Privacy-Aware Synthetic Video Benchmarking and Relational Evaluation for Worker-Under-Suspended-Load Detection

Anshu Singh* Alejandro Seif*
Government Technology Agency Singapore
{singh_anshu, alejandro_seif}@tech.gov.sg

Abstract

Publicly shareable construction-video benchmarks remain scarce, especially for safety-critical hazards that are rare, dangerous to stage, and difficult to release. We study worker under suspended load*, a relational hazard that depends on worker–load geometry and temporal persistence rather than object detection alone. We introduce SynthSite, a focused synthetic video benchmark of 55 clips spanning varied load configurations, viewpoints, clutter, occlusions, and surveillance conditions, together with a privacy-aware hybrid generation workflow that supports both publicly shareable benchmark creation and privacy-constrained synthetic video generation.*

We then ask whether worker appearance can be suppressed without undermining downstream hazard recognition. Under five whole-body privacy conditions, we evaluate worker and load retention, localization stability, and clip-level hazard recognition. We find that structure-preserving obfuscations retain substantially more downstream utility than appearance-smoothing baselines, and that preserving a raw visual reference alone does not guarantee the strongest agreement with human hazard labels. These findings suggest that privacy evaluation for construction safety analytics should assess not only appearance suppression, but also preservation of the geometric cues required for hazard reasoning. Our dataset and code are available at https://huggingface.co/datasets/govtech/SynthSite

1. Introduction

Construction sites remain among the most hazardous work environments, with safety-critical incidents arising from heavy machinery, dynamic layouts, and changing worker–equipment interactions. Safety regulations therefore enforce exclusion zones around dangerous operations, including suspended loads and their fall zones [18]. Recent reports continue to document fatal injuries involving workers struck by falling or moving objects in construction and related sectors [17, 27]. These risks make automated safety monitoring an important application area for computer vision.

Many practically important construction hazards, however, are not object categories but *relations* that unfold over time. A worker may enter a machine danger zone, equipment may transition into an unsafe state, or a suspended load may become hazardous only relative to nearby workers. We study one such hazard: *worker under suspended load*, where risk arises when a worker is located within the fall zone beneath a suspended load.

This setting is challenging for two reasons. First, it is inherently *relational* and *spatiotemporal*: a system must localize workers and loads, infer whether a load is suspended, estimate an approximate fall zone, and aggregate evidence over time under clutter, occlusion, and unstable surveillance imagery. Prior work has shown that crane load fall-zone monitoring is feasible [7, 20], while also highlighting the difficulty of robust suspended-load detection under wide appearance variation [7]. Second, publicly shareable *video* benchmarks for this hazard are scarce. In our setting, available real construction CCTV did not provide enough positive suspended-load events or controlled diversity for benchmark construction, and operational footage is often difficult to release because it may expose workers, site layouts, equipment configurations, and other sensitive details.

In this setting, synthetic video is not just a data augmentation tool; it is a practical mechanism for constructing shareable benchmarks for rare and safety-critical events. Recent work on sensitive-data synthesis with foundation models further suggests that privacy-aware generation is shaped not only by model capability, but also by deployment, infrastructure, and model-access constraints [3, 22, 26].

In this work, we introduce SynthSite, a focused synthetic video benchmark for *worker under suspended load*.

*Both authors contributed equally to this work.

Accepted to the 3rd Workshop on Synthetic Data for Computer Vision (SynData4CV), CVPR 2026.

