# 通过恢复受损视觉输入提升视觉-语言-动作模型的鲁棒性

Daniel Yezid Guarnizo Orjuela[†]、Leonardo Scappatura[†]、Veronica Di Gennaro[†]、Riccardo Andrea Izzo[‡]、Gianluca Bardaro[‡]和Matteo Matteucci[‡]电子、信息与生物工程系，米兰理工大学，米兰

[†](danielyezid.guarnizo, leonardo.scappatura, veronica.digennaro)@mail.polimi.it
[‡](riccardo.izzo, gianluca.bardaro, matteo.matteucci)@polimi.it

arXiv:2602.01158v1 [cs.CV] 1 Feb 2026

摘要——**视觉-语言-动作（VLA）模型已成为通用机器人操作的主流范式，将感知与控制统一于单一端到端架构中。然而，尽管这类模型在受控环境中表现出色，其对视觉干扰的脆弱性仍严重阻碍了它们在真实世界中的可靠部署。现有文献广泛探讨了由场景几何导致的**物理遮挡**，但一种关键情形仍鲜少受到关注：**图像损坏。**这类传感器级伪影从电子噪声、坏点到镜头污染物不等，会在视觉信号被解读之前直接损害其完整性。在本研究中，我们量化了这一脆弱性，并表明先进的视觉-语言-动作模型（如 $\pi_{0.5}$ 和SmolVLA）在常见信号伪影影响下会遭受灾难性的性能退化，成功率从90%跌至最低2%。为缓解这一问题，我们提出了损坏恢复变换器（CRT），这是一种即插即用、模型无关的视觉变换器，旨在使视觉-语言-动作模型免受传感器干扰。CRT利用对抗训练目标，从受损输入中恢复干净观测，无需对底层模型进行计算开销高昂的微调。在LIBERO和Meta-World基准测试上开展的大量实验表明，CRT能够有效恢复损失的性能，使视觉-语言-动作模型即使在严重视觉损坏下也能保持接近基线的成功率。**

索引词—**视觉-语言-动作模型、图像损坏、视觉变换器、对抗训练**

## I. 引言

近年来，视觉-语言-动作（VLA）模型推动了机器人控制任务的发展。这些系统将感知、语言理解和动作预测整合到单一架构中，使机器人能够根据视觉观测和自然语言指令执行复杂的操作任务。

视觉-语言-动作模型有望为众多领域带来重大进步，涵盖工业自动化到医疗领域。然而，尽管这些模型在受控研究环境中展现出令人瞩目的能力，要在现实世界中广泛部署仍面临挑战，原因在于存在显著的可靠性问题。这些问题包括：在满足实时推理约束的同时，确保系统在动态且非结构化环境中的安全性和鲁棒性。

要让视觉-语言-动作模型能够在真实世界中部署，就必须具备环境鲁棒性 [1], ，也就是在输入流出现不可预测的变化时，仍能保持稳定且准确性能的能力。真实世界环境本质上十分复杂，面临诸多挑战，例如光照不稳定、天气状况恶劣，或目标物体被障碍物实体遮挡。

当前研究积极应对场景中的物理遮挡问题，通常通过主动感知或多视角推理来解决 [2], [3]。然而，另一种独特且同样关键的视觉退化来源仍未得到充分研究：图像损坏。与物理障碍物不同，这类问题是伪影，包括水滴、坏点和电子干扰等，会在视觉信号被视觉-语言-动作模型解读之前直接损害其完整性。现有研究表明，在干净且无遮挡的视觉数据上训练的标准视觉-语言-动作模型，在关键特征被遮蔽时会出现显著的性能退化，在噪声条件下性能下降 20–30% [1]。

尽管如此，当前文献仍缺乏对这类图像损坏伪影对现代视觉-语言-动作模型端到端成功率影响程度的系统量化分析。

本文通过两项贡献弥补了这一空白。首先，我们对最先进的视觉-语言-动作模型（具体而言是 SmolVLA 和 $\pi_{0.5}$）在多种图像损坏情境下的性能退化进行了严谨的定量分析。我们的研究结果证实，若不采取缓解措施，这些伪影会导致性能显著下降。

其次，我们提出 **损坏修复 Transformer（CRT）**来应对这些脆弱性。我们采用了 [4],提出的高效基于Transformer的架构，该架构最初用于图像重建，并将其改用于高分辨率仿真环境中的机器人学领域。我们将其设计为位于视觉-语言-动作模型上游的模块化单元。这种方法具有战略优势：无需从头重新训练或微调视觉-语言-动作模型（这在计算上将难以承受），CRT 可作为专用的轻量级模块，在观测到达策略网络之前对其进行修复。

