Beyond Zero-Shot: Industrial Spill Detection via Synthetic Data and PEFT

Anonymous ICCV submission

Paper ID *****

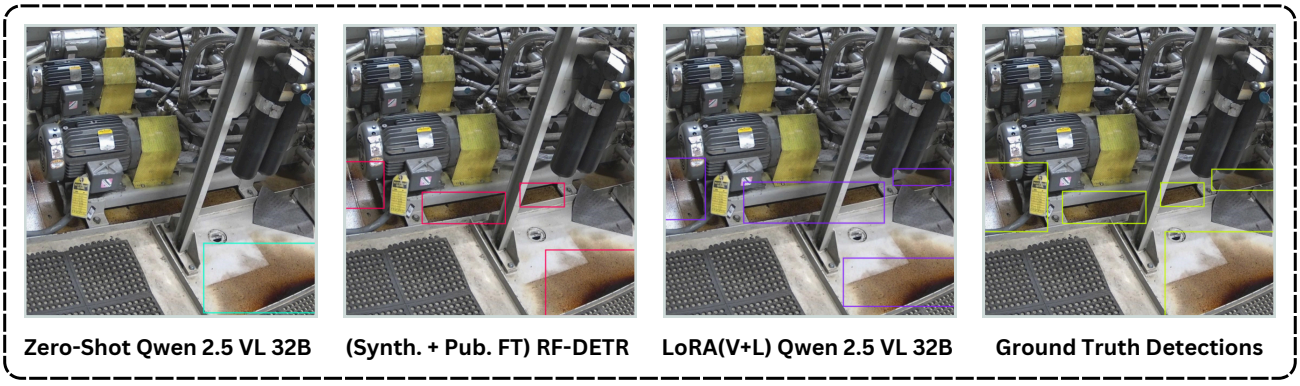


Figure 1. **Performance overview of competing methods:** Visual comparison of detection outputs, showcasing the performance of our proposed model against baseline methods on an authentic CCTV image of an industrial spill.

Abstract

Large-scale Vision-Language Models (VLMs) have transformed general-purpose visual recognition with strong zero-shot capabilities. However, their performance drops sharply in niche, safety-critical domains such as industrial spill detection, where hazardous events are rare, sensitive, and difficult to annotate. This scarcity—driven by privacy constraints, data sensitivity, and the rarity of real incidents—makes conventional fine-tuning of detectors infeasible for most industrial applications.

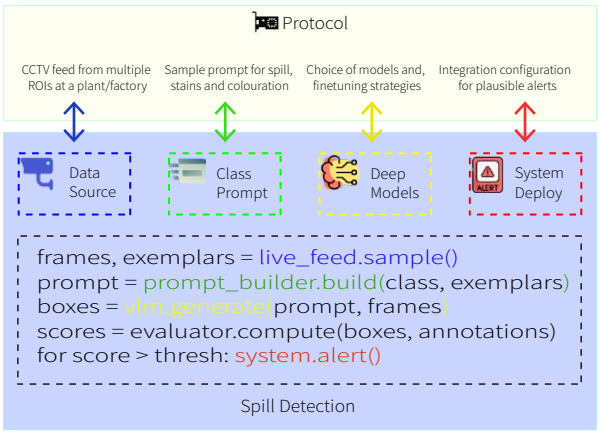
We address this challenge by introducing a scalable framework centered around a high-quality synthetic data generation pipeline. Our first contribution is a domain-specific, photorealistic image synthesis process based on guided Stable Diffusion, IP adapters, and anomaly-focused inpainting. This enables precise control over diverse spill types and visual conditions. Our second contribution is the ***parameter-efficient adaptation*** *of foundation VLMs using* ***Low-Rank Adaptation (LoRA)****, allowing injection of domain knowledge with minimal updates to the pretrained model.*

We show that this synthetic corpus not only enables effective Parameter-Efficient Fine-Tuning (PEFT) of VLMs, but also significantly enhances the performance of state-of-the-art object detectors such as YOLO and DETR. Notably, in settings where synthetic data is unavailable, VLMs still exhibit stronger generalization to unseen spill scenarios compared to these detectors. When synthetic data is available, both VLMs and detectors achieve substantial improvements, with performance becoming comparable.

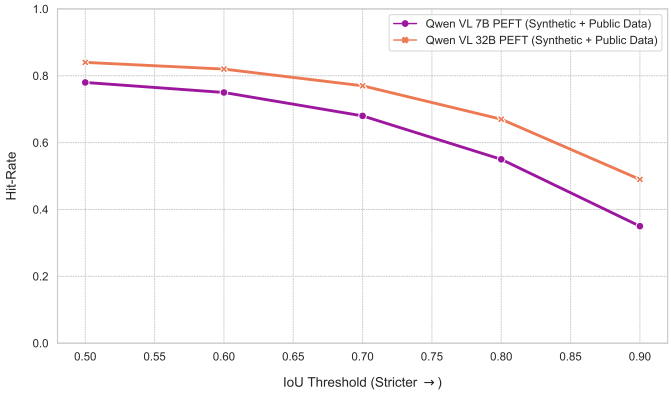
Our results highlight that high-fidelity synthetic data is a powerful tool for bridging the domain gap in safety-critical applications. The combination of synthetic generation and lightweight adaptation provides a cost-effective, scalable solution for deploying vision systems in industrial environments where real data is scarce or impractical to obtain.

