

具身智能VLA的多模态感知技术综述

周江盟^{1,2}, 孙奥兰¹, 瞿晓阳¹, 张旭龙¹, 王健宗¹

1. 平安科技(深圳)有限公司, 广东 深圳 518046;

2. 湖南大学计算机学院, 湖南 长沙 410082

摘要

具身智能中的视觉-语言-动作(VLA)系统对多模态感知的依赖日益增强,但现有研究仍存在感知与泛化能力不足的问题。针对这一挑战,对视觉、语言、音频和触觉等模态的特征表示、对齐与融合方法进行了系统梳理,并提出面向VLA的多模态感知统一框架,重点比较了多层感知机投影与交叉注意力等策略的适用性与局限性。通过对Open X-Embodiment、BridgeData V2等多模态数据集的分析,总结了高质量数据在增强感知与任务泛化中的关键作用。研究表明,多模态感知显著提升了VLA系统的表现,但仍需在持续学习、触觉模态开发与高质量数据采集方面深化探索,以推动更通用和鲁棒的具身智能发展。

关键词

具身智能; 多模态大模型; 机器人

中图分类号: TP18

文献标志码: A

doi: 10.11959/j.issn.2096-0271.2026063

A survey of multimodal perception in embodied VLA systems

Zhou Jiangmeng^{1,2}, Sun Aolan¹, Qu Xiaoyang¹, Zhang Xulong¹, Wang Jianzong¹

1. Ping An Technology (Shenzhen) Co., Ltd., Shenzhen 518046, China;

2. College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China

Abstract

Visual-language-action (VLA) systems in embodied intelligence increasingly rely on multimodal perception, yet existing approaches still suffer from limited perceptual and generalization abilities. To address this challenge, feature representation, alignment, and fusion methods across visual, linguistic, auditory, and tactile modalities were systematically reviewed, and a unified multimodal perception framework for VLA was proposed, with emphasis on comparing strategies such as multilayer perceptron projection and cross-attention. Analysis of datasets including Open X-Embodiment and BridgeData V2 summarized the critical role of high-quality multimodal data in enhancing perception and task generalization. The results indicate that multimodal perception substantially improves VLA performance, while further advances in continuous learning, tactile modality development, and high-quality data collection are essential for building more general and robust embodied intelligence.

Key words

embodied AI, multimodal model, robotics

0 引言

具身智能作为人工智能的重要发展方向，旨在构建能够感知、理解并与真实环境交互的智能体，因此被广泛认为是通用人工智能乃至世界模型的关键要素^[1]。随着大语言模型和多模态技术的飞速发展，人们开始探索利用多模态大模型来控制具身机器人完成物理交互任务。在这样的技术交汇点下，视觉-语言-动作（vision-language-action, VLA）^[2]模型应运而生。VLA模型是具身人工智能领域中的一类多模态模型，旨在处理视觉、语言和动作模态的信息。在此过程中，多模态感知被视为VLA模型中不可或缺的核心要素。面对复杂多变的真实场景，单一模态（如文本或图像）常难以全面刻画环境细节，难以支撑稳定高效的执行任务，因此出现了其他新兴模态（如触觉）。如何有效融合视觉、文本、音频、触觉等多模态信息，实现协同感知，已成为具身智能领域（尤其是VLA模型）的核心问题。

目前，具身智能领域进展显著^[3-5]，但现有综述存在明显局限：一是多按模态类型罗列式梳理，缺乏对技术本质与适用场

景的深度关联；二是未形成“技术特性-任务需求-性能约束”的分析逻辑，难以指导实际系统设计。本文聚焦VLA多模态感知的核心流程，将其划分为“多模态表示-多模态对齐-多模态融合”3个技术环节，结合精细操作、任务规划等典型场景，以及实时性、泛化性等约束条件，系统剖析研究现状。同时整理VLA常用数据集的多模态适配特性，分析多模态泛化、灾难性遗忘等核心挑战的内在关联，为后续研究提供更具针对性的参考。

本文核心观点如下：（1）VLA感知性能的核心是三元适配，而非模态数量堆砌；（2）大模型的隐式跨模态建模能力可弥补轻量融合结构的不足；（3）触觉等物理模态是突破性能瓶颈的关键，但需多技术协同突破；（4）泛化能力不足的根源是模态表示依赖数据分布，需通过自监督学习与持续学习机制来解决。

本文的理论贡献在于构建了一个更具深度的分析框架，为VLA多模态感知研究提供了新的技术参考维度；在应用层面，则显著增强了机器人在复杂动态环境下的自主执行稳健性。展望后续研究，建议重点攻克持续学习中的抗遗忘机制、跨模态的通用对齐融合算法，以及真实场景下的多模态数据集构建等关键课题，以此加速

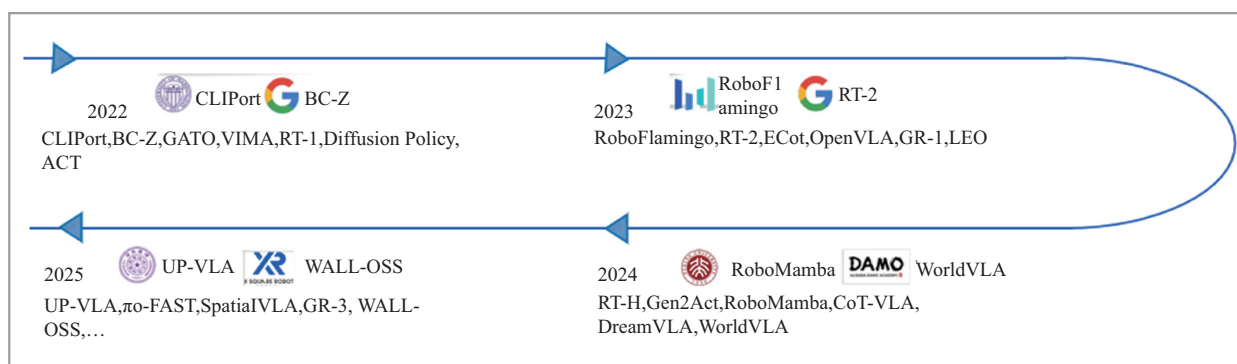


图1 VLA模型发展路线

具身智能向更高通用性与实用化水平演进。

1 多模态感知与表示

视觉-语言-动作（VLA）模型的发展路线如图 1 所示。此类研究旨在通过视觉、听觉、触觉等模态计算机器人控制量，以获得更丰富的语义信息并降低传感器成本。早期方法受限于封闭任务集合和网络容量，难以实现跨任务泛化。随着 Transformer 框架的兴起，大模型展示了卓越的自然语言理解和图像理解能力，经验证，通过增加数据规模和叠加 Transformer 模块可以提升性能。这促进了 VLA 模型的发展，该模型基于 Transformer 框架，利用视觉和语言输入来计算机器人的控制动作。自 2023 年 VLA 概念被首次提出，它经历了从萌芽阶段到探索阶段再到快速发展阶段的演变。在这一过程中，研究者们不仅探

索了如何更好地处理异构信息和多模态动作分布，还尝试通过继承预训练大模型权重、利用视频生成模型和大规模人类操作活动数据等手段提升 VLA 模型的理解能力和泛化能力。此外，分层架构的引入使得 VLA 模型能够在保持高效推理速度的同时，显著增强场景理解和进行复杂操作的能力。为确保对视觉-语言-动作（VLA）模型研究的全面性与代表性，本文采用系统性文献检索策略。具体而言，检索范围覆盖 2018 年 1 月至 2025 年 12 月期间发表于 arXiv、Google Scholar、IEEE Xplore、ACM Digital Library 等平台的相关工作；最终纳入 95 篇核心文献作为本文的分析基础。该覆盖范围完整涵盖了图 1 所示的 VLA 在 2022~2025 年期间的代表性工作（RT-1、CLIPort 等），并在模态类型（含触觉、力、点云等扩展感知^[6]）、融合机制与数据集维度上保持广泛代表性，从而最大限度减少文献遗漏风险，保障综

表1 VLA代表模型中的多模态表示总结

序号	VLA 代表模型			多模态表示类型				模态编码技术			
	年份	模型	模型大小	视觉	文本	音频	触觉	视觉	文本	音频	触觉
1	2022	RT-1 ^[7]	35MB	✓	✓	×	×	FiLM-EfficientNet	Transformer	—	—
2	2023	RT-2 ^[8]	5B/12B/55B	✓	✓	×	×	ViT (PaLM-E)	Transformer (PaLM-E)	—	—
3	2023	VIMA ^[9]	2~200MB	✓	✓	×	×	Mask R-CNN+ViT	Transformer (T5)	—	—
4	2024	OpenVLA ^[10]	7B	✓	✓	×	×	ViT (DINOv2, SigLIP)	Transformer (LLaMA)	—	—
5	2024	RoboFlamingo ^[11]	3B/4B/9B	✓	✓	×	×	NFNet	Transformer (Chinchilla)	—	—
6	2025	VLAS ^[12]	7B/13B	✓	✓	✓	×	ViT (CLIP)	Transformer (CLIP)	Transformer (Whisper)	—
7	2025	ForceVLA ^[13]	3B	✓	✓	×	✓	ViT (SigLIP)	Transformer (SigLIP)	—	线性投影
8	2025	WorldVLA ^[14]	7B	✓	✓	×	×	VQ-GAN	BPE tokenizer	—	—

述结论的科学性与可靠性。

在具身智能 VLA 系统中，多模态感知是智能体理解环境的基础，而多模态表示是实现感知的第一步。其核心任务是从不同的模态输入信号中提取信息、构建语义表示，并为后续的多模态对齐与融合提供支撑。本文将具身智能中的模态类型总结为 4 种：视觉、文本、音频、触觉。不同模态通常借助神经网络编码器进行向量化编码，以获得各自的特征表示向量。以下将从不同模态出发，系统梳理当前的具身智能模型中的模态组成以及每一种模态在主流 VLA 模型中常用的编码器。此外，本文将目前主流的 VLA 大模型涉及的模态表示类型对应的编码技术做了详细总结，见表 1。

1.1 视觉模态

视觉感知是多模态具身智能中最基础的模态之一，被广泛应用于环境建模、目标识别、空间理解等任务。视觉模态通过视觉编码器将图像或视频输入转换为高维语义向量，为下游任务提供支持，如语义检索、物体操作规划等。在多模态 VLA 模型中，视觉信息通常通过视觉 Transformer (vision Transformer, ViT) 架构进行处理，从而获得编码后的视觉表示。通过对图像进行特征提取，视觉编码器能够从静态或动态图像中获取环境的详细信息，并将这些信息与其他模态（如语言或音频）进行融合，用于支持任务执行。常见的视觉编码器包括 CLIP^[15]、SigLIP^[16] 和 DINOv2 等视觉大模型^[17]。

