/TemplateVersion (2026.1)

ReconVLA：重构视觉-语言-动作模型作为有效的机器人感知器

Wenxuan Song[1], Ziyang Zhou[1], Han Zhao[2,3], Jiayi Chen[1], Pengxiang Ding[2,3], Haodong Yan[1], Yuxin Huang[1], Feilong Tang[4], Donglin Wang[2], Haoang Li[1]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]Westlake University [3]Zhejiang University [4]Monash University

Abstract

最近在视觉-语言-动作（Vision-Language-Action，VLA）模型方面的进展使得机器人代理能够将多模态理解与动作执行相结合。然而，我们的实证分析揭示，目前的 VLA 在将视觉注意力分配到目标区域时存在困难。相反，视觉注意力总是分散的。为了引导视觉注意力正确地定位到目标上，我们提出了 ReconVLA，一种具有隐式定位范式的重建型 VLA 模型。该模型基于视觉输出，一个扩散变压器旨在重建图像的凝视区域，该区域对应被操作的目标物体。这个过程促使 VLA 模型学习细粒度的表示，并精确分配视觉注意力，从而有效利用任务特定的视觉信息并进行精确操控。此外，我们整理了一个大规模预训练数据集，包括从开源机器人数据集中超过 10 万条轨迹和 200 万数据样本，进一步提升模型在视觉重建中的泛化能力。模拟和现实世界中的大量实验展示了我们隐式定位方法的优越性，显示了其精确操控和泛化的能力。我们的项目页面是 https://zionchow.github.io/ReconVLA/ 。

1 介绍

在视觉-语言模型（VLMs）(Awadalla et al. 2023; Liu et al. 2024b) 上最近的进展展示了其有效弥合感知与语言模态之间差距的潜力。基于这些进展，视觉-语言-动作（VLA）模型 (Brohan et al. 2023; Zitkovich et al. 2023; Octo Model Team et al. 2024; Niu et al. 2024; Song et al. 2024; Kim et al. 2024) 通过整合多模态理解，将这种能力扩展到了动作执行。得益于数十亿参数和在大规模机器人数据集 (O' Neill et al. 2024; Fang et al. 2024) 上的预训练，这些模型在实现可推广技能方面显示出希望。准确的视觉定位对于实现 VLA 的精确抓取尤为重要，特别是在杂乱环境和长时间任务中。为了分析预测动作过程中的视觉定位行为，我们可视化了视觉输入的注意力图。结果显示，传统的 VLA 模型通常表现出分散的视觉注意（Figure 2 第 1 行），未能精确聚焦于目标对象，这可能进一步导致操作错误的对象。该发现提出了一个关键问题：VLA 模型如何能够优化视觉注意分配并进一步提高视觉定位能力？

先前的视觉指向方法针对 VLAs 通常以一种链式思维（CoT）方式显式输入标记图像 (Huang et al. 2025; Li et al. 2025) 或输出边界框 (Zawalski et al. 2024; Deng et al. 2025) 。这些方法增强了目标区域的感知和提高空间意识，但并未从根本上改进注意力分配。受重建型视觉指令调优 (Wang et al. 2024) 的启发，我们引入了一个辅助的视觉重建模块，该模块实现为一个轻量级扩散变压器 (Peebles and Xie 2022) 。该模块以 VLA 模型的视觉输出为条件，旨在从噪声中重建目标操控区域。这个过程促使 VLA 模型学习具有特定区域信息的细粒度表示，从而将视觉注意力集中在正确的区域。如 Figure 1 所示，该机制类似于人眼的注视行为，眼睛感知一个小而集中的区域时带有清晰的锐度，而周围区域则保持模糊 (Stewart 2020) 。因此，目标操控区域被命名为注视区域。

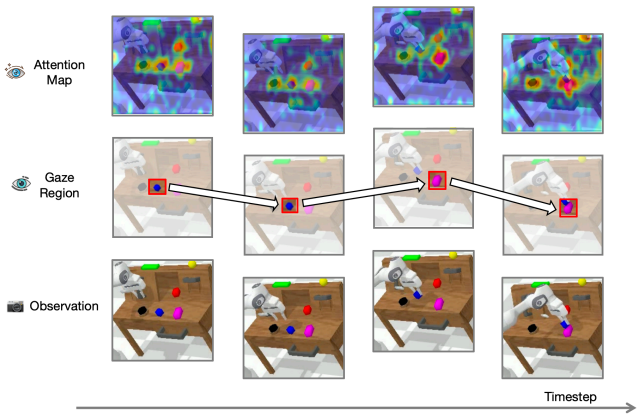

Figure 1: 观察、凝视区域和注意力图的可视化。对于一个要求机械臂将蓝色积木抬起并放到粉色积木上的长时间视野任务 "堆积积木"。虽然存在多个干扰因素，我们的模型自适应地调整凝视区域，引导视觉注意力分配到正确的目标。通过准确的视觉定位，它依次操作不同的目标物体，成功完成任务。

然而，与其 VLM 主干 (Liu et al. 2024b) 类似，传统的 VLA 模型在视觉-语言理解任务上进行了微调，并以自回归的方式生成动作，缺乏视觉生成能力。为了克服这一限制，我们策划了一个包含超过 10 万条轨迹和 200 万个数据样本的预训练数据集。我们选择了多个开源机器人数据集 (Walke et al. 2024; Liu et al. 2024a; Mees et al. 2021) ，并设计了一种基于 Grounding DINO (Liu et al. 2024c) 的自动数据处理，以生成成对的完整图像和目标操控区域的图像。在这个大规模数据集上的预训练显著增强了模型在视觉生成中的泛化能力。