1. Introduction

Continuous vigilance is indispensable in industrial operations, where undetected hazards—such as fluid leaks, chemical spills, or mechanical faults—can escalate rapidly, causing economic damage, environmental harm, and risks to human safety [14]. Historically, industrial monitoring has relied on manual inspections or fixed physical sensors. However, both approaches have serious limitations: human inspections are error-prone and infeasible at scale, while static sensors provide limited spatial awareness—they may indi-



(a) **An overview of the Industrial Spill Detection Framework.** The system integrates live CCTV feeds and a user-defined text prompt as inputs. A Vision-Language Model (VLM), selected and fine-tuned according to a chosen strategy, processes this information to detect and localize potential spills, triggering an alert if a detection exceeds a confidence threshold.



(b) Localization accuracy trends across IoU thresholds. Models adapted on synthetic data with LoRA-V or LoRA-(V+L) retain higher spatial precision under stricter overlap criteria, validating the quality of synthetic supervision.

Figure 2. From data creation to spatial precision: **(a)** Our synthetic generation pipeline; **(b)** Resulting improvements in fine-grained localization for adapted models.

cate *that* a hazard exists, but not *where* or *what* it is.

Computer vision has emerged as a promising solution, offering automated, context-aware surveillance. Modern object detectors such as YOLOv11 [**?**], TOLO [**?**], and RF-DETR [**?**] push the frontier in detection speed and accuracy. These models perform remarkably well in structured environments, but their real-world deployment in industrial settings remains bottlenecked by data scarcity. Industrial incident data is rare, often proprietary, and difficult to annotate, making it infeasible to collect large-scale, diverse datasets. Consequently, even state-of-the-art detectors tend to overfit to the synthetic or clean domains they are trained on, struggling to generalize across variable lighting, occlusions, or facility layouts [14].

To address these limitations, research has explored multiple paths. Classical anomaly detection methods—such as Gaussian pyramid differencing [1]—offer fast but semantically blind solutions. More recently, Vision-Language Models (VLMs) [5, 10] have shown promise for zero-shot recognition in open-world settings. While pretrained VLMs excel at broad generalization, their localization ability and understanding of domain-specific visual cues remain limited without task adaptation.

In this work, we present a framework that addresses the core data bottleneck through generative AI and Parameter-Efficient Fine-Tuning (PEFT). Instead of relying on scarce real-world data, we develop a high-fidelity synthetic data generation pipeline that combines Stable Diffusion XL, IP adapters, along with inpainting to simulate a wide range of realistic spill scenarios.

This synthetic dataset is then used to adapt VLMs via PEFT strategies like Low-Rank Adaptation (LoRA) [4] enabling us to incorporate domain expertise by updating only a small fraction of model weights.

Importantly, we show that this synthetic corpus benefits *both* vision-language and object detection models. Our experiments reveal that fine-tuning RF-DETR and YOLOv11 on the synthetic+web corpus substantially improves their real-world detection accuracy. However, in low-data regimes where no synthetic data is available, VLMs exhibit stronger generalization than these task-specific detectors. Once synthetic data is introduced, both approaches become competitive, underscoring the broad utility of our pipeline.

Our contributions include:

1. We introduce a novel, controllable pipeline for generating diverse and photorealistic industrial spill imagery using guided Stable Diffusion with IP adapters and anomaly inpainting.
2. We apply and evaluate multiple Parameter-Efficient Fine-Tuning (PEFT) strategies—including LoRA, QLoRA, and In-Context Learning—to adapt VLMs to the industrial domain.
3. We demonstrate that our synthetic dataset improves performance across both VLMs and state-of-the-art object detectors (YOLOv11, RF-DETR), highlighting its broad applicability and critical importance in safety-critical, data-scarce domains.

2. Related Work

The pursuit of automated industrial hazard detection has evolved across three interlinked research frontiers: (i) the progression from supervised detectors to vision-language foundation models, (ii) the use of synthetic data to overcome annotation bottlenecks, and (iii) the emergence of parameter-efficient fine-tuning (PEFT) techniques for scalable domain adaptation. Our work situates itself at the intersection of these areas.

From Supervised Detectors to Foundation Models. Traditional industrial visual monitoring systems began with classical techniques like background subtraction and image differencing [11], which were computationally lightweight but highly sensitive to environmental changes. The advent of deep learning introduced powerful supervised object detectors, including two-stage approaches like Faster R-CNN [13] and one-stage detectors like YOLOv3 [12] and DETR [**?**]. More recent successors—such as YOLOv7–v12 [**?** **?**], TOLO [**?**], and RF-DETR [**?**]—have pushed state-of-the-art performance in real-time detection and long-tail robustness. Yet, these models depend heavily on large-scale annotated datasets, which are typically unavailable in sensitive or safety-critical environments.

