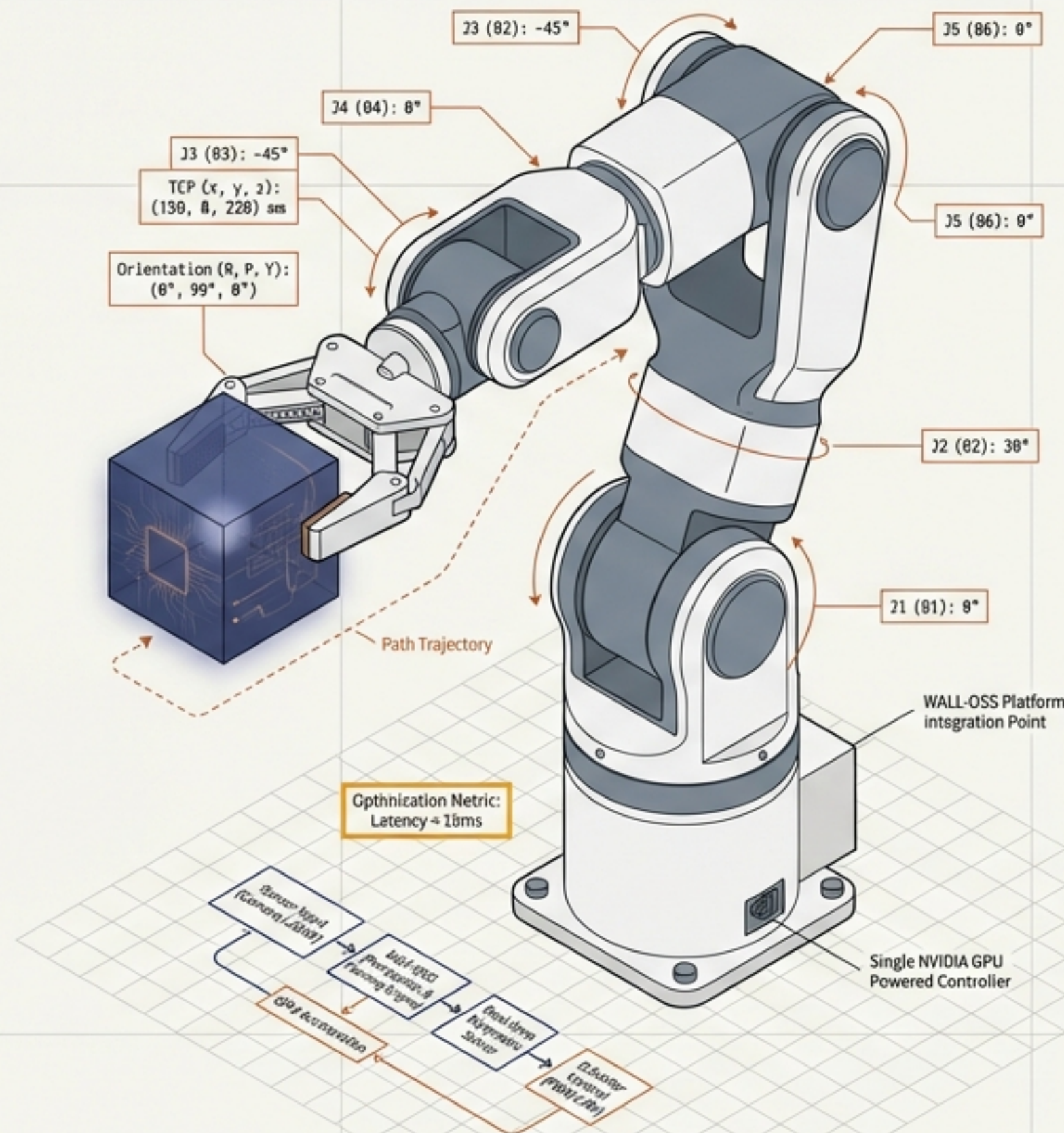


具亮计划 2026 答辩汇报

打通具身智能落地的最后一公里 基于单卡的 WALL-OSS 机械臂全栈部署闭环

张振尧

零零后（北京）科技有限公司创始人 · 具身智能工程师



团队画像与全栈工程底座

从底层物理运动学到云端多模态大模型的完整闭环能力

Card 1



学术背景

北京化工大学
(机械制造与自动化)

