SynSpill: 使用合成数据改进工业洒溢检测

Aaditya Baranwal†
University of Central Florida
Orlando, FL, USA
aaditya.baranwal@ucf.edu

Abdul Mueez
University of Central Florida
Orlando, FL, USA
abdul.mueez@iitj.ac.in

Jason Voelker
Siemens Energy
Orlando, FL, USA
jason.voelker@siemens-energy.com

Guneet Bhatia
Siemens Energy
Orlando, FL, USA
guneet.bhatia@siemens-energy.com

Shruti Vyas
University of Central Florida
Orlando, FL, USA
shruti@ucf.edu

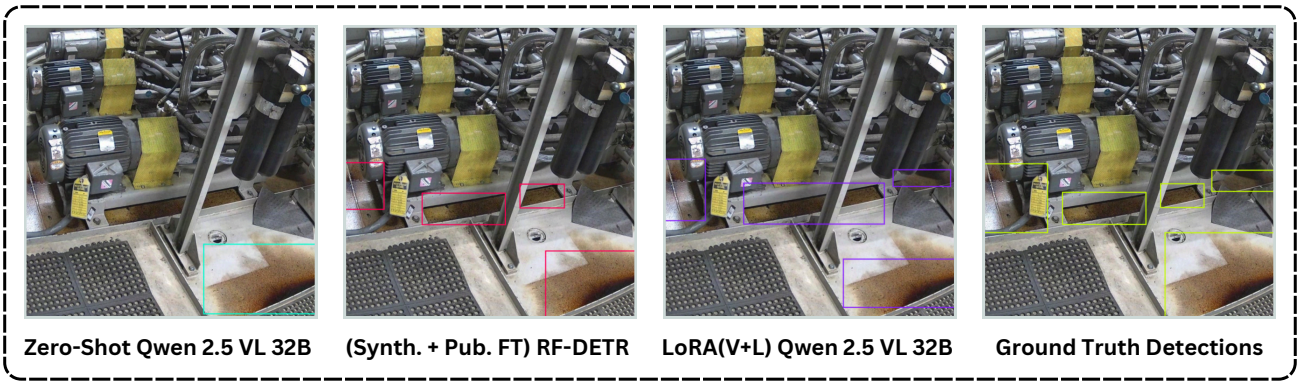


Figure 1. Comparative detection performance of competing methods on a real-world CCTV image of an industrial spill. We visualize and contrast the predictions of three models: (1) a Zero-Shot Qwen2.5-VL-32B baseline without adaptation, (2) a PEFT-adapted Qwen2.5-VL-32B using LoRA on synthetic and web-scraped public data, and (3) a finetuned RF-DETR Base model trained on the same hybrid dataset. The ground-truth annotation is overlaid for reference. This qualitative comparison highlights the relative gains.

Abstract

大规模视觉语言模型（VLMs）通过强大的零样本能力，已经改变了通用视觉识别。然而，在工业泄漏检测等小众且安全至关重要的领域，它们的性能显著下降，这些领域里的危险事件罕见、敏感且难以标注。这种稀缺性是由隐私问题、数据敏感性以及真实事件发生频率低所驱动的，导致传统微调检测器对于大多数工业环境而言不可行。我们通过引入一个以高质量合成数据生成流程为核心的可扩展框架来应对此挑战。我们证明了这种合成语料库能够实现 VLMs 的有效参数高效微调（PEFT），并大幅提高了如 YOLO 和 DETR 等最先进目标检测器的性能。值得注意的是，即使在没有合成数据（SynSpill 数据集）的情况下，VLMs 在未见过的泄露场景中仍然比这些检测器泛化得更好。当使用 SynSpill 时，VLMs 和检测器都取得了显著改进，其性能变得相当。我们的结果强调，高保真合成数据是一种跨越安全关键应用领域差距的强大手段。合成生成和轻量级适应相结合，为在工业环境中部署视觉系统提供了一条具成本效益且可扩展的途径，在真实数据稀缺或难以获取的情况下尤为适用。项目页面：synspill.vercel.app

1. 介绍

持续的警惕在工业操作中是不可或缺的，因为未被发现的危险（例如流体泄漏、化学品溢出或机械故障）可能迅速升级，造成经济损失、环境危害和对人类安全的风险 [48] 。历史上，工业监测依赖于人工检查或固定的物理传感器。然而，这两种方法都有局限性：人工检查容易出错，并且在规模上不可行，而静态传感器提供

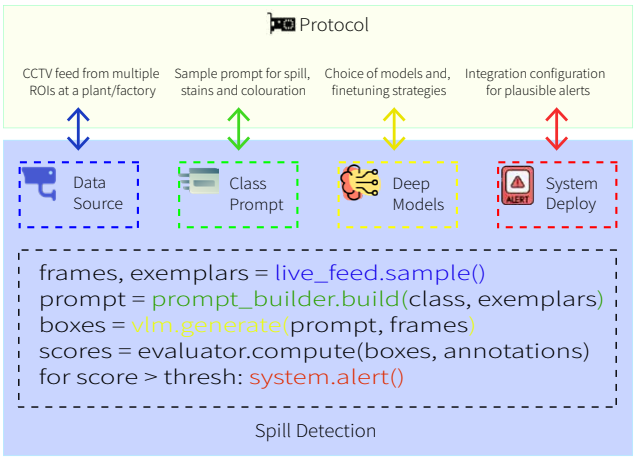


Figure 2. 工业泄漏检测框架概述。系统摄取实时 CCTV 视频流和用户定义的文本提示。通过选择的策略选择并微调的视觉-语言模型（VLM）分析输入，以检测和定位潜在的泄漏，当检测信心超过预定义的阈值时触发警报。