通过利用上述技术，我们开发了重构视觉-语言-动作模型（ReconVLA）。它将当前图像、语言指令和机器人本体感知作为输入。在训练期间，输入图像的注视区域

通过一个冻结的视觉分词器被处理成潜在标记，该分词器保留详细的视觉信息并实现高保真度重建。为了更好地学习潜在信息，我们训练一个扩散变压器，通过重构标记引导学习恢复潜在标记。扩散去噪有效地对视觉观察的条件分布进行建模。

在长时间任务的实验中，我们的隐式定位方法比其他视觉定位范式更为有效。此外，视觉注意力的可视化证明了我们的 ReconVLA 展示了指导性视觉注意力并导致精确操作。接着，消融研究证明了通过大规模预训练的泛化能力。与其他流行方法的全面比较显示我们的 ReconVLA 提供了优越的性能。最后，我们进行真实世界的实验，并评估对未见过物体的泛化能力。这表明我们的 ReconVLA 有潜力促进 VLAs 的真实世界部署。

总之，我们的主要贡献如下：

- 我们提出了 ReconVLA，这是一种具有隐式定位范式的重建性 VLA 模型。凝视区域的重构促使模型进行精确的视觉注意分配和细粒度的表示学习，从而增强其视觉定位能力并执行精确操作。
- 我们构建了一个大规模的机器人预训练数据集，包含超过 10 万个轨迹，200 万数据样本。在该数据集上进行预训练增强了模型在视觉重建能力方面的泛化能力。
- 大量的模拟和真实世界实验展示了我们隐性基准方法的优越性，以及对未见目标进行精确操作和概括的能力。

在预训练的 VLMs 基础性进展的基础上，VLAs 学习生成由动作监督的可执行动作。在它们之中，RoboFlamingo 使用明确的策略头来建模顺序历史信息。OpenVLA 是第一个具有大规模机器人预训练的开源 VLA 模型。VLAS 拓展了音频的模态。UniVLA 从网络规模视频中学习以任务为中心的潜在动作，并适应不同的下游任务。这些模型仅监督动作输出，而我们的模型对视觉输出进行监督作为辅助任务，从而增强视觉感知。

用于操作的生成方法。 之前的研究已经探索了用于机器人控制的图像或视频生成模型。Unipi (Du et al. 2023) 首先生成未来的图像并从生成的图像中提取动作。SuSIE (Black et al. 2024) 使用图像编辑扩散模型生成子目标，并使用与语言无关的策略执行它们。CLOVER (Bu et al. 2024) 产生视觉计划，通过误差测量来引导闭环策略。GR-1 (Wu et al. 2024) 首次将生成方法与 VLAs 结合，提出了一种类 GPT 模型用于视觉机器人操作，通过大规模视频预训练来预测未来图像和机器人动作。3D-VLA (Zhen et al. 2024) 进一步整合深度信息作为视觉-语言-动作推理和规划的指导。GEVRM (Zhang et al. 2025) 以闭环方式为目标条件策略生成未来图像。这些方法 (Tian et al. 2024; Guo et al. 2024; Wang et al. 2025; Cen et al. 2025) 通过预测未来帧从动态中学习，从而增强模型的规划能力。而相比之下，我们的方法重建当前图像的目标区域以实现精确的感知和操作。

用于操作的视觉对齐方法。 显式锚定方法通常将被锚定的图像作为额外输入以作为辅助观察 (?? (a))。RoboGround (Huang et al. 2025) 使用 LISA (Lai et al. 2024) 作为高级分割器，根据指令提取目标对象和背景，并将其作为观察的一部分输入到 VLA 模型中。类似地，VIP (Li et al. 2025) 使用 YOLOv11 (Khanam and Hussain 2024) 对目标对象进行分割，然后放大并提供给基于 transformer 的策略。然而，这些模型依赖于外部专家模型，并没有从根本上增强策略本身的视觉锚定能力。ECoT (Zawalski et al. 2024) 和 GraspVLA (Deng et al. 2025) (?? (b)) 采用链式思维方法，顺序输出边界框和动作，同时训练锚定能力并通过因果注意为动作输出提供更丰富的信息。与这些先前方法相比，我们的 ReconVLA 直接从视觉输出中重建目标操作区域 (?? (c))，从而在鼓励模型学习目标区域的细粒度表征的同时隐式执行锚定。该过程模仿人眼能够自发聚焦于视野内显着区域的能力。

2 方法

为了建立我们方法的基础，我们首先在机器人操控的背景下形式化 VLA 模型的典型公式化和架构。给定一对图像和文本指令 (I, S) ，VLA 模型 Λ 预测动作 $\mathcal{A} = \Lambda(I, S)$ 。

