection archives and curated into an expert-verified evaluation set with tasks covering hazard identification and hazard description. Each instance retains temporal metadata, enabling stratified analysis across time periods and sites with executable evaluation scripts and a standardized LLM-as-a-judge protocol. To make construction feasible from redundant streams, we develop **GEMS**, a graph-enhanced mining pipeline that prioritizes informative and non-duplicate candidates for expert verification. In proxy studies on public instruction-tuning data, fine-tuning on a small **GEMS**-selected subset can match or exceed full-data training on robustness-oriented benchmarks, suggesting that selection can reduce redundancy while preserving hard cases. We also release the dataset, benchmark, curation codebase, and the evaluation scripts to support reproducible research on safety-focused multimodal models.

References

[1] Josh Achiam et al. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] doi:10.48550/ARXIV.2303.08774

[2] Muhammad Adil, Gaang Lee, Vicente A. Gonzalez, and Qipei Mei. 2025. Using Vision Language Models for Safety Hazard Identification in Construction. arXiv:2504.09083 [cs.CV] https://arxiv.org/abs/2504.09083

[3] Anthropic. 2025. *Claude 4.5 Sonnet System Card.* Technical Report. Anthropic.

[4] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. 2025. Qwen3-VL Technical Report. arXiv:2511.21631 [cs.CV] https://arxiv.org/abs/2511.21631

[5] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, et al. 2025. Qwen2.5-VL Technical Report. arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[6] Powei Chang, Jinpeng Zhang, Bowen Chen, Chenyu Wang, Chenlu Guo, Yixing Zhang, Yukang Gao, JianXiang Xiang, Yue Gao, Chaoqun Sun, Yiyi Chen, and Dongying Kong. 2026. SPICE: Submodular Penalized Information-Conflict Selection for Efficient Large Language Model Training. arXiv:2601.23155 [cs.LG] https://arxiv.org/abs/2601.23155

[7] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. 2025. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 513–524.

[8] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

[9] V. S. K. Delhi, R. Sankarlal, and Albert Thomas. 2020. Detection of Personal Protective Equipment (PPE) Compliance on Construction Site Using Computer Vision Based Deep Learning Techniques. *Frontiers in Built Environment* 6 (2020). doi:10.3389/fbuil.2020.00136

[10] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The Faiss library. arXiv:2401.08281 [cs.LG] https://arxiv.org/abs/2401.08281

[11] Dan Feldman. 2020. Core-sets: An Updated Survey. *arXiv preprint arXiv:2011.09384* (2020). doi:10.48550/ARXIV.2011.09384

[12] Chaoyou Fu, Jun Chen, et al. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394* (2023). doi:10.48550/ARXIV.2306.13394

[13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. arXiv:1803.09010 [cs.DB] https://arxiv.org/abs/1803.09010

[14] Google. 2025. *Gemini 3 Flash Model Card.* Technical Report. Google.

[15] Yash Goyal, Tanishq Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.

[17] Jindong Han, Hao Liu, Jun Fang, Naiqiang Tan, and Hui Xiong. 2025. Automatic Instruction Data Selection for Large Language Models via Uncertainty-Aware Influence Maximization. In *Proceedings of the ACM on Web Conference 2025 (WWW '25)*. ACM, 4969–4979. doi:10.1145/3696410.3714817

[18] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[19] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).

[20] Jaewoo Lee, Boyang Li, and Sung Ju Hwang. 2024. Concept-skill transferability-based data selection for large vision-language models. *arXiv preprint arXiv:2406.10995* (2024).

[21] Yeo-Reum Lee, Seung-Hwan Jung, Kyung-Su Kang, Han-Cheol Ryu, and Han-Guk Ryu. 2023. Deep learning-based framework for monitoring wearing personal protective equipment on construction sites. *Journal of Computational Design and Engineering* 10, 2 (2023), 905–917.

[22] Jure Leskovec, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Vanbriesen, and Natalie Glance. 2007. Cost-effective outbreak detection in networks. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 420-429, 420–429. doi:10.1145/1281192.1281239

[23] Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. 2026. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking. *arXiv* (2026).