有限的空间感知能力，它们可能表明存在危险，但不知道危险在哪里或是什么。

计算机视觉已成为一个有前景的解决方案，提供自动化、上下文感知的监控。现代物体检测器如 YOLO 和 DETR [10, 13, 16, 44, 52] 在检测速度和准确性上推动了前沿。这些模型在结构化环境中表现极佳，但在工业环境中的实际应用仍因数据稀缺而成为瓶颈。工业事故数据稀少，通常是专有的且难以标注，因此难以收集大规模、多样化的数据集。因此，即使是最先进的检测器也往往过拟合于它们所训练的领域，难以在不同的光照条件或设施布局中进行泛化 [43, 48] 。

为了应对这些局限性，研究探索了多种路径。经典的异常检测方法，如高斯金字塔差分 [1] ，提供快速但语义盲目的解决方案。最近，视觉-语言模型（VLMs）[14, 36] 在开放世界设置中展示了零样本识别的潜力。虽然预训练的 VLMs 在广泛的概括上表现出色，但它们的定位能力和对特定领域视觉线索的理解在没有任务适应的情况下仍然有限。

我们提出了一个通过生成式人工智能和参数高效微调（PEFT）解决核心数据瓶颈的框架。我们并不依赖稀缺的现实世界数据，而是开发了一个高保真度的合成数据生成流程，该流程结合了 Stable Diffusion XL、IP 适配器和图像修复技术，以模拟各种逼真的泄漏情景。

然后使用此合成数据集通过类似低秩适应（LoRA）[12] 的 PEFT 策略来调整 VLMs，只需更新模型权重的一小部分即可实现领域专业知识的融入。

重要的是，我们展示了这个合成语料库对视觉-语言和目标检测模型都有益。我们的实验表明，在合成 + 网页语料库上微调 RF-DETR 和 YOLOv11 大幅提升了它们的真实世界检测准确性。然而，在没有合成数据的低数据环境中，视觉语言模型展示出了比这些特定任务检测器更强的泛化能力。一旦合成数据被引入，这两种方法都变得具有竞争力，这强调了我们流程的广泛实用性。

我们在这项工作中的贡献包括：

1. 在工业环境中评估用于溢出检测的视觉-语言模型。我们实施了少量样本的方法，例如使用 LoRA 进行领域适应的上下文学习（ICL）和参数高效微调（PEFT）。
2. 引入了一种新颖的、可控的管道 AnomalInfusion，用于使用带有 IP 适配器和异常修复的引导稳定扩散来生成多样化且逼真的工业泄漏图像。
3. 我们的合成数据集提高了 VLM 和最新的目标检测器（整个论文中的 YOLOv11 和 RF-DETR Base）的性能，突出了其在安全关键、数据稀缺领域的广泛适用性和重要性。

2. 相关工作

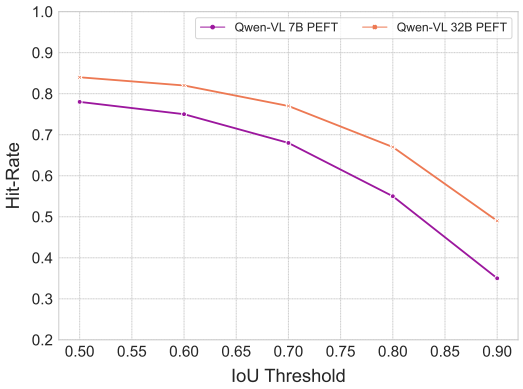
自动化工业危害检测的追求经历了三个相互关联的研究前沿的发展：(i) 从监督检测器到视觉-语言基础模型的演进，(ii) 使用合成数据克服标注瓶颈，(iii) 旨在实现可扩展领域适配的参数高效微调（PEFT）技术的出现。我们的工作定位于这些领域的交汇点。

从监督检测器到基础模型。传统的工业视觉监控系统始于经典技术，如背景减法和图像差分 [37] ，这些方法计算量小，但对环境变化非常敏感。深度学习的出现引入了强大的监督对象检测器，包括像 Faster R-CNN [40] 这样的两阶段方法，以及像 YOLOv3 [39] 和 DETR [3] 这样的单阶段检测器。更近期的继任者，如 YOLOv7–v12 [10, 47] DETR、DINO、SAM 和 RF-DETR [19, 27, 44, 52] ，在实时检测和长尾鲁棒性方面推进了最新性能。然而，这些模型高度依赖于大规模的标注数据集，而在敏感或安全关键的环境中，这些数据集是无法获取的。

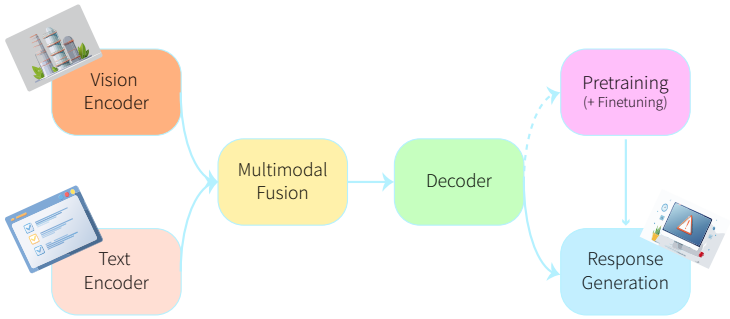
为了应对数据匮乏下的泛化问题，最近的努力转向了视觉-语言模型（VLMs），如 CLIP [36] ，ALIGN [14] 和 Florence [50] ，它们是在网页规模的图文对上进行预训练的。这些模型在图像级分类和视觉定位方面表现出强大的零样本迁移。在此基础上，像 GLIP [24] ，GroundingDINO [27] 和 Grounded-SAM [18] 这样的开放词汇检测器结合了弱监督下的检测和分割。然而，由于在训练语料中严重的领域转移、罕见的纹理、光照条件和异常类型的代表性不足，这些系统在工业环境中仍然表现下降。