Card 2



黑客松战绩

2025 NVIDIA Sky Hackathon
年度总决赛冠军

Card 3



实战经验

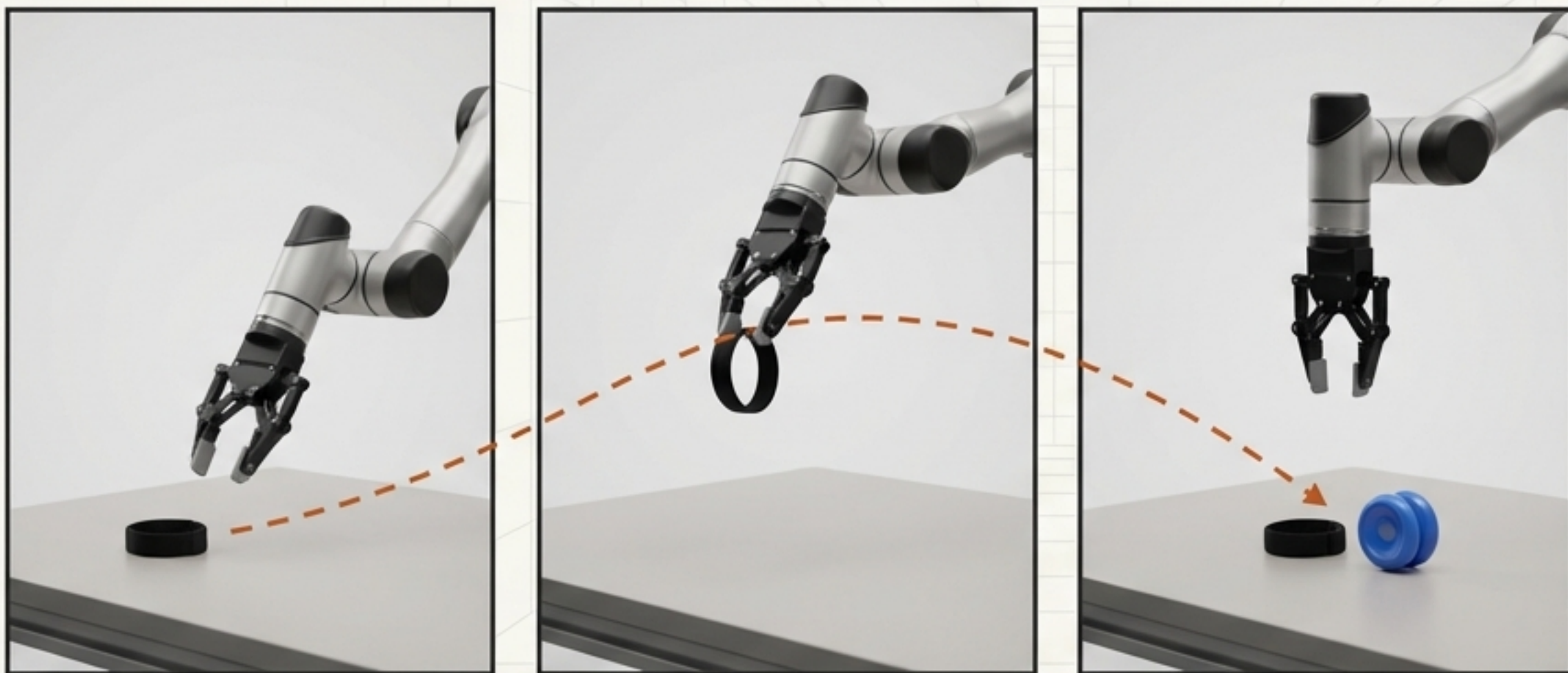
国际人形机器人技能大赛核心参与者



Task Prompt: Grab the black wristband to the blue yo-yo ball

任务实况：高频连续动作的真机闭环展示

指令执行结果验证 (The Proof of Concept)



模型表现：10步 ODE 连续控制求解，动作极为平滑连贯，零顿挫。

平台配置

LeRobot SO-101 单臂

控制模式

6 自由度主从遥操作采集
(Leader-Follower)