To address generalization under data scarcity, recent efforts have turned toward **Vision-Language Models (VLMs)** such as CLIP [10], ALIGN [5], and Florence [**?**], which are pre-trained on web-scale image-text pairs. These models exhibit strong zero-shot transfer for image-level classification and visual grounding. Building upon this, open-vocabulary detectors like GLIP [8], GroundingDINO [9], and Grounded-SAM [7] combine detection and segmentation under weak supervision. However, these systems still degrade in industrial environments due to severe domain shifts—uncommon textures, lighting conditions, and anomaly types that are underrepresented in pretraining corpora.

Synthetic Data and the Sim-to-Real Gap. Synthetic data has emerged as a practical solution to the annotation bottleneck. Earlier approaches used game engines or 3D simulators (e.g., CARLA [**?**], AI2THOR [**?**]) to render synthetic environments for training. However, limited realism in such renderings often introduced a ”sim-to-real” gap [**?**]. Diffusion models [**?**] have revolutionized this space, enabling high-fidelity generation of semantically aligned imagery. Text-to-image models like DALLE-2 [**?**] and Stable Diffusion [**?**] now support conditioning on detailed prompts or images via ControlNet [**?**], IP-Adapters [**?**], or DreamBooth [**?**], improving visual grounding in generated outputs. Several works have shown the efficacy of synthetic imagery for detection and segmentation: GenAug [**?**], Domain Randomization [**?**], Task2Sim [**?**], and DreamFusion [**?**]. Others, such as StyleGAN-based simulators [**?** **?**], have been used for compositional domain transfer. Yet, most existing works focus on **training small models** or **augmenting real datasets**, rather than fully adapting **large foundation models** using synthetic data alone in domains with zero or near-zero real samples. Applications in industrial anomaly detection remain underexplored despite recent interest [**?**].

Efficient Domain Adaptation via PEFT. Full fine-tuning of foundation models is computationally expensive and risks catastrophic forgetting [**?**]. To address this, Parameter-Efficient Fine-Tuning (PEFT) has emerged as a compelling alternative. LoRA [4], QLoRA [3], AdapterFusion [**?**], and BitFit [**?**] allow for tuning less than 1% of a model's weights while retaining strong downstream performance. While widely adopted in NLP, vision applications of PEFT remain nascent. Recent works like VPT [6], SSF [**?**], and AdaptFormer [**?**] apply PEFT to vision transformers, but few tackle VLMs in zero-shot detection regimes or study PEFT's behavior under extreme data scarcity.

Our work advances the field by synthesizing these three threads. We present a unified framework that: (1) leverages high-fidelity diffusion-based generation to create an industrial spill dataset without real incident data, (2) adapts VLMs using parameter-efficient strategies like LoRA, and (3) shows that the resulting models are not only competitive with fully fine-tuned detectors but also generalize better in low-data or zero-shot regimes. To our knowledge, this is the first systematic study of adapting foundation vision-language models for industrial hazard detection using purely synthetic data, bridging the gap between generalist pretraining and specialized safety-critical deployment.

3. Experimental Setup

Designing a study that speaks simultaneously to plant-floor practitioners and AI specialists requires a careful balance: the system must be reproducible with modest resources, yet documented with sufficient technical detail to withstand scrutiny. Our experiments assess how vision-language models (VLMs) can be adapted for industrial spill detection using synthetic and publicly available data.

3.1. Vision{Language Models

Vision-language models (VLMs) align image and text modalities through joint pretraining on large-scale image–caption datasets [5, 10]. At inference time, these models take an image and a text prompt (e.g., “Where is the spill?”) and return either textual answers or structured outputs.

We adopt the open-source **Qwen2.5-VL** family in three parameter scales: 3B, 7B, and 32B. These models integrate a Swin-style vision transformer [**?**] with a causal language decoder, forming a shared vision-language latent space. The 3B model is suitable for single-GPU setups,

while the 32B variant offers higher capacity and grounding fidelity.

3.2. Structured Prompting

Each image is paired with a structured prompt designed to elicit precise detection behavior in high-stakes environments. The system prompt simulates an industrial inspector's reasoning process; the user prompt requests bounding-box outputs in COCO JSON format for one of eight anomaly classes (e.g., `oil-spill`, `rust`, `fluid-stain`).

> **System:** *You are a certified industrial safety inspector specializing in hazardous spill, leak, and stain detection across factories and energy plants. Only report verifiable safety hazards. Do not guess or speculate.*
>
> **User:** *Detect and return the bounding-box coordinates of the <class> in COCO JSON format, if present.*

At decoding, we set temperature $\tau = 0.10$, nucleus sampling $p = 0.001$, and a repetition penalty of 1.2. These values constrain output verbosity and promote single-line COCO responses.

3.3. Parameter-Efficient Fine-Tuning (PEFT)

We compare multiple adaptation strategies:

Zero-Shot Inference. The model is evaluated as-is, without any additional tuning or examples. This baseline assesses the native generalization capability of pretrained VLMs on spill detection.