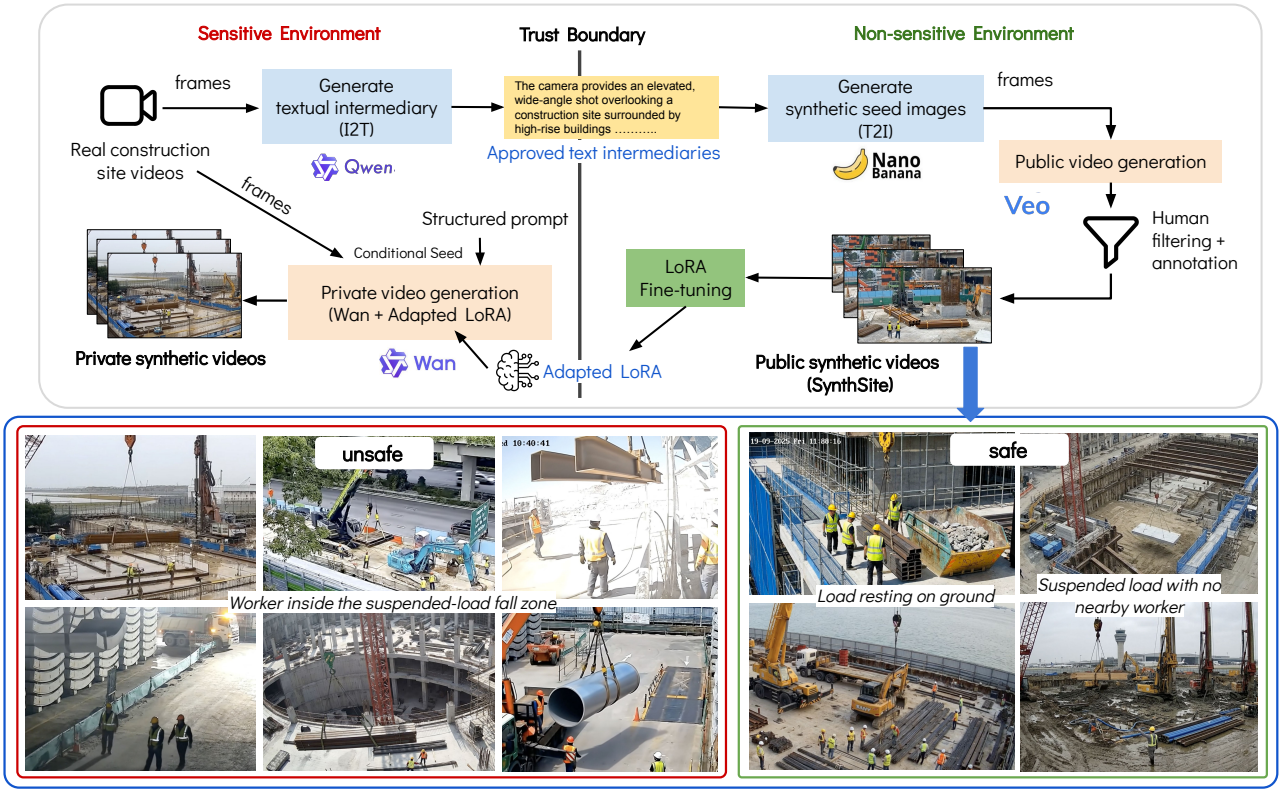

Figure 1. **Privacy-aware hybrid workflow and representative SynthSite clips.** Raw construction footage remains inside the sensitive environment, where a local vision–language model (Qwen2.5-VL) generates approved textual intermediaries. These intermediaries cross into the non-sensitive environment, where Nano Banana 2 produces synthetic seed images and Veo 3.1 Fast generates candidate publicly shareable synthetic videos. Curated and annotated public videos form SynthSite, while a subset is used to train a LoRA that is imported back into the sensitive environment and combined with Wan2.2-I2V-A14B, structured prompts, trigger words, and optional local reference frames to generate privacy-constrained internal synthetic videos. The example clips illustrate both unsafe cases, where workers occupy the suspended-load fall zone, and safe cases, where loads are grounded or workers remain outside the fall zone, across diverse machinery, viewpoints, clutter, occlusion, and surveillance conditions.

SynthSite contains 55 clips spanning varied suspended materials, viewpoints, clutter, occlusions, and surveillance conditions relevant to construction monitoring. To support benchmark creation, we develop a **privacy-aware hybrid generation workflow** that keeps raw construction footage inside a trusted environment while allowing only approved textual intermediaries to cross the trust boundary for downstream synthesis. This workflow supports both publicly shareable benchmark videos and privacy-constrained internal synthetic video generation.

Beyond benchmark construction, we study a practical question for privacy-conscious safety analytics: *how much worker appearance can be suppressed before relational hazard recognition degrades?* Semantically, worker identity is not required for this task, since hazard reasoning depends on worker–load geometry rather than personal identity. In practice, however, visual obfuscation can still degrade worker retention and geometric stability. We therefore define a **relational privacy evaluation** under whole-body worker obfuscation, measuring clip-level hazard recognition together with worker retention, load retention, and localization stability.

Our contributions are threefold:

- We introduce SynthSite, a focused synthetic construction video benchmark for *worker under suspended load*, a safety-critical relational hazard that is difficult to capture, rare in available real CCTV, and underrepresented in public construction-vision benchmarks.
- We propose a **privacy-aware hybrid generation workflow** that separates sensitive and non-sensitive synthesis stages using approved textual intermediaries, enabling publicly shareable benchmark creation and privacy-constrained internal synthetic video generation under realistic deployment constraints.
- We present a **relational privacy evaluation** for construction safety analytics, showing how whole-body worker

obfuscation affects downstream hazard reasoning, and showing that agreement with the raw pipeline and agreement with human hazard labels are not identical objectives.

2. Related Work

Construction safety benchmarks and suspended-load monitoring. Public construction-safety benchmarks remain dominated by object-centric image datasets such as MOCS [2] and SODA [9]. More recent video and simulation efforts move toward dynamic monitoring, but primarily target PPE compliance, vehicle motion, or general proximity events rather than suspended-load hazards [23, 29]. Prior work on crane load fall-zone monitoring shows that suspended-load reasoning is feasible but challenging under clutter, occlusion, and wide variation in load appearance [7, 20]. Relative to these lines of work, SynthSite focuses specifically on *worker under suspended load*, a relational video setting that depends on suspension-state inference, worker–load geometry, and temporal reasoning.

Privacy-aware synthesis workflows for sensitive video. Recent work has framed sensitive-data synthesis as a workflow problem shaped not only by model capability, but also by infrastructure, policy, and model-access constraints [3, 26]. Release-oriented medical-video synthesis further illustrates the need to control what leaves the sensitive environment [22]. Our work follows this line by keeping raw construction footage inside a trusted environment and exporting only approved textual intermediaries for downstream generation, enabling benchmark creation without directly releasing operational sensitive video.

