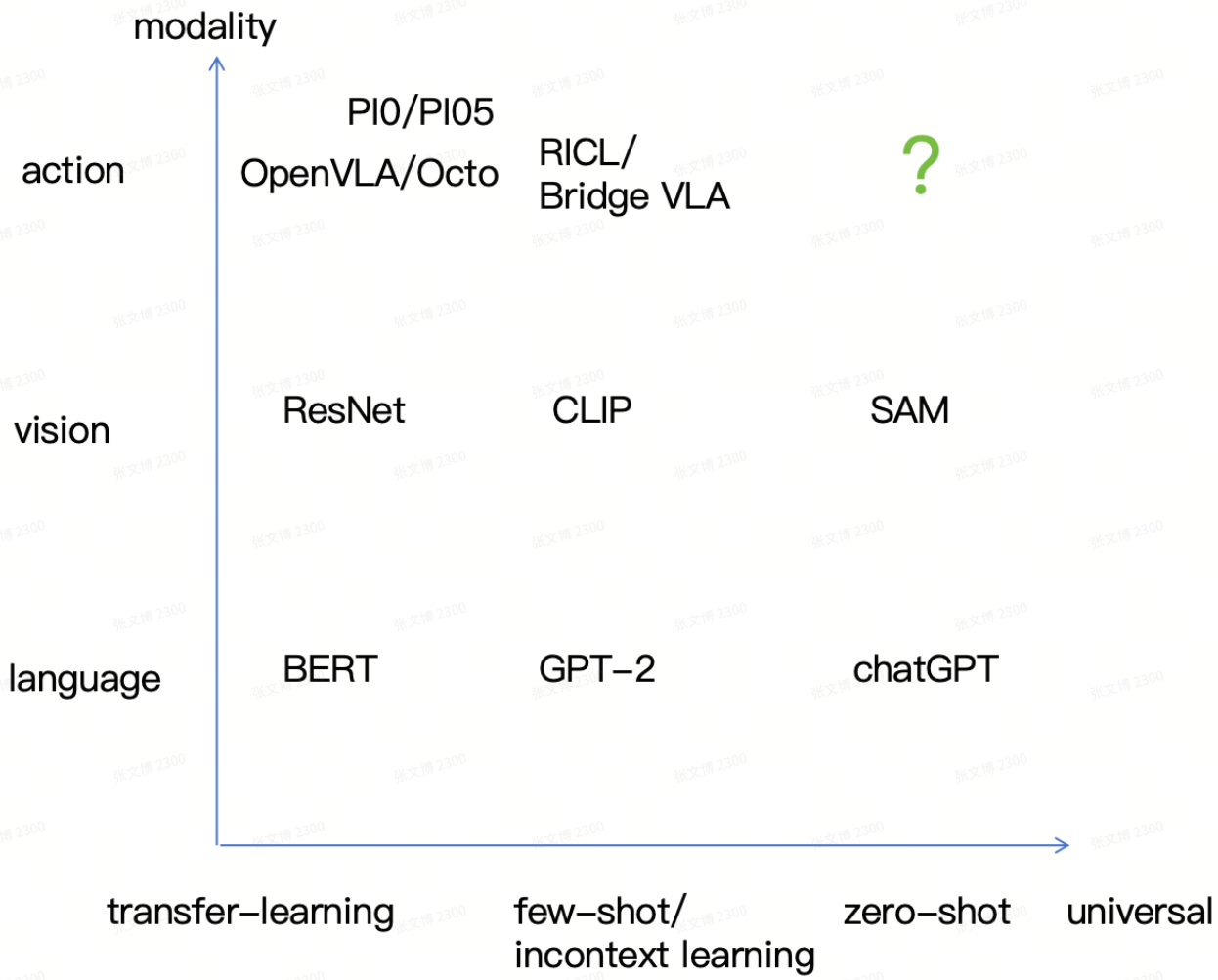
introduction



在计算机视觉（CV）与自然语言处理（NLP）领域，大模型具备零样本学习（zero-shot）能力。然而，现有的视觉语言模型（VLA）均需进行微调。其主要原因在于，语言和视觉领域的数据更为丰富，足以覆盖下游的多数应用环境。与之不同的是，机器人领域的环境变化相较于数据量的变化更为显著。在测试阶段，动作空间、摄像头视角、控制模式以及视觉特征等方面均存在较为明显的领域差异（domain gap）。

尽管存在领域差异，但机器人操作存在共享的表征。文献Scaling Proprioceptive-Visual Learning with Heterogeneous Pre-trained Transformers采用共享的骨干网络以及不同的感知和动作头进行预训练；UniVLA则采用潜在动作（latent action）统一动作表示。这些工作从结构设计层面解耦了共享表征与具体具身配置（embodiment configuration）之间的差异。这促使我们提出如下研究问题：

