研究报告

1. 执行摘要

以自我为中心的视频理解是具身人工智能系统面临的一项根本性挑战，这些系统必须从第一人称视角感知、推理并在实际环境中行动。本研究报告综合了一场研讨会演讲的发现，涵盖了以自我为中心视频理解的五篇最新论文——涉及独特视频描述、场景条件视觉指令生成、包含视野外物体的 3D 物体追踪、包含超出相机视野的手部姿态预测，以及 HD Epic 高细节自我中心视频数据集及其对应的 VQA 基准——以及近期视频语言模型、基于扩散的轨迹预测、以自我为中心视频的 3D 场景理解和暴露 VLM 在物理和运动理解方面关键局限性的评估基准的综合文献综述。

该研究领域解决的核心问题是，当前视觉语言模型（VLM）在细粒度运动级感知任务上仅能达到接近随机的性能。MVP 基准表明，在物理理解任务上，最好的开源视频语言模型仅达到 40.2%的准确率，而人类表现为 92.9%，差距达 52.7 个百分点。这种失败不仅仅是性能差距——它代表了 VLM 在从以自我为中心视频中表示和推理 3D 空间、物体运动和物理因果关系方面的根本性局限。

在文献记录的最具体进展中，判别性提示描述（CDP）通过预测判别性属性来分离视觉相似的片段，使自我中心视频的文本到视频召回率在第 1 位提升了 33 个百分点。Lift-Match-Keep（LMK）流程在物体离开相机视野 120 秒后实现了 57%的正确物体定位，而其他 3D 方法仅为 33%，2D 追踪仅为 17%。EgoH4 表明，身体姿态约束即使在双手离开相机画面时也能实现 3D 手部预测，比基线在轨迹准确度上提升 3.4 厘米，在姿态准确度上提升 5.1 厘米。这些进展为解决以自我为中心视频理解的核心挑战提供了具体机制，尽管在单个组件能力与能够可靠实际部署的集成系统之间仍存在显著差距。

主要未解决的瓶颈是缺乏将 3D 信息摄入 VLM 的可行方法，以及将各个组件集成到统一流程中。会议观察到的"目前尚无明确方法将 3D 信息摄入 VLM"仍然有效，尚未在集成形式中展示结合独特描述、3D 追踪和手部预测的端到端系统。这一点对项目很重要，因为在实际以自我为中心视频环境中运行的具身人工智能系统正是需要这些能力——保持对当前不可见物体的意识、预测手部轨迹以预判操作，以及生成能够区分视觉相似动作的准确描述。

2. 问题设置与来源背景

本报告的基础材料来源于一场关于以自我为中心视频理解最新进展的研讨会演讲，作为计算机视觉研讨会的一部分。演讲涵盖了五个不同但相关的研究方向，每个方向都针对当前以自我为中心视频理解系统的特定局限性或能力差距。

第一个方向涉及重复性以自我为中心动作的独特视频描述。日常生活长视频包含许多重复的动作、事件和镜头。当使用标准现成描述器对这些视频进行描述时，Ego4D 中 66%的片段与至少另一个片段共享描述，严重影响基于文本的视频检索。用户必须线性扫描多个视觉相似的片段才能找到特定内容。这个问题在以自我为中心视频中尤为突出，因为第一人称视角固有地涉

及重复动作——烹饪流程、工具使用、导航模式——这些动作只在细微的上下文细节上有所不同。

第二个方向解决以自我为中心视频中的 3D 物体追踪，特别关注离开相机视野的物体。与保持对当前不可见物体的空间意识的人类不同，现有方法在物体被遮挡或移出视野时会失败。在以自我为中心视频中追踪离开相机视野的物体需要将 2D 观测提升到 3D 世界坐标，并在较长时间内保持物体身份。该方法使用深度估计和 SLAM 将 2D 边界框提升到 3D 世界坐标，然后使用 3D 欧几里得距离、视觉外观（ReID 特征）和交互上下文（靠近手的物体很可能正在被操作）随时间匹配物体。

第三个方向涉及手部姿态预测，特别是预测手部离开相机视野时的手部位置。现有手部姿态预测方法仅在手部可见时预测手部位置，而忽略了即使手部离开画面，也可以从全身姿态推断手部运动。手部运动的随机性需要概率预测，而 2D 位置会受到自我运动（相机移动）的严重影响。EgoH4 方法使用基于扩散的轨迹预测，观察 2 秒预测 1 秒，全身姿态为手部位置提供运动学约束。

第四个方向审视 VLM 在物理和运动理解方面的局限性。现有视频问答基准存在"分数膨胀"问题，模型通过利用基于表面视觉或文本线索的捷径解决方案获得高分，而非真正的物理理解。当前的 VLM 在细粒度运动级感知任务上表现不佳。MVP 基准使用最小变化对，其中每个样本都有一个视觉相似的视频对，具有相同的问题但相反的答案——模型必须正确回答两者才能获得分数。

第五个方向涉及场景条件视觉指令生成。从输入图像（提供场景上下文）和文本指令生成逐步视觉指令，同时保持与输入场景的一致性并生成时间连贯的动作序列，这代表了与描述不同的挑战——需要生成能力而非判别能力。

重要的来源特定上下文包括 HD Epic 数据集设计约束：它被设计为仅用于验证，不应用于微调。这一约束反映了数据集质量的更广泛挑战——视频数据偏向于公开可用内容（说话的头部、电影），而非真实世界的自我场景。数据集提供三层标注：具有精确时间和权重的高细节原料标注、具有动作-物体-手部-推理的详细叙述，以及包括物体追踪和厨房固定装置的 3D 标注。

3. 来自基础材料的接地发现

3.1 重复动作描述问题

基础笔记记录了判别性提示技术可以显著提高重复性自我视频的视频描述唯一性。关于独特描述的 ACCV 最佳论文使用判别性提示技术为重复性自我视频中的每个片段生成独特描述。该方法使用训练网络来选择适当的判别性提示，将一对一映射从 37%提高到 76%。当无法进行唯一描述时，系统推进时间范围以找到区分性上下文——"X 然后 Y"与"X 然后 Z"，其中 Y 和 Z 是不同的后续事件。

接地发现表明，这种方法是描述器无关的，可以集成到基于 VLM 的视频理解流程中。关键洞察是，区分视觉相似的片段需要查看更广泛的上下文——不仅仅是片段本身，而是之前和之后发生了什么。当纯粹片段内判别失败时，这种时间扩展策略提供了后备方案。