Privacy-preserving video analytics. Visual anonymization is a privacy–utility trade-off: suppressing appearance can also distort the cues required for downstream understanding [8, 15]. Recent work in privacy-preserving video understanding likewise shows that anonymization can degrade spatiotemporal reasoning [5, 11]. We study this issue in a largely unexplored construction setting, asking whether whole-body worker obfuscation can hide identity while preserving the worker–load geometry needed for relational hazard detection.

3. SynthSite: Synthetic Worker-Under-Suspended-Load Video Dataset

3.1. Dataset Overview

We introduce SynthSite, a focused synthetic video dataset for the safety-critical construction hazard of *worker under suspended load*. The dataset contains 55 clips, with 28 *safe* and 27 *unsafe* videos. Each clip is 5–10 seconds long, generated at 15 FPS with a native resolution of 1280×720.

Table 1. Scene-level diversity of the 55 clips in SynthSite, estimated automatically from the first frame of each video using Gemini 2.5 Flash.

Attribute	Category	Count	Percent (%)
Proximity	Far	7	13
	Mid	40	73
	Close	8	14
Workers	1	8	15
	2–5	27	49
	6+	20	36
Machinery	Earthmoving	1	2
	Lifting	32	58
	Mixed	22	40
Lighting	Bright	28	51
	Overcast	26	47
	Dim/Night	1	2
Scene	Open site	45	82
	Scaffold	9	16
	Roadwork	1	2

SynthSite is designed to capture variation relevant to construction surveillance and relational hazard reasoning. It includes diverse heavy machinery and suspended materials, such as steel I-beams, precast tunnel segments, and rebar cages, together with practical CCTV-like challenges including illumination changes, wide viewpoints, clutter, scale variation, and partial occlusion (Figure 1). As summarized in Table 1, the clips vary in worker–load proximity, worker density, machinery context, lighting, and scene type. We position SynthSite as a focused benchmark for a narrowly defined relational hazard rather than as a large-scale surveillance corpus.

Because *worker under suspended load* is a relational safety label rather than a simple object category, benchmark construction required manual quality control and rubric-guided annotation (Sections 3.3 and 3.4). We therefore prioritize controlled curation over scale.

3.2. Privacy-Aware Hybrid Video Generation Workflow

Real construction surveillance footage is often operationally sensitive, while the highest-fidelity video generators are typically closed-source and externally hosted. This creates a practical tension: strong generation quality is desirable, but raw footage cannot be moved freely across trust boundaries. We therefore design a *privacy-aware hybrid video generation workflow* that supports both publicly shareable benchmark creation and internally controlled synthetic video generation (Figure 1).

We use the term *privacy-aware* in an operational rather than formal sense. Raw footage remains inside a trusted environment, and only approved textual intermediaries are allowed to cross the boundary. In our instantiation, these intermediaries are structured scene descriptions or reverse prompts that summarize hazard-relevant scene composition, worker–equipment relations, and environmental context (Appendix A.2). We do not claim a formal privacy guarantee; rather, the workflow is designed to reduce direct exposure of sensitive visual data while preserving sufficient structure for downstream synthesis [3, 22, 26].

In the configuration used in this paper, raw frames remain in the **sensitive environment**, where a locally deployed vision–language model converts them into approved textual intermediaries. These are transferred to the **non-sensitive environment**, where external image and video generators produce candidate publicly shareable synthetic videos. After human filtering and annotation, a curated subset of these videos forms SynthSite. The same shareable synthetic videos are also used to adapt a locally deployable open video model via LoRA [14], after which the adapted model is imported back into the sensitive environment for internally controlled generation.

For the benchmark-generation branch, each candidate clip is guided by two complementary conditioning signals: a synthetic seed image, which anchors scene layout and coarse geometry, and a structured hazard prompt, which specifies suspended objects, worker placement, viewpoint, and nuisance factors such as clutter, occlusion, and illumination. For the internal-generation branch, the adapted local model uses structured prompts together with optional local reference frames that remain inside the sensitive environment. Because *worker under suspended load* is a relational and spatiotemporal hazard, generation is treated as a guided proposal process followed by human filtering (Section 3.3) and annotation, rather than as a one-shot reliable synthesis mechanism.

Detailed implementation choices, including model allocation across environments, textual intermediary design, compute setup, and LoRA adaptation settings, are provided in Appendix A.

3.3. Quality Control

Before hazard annotation, all generated clips are manually screened for generation failures that would invalidate downstream labeling or relational hazard reasoning. Roughly half of the generated candidate clips were discarded during this stage. We discard clips containing severe synthesis failures, including hallucinated artifacts, duplication or fusion of key entities, abrupt appearance or disappearance of workers, loads, or machinery, temporally inconsistent depiction of workers or loads, physically implausible motion, or severe temporal corruption such as abrupt scene transitions.