视觉协同

全局正面视角 (Face View)
+ 手眼双视角 (Left Wrist
View) 融合推理

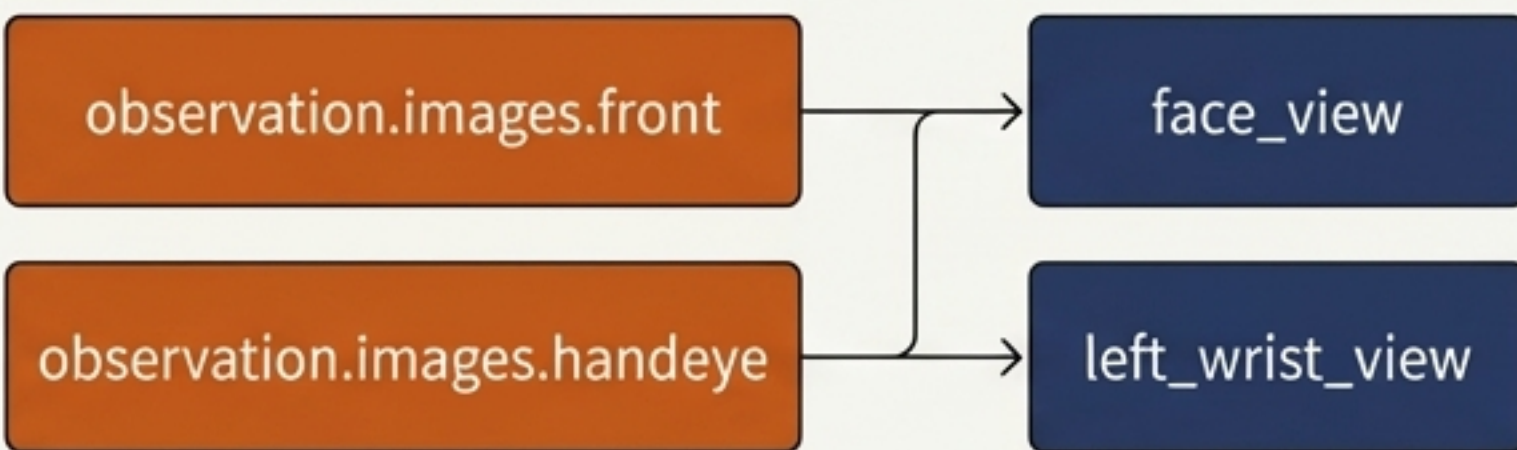
高质量数据管线与规范化对齐

将异构物理世界数据无缝桥接至 VLA 大模型标准体系

自研高质量数据集

📁 HuggingFace ID	shay/lerobot_demo1
📊 闭环轨迹数量	30 条 (95% 训练 / 5% 测试)
📊 总物理帧数	8,199 帧
📊 动作采样率	30 FPS
📊 任务特征	极度聚焦单一抓取任务，摒弃冗余噪声

异构数据键值映射 (Key Mapping)



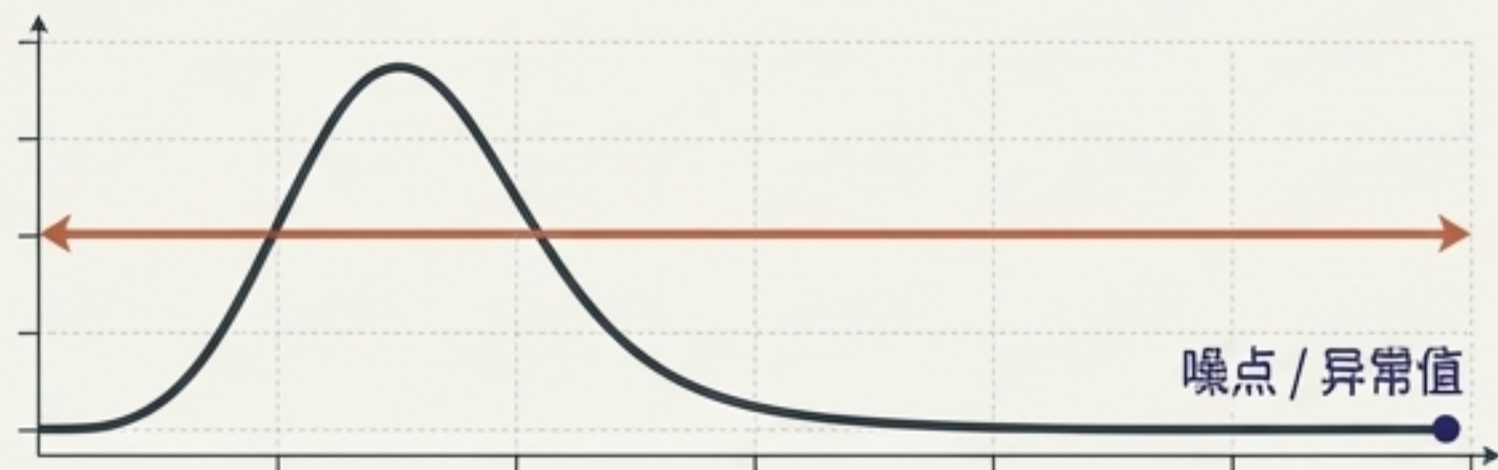
```
KEY_MAPPINGS = {  
    'observation.images.front': 'face_view',  
    'observation.images.handeye': 'left_wrist_view'  
}
```

