光电智能识别与感知

DOI: 10.16818/j.issn1001－5868.20260514001

面向空间操作任务的视觉表征对齐与时延补偿轻量化方法

李 达[1],秦传波[2],吴曾明[1],李锰昭[1],郭佳璐[1],曹明轩[1*]

(五邑大学1. 机械与自动化工程学院; 2. 电子与信息工程学院,广东 江门 529020)

摘 要: 针对视觉-语言-动作(Vision-Language-Action,VLA)模型在空间操作任务中局部几何特征表征不足,以及真实机器人感知-推理-执行链路中预测区间与执行区间时空错位,文章提出一种以中间层视觉表征对齐为主、以时延补偿为辅的轻量化方法。首先,采用参数冻结的DINOv2模型作为视觉教师网络,在VLA视觉-语言骨干网络的中间层引入对齐投影层,通过余弦距离损失约束中间层视觉特征与教师特征的一致性,增强模型对局部空间结构的编码能力;其次,设计时延补偿机制,以提升系统在真实执行链路中的鲁棒性。仿真实验表明,π_0基础模型引入视觉表征对齐后,任务成功率由72.60%提升至75.07%,进一步结合时延补偿训练后提升至80.87%;在$\pi_{0.5}$基础模型上,引入视觉表征对齐后,任务成功率由94.27%提升至96.53%。真实机械臂实验表明,相比$\pi_{0.5}$基础模型,所提部署策略在三类空间操作任务上的成功率分别由70.0%,55.0%和50.0%提升至95.0%,90.0%和85.0%,同时将成功回合的平均完成时间由17.97 s缩短至16.24 s。结果表明,中间层视觉表征对齐能够在不同基础模型上稳定增强空间感知能力,时延补偿更适合作为面向真实执行链路的辅助补偿手段,所提方法兼顾了空间感知增强与真实部署适应性。

关键词: 多模态大模型; 视觉表征; DINOv2; 轻量化; 时延补偿

中图分类号: TP391.4 **文章编号:** 1001－5868(2026)05－0804－10

Lightweight Visual Representation Alignment and Delay Compensation for Spatial Manipulation Tasks

LI Da[1], QIN Chuanbo[2], WU Zengming[1], LI Mengzhao[1], GUO Jialu[1], CAO Mingxuan[1]

(1. School of Mechanical and Automation Engineering; 2. School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, CHN)

Abstract: To address the insufficient representation of local geometric features in Vision-Language-Action (VLA) models for spatial manipulation tasks, as well as the spatiotemporal mismatch between prediction and execution intervals in real robotic perception-inference-execution pipelines, this study proposes a lightweight method based primarily on intermediate visual representation alignment, with delay compensation as an auxiliary component. First, a frozen DINOv2 model is adopted as the visual teacher network, and an alignment projection layer is introduced into an intermediate layer of the VLA vision-language backbone. A cosine distance loss is used to constrain the consistency between intermediate visual and teacher features, thereby enhancing the model's ability to encode local spatial structures. Second, a delay compensation mechanism is

收稿日期:2026－05－14.

基金项目:2024年广东省普通高校科研重点领域专项(2024ZDZX1008).

* 通信作者:曹明轩

designed to improve the robustness of the system in real execution pipelines. Experimental results in the LIBERO-Spatial simulation environment show that, after visual representation alignment is introduced, the task success rate of the π_0-based model increases from 72.60% to 75.07%, further increasing to 80.87% when combined with delay compensation training. For the $\pi_{0.5}$-based model, visual representation alignment improves the task success rate from 94.27% to 96.53%. Real-robot experiments show that, compared with the $\pi_{0.5}$ baseline model, the proposed deployment strategy improves the success rates of three spatial manipulation tasks from 70.0%, 55.0%, and 50.0% to 95.0%, 90.0%, and 85.0%, respectively, while reducing the average completion time of successful trials from 17.97 s to 16.24 s. The results demonstrate that intermediate visual representation alignment can consistently enhance spatial perception across different base models, while delay compensation is more suitable as an auxiliary compensation strategy for real execution pipelines. The proposed method achieves a balance between spatial perception enhancement and real-world deployment adaptability.

Key words: multi-modal large models; visual representation; DINOv2; lightweight; delay compensation

0　引言

近年来，大规模视觉-语言-动作（Vision-Language-Action，VLA）模型在跨模态语义理解与通用操作任务中取得了显著进展[1-3]。VLA模型通常以视觉图像、语言指令和机器人本体状态作为输入，直接生成机器人连续动作或动作块，从而实现从感知理解到行为控制的端到端映射。其中，π_0[4]及其改进版本$\pi_{0.5}$[5]是近年来用于通用机器人操作任务的代表性VLA基础模型：前者通过视觉-语言骨干网络与动作专家模块建模连续动作生成过程，后者在此基础上进一步优化了数据利用方式与模型泛化能力。然而，这类模型大多以预训练视觉-语言模型[6]为骨干，其预训练过程更侧重于跨模态语义对齐与高层语义建模，使得视觉表征偏向全局语义信息，对边界细节和局部几何线索关注不足，最终导致模型缺乏足够的几何感知能力[7-8]。此外，真实机器人系统中的感知-推理-执行链路不可避免地存在时延，易造成模型预测区间与实际执行区间的错位，降低控制鲁棒性。如何在有限计算资源约束下增强VLA模型面向空间操作任务的视觉表征能力，并在真实部署中兼顾时延适应性，已成为相关研究中的关键问题。

在视觉表征增强方面，现有研究已开始关注多模态模型在空间理解中的几何表征不足的问题。Huang等[9]通过引入三维表征监督，提高了多模态模型对场景空间结构的理解能力，表明几何先验能够弥补预训练表征在空间建模方面的不足。Li等[10]进一步面向VLA模型，对骨干网络中间层视觉特征施加空间约束，以增强模型对局部结构和空间关系的编码能力。相关视觉感知研究也表明，注意力机制、多尺度特征融合和多层次特征建模能够提升模型在复杂场景下的特征表达能力[11-13]。然而，现有多模态空间表征增强方法通常依赖三维监督、多视图信息或大型几何教师模型，训练开销较大，部署灵活性受限。