3.2 场景条件视觉指令生成

ShowHowTo 方法代表了不同模式的视觉指令生成——接收输入图像和食谱步骤，并使用可用原料和设备生成图像。该方法使用从 HowTo100M 数据集筛选出的 578K 序列和 450 万步骤。值得注意的是，基础笔记记录了这种方法在某些情况下比原始视频获得更好的人类偏好评分，表明生成的指令实际上可以超过真实世界演示的质量。

接地发现识别出一个关键挑战：当共同先验与动作序列矛盾时会发生状态一致性错误。例如，如果模型学到"鸡蛋通常打入碗中"，但食谱显示直接打入锅中，先验可能会覆盖视觉证据。这一挑战与通过动态帧驱逐保持长视频一致性的 StableWorld 文献直接相关。

3.3 超出相机视野的 3D 物体追踪

Out of Sight, Not Out of Mind 方法使用深度估计和 SLAM 将 2D 物体追踪提升到 3D 场景理解。该系统即使在物体处于相机视野外时也能在 3D 空间中追踪动态物体，达到约 3 厘米的准确度。这种准确度足以用于物体定位应用，并支持关于物体可见性、遮挡、可及性和空间关系的查询。

基础笔记识别出一个关键局限性：当前深度估计器在精确 3D 重建方面存在已知局限性。此外，生成模型在处理发动机零件和其他难以生成的物体（缺乏独特视觉特征或外观高度可变的物体）时存在困难。

3.4 不可见手部姿态预测

EgoH4 方法使用身体姿态估计预测相机视野外的手部位置。该方法使用基于扩散的轨迹预测，观察 2 秒预测 1 秒。当手部可见时，投影的手部姿态通过 3D 到 2D 重投影损失提供额外的监督。

接地发现注意到该方法在手动手部关节而非工具交互上进行评估，这代表了验证范围的空白。将手部姿态预测扩展到更多样的身体配置和工具交互仍然是基础笔记中识别的开放挑战。

3.5 VLM 在 HD Epic VQA 基准上的性能

HD Epic 数据集用于创建 VQA 基准，证明 VLM 在物体运动问题上表现接近随机。基础笔记记录了一个关键洞察：更强的 VLM 提高似然性，但实际上可能掩盖接地失败。生成自信但错误答案的模型比输出接近随机猜测的模型更不利于调试，因为随机输出揭示了模型失败的地方，而自信的错误答案掩盖了这些失败。

3.6 未解决的问题与分歧

基础笔记识别出几个未解决的问题：VLM 视频采样的最佳策略（均匀 FPS 与基于问题类型的自适应）仍然悬而未决；是否在特定领域数据集上微调 VLM 存在争议；以及 3D 到 VLM 特征条件的最佳方法仍然是一个开放的研究问题。这些分歧反映了如何将 3D 场景理解集成到基于语言模型的视频理解流程中的更广泛不确定性。

4. 基于文献的深入分析

4.1 保留的详细论文分析

论文 1：判别性提示描述（CDP）

论文： "It's Just Another Day: Unique Video Captioning by Discriminitive Prompting"（Perrett 等，2024） **arXiv：** https://arxiv.org/html/2410.11702v1 **会议：** BMVC 2024

问题与任务设置

日常生活长视频包含许多重复的动作、事件和镜头。当使用标准现成描述器对这些视频进行描述时，Ego4D 中 66%的片段与至少另一个片段共享描述，严重影响基于文本的视频检索。用户必须线性扫描多个视觉相似的片段才能找到特定内容。这个问题对以自我为中心视频尤其突出，因为第一人称视角固有地捕获重复的流程——烹饪序列、工具操作、导航模式——这些只在细微的上下文细节上有所不同。

方法论

CDP 通过三个关键机制解决独特描述问题：

1. **判别性提示库**：一组通用提示（如"握着"、"看着"、"另一个人"），从训练数据中的频繁 N-gram 中选择。固定提示提供可解释性，可以设计为多样化。

2. **组合搜索**：对于每个片段，方法找到最大化唯一性边际的提示或组合（最多 α=3 个提示）——相对于次优匹配的片段到描述匹配的改进。

3. **CDPNet**：一个轻量级变压器编码器（2 层、4 头、1024 前馈维度），在不穷举运行描述器的情况下预测视觉-文本相似性。这将推理从 300 秒（穷举搜索）减少到 5.8 秒。

时间扩展策略： 当无法找到唯一描述时（边际≤λ 阈值），方法推进时间范围直到实现唯一性，描述"X 然后 Y"与"X 然后 Z"，其中 Y 和 Z 是不同的后续事件。

主要证据

在自我中心基准（Ego4D 与 LaViLa VCLM）上，CDP 实现：

配置	平均 R@1	Cycle@1	改进
LaViLa VCLM（T=+0 秒）	37%	22%	基线
CDP（T=+0 秒）	45%	26%	+8% / +4%
LaViLa VCLM（T=+5 秒）	38%	23%	基线
CDP（T=+5 秒）	57%	39%	+19% / +16%
LaViLa	43%	27%	基线

VCLM（T=+30 秒）			
CDP（T=+30 秒）	76%	62%	+33% / +35%

在 Timeloop Movies（Video-LLaMA）上：

配置	平均 R@1	Cycle@1	改进
Video-LLaMA（T=+10 秒，5 个片段）	38%	18%	基线
CDP（T=+10 秒，5 个片段）	63%	45%	+25% / +27%

关键发现：CDP 在更好的基线模型上提供更大的改进，表明该方法将在描述能力提高时保持相关性。

相关性

直接适用于重复性自我视频的唯一描述。时间推进策略与会议观察一致，即"当唯一描述不可能时，推进视频以找到区分性上下文"。该方法是描述器无关的，可以集成到基于 VLM 的视频理解流程中。观察所有相似片段并预测区分属性的方法为提高重复视频中的描述唯一性提供了具体机制。

局限性

- 需要访问具有相同描述的所有片段进行比较
- 固定提示库可能不适用于所有视频类型
- 组合搜索的计算开销（通过 CDPNet 近似得到缓解）

论文 2：TA-Prompting — 视频 VLM 的时间锚点

论文： "TA-Prompting: Enhancing Video Large Language Models for Dense Video Captioning via Temporal Anchors"（2026） **arXiv：** https://arxiv.org/html/2601.02908