优雅的代码级重构，彻底解决不同机器人硬件平台与模型输入层的数据隔离问题。

动作数据归一化策略：拒绝机械臂抖动的数学基石

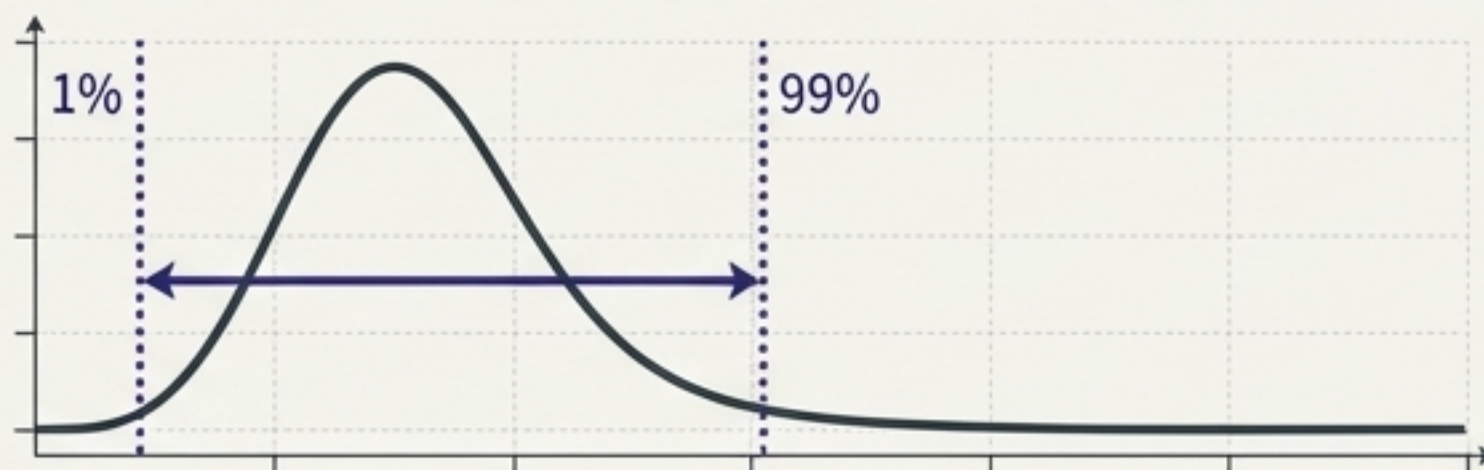
彻底规避极端异常值对物理输出的致命干扰

传统 Min-Max 归一化（缺陷方案）



极易被极端异常值拉扯，导致核心高频动作特征被严重压缩、失真，引发真机抖动。

q1-q99 分位数归一化（我们的方案）



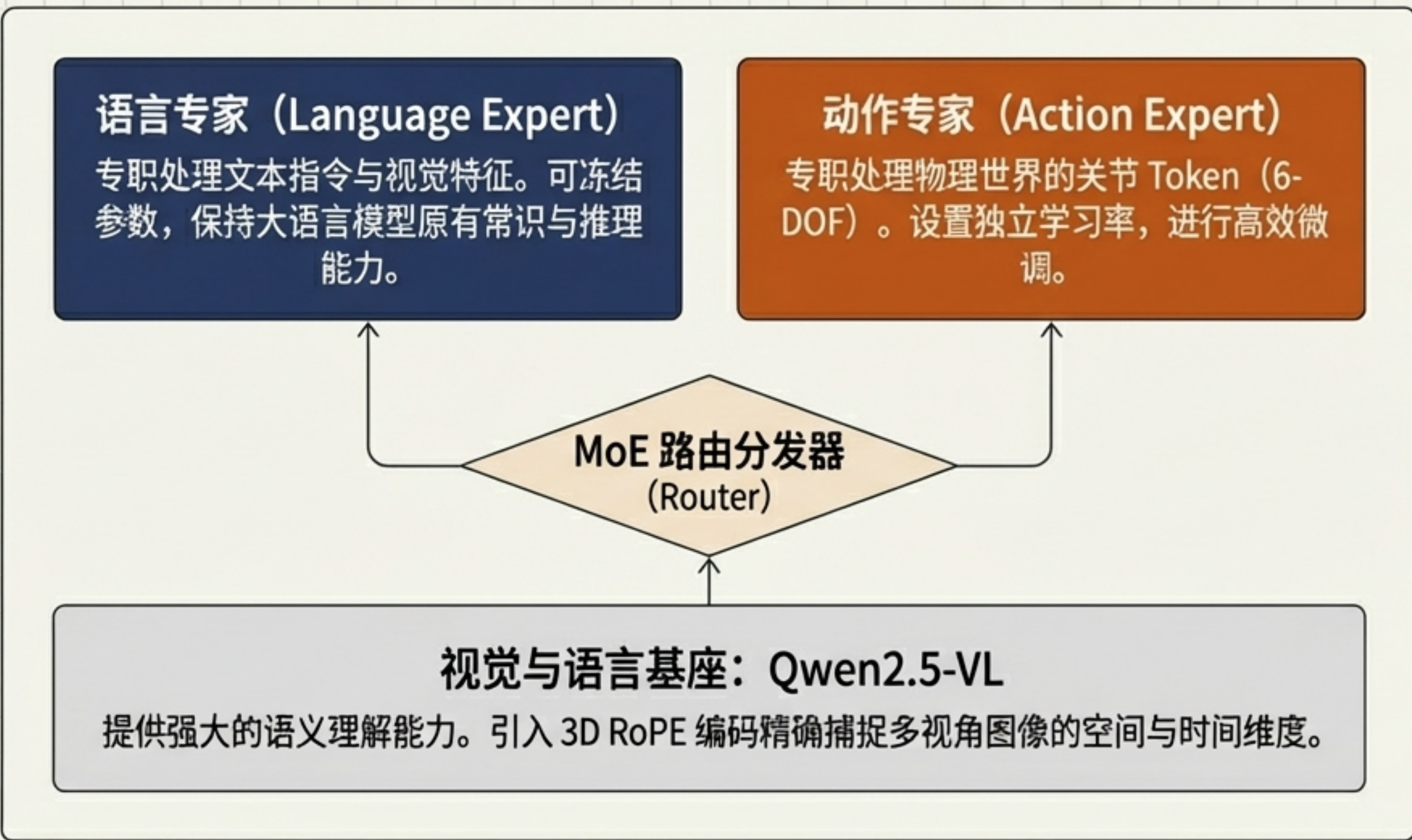
精准切除长尾极值，严格锁定有效数据区间，保证动作输出的绝对稳健与平滑。

硬核工程突破：在线流式分位数估计

- **痛点：**传统计算分位数需将所有数据载入内存进行全局排序，对显存极度不友好。
- **创新：**引入 OpenPI 的 RunningStats 架构，使用 5000 个动态 bin 自定义实现在线流式直方图估计。
- **结果：**兼顾极低内存占用与极高估算精度，完美适配大数据集的数据加载管线。