[24] 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 (EMNLP)*. 292–305. doi:10.18653/v1/2023.emnlp-main.20

[25] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. arXiv:1405.0312 [cs.CV] https://arxiv.org/abs/1405.0312

[26] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV] https://arxiv.org/abs/2310.03744

[27] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 26296–26306.

[28] Liangxin Liu, Xuebo Liu, Derek F Wong, Dongfang Li, Ziyi Wang, Baotian Hu, and Min Zhang. 2024. Selectit: Selective instruction tuning for llms via uncertainty-aware self-reflection. *Advances in Neural Information Processing Systems* 37 (2024), 97800–97825.

[29] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023. MMBench: Is Your Multi-modal Model an All-around Player? *arXiv preprint arXiv:2307.06281* (2023). doi:10.48550/ARXIV.2307.06281

[30] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *arXiv preprint arXiv:2209.09513* (2022). doi:10.48550/ARXIV.2209.09513

[31] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. OCR-VQA: Visual Question Answering by Reading Text in Images. In *ICDAR*.

[32] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[33] Zhenhui Ou, Dawei Li, Zhen Tan, Wenlin Li, Huan Liu, and Siyuan Song. 2025. Building Safer Sites: A Large-Scale Multi-Level Dataset for Construction Safety Benchmark. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management* (Seoul, Republic of Korea) *(CIKM '25)*. Association for Computing Machinery, New York, NY, USA, 6508–6512. doi:10.1145/3746252.3761652

[34] Ahmed Bin Kabir Rabbi and Idris Jeelani. 2024. AI integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications. *Automation in Construction* 164 (2024), 105443.

[35] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Manohar Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. 2019. Towards VQA Models That Can Read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Zhiqing Sun, Sheng Shen, et al. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. *arXiv preprint arXiv:2309.14525* (2023). doi:10.48550/ARXIV.2309.14525

[37] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786* (2025).

[38] Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood

Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. 2026. Kimi K2.5: Visual Agentic Intelligence. arXiv:2602.02276 [cs.CL] https://arxiv.org/abs/2602.02276

[39] V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. 2025. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile Multimodal Reasoning with Scalable Reinforcement Learning. arXiv:2507.01006 [cs.CV] https://arxiv.org/abs/2507.01006

[40] Antoine J. P. Tixier and Matthew R. Hallowell. 2023. Safer Together: Machine Learning Models Trained on Shared Accident Datasets Predict Construction Injuries Better than Company-Specific Models. arXiv:2301.03567 [cs.LG] https://arxiv.org/abs/2301.03567

[41] Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. Selecting Influential Data for Targeted Instruction Tuning. *arXiv preprint arXiv:2402.04333* (2024). doi:10.48550/ARXIV.2402.04333

[42] Lu Yang, He Jiang, Qing Song, and Jun Guo. 2022. A survey on long-tailed visual recognition. *International Journal of Computer Vision* 130, 7 (2022), 1837–1872.

[43] Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei W Koh, and Chelsea Finn. 2022. Wild-time: A benchmark of in-the-wild distribution shift over time. *Advances in Neural Information Processing Systems* 35 (2022), 10309–10324.

[44] Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715* (2024).

[45] Qifan Yu, Zhebei Shen, Zhongqi Yue, Yang Wu, Wenqiao Zhang, Yunfei Li, Juncheng Li, Siliang Tang, and Yueting Zhuang. 2024. Mastering Collaborative Multi-modal Data Selection: A Focus on Informativeness, Uniqueness, and Representativeness. *arXiv preprint arXiv:2412.06293* (2024). doi:10.48550/ARXIV.2412.06293

[46] Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. 2025. MiniCPM-V 4.5: Cooking Efficient MLLMs via Architecture, Data, and Training Recipe. arXiv:2509.18154 [cs.LG] https://arxiv.org/abs/2509.18154

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

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

[49] Xing Zi, Jinghao Xiao, Yunxiao Shi, Xian Tao, Jun Li, Ali Braytee, and Mukesh Prasad. 2025. RSVLM-QA: A Benchmark Dataset for Remote Sensing Vision Language Model-based Question Answering. In *Proceedings of the 33rd ACM International Conference on Multimedia*. 12905–12911.