CLIP 是由 OpenAI 提出的一种视觉-语言预训练模型^[15]，它将图像和文本映射到同一个嵌入空间，进行对比学习。CLIP 使用 ViT 或 ResNet 作为视觉编码器，将

图像特征与语言特征对齐，进而支持图像-文本检索等多模态任务。该模型在视觉和语言的多模态任务中表现出色，能够有效处理文本与图像之间的关系。

SigLIP 是基于 CLIP 的增强版本^[16]，通过优化损失函数，提升了模型在多模态任务中的表现，使其在处理图像和文本的联合表示时更加高效和准确。这使得 SigLIP 在具身智能应用中表现更为优异，增强了视觉-语言模型对多模态任务的适应能力，特别是在具有复杂物理交互的任务（如机器人操作）中。

DINOv2 则是一种自监督学习模型^[17]，采用 ViT 架构，通过自监督学习从无标签的图像中提取特征。DINOv2 能够有效地捕捉图像中的全局和局部信息，并生成高质量的图像表示。它特别适用于那些缺少标注数据的场景，在无监督任务中展现出强大的能力。DINOv2 在多模态任务中的应用，能够有效提升视觉特征的代表精度。

以上 3 类视觉编码器的核心差异与适用场景对比如下：CLIP 凭借视觉-语言预训练优势，在跨模态泛化任务中表现突出，但细粒度特征提取能力不足，适用于场景级理解；SigLIP 通过优化损失函数提升了噪声鲁棒性，更适配真实环境机器人操控；DINOv2 无需标注数据，在少样本场景中具备优势，但缺乏语言语义锚点，需配合后续对齐模块使用。技术选择本质是对“泛化性-精准性-数据依赖”的权衡。

这些视觉大模型通过不同的架构和训练策略，提升了 VLA 模型在多模态任务中的表现，使其能够更好地处理复杂的视觉信息，并与语言、触觉等其他模态进行融合。

1.2 文本模态

在具身智能中，文本模态主要指代通

过自然语言指令向智能体传达任务目标或操作需求的语言信息。文本指令以自然语言的形式向智能体提供明确的任务说明，并指导其执行相应操作。在这一领域，大语言模型被广泛应用于理解自然语言输入，并与视觉、动作模态进行深度融合。例如，Meta 提出的 LLaMA 系列开源大语言模型^[18]，以其高效的参数设计、强大的性能和灵活的适应性，在多模态系统中担任核心语言理解的角色。OpenAI 的 GPT 系列^[19]则作为典型的解码器结构大语言模型，在语言生成、问答和指令遵循等任务中表现突出，成为多个多模态系统的默认语言模块。而 BERT^[20]则采用掩码语言建模机制，基于双向 Transformer 与掩码语言建模（masked language modeling），在句法结构理解与语义匹配任务中具有优势，特别适用于语言-视觉对齐场景（如图文检索或指令-图像配对）。值得注意的是，文本在分层 VLA 架构中常被用作“通信原语”（communication primitive），用于连接高层任务规划器与底层动作控制器。例如，在 RT-1 系统中，采用双模型分层设计：上层模型将用户输入的复杂自然语言指令（如“把红色积木放到蓝色盒子左边”）分解为一系列原子级子指令（如“识别红色积木”、“移动至其上方”、“抓取”等）；下层模型则将这些原子指令转化为具体的关节动作序列。这种解耦设计不仅提升了系统处理长时序、复合任务的能力，还便于诊断语言理解是否出错，若执行失败，可回溯至高层指令解析环节进行修正。

此外，文本语言因其高可解释性与低标注成本，成为预训练视觉-语言模型（vision-language models, VLMs）最擅长处理的输入形式。GraspVLA 便利用冻结的预训练 VLM 直接从语言指令中推断目标物体的抓取语义（如“抓起那个易碎的

玻璃杯”），并生成空间抓取位置的隐式表征，从而显著降低底层动作生成模块对精确语言-动作映射的依赖，增强语义泛化能力。然而，尽管文本在高层意图表达上极具优势，但对于底层动作生成而言，纯文本指令仍存在“语义-动作鸿沟”。研究表明，直接从自然语言生成精细动作轨迹（如手指关节角度序列）的效率较低且不稳定。因此，近期工作开始探索更适配动作生成的中间表示形式。其中，一类方法引入未来动作轨迹作为条件信号：高层系统（S2）先预测一段粗粒度的长时域动作轨迹，再将其作为条件输入给底层系统（S1），指导其生成平滑、可行的执行动作；另一类方法则采用隐空间特征向量作为跨层通信原语，通过端到端梯度传播实现紧密耦合。

1.3 音频模态

音频模态则是指包含语音、语调、情绪等副语言信息的音频信号。在具身智能中，副语言信息为系统提供了丰富的语境、交互反馈和真实语音命令输入，进一步增强了系统的交互能力。近年来，随着研究的深入，更多学者开始探索将音频信号引入多模态具身智能系统，以提升其语境理解和任务执行的鲁棒性。OpenAI 提出的 Whisper^[21]基于 Transformer 架构，支持多语言和多口音的语音转文本，具备较强的鲁棒性和跨域能力，超越了传统的自动语音识别系统。Meta 开发的 HuBERT^[22]通过自监督学习，将音频信号离散化为“隐藏单元”，提升了语音表示的可迁移性与精度。另一重要的自监督语音学习模型 Wav2Vec 2.0^[23]，通过编码原始音频波形并预测遮蔽片段中的特征，深入理解音频中的语义和结构。

在近期的 VLA 模型中, 文献[12]提出的视觉-语言-动作模型 VLAS (Vision-Language-Action models with Speech) 首次尝试将语音模态与传统的视觉语言模态整合, 在 VLA 模型中加入音频感知模态, 以提升 VLA 模型在任务执行过程中的联合感知能力和环境理解水平。与以往依赖于文本指令或通过独立语音识别系统转换文本的 VLA 方法不同, VLAS 引入了端到端的语音-文本对齐机制, 允许模型在处理原始语音信号时保留语音韵律、情感等非语义信息。在 VLA 系统的技术集成中, 音频模态正从“独立预处理”向“端到端融合”演进。文献^[12]提出的 VLAS 首次实现语音模态与视觉-语言模态的深度整合, 其核心创新在于摒弃传统“语音转文本”的中间步骤, 通过端到端的语音-文本对齐机制, 在保留语音韵律、情绪等非语义信息的同时, 实现音频与视觉、语言特征的直接交互。这种集成方式使 VLA 系统在嘈杂环境中任务成功率提升 15%~20%, 尤其在精细操作场景中, 音频信号能辅助判断物体接触状态, 与视觉、触觉形成互补。此外, 大规模多模态数据集为音频模态的融合提供了数据支撑: RH20T^[12]⁴⁾包含同步的视觉、音频和动作信息, 涵盖 147 种任务的密集接触操作序列; ARIO 数据集^[25]整合了声音、文本、触觉等 5 种模态的 3M 条轨迹, 为音频与其他模态的协同训练提供了标准化数据, 推动 VLA 系统实现“视听触”多模态联合感知。

1.4 触觉模态

触觉模态通常通过机器人末端的力觉传感器或触觉皮肤装置进行感知, 用于建模机器人与环境之间的物理交互状态。在具身任务中, 触觉感知尤其重要于精细操

作场景, 如抓握、插接、装配等, 能够感知接触强度、物体硬度、滑动状态等物理信息, 从而指导动作调整。

当前 VLA 系统中, 触觉模态的编码与融合形成了两类主流技术路径: 一类是“图像化编码+视觉迁移”, 适用于输出图像形式数据的触觉传感器。例如 TLA 和 VTALA^[26]通过微调 Qwen-VL 的 ViT 模块, 直接从视触觉图像中提取特征, 借助预训练视觉编码器的强大表示能力, 实现高效跨模态迁移, 在物体区分、滑动检测等精细任务中性能显著提升; AnyTouch^[27]则提出统一 4 类视触觉传感器数据的对齐框架, 为不同硬件的触觉信号标准化提供了思路。另一类是“原始信号编码+专用融合”, 通过轻量化模型将力/力矩等连续信号转化为特征嵌入。Tactile-VLA^[28]则采用多层感知机作为编码器, 将连续触觉测量值处理为融合特征令牌, 在输入层面与视觉、语言令牌拼接, 实现多模态深度融合。

在已有研究中, ForceVLA^[13]是目前具有代表性的一次尝试。该工作通过传感器采集力学数据 ϕ_F , 然后通过线性投影将力学传感器采集的输入数据 f_{raw} 转化为表示力学模态信息的嵌入向量 E_F

$$E_F = \phi_F(f_{\text{raw}}) \quad (1)$$

然后, 在其与其他模态的向量融合后, 再输入混合专家模块, 从而动态地处理力学响应。该工作创新性地引入了力觉模态作为新的感知通道, 通过构建“力觉-视觉-语言”3 种模态感知结构, 使机器人在应对多变的接触情境时具有更好的动作自适应能力。力觉模态的引入, 不仅提升了系统对微小物理变化的阶段性响应能力, 也推动了多模态融合策略向更复杂的物理场景泛化。

1.5 动作模态

动作模态是 VLA 框架中区别于视觉、语言、音频与触觉的核心组成部分，它描述了机器人在具身环境中执行任务时产生的控制指令、操作策略以及行为序列。与触觉模态提供的“被动感知反馈”不同，动作模态是一种“主动交互模态”，反映了智能体的运动意图、行为规划及实时控制过程。动作模态不仅是智能体与环境交互的输出形式，还是状态迁移与长期规划的基础载体。因此，在具身智能系统中，动作模态的结构化表示、编码机制以及与多模态信息的融合方式，对于智能体能否成功执行任务、实现泛化学习与复杂策略推理至关重要。

1.5.1 动作模态的类型

根据控制粒度、语义层次与时间结构的差异，动作模态通常可分为 3 类：低维连续控制信号、离散语义动作 token 以及动作轨迹序列。

首先，低维连续控制信号是机器人系统中最常见的动作表示，包括机械臂关节角度、末端执行器姿态变化、速度控制、夹爪开合力度等^[7-8]。RT 系列即采用连续控制向量，使模型能够在每个时间步直接预测机器人下一步应执行的动作，实现端到端的视觉动作映射。连续控制空间具有高精度和实时性强的特点，被广泛用于真实机器人操控任务。其他方法（如 BC-Z^[29]和 PerAct^[30]）也采用连续动作空间，以适配高精度的抓取与姿态控制任务。

其次，离散语义动作 token 将动作抽象为可组合的高层操作符，如“抓起”“放置”“打开”“旋转”等，适用于多步骤任务规划或具有程序结构的操作描述。VIMA^[9]是此类方法的典型代表，其通过