WALL-X 核心架构解构：紧耦合 MoE 设计

破解多模态微调中的灾难性遗忘（Catastrophic Forgetting）痛点



为什么引入混合专家架构？

在传统单一大模型中，强行混入物理动作数据会覆盖原本的自然语言权重，导致模型“变笨”。

MoE 架构让语言与动作各司其职。通过特定的 Token Type 进行路由，我们在物理动作学习中完美保留了基座模型强大的跨模态认知。

动作预测对决：生成范式的关键技术抉择

为什么我们在高频连续抓取任务中选择 Flow Matching？

	Flow Matching（连续流匹配）	FAST Tokenizer（离散自回归）
动作建模方式	连续坐标回归，优化目标为 MSE	动作离散化为 Token，优化目标为 Cross-Entropy
推理生成机制	10步 ODE（Euler）迭代求解，预测流场	逐 Token 串行自回归生成
动作域注意力	动作 Token 间采用双向注意力，统筹全局轨迹	采用因果注意力，仅能依赖历史状态
真实物理表现	收敛极快。在高频连续动作下轨迹极其平滑连贯。	语义对齐能力极强，但在高频精细控制中容易出现顿挫感。

工程决策结论：在单一高频的精密抓取场景下，Flow Matching 提供了不可替代的物理级平滑度，是我们实现真机闭环的首选。

算力降维打击：跨越 8 卡集群与单卡桌面的鸿沟

如何将千万级实验室配置硬核下放至极客桌面环境

官方推荐的富裕方案（Goliath）

计算硬件	8x GPU 服务器（如 8x A100 / 4090）
并行策略	FSDP2（全分片数据并行）
Batch Size / 单卡	8
梯度累积步数	32 步
全局有效 Batch Size	2048（8x8x32）