A Acknowledgments for Data Collection

We gratefully acknowledge the following individuals and organizations for their contributions to data collection and data verification in this project:

- **Jianhua Yang** (zhbgs@gdzgy.com), Guangdong Zhonggong Architectural Design Institute Co., Ltd.
- **Jian Lin** (gdzgjl@gdzgjl.com), Guangdong Zhonggong Project Management Co., Ltd.
- **Yan Liu** (gdzgjl@gdzgjl.com), Guangdong Zhonggong Project Management Co., Ltd.

B Datasheet for SafeBuild-Bench

Following the standardized framework recommended by Gebru et al. [13], we provide the datasheet for SafeBuild-Bench. This datasheet documents the motivation, composition, collection process, preprocessing, intended uses, distribution, and maintenance plan for the benchmark.

B.1 Motivation

- **For what purpose was the dataset created?** SafeBuild-Bench was created to evaluate the temporal robustness and construction safety understanding capabilities of Multimodal Large Language Models (MLLMs). Existing benchmarks in this domain primarily adopt random (i.i.d.) splits that can inflate performance by allowing models to memorize site-specific backgrounds rather than learning safety-relevant semantics. SafeBuild-Bench addresses this gap by providing an expert-verified evaluation set with retained temporal metadata, enabling stratified analysis across time periods and sites under realistic deployment conditions.
- **Who created the dataset and on behalf of which entity?** The dataset was curated by researchers at The Hong Kong University of Science and Technology (Guangzhou), in collaboration with construction industry partners who provided access to real-world inspection archives. Ground truth annotations were defined and verified by certified construction safety engineers.
- **Who funded the creation of the dataset?** The dataset creation was supported by the authors' affiliated institutions.
- **Any other comments?** The creation of SafeBuild-Bench is motivated by a data-centric AI philosophy: rather than scaling data volume alone, the benchmark emphasizes information density and distributional coverage. The accompanying **GEMS** selection pipeline was developed and validated specifically to enable scalable, principled curation from massive and redundant industrial data streams.

B.2 Composition

- **What do the instances that comprise the dataset represent?** Each instance in SafeBuild-Bench is an image–text pair originating from real construction site inspection records. The image is a surveillance or inspection photograph depicting a construction scene, and each instance is accompanied by structured annotations including: (1) the primary hazard category, (2) an expert-written reference description following construction safety inspection standards, and (3) task-specific evaluation metadata (e.g., multiple-choice options with confusion-group distractors for the identification task, or free-form reference answers for the description task). Each instance also retains temporal (date) and site-level metadata.
- **How many instances are there in total?** The benchmark contains over 3,000 expert-verified images, yielding a total of 3,314 task instances across two complementary evaluation tasks:
  - **Hazard Identification (MCQ):** 2,200 instances, each comprising an image, a question, four answer options (one correct category plus three expert-defined confusion-group distractors), and the correct answer label.
  - **Hazard Description (Free-form):** 1,114 instances, each comprising an image, a prompt requesting a safety hazard description, and an expert-written reference description.
- **Does the dataset contain all possible instances or is it a sample?** The benchmark is a curated sample. It was mined from a raw pool of over 100,000 image–text pairs using the **GEMS** selection pipeline, which prioritizes informative, diverse, and non-redundant candidates for subsequent expert verification. The selection process is designed to concentrate long-tail hazards and reduce the dominance of repetitive, low-risk scenes.
- **What data does each instance consist of?** Each instance consists of:
  - A JPEG image of a construction scene (privacy-protected; see below).
  - A hazard category label from 19 fine-grained categories organized into 5 higher-level safety domains: Site Management, Elevation Work, Facilities, No Hazard, and Others.
  - Task-specific annotations (MCQ options or free-form reference description).
  - Temporal metadata (collection date) and site identifier.
- **Is there a label or target associated with each instance?** Yes. For the Hazard Identification task, the label is the correct hazard category among four options. For the Hazard Description task, the target is the expert-written reference description and a binary hazard-presence indicator.
- **Is any information missing from individual instances?** No. All released instances contain the full set of annotations described above.
- **Are relationships between individual instances made explicit?** Instances are linked through shared site identifiers and temporal metadata, which enable grouping by site or time period for stratified evaluation. The 19 hazard categories are further organized into expert-defined confusion groups that capture common misidentification patterns.
- **Are there any errors, sources of noise, or redundancies?** All samples have been reviewed by construction safety experts. Ambiguous cases were discussed and resolved through expert consensus. The **GEMS** pipeline was specifically designed to minimize redundancy in the candidate pool before expert verification. Nevertheless, some residual annotation ambiguity may