合成数据与模拟到现实的差距。合成数据已经成为解决标注瓶颈的实际方案。早期的方法使用游戏引擎或 3D 模拟器（例如，CARLA [8] ，AI2THOR [21] ）生成用于训练的合成环境。然而，这些渲染中有限的真实感常常引入“模拟到现实”的差距 [6] 。扩散模型 [11] 彻底改变了这个领域，能够生成具有高保真度的语义对齐图像。文本到图像模型，如 DALLE-2 [38] 和 Stable Diffusion [41] ，现在可以通过 ControlNet [53] 、IP-Adapters [49] 或 DreamBooth [42] 对详细提示或图像进行条件控制，从而提高生成中视觉基础的准确性。

几项工作展示了合成图像在检测和分割中的有效性：



(a) 定位准确度作为 IoU 阈值的函数。



(b) PEFT 流程的高级概述

Figure 3. 适应策略及其对空间精度的影响。左侧子图说明了 LoRA 适应的模型在更严格的 IoU 阈值下保持更高的定位精度，验证了合成监督的有效性。右侧子图展示了 LoRA 在视觉和语言骨干网络中应用的组件级别分解。

比如 GenAug [9] 、域随机化 [6] 、Task2Sim [31] 和 DreamFusion [35] 。其他诸如基于 StyleGAN 的模拟器 [17, 26] , 曾被用于组合域转移。然而，现有的大多数工作侧重于训练小型模型或增强真实数据集，而不是在零或近零真实样本的领域中单独使用合成数据完全适应大型基础模型。工业异常检测中的应用仍未被充分探索，尽管最近引起了兴趣 [32] 。

高效领域适应通过 PEFT。模型的全面微调在计算上非常昂贵且容易导致灾难性遗忘 [20] 。为了解决这个问题，参数高效微调（PEFT）成为一种引人注目的替代方案。LoRA [12] 、QLoRA [7] 、AdapterFusion [34] 和 BitFit [51] 允许调整模型的权重不到 1 %，同时保持强劲的下游性能。虽然在 NLP 中被广泛采用，但 PEFT 在视觉应用中仍然是初步的。最近的工作如 VPT [15] 、SSF [25] 和 AdaptFormer [4] 将 PEFT 应用于视觉变换器，但很少有零样本检测方案或检验 PEFT 在极端数据稀缺情况下的表现。

我们的工作通过整合这三方面的研究推进了该领域的发展。我们提出了一个统一的框架，该框架：(1) 利用高保真的基于扩散的方法生成工业泄漏数据集，而无需真实事件数据，(2) 使用像 LoRA 这样的参数高效策略来调整视觉语言模型 (VLMs)，以及 (3) 展示了所生成的模型不仅在与完全微调的检测器相比时具有竞争力，而且在低数据或零样本的情况下具有更好的泛化能力。据我们所知，这是首次系统地研究如何使用纯粹的合成数据来调整基础视觉语言模型以进行工业危险检测，弥合了通用预训练与专业安全关键部署之间的差距。

3. 实验设置

设计一项同时对工厂一线实践者和人工智能研究人员有意义的研究需要谨慎权衡：系统必须能够以有限的资源进行复现，同时要具备足够的技术细节文档以经受审查。我们的实验评估了如何使用合成数据和公开可用的数据将视觉语言模型（VLMs）调整用于工业泄漏检测。

3.1. 视觉语言模型

VLMs 通过在大规模图像-字幕数据集上的联合预训练来对齐图像和文本模态 [14, 36] 。在推理时，它们处理图像和文本提示（例如，“哪里有泄漏?”），以生成文本答案或结构化输出，例如边界框。

我们采用开源的 Qwen2.5-VL [2] 系列模型，具有三个参数规模：3B、7B 和 32B。这些模型结合了 Swin 风格的视觉变换器 [28] 和因果语言解码器，以形成一个共享的视觉语言潜在空间。3B 模型适用于单 GPU 设置，而 32B 变体则提供更高的基础准确性。

3.2. 结构化提示

每张图像都配有一个结构化提示，旨在引发在高风险环境中进行精确且保守的检测行为。系统提示模拟了一名工业检查员的推理，而用户提示则请求以 YOLO v11/COCO JSON 格式输出边界框，用于八种异常类别之一（例如，oil-spill 、floor-stain 、chemical-discoloration ）：

。

系统：“您是一名认证的工业安全检查员，专门从事工厂和能源厂中的危险泄漏、漏液和污染检测。只能报告可验证的安全隐患。不得猜测或推测。”

用户：“如果存在，请检测并返回 < 类 > 的边界框坐标，以 COCO JSON 格式。”

为了鼓励简洁和可靠的输出，我们设置了解码参数如下：温度 $\tau = 0.10$ ，核采样 $p = 0.001$ ，以及重复惩罚为 1.2。

3.3. 参数高效微调 (PEFT)

我们评估了三种适应策略，以评估在不同资源和数据约束下的性能：

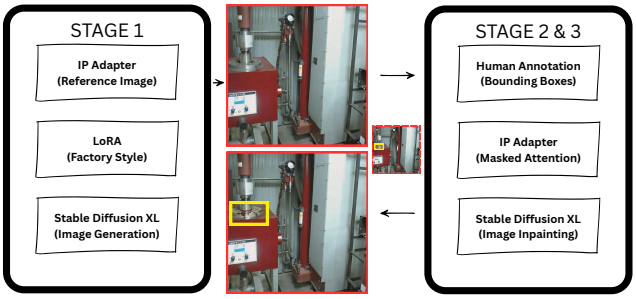


Figure 4. 端到端合成数据生成工作流程。第一阶段通过 Stable Diffusion XL [41] 、IP Adapter [49] 和 LoRA [30] ，在参考图像和结构图的指导下构建工厂风格的背景。第二和第三阶段添加潜在危险：在合理位置手动放置边界框，并使用带有特定泄漏提示和条件的 SDXL 进行修复。