“程序化动作表示”让模型从更高层次理解与生成操作步骤，从而获得更强的可组合性与跨任务泛化能力。类似思想也被用于 SayCan^[31]中，SayCan 通过将语言先验与高层动作技能结合，实现“语言-技能-控制”3 层映射。InnateCoder^[32]进一步强化了基于语言的动作 token 组织，使模型能够在复杂操作中生成可解释且可追踪的行为序列。

第三，动作轨迹表示视动作作为一个序列化的时间过程，适用于连续多步骤、复杂动力学或接触密集的操控任务。轨迹通常由多个连续动作向量组成，能够完整描述机器人从起始状态到目标状态的整个运动过程。ACT^[33]使用 transformer 结构捕捉动作轨迹中的长期依赖，而 Diffusion Policy^[34]、Trajectory Transformer^[35]、CP3ER^[36]则利用扩散模型直接生成高维动作轨迹，使模型能够在视觉与语言条件下生成稳定且连续的控制信号。部分研究，如 LMP (latent motor plans)^[37]，将轨迹进一步抽象为潜在技能，从而实现高效的分层控制与行为规划。

这 3 类动作类型覆盖了从低层实时控制到高层语义规划，再到长时序运动生成的完整需求，构成动作模态研究的基础分类。

1.5.2 动作模态的编码与表示

为了让 VLA 模型能够有效利用动作信息，动作模态需要被编码为与视觉、语言、触觉等模态一致的嵌入表示。本节从连续动作编码、动作 token 化以及轨迹建模 3 个角度展开论述。

• 连续动作向量的嵌入。在连续控制场景中，动作向量通常通过线性层或多层感知机 (multilayer perceptron, MLP) 投影到 Transformer 的语义空间，以便与视

觉或语言特征在统一表征维度下交互。RT-1 和 RT-2 采用此机制，将动作编码为固定维度的表示，从而实现决策预测与观察序列的紧密耦合。在 BC-Z、PerAct 等高精度操控任务中，连续动作向量被证明能够在弱监督或无监督条件下呈现出良好的泛化能力。世界模型路线（如 PlaNet^[38]、TD-MPC2^[39]）也常使用连续动作嵌入驱动潜在动态模型。

- 动作 token 化。为了让模型能够在统一的 Transformer 序列中表示动作，许多方法将动作离散化为 token，以类似于语言 token 的方式处理。VIMA^[9]将动作进行程序化表示，使其可与语言和视觉 token 一并输入模型。CLIPort^[40]和 Skill Transformer^[41]也采用 token 化方式，使技能模块能够通过 transformer 结构进行组合、重用与泛化。语言到动作模型（如 SayCan^[31]、RoboCLIP^[42]）则展示了动作 token 与语言先验结合在高层规划任务中的强大能力。

- 轨迹建模。轨迹是一类序列化动作表示，需要捕捉跨时间步的相关性^[43]。ACT^[33]使用自回归 transformer 结构对轨迹编码；Diffusion Policy^[34]、PAO-DP^[44]则采用扩散生成模型从条件模态直接生成轨迹。Decision Transformer^[45]展示了基于序列建模的强化学习形式，将轨迹表示为 token 序列进行“序列决策”。世界模型（如 DreamerV2^[46]、Genie^[47]、Video-GPT^[48]）则将轨迹视为环境动力学的一部分，通过学习

$$\mathbf{z}_{t+1} = f(\mathbf{z}_t, \mathbf{a}_t) \quad (2)$$

实现动作条件下的长期未来预测。其中， $\mathbf{z}_t$ 表示环境在时间 t 的潜在状态 (latent state)， $\mathbf{a}_t$ 为该时刻执行的动作向量， $f(\cdot)$ 为世界模型学习得到的隐变量动力

学函数，用于预测未来潜在状态 $\mathbf{z}_{t+1}$ 。此外，技能抽象方向（如 HiLLoC^[49]、LMP）通过将轨迹压缩为潜在技能嵌入，使模型能够在任务空间中进行高效规划与泛化。

1.5.3 动作模态与多模态融合

动作模态并非孤立存在，而是与视觉、语言、触觉等模态深度交互，共同构成完整的多模态具身智能系统。融合方式可分为视觉 - 动作、语言 - 动作以及分层动作表示 3 个典型路径。

- 视觉 - 动作融合。视觉是动作生成的核心条件之一^[50]，视觉 - 动作融合主要关注如何从视觉特征推断下一步控制信号。RT-1、RT-2 通过视觉编码器获取场景表示，再根据任务目标生成动作。PerAct 使用体素化 3D 注意力机制实现视觉→动作映射，从而提高抓取与操控鲁棒性。CLIPort 则将视觉特征映射到“可操作区域”，再预测对应动作，展示了跨任务泛化能力。

- 语言 - 动作融合。语言为动作提供高层语义约束，使模型能够执行“指令驱动”的复杂任务。SayCan 通过将语言嵌入作为动作策略条件，使模型能够实现从文本任务描述到高层技能再到低层动作的 3 层衔接。RoboCLIP^[42]通过语言-视觉-动作对齐实现零样本操控。Skill Transformer^[41]与 LMP^[37]展示了语言与技能嵌入之间的关联，使模型能够处理多步骤指令与组合任务。

- 分层动作表示。在具身智能任务中，高层行为意图与低层控制命令通常具有不同的时间尺度，因此动作模态常以分层方式组织：高层策略（如技能）承担动作序列的语义规划，而低层控制指令负责具体的运动执行。BC-Z^[29]通过技能库定义高层行为，再由低层控制网络完成执行；

Long-VLA^[49]通过输入级自适应策略划分移动交互阶段、应用相位-觉察输入掩码聚焦阶段的相关线索，实现长时序行为压缩；LMP^[37]将长时序行为压缩为潜在技能，以便在高维决策空间中实现高效规划；GROOT^[51]则展示了多技能调度在复杂任务中的应用，增强了跨任务迁移与可组合性。在世界模型方法中，Dreamer、Genie、VideoGPT等常将高层策略建模为潜在技能，以在较长时间范围内引导未来动作序列生成，实现任务级规划与长时预测。整体来看，分层动作结构比单纯的低层控制更适合复杂具身任务，并能够更有效地与视觉、语言等模态协同，为高层次行为组织能力提供支撑。

综上，动作模态与多模态融合机制共同构成具身智能的“执行核心”，实现视觉理解、语言指令解析与物理交互的闭环连接，也是VLA系统区别于传统VLM的关键特征。

2 多模态对齐与融合

2.1 面向VLA的多模态感知统一框架

针对现有综述多按模态类型罗列、缺乏系统性分析逻辑的局限，本文提出一个

面向视觉-语言-动作（VLA）系统的多模态感知统一框架。该框架并非旨在提出一种新的网络架构，而是从功能解耦与信息处理流程的角度，为理解和设计VLA感知系统提供一个通用的、任务导向的分析透镜。

如图2所示，本框架将VLA的多模态感知过程明确划分为3个依次递进且紧密耦合的核心环节：多模态表示（multimodal representation）→多模态对齐（multimodal alignment）→多模态融合（multimodal fusion）。这一划分不仅反映了信息处理的内在逻辑，也为后续的技术选型与性能评估提供了清晰的坐标系。

多模态表示（multimodal representation）：作为具身智能感知的起点，其目标是将来自异构传感器（如摄像头、麦克风、力觉传感器）和指令源（文本、语音）的原始信号，通过专用编码器（如ViT、Whisper、MLP）转化为高维、结构化的特征向量。此阶段关注的是单模态内部的语义提取能力，其质量直接决定了后续处理的上限。例如，采用DINOv2还是CLIP作为视觉编码器，会带来不同粒度的环境理解。

多模态对齐（multimodal align-

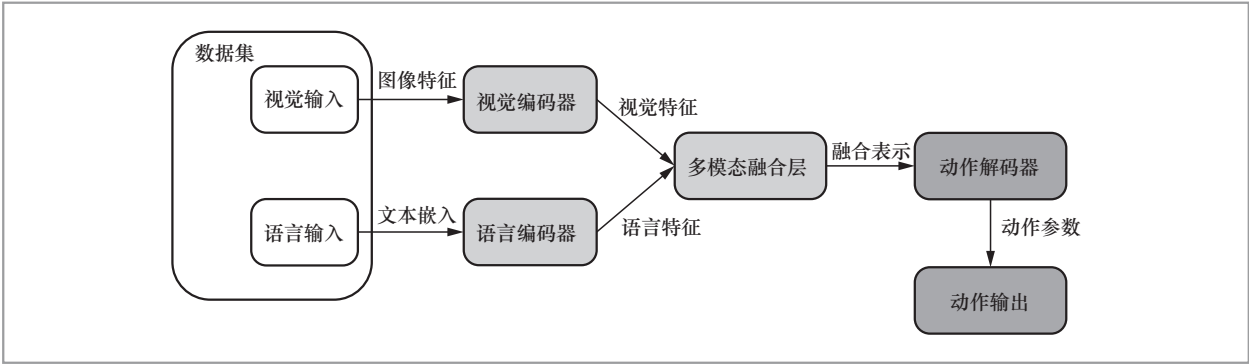


图2 VLA中的多模态感知与融合

nment)：由于各模态的数据分布、特征空间和语义粒度存在天然差异，直接融合会导致“鸡同鸭讲”。对齐环节旨在将不同模态的特征映射到一个共享的、语义一致的潜在空间中^[52]。这可以通过轻量级的 MLP 投影实现几何对齐，也可以借助 CLIP 等预训练模型提供的强大语义锚点实现深度语义对齐。对齐的质量决定了多模态信息能否被模型正确关联，也是泛化能力的关键。

多模态融合 (multimodal fusion)：在完成对齐后，融合环节负责将来自不同模态的特征进行交互、整合与增强，形成一个统一的、富含上下文信息的联合表征，以供下游的动作决策模块使用。融合策略的选择 (如拼接、交叉注意力、FiLM) 直接体现了系统在计算效率、语义深度与实时性之间的权衡。

本框架的独特性在于揭示了上述技术间的耦合关系。该框架清晰地表明，表示、对齐、融合三者并非孤立，而是相互依赖。例如，强大的预训练表示 (如 CLIP) 可以简化对齐甚至融合的设计；反之，复杂的融合机制 (如交叉注意力) 可以在表示较弱时进行补偿。同时本文提供了系统设计指南：通过将复杂系统分解为这 3 个环节，研究者和工程师可以更有针对性地诊断性

