

机器人通用视觉基础模型：理解、比较与应用

2026-10-19

目录

1. 从机器人任务理解视觉模型

2. 核心模型：从“能做什么”到“如何工作”

3. 该选谁？关键模型差异与选型路径

4. 四类典型机器人工作流

5. 实际使用：从模型输出到可靠系统

6. 常见问题与一句话总结

官方参考资料

2

2

5

5

6

7

7

[课程首页](#) · [English](#) · [PDF](#)

项目	说明
主题	视觉基础模型
修订日期	2026-10-19
许可	CC BY-NC-SA 4.0 · Xiong Zeng and contributors

1

1. 从机器人任务理解视觉模型

课程边界：只讨论模型原理、输入输出、使用流程、组合策略和适用边界。不展开相机标定、复杂滤波推导、网络训练，不设置代码或实验章节。

阅读主线：语言或视觉提示 → 找到并分割目标 → 获取距离或形状 → （任务需要时）估计刚体 6D 位姿 → 机器人策略或控制器。这些能力往往来自不同模型；并不是每条机器流水线都必须经过全部步骤。

1.1 五种常见视觉输出

你需要回答的问题	视觉任务	典型输出	不能据此断言
“红色螺丝刀在哪儿？”	检测 / 开放词汇定位	边界框、类别、候选实例	精确轮廓与距离
“哪些像素属于它？”	实例分割 / 视频跟踪	Mask、跨帧实例身份	物体三维朝向
“它离相机多远？”	深度估计 / 立体匹配	深度图或视差图	隐藏表面的完整形状
“它大致长什么样？”	三维重建	三维网格、纹理、空间布局	工程级 CAD 精度与质量惯量
“它的位置和朝向是什么？”	6D 位姿估计	三维平移 + 三维旋转	接触力、是否抓牢、是否滑移

重点：Mask ≠ 深度；深度 ≠ 完整物体模型；形状 ≠ 物体当前位姿；位姿 ≠ 接触状态。相机提供可见表面信息，指尖力、遮挡后的接触和材料参数通常还需要其他传感器或动力学推断。

1.2 先问四个问题，再选择模型

1. **任务需要多精确的几何？** 只想知道“哪个目标”用检测/Mask 即可；抓取朝向、插装对准通常需要更明确的几何状态。
2. **有哪些传感器？** RGB、可靠的 RGB-D、同步标定双目，所适用的模型完全不同。
3. **有哪些物体先验？** 有没有 CAD、准确尺寸、参考图像，决定 6D 位姿方法如何初始化。
4. **何时需要输出？** 离线几何恢复可以较慢；闭环控制更关心观测新鲜度、吞吐量、延迟和失败恢复。

2. 核心模型：从“能做什么”到“如何工作”

2.1 SAM 2：用提示确定目标，用视频记忆传播 Mask

一句话理解：SAM 2 解决“我已指出这个目标，请继续告诉我它在视频中占据哪些像素”。它是可提示的图像/视频分割器，不是深度模型或 6D 位姿估计器。[官方资料](#)

关键思想。第一帧用点击点、框或已有 Mask 指定对象，模型利用当前图像特征生成分割；处理后续视频帧时，通过**流式记忆（streaming memory）**读取此前目标外观及分割信息，使目标在运动、形变和部分遮挡时仍能被传播。用户也能在某一帧追加提示以修正 Mask；记忆并不意味着永久正确的实例身份。

输入与输出。输入是 RGB 图像/视频以及正负点、边界框或掩码等提示；输出是每帧实例 Mask，以及用于管理目标的对象 ID/视频推理状态。Mask 是像素集合，不能直接给出相机到物体的米制距离。

典型使用流程。① 在起始帧指定螺丝刀；② 传播后续帧的 Mask；③ 当发生遮挡、相似工具交叉或对象出画时检查身份；④ 发现漂移后重新点击、给框或切换到全局重新检测。