在时延补偿方面，针对真实机器人系统中预测区间与执行区间错位的问题，Black等[14]提出实时动作分块方法，通过固定已生成的动作序列、修正剩余动作以缓解该问题，但该方法会引入额外的运行时计算开销。Tang等[15]则围绕异步推理场景提出未来状态感知框架，构建状态-动作偏移样本，有效缓解时延引发的性能退化，但其补偿形式与实际系统时延尺度之间的对应关系，仍需结合具体部署场景进一步验证。

基于上述分析，本文面向空间操作任务，提出一种以中间层视觉表征对齐为主、以小步长时延补偿为辅的轻量化改进框架，旨在有限计算资源约束下提升模型的空间感知能力，并提高其对真实链路时延的适应能力。

本文主要研究工作如下：首先，提出一种基于DINOv2[16]的轻量化中间层视觉表征对齐方案，与采用三维视觉教师监督的方法相比，该方案在保持任务成功率相近的同时，将训练显存占用降低14.5%，训练时长缩短46.1%；其次，设计与实际系统时延尺度相匹配的小步长时延补偿策略，提高模型对预测-执行错位的适应能力；最后，在π_0和$\pi_{0.5}$两

个代表性VLA基础模型上，系统验证了所提方法的有效性，并通过真实机械臂实验证明了其工程价值。

1 方法

1.1 问题定义与基础策略

1.1.1 问题定义

本文考虑视觉-语言条件下的机器人连续控制问题，在给定多视角观测图像、语言指令与机器人本体状态的条件下，生成长度为H的未来连续动作块，以完成特定空间交互操作。针对空间操作任务中视觉表征对目标边界、局部结构及空间关系编码不足的问题，本文重点研究VLA模型的视觉表征增强方法，并进一步考虑异步执行条件下时延因素带来的影响。

1.1.2 基础VLA策略

本文以π_0的建模形式定义基础策略。由于所采用的视觉表征对齐与时延补偿机制均作用于基础策略之上，不改变动作生成的训练与推理流程，所以以下以π_0的流匹配框架为例进行说明；对于性能更强的$\pi_{0.5}$基础模型，本文在实验部分进一步验证所提方法的适用性。设时刻t的多模态观测为：

$$O_t=\{I_t^1,\cdots,I_t^n,l_t,s_t\} \tag{1}$$

式中，I_t^i为时刻t第i个视角的图像观测，l_t表示任务语言指令，s_t表示机器人状态。策略网络旨在学习条件分布$P_\theta(A_t|O_t)$，并生成未来动作块：

$$A_t=[a_t,a_{t+1},\cdots,a_{t+H-1}],A_t\in\mathbb{R}^{H\times d_a} \tag{2}$$

在模型架构上，图像I_t^i与语言指令l_t首先输入到基于PaliGemma[17]的视觉-语言骨干网络，以提取多模态编码特征；随后，机器人状态s_t、时间变量τ及带噪动作块联合作为网络输入，由独立的动作专家网络以观测上下文为条件完成动作建模。

在训练阶段，采用流匹配(Flow Matching)训练范式[18]。给定真实动作块A_t与采样得到的高斯噪声$\varepsilon\sim\mathcal{N}(0,I)$，在时间变量$\tau\in[0,1]$上构造带噪插值动作：

$$A_t^\tau=\tau A_t+(1-\tau)\varepsilon \tag{3}$$

式中，流时间变量τ的采样引入Beta分布偏置，旨在引导模型充分学习带噪动作向目标动作演化的连续映射过程。动作专家网络负责接收带噪动作A_t^τ与当前观测O_t，并输出预测的条件速度场$v_\theta(A_t^\tau,O_t)$。而对应的目标速度场则定义为目标动作与初始状态之间的演化梯度：

$$u(A_t^\tau|A_t)=A_t-\varepsilon \tag{4}$$

模型通过最小化两者的均方误差完成优化：

$$\mathcal{L}_{\mathrm{FM}}(\theta)=E[\|v_\theta(A_t^\tau,O_t)-u(A_t^\tau|A_t)\|_2^2] \tag{5}$$

在推理阶段，策略从纯噪声$A_t^0\sim\mathcal{N}(0,I)$出发，采用欧拉法进行离散化迭代更新：

$$A_t^{\tau+\eta}=A_t^\tau+\eta v_\theta(A_t^\tau,O_t) \tag{6}$$

式中，η为积分步长。迭代结束后，系统获得连续动作块A_t，用于机器人底层控制。

1.2 视觉表征增强与时延补偿总体框架

本文整体结构如图1所示。其中，视觉表征分支用于增强模型对目标边界、局部结构及空间关系的编码能力；时延补偿分支则面向异步执行场景，通过训练阶段的偏移样本构造与推理期间的状态前推近似，来缓解预测区间与执行区间错位带来的