问题与任务设置

现有 VideoLLM 难以进行精确的事件边界检测，导致描述与视频内容结合不良。语言先验经常覆盖视觉证据，导致时间上不对齐的描述。这个问题与以自我为中心视频理解直接相关，因为日常活动涉及必须分割成有意义的单元以进行描述、总结和检索的连续动作流。

方法论

1. **时间锚点**：学习性表示，精确地定位视频中的事件。不是使用文本标记来描述时间，而是锚点直接对应于视频片段，实现视觉接地。

2. **事件连贯采样（ECS）**：在推理过程中，选择在时间事件上具有足够连贯性且与视频具有跨模态相似性的事件描述。这取代了束搜索，同时保持效率。

方法细节：

- 在 VTimeLLM 的边界感知数据上进行预训练
- 为密集视频描述进行微调
- 使用均匀采样帧的 CLIP 特征

主要证据

在 ActivityNet-Caption 密集视频描述上，TA-Prompting 优于包括 VTimeLLM 和 LITA 在内的最新 VideoLLM。ECS 以相似的计算时间实现了比束搜索更好的性能。该方法表明，通过时间锚点的视觉接地显著提高了描述保真度，相比基于语言先验的方法。

相关性

解决了对 VLM 描述质量和更好时间理解的关注。时间锚定方法为改进事件定位提供了具体机制，可应用于改进基于 VLM 的视频理解系统。TA-Prompting 使用通过时间锚点的视觉接地，为改进 VLM 事件定位提供了一条途径，而不单独依赖语言先验。

局限性

- 存在基于 LLM 主干的幻觉描述风险
- 均匀采样帧的 CLIP 特征可能错过运动细节
- 锚点和 LLM 的联合优化增加了复杂性

论文 3：OSNOM — 视野之外，记忆之中

论文： "Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind"（2024） **arXiv：** https://arxiv.org/html/2404.05072 **会议：** CVPR 2025

问题与任务设置

在以自我为中心视频中追踪离开相机视野的物体需要将 2D 观测提升到 3D 世界坐标，并在较长时间内保持物体身份。与保持对当前不可见物体空间意识的人类不同，现有方法在物体被遮挡或移出视野时会失败。

方法论：Lift-Match-Keep（LMK）流程

1. **提升**：使用来自 SLAM/ARIA 传感器的深度估计和相机姿态将 2D 边界框提升到 3D。物体被投影到世界坐标。

2. **匹配**：使用以下方式随时间匹配物体：

- 3D 欧几里得距离 - 视觉外观（ReID 特征） - 交互上下文（靠近手的物体很可能正在被操作）

3. **保持**：物体轨迹在世界坐标内存中保持，支持关于当前不可见物体的查询。

数据集： 来自 EPIC-KITCHENS 的 100 个长视频（平均 12 分钟）、45 个不同厨房、总共 25 小时、7939 个掩码对应 2939 个物体。

主要证据

视野外持续时间	LMK	其他 3D 方法	2D 追踪
120 秒	57%	33%	17%

关键发现：交互上下文（手部接近度）显著提高匹配准确度，因为正在操作的物体保持时空连续性。Lift-Match-Keep 流程为在物体离开相机视野时保持物体轨迹提供了具体方法，直接支持空间推理查询。

相关性

为支持物体可见性、遮挡、可及性和空间关系查询的 3D 场景理解提供了具体流程。该方法达到约 3 厘米的准确度，足以用于物体定位应用。LMK 方法直接适用于会议接地中记录的"视野之外，记忆之中"挑战。

局限性

- 需要深度估计（深度估计器存在已知局限性）
- SLAM/ARIA 传感器数据可能并非在所有以自我为中心数据集中都可用
- 交互上下文（手部接近度）改善匹配但需要手部检测

论文 4：Ego3DT — 零样本 3D 物体追踪

论文： "Ego3DT: Tracking Every 3D Object in Ego-centric Videos"（Hao 等，2024）
arXiv： https://arxiv.org/html/2410.08530v1 **会议：** ACM MM 2024

问题与任务设置

以自我为中心视频中的零样本 3D 物体追踪需要能够在没有任务特定训练或数据集特定标注的情况下追踪任意物体的方法。挑战是在扩展视频序列中通过遮挡、视角变化和外观变化保持物体身份。

方法论

1. **2D 分割和开放词汇检测**：使用 GLEE 进行开放词汇物体检测和使用 SAM 进行分割，提供无需数据集特定训练的身份检测。

2. **窗口级 3D 场**：使用 DUSt3R 从相邻视频帧进行 3D 场景重建，将 2D 分割坐标映射到 3D 空间。

3. **动态分层关联**：用于处理物体出现、消失和遮挡的稳定追踪轨迹的分层机制。使用匈牙利算法进行初始化和用于帧间对齐的 3D 场景配准。

数学公式：

- 物体检测：$O^{Det}_{2D} = Det(X)$
- 语义分割：$O^{Seg}*{2D} = Seg(O^{Det}*{2D})$
- 3D 估计：$O^{3D} = G(X, O^{Seg}_{2D})$ 其中 G 是 DUSt3R
- 匹配：$Y = M(O^{3D}) = PointMatch(A(O^{3D}))$

主要证据

- HOTA 改进：比先前方法提高 1.04 倍到 2.90 倍
- 零样本方法实现无需任务特定训练的部署
- 成功通过遮挡和视角变化追踪物体

相关性

补充 OSNOM：Ego3DT 的零样本方法意味着它可以在没有数据集特定训练的情况下应用，可能在不同的以自我为中心视频领域更广泛适用。开放词汇检测和 3D 重建的结合为 3D 物体意识提供了可扩展的方法。

局限性

- 依赖预训练的深度估计和重建模型
- 可能难以处理无纹理或反射表面
- 分层关联增加计算开销

论文 5：EgoH4 — 不可见的自我手预测

论文： "The Invisible EgoHand: 3D Hand Forecasting through EgoBody Pose Estimation"（2025） **arXiv：** https://arxiv.org/html/2504.08654v1

问题与任务设置

现有手部姿态预测方法仅在手部可见时预测手部位置，忽略了即使手部离开画面，也可以从全身姿态推断手部运动。手部运动的随机性需要概率预测，而 2D 位置会受到自我运动（相机移动）的严重影响。这一局限性阻止了预测性系统在手部移出相机视野时预测手部轨迹。