CRT 的集成有效缩小了基于干净输入和受损输入生成的动作之间的差异，从而有效消除了图像损坏对视觉-语言-动作模型的影响。为便于复现，我们公开了代码，

模型和数据集[1]。

## II. 相关工作

### A. 视觉-语言-动作模型（视觉-语言-动作模型）

视觉-语言-动作（VLA）模型以大语言模型（LLM）和视觉-语言模型（VLM）的架构基础为依托，将这些能力扩展至物理领域。这由此形成端到端系统，使其能够进行推理并直接执行动作[1], [5]。视觉-语言-动作模型基于互联网规模的多模态数据以及与之配对的机器人轨迹进行训练，例如 OpenX-Embodiment 数据集 [6]。RT-2 [7], 等开创性研究表明，可以直接对视觉-语言模型进行微调，使其通过将连续运动转换为离散文本词元来输出机器人动作。这样一来，模型便能利用大语言模型内在的语义推理能力，实现对新物体和新指令的零样本泛化。在此基础上，OpenVLA[8]推出了一个拥有 70 亿参数的开源架构，将稳健的视觉编码器（如 DINOv2 [9] 和 SigLIP[10]）与 Llama-2 [11] 语言骨干网络相结合，通过离散动作词元预测，为通用机器人操作确立了稳健的基线。

尽管取得了成功，将连续的运动控制视为离散令牌预测任务，往往会损害操作所需的流畅性和精确度。因此，近期的架构已转向采用基于扩散或流匹配的头部，以在连续空间中表示多模态动作分布。例如， $\pi_0$ [12]采用流匹配 [13] 来预测连续动作块，而 SmolVLA [14]则将交叉注意力层与流匹配动作专家相结合。尽管这些模型在语义泛化和任务执行方面表现出色，但其效能高度依赖视觉编码器特征空间的完整性。由于这些模型主要在精选数据集上训练，因此遇到视觉损坏时，性能往往会显著下降，这凸显了其鲁棒性的严重不足。

### B. 机器人感知中的视觉退化

视觉退化是可靠机器人控制面临的持续挑战。我们根据这些挑战的起因将其分为两类：物理遮挡 （例如，依赖于场景）和图像损坏（例如，依赖于传感器）。

**物理遮挡**：在非结构化的真实世界环境中，目标物体常常被其他障碍物挡住。文献中主流的解决方案是主动感知，即机器人自主调整视点以获取缺失的信息。VISO-Grasp [15]和下一最佳视点（NBV）规划器 [16]等近期框架利用视觉-语言模型推理空间关系并识别遮挡物，从而引导相机移动到能够显露目标的有利视点。类似地，WoW [17] 等模型通过空间推理来推断并移除障碍物。这些方法依赖于一个前提：障碍物是外部因素；通过改变传感器相对于场景的位姿，可以通过多视图融合或场景重新定位显露目标。

**图像损坏**：与物理遮挡不同，基于传感器的损坏源于感知硬件或信号传输流程中固有的伪影。这些问题包括水滴、裂纹、坏点或电子干扰。尽管现有计算机视觉文献为类似问题提供了解决方案 [18]–[20], ，但这些方案很少被应用于端到端机器人系统。关键在于，适用于物理遮挡的标准技术在这种情况下并不起作用。主动感知也会失效，因为损坏（例如镜头上的污点）固定在相机画面中，无论机器人的位置如何，都会持续遮挡部分视野。鉴于标准视觉-语言-动作模型在这些情况下的性能会下降 20–30% [1],, 亟需一种模块化修复方法，在图像观测到达策略网络之前对其进行清理。

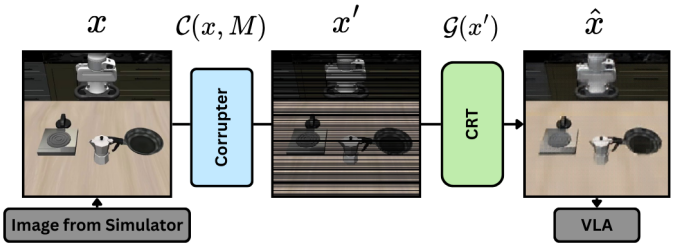



图1. **系统流程**。 CRT模块置于视觉-语言-动作模型上游。它会拦截受损的仿真器观测 $x'$，将其修复为$\hat{x}$，并将干净的估计结果传递给策略网络，用于动作预测。

## III. 方法

### A. 总体架构

为解决图像损坏问题，我们在标准机器人控制流程中引入了一个专用的修复模块——**损坏修复变换器（CRT）**。借助 LeRobot库 [5],, 我们将修复模块置于视觉-语言-动作模型的正前方。在此配置中，从相机获取的每个原始观测帧都会先经过处理以减少损坏，然后再输入策略网络