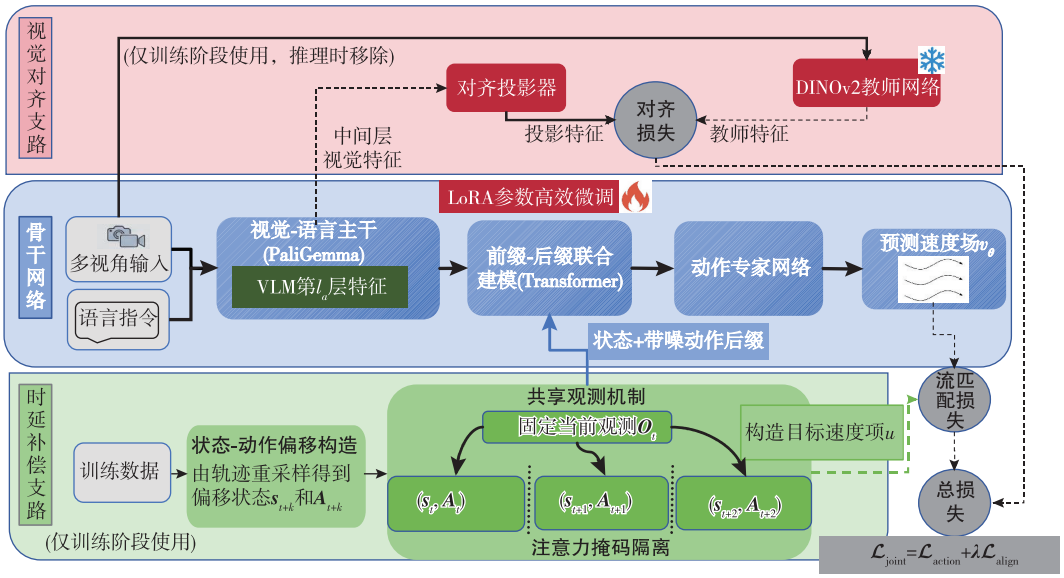


图1 融合视觉表征对齐与状态-动作偏移训练的VLA轻量化联合微调框架

时延问题。本文从空间感知误差与时序执行误差出发，将中间层视觉表征对齐和小步长时延补偿统一于轻量化部署框架中，以提升模型的空间定位能力以及对真实链路时延的适应性。

1.3　基于DINOv2的中间层视觉表征对齐

为增强VLA模型对局部结构与空间关系的刻画能力，本文在视觉-语言骨干网络第l_a层提取N个视觉特征令牌（token），记第i个令牌为$\boldsymbol{h}_i\in\mathbb{R}^{d_h}$，$d_h$为特征维度。随后，利用对齐投影器$P(\cdot)$将其映射到教师特征空间；对于同一输入图像，参数冻结的DINOv2教师网络输出对应的教师特征$\boldsymbol{z}_i\in\mathbb{R}^{d_z}$。在此基础上，通过最小化投影特征与教师特征之间的余弦距离，构造表征对齐损失：

$$\mathcal{L}_{\text{align}}=\frac{1}{N}\sum_{i=1}^{N}\left\{1-\cos\left[P(\boldsymbol{h}_i),\boldsymbol{z}_i\right]\right\}\tag{7}$$

式中，$\cos(\cdot)$表示余弦相似度。通过该约束，模型对目标边界、局部结构及相对位置等信息的编码能力得到增强，从而提升空间操作任务中的目标定位与接触感知能力。具体实现中，本文对所有相机视角分别提取DINOv2教师特征，并按照VLA图像输入顺序进行拼接对齐，图像分辨率统一调整为224×224。当两侧特征网格尺寸不一致时，对教师特征进行插值匹配，并通过有效区域掩码排除补边区域。与Spatial Forcing[10]采用VGGT[19]三维视觉教师网络不同，本文采用DINOv2作为教师网络，其原因在于：DINOv2通过大规模自监督预训练获得了较强的物体边界、区域一致性与语义结构表征能力，其特征在一定程度上能够反映场景的局部几何线索。采用它无需额外引入多视图对齐或三维重建过程，更适用于资源受限场景下的参数高效微调。

1.4　状态-动作偏移数据构造与状态前推补偿

为缓解真实机器人感知-推理-执行链路不同步所引起的预测-执行错位问题，本文采用基于状态-动作偏移训练与状态前推近似的小步长时延补偿机制。

设当前时刻观测为$\boldsymbol{O}_t$。为使策略适应执行时刻的状态变化，训练时从已有轨迹数据中采样状态偏移步长k，其中$k\in\{0,1,2,\cdots,k_{\max}\}$，$k_{\max}$表示最大偏移控制步数，其取值范围依据系统真实推理时延所对应的控制步数设定。在保持视觉图像与语言指令不变的条件下，从同一条轨迹中提取未来状态$\boldsymbol{s}_{t+k}$作为偏移状态输入，并构造其对应的未来动作块作为监督目标：

$$\boldsymbol{A}_{t+\delta}=[\boldsymbol{a}_{t+k},\boldsymbol{a}_{t+k+1},\cdots,\boldsymbol{a}_{t+H+k-1}]\tag{8}$$

通过这种时序非对称采样方式，模型能够在相同视觉输入下学习执行时刻状态变化与对应动作之间的关系，提高模型在视觉观测与执行状态存在轻微时序错位时的动作预测适应性。

在异步推理部署中，策略生成下一动作块需要一定推理时间。该推理时延约等于μ个控制步的时长，则在此期间，上一时刻输出的动作块$\boldsymbol{A}_{t-1}$中仍有一部分动作会继续执行，记为剩余动作$\boldsymbol{A}_{t-1}^{\text{rem}}$。基于此，本文利用$\boldsymbol{A}_{t-1}^{\text{rem}}$对当前状态$\boldsymbol{s}_t$进行前推近似：

$$\hat{\boldsymbol{s}}_{t+\mu}=F_{\text{roll}}(\boldsymbol{s}_t,\boldsymbol{A}_{t-1}^{\text{rem}})\tag{9}$$

式中，F_{roll}表示基于剩余动作的状态前推算子。其作用是在不引入额外动力学模型的条件下，按照控制步顺序将尚未执行的动作依次作用于当前机器人状态，从而近似估计新动作块实际开始执行时的系统状态。推理时，策略以当前观测$\boldsymbol{O}_t$与估计执行时刻状态$\hat{\boldsymbol{s}}_{t+\mu}$作为联合条件生成下一动作块，以此缓解感知-执行链路中的时延失配。

1.5　观测共享机制与参数高效微调

由于不同偏移分支共享同一观测$\boldsymbol{O}_t$，若对每个偏移分支独立进行观测特征编码，同一观测会被重复计算，从而显著增加训练成本。为此，本文采用观测共享机制，将同一观测对应的多个偏移分支合并到单次前向传播中。设最大偏移步长为$k_{\max}$，则输入序列可表示为：

$$[\boldsymbol{O}_t,(\boldsymbol{s}_t,\boldsymbol{A}_t),(\boldsymbol{s}_{t+1},\boldsymbol{A}_{t+1}),\cdots,(\boldsymbol{s}_{t+k_{\max}},\boldsymbol{A}_{t+k_{\max}})]$$