方法论

1. **身体姿态约束**：利用 EgoBody 姿态估计来约束手部运动。全身姿态为手部位置提供运动学约束。
2. **联合优化**：联合去噪手部和身体关节，身体关节作为手部运动的约束。
3. **可见性预测器**：估计手部可见性的分类器，提高处理不可见手部的能力。
4. **基于扩散的预测**：使用基于扩散的变压器模型生成多个似真的手部路径。
5. **3D 到 2D 重投影损失**：最小化手部在视野内时的误差。

数据集： Ego-Exo4D，包含 156K 训练序列和 34K 测试序列。使用 3D 标注，即使手部在相机视野外（感谢多个外视角相机）。

主要证据

指标	基线	EgoH4	改进
手部轨迹 ADE	-	-	+3.4 厘米
手部姿态 MPJPE	-	-	+5.1 厘米

不同视野外比例的性能：

视野外比例	ADF（基线）	ADF（我们的）	FDE（基线）	FDE（我们的）
(0.0, 0.2]	0.284	0.236	0.329	0.284
(0.8, 1.0]	0.363	0.335	0.459	0.434

相关性

直接解决了会议中的"不可见的自我手"讨论。该方法提供了超出相机视野的手部姿态预测，使用身体姿态作为额外监督。这使得预测性系统能够预测手部将重新进入画面的位置，或接下来将操作什么工具。

局限性

- 需要准确的身体姿态估计作为输入
- 性能取决于运动学模型的质量
- 在手动手部关节而非工具交互上进行评估

论文 6：MADiff — 运动感知 Mamba 扩散

论文： "MADiff: Motion-Aware Mamba Diffusion Models for Hand Trajectory Prediction on Egocentric Videos"（2024） **arXiv：** https://arxiv.org/html/2409.02638

问题与任务设置

以自我为中心视频中的手部轨迹预测必须考虑纠缠的手部和相机运动。标准方法失败，因为它们将相机运动视为噪声而非约束手部运动预测的结构化信号。

方法论

1. **运动驱动选择性扫描（MDSS）**：将相机自我运动整合到去噪过程中，捕获具有时间因果性的纠缠手部和相机运动模式。

2. **语义特征提取**：使用融合视觉和语言特征的基础模型来理解手部-场景关系，无需显式的可供性标签。

3. **潜空间去噪**：在压缩的潜空间中进行操作以提高效率。

主要证据

- 在五个公共数据集上实现最新结果
- 实时推理能力（在边缘设备上测试）

- MDSS 有效地将手部运动与相机运动分离

不同视野外比例的性能：

视野外比例	ADF（基线）	ADF（MADiff）	FDE（基线）	FDE（MADiff）
(0.0, 0.2]	0.284	0.236	0.329	0.284
(0.8, 1.0]	0.363	0.335	0.459	0.434

相关性

证明了扩散模型有效地处理手部运动的随机性，同时考虑相机自我运动。该方法为在显著相机运动期间保持手部追踪准确度提供了经过验证的方法。

局限性

- 需要相机姿态估计作为输入
- 潜空间压缩可能丢失细粒度细节

论文 7：MMTwin — 多模态双胞胎扩散

论文： "Novel Diffusion Models for Multimodal 3D Hand Trajectory Prediction"（2025） **arXiv：** https://arxiv.org/html/2504.07375v1

问题与任务设置

手部轨迹预测需要整合多种输入模态——光流、相机轨迹和视觉特征——每种都为手部运动和场景结构提供互补信息。

方法论

1. **双胞胎潜扩散模型**：用于相机自我运动预测和手部轨迹预测的两个独立模型。
2. **相机自我运动预测模型**：从光流和轨迹预测未来相机运动。
3. **手部轨迹预测模型**：使用预测的相机运动作为上下文预测手部航点。
4. **混合 Mamba-变压器模块**：更好的多模态特征融合。

主要证据

通过实验验证，显示相机运动和手部运动的单独但协调的模型优于必须联合学习两种关系的集成方法。

相关性

为基于扩散的手部预测提供了替代架构，明确将相机自我运动建模为单独的预测问题。双胞胎模型方法可能更加模块化，更容易适应新领域。

局限性

- 需要多种输入模态
- 双胞胎模型协调增加复杂性
- 可能比单模型方法需要更多参数

论文 8：MVP 基准 — 捷径感知视频问答

论文： "A Shortcut-aware Video-QA Benchmark for Physical Understanding via Minimal Video Pairs"（2025） **arXiv：** https://arxiv.org/html/2506.09987v1

问题与任务设置

现有视频问答基准存在"分数膨胀"问题，模型通过利用基于表面视觉或文本线索的捷径解决方案获得高分，而非真正的物理理解。当前的 VLM 在细粒度运动级感知任务上表现不佳。这个问题对以自我为中心视频理解尤其突出，因为物理交互——物体操作、工具使用、因果事件——需要推理 3D 空间、物体永久性和因果关系，而这些无法通过表面视觉模式捕获。

方法论

1. **最小变化对**：每个样本都有一个视觉相似的视频对，具有相同的问题但相反的答案。模型必须正确回答两者才能获得分数。

2. **物理理解重点**：问题需要推理物体永久性、空间关系和因果关系。

数据集： 55K 个需要真正物理理解的最小变化对示例。

主要证据

模型	准确率
人类表现	92.9%
最佳开源 Video-LLM	40.2%
差距	52.7 个百分点

关键发现：在标准基准上表现良好的模型在 MVP 上表现不佳，确认标准基准允许捷径利用。即使像 Gemini Pro 这样的强大模型在 HD-EPIC 26.6K 问题基准上也仅达到 37.6%的准确率，突显了当前视觉语言模型在细粒度动作理解、3D 感知和物体运动问题上的显著不足。

相关性

MVP 基准暴露了 VLM 在物体运动理解方面的失败，而非允许来自捷径解决方案的膨胀分数。52.7 个百分点的差距确认了在 VLM 能够可靠处理 3D 空间推理之前还有大量工作要做。这一发现直接验证了会议接地中表达的关于 VLM 在物体运动问题上局限性的担忧。

局限性

- 仅限于多选题格式
- 数据集规模可能无法覆盖所有物理推理类型
- 捷径感知需要仔细的基准设计

论文 9：几何引导的相机运动理解

论文： "Geometry-Guided Camera Motion Understanding in VideoLLMs"（2026）
arXiv： https://arxiv.org/html/2603.13119v1