令 $x \in \mathbb{R}^{H\times W\times 3}$ 表示来自机器人相机的干净观测帧。其空间尺寸 $H\times W$ 根据仿真器环境进行调整：LIBERO [21]基准测试的分辨率为 $360\times 360$，Meta-World [22]的分辨率为 $480\times 480$。

我们定义一个损坏过程 $\mathcal{C}$，使观测输入表示为 $x' = \mathcal{C}(x,\ M)$，其中 $M$ 表示施加于干净帧的具体伪影类型。如图 1 所示，CRT 的目标是逆转这一过程，从损坏的输入 $x'$ 中恢复原始信号 $x$ ，确保视觉-语言-动作模型基于修复后的观测运行。

### B. 损坏修复*Transformer*（*CRT*）

我们提出了损坏恢复变换器（CRT）$\mathcal{G}$，这是一种基于Transformer的重建模型，经过训练可将受损的图像观测恢复为其原始版本。

[1]https://huggingface.co/collections/AIRLab-POLIMI/ corruption-restoration-transformer

为此，我们借鉴了 [4]，提出的图像恢复框架，并针对机器人感知领域对该方法进行了专门设计。该架构能够以较高的计算效率处理密集预测任务，这与目标领域的需求相契合。我们确定了三项关键的架构机制，以确保CRT具备视觉-语言-动作模型所需的精度和效率：

- **移位补丁标记化 (SPT)**：用于保留相邻像素之间的空间关系。SPT 直接解决了标准视觉 Transformer (ViT) [23] 缺乏局部归纳偏置的问题。
- **旋转位置嵌入 (RoPE)**：用于编码绝对位置和相对位置，使模型能够泛化到动作预测中使用的更长动作序列 [24]。
- **局部自注意力 (LSA)**：用于强化注意力分数，并聚焦于局部纹理细节。

来自 [4] 的基线架构最初是为低分辨率（$64 \times 64$ 像素）任务设计的，而我们的应用需要处理包含密集语义信息的更高分辨率视觉输入（最高达 $480 \times 480$ 像素）。因此，必须显著扩展模型容量，以满足视觉-语言-动作模型的要求。我们增大了潜向量的尺寸，以保留对视觉-语言-动作模型执行机器人操作至关重要、但很容易被损坏伪影抹去的精细纹理和物体细节。此外，我们还增加了 Transformer 块的数量，从而增加了堆叠的编码器层数。在 Transformer 中，较早的层通常处理局部模式（边缘、纹理），而较深的层则将这些信息汇总为全局语义概念。增加深度为模型提供了必要的容量，以便对整幅图像中损坏伪影的复杂统计特性进行建模。最后，为了配合深度的增加，我们也增加了注意力头的数量。这使模型能够并行推理多种关系（例如，区分场景边缘与损坏线条）。

### C. 对抗训练

虽然标准的逐像素重建损失（例如 $L_1$或均方误差（MSE））能够有效恢复低频全局结构，但在处理高频细节时往往会导致输出模糊。在机器人操作中，保留这些细节对于下游视觉-语言-动作模型正确理解物体可供性和边界至关重要。为此，我们采用并扩展了 [4]，提出的对抗训练框架，将 CRT 模块视为生成对抗网络中的生成器 $\mathcal{G}$。图 2 概述了我们的方法。

我们引入一个判别器网络 $\mathcal{D}$，用于区分真实干净观测 $x$ 与重建观测 $\hat{x} = \mathcal{G}(x')$。判别器的架构与生成器基于 Transformer 的设计相仿，但针对分类任务进行了优化。它通过补丁嵌入层和一系列 Transformer 块处理输入，

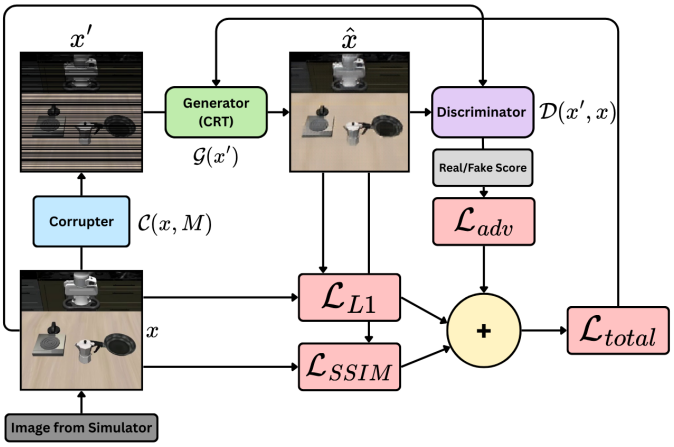



图 2. **CRT 对抗训练**。CRT 生成器（$\mathcal{G}$）接收受损输入 $x'$，并生成重建后的观测$\hat{x}$。网络通过多目标损失函数进行优化，该函数结合了逐像素保真度$\mathcal{L}_{L1}$、结构相似性 $\mathcal{L}_{SSIM}$以及来自判别器 $\mathcal{D}$ 的对抗反馈 $\mathcal{L}_{adv}$，以确保有效重建。