We retain minor imperfections that do not prevent reliable hazard interpretation, including mild geometric distortion, low resolution, corrupted text, limited worker motion, and occasional dropped or inconsistent frames. This is intentional: the goal is not photorealistic perfection, but surveillance-style clips in which the worker–load relationship remains visually interpretable. Quality control is therefore applied as a task-oriented filtering stage, excluding clips that break spatial or temporal plausibility while retaining artifacts that do not invalidate the safe/unsafe label.

3.4. Annotations

Each clip is assigned a binary safety label (*safe*/*unsafe*) using a rubric-based event-occurrence rule and custom annotation platform, both described in Appendix B. A clip is labeled *unsafe* when a visibly suspended load is present and a worker remains within the fall zone beneath the load for at least 1 second (15 consecutive frames); otherwise, it is labeled *safe*. The rubric also defines common hard negatives, including clips where (i) no worker is present within the fall zone, (ii) the load is not physically suspended, or (iii) the scene creates a misleading impression of suspension due to perspective or occlusion. Clips with severe perspective ambiguity, where the worker–load relationship cannot be determined reliably, are excluded.

Annotations were collected from a pool of eight annotators briefed using a construction-safety-informed rubric. Each clip was independently reviewed by two annotators using the custom annotation platform, and disagreements were resolved by a third annotator. Cohen's κ was 0.308. Because the task requires fine-grained spatial reasoning under clutter, occlusion, and perspective distortion, we rely on adjudication and exclusion of unresolved ambiguous clips to improve label reliability.

4. Relational Privacy Evaluation

We study whether *worker under suspended load* can be recognized while suppressing worker-identifying appearance cues. Under our annotation rubric (Appendix B), a clip is labeled *unsafe* when a visibly suspended load is present and a worker remains within the fall zone beneath the load for at least 1 second (15 consecutive frames); otherwise, it is labeled *safe*. The task is therefore inherently *relational* and *spatiotemporal*: prediction depends on the worker–load relationship over time rather than on worker identity.

4.1. Whole-Body Obfuscation Conditions

We evaluate five conditions: a **raw-video reference** and four whole-body obfuscation settings: **Canny-edge body**, **cartooned body**, **pixelated body**, and **blurred body** (Figure 2). We use whole-body obfuscation rather than face-only redaction because worker faces are typically small,

Figure 2. **Relational privacy evaluation under worker obfuscation.** The same worker-under-suspended-load scene is shown under five evaluation conditions: a raw-video reference and four whole-body obfuscation settings: Canny-edge, cartooned, pixelated, and blurred worker representations.

oblique, or occluded in construction surveillance. In such settings, identity-related cues may still be conveyed through soft biometrics such as clothing, build, and gait [15, 32].

These transformations span practically relevant anonymization families with different privacy–utility trade-offs [8]. Canny-edge and cartooning apply stronger appearance abstraction while preserving coarse body structure, whereas pixelation and blur serve as standard visual anonymization baselines [10, 28]. All four are designed to preserve worker extent and position sufficiently for instance-level localization, which is essential because hazard recognition depends on worker–load geometry rather than photometric identity. To ensure a fair comparison, all obfuscations are applied only to worker regions obtained from a shared person-segmentation stage, with identical masks reused across methods; implementation details are provided in Appendix C.

Although the benchmark is synthetic, worker obfuscation remains relevant as an additional privacy layer. Synthetic visual data are not automatically privacy-safe: prior work has highlighted disclosure risks in synthetic-data release and memorization risks in high-capacity generative models [4, 24]. We therefore study whether appearance can be suppressed further while preserving the relational cues needed for hazard recognition.

4.2. Relational Hazard Detection Pipeline

We formulate *worker under suspended load* as a geometry-driven relational detection problem. Rather than attempting full 3D physical reconstruction from monocular surveillance video, we use a practical 2D proxy based on detections, motion cues, and temporal persistence. Given a video clip $\mathcal{V} = \{I_t\}_{t=1}^{T}$, the pipeline has three stages: (i) detect and track workers and hoisted loads, (ii) infer frame-level worker–load risk relations, and (iii) aggregate these into a clip-level safe/unsafe decision. Full implementation details are provided in Appendix D.

Detection and suspension inference. For each frame I_t, we detect workers and hoisted loads using YOLO-World [6] and associate detections over time with ByteTrack [31]. We use an open-vocabulary detector because task-specific lifted-load annotations are scarce, making supervised detector training outside the scope of this study. In a small pilot, YOLO-World was more suitable than Grounding DINO for this recall-sensitive surveillance setting in both usable lifted-load detections and inference latency; Appendix E provides prompt-design and detector-selection details.