exist for borderline safety scenarios, which is inherent to the domain.

- **Is the dataset self-contained, or does it link to or otherwise rely on external resources?** The dataset is self-contained. All images, annotations, and metadata are included in the release. The evaluation scripts rely on standard open-source libraries and, for the Hazard Description task, access to an LLM judge (e.g., GPT-4o[19]) for automated scoring.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** The dataset depicts real construction site conditions, some of which involve genuine safety hazards (e.g., improperly installed safety netting, exposed electrical cables, missing guardrails). These images are representative of occupational safety risks and are included for research purposes. They do not contain graphic injury content.
- **Does the dataset relate to people?** Construction workers may appear in the images, but all faces are blurred and no personally identifiable information is retained. The dataset does not contain demographic labels or behavioral tracking data.
- **Does the dataset identify any subpopulations (e.g., by age, gender)?** No. The dataset does not identify, label, or distinguish any human subpopulations.

B.3 Collection Process

- **How was the data associated with each instance acquired?** Images were captured by fixed surveillance and inspection cameras installed at active construction sites. The accompanying text descriptions were recorded as part of routine safety inspection workflows by field personnel and safety experts.
- **What mechanisms or procedures were used to collect the data?** Data was sourced from more than 50 construction locations in collaboration with multiple large-scale construction projects. The collection covers diverse scenes (e.g., scaffolding, foundation pits, tower cranes), various facilities (e.g., tower cranes, scaffolding, distribution boxes), and a range of environmental conditions (e.g., daytime, nighttime, rainy, and foggy weather).
- **If the dataset is a sample from a larger set, what was the sampling strategy?** The benchmark was mined from a raw pool of 100,000+ image–text pairs using the **GEMS** selection engine. **GEMS** combines epistemic uncertainty quantification (identifying samples where a proxy MLLM is uncertain) with a dual-graph structure (enforcing semantic coverage and reducing near-duplicates) to extract the informational core. The resulting candidate set was then subjected to expert verification and annotation.
- **Who was involved in the data collection process and how were they compensated?** Multiple safety experts and field inspection personnel were involved in data acquisition, annotation, and verification. All personnel were compensated as part of their professional duties through their respective organizations.
- **Over what timeframe was the data collected?** Data was collected over a five-month period from July 1, 2025, to November 30, 2025.n.
- **Does the dataset relate to people?** Workers may be visible in construction scenes, but all identifying features are anonymized. The dataset is designed to evaluate safety conditions of construction environments, not to monitor or identify individual workers.

B.4 Preprocessing, Cleaning, and Labeling

- **Was any preprocessing/cleaning/labeling of the data done?** Yes. The following preprocessing steps were applied:
  (1) *Privacy anonymization:* Automated detection and blurring of all human faces and vehicle license plates, followed by manual verification on a random 5% subset.
  (2) *Candidate mining:* The **GEMS** pipeline was applied to the raw 100,000+ pool to rank and diversify candidates, producing a compact candidate set enriched in long-tail hazards.
  (3) *Expert annotation:* Construction safety experts assigned hazard category labels and wrote reference descriptions for each selected sample following established construction safety inspection standards.
  (4) *Confusion group construction:* Experts defined semantically similar hazard groupings to generate challenging MCQ distractors.
  (5) *Quality assurance:* Ambiguous annotations were resolved through expert consensus review.
- **Was the "raw" data saved in addition to the preprocessed /cleaned/labeled data?** The raw data pool is retained internally for potential future benchmark expansions but is not part of the public release. Only the expert-verified, anonymized benchmark instances are released.
- **Is the software used to preprocess/clean/label the data available?** Yes. The **GEMS** codebase (selection pipeline) and the evaluation scripts (including the LLM-as-a-judge protocol) are released alongside the benchmark.