并通过最终的多层感知机（MLP）头输出有效性分数。训练目标采用多目标优化策略，结合了三个不同损失分量的加权组合：

**对抗损失**：为增强感知真实感，我们采用标准的二元交叉熵目标。判别器旨在最大化正确分类真实图像和重建图像的概率，而生成器则旨在最小化 $\mathcal{D}$ 将 $\hat{x}$ 识别为伪造图像的概率：

$$\mathcal{L}_{adv} = \mathbb{E}_x[\log \mathcal{D}(x)] + \mathbb{E}_{x'}[\log(1 - \mathcal{D}(\mathcal{G}(x')))] \quad (1)$$

**重建损失**：为确保重建图像忠实保留真实内容，我们采用逐像素的 $L_1$ 损失，与 $L_2$ [25]相比，该损失能够促进稀疏性并减少模糊：

$$\mathcal{L}_{L1} = \mathbb{E}_{x,x'}[\|x - \mathcal{G}(x')\|_1] \quad (2)$$

**结构相似性指数（SSIM）损失** [26]：用于保留亮度、对比度和结构等结构信息：

$$\mathcal{L}_{SSIM} = 1 - \text{SSIM}(x, \mathcal{G}(x')) \quad (3)$$

生成器 $G$ 的最终目标函数是这些项的加权和：

$$\mathcal{L}_{total} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{SSIM}\mathcal{L}_{SSIM} + \lambda_{adv}\mathcal{L}^G_{adv} \quad (4)$$

其中，$\mathcal{L}^G_{adv}$ 表示对抗损失中生成器对应的部分。经过经验优化，我们将超参数设为 $\lambda_{L1} = 10.0$、$\lambda_{SSIM} = 1.0$ 和 $\lambda_{adv} = 0.05$。对抗项的权重相对较低，这确保训练仍以内容重建为基础，同时提供足够的梯度信号，以优化纹理细节并消除伪影。

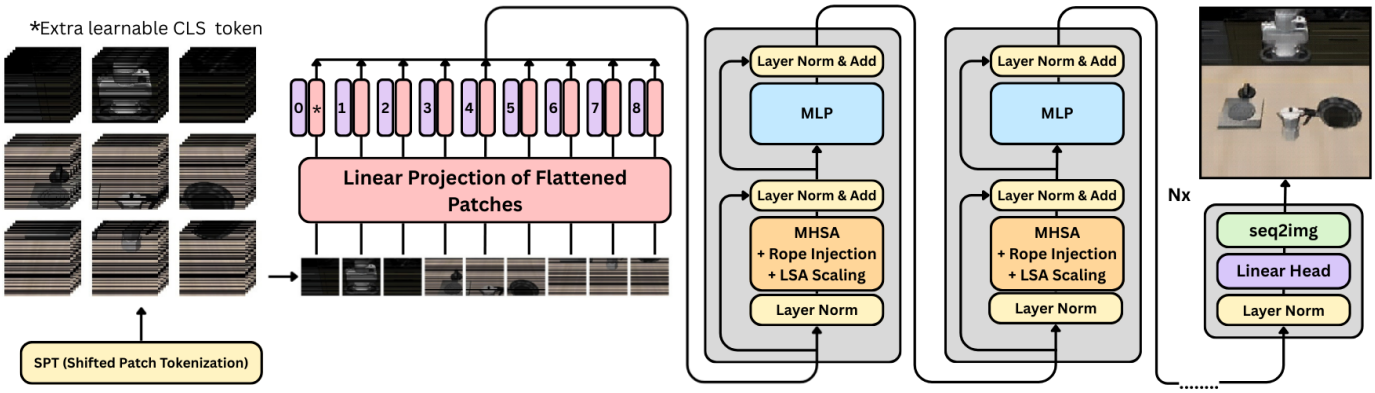



图 3. **损坏恢复变换器（CRT）的架构**该模型采用了四种专门机制：1）移位补丁标记化（SPT），用于恢复局部空间依赖关系；2）旋转位置嵌入（RoPE），用于实现稳健的相对空间编码；3）局部自注意力（LSA），用于增强纹理细节。深层 Transformer 主干网络使模型能够将复杂的损坏伪影与底层语义场景解耦，最终将潜在标记重新塑形为修复后的 RGB 观测。