上述序列仅用于表示观测共享下的多偏移分支组织方式；在实际训练中，各分支动作块还需进一步施加流匹配噪声扰动。不同偏移分支共享同一观测输入，但分支之间不发生交叉注意力交互。观测$\boldsymbol{O}_t$对所有状态-动作分支可见，不同偏移量对应的状态-动作分支之间通过分块对角掩码[20-21]（block-diagonal mask）实现信息隔离，从而避免不同偏移分支之间的信息相互干扰。该设计避免了对同一视觉观测进行重复编码，从而降低多偏移训练中的前向计算冗余。

参数更新方面，本文采用低秩自适应（Low-Rank Adaptation，LoRA）[22]的参数高效微调策略，并配合少量任务适配模块进行协同更新，在保持大部分预训练参数不更新的前提下，为视觉表征对齐与时延补偿提供必要的参数调节空间，从而在计算成本、训练稳定性与任务适应性之间取得平衡。

1.6　联合优化目标与部署策略

综合考虑动作生成精度与空间表征约束，本文

的联合优化目标由流匹配动作监督损失和表征对齐损失两部分组成：

$$\mathcal{L}_{\text{joint}} = \mathcal{L}_{\text{action}} + \lambda \mathcal{L}_{\text{align}} \tag{10}$$

式中，λ为对齐损失权重。在推理阶段，系统移除训练阶段专用的空间监督分支与偏移样本构造分支，仅保留主干策略网络进行动作生成。状态前推补偿则作为面向真实执行链路的辅助补偿策略，用于存在异步时延时近似估计执行时刻状态。

2 实验设置与结果分析

2.1 实验平台与评价方式

实验所采用的硬件与软件环境如表1所示，主要训练与模型配置参数如表2所示。

表1 实验环境参数

参数	配置
操作系统	Ubuntu 22.04
GPU	GeForce RTX 4090(24 GB)
CPU	Intel Core i7-14700K CPU
系统内存	64 GB
CUDA版本	12.4
Python版本	3.10
PyTorch版本	2.1.2
机器人平台	SO101
仿真平台	LIBERO-Spatial

表2 主要训练与模型配置参数

参数	配置
总训练步数	LIBERO-Spatial：12000 SO101：8000
峰值学习率	LIBERO-Spatial：5×10^{-5} SO101：3×10^{-5}
等效批次大小	8
LoRA秩	16
教师网络	DINOv2-Large(参数冻结)
优化器	AdamW
动作块长度H	50
推理采样步数	10
真机训练轨迹总数	180

视觉对齐采用参数冻结的DINOv2-Large作为教师网络。基准实验中对齐损失权重$\lambda=0.1$，同时设置$\lambda=0.05$和$\lambda=0.5$两组对照；时延补偿模块的消融实验主要考察观测共享机制、最大时延步长配置对模型性能的影响。

仿真实验在LIBERO[23]仿真环境下开展，选用其中的LIBERO-Spatial任务集作为评测基准，该任务集包含10个空间操作任务。每个任务独立测试50个回合，总计500次测试；首先计算各任务的成功率，再取10个任务成功率的算术平均值作为模型的主要评价指标。对于基础模型、关键视觉层对齐配置和主要时延补偿配置，每个配置均采用随机种子1 000，1 001，1 002进行3次独立重复实验，取平均结果作为最终评估值；对于网络层级扫描、损失权重扫描及部分探索性消融实验，采用单一随机种子1 000进行评估。

真实机器人实验在SO101机械臂[24]平台上开展，选取3个具有代表性的空间操作任务，包括方块抓取与盒内放置、多高度物体放置以及目标载体移动干扰下的跟随放置。4组配置下每个任务测试20次，总计240次实验，重点考察模型抓取稳定性、放置准确性及扰动条件下的补偿能力。

在本文设定的训练配置下，所采用的两类基础模型的参数规模及可训练参数占比如表3所示。其中，可训练参数主要包括低秩适配层、对齐投影器及少量状态适配层等。

表3 基础模型配置与训练参数占比

基础模型	总参数量/B	可训练参数量/M	可训练参数占比/%
π_0	3.54	40.16	1.13
$\pi_{0.5}$	3.77	155.52	4.12

2.2 视觉表征对齐中网络层级选择对策略性能的影响

为分析视觉表征对齐中的网络层级选择对策略性能的影响，本文以π_0为基础模型开展对比实验。为控制单一变量，本次对比实验固定教师网络端的特征提取层为DINOv2-Large的最后一层输出，仅遍历学生网络端的对齐网络层级，结果如图2所示。本网络层级选择实验基于随机种子1 000的单次评估，仅用于比较不同对齐网络层级的相对性能变化趋势，不构成严格统计意义上的最优网络层级结论。

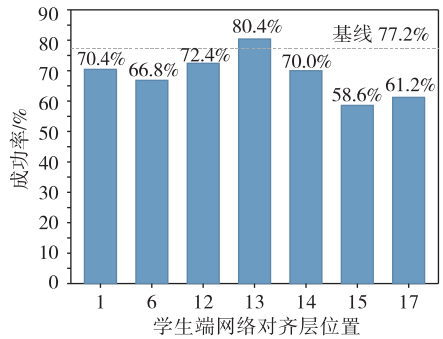


图2 π_0模型在不同视觉表征对齐层级下的LIBERO-Spatial任务成功率变化(单个随机种子筛选实验)

实验结果表明，视觉表征对齐的性能对网络层级选择较为敏感，并非越浅或越深越有效。在当前实验设置下，第 13 层对齐配置表现最佳，较基线模型提升 3.2%；而浅层和深层配置均未带来稳定增益，部分配置的性能甚至明显低于基线。这说明不同深度的特征层所承载的语义信息存在明显差异。其原因可能在于：浅层特征主要包含边缘、纹理等低层视觉信息，对目标结构与空间关系的表征尚不充分；深层特征虽具有更强的语义信息，但随着高层语义的全局融合，局部几何细节会被大幅削弱。因此，在本文实验设置下，相对较深但非最深的中间层更适合作为视觉表征对齐位置。