For each worker, we define the worker footpoint as the bottom-center of the bounding box, denoted by $(x^{\text{foot}}_{\text{worker}}, y^{\text{foot}}_{\text{worker}})$. For each load, we use x_{load} for the horizontal center of the bounding box and $y^{\text{bottom}}_{\text{load}}$ for its bottom edge. Image coordinates increase downward in the vertical direction. To distinguish a suspended load from one resting on the ground, we use nearby workers as local ground proxies. For a load with width w_{load} and fixed horizontal expansion factor α, only workers whose footpoints satisfy

$$|x^{\text{foot}}_{\text{worker}} - x_{\text{load}}| < \alpha\, w_{\text{load}} \tag{1}$$

are considered. For each such worker–load pair, we compute a normalized clearance ratio

$$r^{\text{clr}} = \frac{y^{\text{foot}}_{\text{worker}} - y^{\text{bottom}}_{\text{load}}}{h_{\text{worker}}}, \tag{2}$$

where h_{worker} is the detected worker height in pixels. A load is treated as suspended when the maximum nearby-worker clearance exceeds a threshold. Normalizing by worker height improves robustness to viewpoint and scale variation. This suspension proxy is intentionally local and conservative, and is less reliable when no suitable nearby worker is available.

Fall-zone proxy and temporal decision. Once a load is inferred to be suspended, we estimate a trapezoidal fall-zone proxy beneath it. The region widens from the load

toward the ground to reflect increasing lateral displacement during a possible fall. Its extent is computed from an estimate of load height above ground, recent horizontal motion, and a minimum drift floor that accounts for residual crane swing. A worker is considered exposed when the worker footpoint lies inside this trapezoidal region, subject to a vertical compatibility gate.

For each frame, we define a binary hazard indicator

$$h_t = \mathbb{I}\begin{pmatrix} \exists\, i, j \text{ s.t. load } j \text{ is suspended} \\ \qquad\qquad \text{and worker } i \text{ lies in its fall-zone proxy} \end{pmatrix} \tag{3}$$

where $\mathbb{I}(\cdot)$ is the indicator function. We then aggregate h_t over a 1-second sliding window and label a clip as *unsafe* when sufficiently persistent hazardous evidence is observed within the window. To remain robust to transient detector dropouts, this temporal decision is a relaxed proxy rather than a strict enforcement of 15 consecutive positive frames; small negative gaps inside otherwise positive runs may be filled before counting. This proxy provides a lightweight basis for evaluating whether worker obfuscation preserves the relational cues needed for downstream hazard recognition.

4.3. Evaluation Setup and Metrics

We evaluate the 55 clips in SynthSite from two complementary perspectives: (*i*) *raw-reference stability*, i.e., whether worker obfuscation preserves the behavior of the relational hazard pipeline relative to the raw-video reference, and (*ii*) *agreement with human annotations*, i.e., whether clip-level hazard predictions remain aligned with the human safe/unsafe labels. The same relational hazard pipeline and decision thresholds are used for all privacy conditions; only the video input is changed.

Raw-reference stability. We measure **worker retention** and **load retention** by matching detections in each obfuscated video to detections in the corresponding raw video using an IoU-based criterion. These metrics quantify how well the two detection primitives required for downstream hazard reasoning are preserved under obfuscation. We also measure **worker localization jitter** and **load localization jitter** to quantify the geometric stability of matched detections:

$$\text{Jitter} = \frac{\|c_{\text{raw}} - c_{\text{priv}}\|_2}{d_{\text{raw}}}, \tag{4}$$

where c_{raw} and c_{priv} are the center coordinates of matched bounding boxes in the raw and obfuscated settings, respectively, and d_{raw} is the diagonal length of the raw bounding box. Finally, we report clip-level F_2 against the raw-video predictions, which measures hazard-decision consistency under privacy transformations. Raw-reference stability therefore reflects consistency with the raw proxy, not correctness with respect to the human annotations.

Agreement with human annotations. We separately evaluate clip-level predictions against the human safe/unsafe annotations, treating *unsafe* as the positive class. Because this is a recall-sensitive safety setting, we emphasize clip-level F_2 as the primary metric.

4.4. Results and Discussion

Table 2 summarizes raw-reference stability and agreement with human annotations, while Figure 3 shows representative examples. Overall, cartooning yields the strongest agreement with human annotations ($F_2 = 0.767$), whereas Canny-edge gives the highest raw-reference stability ($F_2 = 0.964$). This distinction is important: preserving the behavior of the raw pipeline is not identical to preserving the strongest human-aligned hazard decisions.

The primary degradation is reduced worker retention, not geometric drift. Once detections are retained, their geometry remains relatively stable across privacy settings. Worker localization jitter stays within 1.35%–2.17% of the raw bounding-box diagonal, and load localization jitter within 1.03%–2.31%. By contrast, worker retention varies much more strongly, from 88.9% for cartooning to 41.4% for blur. This indicates that the main effect of obfuscation is not large localization error, but reduced *worker detectability*. In other words, the relational hazard logic remains largely intact once a worker is still detected; the dominant failure mode is that some privacy transformations make the worker harder to retain as a stable instance.

Structure-preserving obfuscations best preserve relational hazard cues. Blur likely performs worst because it suppresses local contrast, articulation, and silhouette detail, reducing instance separability and stable worker retention. Cartooning likely performs best because it suppresses appearance texture while preserving coherent human structure, yielding the highest worker retention and the strongest agreement with human annotations. Pixelation is also informative: although its worker retention is substantially lower than cartooning, its human-annotation F_2 remains close. This suggests that clip-level unsafe classification depends on preserving sufficient relational evidence over time rather than perfectly retaining every worker instance. Overall, the results suggest that, for *worker under suspended load*, preserving worker geometry matters more than preserving photorealistic appearance.