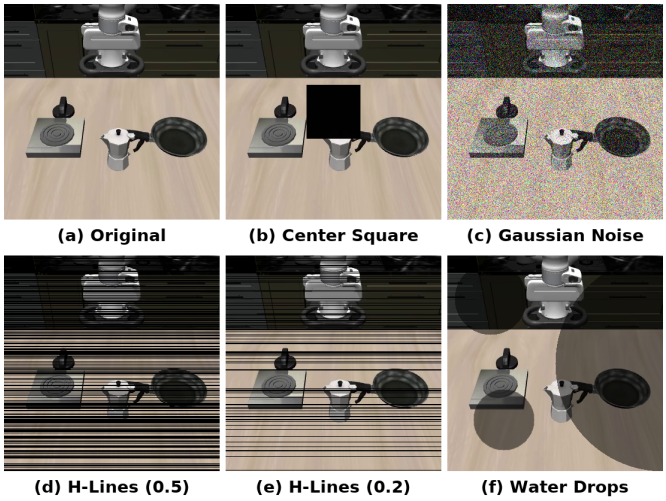



图 4. **视觉损坏类型**。 从左上方开始：(a) 无损基线，(b) 居中方块，(c) 高斯噪声，(d) 水平线（0.5），(e) 水平线（0.2），(f) 水滴。

### *D.* 数据集

在 CRT 训练中，我们基于 LIBERO-10 和 Meta-World MT50 基准测试中的成功执行轨迹，构建了两个不同的配对数据集。我们为每个基准测试分别训练了专用的重建模型，确保 CRT 能够适应各个仿真环境不同的视觉和语义分布。

为了构建真实值参考数据集，我们利用预训练视觉语言动作模型作为专家示范者。对于 LIBERO-10 和 Meta-World MT50 基准测试中定义的每项任务，我们都在无噪声环境中运行相应的视觉语言动作策略。我们严格筛选这些策略执行轨迹，仅保留智能体成功达到目标状态的轨迹。为确保视觉语义分布均衡，我们为每项不同的任务恰好采样一条成功轨迹（即 LIBERO 取 10 条，Meta-World 取 50 条）。

表 I**训练超参数**。 我们报告了使用单张 NVIDIA RTX QUADRO6000 训练 CRT 时采用的超参数。

| **参数** | **LIBERO** | **Meta-World** |
|---|---|---|
| 训练轮数 | 30 | 37 |
| 学习率 | 1e-4 | 7e-4 |
| 批大小 | 12 | 8 |
| 累积步数 | 12 | 32 |
| 训练时间 | 1天 8小时 14分钟 | 2天 22小时 36分钟 |

我们将这些轨迹分解为单帧，构成真值观测集 $x$，同时保留原生仿真分辨率（即 LIBERO 为 $360 \times 360$，Meta-World 为 $480 \times 480$）。随后，对于每个提取出的干净帧 $x$，我们应用第 IV-A 节中所述的损坏生成过程$\mathcal{C}(x, M)$，生成对应的受损 $x'$。这一流程生成了完全对齐的 $(x', x)$ 配对数据，可用于监督训练。

最终得到的数据集包含 48,660 对图像：其中 27,310 帧来自 LIBERO-10，21,350 帧来自Meta-World MT50。每个数据集均按 80/20 的比例划分为训练集和验证集。训练超参数列于表 I 中。

## IV. 实验与结果

我们使用两个广受认可的机器人操作基准评估了CRT模块的有效性：LIBERO和Meta-World。这些环境提供了多样化的任务复杂度与语义变化，而它们与LeRobot库的集成则确保了可复现性的标准化框架。我们将我们的流程与两种最先进的视觉-语言-动作模型进行了集成：SmolVLA在LIBERO和Meta-World任务套件上均进行了评估，而 $\pi_{0.5}$ 仅在LIBERO套件上进行了评估，因为在本研究进行时尚未获得适用于Meta-World的微调检查点。