能瓶颈 (如表示不足、对齐失败、融合低效)，并据此选择合适的技术路径。

在具身智能中，来自视觉、语言、触觉、音频等模态的数据具有显著的异构性，如何在感知阶段实现不同模态间的语义对齐与有效融合，是构建稳健多模态系统的关键挑战。融合质量直接影响智能体对环境的理解深度以及后续行为的准确性。在当前典型的 VLA 架构中，感知模块往往需要将视觉与语言两类核心模态进行对齐与融合，形成统一的语义表示以指导动作生成，如图 2 所示。此外，越来越多的研究尝试在视觉-语言模型的基础上引入音频、触觉等新模态以提升系统的多感知能力，进一步推动多模态对齐与融合成为 VLA 系统的核心技术之一。本文在此基础上将系统梳理当前 VLA 中常用的多模态对齐与融合技术，并总结主流的 VLA 模型常用的对齐与融合技术^[4,7-13]，见表 2。

2.2 多模态对齐

多模态对齐旨在解决不同模态在数据分布、特征维度以及语义粒度上的差异，将其有效映射到统一的潜在空间，以便后续联合建模，如果没有适当的对齐，VLA

表2 VLA代表模型中对齐与融合技术总结

序号	VLA 代表模型			多模态对齐技术		多模态融合技术		
	年份	模型	模型大小/B	多层感知机投影	预训练编码器对齐	注意力机制融合	拼接嵌入	FiLM 融合
1	2022	RT-1 ^[7]	0.035	×	✓	×	×	✓
2	2023	RT-2 ^[8]	5/12/55	×	✓	×	✓	×
3	2023	VIMA ^[9]	2~20	✓	×	✓	×	×
4	2024	OpenVLA ^[10]	7	✓	✓	×	✓	×
5	2024	RoboFlamingo ^[11]	3/4/9	×	×	✓	×	×
6	2025	VLAS ^[12]	7/13	✓	✓	×	✓	×
7	2025	ForceVLA ^[13]	3	✓	✓	×	✓	×
8	2025	WorldVLA ^[14]	7	×	✓	×	✓	×

模型的推理准确率可能会下降。当前 VLA 中常用的模态对齐方法主要可分为以下两种：

2.2.1 多层感知机投影

多层感知机（MLP）投影是实现多模态对齐的经典方法之一。其核心思路是通过多层感知机的神经网络线性层将不同模态（如视觉和语言）的编码特征统一映射到相同维度的向量空间中。

具体而言，视觉模态的特征通常通过卷积神经网络（Convolutional Neural Network，CNN）或 ViT 等视觉编码器提取，而语言模态的特征通过 GPT、BERT 等语言模型获得，再分别通过 MLP 进行投影，从而对齐它们的特征维度。许多 VLA 模型^[8-13]（如 OpenVLA）会在视觉编码器和大语言模型 Llama 2 之间插入 MLP 投影层，把编码后的视觉向量进行投影，转换到和语言向量相同的维度，从而向语言模态对齐。

MLP 投影的优点在于结构简单、计算开销低，易于在多模态系统中部署。特别是在需要快速处理高维特征的场景下，多层感知机能够有效地将不同模态映射到统一表示空间。然而，由于其仅在几何层面进行映射，缺乏对模态间深层语义关系的显式建模，因此在复杂的推理任务中往往需要结合注意力机制或更强的交互模块以弥补不足。

2.2.2 预训练编码器对齐

另一种被广泛采用的对齐方式是依托预训练多模态编码器提供的共享语义空间，实现不同模态的对齐。具体做法如下，利用已经在大规模多模态数据（如图文对）上经过对比学习等方式预训练好的模型

（如 CLIP），通过其视觉编码器和文本编码器分别将图像和文本输入编码为同一语义空间中的向量。这样就不需要在下游任务中重新学习模态对齐，可直接获得语义一致的特征表示。

以 OpenVLA 为例，该模型采用 CLIP 作为视觉模态的编码器之一，CLIP 在预训练阶段已通过大规模图文配对数据和对比损失，使视觉与文本特征在同一向量空间中对齐。因此，VLA 系统可以直接利用 CLIP 提供的编码向量，方便地进行模态融合及后续的策略学习和动作预测。这种方式显著降低了多模态对齐的学习成本，在多任务迁移和跨模态泛化方面表现尤为突出。

2.3 多模态融合

在完成模态对齐后，如何有效地将来自不同模态的表示进一步整合，以形成丰富的联合表征，是多模态感知的另一个核心问题。根据当前主流实践，融合方法主要分为 3 类。

2.3.1 拼接嵌入

拼接嵌入（Concatenation）是一种实现最为直接的多模态融合方式^[7,9,11-13]。在该策略中，来自不同模态的编码向量被直接沿特征维度拼接，形成一个联合模态表示，然后输入至后续的策略网络或动作生成模块。这种方式不涉及显式的注意力或交互机制，因此结构简单，计算开销低。

尽管拼接融合缺乏对模态间细粒度语义关系的建模能力，但在大规模模型参数和充足训练数据的支持下，仍能在多模态任务上取得接近甚至媲美交叉注意力结构的性能。常见的 VLA 模型（如 OpenVLA）以拼接嵌入作为核心多模态融合策略^[9]，

实现了对视觉和语言模态的高效整合，并在多模态指令理解、环境感知以及跨任务迁移等任务中表现优异。

2.3.2 注意力机制融合

注意力机制融合则通过引入自注意力机制^[53]或交叉注意力机制^[54]，更加细粒度地捕捉不同模态间的依赖关系和交互信息。

自注意力通常将不同模态的序列直接混合输入统一的 Transformer，通过序列间的全局注意力隐式学习模态关系。而交叉注意力则显式地指定一个模态作为 query，另一个模态作为 key 与 value，通过注意力计算实现高精度的语义对齐与信息聚合。

这类方法能够精准地对齐并融合多模态信息，尤其适用于含复杂操作指令、条件判断或多步骤规划的场景。例如，RoboFlamingo^[11]通过在语言序列中插入交叉注意力层，使其能够显式关注视觉特征，从而大幅提升 few-shot 推理和复杂交互理解能力。需要注意的是，这类机制的计算开销通常较高，适合在算力资源较为充足的环境下部署。

2.3.3 FiLM 融合

特征级线性调制（feature-wise linear modulation, FiLM）是一种轻量级的多模态条件融合方法^[55]。FiLM 先从条件模态（如语言）生成一组缩放和偏移参数，再对另一模态（如视觉）的中间特征进行逐通道的线性调制，显式地在每个通道层级引入多模态信息，使模型能够有选择地增强或抑制视觉特征以匹配语言语义，从而在特征层面对模态进行融合。其数学原理如下：

$$\text{FiLM}(F|h) = \gamma(h) \odot F + \beta(h) \quad (3)$$

假设将视觉模态作为条件模态，则 $F \in \mathbf{R}^{C \times H \times W}$ 是被调制的视觉特征图，如 CNN 或 Transformer 的输出，而 h 则是从另一模态（如语言）获得的全局上下文向量。 $\gamma(h) \in \mathbf{R}^C$ 和 $\beta(h) \in \mathbf{R}^C$ 分别是根据条件模态生成的逐通道缩放和偏移参数，通常通过多层感知机层来预测。

Google 提出的 RT-1^[7]在视觉编码器 EfficientNet 中引入了 FiLM 层，这些 FiLM 层通过语言模块生成的参数对视觉特征逐通道进行操作，从而让视觉输入与语言指令更紧密地对齐。

尽管视觉与语言构成了当前 VLA 系统的主干，但触觉与音频模态在实现高保真、鲁棒的具身交互中具有不可替代的作用。它们并非简单的感知补充，而是提供物理接触状态、材料属性、操作反馈与环境动态性等关键信息的直接通道。然而，若仅将触觉/音频视为“额外输入”，而不明确其在 VLA 信息流中的结构化角色，易导致模态冗余或集成低效。本文基于上述“表示-对齐-融合”统一框架，系统阐述触觉与音频模态的技术整合路径。在开环 VLA 系统中，动作由视觉-语言指令一次性生成，但缺乏执行过程中的实时校正能力。而触觉信号（如力/扭矩、滑动检测、接触点分布）和音频信号（如碰撞声、摩擦声、物体形变声）天然具备时序性和事件触发性，可作为动作执行质量的即时反馈，驱动闭环控制。例如，在精细抓取任务中，即使视觉无法分辨玻璃杯是否打滑，高频触觉传感器可检测微滑动并触发握力调整；在非预抓取操作（如推动物体）中，音频特征可辅助判断物体是否卡住或倾倒。

在 VLA 架构中，触觉与音频模态通常通过表示、对齐与融合 3 个阶段嵌入。首

先，在表示阶段，原始数据被编码为特征向量：触觉数据（如 GelSight 图像或力阵列）通过 CNN 或 MLP 提取结构化特征，而音频信号则经短时傅里叶变换或 wav2vec2 等预训练编码器转化为时频或语义表示。随后，在对齐阶段，这些特征需与视觉-语言联合表征进行跨模态对齐。例如，在 DexGraspVLA^[56]中，触觉特征通过投影层映射到 CLIP 共享语义空间，使得“看到光滑表面”能与“预期需要更大摩擦力”以及“检测实际滑动”的触觉反馈形成语义一致性；类似地，音频事件（如“咔哒”声）可与语言指令中的“确认卡扣到位”精确对齐。最后，在融合阶段，对齐后的触觉与音频特征作为条件信号注入动作生成模块，在扩散策略或自回归动作头中作为时间相关的上下文条件，对动作生成的概率分布进行动态调制。以 GO-1^[57]为例，其在底层控制器中引入触觉残差作为扩散模型的 conditioning 输入，显著提升了在未知材质上的抓取成功率。

2.4 对齐与融合技术的对比

随着 VLA 模型在规模、任务复杂度与多模态类型上不断扩展，其在“模态对齐”与“模态融合”层面的技术路线也呈现出高度多样化，从最初的轻量级 MLP 投影到利用大规模预训练语义空间的 CLIP/BLIP 类方法，再到提供细粒度语义交互的交叉注意力结构。为了揭示不同方法的技术优势、性能差异以及应用场景，本节从横向视角总结主流 VLA 体系中的对齐与融合技术，并结合近期在真实机器人评测中的基准测试表现，对技术路线的演化趋势进行系统分析。

2.4.1 模态对齐技术的横向比较与适用场景

模态对齐旨在将视觉、语言、音频、触觉等多模态输入映射到语义一致的特征空间中。当前对齐技术大致可分为 3 类：MLP 投影、依赖预训练多模态编码器的对齐策略，以及基于交叉注意力的深语义对齐。这 3 类方法在计算成本、语义精度与泛化能力上各具特色。