In-Context Learning (ICL). We prepend $k \in \{5, 10, 15\}$ labeled examples (image + bounding box) as context [2**?**]. No model weights are updated. Formally, the prediction is conditioned on a support set $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^{k}$, modifying the conditional distribution:

$$\mathbb{P}(y \mid \mathbf{x}, S). \tag{1}$$

This simulates scenarios where a few solved examples are available at deployment time.

Low-Rank Adaptation (LoRA). We also evaluate LoRA [4], which injects domain knowledge into the pretrained model by updating only low-rank matrix adapters. Instead of fine-tuning the full weight matrix $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, LoRA learns two small matrices A and B:

$$W' = W + \alpha AB, \quad A \in \mathbb{R}^{d_{\text{out}} \times r}, \quad B \in \mathbb{R}^{r \times d_{\text{in}}}. \tag{2}$$

Here $\alpha = \frac{1}{r}$ We test three configurations:

- **LoRA-L:** Only the language pathway is adapted.
- **LoRA-V:** Only the vision encoder is adapted.
- **LoRA-(V+L):** Both pathways are adapted jointly.

3.4. Evaluation Metric

We report **mean hit rate** at IoU threshold 0.5 across all anomaly classes. A prediction is a *hit* if the predicted bounding box has sufficient overlap with ground truth:

$$\text{IoU} = \frac{|B_{\text{pred}} \cap B_{\text{gt}}|}{|B_{\text{pred}} \cup B_{\text{gt}}|}, \quad \text{Hit if IoU} \geq 0.5. \tag{3}$$

3.5. Results Overview

Tables 1 and 2 summarize the mean hit rates for different adaptation strategies and model sizes on the public and Siemens Energy (SE) proprietary test sets.

Key observations:

- **Model scaling improves performance consistently**, with Qwen2.5-32B outperforming smaller variants across all adaptation types.
- **ICL offers diminishing returns beyond 10 shots**, suggesting the model saturates quickly under prompt conditioning.
- **LoRA-(V+L)** outperforms all other configurations, showing the synergy of joint adaptation.
- **LoRA-V (vision-only tuning)** performs notably better than LoRA-L across all sizes, indicating the importance of adapting visual representations in this domain.
- **Qwen2.5-3B with LoRA-V achieves 0.42 on public data**, outperforming 7B zero-shot and approaching 7B ICL performance — a strong result for resource-limited settings.

Table 1. Mean hit-rate @ IoU = 0.5 on Public Evaluation Set.

Method	3B	7B	32B
Zero-Shot	0.25	0.35	0.42
ICL (5 shots)	0.38	0.51	0.59
ICL (10 shots)	0.39	0.53	0.62
ICL (15 shots)	0.37	0.52	0.63
LoRA (L)	0.36	0.52	0.58
LoRA (V)	0.42	0.58	0.65
LoRA (V+L)	**0.46**	**0.63**	**0.71**

4. Preliminary VLM Results & Analysis

Before introducing our full methodology, we establish baseline performance using existing Vision-Language Models (VLMs) adapted with a small number of real-world examples. This early-phase study examines the practical limits of few-shot and parameter-efficient tuning techniques in highly data-constrained environments. Table 1 summarizes performance on a public test set, while Table 2 reflects transfer to an in-house industrial dataset.

Table 2. Mean hit-rate @ IoU = 0.5 on Siemens Energy Internal Test Set.

Method	3B	7B	32B
Zero-Shot	0.11	0.15	0.24
ICL (5 shots)	0.21	0.26	0.33
ICL (10 shots)	0.24	0.29	0.34
ICL (15 shots)	0.23	0.28	0.36
LoRA (L)	0.19	0.23	0.31
LoRA (V)	0.26	0.31	0.41
LoRA (V+L)	**0.29**	**0.34**	**0.49**

4.1. Initial Observations

Scale provides potential, but adaptation unlocks it. Larger models predict better—but only if guided. Qwen2.5-VL-7B outperforms 3B in every configuration, yet raw scale alone is insufficient: a 3B model adapted with LoRA-(V+L) rivals or surpasses the 7B model under zero-shot or poorly tuned settings. This affirms that *capacity without domain adaptation underutilizes the foundation model's potential*.

Few-shot prompting helps—up to a point. In-Context Learning (ICL) with just five labeled examples lifts the 7B model's hit-rate by over 45% from its zero-shot baseline. However, this benefit quickly saturates: increasing the context to 10 or 15 shots yields only marginal improvements, and sometimes regressions. These diminishing returns point to ICL's brittle reliance on prompt composition and token window limits [**?**], making it poorly suited for scalable, long-term deployment in high-stakes domains.

Visual adaptation is key. LoRA-V (vision-only fine-tuning) consistently outperforms LoRA-L (language-only) across model sizes. This suggests the primary challenge is not textual ambiguity ("what is a spill?") but visual specificity—learning the color gradients, surface reflections, and irregular shapes characteristic of real-world anomalies. Joint tuning with LoRA-(V+L) offers modest additional gains, but vision pathway adaptation carries the most impact per parameter.