为进一步验证中间层对齐的有效性，本文在 LIBERO 抓取任务中对特征相似度分布进行可视化对比，结果如图 3 所示。

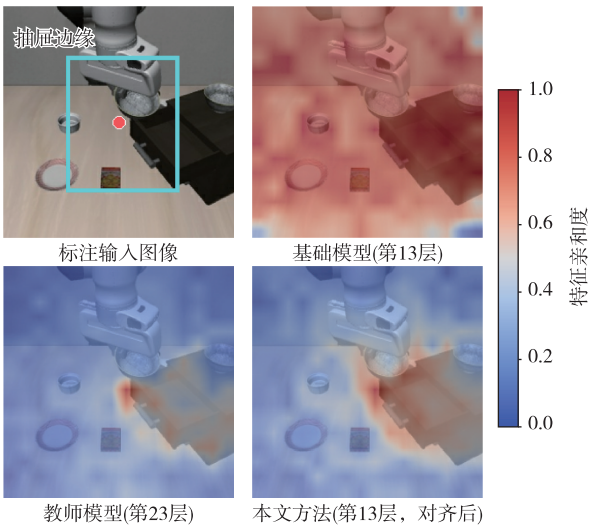


图 3　不同模型在仿真实验中的空间特征热力图对比

在相同观测条件下，分别基于基础模型、DINOv2 教师模型和本文方法提取特征，计算目标区域特征与图像其余区域特征之间的余弦相似度，生成空间特征热力图。可以看出，基础模型的高相似区域分布分散，易受无关背景干扰；经过中间层视觉表征对齐后，模型在目标碗、抽屉边缘等任务相关区域呈现出更集中的高相似分布，且整体分布模式与教师模型接近。这表明，所提出的中间层视觉表征对齐能够有效增强模型对任务相关局部结构的表征能力，为后续动作生成提供更可靠的空间信息。

在确定对齐网络层级后，本文进一步考察对齐损失权重 λ 对模型性能的影响，在第 13 层对齐条件下，当 $\lambda=0.05$ 时，模型成功率为 78.4%，较基线提升 1.2%；当 $\lambda=0.1$ 时，模型成功率达到 80.4%，较基线提升 3.2%；当 $\lambda=0.5$ 时，成功率下降至 63.8%，较基线下降 13.4%。该结果表明，适中的对齐监督有助于强化中间层的空间表征能力，而过大的对齐权重会打破联合优化的平衡，削弱流匹配动作监督的主导作用，导致策略过度拟合教师特征而偏离动作生成目标。因此，后续实验统一采用 $\lambda=0.1$ 作为对齐损失权重。

2.3　不同教师模型的性能与训练开销对比

为比较不同教师监督在本文轻量化微调框架下的适用性，本实验保持学生模型、训练数据、随机种子、训练步数、对齐网络层级及损失权重完全一致，仅替换教师网络，采用 3 个随机种子进行 3 次独立重复实验，实验结果如表 4 所示。在轻量化微调配置下，采用 DINOv2 作为教师模型，其训练开销明显低于 VGGT，峰值显存占用由 21.89 GB 降至 18.71 GB，单步训练耗时由 2.42 s 缩短至 1.46 s，总训练时长由 7.92 h 缩短至 4.27 h，两者的平均任务成功率保持在相当水平。

表 4　轻量化微调配置下 VGGT 与 DINOv2 教师模型的开销与性能对比

教师模型	峰值显存/GB	单步训练耗时/s	12 000 步总训练时长/h	成功率/%
VGGT-1B	21.89	2.42±0.03	7.92±0.05	75.2±3.86
DINOv2-Large	18.71	1.46±0.03	4.27±0.06	75.07±4.64

注：基础模型为 π_0。

实验结果表明，在本文所采用的轻量化微调框架下，DINOv2 能够在不牺牲模型执行性能的前提下，降低训练阶段的计算与时间开销。综合考量训练成本与任务表现，后续实验统一采用 DINOv2 作为默认的教师网络。

2.4　时延偏移步长对模型性能的影响

在确定视觉对齐层为第 13 层、损失权重 $\lambda=0.1$ 后，进一步考察观测共享机制下不同时延偏移配置对模型性能的影响，结果如图 4 所示，图中误差线表示 3 个随机种子的标准差。本节针对所有关键配置，均分别采用随机种子 1 000，1 001，1 002 开展 3 次独立重复实验，针对每个配置单独统计其平均成功率与标准差。因此，本节结果的绝对数值与 2.2 节的单次评估结果存在差异，但关于第 13 层对齐配置的核心结论仍保持一致。与基础模型相比，仅引入第 13 层空间对齐后，模型平均成功率由 72.60% 提升至 75.07%；进一步引入观测共享机制，并采用步长为 2 的时延偏

移训练后，平均成功率进一步提升至80.87%，较基线提高8.27%。这表明视觉表征对齐与时延补偿模块具有良好的协同增益效果。

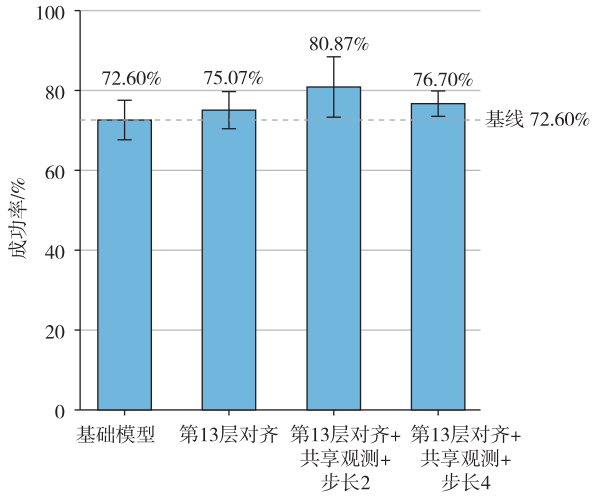


图4 π_0在LIBERO-Spatial任务集上关键配置的成功率对比

当时延偏移步长增大至4时，平均成功率回落至76.70%。虽较步长为2的配置明显下降，但仍高于基线水平。这说明适度的时延偏移有助于模型学习执行时刻的控制需求，而过大的时延偏移会扩大当前观测与未来状态之间的分布差异，增加训练难度并削弱控制稳定性。因此，在π_0基础模型与本文实验设置下，步长为2是训练阶段更合适的时延补偿配置。