问题与任务设置

运动相关信息在 VLM 预训练期间被捕获不足。VideoLLM 难以区分物体运动和相机运动，导致需要推理两者的物理理解任务失败。

方法论

1. 显式编码相机运动参数
2. 教授模型区分物体运动和相机运动
3. 使用几何一致性作为监督

主要证据

探测实验揭示，运动相关信息在预训练期间被捕获不足。几何引导方法为将空间意识纳入 VideoLLM 提供了途径。

相关性

解决了一个根本性差距：如何将几何和空间信息摄入 VLM。运动相关信息捕获不足的发现确认了需要架构改变或明确纳入几何推理的训练方法。

局限性

- 训练期间需要额外的监督信号
- 可能增加计算开销
- 需要与现有 VLM 架构仔细集成

论文 10：ShowHowTo — 场景条件视觉指令生成

论文： "ShowHowTo: Generating Scene-Conditioned Step-by-Step Visual Instructions"（CVPR 2025） **URL：** https://soczech.github.io/showhowto/

问题与任务设置

从输入图像（提供场景上下文）和文本指令生成逐步视觉指令，同时保持与输入场景的一致性并生成时间连贯的动作序列。挑战是在生成的帧中保持场景身份、物体存在和布局，同时准确描绘请求的动作序列。

方法论

1. **大规模训练数据**：从 100 万个教学视频自动收集，产生 60 万个图像-文本对序列。
2. **视频扩散模型**：基于输入场景和文本指令生成图像序列。

3. **场景一致性机制**：在生成的步骤中保持输入的物体存在和场景布局。

主要证据

在三个评估维度上实现最新性能：

- 步骤准确率：正确动作进展
- 场景准确率：一致的场景元素
- 任务准确率：整体任务成功完成

值得注意的是，ShowHowTo 在某些情况下比原始视频获得更好的人类偏好评分，表明生成的指令可以超过真实世界演示的质量。

相关性

直接解决了会议中的"如何展示"讨论。该方法证明了场景条件生成可以保持工具和原料一致性。该方法与必须从文本指令生成动作演示或教程的具身人工智能系统的视觉指令生成相关。

局限性

- 目前仅限于烹饪领域
- 生成的图像可能有伪影
- 需要高质量的输入场景图像

论文 11：StableWorld — 长交互视频生成

论文： "StableWorld: Towards Stable and Consistent Long Interactive Video Generation"（2026） **arXiv：** https://arxiv.org/html/2601.15281v1

问题与任务设置

交互视频生成存在误差累积问题，生成的帧逐渐偏离初始状态，将误差传播到后续帧。自回归方法导致与似真未来状态的显著偏离。这个问题与接地发现直接相关："当共同先验与动作序列矛盾时发生状态一致性错误"。

方法论

1. **动态帧驱逐**：监控帧之间的几何一致性，驱逐累积了误差的降级帧。
2. **一致帧保留**：保留几何一致的帧用于下一步生成。
3. **记忆机制**：通过记忆保持长期一致性。

主要证据

- 减少扩展生成中的身份漂移
- 改善时间一致性（更少的闪烁、抖动、漂移）
- 在更长序列上保持输出质量

相关性

解决了关于"当共同先验与动作序列矛盾时状态一致性错误（生与熟）"的担忧，这是会议接地中记录的内容。帧驱逐机制为保持一致性提供了具体方法，可以改编用于基于 VLM 的视频理解或指令生成系统。

局限性

- 帧驱逐可能丢失相关上下文
- 记忆机制增加复杂性
- 性能取决于几何一致性指标

论文 12：HD-EPIC 数据集

论文： "HD-EPIC: A Highly-Detailed Egocentric Video Dataset"（CVPR 2025） **URL：** https://hd-epic.github.io/site

数据集特征

- 来自不同家庭厨房的 41 小时无脚本多日录制
- 440 万帧、69 个食谱、59.4K 个细粒度动作
- 平均每分钟 263 个标注
- 通过厨房场景数字孪生的 3D 标注

三层标注：

1. 具有精确时间和权重的高细节原料标注
2. 具有动作-物体-手部-推理的详细叙述
3. 包括物体追踪和厨房固定装置的 3D 标注

VQA 基准： 七个类别共 26.6K 个问题。

相关性

该数据集提供了验证基准，证明 VLM 在物体运动问题上表现接近随机。它被设计为仅用于验证，不应用于微调，反映了数据集分离对可靠评估的重要性。

局限性

- 特定于厨房活动领域
- 扩展需要显著的标注工作
- 通过数字孪生创建 3D 标注成本高昂

4.2 综合主题评估

主题 1：为什么以自我为中心视频描述中的可区分性很重要，文献对解决方案的支持程度如何

为什么这个主题很重要

以自我为中心视频固有地包含重复动作——烹饪流程、工具使用、导航——这些只在细微的上下文细节上有所不同。没有判别性描述，基于文本的检索严重失败：Ego4D 中 66%的片段与至少另一个片段共享描述。对于必须理解和响应视频内容的具身人工智能系统，这种描述歧义直接影响识别特定时刻、引用特定动作或检索相关过去事件的能力。

文献说了什么

CDP（参见 4.1 节论文 1 了解完整方法论）提供了最具体经过验证的解决方案。该方法使用判别性提示库，其中包含从频繁 N-gram 中选择并设计为多样化的通用提示（如"握着"、"看着"、"另一个人"）。对于每个片段，方法找到最大化唯一性边际的提示或组合（最多 α=3 个提示）——相对于次优匹配的片段到描述匹配的改进。CDPNet 是一个轻量级变压器编码器，将推理从 300 秒（穷举搜索）减少到 5.8 秒。

定量结果很强：CDP 在 T=+30 秒时达到 76%的 R@1，而基线为 43%（33 个百分点的改进）。该方法在更好的基线模型上提供更大的改进，表明它将在描述能力提高时保持相关性。

这些发现之间的综合揭示了实现描述可区分性的三种机制：（1）通过提示选择的片段内判别，（2）片段内判别失败时的时域扩展，以及（3）基于时间锚点的事件定位以改进接地。TA-Prompting（4.1 节论文 2）提供了补充证据，表明时间锚点——直接对应于视频片段的学习性表示——改进了事件定位，相比基于语言先验的方法。这两种方法是互补的：CDP 针对说什么（判别性内容选择），而 TA-Prompting 针对何时说（精确时间定位）。