一个常规的 VLA 主要由一个大型语言模型 LLM 、一个视觉编码器 $\mathcal{E}$ 、分词器 $\mathcal{T}$ 和动作去分词器 $\mathcal{Q}$ 组成。元组 (I, S) 分别通过 $\mathcal{E}$ 和 $\mathcal{T}$ 处理成图像标记 $\boldsymbol{h}_I$ 和文本标记 $\boldsymbol{h}_S$ 。这些标记随后输入到 LLM 中以生成动作标记 $\boldsymbol{a}$ 。最终，动作去分词器 $\mathcal{Q}$ 将 $\boldsymbol{a}$ 映射为用于机器人控制的可执行动作 $\mathcal{A}$ 。整个过程可表示为：具体来说，动作标记是以自回归方式生成的：其中 i 表示第 i 个动作标记，N 表示动作标记的总数。

2.1 重构视觉-语言-动作模型

通过观察分散的注意力，我们旨在引导 VLA 的视觉注意力集中在正确的目标上。我们的理念是通过设置一个重构的视觉信号来构建辅助视觉监督。监督信号作为条件，引导扩散去噪过程以重构目标操控区域。正式地，我们在这个框架中提出了重构视觉-语言-动作模型 (ReconVLA)。

在操作物体时，人类能够看到场景的全局视图。然而，视觉感知主要集中在其中的一小部分，即预期将要操作的区域。这种行为被称为凝视。类似地，我们的 ReconVLA 的重建目标是目标操作区，我们称之为凝视区。凝视区不仅有助于模型在多个可操作区域中聚焦于正确的目标，还增强了对这些区域的细节感知。此外，这一机制在长时间任务中通过关注和切换至不同子目标，隐含地促进了子任务的规划。

损失函数。 总体训练目标 ReconVLA 包括 (i) 由示范数据监督的自回归动作预测，以及 (ii) 由注视区域的视觉特征监督的另一个重构项，即 $\mathcal{L}_{\text{ReconVLA}} = \mathcal{L}_{\text{VLA}}^{\text{action}} + \mathcal{L}_{\text{VLA}}^{\text{visual}}$ ，其中 $\mathcal{L}_{\text{VLA}}^{\text{action}}$ 是交叉熵损失，$\mathcal{L}_{\text{VLA}}^{\text{visual}}$ 则是重构标记 $\boldsymbol{h}_R$ 和重构目标 I' 之间的一种度量。

为了从具有空间信息冗余的 RGB 输入中构建区域特定的视觉监督信号，我们设计了一个去噪过程，以用注视区域的低层特征重建标记。这个过程鼓励模型充分捕捉内在特征，而不是简单克隆显式的 RGB 值。

?? 展示了我们的 ReconVLA 如何利用视觉标记器 $\mathcal{F}$ 提取目标场景标记 $\boldsymbol{z}_0 = \mathcal{F}(I')$ 。具体来说，我们在 (Rombach et al. 2022) 中使用了一种连续变分自编码器 (VAE) 作为视觉标记器 $\mathcal{F}$ ，因为它具备视觉保真度以及捕捉细粒度图像特征的能力。去噪器 $\mathcal{D}$ 试图预测噪声，并基于重构标记 $\boldsymbol{h}_R = \text{LLM}(\boldsymbol{h}_I)$ 从噪声标记 $\boldsymbol{z}_t$ 中恢复 $\boldsymbol{z}_0$ 。重构的目标函数根据扩散过程 (Ho, Jain, and

Abbeel 2020) 进行形式化：

$$\mathcal{L}_{\text{VLA}}^{\text{visual}}(\boldsymbol{h}_R, I') = \mathbb{E}_{t,\boldsymbol{\epsilon}}\left[||\mathcal{D}(\boldsymbol{z}_t; \boldsymbol{h}_R, t) - \boldsymbol{\epsilon}||^2\right], \quad (1)$$

其中 t 表示扩散时间步长。去噪器 $\mathcal{D}$ 由一组 Transformer 编码器模块 (Vaswani 2017) 组成，使用自注意力模块以捕捉噪声标记和重构标记之间的相关性。

为了确保 VLA 模型能够处理与指令目标相对应的视觉标记，有必要保证图像标记能够关注到指令标记。因此，我们在图像标记之前插入一组指令标记，使得图像标记能够通过因果注意力从这些前缀文本中融合信息。实验结果表明，这种交错格式可以在不降低模型固有语言建模能力的情况下实现我们的目标。

在本文中，我们基于一个预训练的视觉语言模型 LLaVA-7b 构建我们的 ReconVLA，使用 Qwen2-7b 作为 LLM 骨干，并使用 siglip-so400m-patch14-384 作为视觉编码器。

2.2 视觉预训练

VLA 模型的重建能力本质上是有限的，因为它的 VLM 骨干网络主要是针对视觉-语言理解任务进行训练的。为了增强它对特定区域的定位和重建能力，我们设计了一种针对一个大型机器人数据集的重建任务的预训练过程。

为了构建一个基本的重建能力，我们将大型开放源代码机器人数据集 BridgeData V2 (Walke et al. 2023) 与高质量的模拟数据集 LIBERO (Liu et al. 2024a) 和 CALVIN (Mees et al. 2021) 结合起来，构建了预训练数据集。给定一对图像-文本对，我们微调 Grounding DINO (Liu et al. 2024c)，这是最先进的开放词汇对象检测器，以分割出机器人指示互动的注视区域。裁剪后的图像与原始图像成对组织。通过这种方式，我们获得了一个标注的视觉预训练数据集，其中包含超过 100k 的轨迹和 200 万的样本。