零样本推理。 该模型原样使用，没有进行额外的调优。该基线评估了预训练的 VLM 在泄漏检测方面的即插即用通用性。

上下文学习 (ICL)。 在输入提示前附加 $k \in \{5, 10, 15\}$ 标记的示例（图像 + 边界框）。模型权重保持不变。预测是以一个支持集 $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{k}$ 为条件的：

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

在推理时利用少量可用的已解决示例。

低秩适应 (LoRA)。 LoRA [12] 引入了轻量化的可训练适配器，通过仅修改少量参数来注入特定领域的知识。LoRA 不是微调完整的权重矩阵 $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ ，而是学习两个低秩矩阵 A 和 B ：

$$W' = W + \alpha AB. \tag{2}$$

其中 $A \in \mathbb{R}^{d_{\text{out}} \times r}$， $B \in \mathbb{R}^{r \times d_{\text{in}}}$， $\alpha = \frac{1}{r}$ 。
我们评估了三种 LoRA 配置：

- LoRA-L：仅适应语言通路。
- LoRA-V: 仅适用于视觉编码器。
- LoRA-(V+L)：两个路径共同适应。

对于所有配置和数据集，LoRA 均在 800 步（带有提前停止）内进行训练，批次大小为 8，梯度累积为 4 步（有效批次大小为 32）。在我们的实验中，秩固定为 8，学习率设置为 $5 * 10^{-5}$ 。

3.4. 数据集

我们在两个真实世界数据集上进行评估，其中一部分在训练和生成阶段被保留，以确保对模拟到真实泛化的严格测试：

- 公共泄漏数据：一个通过网络爬虫获取的数据集，来源于 Roboflow 公共数据集，并经过彻底的去重处理。这些来源经过精心挑选，以尽可能接近我们的工厂场地泄漏。共收集了 1520 张图像，其中 520 张用于评估和测试，其余 100 张分别用于 VLM 和 SOTA 目标检测器的 PEFT 和 FT。
- 专有工厂数据：这是一个具有挑战性的内部数据集，从实际工业场所收集，特点是在复杂环境中存在细微的泄漏、污迹和遮挡。总共 150 张图像，其中 50 张用于 ICL，其余 100 张用于评估。
- 合成泄漏（SynSpill）数据集：一个由 2000 张通过我们的 AnomalInfusion 管道（见第 4 节）生成的照片级真实感合成泄漏图像组成的精选数据集，全部用于 PEFT 和 FT 目的。

3.5. 评估指标

我们报告了在 IoU 阈值为 0.5 时所有异常类别的平均命中率。如果预测的 bbox 与真实值有足够的重叠，则计为命中：

这捕捉到检测准确性和空间精度，与安全关键监测中的要求保持一致。

4. 合成数据生成

尽管监督学习和微调推动了视觉系统的显著进步，它们的成功最终取决于标注数据的可用性。在像工业泄漏检测这样的安全关键领域，这提出了一个核心挑战：危险事件是稀有的、敏感的，并且由于隐私、安全和操作限制而难以捕获。因此，模型必须学习检测它们在真实世界训练数据中可能从未遇到的异常。

我们通过认识到尽管危险事件很少发生，但可以合成它们的视觉特征来解决这一悖论。我们引入了一种可扩展的三阶段合成数据管道，该管道生成以真实工业图像为基础的具有照片逼真度和结构合理性的泄漏场景。仅使用一小组 150 张无标注的工厂图像作为风格参考，我们的方法生成了包含 2000 张合成图像的语料库，并提供精确的注释，旨在用作危险检测的直接适应数据集。

4.1.

阶段 1：领域锚定场景生成

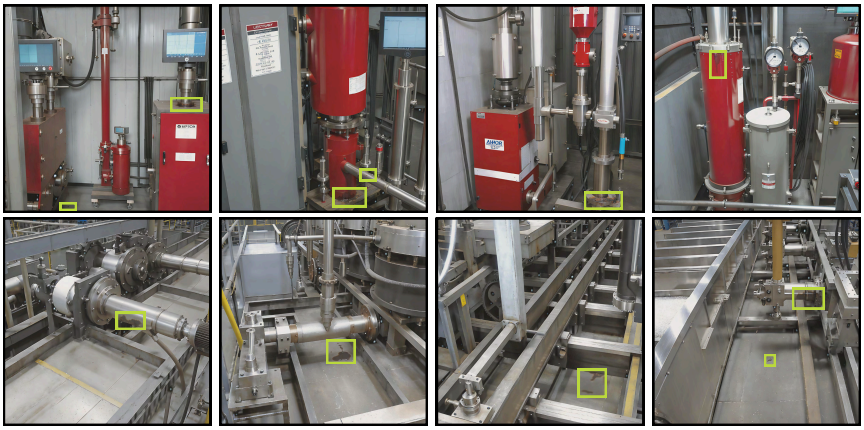
我们的流程基础是生成多样化的、具有上下文感知的背景图像。我们采用一个使用 Stable Diffusion XL (SDXL) 的三重引导生成过程，以确保输出既是照片真实的，又与领域一致。这三个引导信号是：
文本提示：定义语义内容、建筑元素（例如，混凝土地板、金属走道）、照明（漫射荧光照明）和场景类型（控制室、工厂走廊）。IP-Adapter 调节：从 150 张真实工厂图像中转移风格先验，将生成过程锚定在真实的纹理、颜色和空间布局中。LoRA 风格注入：应用低秩适配模块以增强工业现实主义和一致性。

这些机制共同生成了一组多样的“干净”工厂场景，这些场景不仅仅是合成的，而且在视觉上忠实于实际的工厂环境。

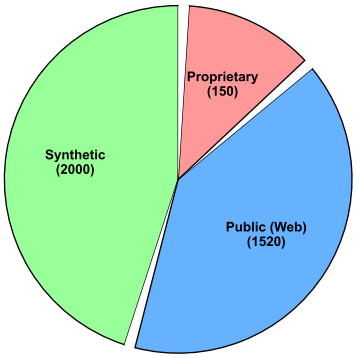
4.2.