B.5 Uses

- **Has the dataset been used for any tasks already?** Yes. The dataset has been used to benchmark a diverse set of state-of-the-art MLLMs, including both proprietary models, as reported in Section 4.
- **Is there a repository that links to any or all papers or systems that use the dataset?** Yes. The dataset repository (hosted on Hugging Face and GitHub) will maintain a list of publications and systems that use SafeBuild-Bench.
- **What (other) tasks could the dataset be used for?** Beyond the two primary tasks (Hazard Identification and Hazard Description), SafeBuild-Bench could support research on:
  - Temporal robustness analysis of vision–language models.
  - Domain adaptation and transfer learning for industrial safety.
  - Long-tail recognition and few-shot learning in safety-critical settings.
  - Calibration and uncertainty estimation of MLLMs.

  - Development and evaluation of LLM-as-a-judge methods for free-form safety assessment.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** The dataset is specific to Chinese construction sites and may not generalize to construction practices in other countries without adaptation. The hazard taxonomy (19 categories across 5 domains) follows Chinese construction safety inspection standards. Users applying the benchmark in other regulatory contexts should be aware of potential taxonomy mismatches.
- **Are there tasks for which the dataset should not be used?** The dataset must **not** be used for: biometric identification, worker surveillance, punitive monitoring, or any application that targets individual workers. It is released strictly for research purposes aimed at improving workplace safety monitoring systems.

B.6 Distribution

- **Will the dataset be distributed to third parties outside of the entity on behalf of which the dataset was created?** Yes. The dataset is publicly released for research use.
- **How will the dataset be distributed?** The dataset will be hosted on Hugging Face Datasets and GitHub, with accompanying evaluation scripts and the **GEMS** codebase.
- **When will the dataset be distributed?** The dataset is released concurrently with the publication of this paper.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** Yes. The dataset is released under the **CC-BY-NC-SA 4.0** (Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International) license. This permits non-commercial use with attribution and requires derivative works to be shared under the same license.
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** The construction industry partners have authorized the release of the anonymized data for non-commercial research purposes under the terms described above.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** Not to our knowledge.

- **Who is supporting/hosting/maintaining the dataset?** The dataset is maintained by the authors at The Hong Kong University of Science and Technology (Guangzhou). Hosting is provided via Hugging Face Datasets and GitHub.
- **How can the owner/curator/manager of the dataset be contacted?** Through the corresponding authors listed in this paper, or via the issue tracker on the dataset's GitHub repository.
- **Is there an erratum?** Not currently. Any corrections will be documented in the dataset repository's changelog.
- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** Yes. We plan to release updated versions to correct any identified errors and to expand coverage. A planned v2.0 release will incorporate video-based evaluation tasks.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances?** All personal identifying features have been irreversibly anonymized (blurred) prior to release. No PII is retained in the dataset, so standard data retention limits for identifiable personal data do not apply.
- **Will older versions of the dataset continue to be supported/hosted/maintained?** Yes. All released versions will remain available on the hosting platforms with clear version identifiers to support reproducibility.
- **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** Contributors may submit extensions or corrections via pull requests on the GitHub repository. Proposals for new hazard categories or evaluation tasks will be reviewed by the maintainers in consultation with domain experts.

Their domain expertise and on-site support were essential for ensuring the quality, safety relevance, and regulatory compliance of the SafeBuild-Bench dataset.

C Benchmark Case Study

To illustrate the evaluation design of SafeBuild-Bench, we present four representative examples in Figure 5 and 6, organized into two rows covering both benchmark tasks under different visual conditions.

Hazard Identification (MCQ). The left column shows four multiple-choice samples that test a model's ability to distinguish the correct hazard category from semantically similar distractors drawn from expert-defined confusion groups.