MLP 投影是最常见的对齐方式之一，因其计算开销低、易于训练，并且对新增模态（如音频、触觉）具有较好的扩展能力而被广泛采用。OpenVLA^[10]、VLAS^[12]与 ForceVLA^[13]等近年来的模型均在视觉与语言之间采用 MLP 映射技术。在大模型强大的表达能力支持下，OpenVLA 等模型在 BridgeData V2 上展现出较高成功率，均在 60% 以上^[10]，说明使用轻量 MLP 对齐和容量大的大模型也具备实用性。此外，MLP 投影也被用于视觉与触觉等物理模态的对齐，其跨模态适应性在 VTLA、VTG-Grasp 等方法中被进一步验证^[26]。此外，在更基础的视觉自监督方面，MAE^[58]、iBOT^[59]、SimCLR^[60]、EVA-02^[61]等视觉预训练模型为 VLA 的高质量视觉表征提供了可靠基础^[62]。这些模型无需人工标注即可学习鲁棒的视觉语义，为跨任务、跨场景的泛化能力提供底层支持。

交叉注意力是一种能够显式对齐不同模态语义内容的结构，适用于复杂任务规划、多步骤推理及长指令解析。RoboFlamingo^[11]、Flamingo^[63]、CoCa^[64]等模型均依赖交叉注意力，使语言 token 能够主动查询视觉补丁，从而大幅提升零样本和少样本表现。研究表明，在 Meta-World 等数据集上，交叉注意力可带来 10%~20% 的性能提升^[11, 63]。然而，该结构的计算成本显著高于 MLP 与预训练对齐

策略，因此在需要实时控制或高吞吐训练的任务中尚难以大规模部署。

总体而言，MLP 投影适合工程部署与模态扩展，预训练对齐适合泛化需求强的 VLA 模型，而交叉注意力对齐则适用于语义推理要求高、任务结构复杂的场景。

2.4.2 模态融合技术的横向比较与适用场景

模态融合是 VLA 感知阶段的核心环节，其目标是将已对齐的特征整合为统一的语义表示。当前技术路线包括 拼接融合、注意力融合与特征调制类融合（如 FiLM）。

拼接融合因其简单、高效且易于扩展多模态通道而成为大规模 VLA 的主流方式。OpenVLA、VLAS 与 ForceVLA 均在视觉 - 语言 - 触觉/音频融合阶段采用拼接策略。实验显示，即便缺乏显式交叉注意力，拼接融合在大模型表达能力的加持下，仍然能够在 BridgeData V2 上达到与交叉注意力模型相近的成功率。MineDojo^[65]、RT-X^[66]、R3M^[67]等模型方法也证实，在数据量足够大的情况下，拼接融合的效果可以与更复杂的融合结构媲美。

注意力融合包括“统一 Transformer 融合”与“交叉注意力融合”两种形式，是当前视觉 - 语言模型（VLM）中的核心技术路径。CoCa、LLaVA、Kosmos、

Gato 等模型^[63, 68-69]通过注意力机制建模多模态之间的细粒度语义关系，使高层推理能力显著增强。在 RoboFlamingo^[11]中，这一策略使其少样本能力得到显著提升，同时在长序列视觉条件任务中表现优于拼接模型。因此，注意力融合适用于多步骤规划、语义复杂指令、多轮操控决策等任务。然而，由于每次注意力计算的复杂度随序列长度平方增长，注意力融合在实时机器人控制（如 30~50 Hz 机械臂控制）中仍存在部署难度。

特征调制类方法以 FiLM^[55]、Conditional BatchNorm^[70]、HyperNetworks^[71]与 AdapterFusion^[72]为代表，通过从语言或其他模态生成一组缩放系数对视觉或触觉特征进行逐通道调制。RT-1 是应用该策略的典型模型，它在视觉感知流程中插入 FiLM 层，实现了极低计算开销下的模态融合，使其能够在真实机器人上维持接近 30 Hz 的实时推理速度。类似地，在可变形物体操控、接触丰富的多步骤任务中，基于调制的方法仍具有优越能效比，VTLA^[26]、VIMA^[9]与 ACT^[33]等模型的部分模块也采用了此类策略。

为解决性能评估缺乏统一标准的问题，本文基于同一数据集（RoboEval^[73]）、同一核心指标（任务成功率），对 3 类融合技术进行横向对比，结果见表 3。

在涉及精细物理交互与多模态协同的双臂任务中，FiLM 融合比拼接嵌入和交叉注意力融合的任务成功率都要高，核心原因是其显式捕捉了模态间的细粒度语义依赖，对环境变化、任务变异的适应性更强。上述结果表明融合技术的选择需紧密结合任务需求：复杂多步骤任务优先选择交叉注意力融合，实时控制任务适配 FiLM 融合，大规模通用 VLA 模型可采用拼接嵌入以平衡性能与效率。

表3 3类融合技术横向对比

融合策略	VLA 代表模型	任务成功率
拼接嵌入	OpenVLA	11.6%
交叉注意力融合	RoboFlamingo	21.1%
FiLM 融合	ACT	39.7%

综上，拼接融合适合大模型与多模态扩展；注意力融合适合高语义需求场景；而 FiLM 融合适合资源受限或实时推理场景。

2.4.3 技术路线的演进与趋势分析

从 RT-1（2022）到 WorldVLA（2025）^[14]，VLA 在对齐与融合技术上形成了清晰的演进脉络，可归纳为以下 3 个趋势：

第一，从结构驱动向预训练驱动演进。最初的 RT-1、VIMA 依赖结构创新（如 FiLM、交叉注意力）实现多模态对齐与融合。然而自 RT-2 起，预训练视觉-语言模型^[15]（CLIP、PaLI、BLIP-2）^[74-78]成为主流，使 VLA 直接继承大规模图文语义空间，显著提升零样本与跨任务泛化能力。

第二，从显式交互结构向“大模型+轻融合”发展。2024—2025 年出现的 OpenVLA 等展示了大模型在感知与对齐上的强大潜力：即使采用轻量级的 MLP 与拼接结构，也能取得与传统交叉注意力模型相近甚至更优的性能。这一趋势源于大模型自身就具备足够的隐式跨模态建模能力，使结构复杂度不再是感知性能的决定因素。

第三，向世界模型与更丰富模态的融合发展。随着具身智能任务复杂度的提升，越来越多研究探索基于视频生成与环境预测的世界模型路线^[79]。DreamerV2、Genie 以及 WorldVLA 等模型^[14]展示了未来趋势，即 VLA 将不仅是“多模态感知+动作预测”，而是“多模态世界建模+长时预测+动作规划”。同时，DROID、MineDojo 与 Ego4D 等数据集^[26,65,80-81]拓展了触觉、音频、物理状态等模态的应用边界，展现出 VLA 向‘全模态具身智能’发展的广阔前景。

2.5 模态对齐策略的量化对比

为系统评估不同模态对齐策略在 VLA 系统中的权衡关系，本文构建了一个涵盖对齐质量、计算开销与数据效率的统一评估框架，并在 RT-1-X、OpenVLA、BridgeData 等公开基准及典型实现上进行了量化对比。

在评估多模态模型性能时，主要从 3 个关键维度进行考量：对齐质量、计算开销和数据效率。对齐质量采用下游任务成功率作为代理指标，基于 RT-1-X 基准中的“端到端任务完成率”来衡量，涵盖多物体操作和长指令遵循等复杂场景，高质量的对齐确保了语言指令与视觉状态在潜在空间中的语义一致性，有效减少了动作决策时的歧义。计算开销则通过推理延迟（ms/step）和参数量增量（M）来评估，前者直接决定了模型在机器人部署中的实时性表现，后者反映了模型的复杂度和内存占用情况，所有对比均基于相同的 A100 硬件平台和输入规格进行复现以确保公平性。数据效率采用定性分级评估，分为“低中高”3 个等级：低效率意味着可以使用 CLIP 等冻结的预训练模型而无需额外对齐数据；中效率需要少量配对数据进行 MLP 微调；高效率则需要大量多模态轨迹来训练端到端的注意力机制。

3 数据集与仿真环境

在具身智能中的多模态感知研究中，数据集不仅是验证算法有效性的重要工具，更是驱动模型学习多模态协同表征与推理的关键基石。相比于传统的单模态任务，VLA 模型通常需要同时处理视觉、语言以及动作（或更多模态）信息，因此对数据集提出了更高要求。高质量的具身智能数

数据集应该包含多模态、多任务^[82]，能够让具身模型学习来自不同模态的数据信息，并且能够反映真实环境中机器人任务分布的复杂性与多样性。此外，还需要区分真实场景和仿真场景的数据集，因为从仿真到真实数据场景的迁移对实际部署具身智能模型至关重要。

3.1 高质量VLA数据集的衡量指标与判断依据

为客观评估数据集质量，需从“多模态协同性”“数据可靠性”“任务适配性”3个核心维度构建衡量指标体系，具体如下：

（1）模态完备性：指数据集涵盖VLA任务所需核心模态（视觉、语言、动作）及扩展模态（音频、触觉、力觉）的全面程度，直接决定模型多模态感知的覆盖范围。完备的多模态数据能为模型提供互补信息，例如触觉+视觉数据可解决遮挡场景下的精细操作问题，音频+语言数据能提升噪声环境中的指令理解鲁棒性。

（2）跨模态对齐误差：指不同模态数据在时间戳、语义关联上的匹配精度，是多模态融合的基础。对齐误差通常以时间同步偏差（如视觉帧与力觉信号的时间差不大于0.1s）或语义一致性（如语言指令与视觉目标的匹配度）衡量，低对齐误差能避免模型出现语义混淆，提升任务执行准确率。

（3）数据噪声水平：包括传感器噪声（如视觉模糊、触觉信号干扰）、标注噪声（如语言指令歧义、动作标签错误）等，直接影响模型训练的稳定性。高质量数据集需通过传感器校准、多轮标注校验等方式控制噪声，例如真实场景数据集需过滤机械臂运动中的电磁干扰信号，仿真数据集

需优化渲染精度以降低视觉噪声。

（4）场景与任务多样性：以场景熵值、任务覆盖度为量化指标，反映数据集对真实环境复杂性的刻画能力。场景多样性越高（如涵盖不同光照、障碍物分布的场景），任务覆盖度越广（如包含抓取、装配、搬运等多种操作），模型的泛化能力越强，越能适应未见过的新场景。

（5）泛化支撑能力：指数据集训练的模型在跨场景、跨任务或跨机器人平台上的性能迁移效果，可通过模型在陌生数据集上的任务成功率衡量。高质量数据集需具备一定的分布代表性，避免过度聚焦单一场景，才能支撑模型从仿真到真实、从特定任务到通用任务的迁移。