在预训练过程中，我们在重建损失和动作损失上都进行梯度反向传播，以保持优化目标的一致性。这个过程赋予我们的 VLM 广泛的视觉重建能力，并促进模型在不同环境和任务中的部署。预训练完成后，我们在特定任务上微调模型，以精确地将视觉语言理解和视觉重建能力与在相应动作空间上的操作能力对齐。

在本节中，我们设计实验以回答以下问题：

- 我们的隐式落地方法是否优于其他视觉落地范式？（参见 ?? ）
- 注视机制是否有助于视觉基础并进一步提高精确操作？（见 ?? ）
- 我们提出的预训练阶段是否改善了视觉生成的泛化能力，以及其他提议的关键设计如何影响 ReconVLA 的整体性能？（参见 ?? ）
- ReconVLA 能否有效管理与其他竞争方法相比的长期任务？（参见 ?? ）
- ReconVLA 能够在现实任务中对未见过的目标实现广义操控吗？（见 ?? ）

2.3 模拟环境

CALVIN 基准测试 (Mees et al. 2021) 构建于 PyBullet (Coumans and Bai 2016–2019) 模拟器之上，涉及一个操控场景的 Franka Panda 机器人手臂。CALVIN 包含 34 个任务和 4 个不同的环境（A、B、C 和 D）。

Paradigm	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg. Len
Baseline	88.8	76.1	63.7	57.0	49.0	3.36
EG	94.4	82.5	70.9	62.2	50.2	3.61
CG	47.0	14.3	1.6	0.0	0.0	0.63
IG (ours)	95.6	87.6	76.9	69.3	64.1	3.95

Table 1: 比较不同范式，包括显式基础 (EG)、串联推理基础 (CG) 和我们的隐式基础 (IG)。比较在 CALVIN ABC → D 上进行。

CALVIN 长期挑战是一个顺序任务，包括五个子任务。我们报告每个子任务的成功率和所有五个任务的平均完成长度。该方法通过 500 次试验进行评估，以确保公平比较。CALVIN 的指标是每个子任务的成功率和所有连续 5 个子任务的平均长度。

我们在相同的基线上在 ?? 中实现不同的视觉定位范式，以进行公平的比较。

显式基础 (EG)。我们选择经过微调的 YOLOv11 (Khanam and Hussain 2024) 作为检测器来识别每个时间步的目标对象。然后我们从图像中裁剪出识别出的对象区域并调整大小。接下来，将调整大小后的图像和原始图像共同输入 VLA 模型，以指导对象操作。

思维链基础 (CG)。对于数据准备，我们使用检测器对图像进行预处理，以获得边界框的坐标。然后，我们重新构建训练数据集，并将输出修改为 CoT 格式：Bbox [x1 x2 y1 y2] + 动作序列。输入仍然是原始图像。通过这种方式，VLA 模型学习定位目标对象并输出具有定位信息的动作。

正如在 Table 1 中所示，EG 获得了比基线相对更高的成功率。这表明将显式定位作为输入有助于更好地理解空间关系。然而，简单地拼接完整和裁剪的图像引入了视觉信息冗余，限制了模型的性能。CG 的性能更差。这表明坐标形式的边界框不足以有效指导模型精准地操作目标位置。此外，直接同时输出精确的坐标和动作值给 VLA 模型的训练带来了挑战。

我们的隐式定位方法获得了最高的成功率，这证明了我们的方法优于其他范式。从训练机制的角度看，这一优势来源于我们的隐式定位学习框架，它使模型能够精确地关注目标物体的视觉信息，从而实现精确的操作。从架构的角度来看，我们的模型直接监督视觉输出，消除了对额外输入或输出的需求。这一设计带来了一个简单而有效的训练和推理流程。

为了更好地探索注视机制的影响，我们进行了视觉注意力的定性实验及其对细粒度操作任务的影响。

注意力可视化 Figure 2 证明 $\mathcal{L}_{\text{VLA}}^{\text{visual}}$ 的实现能够使注意力与视线区域紧密对齐，这对应于目标对象。对于“将西瓜放入黄色碗中”这一指令，基线的注意力高度分散，三视图图像的注意力大多集中在不相关的位置上，导致任务失败。相比之下，ReconVLA 成功地将注意力集中在正确的目标上，即西瓜。这表明我们的方法带来了精确的视觉定位，促进了任务的成功。