### *A.* 图像损坏

为了评估视觉-语言-动作模型对传感器级干扰的鲁棒性，我们选取了图4所示的五种不同伪影

表 II定量 分析 CRT 在 LIBERO-10 和 META-WORLDMT50 基准测试。 干净与受损观测下的平均成功率（SR）及相对性能退化（Δ）对比。

| 基准测试 | 模型 | 基线 平均成功率 (↑) | Δ | 居中方块 平均成功率 (↑) | Δ | 高斯噪声 平均成功率 (↑) | Δ | 水平线 (0.5) 平均成功率（↑） | Δ | 水平线 (0.2) 平均成功率（↑） | Δ | 水滴 平均成功率（↑） | Δ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LIBERO-10 | $\pi_{0.5}$ | **90.0**% | – | 65.0% | -27.78% | **87.0**% | -3.33% | 2.0% | -97.78% | 62.0% | -31.11% | 68.0% | -24.44% |
| | $\pi_{0.5}$ +**CRT** | 89.0% | -1.11% | **89.0**% | -1.11% | 79.0% | -12.22% | **87.0**% | -3.33% | **89.0**% | -1.11% | **87.0**% | -3.33% |
| | SmolVLA | **43.0**% | – | **16.0**% | -62.79% | **13.0**% | -69.77% | 0.0% | -100% | 1.0% | -97.67% | 11.0% | -74.42% |
| | SmolVLA +**CRT** | 33.0% | -23.26% | 11.0% | -74.42% | 4.0% | -90.70% | **3.0**% | -93.02% | **15.0**% | -65.12% | **15.0**% | -65.12% |
| Meta-World | SmolVLA | **58.0**% | – | 33.0% | -43.10% | 27.6% | -52.41% | 20.6% | -64.48% | 20.0% | -65.52% | 40.0% | -31.03% |
| | SmolVLA +**CRT** | 47.0% | -18.96% | **40.6**% | -30.0% | **34.6**% | -40.34% | **32.2**% | -44.48% | **41.4**% | -28.62% | **43.34**% | -25.17% |

类型。这些伪影旨在模拟现实世界机器人部署中常见的特定硬件故障、环境条件和信号传输错误。

- **居中方块**：在大多数操作任务中，工作空间和主要关注对象位于画面中心。我们通过在图像中心施加一个黑色方块来模拟灾难性传感器故障，从而挑战模型，并破坏主要交互区域。

- **高斯噪声**：这模拟了低照度条件、高ISO设置或热干扰下典型的电子传感器噪声。我们向每个像素注入标准差为0.20的零均值高斯噪声，从而降低细粒度视觉特征。

- **水平线**：表示来自隔行扫描视频传输、滚动快门效应或电磁干扰的扫描线伪影。黑色水平线被随机放置在图像上，破坏场景的20%（低强度）或50%（高强度）。

- **水滴**：再现由液体飞溅、冷凝或湿度造成的镜头污染。这些情况在工业或户外环境中经常遇到。在多个随机位置施加带有高斯模糊的半透明圆形区域。

*B.* 结果

我们采用三阶段评估方案，在多种操作任务中验证 CRT 模块的有效性：

1) 首先，我们在干净、未受损的观测上评估 SmolVLA 和 $\pi_{0.5}$ ，以建立基线并确认仿真环境的有效性。2) 随后，我们引入损坏过程 $\mathcal{C}(x,\ M)$，分别施加五种损坏类型中的每一种。我们记录了各模型在未采用任何缓解策略时的性能下降。3) 最后，我们将 CRT 模块集成在视觉-语言-动作模型的上游。在相同的损坏条件下重复评估，以衡量系统恢复性能的能力。

我们的评估结果总结于表 II。分析重点关注三个关键方面：视觉-语言-动作模型对损坏的抵御能力、CRT 的有效性，以及不同视觉-语言-动作模型对恢复过程的敏感程度。

**视觉-语言-动作模型易受图像损坏影响**：与以往研究一致，我们的结果证实，最先进的视觉-语言-动作模型对传感器伪影高度敏感。如表 II 所示，在图像损坏的情况下，性能会显著下降。例如， $\pi_{0.5}$ 在 LIBERO-10 上的成功率在强度为 0.5 的水平线损坏下，从 90% 降至仅 2%。类似地，SmolVLA 在 Meta-World MT50 上的表现，在相同条件下从 58% 减半至 20.6%。这些结果证实了这样一种观点：几何泛化（例如主动感知）不足以应对底层视觉特征遭到结构性破坏的情况。

**CRT 的有效性**：集成 CRT 后，性能显著恢复，尤其是对于 $\pi_{0.5}$ 架构。如表 II 所示，配备 CRT 的 $\pi_{0.5}$ 几乎完全恢复了性能。在最严重的“水平线 (0.5)”损坏下，成功率从 2% 回升至 87%，基本达到无损基线水平。“水滴”（68% 升至 87%）和“高斯噪声”（87% 升至 79%）下也观察到类似的恢复效果。对于 Meta-World 上的 SmolVLA，CRT 在所有损坏类型下的表现均始终优于未采用缓解措施的模型。值得注意的是，CRT 将“水平线”下的成功率提高了 11% 以上，从 20.6% 升至 32.2%；对于“居中方块”遮挡也表现出较强的韧性，成功率提升了 7.6%。

**基线保持与权衡**：CRT的一项关键要求，是在已有的干净观测上保持性能。对于 $\pi_{0.5}$ 模型，CRT几乎不会带来额外开销，在干净数据上的成功率为89%，接近90%的基线。然而，对于较轻量的SmolVLA架构，我们观察到了性能权衡：在Meta-World上，加入CRT后，基线性能从58%降至47%。这表明，尽管CRT能够成功恢复全局语义结构，但较小的视觉-语言-动作模型对特征扰动的鲁棒性较弱，也缺乏在重建过程中对细微分布偏移进行泛化的能力。尽管如此，在不利条件下，CRT带来的可靠性提升十分显著，远远超过了这一权衡。此外，CRT模块是一种轻量级神经网络，对机器人控制回路的额外开销极小。实测表明，CRT仅需10–50毫秒的推理时间和约1 GB显存。相比之下，$\pi_0$ 或SmolVLA等视觉-语言-动作模型需要更多计算资源，令牌生成的延迟通常达数百毫秒。