支持有多强

支持很强且直接经过验证。CDP 的改进在多个基准（Ego4D、Timeloop Movies）和多个基线模型（LaViLa VCLM、Video-LLaMA）上得到证明。该方法是描述器无关的，可以集成到 VLM 流程中。TA-Prompting 的 ActivityNet-Caption 结果进一步支持时间接地改进转化为更好的描述保真度。

什么不能干净地转移

CDP 需要访问具有相同描述的所有片段进行比较——这假设了片段语料库而非单查询场景。固定提示库可能不适用于所有视频类型，特别是词汇在训练 N-gram 中代表性不佳的视频。TA-Prompting 的幻觉描述风险取决于 LLM 主干质量。

实际意义

对于具身人工智能系统，将判别性提示集成到视频理解流程中将显著改善检索和时刻识别。当片段内判别失败时，时间扩展策略提供了后备方案，尽管牺牲了精确度换取唯一性。结合 CDP 的提示选择机制与 TA-Prompting 的时间锚定，可以为以自我为中心视频理解提供内容级和时间级的可区分性。

主题 2：3D 场景理解作为以自我为中心视频核心能力的证据

为什么这个主题很重要

基础笔记识别出"目前尚无明确方法将 3D 信息摄入 VLM"仍然是一个开放挑战。然而 3D 场景理解——保持对当前不可见物体的意识、推理空间关系、理解可及性和遮挡——对具身智能是根本性的。无法追踪相机视野外物体的系统无法有效地预测、规划或推理物理操作。

文献说了什么

OSNOM 提供了 3D 物体追踪的最经验证流程。Lift-Match-Keep 方法在物体离开视野 120 秒后达到 57%的正确定位，而其他 3D 方法为 33%，2D 追踪为 17%。关键洞察是交互上下文——靠近手的物体很可能正在被操作——显著改善匹配准确度，因为被操作的物体保持时空连续性。

Ego3DT 提供了使用开放词汇检测（GLEE）和 3D 重建（DUSt3R）的互补零样本方法。该方法比先前方法实现 1.04 倍到 2.90 倍的 HOTA 改进，无需任务特定训练，使其更易于部署。

几何引导相机运动论文提供了证据，表明运动相关信息在 VLM 预训练期间被捕获不足。探测实验揭示 VideoLLM 难以区分物体运动和相机运动，这是从以自我为中心视频进行准确 3D 场景理解的先决条件。

支持有多强

对 3D 追踪本身的支持很强——OSNOM 和 Ego3DT 都展示了具有强定量结果的经过验证方法。然而，对将 3D 理解集成到 VLM 中的支持很弱。几何引导论文识别了差距但没有展示端到端集成。会议笔记观察"目前尚无明确方法将 3D 信息摄入 VLM"仍然是准确的。

什么不能干净地转移

OSNOM 需要 SLAM/ARIA 传感器数据，这些数据可能并非在所有以自我为中心数据集中都可用。Ego3DT 依赖可能难以处理无纹理或反射表面的预训练深度估计。两种方法都未在可通过自然语言查询 3D 场景状态的集成 VLM 流程中得到验证。

实际意义

对于具身人工智能，3D 追踪提供了具体能力（物体定位、遮挡推理、可及性查询），但与 VLM 的集成仍然是一个开放挑战。系统可能需要维护独立的 3D 追踪模块和 VLM 推理模块，而非统一架构。

主题 3：扩散模型作为手部轨迹预测的主导方法

为什么这个主题很重要

以自我为中心视频中的手部姿态预测需要处理随机未来运动、考虑相机自我运动，以及预测手部不可见时的位置。基础笔记特别强调"不可见的自我手"挑战——预测超出相机视野的手部位置——作为预测系统的重要能力。

文献说了什么

三篇论文提供了手部轨迹预测的经验证基于扩散的方法：

EgoH4 使用基于扩散的预测与身体姿态约束，以实现在手部离开画面时的预测。该方法联合去噪手部和身体关节，身体关节作为手部运动的运动学约束。可见性预测器估计手部可见性，提高处理不可见手部的能力。结果显示在手部轨迹 ADE 上+3.4 厘米的改进，在手部姿态 MPJPE 上+5.1 厘米。

MADiff 引入运动驱动选择性扫描（MDSS），将相机自我运动整合到去噪过程中。该方法将手部运动与相机运动分离，具有时间因果性，在五个公共数据集上实现最新性能，具有实时推理能力。

MMTwin 使用双胞胎潜扩散模型，具有用于相机自我运动预测和手部轨迹预测的单独模型，通过混合 Mamba-变压器模块协调。

支持有多强

强且经过充分验证。三种方法都展示了相对于基线的定量改进。EgoH4 特别解决了视野外预测挑战，MADiff 展示了实时能力，MMTwin 提供了关于分离相机和手部运动预测的架构见解。

什么不能干净地转移

EgoH4 需要准确的身体姿态估计作为输入，在手动手部关节而非工具交互上进行评估。MADiff 需要相机姿态估计。MMTwin 需要多种输入模态并增加协调复杂性。三种方法都未在集成具身人工智能系统中得到验证，以使用手部预测进行操作规划。

实际意义

基于扩散的方法是手部轨迹预测的明显赢家。对于具身人工智能，集成这些方法将实现预测性操作规划——预测手部接下来将做什么以准备工具、清理工作空间或避免碰撞。核心能力差距是三种方法都在手动手部关节上得到验证；工具持有手部（餐具、器具、物体）的性能尚未确定。这很重要，因为具身人工智能操作主要涉及工具使用，而非裸手移动。

主题 4：VLM 在物理理解上的失败是根本性的，而非仅仅是性能差距

为什么这个主题很重要

基础笔记记录了 VLM 在 HD Epic VQA 基准的物体运动问题上表现接近随机。更关键的是，笔记观察到"更强的 VLM 提高似然性，但实际上可能掩盖接地失败"。这种区别非常重要：接近随机的模型揭示其失败模式，而生成自信但错误答案的模型掩盖它们。

文献说了什么

MVP 基准提供了最系统的证据。使用最小变化对（具有相同问题但相反答案的视觉相似视频对），基准迫使模型展示真正的物理理解，而非利用捷径。结果显示人类表现（92.9%）与最佳开源 Video-LLM（40.2%）之间有 52.7 个百分点的差距。

关键的是，在标准基准上表现良好的模型在 MVP 上表现不佳，确认标准基准允许捷径利用。MVP 发现验证了基础笔记的担忧：VLM 在标准基准上可能看起来有能力，同时在真正的物理理解方面仍然根本受限。