Better detection leads to better localization. Improvements in hit-rate also correspond to tighter bounding boxes. Table **??** (see Appendix) shows that adapted models maintain higher performance even at stricter IoU thresholds, suggesting that recognition and localization improve hand-in-hand when adaptation is effective.

4.2. Quantifying Gains and Exposing Gaps

To formalize adaptation benefit, we define the *relative uplift* of method m over the zero-shot baseline:

$$\Delta_m = \left(\frac{\mathrm{HR}_m - \mathrm{HR}_{\mathrm{ZS}}}{\mathrm{HR}_{\mathrm{ZS}}} \right) \times 100\%.$$

For Qwen2.5-VL-7B on the public test set, LoRA-(V+L) yields a relative uplift of over 80%. On paper, this is a remarkable gain with minimal training overhead.

However, the absolute hit-rate remains limited. A score of 0.63 may be adequate for exploratory detection tasks, but falls short for high-stakes, safety-critical deployment. More alarmingly, the same method on our Siemens Energy test set—despite using the same adaptation corpus—drops to 0.34. This underlines a core limitation: **current adaptation methods—even parameter-efficient ones—are fundamentally capped by the available data.**

4.3. The Limits of Real-World Exemplars

While these findings validate the generalization capabilities of VLMs and the promise of PEFT strategies, they also expose a hard ceiling. Despite careful prompt design, architecture tuning, and adaptation, all performance gains converge toward a plateau defined by the quantity and diversity of real-world examples.

The implication is clear: no amount of smart prompting or low-rank adaptation can replace the breadth of a rich training distribution. In our experiments, the model tends to detect anomalies that visually resemble examples it has seen during adaptation. It generalizes locally—not structurally. Any deviation in color, material, lighting, or spill geometry degrades performance. This is unacceptable in safety-critical contexts.

4.4. The Case for Synthetic Data

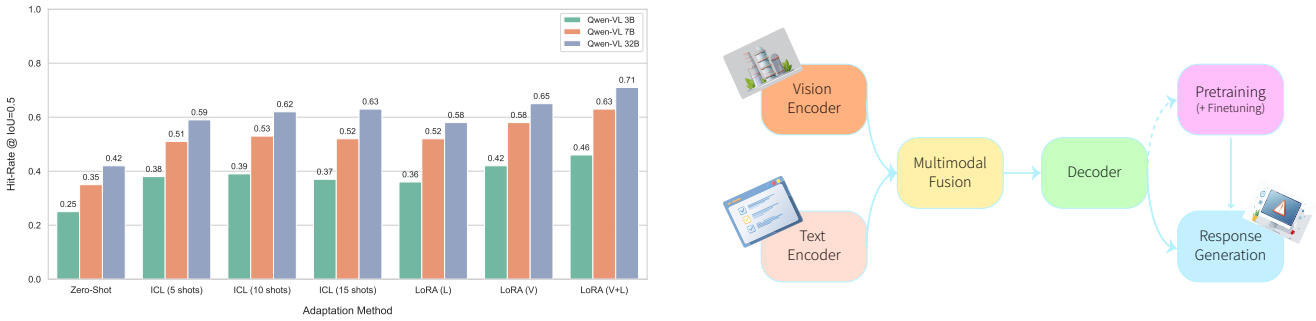
These findings necessitate a shift in paradigm. If acquiring thousands of real spill images is infeasible due to cost, rarity, or privacy constraints, then an alternative is required. We argue that **generative models—particularly modern diffusion systems—offer a practical and scalable solution to the data bottleneck**.

By carefully engineering synthetic imagery that spans the visual diversity of industrial anomalies, we can teach VLMs to recognize both the common and the rare. Such data can be produced in arbitrary quantity, with perfect labels, under varied environmental conditions—without compromising operational safety or privacy.

The next section outlines our synthetic data generation pipeline, its integration with VLM adaptation, and its impact on downstream performance.

5. Synthetic Data Generation

The effectiveness of supervised learning and fine-tuning strategies is ultimately constrained by the data available. In safety-critical domains such as industrial spill detection, this poses a major obstacle: real-world hazardous incidents are both rare and difficult to document due to privacy, safety, and operational constraints. As a result, there is a growing



(a) Performance of adaptation methods across Qwen-VL model sizes.

(b) Architecture-level overview of LoRA adaptation strategies.

Figure 3. Comparison of adaptation strategies and their corresponding model targets.