2.5 $\pi_{0.5}$基础模型的视觉对齐与时延补偿分析

为了确保不同基座模型之间对比的公平性，$\pi_{0.5}$在仿真实验中的训练步数与π_0保持严格一致，同样设置为12 000步，基于$\pi_{0.5}$的仿真对比实验如图5所示。$\pi_{0.5}$在训练方案与数据利用方式上较π_0更为完备，因此其在LIBERO-Spatial仿真实验中平均任务成功率达到了94.27%。

实验结果表明，仅引入第13层视觉对齐，模型平均成功率即可提升至96.53%，说明视觉表征对齐在高性能基础模型上仍具有稳定增益。当直接引入步长为2的时延偏移训练时，$\pi_{0.5}$的平均成功率下降至91.80%。该结果说明，时延补偿模块并非在所有基础模型上都能带来稳定的正向增益，其效果与基础策略的状态注入方式、偏移步长设置及实际执行链路的异步程度密切相关。对于基线性能已较高的$\pi_{0.5}$模型，时延补偿更适合作为真实执行链路中的补偿策略，而非单纯用于提升仿真基准测试性能的常规手段。

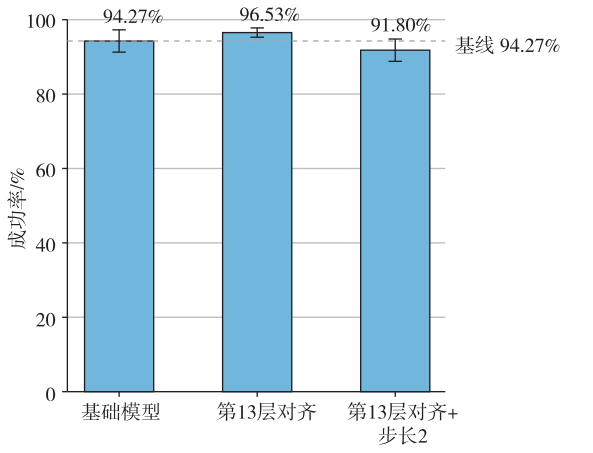
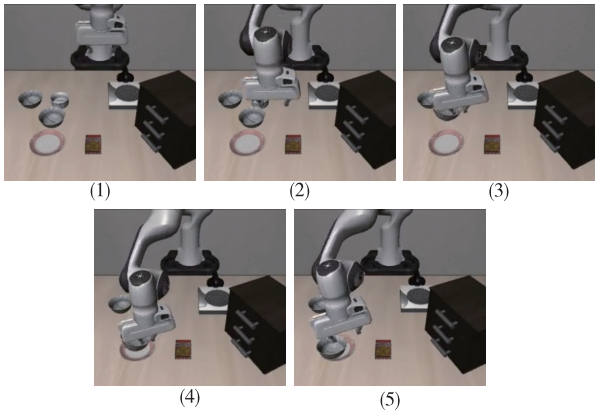
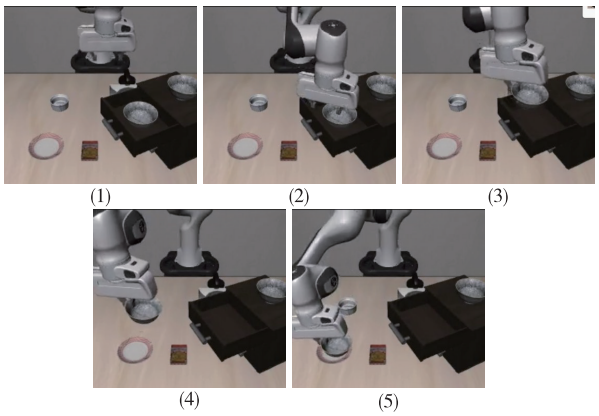


图5 $\pi_{0.5}$在LIBERO-Spatial任务集上关键配置的成功率对比

为进一步展示本文方法在代表性空间操作任务中的实际执行效果，图6给出了仿真环境下关键执行过程的可视化结果。可以看出，模型能够较为准确地区分盘子、小碗与目标黑碗之间的局部空间关系，并在抓取、搬运与放置过程中保持较好的动作连贯性。



（a） 抓取桌面碗至盘中



（b） 抓取柜中碗至盘中

图6 本文方法在LIBERO-Spatial代表性任务中的执行关键帧

2.6　真实机器人部署验证

基于上述实验结果，本文在真实机械臂部署中综合考虑基础模型性能、状态利用方式、在线时延与动作连贯性，最终确定以$\pi_{0.5}$为基础模型、结合第13层视觉对齐与小步长偏移的部署策略。本文对真实部署系统的时延特性进行了量化测试，结果表明：系统平均控制周期约为35.32 ms，端到端感知-推理-执行链路平均延迟约为81.88 ms，该总延迟约等于2.3个控制周期的时长。

因此，当新动作块开始执行时，机器人实际状态相较于当前观测时刻已向前更新约2～3个控制步。基于此，本文在训练阶段采用步长为2的时延偏移配置，以建模真实链路中的时延特性；在部署阶段，基于上一动作块中尚未执行完的2个控制步，通过状态前推近似估计执行时刻状态，使训练阶段的时延设定与真实执行链路中的时延尺度基本一致，从而减轻预测区间与实际执行区间错位对控制稳定性的影响。上述策略兼顾了模型从仿真迁移到真实系统时对性能、稳定性和计算成本的要求。

真实实验采用双视角视觉配置，包括一台安装于工作区上方的海康威视USB全局相机与一台安装于机械臂末端的杰锐微通USB腕部相机。全局相机提供任务工作区的整体观测，腕部相机补充抓取接触区域附近的局部视觉信息，如图7所示。