几何引导相机运动论文提供了机械解释：探测实验揭示运动相关信息在预训练期间被捕获不足。这不是参数数量或架构问题——这是关于模型学习表示什么的根本性差距。

HD-EPIC VQA 结果确认了这一模式：即使像 Gemini Pro 这样的强大模型在 26.6K 问题基准上也仅达到 37.6%的准确率，在细粒度动作理解、3D 感知和物体运动问题上表现尤其差。

支持有多强

非常强且控制良好。MVP 基准设计（最小变化对）专门控制捷径利用。发现在多个基准和模型族中一致。来自几何引导探测的机械解释为为什么差距存在提供了理论支持。

什么不能干净地转移

MVP 基准限于多选题格式，可能无法覆盖所有物理推理类型。几何引导方法在训练期间需要额外监督，这对现有 VLM 可能不可行。防止 VLM 中捷径的解决方案尚未得到验证。

实际意义

当前 VLM 不能作为需要物理理解的具身人工智能系统的可靠基础。这不是可以通过更大模型或更多训练关闭的临时性能差距，而是 VLM 表示运动、空间和因果关系方式的根本性局限。系统可能需要维护独立的物理推理模块，而非依赖 VLM 进行所有理解。

主题 5：场景条件生成作为判别性理解的补充

为什么这个主题很重要

基础笔记讨论了两个互补方向：判别性理解（描述正在发生的事情）和生成性指令（展示如何做某事）。ShowHowTo 解决生成方向，从输入图像和文本指令生成逐步视觉指令。基础笔记记录了这种方法在某些情况下比原始视频获得更好的人类偏好评分——一个引人注目的发现，表明生成的指令可以超过真实世界演示的质量。

文献说了什么

ShowHowTo 使用从 100 万个教学视频自动收集的训练数据，产生 60 万个图像-文本对序列。视频扩散模型基于输入场景和文本指令生成图像序列。场景一致性机制在生成的步骤中保持输入的物体存在和场景布局。

StableWorld 为在长视频生成中保持一致性提供了互补证据。动态帧驱逐机制——监控几何一致性并驱逐降级帧——解决了导致生成帧偏离初始状态的误差累积问题。

支持有多强

中等。ShowHowTo 在三个评估维度上展示了最新性能，人类偏好发现引人注目。StableWorld 提供了改善时间一致性的定量证据。然而，两种方法都限于烹饪领域，ShowHowTo 的生成图像可能有伪影。

什么不能干净地转移

当前方法特定于烹饪，可能无法泛化到其他领域而无需显著重新训练。基础笔记识别出当共同先验与动作序列矛盾时的状态一致性错误——StableWorld 解决了但未完全解决问题。

实际意义

对于具身人工智能，场景条件生成可以实现任务演示、教程生成和动作可视化。然而，当前方法是领域受限的且易产生伪影，在可靠部署之前需要显著开发。

5. 当前项目的综合评估

研究调查揭示了以自我为中心视频理解现状的清晰图景：单个组件能力正在快速成熟，但集成到统一系统仍然是关键瓶颈。

最有根据的方向

判别性提示以获得独特描述是最成熟和经验证的方法。CDP 展示了在自我中心基准上 R@1 的 33 个百分点改进，是描述器无关的，并随基线模型质量扩展。对于在重复动作领域需要视频理解的项目，集成判别性提示将提供直接和实质性的好处。

3D 物体追踪在保持相机视野外的物体意识方面得到充分验证。OSNOM 在 120 秒后达到 57% 的定位，交互上下文洞察（靠近手的物体保持时空连续性）提供了改善匹配准确度的实用机制。Ego3DT 的零样本方法将此能力扩展到任意物体类别，无需任务特定训练。

基于扩散的手部轨迹预测已在多篇论文和数据集上取得强结果。EgoH4 的身体姿态约束实现视野外预测，MADiff 的运动解缠处理相机自我运动，MMTwin 的双胞胎模型架构提供了模块化方法。对于需要预测性操作理解的系统，这些方法已准备好集成。

支持较弱的假设

将 3D 信息摄入 VLM 的假设支持较弱。几何引导论文识别了差距但未展示集成。会议笔记的观察"目前尚无明确方法将 3D 信息摄入 VLM"仍然准确。系统应为独立的 3D 追踪和 VLM 推理模块规划，而非统一架构。

VLM 将改进以处理物理理解的假设也支持较弱。MVP 基准证明，即使强大的 VLM 也利用捷径而非开发真正的物理推理。几何引导训练有前景，但需要可能对现有模型不可行的额外监督。

值得进一步测试的分支

判别性描述、3D 追踪和手部预测的端到端集成到统一流程中尚未得到验证。尽管单个组件已验证，但它们在集成系统中的交互效应、误差传播和延迟特征仍然未知。

具身人工智能任务的捷径抵抗评估值得系统测试。MVP 基准设计（最小变化对）可应用于以自我为中心视频基准，以暴露 VLM 是真正理解物理交互还是利用捷径。

6. 未解决的问题与决策关键差距

差距 1：3D 到 VLM 集成没有经过验证的方法

基础笔记识别出"目前尚无明确方法将 3D 信息摄入 VLM"仍然是一个开放挑战，文献调查确认这一差距仍然存在。几何引导相机运动论文提供了一条途径但未展示端到端集成。项目的一个决策关键问题是追求统一架构（具有几何引导训练的 VLM）还是模块化架构（独立 3D 追踪 + VLM 推理）。

差距 2：VLM 捷径掩盖了真正的能力和失败

MVP 基准证明，在标准基准上表现良好的模型在控制捷径时表现不佳。对于评估具身人工智能系统，这意味着标准 VQA 基准可能无法揭示模型真正失败的地方。项目需要评估方法论来暴露捷径利用，而非允许膨胀分数。

差距 3：手部预测验证限于手动手部关节

EgoH4 在手动手部关节而非工具交互上进行评估。对于必须理解操作的具身人工智能系统（使用工具、操作器具、处理物体），手部关节预测与工具交互预测之间的差距是显著的，需要验证。

差距 4：场景条件生成仅限于烹饪领域

ShowHowTo 和 StableWorld 在烹饪领域得到验证。将场景条件生成扩展到其他具身人工智能领域（车间、车库、户外）将需要数据集收集和模型重新训练。此扩展的范围尚未被表征。