因此，CRT在几乎不增加额外开销的同时，提升了底层模型的鲁棒性。

## V. 结论

在这项工作中，我们发现当前视觉-语言-动作（VLA）模型存在一个关键弱点：对图像损坏缺乏鲁棒性。以往研究积极通过主动感知应对物理遮挡，而我们证明，镜头污染物、电子噪声和信号丢失等传感器级伪影需要从根本上采取不同的方法。我们的系统性评估表明，即使是 $\pi_{0.5}$ 和 SmolVLA 这样的最先进模型，在受到这些干扰时也会遭遇灾难性失效；在严重情况下，成功率下降近 90%。为弥补这一不足，我们提出了损坏修复 Transformer（CRT），这是一种专用模块，可在视觉-语言-动作模型处理输入之前恢复视觉信息的完整性。CRT 结合了先进的架构机制（如 SPT、RoPE、LSA）与对抗训练目标，能够从受损输入中有效重建干净观测，无需对底层视觉-语言-动作模型进行计算成本极高的微调。在 LIBERO 和 Meta-World 基准测试上的实验结果证实，CRT 能够有效应对传感器伪影。对于 $\pi_{0.5}$ 这样的大型视觉-语言-动作模型，CRT 可充当近乎完美的重建层，即使在严重损坏的情况下，也能将性能恢复到与无损基线相差仅 1–3% 的水平。我们还观察到，SmolVLA 这类较小的架构对重建过程引入的细微分布偏移更为敏感，这凸显了恢复效果与输入保真度之间的权衡。这一局限为未来研究指明了明确方向。鉴于 CRT 的模块化特性，整个流程可以调整为仅在检测到损坏时才启动恢复过程。这样既能保持视觉-语言-动作模型原有的性能，也能弥补图像损坏造成的问题。总而言之，CRT 为构建更鲁棒的视觉-语言-动作模型提供了一条可扩展的途径，使其能够部署于不可预测且存在图像损坏的真实世界环境中。