适用与局限。已知目标、持续分割、制作视频伪标签、给位姿估计器提供对象区域时很合适；但当用户只有“找出红柄工具”这样的文本任务时，需要另外的语言定位模块。手和物体互相遮挡、视觉相似实例交换、运动模糊会导致 Mask 漂移；SAM 2.1 是更新的权重/实现系列，不是增加了度量深度能力。

2.2 SAM 3 / SAM 3.1：按概念发现目标，再分割和跟踪

一句话理解：SAM 3 不要求先由人指出唯一实例，而是能够根据**短文本概念或视觉示例**找出图像/视频中符合条件的实例，同时输出 Mask 并持续跟踪。[官方论文](#)·[官方仓库](#)

关键思想。SAM 3 把“概念是否存在及在哪里”的**检测器**和“已找到实例如何在视频里延续”的**记忆式跟踪器**组织在同一视觉特征体系中。概念提示可以是“red-handled screwdriver”等短名词短语，也可以是图像中的示例对象；系统返回**所有符合概念的候选实例**，而非默认选出机器人唯一要抓的那个。出现多个目标时，仍需任务级选择和 ID 管理。

与 SAM 2 的实质区别。SAM 2 更适合“实例已指定，维持像素跟踪”；SAM 3 增加概念驱动的**发现、分割与跟踪**。SAM 3.1 进一步用**对象复用（object multiplexing）**改进多个对象同时跟踪的计算效率；它不是独立的三维模型，也不把 Mask 直接升级成 6D 位姿。[SAM 3.1 官方说明](#)

输入与输出。输入为图像/视频 + 文本概念、视觉示例或交互提示；输出包括候选检测、实例 Mask、视频中的实例身份。复杂长指令可能要先经 VLM/Agent 化简为可定位的视觉概念，再用空间关系、任务记忆或机器人当前状态筛选目标。

典型使用流程。① 将“把桌上的红柄螺丝刀拿起来”转换为“red-handled screwdriver”提示；② 获取所有匹配 Mask；③ 根据桌面位置和任务约束选定实例；④ 开始视频跟踪；⑤ 身份不可靠时重新用概念检索。

适用与局限。适合语言条件抓取、开放类别物体发现、语义驱动目标选择。短语歧义、细小目标、相似物体、遮挡仍会出错；单个 Mask 不携带 CAD、深度、精确物体坐标系或接触状态。论文中的视频 FPS 依赖硬件、分辨率与对象数量，不能直接当成机器人闭环频率。

2.3 FoundationPose：刚体物体的 6D 位姿初始化、细化与跟踪

一句话理解：SAM 告诉我们“哪些像素属于螺丝刀”，FoundationPose 则回答“这把已知或有参考视图的刚体螺丝刀相对于相机的位置与朝向”。[论文与代码](#)

什么是 6D 位姿？三维位置加三维旋转（通常记为 $SE(3)$ 变换）相对于一个已定义尺度和坐标轴的物体模型才有意义。对长工具而言，刀尖、握柄和模型原点不同；“物体中心误差不大”未必意味着刀尖对得准。

两种使用条件，不等于两个无条件通用模型。

模式	需要准备什么	应如何理解
Model-based	已知 CAD/网格、实际尺度、物体坐标约定	根据已知形状解释当前视觉观测；通常更适合重复使用的工具。
Model-free	对该物体采集的少量参考图像，并建立对应的对象表征	不要求现成 CAD，但仍需对象参考信息；公开流程可包含预先构建神经物体表示，不能理解为单张陌生图像就能直接得到精确 6D 位姿。

关键思想。首先根据可见对象区域和物体先验产生若干可能的姿态；**scorer** 判断候选的图像/几何一致性，**refiner** 对较好的候选进一步细化；后续帧把上一时刻的估计作为局部起点进行 **tracking**。这比每帧从全空间重新搜索更经济，但局部跟踪也容易把一次错误持续传播。论文通过共享后端，使 CAD 模式和少量参考图像模式能够使用相似的估计框架。