3.2 主流VLA数据集的质量分析与特性梳理

基于上述指标，本节总结了主流VLA模型常用的数据集，并从多模态感知的角度对近年出现的数据集进行总结分类，具体见表4。CALVIN数据集在2022年被提出^[83]，是一个开源的仿真数据集，专注于多步骤桌面操作任务，包含视觉、语言与触觉模态，覆盖4个场景、总计约240万时间步。近期，引入音频模态并融合到视觉-语言-动作模型的VLAS模型使用了CALVIN数据集作为基准测试数据集。

2023年，多模态数据集迅速扩展。一方面，RH20T^[24]引入真实传感器采集数据，包括丰富的触觉、音频等模态，以及人机对齐的演示视频。另一方面，Open X-Embodiment^[66]涵盖了22台机器人、527种技能、160 266个任务以及200万条任务片段，是一个具有代表性的机器人数据集。同时，LIBERO^[84]在仿真平台中提供超过100个任务和语言指令的演示，为持续学习和任务泛化提供了

表4 VLA常用数据集

数据集	年份	感知模态	数据来源	数据类型	数据量	任务数
CALVIN ^[83]	2022	视觉、语言、触觉	仿真	本体状态、多模态感知数据、语言指令等	240万	34
RH20T ^[24]	2023	视觉、语言、音频、触觉	真实	本体状态、点云、多模态感知数据等	11万	140+
Open X-Embodiment ^[66]	2023	视觉、语言	真实、仿真、开源数据集	运动信息、多模态感知信息等	200万+	16万+
LIBERO ^[84]	2023	视觉、语言	仿真	图像、本体状态、任务语言描述	—	130
Touch100k ^[85]	2024	视觉、语言、触觉	真实、仿真、开源数据集	视觉-触觉配对图像、触觉感知图、任务语言描述	10万	—
ARIO ^[25]	2024	视觉、语言、音频、触觉	真实、仿真、开源数据集	运动信息、多模态感知信息等	300万+	30万+
BridgeData V2 ^[86]	2024	视觉、语言	真实	图像、本体状态、语言指令等	6万+	—
SynGrasp-1B ^[87]	2025	视觉、语言、触觉	仿真	图像、抓取轨迹、语言指令等	10亿	—

稳固基础。

2024 年，触觉话题获得更多关注。Touch100k^[85]针对 GelSight 触觉场景构建视觉-触觉-语言三模态数据库，为触觉驱动表示学习提供大规模真实数据。与此同时，ARIO^[25]汇聚了多房间、多场景的触觉与语言数据样本，为抓取与任务规划等场景提供支持；BridgeData V2^[86]则回归真实 WidowX 机械臂平台，采集 6 万多条演示轨迹与自然语言指令，大幅提升真实环境中的多模态感知与任务分布代表性。斯坦福大学等团队提出的 OpenVLA 模型选择了 BridgeData V2 作为基准测试集，对模型进行了全面测试。

最后，2025 年建立的 SynGrasp-1B^[87]代表了仿真大数据范式的升级，它在逼真的环境中合成了 10 亿帧抓取动作，用于训练通用 VLA 抓取模型并验证其仿真到真实的泛化能力。

3.3 场景熵值与任务覆盖度

为系统评估 VLA 模型在复杂、开放环境中的泛化能力与任务理解广度，本文引入两个量化指标：场景熵值（scene entropy）与任务覆盖度（task coverage）。二者分别从环境复杂性与指令多样性角度刻画测试集的挑战性，具体定义如下。

3.3.1 场景熵值

真实世界操作场景常包含多个物体、杂乱布局、遮挡与背景干扰^[88]。为量化场景的视觉-语义复杂度，借鉴信息论中熵的概念衡量场景中物体类别与空间分布的不确定性。设测试场景集合为 $S=\{s_1,s_2,\cdots,s_N\}$ ， N 代表测试场景的总数量。 $O_i=\{o_1,o_2,\cdots,o_{M_i}\}$ 表示场景 s_i 中检测到的物体集合， M 为场景 s_i 中检测到的物体的数

量。每个物体 o_j 具有类别标签 $c_j \in C$ (C 为预定义物体类别集)。定义场景 s_i 的类别分布概率为:

$$p_k^{(i)} = \frac{|\{o_j \in \mathcal{O}_i | c_j = k\}|}{|\mathcal{O}_i|}, \quad \forall k \in C \quad (4)$$

则该场景的场景熵值定义为:

$$H(s_i) = - \sum_{k \in C} p_k^{(i)} \log p_k^{(i)} \quad (5)$$

其中, 约定 $0 \log 0 = 0$ 。整个测试集的平均场景熵值为:

$$\bar{H} = \frac{1}{N} \sum_{i=1}^N H(s_i) \quad (6)$$

3.3.2 任务覆盖度

VLA 系统的泛化能力不仅取决于环境复杂度, 还依赖于其能正确执行的自然语言指令范围。为衡量测试指令集的语义广度, 定义任务覆盖度以评估指令在动作语义空间中的分布密度。设测试指令集为 $T = \{t_1, t_2, \dots, t_K\}$, K 代表测试指令集中指令的总数量。首先, 使用预训练大语言模型 (如 Sentence-BERT^[20]) 将每条指令 t_j 编码为语义向量 $v_j \in \mathbf{R}^d$ 分别是根据条件, d 代表语义向量的维度。然后, 在该语义空间中构建一个离散化网格或使用聚类中心 (如 K-means, K 个簇) 代表潜在的任务语义类别。 n_c 为第 c 个簇中包含的指令数量, 其中 $C = 1, 2, \dots, K$ 。任务覆盖度定义为非空簇的比例。

$$\text{TaskCoverage} = \frac{1}{K} \sum_{c=1}^K \mathbb{I}(n_c > 0) \quad (7)$$

其中, $\mathbb{I}(\cdot)$ 为指示函数。该值越接近 1, 表示指令覆盖的语义区域越广。若系统输出为动作轨迹 a_j , 也可在经 VAE 编码得到的动作嵌入空间中计算类似覆盖率, 以反映行为多样性。

4 现有挑战与机遇

尽管当前已有多种融合与对齐机制在 VLA 系统中得到实践验证, 并在多个具身任务上取得良好性能, 但 VLA 中的多模态感知依然面临诸多核心挑战。

4.1 多模态的泛化问题

多模态模型虽然在特定基准或特定领域内表现出较好的性能, 但在多样化上下文或任务中的泛化能力仍然有限^[89-90]。研究人员必须探索方法以提升这些模型的迁移性和适应性, 从而应用于现实世界。尽管已有世界模型的概念和相关的研究工作被提出, 但是这些模型可处理的任务范围相对固定, 其泛化能力仍不理想。

尽管现有模型在熟悉的环境或与其训练场景相似的任务中表现出色, 但遇到完全不同的任务或不熟悉的变体时, 其性能会显著下降——有时甚至下降了多达 40%^[2]。这源于模型对狭窄分布的训练数据过拟合以及对任务多样性的表示不充分。VLA 的泛化能力差主要体现为当前的 VLA 模型在零样本或少样本学习场景中的表现不佳, 影响了它们面对不同任务和输入干扰的适应性和扩展性。

4.2 灾难性遗忘

灾难性遗忘是多模态感知系统中持续学习面临的主要瓶颈^[91]。灾难性遗忘即是在多任务、多模态环境中, 模型在适应新模式或新任务的微调阶段, 往往会遗忘原本在预训练阶段习得的知识, 特别是在长期训练或动态任务切换过程中表现明显, 从

而导致模型在各个模态上的基本性能严重衰退。许多研究指出，当前 VLA 模型的灾难性遗忘是一个普遍存在的迫切问题。

目前已经有研究提出 ReVLA^[91]，通过一种模型合并策略恢复在微调过程中丢失的域外视觉鲁棒性，在整合新知识的同时保留旧知识，以抵抗灾难性遗忘问题。该研究能够将域外抓取任务的成功率提高到 77%^[2]，但会增加计算量和模型复杂性。如何更好地解决 VLA 中由多模态感知引起的灾难性遗忘问题，是未来一个有潜力的研究方向。

4.3 仿真与真实环境差异

仿真与现实之间的迁移是目前 VLA 中仍难以解决的问题^[92–94]。尽管大量 VLA 模型在仿真平台中表现优异，但迁移到真实环境的系统后，往往会因感知精度下降、传感器分辨率不同或环境复杂性提升而出现显著性能衰减。如何提升模态感知与融合在现实场景下的稳健性，是具身智能落地应用必须突破的关键难题。

4.4 关于触觉模态的探索

现有的主流 VLA 模型^[7]都没有触觉模态，都是通过视觉来感知外部物理信息，但视觉传感器可能会受到遮挡以及无法测量接触力的限制，而接触力对于成功执行动作至关重要。如果没有触觉感知，在进行任务（如抓取）时就无法感知把握物体的力度，这也是限制任务成功率的一大原因。

在 VLA 概念兴起之前，已经有一些关于结合触觉的多模态相关研究^[95]，探索了结合视觉和触觉数据的方法。例如，研究^[95]结合视觉和触觉数据以估计物体的姿态，用于手内操作。研究^[96]针对抓取任务

研究了一种学习到的重抓取策略，该策略利用视觉和触觉反馈迭代地调整抓取动作，显著提升了抓取性能，并进一步扩展至可变形物体等更复杂场景的处理中，研究^[97]引入了一个基于 Transformer 的框架，使用触觉和视觉数据进行安全抓取，通过探索性动作收集触觉反馈并预测抓取结果，从而实现更安全的参数选择。

关于触觉模态的探索还远远不够，人体通过灵敏的皮肤来感知触觉，那么机器人或许应该开发出一套针对具身智能触觉感知的电子皮肤，从而增强 VLA 模型在触觉模态方面的感知能力。

4.5 数据集稀缺问题

高质量的数据稀缺制约了模型的泛化能力。与大规模图像或文本任务不同，具身智能中具有高质量对齐标签的多模态数据仍较为稀缺，尤其在涉及真实操作动作或新兴模态（如触觉、音频）时更为明显。当引入一种全新的模态时，模型需要大量的此类数据进行训练，这对于模型的表现和泛化能力至关重要。如果缺乏大量的优质数据集，新模态的感知与融合将难以取得理想的结果。

5 总结

本文从多模态感知角度系统回顾了具身智能 VLA 模型中的多模态表示、对齐与融合技术，总结出多模态感知是实现 VLA 稳健任务执行的重要部分。现有研究在引入新模态、泛化及灾难性遗忘等方面仍面临挑战，触觉、音频等模态的稀缺限制了感知范围。不同于以往仅关注视觉语言大模型的综述，本文首次整合具身场景下多

模态感知框架，为理解具身智能的VLA模型提供了新视角。其理论意义在于为VLA中多模态的感知技术研究提供了参考，应用上助力机器人在复杂环境下更稳健自主地执行任务。未来可进一步探索更通用的模态对齐与融合、抗遗忘的持续学习及真实多模态数据采集，推动多模态感知在更广泛的场景中落地。

参考文献：

[1] Ma Y E, Song Z X, Zhuang Y Z, et al. A survey on vision-language-action models for embodied AI[PP]. arXiv preprint, 2024, arXiv: 2405.14093.