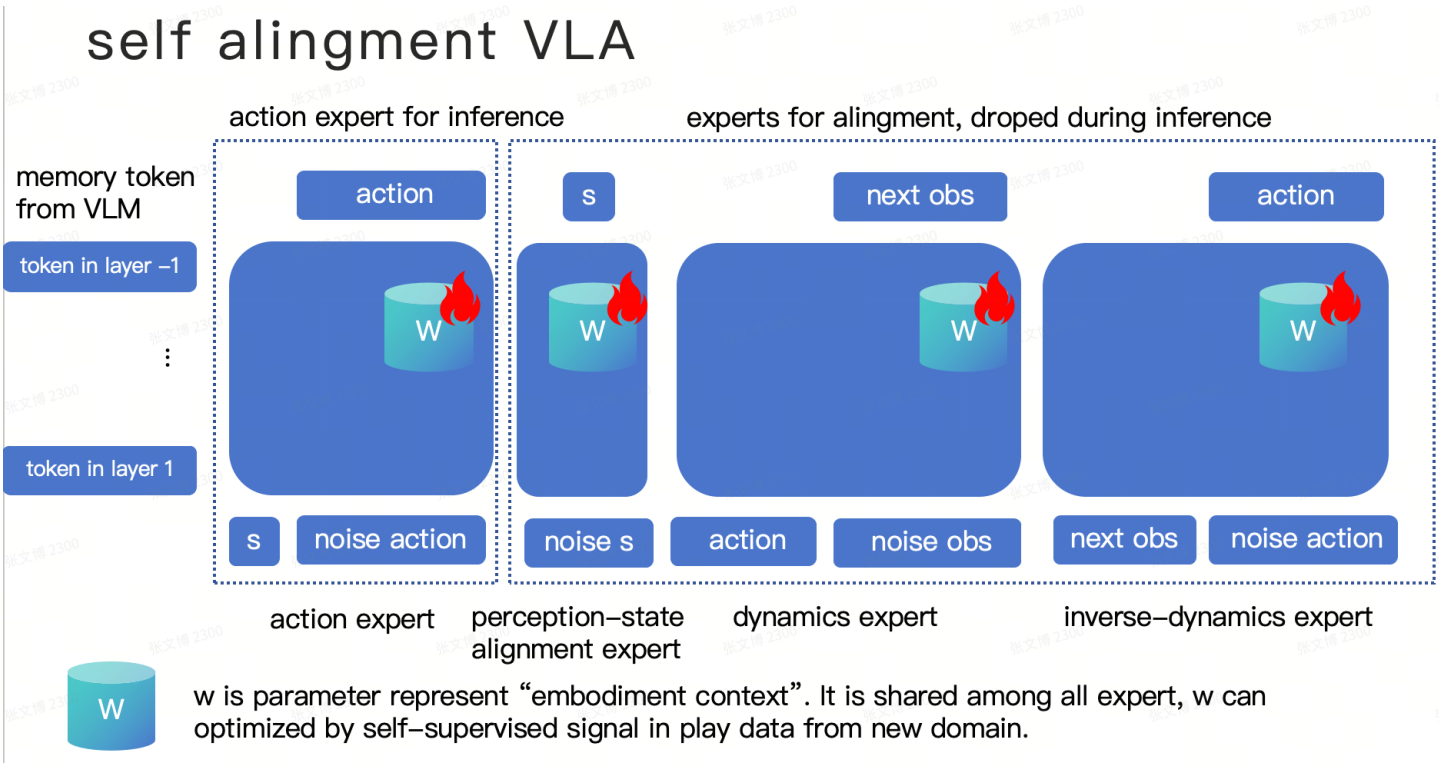
在当下，预训练阶段收集所有机器人操作数据并不现实，是否能够设计一种结构，解耦与具身无关（embodiment-agnostic）和与具身相关（embodiment-relevant）的表征？其中，与具身无关的

表征通过大规模预训练习得，与具身相关的表征在预训练阶段学会如何利用，而后在下游任务中借助游戏数据（play data）进行自对齐（self alignment），从而生成具备零样本学习能力的操作策略（manipulation policy）？

最为直观的做法是借鉴视觉语言模型（VLM）和大语言模型（LLM）的方法，利用新领域的历史数据进行上下文学习（in-context learning），如文献RICL（http://arxiv.org/abs/2508.02062）所采用的方法。然而，这种方法存在两个问题：其一，上下文学习仍需收集下游有监督的数据，无法实现零样本学习；其二，即便收集了数据，模型也仅能学会已收集演示样本的任务，无法激活预训练模型在其他任务上的能力。

我们需要一种能够更有效利用下游无标签游戏数据的方法。实际上，文献Cross-Embodiment Robot Manipulation Skill Transfer using Latent Space Alignment 曾采用过类似思路。该文献在预训练阶段植入多个自监督训练信号，并在下游自适应阶段利用这些自监督信号更新编码器和解码器。但其不足之处在于，未使用大规模数据进行预训练，导致与具身无关的操作表征缺乏通用性，整体迁移的领域范围有限。