输入与输出。典型流程需要 RGB-D 观测、相机内参、目标 Mask，以及 CAD 或参考视图准备得到的物体表示；输出是在明确对象坐标系下的刚体 6D 位姿，连续帧可更新。若下游要在机器人基座坐标系控制，还需使用相机标定结果与采集时刻的机械臂状态完成坐标转换；这里不重复标定推导。

典型使用流程。① 确认 CAD/参考视图和真实物体一致；② 通过 SAM 等获得 Mask 与有效 RGB-D；③ **初始化**多个姿态候选；④ **评分、细化**得到当前位姿；⑤ 使用相邻帧做跟踪；⑥ 丢失或发生大姿态跳变时**重新初始化**，而不是无限期使用旧位姿。

适用与局限。适合刚性工具抓取、工具定向、装配中的视觉对齐。长时间遮挡、强反光/透明材质、细杆物体、深度缺失、CAD 与实物不符都会增加失效风险；**几何对称**可能使某些旋转角从视觉上不可区分。FoundationPose 估计的是物体位姿，不直接判断手指接触力或是否稳定抓持。对于严重非刚体，不应强行假设只有一个 6D 位姿。

2.4 FoundationStereo：用真实双目对应关系恢复深度

一句话理解：FoundationStereo 解决“左右相机看见的同一位置如何对应”，先输出**稠密视差图**，再利用标定过的双目几何恢复以米为单位的深度。[官方资料](#)

关键思想。对整流后的左右图像，模型尝试判断左图某个像素在右图中的对应位置。两者的水平位移叫视差：一般同等相机配置下，近物体视差较大、远物体视差较小。深度与视差的关系可概括为 $Z = f \times B / d$ （焦距 f 、基线 B 、视差 d ）。模型通过图像特征与上下文匹配，力图让立体对应关系跨不同场景泛化；真正的米制尺度仍然来自**焦距和基线的标定**。

输入与输出。需要一对**同步、标定并完成立体校正**的左右 RGB 图像；主要输出视差，随后可转换为深度和点云。它不进行语言分割，不识别某一把螺丝刀的规范坐标系，也不直接求物体的 6D 位姿。

典型使用流程。① 确认确实是固定相对几何的双目；② 校正左右图像与有效视场；③ 推理视差；④ 根据双目参数转换深度；⑤ 与 Mask 联用提取目标可见表面；⑥ 如任务确实需要物体朝向，再接位姿估计器。

Fast-FoundationStereo 有什么变化？它强调在保持较强跨域能力的同时降低立体匹配计算量，公开方法涉及知识蒸馏、按计算预算搜索网络设计和结构化剪枝；核心输出仍是**双目视差**，不是一个新类别的视觉任务。[官方资料](#)

适用与局限。已有硬件双目但原生深度不可靠时非常有用；如果 RGB-D 相机已经稳定提供精度足够的深度，不必为了“使用基础模型”重复运行立体网络。低纹理、镜面、透明、重复花纹、远距离小视差和双目失配会损害结果。头部 RGB-D 与随手臂运动的腕部 RGB **不能未经同步、外参建模与几何处理直接当成标准双目对**；论文报告的速度也不能替代本机端到端测量。

2.5 SAM 3D Objects：从单张图像生成物体形状先验

一句话理解：SAM 3D Objects 侧重“这个目标可能具有怎样的**完整三维形状、外观和布局**”，并非直接返回经过标定、可用于精密装配的 CAD。[官方资料](#)

关键思想。单张图片只显示物体部分表面，其背面几何通常不可见。SAM 3D Objects 利用从大量数据学到的形状与外观先验，为被选中的目标生成合理的完整三维几何、纹理及空间布局。生成能力能够在遮挡和杂乱场景下补全形状，但这种“看起来合理”不代表所有不可见细节都是**真实测量**。