在所有任务中，“堆叠方块”任务是最具挑战性的，它要求机器人提起一个方块并将其精准地堆叠在另一个方块上。尽管我们的基线在此任务上仅实现了 59.3 %

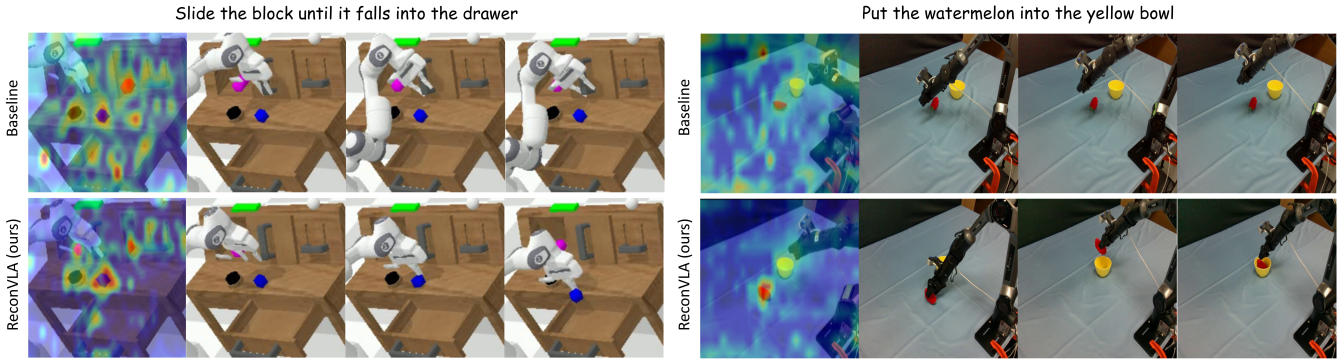

Figure 2: 对 CALVIN (Mees et al. 2021) 和现实世界中注意力图的定性比较。第一行：基线显示出分散的注意模式，或主要集中在不正确的区域，导致不准确的操作。第二行：通过辅助的视觉监督信号，ReconVLA 迫使模型专注于特定图像内容，并赋予更高的注意值，从而精确地移动到目标区域，成功完成任务。

Recon.	Gaze Region	Pretrain	Splits	Task completed in a row (%) 1	2	3	4	5	Average Length
✓	✓	✓	ABC → D	95.6	87.6	76.9	69.3	64.1	3.95
✓	✓	×	ABC → D	96.8	86.9	76.9	64.9	58.2	3.85
✓	×	×	ABC → D	89.8	80.3	67.7	56.6	46.5	3.42
×	×	×	ABC → D	88.8	76.1	63.7	57.0	49.0	3.36

Table 2: 使用重建部分、注视区域和预训练进行的所提出技术的消融结果。

的成功率，我们的凝视机制成功率高达 79.5 % ，增加了 20.2 % 。这一显著的改进突显了我们凝视机制通过精准视觉定位提升的动作准确性。

我们在 Table 2 中使用重建部分、注视区和在大规模机器人数据集上的预训练来进行所提技术的消融研究。我们观察到预训练显著提高了成功率。这是因为在未见过的测试环境中，锚定目标对象并执行重建本质上是具有挑战性的，对模型的生成能力提出了泛化挑战。对大规模数据集的预训练在视觉重建过程中大大增强了模型的泛化能力。此外，重建要操作的注视区域，而不是整个图像，证明更加有效。这引导模型的视觉注意力集中于目标对象，从而避免了对错误目标的操控。值得注意的是，训练重建整个图像的模型仍然比基线性能更优，这可以归功于整体视觉注意力的增强。然而，在未见过的场景中，带有像素冗余的整个图像重建是极具挑战性的，这进一步限制了性能提升。

我们将我们的模型与预测未来图像的生成方法（UniPi、SuSIE、CLOVER、3D-VLA、GR-1、Vidman、GEVRM）以及大型 VLA 模型（RoboFlamingo、VLAS、OpenVLA、UniVLA）进行比较，如 Section 1 中所介绍的。

在基础的 ABCD → D 任务中，我们的 ReconVLA 实现了具有竞争力的表现，成功完成了平均 5 个连续任务中的 4.23 个，第一个任务的成功率为 98.0 %。这表明，我们的凝视机制提供了一种灵活的规划能力，以在长远任务中实现更好的操作安排。ABC → D 任务挑战对未见过背景的泛化。我们的方法超越了所有生成方法，包括流行的 GR-1，在最后一个子任务中超过 20 % 的成功率。这表明，除了预测未来图像的生成模型外，增强对当前观察的感知对机器人操作同样具有价值。在可比较的参数量下，我们的方法在最后一个子任务中比 OpenVLA 提高了 20.6 %，比 UniVLA 提高了 7.6 %，这表明了我们隐式基础学习策略的有效性。

我们使用一个具有 1 个自由度平行夹爪的 6 自由度 AgileX PiPer 机器人手臂进行了实际实验。此外，我们使用 RealSense D515 深度相机作为基础定位相机，并使用 ORBBEC Dabai 深度相机作为手持相机获取视觉输入。我们选择了四个代表性任务：将水果放入碗中、堆叠碗、翻转杯子和清理餐桌。为了增强模型的泛化能力，每个任务都包括目标对象和背景颜色的变化。我们平均每个任务收集 150 条轨迹。为了评估，每个模型在每项任务上进行 20 次尝试，并使用成功率作为性能指标。对于未见过的任务，我们用未见过的对象替换目标对象。[模型名称] 在四个实际任务中始终优于流行的 OpenVLA 和强劲的 PD-VLA，且在每种情况下都取得了最高的成功率。特别是，[模型名称] 在“将水果放入碗中”和“堆叠碗”任务中达到了接近或超过 90% 的成功率。OpenVLA 在执行精细操控任务（例如翻转杯子和清理餐桌）时效果有限，而我们的模型通过精确的视觉定位实现了显著的性能提升。

在看不见的任务中，目标对象在训练数据中缺失，OpenVLA 和 PD-VLA 方法几乎显示出 0 % 的成功率。得益于大规模混合数据的预训练，我们的 ReconVLA 仍然可以成功地识别目标对象并完成预期的动作，展示了我们方法在视觉推广能力上的优势。

