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

Yi Cui[1,*] Zilin Wang[1,*] Yijie Xu[1] Qianyi Cai[1] Huizai Yao[1] Shuai Jiang[2] Bingzhuo Zhong[1,†] Hui Xiong[1,3,†]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]Northwestern Polytechnical University [3]CSE, HKUST

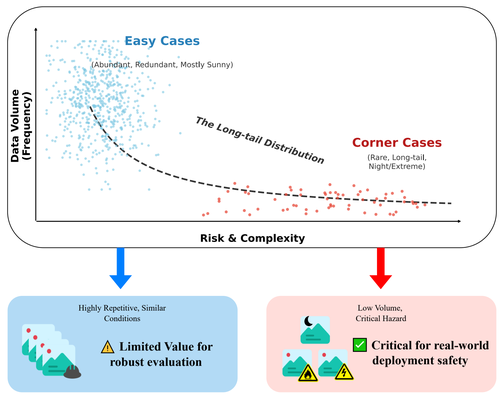
[*]Equal contribution [†]Corresponding authors

香港科技大學(廣州) THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

KDD 2026 August 9–13, 2026 Jeju, Korea

Motivation: Data-Rich, Information-Poor

- **Redundancy dominates.** Most frames repeat the same compliant scene; high-impact hazards form a sparse long tail.
- **Sites drift.** Weather, build progress and lighting shift the distribution month to month.
- **Aggregate scores mislead.** One number rewards models that exploit stable backgrounds over safety semantics.



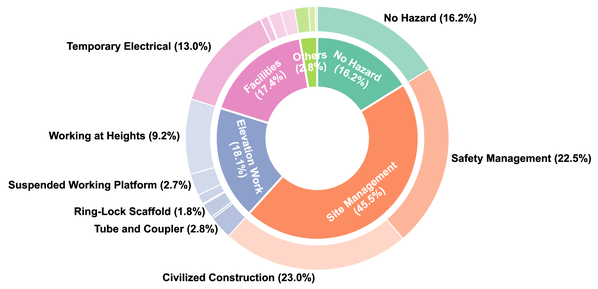
Volume lives in easy cases; risk lives in the tail.

A useful benchmark must keep **time and site metadata**, so scores can be sliced by deployment condition instead of collapsed into one.

Contributions

1. **SafeBuild-Bench** — expert-verified construction-safety benchmark mined from 100K+ image–text pairs across **50+ sites**, Jul–Nov 2025: **3,314** instances over 3,000+ images.
2. **Metadata-driven evaluation** — month- and site-stratified protocols, executable scripts, and an LLM-as-a-judge scheme audited against two humans.
3. **GEMS** — graph-enhanced selection that turns a redundant stream into a compact expert-review queue, validated under tight data budgets.

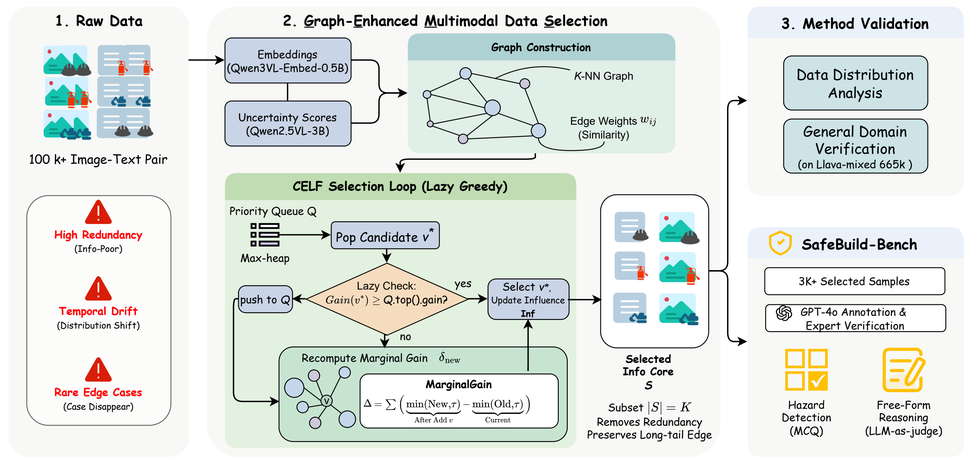
The Benchmark: Two Tasks, 19 Hazards



19 fine-grained hazard categories in 5 safety domains.

- **Identification (MCQ)** — 2,200 items. Distractors come from expert-defined *confusion groups*, so they are hazards genuinely mixed up on site. Accuracy + Macro-Recall.
- **Description (free-form)** — 1,114 items, scored for Hazard Detection Rate and Description Quality.
- Faces and plates anonymized; date and site ID kept on every instance.

Overall Framework: GEMS



Fused embeddings → k-NN graph → CELF lazy-greedy selection → expert-verified benchmark.

Mine, then verify. GEMS ranks candidates from a redundant stream so expert review is spent where it changes the benchmark.

① Multimodal Encoding

Qwen3-VL-Embedding maps each inspection record to one joint vector; FAISS builds a sparse k-NN graph over $N > 100$K nodes.

② Confusion Signal

A Qwen2.5-VL-3B proxy scores each record by length-normalized NLL under a fixed hazard prompt — high means the proxy doubts its own answer.

③ Submodular Selection

Influence spreads along edges weighted by uncertainty and saturates at τ. CELF lazy greedy gives $(1 - 1/e)$.