[2] Sapkota R, Cao Y, Roumeliotis K I, et al. Vision-language-action (VLA) models: concepts, progress, applications and challenges[PP]. arXiv preprint, 2025, arXiv: 2505.04769.

[3] 赵博涛, 亢祖衡, 瞿晓阳, 等. 基于多模态大模型的具身智能体研究进展与展望[J]. 大数据, 2025, 11(3): 108-138.

Zhao B T, Kang Z H, Qu X Y, et al. Review and emerging trends of embodied agent based on multimodal large language models[J]. Big Data Research, 2025, 11(3): 108-138.

[4] 贾子琦, 王健宗, 张旭龙, 等. 基于大模型的具身智能任务规划研究: 从单智能体到多智能体[J]. 大数据, 2025, 11(2): 73-90.

Jia Z Q, Wang J Z, Zhang X L, et al. Large language model-based embodied intelligence task planning: from single-agent to multi-agent[J]. Big Data Research, 2025, 11(2): 73-90.

[5] Chen S B, Wu Z H, Zhang K, et al. Exploring embodied multimodal large models: development, datasets, and future directions[J]. Information Fusion, 2025,

122: 103198.

[6] Hassan S, Wang L X, Mahmud K R. Integrating vision and olfaction via multi-modal LLM for robotic odor source localization[J]. Sensors, 2024, 24(24): 7875.

[7] Brohan A, Brown N, Carbajal J, et al. RT-1: robotics Transformer for real-world control at scale[PP]. arXiv preprint, 2022, arXiv: 2212.06817.

[8] Brohan A, Brown N, Carbajal J, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control[PP]. arXiv preprint, 2023, arXiv: 2307.15818.

[9] Jiang Y F, Gupta A, Zhang Z C, et al. VIMA: general robot manipulation with multimodal prompts[PP]. V2. arXiv preprint, 2022, arXiv: 2210.03094.

[10] Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: An open-source vision-language-action model[PP]. arXiv preprint, 2024, arXiv: 2406.09246.

[11] Li X H, Liu M H, Zhang H B, et al. Vision-language foundation models as effective robot imitators[PP]. arXiv preprint, 2023, arXiv: 2311.01378.

[12] Zhao W, Ding P X, Zhang M, et al. VLAS: vision-language-action model with speech instructions for customized robot manipulation[PP]. arXiv preprint, 2025, arXiv: 2502.13508.

[13] Yu J W, Liu H R, Yu Q J, et al. ForceVLA: enhancing VLA models with a force-aware MoE for contact-rich manipulation[PP]. arXiv preprint, 2025, arXiv: 2505.22159.

[14] Cen J, Yu C H, Yuan H J, et al. World-VLA: towards autoregressive action world model[PP]. arXiv preprint, 2025, arXiv: 2506.21539.

[15] Radford A, Kim J W, Hallacy C, et al.

- Learning transferable visual models from natural language supervision[PP]. arXiv preprint, 2021, arXiv: 2103.00020.
- [16] Zhai X H, Mustafa B, Kolesnikov A, et al. Sigmoid loss for language image pre-training[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 11941–11952.
- [17] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision[PP]. arXiv preprint, 2023, arXiv: 2304.07193.
- [18] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models[PP]. arXiv preprint, 2023, arXiv: 2307.09288.
- [19] OpenAI, Achiam J, Adler S, et al. GPT-4 technical report[PP]. arXiv preprint, 2023, arXiv: 2303.08774.
- [20] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Kerrville: Association for Computational Linguistics 2019: 4171–4186.
- [21] Radford A, Kim J W, Xu T, et al. Robust speech recognition via large-scale weak supervision[C]//Proceedings of the 40th International Conference on Machine Learning. New York: ACM, 2023: 28492–28518.
- [22] Hsu W N, Bolte B, Tsai Y H, et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transac-
- tions on Audio, Speech, and Language Processing, 2021, 29: 3451–3460.
- [23] Baevski A, Zhou Y H, Mohamed A, et al. wav2vec 2.0: a framework for self-supervised learning of speech representations[J]. Advances in neural information processing systems, 2020, 33: 12449–12460.
- [24] Fang H S, Fang H J, Tang Z Y, et al. RH20T: a comprehensive robotic dataset for learning diverse skills in one-shot[C]//Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE Press, 2024: 653–660.
- [25] Wang Z Q, Zheng H, Nie Y S, et al. All robots in one: a new standard and unified dataset for versatile, general-purpose embodied agents[PP]. arXiv preprint, 2024, arXiv: 2408.10899.
- [26] Zhang C F, Hao P, Cao X G, et al. VTLA: vision-tactile-language-action model with preference learning for insertion manipulation[PP]. V1. arXiv preprint, 2025, arXiv: 2505.09577.
- [27] Feng R X, Hu J Y, Xia W K, et al. Any-Touch: learning unified static-dynamic representation across multiple visuotactile sensors[PP]. arXiv preprint, 2025, arXiv: 2502.12191.
- [28] Huang J L, Wang S, Lin F Q, et al. Tactile-VLA: unlocking vision-language-action model’s physical knowledge for tactile generalization[PP]. arXiv preprint, 2025, arXiv: 2507.09160.
- [29] Jang E, Irpan A, Khansari M, et al. BC-Z: zero-shot task generalization with robotic imitation learning[PP]. arXiv preprint, 2022, arXiv: 2202.02005.
- [30] Shridhar M, Manuelli L, Fox D. Perceiver-Actor: a multi-task trans-

- former for robotic manipulation[C]//Proceedings of the 6th Conference on Robot Learning (CoRL 2022). London: PMLR, 2023: 785–799.
- [31] Ahn M, Brohan A, Brown N, et al. Do as I can, not as I say: grounding language in robotic affordances[PP]. arXiv preprint, 2022, arXiv: 2204.01691.
- [32] Moraes R O, Sadmine Q A, Baier H, et al. InnateCoder: learning programmatic options with foundation models[PP]. arXiv preprint, 2025, arXiv: 2505.12508.
- [33] Zhao T Z, Kumar V, Levine S, et al. Learning fine-grained bimanual manipulation with low-cost hardware[PP]. arXiv preprint, 2023, arXiv: 2304.13705.
- [34] Chi C, Xu Z J, Feng S Y, et al. Diffusion policy: Visuomotor policy learning via action diffusion[J]. The International Journal of Robotics Research, 2025, 44 (10/11): 1684–1704.
- [35] Janner M, Li Q Y, Levine S. Offline reinforcement learning as one big sequence modeling problem[C]//Proceedings of the 35th International Conference on Neural Information Processing Systems. New York: ACM, 2021: 1273–1286.
- [36] Li H R, Jiang Z N, Chen Y H, et al. Generalizing consistency policy to visual RL with prioritized proximal experience regularization[PP]. arXiv preprint, 2024, arXiv: 2410.00051.
- [37] Lynch C, Khansari M, Xiao T, et al. Learning latent plans from play[C]//Proceedings of the 3rd Conference on Robot Learning (CoRL 2019). London: PMLR, 2020: 1113–1132.
- [38] Hafner D, Lillicrap T, Fischer I, et al. Learning latent dynamics for planning from pixels[PP]. arXiv preprint, 2018, arXiv: 1811.04551.
- [39] Hansen N, Su H, Wang X L. TD-MPC2: scalable, robust world models for continuous control[PP]. arXiv preprint, 2023, arXiv: 2310.16828.
- [40] Shridhar M, Manuelli L, Fox D. CLIPORT: what and where pathways for robotic manipulation[C]//Proceedings of the 5th Conference on Robot Learning (CoRL 2021). London: PMLR, 2021: 894–906.
- [41] Huang X Y, Batra D, Rai A, et al. Skill Transformer: a monolithic policy for mobile manipulation[C]//Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2023: 10818–10828.
- [42] Sontakke S, Zhang J, Arnold S, et al. Roboclip: one demonstration is enough to learn robot policies[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023: 55681–55693.
- [43] Kienle C, Alt B, Celik O, et al. MuTT: a multimodal trajectory transformer for robot skills[C]//Proceedings of the 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Piscataway: IEEE Press, 2024: 9644–9651.
- [44] Zhang T L, Guan J Y, Zhao L, et al. Preferred-action-optimized diffusion policies for offline reinforcement learning[PP]. arXiv preprint, 2024, arXiv: 2405.18729.
- [45] Chen L L, Lu K, Rajeswaran A, et al. Decision transformer: reinforcement learning via sequence modeling[C]//Proceedings of the 35th International Conference on Neural Information Process -

- ing Systems. New York: ACM, 2021: 15084–15097.
- [46] Hafner D, Lillicrap T, Norouzi M, et al. Mastering Atari with discrete world models[PP]. arXiv preprint, 2020, arXiv: 2010.02193.
- [47] Bruce J, Dennis M, Edwards A, et al. Genie: Generative interactive environments[C]//Proceedings of the 41st International Conference on Machine Learning. Vienna: PMLR, 2024: 1–21.
- [48] Yan W, Zhang Y Z, Abbeel P, et al. VideoGPT: video generation using VQ–VAE and transformers[PP]. arXiv preprint, 2021, arXiv: 2104.10157.
- [49] Fan Y G, Ding P X, Bai S H, et al. Long–VLA: Unleashing long–horizon capability of vision language action model for robot manipulation[PP]. arXiv preprint, 2025, arXiv: 2508.19958.
- [50] Shah D, Sridhar A, Dashora N, et al. ViNT: a foundation model for visual navigation[PP]. arXiv preprint, 2023, arXiv: 2306.14846.
- [51] Bjorck J, Castañeda F, Cherniadev N, et al. GR00T N1: an open foundation model for generalist humanoid robots [PP]. arXiv preprint, 2025, arXiv: 2503.14734.
- [52] Shafiullah N M M, Paxton C, Pinto L, et al. CLIP–fields: weakly supervised semantic fields for robotic memory[PP]. arXiv preprint, 2022, arXiv: 2210.05663.
- [53] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[C]//Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). New York: Curran Associates, Inc., 2017.
- [54] Lin H Z, Cheng X, Wu X Y, et al. CAT: cross attention in vision transformer[C]//Proceedings of the 2022 IEEE International Conference on Multimedia and Expo (ICME). Piscataway: IEEE Press, 2022: 1–6.
- [55] Perez E, Strub F, De Vries H, et al. FiLM: visual reasoning with a general conditioning layer[C]//Proceedings of the AAAI Conference on Artificial Intelligence. California: AAAI Press, 2018.
- [56] Zhong Y F, Huang X C, Li R C, et al. DexGraspVLA: a vision–language–action framework towards general dexterous grasping[PP]. arXiv preprint, 2025, arXiv: 2502.20900.
- [57] AgiBot–World–Contributors, Bu Q W, Cai J S, Chen L, et al. AgiBot World Colosseo: a large–scale manipulation platform for scalable and intelligent embodied systems[PP]. arXiv preprint, 2025, arXiv: 2503.06669.
- [58] He K M, Chen X L, Xie S N, et al. Masked autoencoders are scalable vision learners[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15979–15988.
- [59] Zhou J H, Wei C, Wang H Y, et al. iBOT: image BERT pre–training with online tokenizer[PP]. arXiv preprint, 2021, arXiv: 2111.07832.
- [60] Chen T, Kornblith S, Norouzi M, et al. A simple framework for contrastive learning of visual representations[C]//Proceedings of the 37th International Conference on Machine Learning. Cambridge: PMLR, 2020: 1597–1607.
- [61] Fang Y X, Sun Q, Wang X G, et al. EVA–02: a visual representation for neon genesis[J]. Image and Vision Computing, 2024, 149: 105171.
- [62] Yuan L, Chen D D, Chen Y L, et al. Florence: a new foundation model for