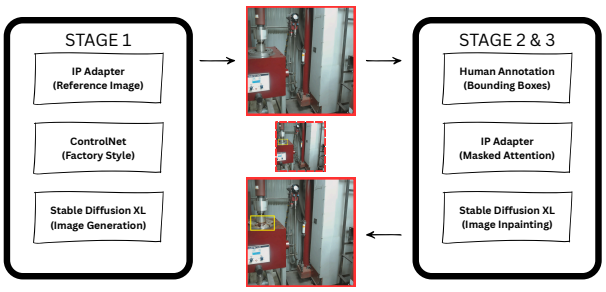


Figure 4. End-to-end synthetic data generation workflow. Stage 1 constructs factory-style backgrounds guided by reference images and structural maps via Stable Diffusion XL, IP Adapter, and ControlNet. Stages 2 and 3 add hazards: bounding boxes are manually placed in plausible locations, and inpainting is performed using SDXL with spill-specific prompts and attention conditioning.

paradox—models must be trained to detect anomalies they are unlikely to ever observe in real data.

We resolve this paradox by showing that while hazardous *events* are rare, their visual *signatures* are synthesizable. We introduce a scalable, three-stage synthetic data pipeline that produces photorealistic, structurally plausible spill scenarios anchored in real industrial imagery. Using only a handful of unlabelled factory images as style guides, our method yields a corpus of 2,000 high-fidelity synthetic images with precise annotations, designed to serve as a drop-in adaptation dataset for VLMs and object detectors.

Stage 1: Domain-Anchored Scene Generation

The foundation of our pipeline is the generation of diverse, realistic, and context-aware background images. We employ a **triple-guided generative process** using Stable Diffusion XL (SDXL) to ensure the outputs are both photorealistic and domain-aligned. The three guidance mechanisms are:

Textual Prompts: These define semantic content, including architectural elements (e.g., *concrete floor*, *metal catwalk*), lighting conditions (*diffuse fluorescent lighting*), and scene types (*industrial equipment corner*, *control room*).

IP-Adapter Conditioning: This injects compositional and stylistic priors from real Siemens Energy plant imagery, anchoring the generation in authentic color palettes, textures, and spatial layouts.

ControlNet: Using edge and depth maps extracted from the same references, ControlNet imposes geometric structure and enforces plausible perspective, preventing distortion of industrial elements.

Together, these signals generate a diverse corpus of "clean" factory environments that are not merely synthetic but visually faithful to real-world plant conditions (Figure **??**).

Stage 2: Expert-Guided Anomaly Localization

A key novelty of our pipeline lies in the accurate placement of anomalies. Rather than inserting spills arbitrarily—which risks creating unrealistic scenarios—we introduce a **human-in-the-loop annotation step**. An experienced annotator identifies semantically plausible regions for hazards based on operational knowledge: beneath valves, near pipe junctions, or along frequently serviced equipment.

Each bounding box encodes an implicit causal narrative ("a valve may leak here"), aligning the synthetic data with how real spills manifest in industrial settings. This step grounds the synthetic corpus in physical plausibility and prevents the generation of spurious training data.

Stage 3: Physically-Plausible Spill Inpainting

Given expert-defined spill locations, we simulate realistic hazards via **differential inpainting**. A soft binary mask is generated for each bounding box and passed to SDXL in inpainting mode to render the spill. This allows the inserted anomaly to seamlessly blend with the surrounding scene.

We again use a combination of conditioning techniques to ensure fidelity:

- **Textual Prompts** specify the material (e.g., *glistening oil stain*, *chemical residue*).
- **IP-Adapter** now uses real-world spill references to guide texture, opacity, and fluid behavior.
- **ControlNet** preserves lighting and geometry cues so the inpainted region respects shadows, reflections, and surface curvature.

The result is a spill that is not just visually plausible in isolation, but contextually integrated into its environment—respecting both physics and optics (see Figure **??**).

Pipeline Summary and Implementation Details

Our full pipeline—background generation, expert-informed box annotation, and guided inpainting—is lightweight and modular, enabling rapid scaling. Table 3 outlines the key hyperparameters used at each stage.

Table 3. Key hyperparameters used in synthetic image generation.

Parameter	Value / Configuration
Scene Generation and Inpainting	
Base Model	Stable Diffusion XL 1.0
Image Resolution	1024 × 1024
Sampler	DDPM-SDE (Karras scheduler)
Sampling Steps	50
CFG Scale	7.5
ControlNet Guidance	0.5–0.7
IP-Adapter Strength	0.6
Inpainting Specifics	
Inpainting Model	SDXL Inpainting
Mask Feathering	50 pixels

Distribution and Dataset Insights

Figure **??** visualizes the distribution of spill types and bounding box positions in our 2,000-image corpus. The dataset exhibits wide variability across lighting, geometry, spill size, and composition—approximating the operational diversity of real-world energy plants. The bounding box annotations are exportable in COCO format and used directly for training both VLMs and traditional object detectors.

Why This Matters

Unlike previous work that relies on purely synthetic environments or generic industrial textures, our method produces data that is:

- **Visually grounded in real operational sites**;
- **Structurally valid at both local and global scales**;
- **Label-consistent and ready-to-train without human refinement**.

By bridging the sim-to-real gap with tailored generation and human-in-the-loop realism checks, we create a dataset that unlocks robust downstream performance, even in the absence of any real-world labels.