Worker obfuscation perturbs the load channel, but less than the worker channel. Although the load itself is

Table 2. Effect of worker obfuscation on relational hazard detection in SynthSite. Columns under *Raw-reference stability* measure consistency relative to the raw-video reference pipeline; columns under *Agreement with human annotations* measure clip-level hazard recognition against human safe/unsafe labels. Higher is better for F_2, worker retention, and load retention; lower is better for worker and load localization jitter.

	Raw-reference stability					**Agreement with human annotations**
Method	$F_2 \uparrow$	**Worker Ret.** $\uparrow$	**Load Ret.** $\uparrow$	**Worker Jitter** $\downarrow$	**Load Jitter** $\downarrow$	$F_2 \uparrow$
RAW	Ref.	Ref.	Ref.	Ref.	Ref.	0.757
Canny-edge	**0.964**	80.2%	85.2%	2.17%	1.48%	0.742
Cartooning	0.963	**88.9%**	86.1%	**1.35%**	**1.03%**	**0.767**
Pixelation	0.913	63.1%	**86.8%**	1.82%	1.45%	0.762
Blur	0.876	41.4%	77.2%	1.44%	2.31%	0.671

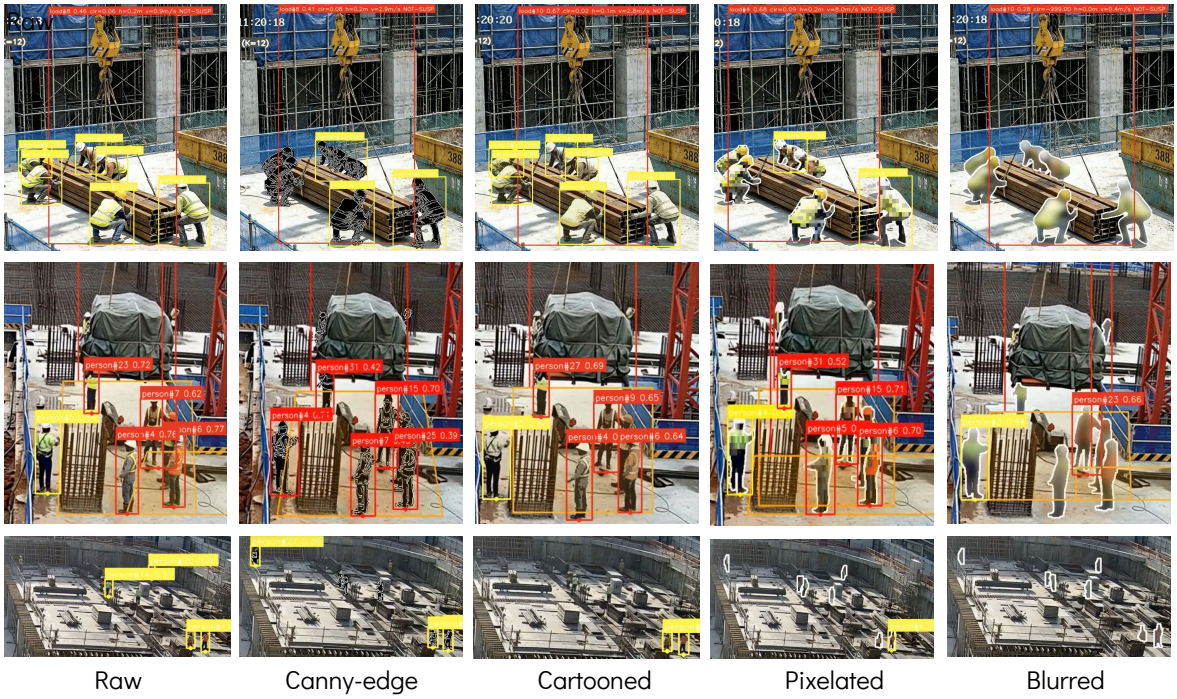

Figure 3. **Qualitative examples from relational privacy evaluation under whole-body worker obfuscation.** Each row shows the same scene under five conditions: raw, Canny-edge, cartooned, pixelated, and blurred worker representations. Orange overlays denote the estimated fall-zone proxy; red boxes indicate unsafe worker–load detections, and yellow boxes indicate safe workers. Consistent with the quantitative results, structure-preserving obfuscations, especially cartooning and Canny-edge, better preserve worker localization and downstream relational hazard reasoning than pixelation and blur. The middle rows illustrate cross-entity effects, where worker obfuscation perturbs load localization and fall-zone geometry through the shared detection pipeline. The bottom row shows a distant-view failure case, where small worker scale degrades relational hazard inference across conditions.

never obfuscated, load detection still varies across privacy settings: load retention ranges from 77.2% to 86.8%, and load localization jitter from 1.03% to 2.31%. One possible explanation is that the open-vocabulary detector is sensitive not only to the load region itself but also to surrounding context such as nearby workers, scale, and scene layout. More broadly, this is consistent with prior work showing that anonymizing human regions can affect downstream utility beyond the directly protected content, including object-detection performance on other scene elements [1]. However, the variation is noticeably smaller than for worker retention, reinforcing that the worker channel remains the primary point of sensitivity under privacy transformation. Thus, worker obfuscation perturbs both sides of the worker–load relation, but the dominant utility loss still arises from reduced worker detectability.