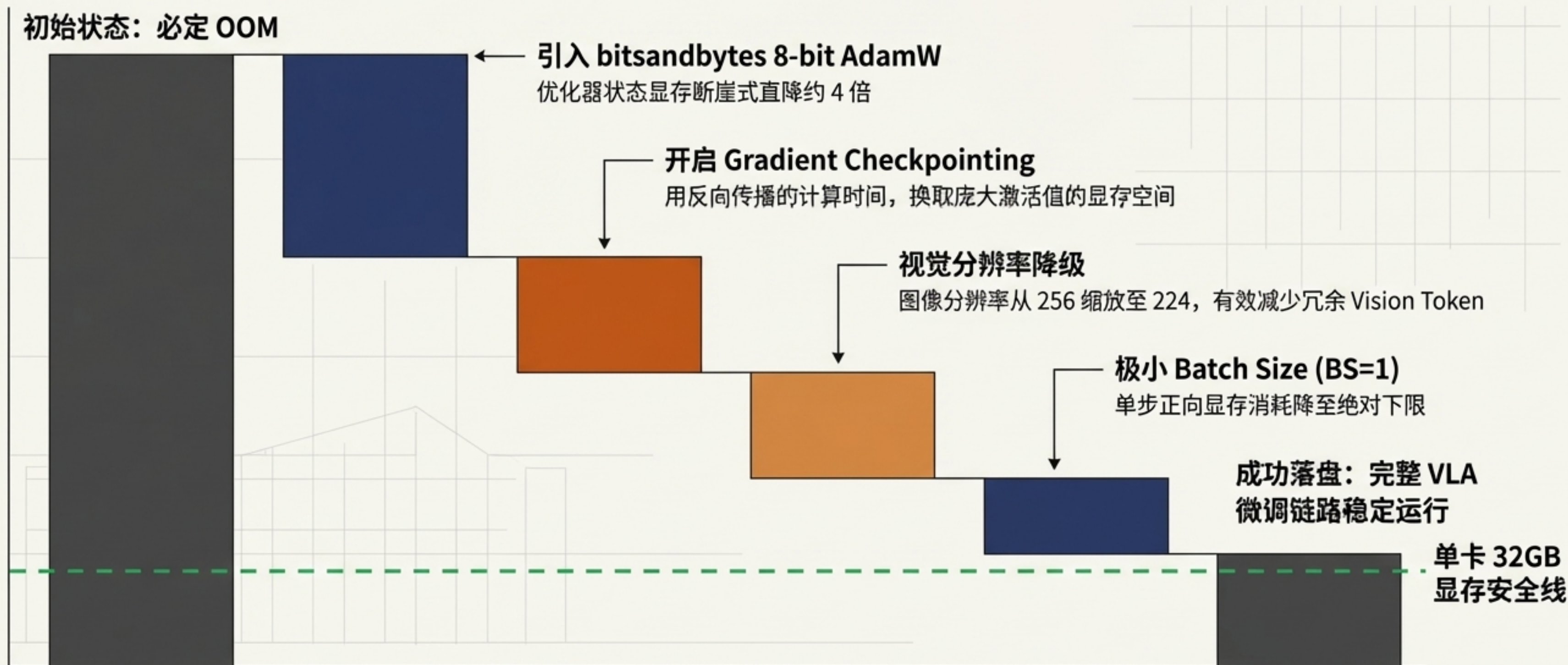
我们的极限压榨方案（David）

计算硬件	1x 消费级单卡（RTX 5090 32GB）
并行策略	无（纯单卡极限微调）
Batch Size / 单卡	极致压缩至 1，避免爆显存
梯度累积步数	超长累积 256 步
全局有效 Batch Size	256（精准对齐梯度更新节奏）

核心破局点：在无法使用分布式数据并行的前提下，利用超长步数的梯度累积，强行拉平了模型收敛所需的有效 Batch，成功在消费级显卡上跑通 80 层大模型的微调全链路！

显存瀑布图：32GB 空间是从哪里抠出来的？