- computer vision[PP]. arXiv preprint, 2021, arXiv: 2111.11432.
- [63] Alayrac J B, Barr I, Barreira R, et al. Flamingo: a visual language model for few-shot learning[C]//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). New York: Curran Associates, Inc., 2022: 23716–23736.
- [64] Yu J H, Wang Z R, Vasudevan V, et al. CoCa: contrastive captioners are image-text foundation models[PP]. arXiv preprint, 2022, arXiv: 2205.01917.
- [65] Fan L X, Wang G Z, Jiang Y F, et al. MineDojo: building open-ended embodied agents with Internet-scale knowledge[C]//Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). New York: Curran Associates, Inc., 2022: 18343–18362.
- [66] Open X-Embodiment Collaboration. Open X-Embodiment: robotic learning datasets and RT-X models[C]//Proceedings of 2024 IEEE International Conference on Robotics and Automation (ICRA). Piscataway: IEEE, 2024: 6892–6903.
- [67] Nair S, Rajeswaran A, Kumar V, et al. R3M: a universal visual representation for robot manipulation[J]. arXiv preprint, 2022, arXiv: 2203.12601.
- [68] Liu H T, Li C Y, Wu Q Y, et al. Visual instruction tuning[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023: 34892–34916.
- [69] Reed S, Zolna K, Parisotto E, et al. A generalist agent[PP]. arXiv preprint, 2022, arXiv: 2205.06175.
- [70] de Vries H, Strub F, Mary J, et al. Modulating early visual processing by language[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6597–6607.
- [71] Ha D, Dai A, Le Q V. HyperNetworks[PP]. arXiv preprint, 2016, arXiv: 1609.09106.
- [72] Pfeiffer J, Kamath A, Rücklé A, et al. AdapterFusion: non-destructive task composition for transfer learning[C]//Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Stroudsburg: ACL, 2021: 487–503.
- [73] Wang Y R, Ung C, Tannert G, et al. RoboEval: where robotic manipulation meets structured and scalable evaluation[PP]. arXiv preprint, 2025, arXiv: 2507.00435.
- [74] Jia C, Yang Y F, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C]//Proceedings of the 38th International Conference on Machine Learning (ICML 2021). [S.l.]: PMLR, 2021: 4904–4916.
- [75] Li J N, Li D X, Xiong C M, et al. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation[C]//Proceedings of the 39th International Conference on Machine Learning (ICML 2022). Baltimore: PMLR, 2022: 12888–12900.
- [76] Li J N, Li D X, Savarese S, et al. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models[C]//Proceedings of the 40th International Conference on Machine Learning (ICML 2023). Hawaii: PMLR, 2023, 202: 19730–19742.

- [77] Chen X, Wang X, Changpinyo S, et al. PaLI: a jointly-scaled multilingual language-image model[PP]. arXiv preprint, 2022, arXiv: 2209.06794.
- [78] Huang S H, Dong L, Wang W H, et al. Language is not all you need: aligning perception with language models[C]//Advances in Neural Information Processing Systems 36 (NeurIPS 2023). New York: Curran Associates, Inc., 2023: 72096–72109.
- [79] Yu L J, Cheng Y, Sohn K, et al. MAGVIT: masked generative video transformer[C]//Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2023: 10459–10469.
- [80] Khazatsky A, Pertsch K, Nair S, et al. DROID: a large-scale in-the-wild robot manipulation dataset[C]//Proceedings of Robotics: Science and Systems (RSS 2024). Delft: RSS Foundation, 2024.
- [81] Grauman K, Westbury A, Byrne E, et al. Ego4d: Around the world in 3,000 hours of egocentric video[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2022). Piscataway: IEEE, 2022: 18995–19012.
- [82] Mandlekar A, Xu D F, Wong J, et al. What matters in learning from offline human demonstrations for robot manipulation[PP]. arXiv preprint, 2021, arXiv: 2108.03298.
- [83] Mees O, Hermann L, Rosete-Beas E, et al. CALVIN: a benchmark for language-conditioned policy learning for long-horizon robot manipulation tasks[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 7327–7334.
- [84] Feng Y H, Gao C K, Liu B, et al. LIBERO: benchmarking knowledge transfer for lifelong robot learning[C]//Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS 2023). New York: Curran Associates, Inc., 2023: 44776–44791.
- [85] Cheng N, Xu J N, Guan C H, et al. Touch100k: a large-scale touch-language-vision dataset for touch-centric multimodal representation[J]. Information Fusion, 2025, 124: 103305.
- [86] Walke H R, Black K, Zhao T Z, et al. Bridgedata v2: A dataset for robot learning at scale[C]//Proceedings of the 7th Conference on Robot Learning (CoRL 2023). Atlanta: PMLR, 2023: 1723–1736.
- [87] Deng S L, Yan M, Wei S L, et al. GraspVLA: a grasping foundation model pre-trained on billion-scale synthetic action data[PP]. arXiv preprint, 2025, arXiv: 2505.03233.
- [88] Mahler J, Liang J, Niyaz S, et al. Dex-Net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics[C]//Proceedings of Robotics: Science and Systems (RSS 2017). Cambridge: RSS Foundation, 2017.
- [89] Stepputtis S, Campbell J, Phielipp M, et al. Language-conditioned imitation learning for robot manipulation tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 13139–13150.
- [90] Cheng Z L, Tu Y G, Li R, et al. EmbodiedEval: evaluate multimodal LLMs as embodied agents[PP]. arXiv preprint, 2025, arXiv: 2501.11858.
- [91] Dey S, Zaech J N, Nikolov N, et al.

ReVLA: reverting visual domain limitation of robotic foundation models[PP]. arXiv preprint, 2024, arXiv: 2409.15250.

[92] Anderson P, Shrivastava A, Truong J, et al. Sim-to-real transfer for vision-and-language navigation[C]//Proceedings of the 4th Conference on Robot Learning (CoRL 2020). [S. l.]: PMLR, 2020: 671–681.

[93] Wang R H, Geng H R, Li T L, et al. The sound of simulation: learning multi-modal sim-to-real robot policies with generative audio[PP]. arXiv preprint, 2025, arXiv: 2507.02864.

[94] Fang Y, Yang Y, Zhu X H, et al. ReBot: scaling robot learning with real-to-sim-to-real robotic video synthesis[PP]. arXiv preprint, 2025, arXiv: 2503.14526.

[95] Li H Y, Dikhale S, Iba S, et al. Vi-HOPE: visuotactile in-hand object 6D pose estimation with shape completion [J]. IEEE Robotics and Automation Letters, 2023, 8(11): 6963–6970.

[96] Calandra R, Owens A, Jayaraman D, et al. More than a feeling: learning to grasp and regrasp using vision and touch [J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3300–3307.

[97] Han Y H, Yu K L, Batra R, et al. Learning generalizable vision-tactile robotic grasping strategy for deformable objects via transformer[J]. IEEE/ASME Transactions on Mechatronics, 2025, 30(1): 554–566.

作者简介



周江盟（2000–），男，湖南大学计算机学院博士生，主要研究方向为多模态大模型、具身智能。



孙奥兰（1996–），女，平安科技（深圳）有限公司高级算法工程师，主要研究方向为具身智能、信号处理、语音生成、大模型等。



瞿晓阳（1988–），男，博士，平安科技前沿机器学习算法分组负责人，清华大学深圳国际研究生院校外导师，中国科学技术大学先进技术研究院校外导师，中佛罗里达大学访问学者，从事机器学习、大数据、体系结构等方面的研究，在语音语义分析、自动化机器学习、零样本和小样本学习、高性能计算与存储等方面经验丰富。近几年，在体系结构方向（如INFOCOM、DAC、TPDS、IPDPS）和人工智能方向（如NeurIPS、IJCAI、ICASSP、Interspeech）等国际顶级会议和顶级期刊发表近50篇文章，其中1篇论文荣获会议最佳学生论文奖提名，担任过多个国际顶级期刊的评委，已授权的专利70项，已出版的专著2本。



张旭龙（1988–），男，博士，平安科技（深圳）有限公司高级算法研究员，复旦大学计算机理学博士，2023年入选上海市东方英才计划青年项目。兼任清华大学深圳研究院及中国科学技术大学先进技术研究院校外导师，中国自动化学会联邦数据与联邦智能专业委员会委员。目前是IEEE、中国自动化学会以及中国计算机学会会员。主要研究方向为大模型、具身智能、跨模态智能计算以及模型上下文协议（MCP）等。发表学术论文80余篇，获得国家发明专利授权20余项。



王健宗（1983–），男，博士，平安科技（深圳）有限公司正高级工程师，智能金融前沿技术研究院院长。美国佛罗里达大学人工智能博士后，美国莱斯大学和华中科技大学联合培养博士，中国计算机学会杰出会员，中国计算机学会大数据专家委员会委员，中国自动化学会联邦数据和联邦智能专业委员会副主任。主要研究方向为具身智能、大模型、联邦学习和深度学习等。荣获中国专利优秀奖、广东省科技进步奖二等奖、吴文俊人工智能科技进步奖一等奖、人民银行金融科技发展奖一等奖、深圳高层次专业人才领军人才、福田英才等。

收稿日期: 2025-09-30
通信作者: 王健宗, jzwang@188.com
基金项目: 深港联合资助项目(A类)(No.SGDX20240115103359001)
Foundation Items: Shenzhen-Hong Kong Joint Funding Program (Category A) (No.SGDX20240115103359001)