Human-grounded evaluation sharpens the interpretation of the benchmark. The results against human annotations confirm that the main privacy conclusion is not merely an artifact of agreement with the raw pipeline. In particular, cartooning slightly outperforms the raw condition against human labels ($F_2 = 0.767$ vs. 0.757), while Canny-edge, despite having the highest raw-reference stability, does not achieve the strongest agreement with human annotations. This indicates that higher agreement with the raw detector is not itself the target objective. What matters is whether the obfuscated representation preserves the cues needed for human-aligned hazard recognition. At the same time, the absolute performance of the raw proxy remains modest, and Appendix F.1 shows that it is recall-oriented but precision-limited. We therefore interpret the current system as a lightweight relational proxy for privacy evaluation rather than as a production-ready suspended-load detector.

4.5. Limitations and Future Work

Privacy evaluation. Our study evaluates *utility under privacy-motivated obfuscation*, but does not directly measure *privacy risk*, such as re-identification or attribute inference. We also consider only one parameterization per obfuscation family. Future work should therefore quantify privacy risk explicitly and characterize the privacy–utility frontier across obfuscation strengths [8, 15, 16].

Suspended-load reasoning. The hazard detector is a lightweight *relational proxy*, not a calibrated physical risk estimator. It operates on monocular 2D detections and remains sensitive to occlusion, small workers, and perspective distortion. In addition, the current system relies on open-vocabulary detection rather than task-specific lifted-load supervision, so performance depends on prompt design and is constrained by the scarcity of labeled suspended-load data. This is reflected in the absolute raw-video performance of the proxy, which is recall-oriented but precision-limited rather than production-ready. Representative failure cases are shown in Appendix F.2. Future work should combine dedicated airborne-load annotation with stronger supervision and richer geometric reasoning, including camera calibration, homography, and depth-aware worker–load estimation [30].

Real-world validation and benchmark scale. We do not report evaluation on real construction CCTV. This is a central motivation for SynthSite: the real footage available to us did not provide enough positive suspended-load events or controlled diversity for benchmark construction. We therefore use synthetic data to construct a targeted benchmark for this hazard, rather than to claim that the benchmark fully captures the distribution of operational surveillance footage. The present study should accordingly be interpreted as a controlled synthetic benchmark study rather than as a claim of deployment readiness. Although the observed trends are encouraging within the 55 clips of SynthSite, larger-scale validation remains necessary. Future work should expand the benchmark and evaluate on operational CCTV footage with representative suspended-load events.

Beyond worker under suspended load. An important next step is to test whether the same privacy–utility pattern holds beyond the current hazard. This includes other *relational* events, such as worker–machinery proximity and geofencing violations, as well as finer-grained worker-centric tasks such as PPE monitoring. In particular, cartooning is promising because it suppresses identity-bearing texture while preserving coarse silhouette and limb structure, but it remains unclear whether this is sufficient for head- and torso-level PPE localization, where the cues are smaller and more appearance-dependent.

5. Conclusion

In this work, we introduced SynthSite, a focused synthetic video benchmark of 55 clips for the safety-critical hazard of *worker under suspended load*, together with a privacy-aware hybrid workflow for constructing shareable benchmark data under trust-boundary and deployment constraints. This provides a practical route to benchmarking rare hazards that are difficult, unsafe, or infeasible to capture and release from real construction surveillance footage.

Our central finding is that downstream utility in this setting depends less on preserving worker appearance than on preserving the geometric cues required for relational hazard reasoning. Structure-preserving obfuscations retain substantially more hazard-recognition utility than appearance-smoothing baselines, and agreement with the raw reference is not identical to agreement with human hazard annotations. We also find that worker obfuscation can perturb the load channel even when the load itself is not transformed, suggesting that relational hazard detection can be sensitive to scene context beyond the directly obfuscated region.

Taken together, these results motivate geometry-aware and human-grounded evaluation of privacy-preserving video analytics for construction safety. More broadly, SynthSite is best viewed as a focused starting point rather than a finished benchmark. Expanding benchmark scale, validating on operational CCTV, and extending the framework to other relational construction hazards remain important next steps. We hope this work helps establish privacy-aware benchmarking as a practical foundation for safety-critical construction video analytics.

Acknowledgements

We thank Prof. Roy Ka-Wei Lee at the Singapore University of Technology and Design (SUTD) for his valuable feedback and guidance in improving this work.

References

[1] Sara Abdulaziz, Giacomo D'Amicantonio, and Egor Bondarev. Evaluation of human visual privacy protection: A three-dimensional framework and benchmark dataset, 2025.

[2] Xuehui An, Li Zhou, Zuguang Liu, Chengzhi Wang, Pengfei Li, and Zhiwei Li. Dataset and benchmark for detecting moving objects in construction sites. *Automation in Construction*, 122:103482, 2021.