④ Expert Verification

Certified safety engineers assign every label and reference description; ambiguous cases go to group review.

NLL only *proposes* candidates — experts decide. Blur, rain and occlusion also raise NLL, so GEMS is a review queue, never an auto-labeler.

Why the Objective Scales

1. Confusion signal

$$u_i = -\frac{1}{L}\sum_t \log P_\phi(y_t \mid y_{<t}, x_i, p)$$

Length-normalized NLL from a frozen 3B proxy.

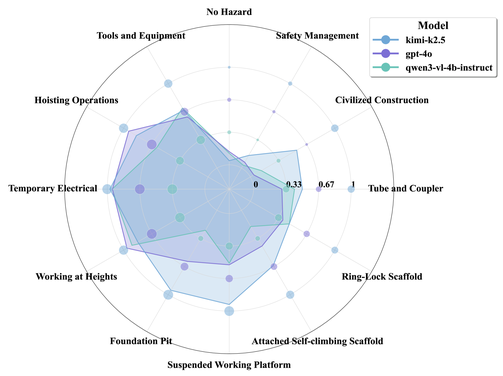
2. Saturated influence

$$F(S) = \sum_{j \in V} \min(\mathrm{Inf}_j(S),\ \tau)$$

Monotone + submodular $\Rightarrow (1 - 1/e)$ under greedy.

The cap τ is what buys diversity. Once a region of the manifold is covered, further samples from it stop paying — near-duplicates are dropped while long-tail hazards survive.

Where Models Fail: Detection, Not Phrasing



Category-wise MCQ accuracy, three representative models.

- **Easier:** categories with a distinctive object — Temporary Electrical, Hoisting Operations.
- **Harder:** categories defined by a rule rather than an object — Tube & Coupler, Ring-Lock Scaffold, Safety Management.
- Models **over-predict hazards** on safe scenes, so no-hazard calibration is weak.

Major description failures are **detection** failures: **303/304** (Kimi-K2.5) and **493/495** (Qwen3-VL-4B) coincide with $\mathrm{HDR} = 0$. Once a hazard is seen, the wording is usually fine.

Results: The Best MLLM Reaches 60.7

15 models • zero-shot • identical prompts • GPT-4o judge

Model	Acc	M-Rec	HDR	DQ	Overall
Kimi-K2.5	48.7	**67.7**	72.6	53.7	**60.7**
Gemini-3-Flash	**55.8**	65.8	67.6	48.4	59.4
Qwen3-VL-Plus	40.3	53.7	**79.9**	**62.5**	59.1
Claude-4.5-Sonnet	48.3	56.2	64.4	50.6	54.9
Qwen3-VL-235B	40.6	51.6	62.1	51.3	51.4
GPT-4o	34.3	43.8	66.5	50.7	48.8
GLM-4.6V	36.8	40.2	59.6	39.3	44.0
LLaVA-1.5-7B	25.3	29.8	38.2	27.2	30.1

Open-source **Kimi-K2.5** tops the board — scale, not access, drives performance. No model is close to reliable.

Data Efficiency: 1% Matches 100%

Method	Size	VizWiz	POPE	MMMU	Avg %
Full Data	100%	47.8	86.4	32.8	100.0
Random	8%	42.3	85.7	32.2	93.3
DataTailor	8%	46.3	82.1	33.9	96.9
GEMS	**1%**	**53.7**	**86.2**	34.8	**98.3**

LLaVA-v1.5-7B fine-tuned on LLaVA-Mixed-665K subsets.

+5.9 VizWiz over full data

101.1% of full data at 20% (150K)

Diversity, not just uncertainty.

At a matched 10% budget, dropping the graph term collapses coverage: **80** distinct regulations for uncertainty-only against **305** for full GEMS, at higher redundancy (NN cosine **0.926** vs **0.825**). Against 100 matched random draws, GEMS is 0.825 vs 0.887 ± 0.003.

Temporal Robustness: Months Swing 8–13 Points

MCQ Acc.	Jul	Aug	Sep	Oct	Nov
Gemini-3-Flash	52.2	60.1	**62.4**	60.6	55.7
Kimi-K2.5	45.6	51.3	**53.9**	50.0	49.8

Month ranges are typically **8–13 points** — Claude-4.5-Sonnet runs 46.4 (Jul) to 54.6 (Oct). This is **variation, not monotonic decay**; the released metadata is what makes it measurable.

Judge reliability (60 cases, 2 humans): hazard agreement **88.3–90.0%**; description quality within one point in **90.0–96.7%** — close to human–human agreement.

Takeaways

- **Construction safety is unsolved.** Best of 15 MLLMs scores 60.7; small open models fall to 30.
- **The gap is perceptual, not linguistic.** ∼99% of major description failures are missed detections.
- **Selection beats scale.** 1% of GEMS-mined data matches full-data robustness — mine, then verify.
- **Report stratified scores.** A missed hazard is a safety risk; a false alarm is why nobody trusts the system. They need separate numbers.

Benchmark, metadata schema, GEMS code, evaluation scripts and the temporal, judge, ambiguity and subset-composition audit packages are all public.

KDD 2026, Jeju Island, Korea **arXiv:2608.00068** · github.com/safebuild/gems · DOI 10.1145/3770855.3817581

Contact: ycui785@connect.hkust-gz.edu.cn