在本文中，我们分析并揭示了传统 VLA 中分散的视觉注意力，这限制了精确操控。然后，我们提出了一种重建的视觉-语言-动作模型（ReconVLA），这是一种在隐式落地范式下训练的新颖框架。我们的 ReconVLA 成功实现了准确的视觉注意力分配，并进一步提高了操控技能。我们还构建了一个大型的预训练数据集用于 ReconVLA 的在多样场景和未见物体上的泛化。在模拟

Category	Method	Splits	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg.
Generative Methods	UniPi (Du et al. 2023) (NIPS'23)	ABC → D	56.0	16.0	8.0	8.0	4.0	0.
	SuSIE (Black et al. 2024) (ICLR'24)	ABC → D	87.0	69.0	49.0	38.0	26.0	2.
	GEVRM (Zhang et al. 2025) (ICLR'25)	ABC → D	92.0	70.0	54.0	41.0	26.0	2.
	GR-1 (Wu et al. 2024) (ICLR'24)	ABC → D	85.4	71.2	59.6	49.7	40.1	3.
	Vidman (Wen et al. 2024) (神经信息处理系统会议'24)	ABC → D	91.5	76.4	68.2	59.2	46.7	3.
	CLOVER (Bu et al. 2024) (NIPS'24)	ABC → D	96.0	83.5	70.8	57.5	45.4	3.
Large VLA Models	VLAS (Zhao et al. 2025) (ICLR '25)	ABC → D	87.2	64.2	40.9	28.1	19.6	2.
	RoboFlamingo (Li et al. 2024) (ICLR'24)	ABC → D	82.4	61.9	46.6	33.1	23.5	2.
	OpenVLA (Kim et al. 2024) (CoRL'24)	ABC → D	91.3	77.8	62.0	52.1	43.5	3.
	UniVLA (Bu et al. 2025) (RSS'25)	ABC → D	95.5	85.8	75.4	66.9	56.5	3.
Reconstructive Methods	ReconVLA (ours)	ABC → D	95.6	87.6	76.9	69.3	64.1	3.

Table 3: 在成功率和平均长度方面，与不同的操控模型在 CALVIN ABC → D 上的比较。

Category	Method	Splits	Success Rate (%) 1/5	2/5	3/5	4/5	5/5	Avg. Len
Generative Methods	3D-VLA (Zhen et al. 2024) (ICML'24)	ABCD → D	44.7	16.3	8.1	1.6	0	0.70
	GR-1 (Wu et al. 2024) (ICLR'24)	ABCD → D	94.9	89.6	84.4	78.9	73.1	4.21
Large VLA Models	VLAS (Zhao et al. 2025) (ICLR'25)	ABCD → D	94.2	84.0	73.2	64.3	54.6	3.70
	RoboFlamingo (Li et al. 2024) (ICLR'24)	ABCD → D	96.4	89.6	82.4	74.0	66.0	4.08
Reconstructive Methods	ReconVLA (ours)	ABCD → D	98.0	90.0	84.5	78.5	70.5	4.23

Table 4: 在成功率和平均长度方面与 CALVIN ABCD → D 中的各种操控模型进行比较。

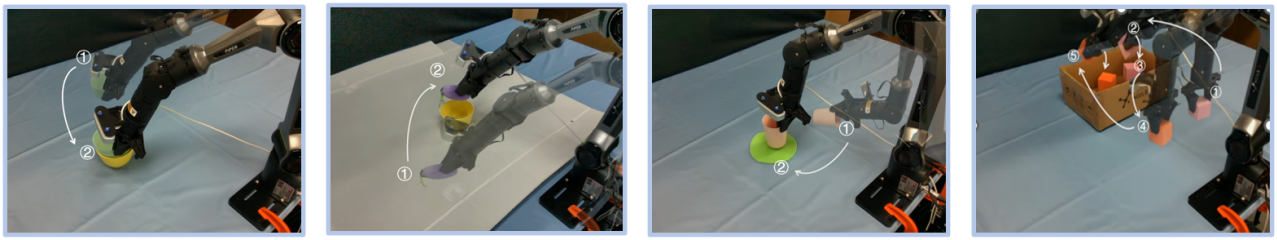

Figure 3: 四个代表性任务的现实环境设置。我们使用一个具有 1 自由度平行抓手的 6 自由度 AgileX PiPer 机器人臂和一个 RealSense D515 深度摄像头作为基座摄像头，以及一个 ORBBEC Dabai 深度摄像头作为手持摄像头。我们选择了四个具有代表性和实际意义的任务：（1）堆叠碗，（2）将水果放入碗中，（3）翻杯子，（4）清理桌子。

和真实世界中的大量实验表明我们隐式落地方法的优越性。

References

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; Jitsev, J.; Kornblith, S.; Koh, P. W.; Ilharco, G.; Wortsman, M.; and Schmidt, L. 2023. OpenFlamingo: An Open-Source Framework for Training Large Autoregressive Vision-Language Models. arXiv preprint arXiv:2308.01390.

Black, K.; Nakamoto, M.; Atreya, P.; Walke, H. R.; Finn, C.; Kumar, A.; and Levine, S. 2024. Zero-Shot Robotic Manipulation with Pre-Trained Image-Editing Diffusion Models. In The Twelfth International Conference on Learning Representations.