阶段 2：专家引导的异常定位

我们的流程的一个关键创新点在于异常是如何引入的。我们并不是随意插入泄漏，这样做有生成不现实样本的风险，而是采用“人机交互”标注步骤。一位经验



(a) 使用三重引导扩散生成的代表性合成图像样本。



(b) 来源分布：合成 (2,000)、公开 (1,520)、专有 (150)。

Figure 5. 用于适应和评估的数据集组成。左侧面板突出显示了通过合成生成实现的变异性，而右侧面板总结了每个来源类别的总图像数量。

Table 1. 用于生成合成图像的关键超参数。

Parameter	Value / Configuration
Scene Generation Specifics	
Base Model	Stable Diffusion XL 1.0 [46]
Image Resolution	1024 × 1024
Sampler	DDPM-SDE-2m-GPU
Scheduler	Karras
Sampling Steps	64
CFG Scale	8
LoRA Strength	0.2–0.4
IP-Adapter	IP Composition+CLIP-ViT-H [22, 33]
IP-Adapter Strength	0.6
Inpainting Specifics	
Inpainting Model	SDXL-Turbo Inpainting [45]
Differential Diffusion	Enabled
Mask Feathering	50 pixels
Mask Opacity	75 %
Denoise Strength	0.5-0.6

丰富的标注者识别出语义上合适的危险区域：如阀门、管道接口或频繁维修的设备附近。

每个边界框编码了隐式的因果逻辑（“阀门可能会在这里泄漏”），这与溢出在实际场景中的表现方式相一致。此步骤确保我们的语料库反映出操作的合理性，而非合成的随机性。

4.3.

阶段 3：物理上合理的溢出修补

给定标注的区域，我们通过微分补全 [23, 29] 模拟危害。每个边界框生成一个软二元掩码，并传递给 SDXL 的补全模式。这使得异常能够与周围场景无缝融合。我们应用分层条件以保持真实性和一致性：

- 提示材料描述类型（例如，闪闪发光的油）。
- IP-Adapter 参考现实世界的溢出纹理。
- LoRA 保持光照、表面、风格和纹理。

泄漏在视觉上在孤立情况下是逼真的，并在背景中进行上下文整合，遵循物理约束和视觉一致性（参见补充材料中的图表）。

4.4.

流程总结与实现细节

整个流程，包括背景生成、专家指导的边界框标注和引导修补，是轻量的、模块化的和可扩展的。表 1 总结了每个阶段使用的核心生成参数。

4.5.

分布与数据集见解

合成语料库涵盖了多样化的照明、几何形状、泄漏形态和环境背景，紧密地反映了现实工业现场的操作变化。注释以 COCO 格式导出，并用于使用相同的边界框坐标生成羽毛蒙版以处理图像中的泄漏，之后生成的带注释图像直接用于训练 VLMs 和物体检测器。我们还从公开的工业数据集聚合了 1,520 张额外的网络抓取图像，以扩大测试的多样性并进行泛化研究。与我们 150 张专有图像集结合，这些形成了一个丰富的评估套件。

不同于以往的工作，即 (i) 生成完全合成的环境，或 (ii) 在没有语义指导的情况下叠加纹理，我们的方法通过结合针对性的生成和专家参与的现实性检查，提供了一种有原则的替代方案，以此来弥合模拟到真实的鸿沟。我们创建了一个数据集，使得在真实数据稀缺、敏感或不可用的领域能够实现稳健的适应。

这个合成语料库构成了接下来所述实验的基础，在这些实验中，我们展示了在该数据上调整过的模型不仅优于它们的零样本对应物，而且在许多情况下，其性能接近或超过在精心挑选的现实世界子集上训练的检测器。

5. 结果与讨论

我们将我们的研究结果分为两个阶段。首先，我们使用有限的真实数据评估视觉语言模型（VLMs）在少样本和参数高效调优模式下的基线性能。然后，我们使用高保真合成数据评估完整的提议管道 VLM 适应性，并将其与 SOTA 对象检测器进行对比基准测试。

Table 2. 在公共评估集上，平均命中率 @ IoU = 0.5。

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

Table 3. 在专有评估集上的平均命中率 @ IoU = 0.5。

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

5.1. 在真实世界数据稀缺情况下的适应

设置：我们评估了在不同大小（3B、7B、32B）的 Qwen2.5-VL 模型上的三种适应策略：零样本推理、上下文学习（ICL）和基于 LoRA 的参数高效微调（PEFT）。性能通过在公共和专有数据集上 IoU ≥ 0.5 的平均命中率进行测量。少样本学习：承诺与限制：使用 5 个示例的 ICL 显著提高了性能，相较于零样本基线，使 7B 模型的命中率提高了超过 45 %。然而，效果很快达到瓶颈。增加更多的示例（10 或 15 个）仅提供了微小的增益或偶尔的退步。这与先前的工作一致，突出了 ICL 在符号限制下的脆弱性 [5] 。模型规模与有效适应：较大的模型始终表现更好，但单纯的规模是不够的。例如，使用 LoRA-(V+L) 微调的 3B 模型在零样本或 ICL 设置中可以与 7B 模型媲美甚至超越。这强调了领域适应不仅仅是参数数量，对于释放模型能力至关重要。视觉最为重要。LoRA-V（仅视觉调优）始终优于 LoRA-L（仅语言），表明视觉特异性、纹理、阴影、梯度在该领域中比文本理解更为关键。LoRA-(V+L) 通过联合适应提供了额外的增益。改进的识别 = 更严格的定位：如图 3a 所示，检测的改进与更高的空间准确性相关。即使在更严格的 IoU 阈值下，适应后的模型仍保持更佳的性能，验证了监督的质量。适应增益与领域间隙：我们使用以下方法量化相对改进：对于公开数据集上的 Qwen2.5-VL-7B，LoRA-(V+L) 提升了 80 %。然而，在专有数据集上，采用相同配置的提升降到了 0.34，表明即使是参数高效的调整最终也受到可用数据覆盖范围的限制。洞察：VLMs 在局部（对于已见过的纹理或光影）上具有泛化能力，但在结构上则不然。在安全关键领域，这种能力是不足的。