This corpus becomes the foundation for the experiments described in the next section, where we demonstrate that models adapted on this synthetic data not only outperform their zero-shot counterparts, but in many cases, approach or exceed the performance of detectors trained on curated real-world subsets.

6. Experiments and Results

We now present a comprehensive evaluation of our proposed approach. Our central claim is that adapting a Vision-Language Model (VLM) using synthetic data via Parameter-Efficient Fine-Tuning (PEFT) offers superior performance over conventional object detectors, even when all models are trained on the same synthetic dataset. This section quantifies that advantage and outlines its practical significance.

6.1. Experimental Setup

Evaluation Datasets. We evaluate on two real-world test sets, both entirely held out from training and generation, ensuring a true test of sim-to-real generalization:

- **Public Oil Spill Dataset:** A standardized, open-source benchmark of diverse spill images.
- **Siemens Energy (SE) Proprietary Data:** A challenging in-house dataset from operational plant environments, featuring subtle leaks and visually complex industrial scenes.

Baseline Models. We fine-tune two state-of-the-art object detectors on our 2,000-image synthetic dataset, both initialized from COCO pre-trained weights:

- **YOLO-SOTA:** A high-speed detector from the YOLO family (e.g., YOLOv11).
- **DETR-SOTA:** A transformer-based variant (e.g., RF-DETR) known for handling complex spatial contexts.

Proposed Method. Our method uses **Qwen-VL 7B and 32B** VLMs adapted using LoRA (V+L), the best-performing PEFT strategy identified earlier. We compare their performance with both zero-shot VLMs and the baselines above.

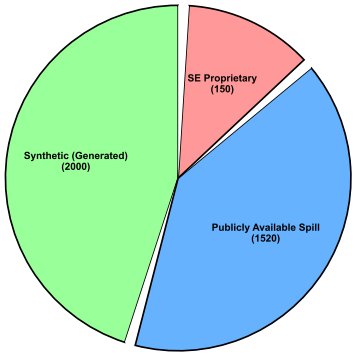
Metrics. Performance is reported using mean Average Precision (mAP) at IoU = 0.5 — a standard metric in object detection.

6.2. Main Quantitative Results

Table 4 summarizes the core results. Several important trends emerge:

- **Zero-shot VLMs** underperform, confirming that domain specialization is essential for industrial safety tasks.
- **YOLO and DETR baselines** perform well after fine-tuning, demonstrating the effectiveness of our synthetic data.

(a) Synthetic background samples generated via triple-guided diffusion.



(b) Image source distribution across synthetic, public, and proprietary datasets.

Figure 5. Overview of the dataset used for model adaptation and evaluation.

Table 4. Main performance comparison (mAP@50).

Model / Method	Public Dataset (mAP@50)	SE Proprietary (mAP@50)
Qwen-VL 7B (Zero-Shot)	0.35	0.15
Qwen-VL 32B (Zero-Shot)	0.42	0.24
Baselines (Fine-Tuning w/ Synthetic + Public Data)		
YOLOv11	0.81	0.64
RF-DETR	0.83	0.67
Proposed Method (PEFT w/ Synthetic + Public Data)		
Qwen-VL 7B + LoRA (V+L)	0.78	0.66
Qwen-VL 32B + LoRA (V+L)	**0.84**	**0.71**

- **Our PEFT-adapted VLMs** outperform all baselines — even the best-performing DETR — by 6–7 mAP points, despite not being trained end-to-end.

This validates our core hypothesis: steering a large foundation model with lightweight adaptation is more powerful than training smaller models from scratch. The VLM's broad prior knowledge, when combined with targeted synthetic data, creates a highly accurate and generalizable system for industrial leak detection.

6.3. Qualitative Insights

As illustrated in Figure **??**, the benefits extend beyond metrics. In complex scenes with occlusions, glare, or ambiguous textures, baseline detectors often falter. Our PEFT-adapted VLM correctly identifies subtle leaks while ignoring misleading artifacts like water stains or shadows — a result of its deeper contextual understanding.

6.4. Discussion and Industrial Impact

Fast, Practical, and Scalable. Our system is designed for deployment, not just academic benchmarks. It runs on single-GPU infrastructure, integrates with standard CCTV streams, and adapts overnight via PEFT — no new sensors, no retraining from scratch. This makes it ideal for real-world operations where speed and agility are critical.

Towards Predictive Maintenance. By logging the timestamp, location, and area of detected leaks, the system lays the groundwork for predictive analytics. Trends in leak patterns can serve as early indicators of mechanical fatigue, enabling timely interventions and reducing unplanned downtime.

Future-Ready Architecture. Our LoRA-based adaptation strategy ensures longevity. As new VLMs are released, they can be slotted into our framework with minimal overhead — no need to recollect data or retrain from scratch. This decouples model upgrades from infrastructure constraints and ensures continued state-of-the-art performance.