Brohan, A.; Brown, N.; Carbajal, J.; Chebotar, Y.; Dabis, J.; Finn, C.; ; et al. 2023. RT-1: Robotics Transformer for Real-World Control at Scale. Proceedings of Robotics: Science and Systems.

Bu, Q.; Yang, Y.; Cai, J.; Gao, S.; Ren, G.; Yao, M.; Luo, P.; and Li, H. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. arXiv preprint arXiv:2505.06111.

Bu, Q.; Zeng, J.; Chen, L.; Yang, Y.; Zhou, G.; Yan, J.; Luo, P.; Cui, H.; Ma, Y.; and Li, H. 2024. Closed-loop visuomotor control with generative expectation for robotic manipulation. Advances in Neural Information Processing Systems, 37: 139002–139029.

Cen, J.; Yu, C.; Yuan, H.; Jiang, Y.; Huang, S.; Guo, J.; Li, X.; Song, Y.; Luo, H.; Wang, F.; et al. 2025. World-

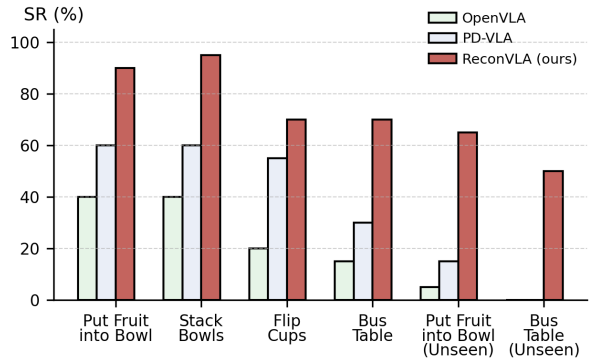

Figure 4: 现实世界的多任务结果。我们报告了 4 个不同任务以及 2 个未见任务的成功率 (SR)。

VLA: Towards Autoregressive Action World Model. arXiv preprint arXiv:2506.21539.

Coumans, E.; and Bai, Y. 2016–2019. PyBullet, a Python module for physics simulation for games, robotics and machine learning. http://pybullet.org.

Deng, S.; Yan, M.; Wei, S.; Ma, H.; Yang, Y.; Chen, J.; Zhang, Z.; Yang, T.; Zhang, X.; Cui, H.; et al. 2025. Graspvla: a grasping foundation model pre-trained on billion-scale synthetic action data. arXiv preprint arXiv:2505.03233.

Du, Y.; Yang, S.; Dai, B.; Dai, H.; Nachum, O.; Tenenbaum, J.; Schuurmans, D.; and Abbeel, P. 2023. Learning universal policies via text-guided video generation. Advances in neural information processing systems, 36: 9156–9172.

Fang, H.-S.; Fang, H.; Tang, Z.; Liu, J.; Wang, C.; Wang, J.; Zhu, H.; and Lu, C. 2024. Rh20t: A comprehensive robotic dataset for learning diverse skills in one-shot. In 2024 IEEE International Conference on Robotics and Automation (ICRA), 653–660. IEEE.

Guo, Y.; Hu, Y.; Zhang, J.; Wang, Y.-J.; Chen, X.; Lu, C.; and Chen, J. 2024. Prediction with Action: Visual Policy Learning via Joint Denoising Process. arXiv:2411.18179.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. arXiv:2006.11239.

Huang, H.; Chen, X.; Chen, Y.; Li, H.; Han, X.; Wang, Z.; Wang, T.; Pang, J.; and Zhao, Z. 2025. RoboGround: Robotic Manipulation with Grounded Vision-Language Priors. In Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 22540–22550.

Khanam, R.; and Hussain, M. 2024. Yolov11: An overview of the key architectural enhancements. arXiv preprint arXiv:2410.17725.

Kim, M. J.; Pertsch, K.; Karamcheti, S.; Xiao, T.; Balakrishna, A.; Nair, S.; Rafailov, R.; Foster, E. P.; Sanketi, P. R.; Vuong, Q.; et al. 2024. OpenVLA: An Open-Source Vision-Language-Action Model. In 8th Annual Conference on Robot Learning.

Lai, X.; Tian, Z.; Chen, Y.; et al. 2024. LISA: Reasoning Segmentation via Large Language Model. arXiv:2308.00692.

Li, X.; Liu, M.; Zhang, H.; Yu, C.; Xu, J.; Wu, H.; Cheang, C.; Jing, Y.; Zhang, W.; Liu, H.; et al. 2024. Vision-Language Foundation Models as Effective Robot Imitators. In The Twelfth International Conference on Learning Representations.

Li, Z.; Ren, L.; Yang, J.; Zhao, Y.; Wu, X.; Xu, Z.; Bai, X.; and Zhao, H. 2025. VIP: Vision Instructed Pre-training for Robotic Manipulation. In Forty-second International Conference on Machine Learning.

Liu, B.; Zhu, Y.; Gao, C.; Feng, Y.; Liu, Q.; Zhu, Y.; and Stone, P. 2024a. Libero: Benchmarking knowledge transfer for lifelong robot learning. Advances in Neural Information Processing Systems, 36.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024b. Visual instruction tuning. Advances in neural information processing systems, 36.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; et al. 2024c. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In European conference on computer vision, 38–55. Springer.