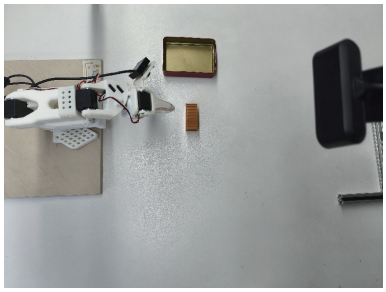


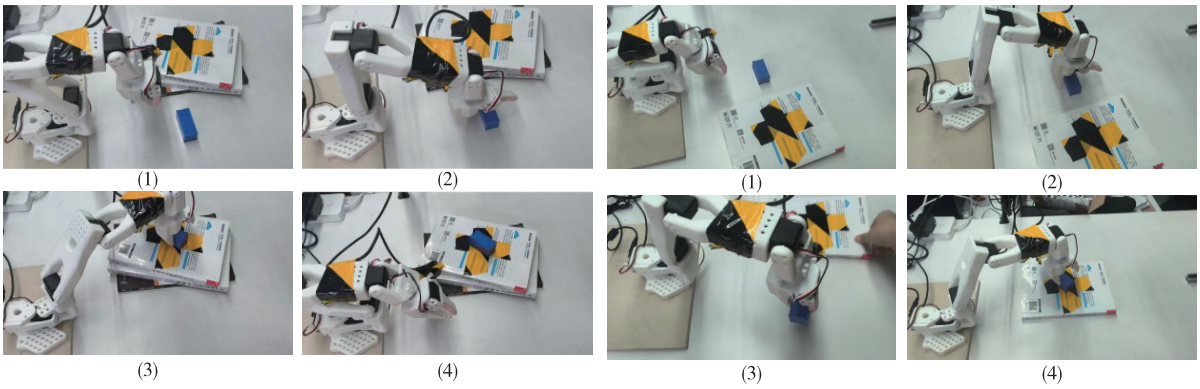
图7　真实机器人实验平台

本次实验以$\pi_{0.5}$为基础模型，在真实机械臂平台上对视觉对齐和小步长时延补偿两个模块进行分模块验证，实验结果见表5，真实环境下关键执行过程的可视化结果见图8。

表5　SO101真机分模块消融实验结果

任务名称	指标	$\pi_{0.5}$基础模型	$\pi_{0.5}$＋视觉对齐	$\pi_{0.5}$＋视觉对齐＋时延训练	$\pi_{0.5}$＋视觉对齐＋小步长时延补偿
方块抓取与盒内放置	成功率/%	70.0	85.0	85.0	95.0
	平均完成时间/s	16.67	16.94	16.81	15.86
多高度物体放置	成功率/%	55.0	90.0	90.0	90.0
	平均完成时间/s	18.08	17.78	17.25	14.78
目标载体移动干扰下的跟随放置	成功率/%	50.0	75.0	60.0	85.0
	平均完成时间/s	19.17	19.25	20.16	18.09

注：平均完成时间仅统计成功回合。



（a）多高度块体放置　　（b）目标载体移动干扰下的跟随放置

图8　SO101平台上多高度物体放置与移动载体跟随放置任务的执行过程

实验结果表明：视觉对齐模块能够在三类任务上稳定提升任务成功率。其中，方块抓取与盒内放置任务中，视觉对齐将基础模型的成功率由70.0%提升至85.0%，进一步引入小步长时延补偿后提升至95.0%，平均完成时间由16.67 s缩短至15.86 s；在多高度物体放置任务中，视觉对齐将

成功率由55.0%提升至90.0%，仅引入时延训练而不进行状态前推补偿时，成功率保持不变，但平均完成时间略有下降；进一步结合小步长时延补偿后，成功率仍保持在90.0%，平均完成时间缩短至14.78 s；对于目标载体移动干扰下的跟随放置任务，视觉对齐将成功率由50.0%提升至75.0%，仅引入时延训练而不进行状态前推补偿时，成功率下降至60.0%，平均完成时间增加至20.16 s；在训练阶段的时延建模与部署阶段的小步长状态前推联合使用后，成功率提升至85.0%，平均完成时间缩短至18.09 s。该结果说明，仅在训练阶段引入时延偏移而缺乏部署期状态前推补偿时，模型难以充分适应真实执行链路中的预测-执行错位；而训练与部署协同的小步长时延补偿策略能够更有效地缓解动态扰动场景下的时延失配问题。

对基础模型的失败案例分析发现，在多高度放置任务中难以准确识别书本表面的目标区域，易导致机械臂末端偏移至错误位置引发误放置，这正是视觉对齐模块能够针对性解决的核心缺陷。

综合三类任务的结果，结合前文仿真实验的图5与本次真实实验数据的表5表明，视觉表征对齐主要提升模型的空间感知与目标定位能力，该效果在仿真与真实场景下均得到体现；小步长时延补偿则进一步改善了真实部署中的动作连贯性与执行效率，尤其适配实际应用场景的时延误差补偿需求。

3 结论

为缓解VLA模型在空间操作中的细粒度几何理解缺失及推理-执行时延错位问题，本文提出一种面向资源受限空间操作部署场景的轻量化视觉表征对齐与小步长时延补偿方案。该方法选取参数冻结的DINOv2作为视觉教师网络，对VLA视觉-语言骨干网络的中间层视觉特征施加对齐监督，并通过观测共享机制与参数高效微调策略，在较低训练成本下提升视觉表征质量。实验结果表明，中间层表征对齐能够在不同基础模型上稳定提升空间操作性能，验证了基于教师监督的中间层特征约束能够有效增强模型对目标边界、局部结构及空间关系关键信息的编码能力。对于基线性能较高的模型，中间层表征对齐方法仍能带来稳定增益，表明其具有较好的跨模型适用性。

进一步分析表明，时延补偿的作用更多体现在提升模型对真实执行链路时延的适应能力，而非单纯用于提升仿真基准测试性能的常规手段。真实机械臂实验进一步验证了该部署配置的有效性，在不引入额外教师推理分支的条件下，该方法能够提升多类空间操作任务的完成能力并改善动作连贯性，表明所提方法兼顾了空间感知增强与工程实用性。

未来工作将围绕以下方向展开：构建更强的教师监督模型，设计更具鲁棒性的状态传播与时延补偿机制，并开展更广泛的基准测试与多平台真实部署验证。

参考文献：

[1] Stone A, Xiao T, Lu Y, et al. Open-world object manipulation using pre-trained vision-language models [EB/OL]. 2023: arXiv: 2303.00905. https://arxiv.org/abs/2303.00905

[2] 李浩然，陈宇辉，崔文博，等. 面向具身操作的视觉-语言-动作模型综述[J]. 自动化学报，2026，52(1)：18-51.
Li Haoran, Chen Yuhui, Cui Wenbo, et al. Survey of vision-language-action models for embodied manipulation [J]. Acta Automatica Sinica, 2026, 52(1): 18-51.