## 参考文献


[1] R. Sapkota、Y. Cao、K. I. Roumeliotis 和 M. Karkee，《视觉-语言-动作模型：概念、进展、应用与挑战》，*arXiv*预印本 *arXiv:2505.04769*，2025年。[2] S. Wang 和 X. Cao，《Peafowl：面向双臂操作的感知增强型多视角视觉-语言-动作模型》，载于未知，2026年。[在线]。可访问：https://arxiv.org/pdf/2601.17885.pdf [3] J. Liu、Y. Qi、J. Zhang、M. Li、S. Wang、K. Wu、H. Ye、H. Zhang、Z. Chen、F. Zhong 等，《Trackvla++：释放视觉-语言-动作模型在具身视觉跟踪中的推理与记忆能力》，*arXiv*预印本 *arXiv:2510.07134*，2025年。[4] N. Verma、D. Kaur 和 L. Chau，《使用增强型视觉变换器进行图像重建》，*arXiv*预印本 *arXiv:2307.05616*，2023年。[5] R. Cadene、S. Alibert、A. Soare、Q. Gallouedec、A. Zouitine 和 T. Wolf，《Lerobot：基于 PyTorch 的真实世界机器人学最先进机器学习》，2024年。[6] A. O'Neill、A. Rehman、A. Maddukuri、A. Gupta、A. Padalkar、A. Lee、A. Pooley、A. Gupta、A. Mandlekar、A. Jain 等，《Open X-Embodiment：机器人学习数据集与 RT-X 模型：Open X-Embodiment 协作项目 0》，载于*2024 IEEE*国际机器人与自动化会议（*ICRA*）。IEEE，2024年，第6892–6903页。[7] B. Zitkovich、T. Yu、S. Xu、P. Xu、T. Xiao、F. Xia、J. Wu、P. Wohlhart、S. Welker、A. Wahid 等，《RT-2：视觉-语言-动作模型将网络知识迁移至机器人控制》，载于机器人学习会议。PMLR，2023年，第2165–2183页。[8] M. J. Kim、K. Pertsch、S. Karamcheti、T. Xiao、A. Balakrishna、S. Nair、R. Rafailov、E. P. Foster、P. R. Sanketi、Q. Vuong 等，《OpenVLA：开源视觉-语言-动作模型》，载于机器人学习会议。PMLR，2025年，第2679–2713页。[9] M. Oquab、T. Darcet、T. Moutakanni、H. Vo、M. Szafraniec、V. Khalidov、P. Fernandez、D. Haziza、F. Massa、A. El-Nouby 等，《DINOv2：无需监督地学习鲁棒视觉特征》，arXiv预印本 arXiv:2304.07193，2023年。[10] X. Zhai、B. Mustafa、A. Kolesnikov 和 L. Beyer，《用于语言-图像预训练的 Sigmoid 损失》，载于 IEEE/CVF国际计算机视觉会议论文集，2023年，第11 975–11 986页。[11] H. Touvron、L. Martin、K. Stone、P. Albert、A. Almahairi、Y. Babaei、N. Bashlykov、S. Batra、P. Bhargava、S. Bhosale 等，《Llama 2：开放基础模型与微调聊天模型》，arXiv预印本 arXiv:2307.09288，2023年。[12] K. Black、N. Brown、D. Driess、A. Esmail、M. Equi、C. Finn、N. Fusai、L. Groom、K. Hausman、B. Ichter 等，《$\pi$ 0：用于通用机器人控制的视觉-语言-动作流模型》，arXiv预印本 arXiv:2410 24164，2024年。.[13] Y. Lipman、R. T. Chen、H. Ben-Hamu、M. Nickel 和 M. Le，《用于生成建模的流匹配》，arXiv 预印本 arXiv:2210 02747，2022年。.[14] M. Shukor、D. Aubakirova、F. Capuano、P. Kooijmans、S. Palma、A. Zouitine、M. Aractingi、C. Pascal、M. Russi、A. Marafioti 等，《SmolVLA：面向经济高效的机器人技术的视觉-语言-动作模型》，arXiv预印本 arXiv:2506.01844，2025年。[15] Y. Shi、D. Wen、G. Chen、E. Welte、S. Liu、K. Peng、R. Stiefelhagen 和 R. Rayyes，《VISO-Grasp：视觉-语言引导的空间物体中心六自由度主动视角规划与抓取，适用于杂乱环境和不可见性场景》，arXiv预印本 arXiv:2503.12609，2025年。[16] Z. Jia、Y. Li、Q. Hao 和 S. Zhang，《Pb-nbv：用于未知物体重建的高效投影式下一最佳视点规划框架》，IEEE机器人与自动化快报，2025年。[17] X. Chi、P. Jia、C.-K. Fan、X. Ju、W. Mi、K. Zhang、Z. Qin、W. Tian、K. Ge、H. Li 等，《WoW：通过具身交互迈向全知世界模型》，arXiv预印本 arXiv:2509.22642，2025年。[18] H. Xu、A. A. Baniya、S. Wells、M. R. Bouadjenek、R. Dazeley 和 S. Aryal，《Totnet：面向体育视频中鲁棒球体检测的遮挡感知时序跟踪》，计算机视觉与图像理解，第104657页，2026年。[19] Q. Li 和 Y. Cao，《透过模糊玻璃观察：单次拍摄去除遮挡》，arXiv预印本 arXiv:2509.01033，2025年。[20] R. Mallick、S. Dong、N. Ruiz 和 S. A. Bargal，《D-feat遮挡：利用扩散特征提高物体识别对局部视觉遮挡的鲁棒性》，载于计算机视觉与模式识别会议论文集，2025年，第1722–1731页。[21] B. Liu、Y. Zhu、C. Gao、Y. Feng、Q. Liu、Y. Zhu 和 P. Stone，《LIBERO：终身机器人学习知识迁移基准测试》，神经信息处理系统进展，第36卷，44 791，2023年，44 776页。– .[22] T. Yu、D. Quillen、Z. He、R. Julian、K. Hausman、C. Finn 和 S. Levine，《Meta-World：多任务与元强化学习的基准测试和评估》，载于机器人学习会议。PMLR，2020年，第1094–1100页。[23] G. J. Ferdous、K. A. Sathi、M. A. Hossain 和 M. A. A. Dewan，《Spt-swin：用于图像分类的移位补丁标记化 Swin Transformer》，IEEE Access，2024年。[24] J. Su、M. Ahmed、Y. Lu、S. Pan、W. Bo 和 Y. Liu，《Roformer：采用旋转位置嵌入的增强型 Transformer》，Neurocomputing，第568卷，第127063页，2024年。,[25] H. Zhao、O. Gallo、I. Frosio 和 J. Kautz，《使用神经网络进行图像复原的损失函数》，IEEE计算成像汇刊，第3卷，第1期，第47–57页，2016年。[26] D. Brunet、E. R. Vrscay 和 Z. Wang，《论结构相似性指数的数学性质》，IEEE图像处理汇刊，第21卷，第4期，第1488–1499页，2011年。