为了克服上面提到的数据瓶颈，我们评估了我们的完整流程：使用通过我们的 AnomalInfusion 框架生成的合成数据来调整 Qwen2.5-VL 模型。

设置：我们在合成数据集（2000 张图像）上使用 LoRA-(V+L) 微调 Qwen2.5-VL（7B 和 32B），并与在相同数据上训练的 YOLOv11 和 RF-DETR 基线进行比较。评估使用 mAP@50。

定量结果：

- 通过 PEFT 适配的 VLMs 在两个数据集上的检测器性能相当或超出。
- Qwen2.5-32B + LoRA 在专有数据集上比 RF-DETR 高出 7 个百分点。
- 零样本视觉语言模型显著落后，进一步确认了进行领域特定适应的必要性。

定性趋势：如图 5 所示，PEFT 调整的 VLMs 对遮挡、光照变化和背景杂乱更加稳定。它们可以检测出微妙的异常，例如传统检测器经常漏检或误判的半透明油渍。

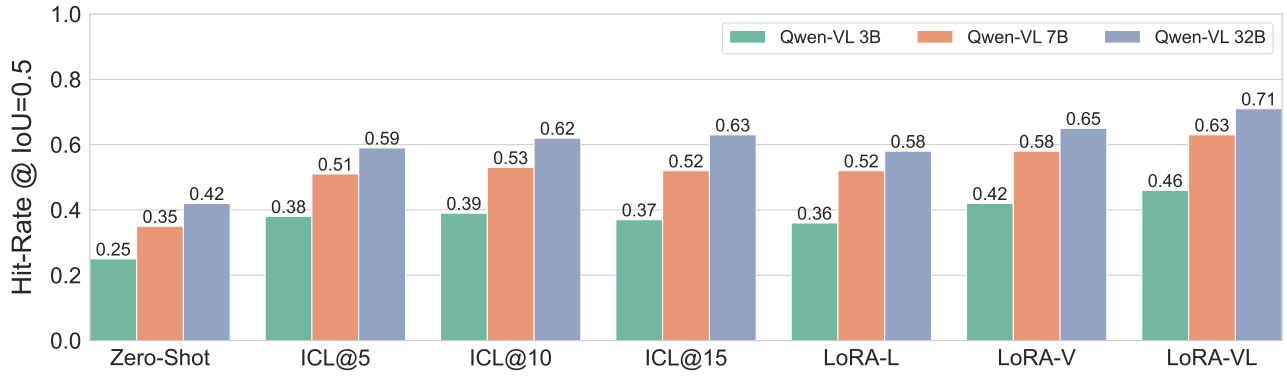
5.2. 对现实世界部署的影响

快速且可部署：我们的 PEFT 管道运行在单个 GPU 上，可以在一夜之间调整模型，并且无需进行端到端的重新训练。它可以轻松地与工厂的 CCTV 监控流集成，使其在实际监控中可行。迈向预测性维护：通过记录检测的地点、时间和严重程度，系统可以为预测故障点提供高层次的数据分析，支持主动的安全干预。面向未来的架构：模块化的 LoRA 调整策略确保了与未来基础模型的兼容性。随着新 VLM 的出现，我们的系统可以在无需大量重新工程的情况下进行适应。

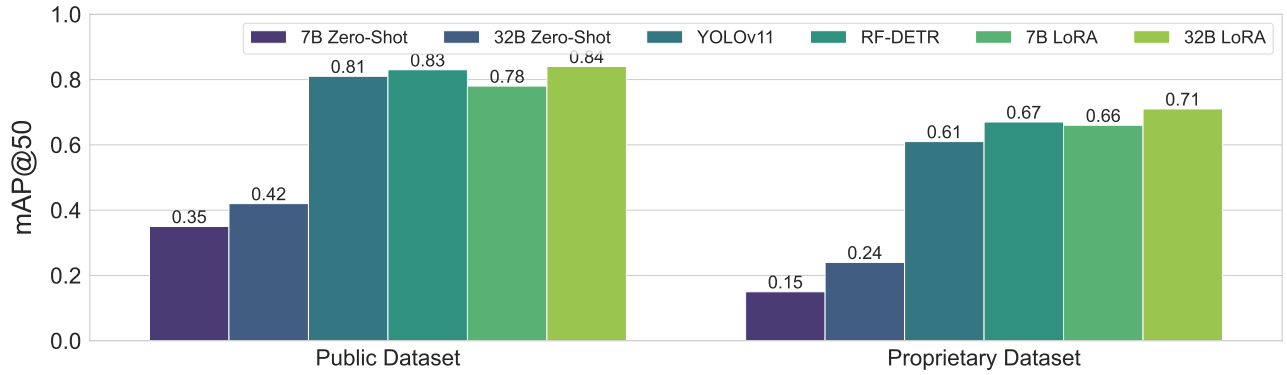
6. 结论

工业危险检测面临一个基本障碍：我们最需要识别的事件过于罕见、敏感或危险，无法支持大规模数据收集。我们通过将高保真合成数据生成与 $VLMs$ 的参数高效适配相结合来应对这一挑战。

我们的三阶段流程仅使用少量未标记的工厂参考，就能生成逼真且高分辨率的工业泄漏图像。这产生了 2,000 个具有丰富注释的场景，反映了真实世界的几何和背景，消除了进行风险或不可行的数据收集的需要。当这些合成例子用于通过参数高效微调（PEFT）来适应 VLM 时，它们在公共和专有测试集上表现出色，超越了在相同数据上训练的传统探测器。这证实了我们的核心见解：基础模型与精心设计的合成数据相结合，比从零开始训练的传统模型提供了更高效和可扩展的安全关键检测路径。SynSpill 数据集将在接收后公开。