[3] Eduardo Brito, Mahmoud Shoush, Kristian Tamm, Paula Etti, and Liina Kamm. SynthGuard: Redefining synthetic data generation with a scalable and privacy-preserving workflow framework. In *Availability, Reliability and Security – ARES 2025*, pages 193–211. Springer, 2025.

[4] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[5] Lixin Chen, Chaomeng Chen, et al. StegaVAR: Privacy-preserving video action recognition via steganographic domain analysis. *arXiv preprint arXiv:2512.12586*, 2025.

[6] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16901–16911, 2024.

[7] Eugene Yan Tao Chian, Yang Miang Goh, Jing Tian, and Brian H.W. Guo. Dynamic identification of crane load fall zone: A computer vision approach. *Safety Science*, 156: 105904, 2022.

[8] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, Steven Bohez, and Pieter Simoens. Privacy-preserving visual analysis: Training video obfuscation models without sensitive labels. *Applied Intelligence*, 54:6041–6052, 2024.

[9] Rui Duan, Hui Deng, Mao Tian, Yichuan Deng, and Jiarui Lin. SODA: A large-scale open site object detection dataset for deep learning in construction. *Automation in Construction*, 142:104499, 2022.

[10] Ádám Erdélyi, Thomas Winkler, and Bernhard Rinner. Privacy protection vs. utility in visual data: An objective evaluation framework. *Multimedia Tools and Applications*, 77(2): 2285–2313, 2018.

[11] Joseph Fioresi, Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Latent anonymization for privacy-preserving video understanding. *arXiv preprint arXiv:2511.08666*, 2025.

[12] Google. Nano banana 2 generative API. Proprietary Generative AI API, 2026.

[13] Google DeepMind. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

[14] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[15] Håkon Hukkelås and Frank Lindseth. DeepPrivacy2: Towards realistic full-body anonymization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1329–1338, 2023.

[16] Bayrem Kaabachi, Jérémie Despraz, Thierry Meurers, Karen Otte, Mehmed Halilovic, Bogdan Kulynych, Fabian Prasser, and Jean Louis Raisaro. A scoping review of privacy and utility metrics in medical synthetic data. *npj Digital Medicine*, 8(1):60, 2025.

[17] Ministry of Manpower, Singapore. Workplace safety & health report, 2024. https://www.mom.gov.sg/-/media/mom/documents/safety-health/reports-stats/wsh-national-statistics/wsh-national-stats-2024.pdf, 2025. Accessed: 2026-03-11.

[18] Occupational Safety and Health Administration. 1926.1425 – keeping clear of the load. https://www.osha.gov/laws-regs/regulations/standardnumber/1926/1926.1425, 2025. Accessed: 2026-03-11.

[19] Ostris. AI-Toolkit: The ultimate training toolkit for fine-tuning diffusion models. https://github.com/ostris/ai-toolkit, 2024. Accessed: 2026-03-12.

[20] Parham Pazari, Nasim Didehvar, and Amin Alvanchi. Enhancing tower crane safety: A computer vision and deep learning approach. *Engineering Proceedings*, 53(1):38, 2023.

[21] Qwen Team. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[22] Hadrien Reynaud, Qingjie Meng, Mischa Dombrowski, Arijit Ghosh, Thomas Day, Alberto Gomez, Paul Leeson, and Bernhard Kainz. EchoNet-Synthetic: Privacy-preserving video generation for safe medical data sharing. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pages 285–295. Springer, 2024.

[23] Meenakshi N. Shrigandhi and Sachin R. Gengaje. CSOD-24: Construction site object detection dataset for safety monitoring at construction site using deep learning. *Journal of Innovative Image Processing*, 7(1):182–206, 2025.

[24] Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data–anonymisation groundhog day. In *31st USENIX Security Symposium (USENIX Security 22)*, pages 1451–1468, 2022.

[25] Wan Team. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[26] Haoxiang Wang, Zinan Lin, Da Yu, and Huishuai Zhang. Synthesize privacy-preserving high-resolution images via private textual intermediaries. In *Advances in Neural Information Processing Systems*, 2025.

[27] Workplace Safety and Health Council, Singapore. Struck-by-falling-objects workplace fatal injuries in 2h2024. https://www.tal.sg/wshc/resources/newsletters/wsh-advisory/struck-

by - falling - objects - workplace - fatal - injuries-in-2h2024, 2025. Accessed: 2026-03-11.

[28] Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. A privacy-preserving action recognition system using action-specific edge detection. In *IEEE International Conference on Image Processing (ICIP)*, 2018.

[29] Zhengyu Wu, Yuxiang Feng, Yiannis Demiris, and Panagiotis Angeloudis. Development of a digital twin-based simulation system and a novel synthetic video dataset for enhancing computer vision in construction site safety. In *Proceedings of the European Conference on Computing in Construction*, 2024.

[30] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10371–10381, 2024.

[31] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision (ECCV)*, pages 1–21. Springer, 2022.

[32] Liang Zheng, Yi Yang, and Alexander G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.