一步步消除致命的 Out-Of-Memory (OOM) 边界



Server-Client 极低延迟真机推理架构

毫秒级响应：将沉重的大模型剥离至云端，保持物理端的极速轻盈



交付的不是 Demo，而是可扩展的工程闭环

大幅降低前沿 VLA 模型的工业落地门槛

1. 物理世界感知与规范

自研高质数据采集管线，创新流式分位数归一化，解决异构硬件接入标准。

2. 多模态大模型演进

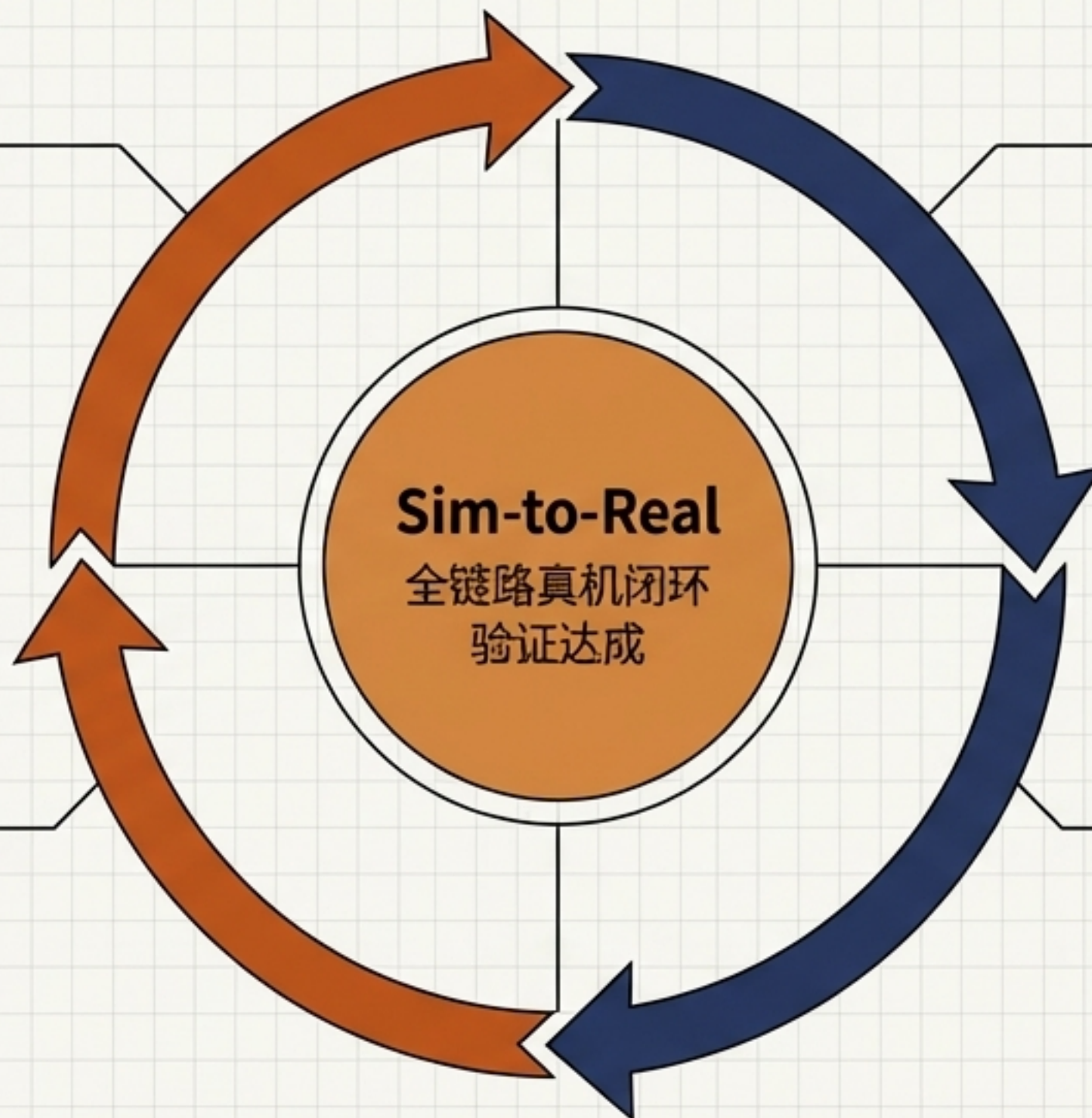
引入 Qwen2.5-VL 基座，基于 MoE 防遗忘机制与 Flow Matching 算法构建高频连续动作空间。