输入与输出。输入一般是 RGB 图片和目标选择/区域提示；输出为候选物体三维形状、纹理及相关场景信息。需要区分：**视觉网格、实际米制尺寸、物体坐标原点、碰撞几何、质量/惯量**是五类不同的物理要求，单图生成并不自动满足。

典型使用流程。① 选择图片中的目标；② 得到候选网格和纹理；③ 如果用于仿真或位姿估计，先检查尺度、法向、拓扑、对称性及对象坐标；④ 用 RGB-D、多视图或已知尺寸核验关键接触区域；⑤ 只把通过检查的网格作为下游几何先验。

与 FoundationPose 的关系。SAM 3D Objects 生成“物体可能长什么样”，FoundationPose 用物体先验估计“它当前在哪里”。两者理论上可串联，但生成网格必须先验证尺度、坐标约定和几何质量；不能把它们描述成官方保证的直接、无校验即插即用接口。陌生物体的粗形状建模比毫米级插装更适合它。

2.6 Depth Anything 3：单目和多视图的通用几何先验

一句话理解：Depth Anything 3 (DA3) 把单幅/多幅图像的**深度与空间几何估计**统一到一个模型家族；但“有深度图”与“有可信的米制深度”依然要区分。[官方资料](#)

关键思想。模型从单目或多视角图像学习场景几何，既可预测深度，也可在一些配置下估计相机相关参数或一致的多视图表示。**不同 checkpoint 承担不同任务：**Main 系列支持多种视图配置；DA3Mono 面向相对单目深度；DA3Metric 面向度量单目深度。不能把某个模型的宣传能力不加区分地套到所有权重上。

怎么用。单目 RGB 场景需要大致前后关系时，可用它补充几何先验；若已有多视角图像，可关注几何一致性。对于厘米/毫米级机器人操作，必须检查尺度、视角变化和跨帧一致性：即使选择 Metric 版本，也不保证优于经过校准且质量良好的硬件深度或真实双目。

2.7 SAM 3D Body：估计人体三维形状，不是估计接触力

一句话理解：SAM 3D Body 从人体图像估计**人体三维网格与姿态**，常用于人类动作理解或机器人模仿学习中的人体侧几何先验。[官方资料](#)

使用方式与限制。输入为包含人体的图像，推理得到人体形状/姿态；如需由演示视频生成可用的机器人参考轨迹，还要解决**跨帧一致性、相机运动、尺度、人体到机器人运动学重定向、手—物接触**等问题。重建的人体网格不等于机器人关节轨迹，更不能直接推出抓取压力、摩擦力或物体被手挡住时的真实姿态。

2.8 Grounding DINO：可组合的文字目标检测器

一句话理解：Grounding DINO 根据开放词汇文本提示提供目标候选**边界框**，它适合“先文本定位，再交给独立分割器”的可替换流水线。[官方资料](#)

如何组合。“red screwdriver” → Grounding DINO 框 → SAM 2 分割/视频传播。与 SAM 3 的概念分割统一接口相比，这条

路线的组件可分别替换，但需要管理框与 Mask 的对应、多个相似目标的筛选，以及额外推理延迟。只需要框就停在框，不必为了模型齐全而继续做三维重建。

3. 该选谁？关键模型差异与选型路径

3.1 最容易选错的五组模型

容易混淆的模型	应该如何区分
SAM 2 vs SAM 3	已指定实例、重点是 Mask 传播 → SAM 2；用语言/示例发现新目标并跟踪 → SAM 3。
SAM 3 vs Grounding DINO	希望概念检测、Mask 与视频跟踪统一 → SAM 3；只需文字找框或保留模块化组合 → Grounding DINO。
FoundationStereo vs DA3	有同步标定的左右相机、需要基于视差的几何深度 → FoundationStereo；只有单目/任意视角图片、需要学习型几何先验 → 选 DA3 的合适变体。
FoundationPose vs SAM 3D Objects	已有 CAD/参考信息，想知道当前物体 6D 位姿 → FoundationPose；缺少物体形状，先生成候选几何 → SAM 3D Objects。
SAM 3D Body vs Objects	人体网格/姿态 → Body；一般物体三维形状与外观 → Objects；二者都不是接触传感器。