差距 5：视频生成中的时间对应仍然具有挑战性

基础笔记识别出"继续改善视频生成中的时间对应"作为建议的下一步。StableWorld 通过帧驱逐解决了这个问题，但未完全解决问题。对于需要高保真生成视频（教程、演示、模拟）的应用，时间对应中的剩余差距需要系统表征。

差距 6：评估基准是领域特定的

HD-EPIC、EPIC-KITCHENS 和 Ego4D 都专注于厨房。其他具身人工智能领域（车间、车库、户外活动）缺乏可比的标注数据集。在厨房设置中验证的方法的跨领域泛化尚未建立。

7. 建议的下一步

行动 1：将判别性提示集成到视频理解流程中

CDP 为重复动作领域的唯一描述提供了立即适用的改进。该方法是描述器无关的，计算上可处理的（通过 CDPNet 近似为 5.8 秒），并在更好的基线模型上展示更大的改进。此行动很重

要，因为描述歧义直接影响检索、时刻识别和基于文本的视频理解。实施路径清晰：将提示库和 CDPNet 集成到现有描述流程中。

行动 2：部署 3D 物体追踪作为独立模块

OSNOM 和 Ego3DT 提供了 3D 物体追踪的经验证方法。鉴于缺乏 3D 到 VLM 集成方法，将这些部署为维护世界坐标物体内存并通过结构化接口暴露查询的独立模块。此方法以架构优雅换取即时能力。使用深度估计和 SLAM 在 3D 空间中追踪物体，使用交互上下文（手部接近度）保持物体身份，并通过独立推理模块暴露空间推理查询（可见性、遮挡、可及性），而非尝试 VLM 集成。

行动 3：在工具交互场景中验证手部预测

EgoH4、MADiff 和 MMTwin 在手部关节上展示了手部轨迹预测。在依赖这些方法进行具身人工智能操作理解之前，在工具交互场景（持有餐具、操作器具、操作物体）中验证其性能。此行动很重要，因为具身人工智能中的操作理解主要涉及工具使用，而非裸手移动。比较有和没有工具接触的基于扩散的预测准确度，并表征失败模式。

行动 4：为具身人工智能基准设计捷径抵抗评估

应用 MVP 最小变化对方法论创建控制捷径利用的具身人工智能特定基准。生成视觉相似但在关键物理交互细节上不同的视频对，并要求模型正确回答两者的问题。此行动很重要，因为标准基准可能给出 VLM 物理理解的膨胀印象。评估设计应遵循 MVP 方法论，但针对具身人工智能任务：物体操作、工具使用、空间推理。

行动 5：表征几何引导训练可行性

研究几何引导训练方法是否可以应用而无需完全重新训练到现有 VLM。几何引导相机运动论文识别出运动信息在预训练期间被捕获不足，但训练方法（额外的几何监督）可能作为微调而非完全重新训练适用。此行动很重要，因为它决定了是否可以用有针对性的训练改进现有 VLM，还是需要架构改变。

行动 6：研究 StableWorld 帧驱逐以保持一致性

StableWorld 的动态帧驱逐机制通过监控几何一致性并驱逐降级帧来解决视频生成中的误差累积。研究此机制是否可以改编用于场景条件指令生成，以解决基础笔记中记录的状态一致性错误（生与熟）。此行动很重要，因为一致性错误破坏了生成指令的可靠性。

8. 关键风险、注意事项与证据边界

风险 1：单个组件验证并不意味着集成系统性能

文献在隔离中展示了判别性描述、3D 追踪、手部预测和 VLM 评估的强结果。然而，没有文献展示结合这些组件的端到端系统。集成可能引入在组件级评估中不可见的误差传播、延迟问题和交互效应。这是证据转移风险：单个组件的证据强度不会转移到集成系统。

风险 2：VLM 物理理解失败是根本性的，无法通过规模解决

MVP 基准（52.7 个百分点差距）和 HD-EPIC 结果（Gemini Pro 37.6%）证明，VLM 在物理理解上的失败不是可以通过更大模型或更多训练关闭的性能差距。几何引导探测确认运动信息在预训练期间被捕获不足。这是方法论风险：当前 VLM 可能在物理推理方面存在架构限制，需要根本性改变，而非增量改进。

风险 3：3D 追踪质量取决于深度估计

OSNOM 达到 57%的定位准确度，但需要已知有限制性的深度估计。Ego3DT 依赖 DUSt3R 进行 3D 重建。在无纹理表面、反光物体或光照差的环境中，深度估计质量会下降。这是数据质量风险：3D 追踪性能受深度估计质量限制，这在不同环境中有所不同。

风险 4：工具交互的手部预测验证差距

EgoH4 在手部关节上得到验证，而非工具交互。对于主要涉及工具使用的具身人工智能系统，这代表了显著的证据边界。持握工具的手部预测准确度可能与裸手预测有实质性差异，由于运动约束、遮挡和物体-手部附着。这是评估风险：手部预测的证据不会直接转移到具身人工智能使用案例的工具交互理解。

风险 5：场景条件生成仅限于烹饪领域

ShowHowTo 和 StableWorld 在烹饪领域得到验证。扩展到其他领域需要新的数据集、重新训练和验证。基础笔记特别涉及以自我为中心视频理解，涵盖烹饪之外的许多领域。这是领域转移风险：场景条件生成的证据不会跨具身人工智能领域泛化。

风险 6：捷径抵抗评估可能揭示比标准基准建议的更差的 VLM 性能

MVP 基准证明，在标准基准上表现良好的模型在控制捷径时表现不佳。将此方法论应用于具身人工智能评估可能揭示 VLM 能力比标准基准表明的差得多。这是评估风险：基于标准基准性能部署 VLM 可能导致在捷径不可用的真实场景中失败的系统。

风险 7：HD-EPIC 仅用于验证，限制训练数据可用性

HD-EPIC 数据集被设计为仅用于验证，不应用于微调。此约束限制高质量 3D 标注的可用性用于训练。对于需要在以自我为中心视频理解上进行训练的项目，这代表数据质量风险：可能需要较低质量的标注进行训练，可能限制模型性能。

风险 8：密集物体追踪标注成本高昂

基础笔记识别出"密集物体追踪（逐帧边界框）标注成本高昂"。将 3D 物体追踪扩展到新领域需要大量标注投资。这是方法论风险：当前 3D 追踪证据来自大量标注的数据集，这些数据集对于新领域可能不可行复制。