(a) 不同大小版本的 Qwen 2.5 VL 在性能增强策略中的表现。



(b) 不同模型类型和适应策略的整体检测性能。

Figure 6. 方法、模型和数据集的表现

Table 4. 主要性能比较（mAP@50）。

Model / Method	Public Dataset (mAP@50)	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

关键的是，我们的方法不仅准确，而且具有适应性。它支持在新环境中的快速部署，并通过数据再生实现无缝更新，而无需从头开始重新训练。这种灵活性使其非常适合不断变化的工业环境。

总而言之，这项工作为现实世界中的工业 AI 提供了一个实用的蓝图：生成你无法收集的数据，适应你无法重新训练的内容，并自信地进行部署。

7. 局限性

尽管我们的方法表现出色且具有实用性，但一些限制源于现实世界的约束，而非概念上的缺陷。

首先，我们的合成数据集尽管多样，但由于高保真生成和人工参与框标注的资源需求，限制在 2000 张图像。其次，我们的评估限于静态图像。许多工业危害随着时间发展，结合视频中的时间线索可以提高检测的稳健性。第三，虽然我们的人工参与泄漏物安置增加了必要的真实性，但引入了主观性，无法在不牺牲

合理性的情况下实现自动化。

最后，我们的实验只专注于视觉数据。在实际应用中，多光谱输入（例如热成像或红外线）可以帮助消除视觉上模棱两可的情况。支持这些模式将需要重新构建生成和适应的流程。

References

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

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025.

[3] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers, 2020.

[4] Shoufa Chen, Chongjian Ge, Zhan Tong, Jiangliu Wang, Yibing Song, Jue Wang, and Ping Luo. Adaptformer: Adapting vision transformers for scalable visual recognition, 2022.

[5] Wentong Chen, Yankai Lin, ZhenHao Zhou, HongYun Huang, Yantao Jia, Zhao Cao, and Ji-Rong Wen. Icleval: Evaluating in-context learning ability of large language models, 2024.

[6] Xiaoyu Chen, Jiachen Hu, Chi Jin, Lihong Li, and Liwei Wang. Understanding domain randomization for sim-to-real transfer, 2022.

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

[8] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. Carla: An open urban driving simulator, 2017.

[9] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures, 2022.

[10] Priyanto Hidayatullah, Nurjannah Syakrani, Muhammad Rizqi Sholahuddin, Trisna Gelar, and Refdinal Tubagus. Yolov8 to yolo11: A comprehensive architecture in-depth comparative review, 2025.

[11] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[12] 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.

[13] Nidhal Jegham, Chan Young Koh, Marwan Abdelatti, and Abdeltawab Hendawi. Yolo evolution: A comprehensive benchmark and architectural review of yolov12, yolo11, and their previous versions, 2025.

[14] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.

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

[16] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023.

[17] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan, 2020.

[18] 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.

[19] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Lily Rolland, Laura Gustafson, Callen Romero, Michael Krainin, David Li, Chao Li, et al. Segment anything. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2426–2437, 2023.

[20] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114 (13):3521–3526, 2017.

[21] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, Aniruddha Kembhavi, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai, 2022.

[22] laion. CLIP ViT-H/14 –LAION-2B (laion/CLIP-ViT-H-14-laion2B-s32B-b79K). https://huggingface.co/laion/CLIP-ViT-H-14-laion2B-s32B-b79K, 2023. Published ca. Sep 2023; Accessed: 2025-07-04.

[23] Eran Levin and Ohad Fried. Differential diffusion: Giving each pixel its strength, 2024.

[24] 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.

[25] Dongze Lian, Daquan Zhou, Jiashi Feng, and Xinchao Wang. Scaling & shifting your features: A new baseline for efficient model tuning, 2023.

[26] Ruibo Liu, Jerry Wei, Fangyu Liu, Chenglei Si, Yanzhe Zhang, Jinmeng Rao, Steven Zheng, Daiyi Peng, Diyi Yang, Denny Zhou, and Andrew M. Dai. Best practices and lessons learned on synthetic data, 2024.

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

[28] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021.

[29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models, 2022.

[30] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023.

[31] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S. Feris. Task2sim : Towards effective pre-training and transfer from synthetic data, 2022.

[32] Keno Moenck, Duc Trung Thieu, Julian Koch, and Thorsten Schüppstuhl. Industrial language-image dataset (ilid): Adapting vision foundation models for industrial settings, 2024.

[33] ostris. IP Composition Adapter (stable diffusion) model. https://huggingface.co/ostris/ip-composition-adapter, 2024. Published: March 20, 2024; Accessed: 2025-07-04.

[34] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. Adapterfusion: Non-destructive task composition for transfer learning, 2021.

[35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion, 2022.

[36] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[37] 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.

[38] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

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

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

[41] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[42] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman.

Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[43] Kaziwa Saleh, Sandor Szenasi, and Zoltan Vamossy. Occlusion handling in generic object detection: A review. In 2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI), page 000477–000484. IEEE, 2021.

[44] Ranjan Sapkota, Rahul Harsha Cheppally, Ajay Sharda, and Manoj Karkee. Rf-detr object detection vs yolov12 : A study of transformer-based and cnn-based architectures for single-class and multi-class greenfruit detection in complex orchard environments under label ambiguity, 2025.

[45] Stability AI. SDXL-Turbo 1.0 fp16 model checkpoint. https://huggingface.co/stabilityai/sdxl-turbo/blob/main/sd_xl_turbo_1.0_fp16.safetensors, 2023. Accessed: 2025-07-04.

[46] Unknown Author. Interior scene xl (stable diffusion xl model). https://civitai.com/models/715747/interior-scene-xl, 2025. Accessed: 2025-07-04.