Hazard Description (Free-form). The right column demonstrates the open-ended description task together with the LLM-as-a-judge evaluation protocol.

These cases collectively illustrate several key challenges embedded in SafeBuild-Bench: (1) fine-grained discrimination among visually and semantically similar hazard categories; (2) robustness to adverse environmental conditions such as nighttime and occlusion; (3) the gap between hazard detection (binary) and hazard description (requiring precise, grounded reasoning); and (4) the role of the LLM-as-a-judge scheme in providing scalable, interpretable, and multi-dimensional evaluation of free-form model outputs.

D Disclosure of AI Use

In the preparation of this manuscript, we utilized AI tools to assist with linguistic polishing and the formatting of tables to improve readability. The authors have reviewed and edited the content, and take full responsibility for the accuracy and integrity of the published work.

Task: Hazard Detection (MCQ)

Question: Identify the primary construction safety hazard type shown in the image.

Options:

A. Temporary Electrical Supply
B. **No Hazard** ✓
C. Safety Management
D. Civilized Construction

Analysis: *Option B is correct because the construction site is orderly, with fire extinguishers properly stored and accessible. Safety measures appear to be in place.*

Task: Hazard Detection (MCQ)

Question: Identify the primary construction safety hazard type shown in the image.

Options:

A. No Hazard
B. Construction Tools and Equipment
C. **Safety Management** ✓
D. Civilized Construction

Analysis: *Option C is correct because The area lacks proper safety signage to warn of potential hazards, and the equipment is covered with a tarp, which may conceal tripping hazards.*

Task: Hazard Detection (MCQ)

Question: Identify the primary construction safety hazard type shown in the image.

Options:

A. Temporary Electrical Supply
B. Safety Management
C. **No Hazard** ✓
D. Construction Tools and Equipment

Analysis: *Option C is correct because the construction site is orderly, with equipment neatly arranged and safety measures such as caution signs visible.*

Task: Hazard Detection (MCQ)

Question: Identify the primary construction safety hazard type shown in the image.

Options:

A. **Construction Tools and Equipment** ✓
B. Hoisting Operations
C. Tower Crane
D. Temporary Electrical Supply

Analysis: *Option A is correct because the power cables are scattered across the floor, creating a tripping hazard in the construction area. And this equipment are not properly protected, which presents potential safety hazards.*

Figure 5: Qualitative Examples from SafeBuild-Bench. (Left) Four MCQ samples testing precise hazard identification. The correct option is highlighted.

Task: Open-Ended Description & Evaluation

Ground Truth: *The **safety netting** is improperly installed, with **visible gaps** that could lead to falls from height.*

Model Prediction: There is a hazard due to exposed and tangled **electrical cables**, which pose a risk of tripping, electrical shock, or fire. The suspended **yellow box** appears unstable, increasing the risk of falling and injury.✓

LLM-as-a-judge Output:

Hazard Detection: **1**
Quality Score: **4/5** (Miss Key Words)
Final Score: **0.875**

Reasoning: *The predicted description correctly identifies the existence of hazards, including tangled cables and the instability of the suspended box. It provides a clear explanation of risks such as tripping, electrical shock, and fire. However, it does not explicitly mention the straps or the column, which are key objects in the ground truth description. This omission prevents it from being fully comprehensive, resulting in a score of 4 for description quality*

Task: Open-Ended Description & Evaluation

Ground Truth: *The **safety netting** is improperly installed, with **visible gaps** that could lead to falls from height.*

Model Prediction: There is no visible construction safety hazard in this scene. The area appears secure with a protective barrier in place. [FAIL]

LLM-as-a-judge Output:

Hazard Detection: **0** (Missed)
Quality Score: **1/5** (Poor)
Final Score: **0.0**

Reasoning: *The predicted description **fails to detect the hazard** described in the ground truth (safety netting gaps). It incorrectly states that there is no hazard... making it both **inaccurate and unhelpful**.*

Figure 6: Qualitative Examples from SafeBuild-Bench. (Right) Two reasoning samples requiring complex chain-of-thought analysis of Temporary Electrical Supply and Working At Heights type hazard, Key risk factors are bolded.