[3] Kim M J, Pertsch K, Karamcheti S, et al. OpenVLA: an open-source vision-language-action model [EB/OL]. 2024: arXiv: 2406.09246. https://arxiv.org/abs/2406.09246

[4] Black K, Brown N, Driess D, et al. π_0: a vision-language-action flow model for general robot control [EB/OL]. 2024: arXiv: 2410.24164. https://arxiv.org/abs/2410.24164

[5] Intelligence P, Black K, Brown N, et al. $\pi_{0.5}$: a vision-language-action model with open-world generalization [EB/OL]. 2025: arXiv: 2504.16054. https://arxiv.org/abs/2504.16054

[6] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision [EB/OL]. 2021: arXiv: 2103.00020. https://arxiv.org/abs/2103.00020

[7] Team O M, Ghosh D, Walke H, et al. Octo: an open-source generalist robot policy [EB/OL]. 2024: arXiv: 2405.12213. https://arxiv.org/abs/2405.12213

[8] Sapkota R, Cao Y, Roumeliotis K I, et al. Vision-language-action (VLA) models: concepts, progress, applications and challenges [EB/OL]. 2025: arXiv: 2505.04769. https://arxiv.org/abs/2505.04769

[9] Huang X, Wu J J, Xie Q Y, et al. MLLMs need 3D-aware representation supervision for scene understanding [EB/OL]. 2025:arXiv: 2506.01946. https://arxiv.org/abs/2506.01946

[10] Li F, Song W, Zhao H, et al. Spatial forcing: implicit spatial representation alignment for vision-language-action model

[EB/OL]. 2025: arXiv: 2510.12276. https://arxiv.org/abs/2510.12276

[11] 代少升，周　漫，余自安，等. 基于改进YOLOv8n的货车盲区目标检测方法[J]. 半导体光电，2025，46(1)：172-179.
Dai Shaosheng, Zhou Man, Yu Zian, et al. Object detection in the blind spot of truck based on improved YOLOv8n[J]. Semiconductor Optoelectronics, 2025, 46(1): 172-179.

[12] 王德智，周芸伊，刘晗庆，等. 多层次架构下融合自注意力的三维激光点云语义分割算法[J]. 半导体光电，2024，45(4)：646-652.
Wang Dezhi, Zhou Yunyi, Liu Hanqing, et al. 3D laser point cloud with self-attention under multi-level architecture semantic segmentation algorithms [J]. Semiconductor Optoelectronics, 2024, 45(4): 646-652.

[13] 罗　元，马文龙，唐小平. 基于深度学习的视觉目标跟踪研究综述[J]. 半导体光电，2026，47(1)：13-27.
Luo Yuan, Ma Wenlong, Tang Xiaoping. Review of visual object tracking based on deep learning[J]. Semiconductor Optoelectronics, 2026, 47(1): 13-27.

[14] Black K, Galliker M Y, Levine S. Real-time execution of action chunking flow policies[EB/OL]. 2025: arXiv: 2506.07339. https://arxiv.org/abs/2506.07339

[15] Tang J, Sun Y, Zhao Y, et al. VLASH: real-time VLAs *via* future-state-aware asynchronous inference [EB/OL]. 2025: arXiv: 2512.01031. https://arxiv.org/abs/2512.01031

[16] Oquab M, Darcet T, Moutakanni T, et al. DINOv2: learning robust visual features without supervision[EB/OL]. 2023: arXiv: 2304.07193. https://arxiv.org/abs/2304.07193

[17] Beyer L, Steiner A, Pinto A S, et al. PaliGemma: a versatile 3B VLM for transfer[EB/OL]. 2024: arXiv: 2407.07726. https://arxiv.org/abs/2407.07726

[18] Lipman Y, Chen R T Q, Ben-Hamu H, et al. Flow matching for generative modeling[EB/OL]. 2022: arXiv: 2210.02747. https://arxiv.org/abs/2210.02747

[19] Wang J, Chen M, Karaev N, et al. VGGT: visual geometry grounded transformer[C]//2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2025: 5294-5306.

[20] Han K, Wang Y, Chen H, et al. A survey on vision transformer[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 87-110.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.

[22] 吴建华，刘振宇，曾　睿，等. 基于LoRA微调大模型在网络安全等级保护测评报告质量检测中的应用研究[J]. 数据与计算发展前沿(中英文)，2025，7(6)：111-123.
Wu Jianhua, Liu Zhenyu, Zeng Rui, et al. Research on the application of fine-tuned large language models based on LoRA in quality evaluation of the security level protection assessment reports [J]. Frontiers of Data & Computing, 2025, 7(6): 111-123.

[23] Liu B, Zhu Y, Gao C, et al. LIBERO: benchmarking knowledge transfer for lifelong robot learning[C]//Advances in Neural Information Processing Systems 36. Neural Information Processing Systems Foundation, Inc. 2023: 44776-44791.

[24] Cadene R, Aliberts S, Capuano F, et al. LeRobot: an open-source library for end-to-end robot learning[EB/OL]. 2026: arXiv: 2602.22818. https://arxiv.org/abs/2602.22818

作者简介：

李　达(2002—)，男，硕士研究生，主要研究方向为具身智能；

曹明轩(1986—)，男，博士，研究生导师，主要研究方向为具身智能、机器视觉及深度学习算法优化、大功率激光芯片热管理、激光选区熔覆设备研发。