Mees, O.; Hermann, L.; Rosete-Beas, E.; and Burgard, W. 2021. CALVIN: A Benchmark for Language-Conditioned Policy Learning for Long-Horizon Robot Manipulation Tasks. IEEE Robotics and Automation Letters.

Niu, D.; Sharma, Y.; Biamby, G.; Quenum, J.; Bai, Y.; Shi, B.; Darrell, T.; and Herzig, R. 2024. LLARVA: Vision-Action Instruction Tuning Enhances Robot Learning. arXiv preprint arXiv:2406.11815.

Octo Model Team; Ghosh, D.; Walke, H.; Pertsch, K.; Black, K.; Mees, O.; Dasari, S.; Hejna, J.; Xu, C.; Luo, J.; Kreiman, T.; Tan, Y.; Chen, L. Y.; Sanketi, P.; Vuong, Q.; Xiao, T.; Sadigh, D.; Finn, C.; and Levine, S. 2024. Octo: An Open-Source Generalist Robot Policy. In Proceedings of Robotics: Science and Systems. Delft, Netherlands.

O' Neill, A.; Rehman, A.; Maddukuri, A.; Gupta, A.; Padalkar, A.; Lee, A.; Pooley, A.; Gupta, A.; Mandlekar, A.; Jain, A.; et al. 2024. Open x-embodiment: Robotic learning datasets and rt-x models: Open x-embodiment collaboration 0. In 2024 IEEE International Conference on Robotics and Automation (ICRA), 6892–6903. IEEE.

Peebles, W.; and Xie, S. 2022. Scalable Diffusion Models with Transformers. arXiv preprint arXiv:2212.09748.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. arXiv:2112.10752.

Song, W.; Zhao, H.; Ding, P.; Cui, C.; Lyu, S.; Fan, Y.; and Wang, D. 2024. Germ: A generalist robotic model with mixture-of-experts for quadruped robot. In

2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 11879–11886. IEEE.

Stewart, E. E. M. e. a. 2020. A review of interactions between peripheral and foveal vision. Journal of Vision, 20(12): 2–2.

Tian, Y.; Yang, S.; Zeng, J.; Wang, P.; Lin, D.; Dong, H.; and Pang, J. 2024. Predictive inverse dynamics models are scalable learners for robotic manipulation. arXiv preprint arXiv:2412.15109.

Vaswani, A. 2017. Attention is all you need. Advances in Neural Information Processing Systems.

Walke, H.; Black, K.; Lee, A.; Kim, M. J.; Du, M.; Zheng, C.; Zhao, T.; Hansen-Estruch, P.; Vuong, Q.; He, A.; Myers, V.; Fang, K.; Finn, C.; and Levine, S. 2024. BridgeData V2: A Dataset for Robot Learning at Scale. arXiv:2308.12952.

Walke, H. R.; Black, K.; Zhao, T. Z.; Vuong, Q.; Zheng, C.; Hansen-Estruch, P.; He, A. W.; Myers, V.; Kim, M. J.; Du, M.; et al. 2023. Bridgedata v2: A dataset for robot learning at scale. In Conference on Robot Learning, 1723–1736. PMLR.

Wang, H.; Zheng, A.; Zhao, Y.; Wang, T.; Ge, Z.; Zhang, X.; and Zhang, Z. 2024. Reconstructive Visual Instruction Tuning. arXiv:2410.09575.

Wang, Y.; Li, X.; Wang, W.; Zhang, J.; Li, Y.; Chen, Y.; Wang, X.; and Zhang, Z. 2025. Unified Vision-Language-Action Model. arXiv preprint arXiv:2506.19850.

Wen, Y.; Lin, J.; Zhu, Y.; Han, J.; Xu, H.; Zhao, S.; and Liang, X. 2024. Vidman: Exploiting implicit dynamics from video diffusion model for effective robot manipulation. Advances in Neural Information Processing Systems, 37: 41051–41075.

Wu, H.; Jing, Y.; Cheang, C.; Chen, G.; Xu, J.; Li, X.; Liu, M.; Li, H.; and Kong, T. 2024. Unleashing Large-Scale Video Generative Pre-training for Visual Robot Manipulation. ICLR.

Zawalski, M.; Chen, W.; Pertsch, K.; Mees, O.; Finn, C.; and Levine, S. 2024. Robotic control via embodied chain-of-thought reasoning. arXiv preprint arXiv:2407.08693.

Zhang, H.; Ding, P.; Lyu, S.; Peng, Y.; and Wang, D. 2025. GEVRM: Goal-Expressive Video Generation Model For Robust Visual Manipulation. In The Thirteenth International Conference on Learning Representations.

Zhao, W.; Ding, P.; Zhang, M.; Gong, Z.; Bai, S.; Zhao, H.; and Wang, D. 2025. VLAS: Vision-Language-Action Model With Speech Instructions For Customized Robot Manipulation. International Conference on Learning Representations (ICLR).

Zhen, H.; Qiu, X.; Chen, P.; Yang, J.; Yan, X.; Du, Y.; Hong, Y.; and Gan, C. 2024. 3D-VLA: A 3D Vision-Language-Action Generative World Model. In ICML.

Zitkovich, B.; Yu, T.; Xu, S.; Xu, P.; Xiao, T.; Xia, F.; Wu, J.; Wohlhart, P.; Welker, S.; Wahid, A.; et al. 2023. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In Conference on Robot Learning, 2165–2183. PMLR.