7. Conclusion

Industrial hazard detection suffers from a core challenge: the events we most need to detect are too rare and sensitive to collect large-scale real-world datasets. This paper tackles

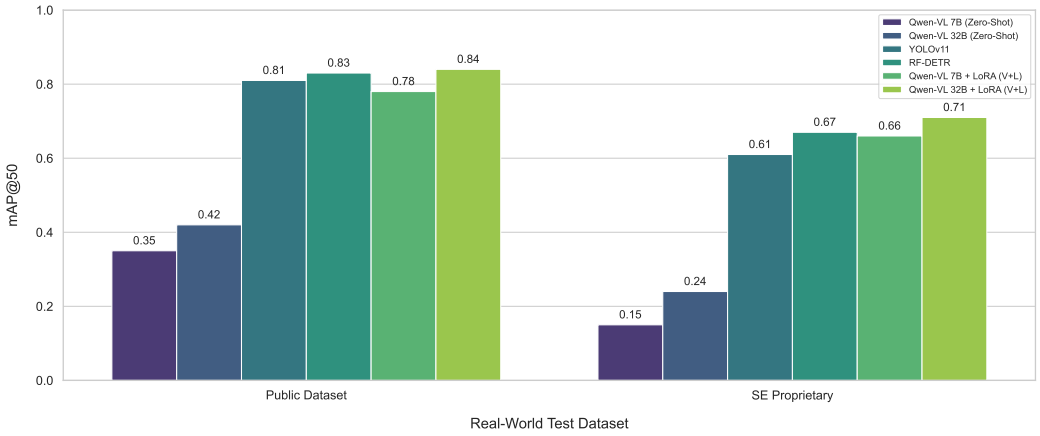


Figure 6. Overall detection performance across different model types and adaptation strategies. The VLM adapted with LoRA-(V+L) on synthetic data achieves the highest mAP, surpassing even traditional detectors fine-tuned on the same data.

that challenge head-on by combining synthetic data generation with efficient model adaptation.

We introduced a practical, three-stage pipeline that generates realistic, high-resolution images of industrial spills. Using just a few unlabelled factory images as references, we created 2,000 richly annotated scenes grounded in real-world context and geometry. These synthetic examples serve as effective training data for vision-language models (VLMs), sidestepping the need for risky or infeasible data collection. Our experiments show that parameter-efficient fine-tuning (PEFT) of a VLM on this synthetic dataset leads to strong performance across public and proprietary test sets—surpassing traditional object detectors trained on the same data. This validates our core idea: foundation models, when adapted with high-quality synthetic data, offer a better path for safety-critical detection than conventional models trained from scratch.

Beyond accuracy, our approach is flexible. It enables rapid deployment in new environments and supports continual updates simply by generating new data. This adaptability is key to building practical, scalable safety systems that can evolve as factories, hazards, and regulations change.

This work presents a viable blueprint for real-world industrial AI: generate what you cannot collect, adapt what you cannot retrain, and deploy with confidence in high-stakes settings.

8. Limitations

While our approach demonstrates strong performance and practical viability, several limitations arise from real-world constraints rather than conceptual shortcomings.

First, our synthetic dataset, though diverse, was capped at 2,000 images due to the time and resource demands of high-fidelity image generation and human-in-the-loop bounding box annotation. Scaling further could improve model generalization, especially for rare spill types or edge-case environments. Second, our evaluation is currently limited to static images. Many industrial hazards evolve over time, and incorporating temporal cues from video could further improve detection reliability. However, video-based synthetic generation and VLM adaptation remain technically and computationally intensive at scale. Third, while our human-in-the-loop spill placement adds critical realism, it introduces subjectivity and cannot yet be fully automated without risking plausibility errors. Automating this step with learned priors or simulation-based physics models is a promising direction, but outside the scope of this study.

Lastly, our experiments focus on vision-only data. In practice, multispectral inputs—such as thermal or infrared—could help disambiguate ambiguous spill types. Extending our framework to handle such modalities would require a different generation and adaptation pipeline.

Overall, these limitations reflect practical trade-offs rather than fundamental flaws. Our framework is designed to be modular and extensible, laying the foundation for broader future work in industrial-scale vision systems.

References

[1] John Adams and Howard Bloom. Gaussian pyramid image and its application to change detection. *Computer Vision and Image Understanding*, 1995. 2

[2] Tom Brown and et al. Language models are few-shot learners. In *NeurIPS*, 2020. 4

[3] Tim Dettmers and et al. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2310.02578*, 2023. 3

[4] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3, 4

[5] Li Jia and et al. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 2, 3

[6] Menglin Jia, Luming Tang, Bor-Chun Chen, Claire Cardie, Serge Belongie, Bharath Hariharan, and Ser-Nam Lim. Visual prompt tuning, 2022. 3

[7] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 3

[8] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 3

[9] Xin Liu and et al. Grounding dino: Marrying object detection with grounded language queries. In *CVPR*, 2023. 3

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3

[11] R. J. Radke, S. Andra, O. Al-Kofahi, and B. Roysam. Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing*, 14(3):294–307, 2005. 3

[12] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 3

[13] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3

[14] Ping Wang, Jiangbo Liu, Yilai Yan, and Zongjian Tang. A survey on deep learning-based industrial defect detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021. 1, 2