3. 极限算力优化微调

通过 8-bit 量化、梯度检查点与极限有效 Batch 对齐，成功在单卡 32GB 显存下打通全栈微调。

4. 云边协同极速部署

构建 Server-Client gRPC 毫秒级通信架构，实现从云端大算力到边缘端物理响应的无缝衔接。



客观复盘与未来技术演进路线图

直面痛点，以极限工程探索物理世界的边界

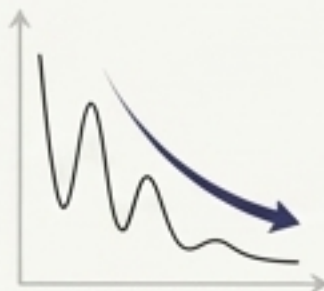
训练稳定性优化

现状痛点

单卡环境下的 Loss 曲线存在明显振荡，收敛平滑度不如多卡集群。

原因剖析

受制于显存瓶颈，单卡全局有效 Batch Size 远小于官方基准，导致梯度噪声偏大。



演进计划

下一步将引入非线性的动态学习率衰减策略与动态梯度裁剪算法，强行压制训练晚期的振荡。

跨越 Sim-to-Real 鸿沟

现状痛点

当前模型对抓取对象的位置、光照过拟合，缺乏泛化能力。

原因剖析

训练集仅包含单一抓取任务，未激活基座大模型庞大的跨领域世界知识库。



演进计划

深度利用 VLM 的强语义先验，开展多任务数据混合训练，实现真正意义上的“一句自然语言，无缝切换物理任务”。

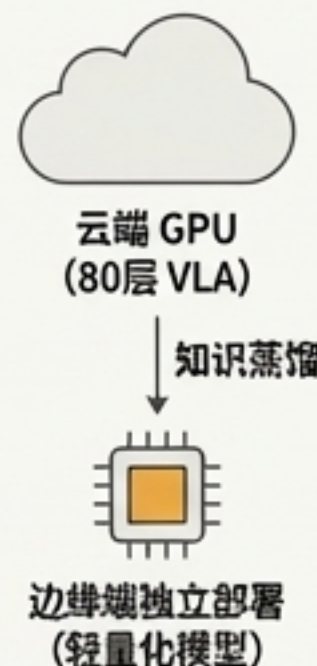
极致边缘部署愿景

现状痛点

推理依然高度依赖厚重且昂贵的云端 GPU 服务算力。

原因剖析

80层 Transformer 架构的参数规模过于庞大，无法直接塞入机器人本体。



演进计划

全面探索大模型蒸馏 (Knowledge Distillation) 技术，将大参数 VLA 知识浓缩化身，彻底打通边缘端 (Edge) 的独立部署能力。