[47] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, 2022.

[48] 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.

[49] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models, 2023.

[50] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision, 2021.

[51] Elad Ben Zaken, Shauli Ravfogel, and Yoav Goldberg. Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models, 2022.

[52] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.

[53] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

8.

补充材料

这个补充部分提供了支持我们工作主要论点和方法的附加细节、可视化和实施细节。其组织结构如下：

8.1.

A. 附加的定性结果 图 7b 展示了我们合成数据集中额外的样本。每一行展示了干净的背景、专家标注的掩码以及最终修复的输出。这些示例展示了我们流水线可以处理的溢出位置和材料（油、水、锈）的多样性。

8.2.

B. 提示工程细节 稳定扩散提示（场景生成）：

- 正："A factory interior, close-up of industrial equipment, image captured via a colored high-quality high-end inspection camera, clear mechanical details, realistic metallic textures, authentic lighting, natural industrial setting, accurate machinery components, subtle equipment variations, realistic wear and tear."
- 负："Text, watermark, low quality, jitter, nsfw, stickers, labels, blurred details, distorted equipment, cartoonish or unrealistic textures, unnatural colors, overly bright lighting, irrelevant objects, human presence, animals, plants, visible text, duplicated or repeated elements, unrealistic proportions, overly polished surfaces, plastic-like or artificial appearance."

修复提示（危险合成）：

- 正 "Realistic oil spill in factory with brown or black stains, industrial scene with dark oil leakage stains, brown-black oily patch on factory floor, factory oil spill with realistic black sludge, realistic factory environment with oil smears, black or brown oil leakage on industrial surface, dirty oil-stained floor in realistic factory, blackened spill area in a manufacturing plant, authentic oil spill marks on brown concrete, industrial realism with black or brown oil spill."
- 负 "Cartoon, anime, illustration, painting, drawing, lowres, blurry, pixelated, overexposed, unrealistic, stylized, clipart, animated, text, watermark, signature, frame, border, extra limbs, distorted hands, shiny, plastic, toy-like, glossy, yellow tint, white overlay, newspaper texture, poster art, human figures, fingers, deformed body parts, 3D render, CGI, artifact, sketch."

VLM 系统提示：如正文详细介绍的那样，系统提示强调了可验证性、物理线索以及避免猜测。对措辞的微小调整（例如，强调“显而易见的危险”）对精确度影响甚微。

8.3.

C. 适应与 LoRA 配置 所有 LoRA 实验都使用以下设置：

- 排名：8
- 缩放因子 α：1/8
- 优化器：AdamW，学习率 5e-5
- 硬件：NVIDIA A100 80GB

标注示例（COCO 风格）：

{
"image_id": 134,
"category_id": 3,
"bbox": [256, 411, 142, 95],
"score": 0.97
}

在表 3 中，我们为 Qwen-7B 跨适应策略提供了更严格的 IoU 阈值下的命中率。这些数据进一步证明，使用 LoRA（特别是在视觉路径上）进行适应的模型在更严格的标准下保持了更精确的定位。

Table 5. Qwen-7B 在不同 IoU 阈值下的命中率

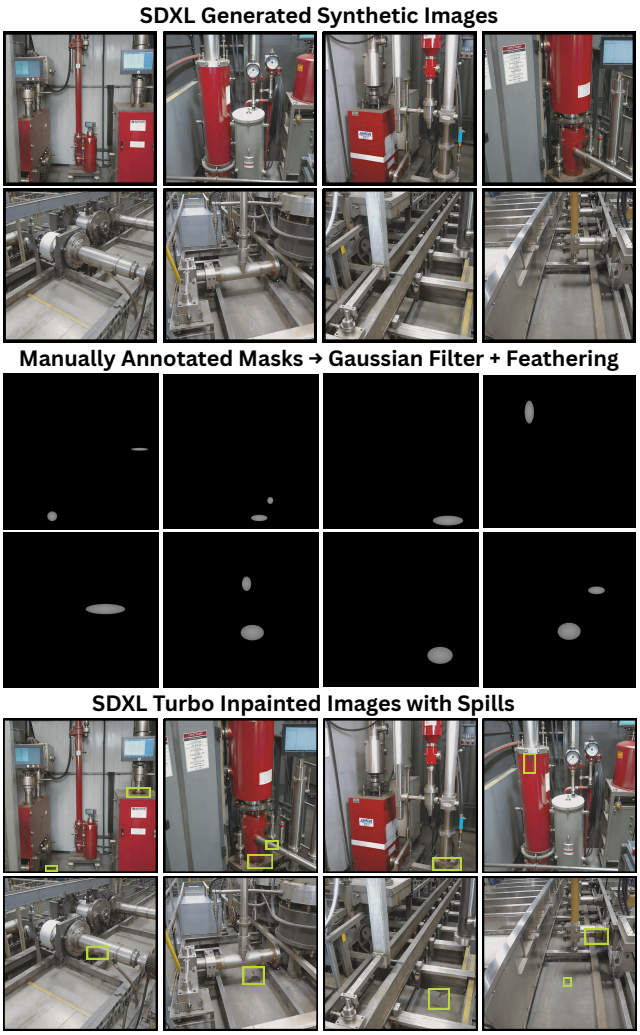
Method	0.5	0.6	0.7	0.8	0.9
Zero-Shot	0.35	0.22	0.15	0.08	0.02
ICL (10-shot)	0.53	0.41	0.29	0.17	0.05
LoRA (V)	0.58	0.49	0.38	0.26	0.10
LoRA (V+L)	0.63	0.54	0.42	0.29	0.13

我们在 CC BY-NC 4.0 许可下发布 SynthSpill 数据集。研究人员可以：

8.4.

F. 图表

(a) 公开（网络抓取）数据集图像



(b) 合成数据图像

Figure 7. 数据集样本