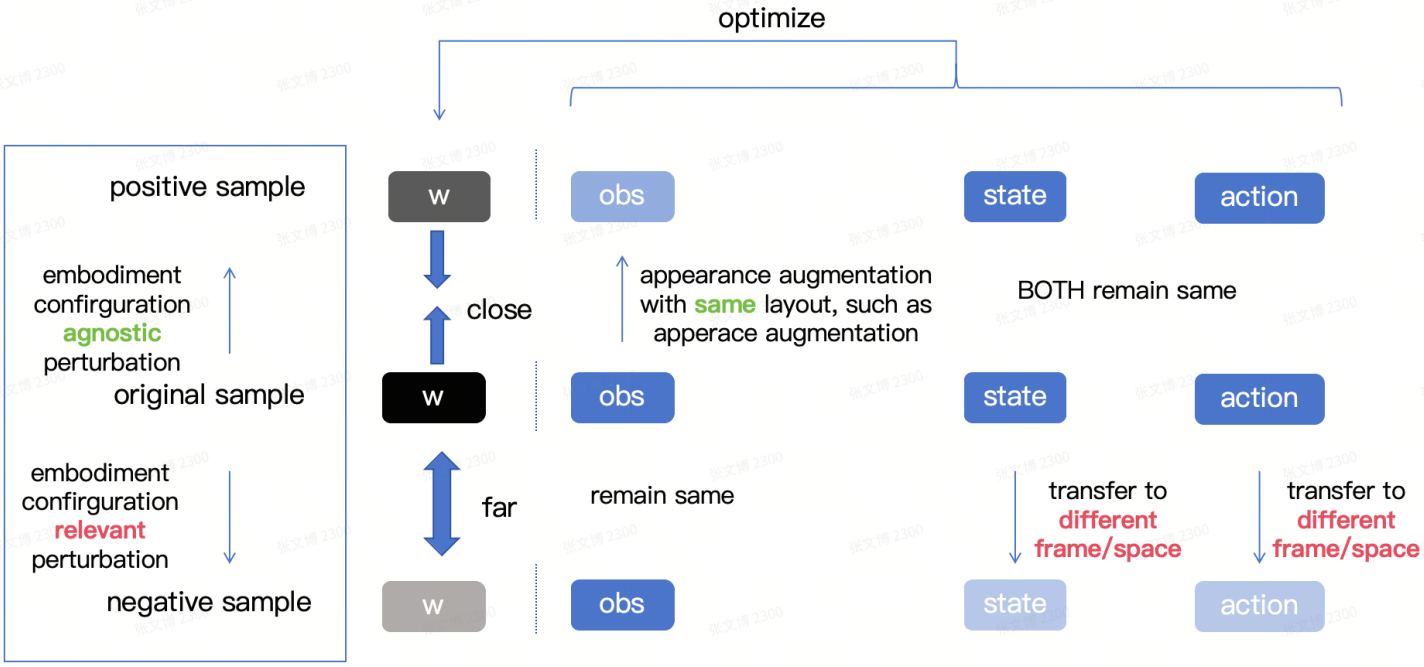
因此，我们的思路是结合大规模预训练的视觉语言模型（VLA）和基于游戏数据的自对齐方法。通过大规模预训练的VLA形成与具身无关的操作表征，利用自监督对齐完成对具体构型的适应。



本研究采用三个基于操作领域知识手工设计的自监督信号：1. 观测 - 状态对齐信号；2. 动力学对齐信号；3. 逆动力学对齐信号。这三个对齐专家（alignment expert）和动作专家（action expert）均以参数W为条件，参数W表征了具体场景的具身上下文（embodiment context）。我们期望通过在新场景下对参数W进行优化，实现对具身上下文的对齐。此外，还可植入TTT自监督信号。

直接训练无法迫使模型将具身上下文表征到参数W中，模型可能会利用其他参数记忆具体场景。因此，我们引入对比损失，将与具身相关和与具身无关的表征进行分离。

contrasive learning



对于负样本，对与具身相关的部分进行扰动，具体表现为状态/动作空间的变化，例如从笛卡尔空间转换到关节空间，坐标系从基坐标系转换到世界坐标系。这可迫使模型在理解状态和预测动作时以参数W为条件，表明在相同的视觉观测下，面对不同的具身上下文，模型的预测能够对应具体action space 。

对于正样本，对观测数据进行observation appearance层面的增强，但不改变observation layout（不采用翻转、剪裁等操作）。此类增强可作为正样本的原因在于，外观的改变并未显著改变具身上下文，观测 - 状态匹配、动力学以及逆动力学等方面未发生较大变化。