3.2 一个按问题依次做决定的选型树

1. 仅需目标语义和位置？有语言提示选 SAM 3（或 Grounding DINO），手工点选目标选 SAM 2；不必引入深度。
2. 需要持续获得目标 Mask？由 SAM 2 或 SAM 3 视频跟踪维持 ID；准备遮挡失效后的重检测。
3. 需要目标距离？先用已有可靠 RGB-D；有标定双目用 FoundationStereo / Fast 版；只有单目可考虑 DA3 对应 checkpoint，但要审查尺度。
4. 需要完整形状？已有 CAD 就优先使用 CAD；陌生物体可考虑 SAM 3D Objects，并核验物理尺寸和关键表面。
5. 确实需要刚体姿态？以 RGB-D + Mask + CAD/参考视图接 FoundationPose；只有 Mask 或只有生成网格，并不自动满足其使用前提。
6. 要做接触丰富的操作？还需要本体状态、执行反馈，以及条件允许时的力觉/触觉；视觉只能提供其中一部分信息。

最小充分模型原则：抓取方向需要 6D 位姿，不等于必须运行 SAM 3D；已有可靠 RGB-D，不等于还要运行 FoundationStereo；VLA 能直接消费图像，也不等于显式 FoundationPose 永远没有价值。模型应由任务输出和输入条件驱动。

4. 四类典型机器人工作流

4.1 工具抓取与定向：需要已知目标的 6D 位姿

任务：根据语言指令拿起螺丝刀，并把刀尖对准指定方向。

推荐链路：任务语言 → SAM 3 发现螺丝刀并选择实例 → Mask + 头部 RGB-D + 对应的 CAD/参考视图 → FoundationPose 初始化 6D 位姿 → 连续跟踪与几何反馈 → 控制器执行。若目标和初始 Mask 已知，可跳过 SAM 3；如 FoundationPose 跟踪可靠，也未必需要每帧都运行完整分割。

为什么这样搭配？SAM 3 管“是哪一个”；深度管“可见表面在哪里”；FoundationPose 管“该刚体朝哪儿”；机器人控制器负责动作。视觉仍不能替代刀尖接触检测，物体滑移、跌落时要重新确认目标并定位。

4.2 双目机器人：没有 RGB-D 但想获得距离

任务：用两只已标定、同步的相机获取工作台的稠密深度。

推荐链路：左右图像校正 → FoundationStereo（实时性紧张时评估 Fast-FoundationStereo）→ 视差转换深度 → 根据目标 Mask 提取可见表面。只有在确实需要物体规范坐标系的 6D 朝向时，再加入具备对象先验的位姿估计。

判断要点：双目来自真实相机几何，单目深度来自学习先验；两者都可能失效，但尺度依据不一样。腕相机和头相机随时间变化的相对位置，不能直接代替标准固定双目输入。

4.3 未知物体：先获得候选形状，再决定是否能用于控制

任务：机器人面对一件没有 CAD 的陌生工具。

推荐链路：SAM 3 / SAM 2 选择目标 → SAM 3D Objects 得到候选形状 → 利用已知尺寸、RGB-D 或多视图检查尺度和几何 → 根据任务选择直接基于 Mask/点云抓取，或建立合适的参考模型供位姿估计使用。

判断要点：单视图看不到的槽、底面、接触边缘可能被生成模型“猜错”；若要精密插装、受力配合或接触丰富的旋转，必须进一步验证关键几何。若只要粗略避障、可视化或对象级检索，不必先强求高精度 CAD。

4.4 人类演示与 VLA：不一定需要逐帧显式 6D 位姿

任务：从人类视频理解动作，或用视觉语言动作策略直接执行操作。

推荐链路：人体动作理解可用 SAM 3D Body 提供人体网格/姿态；若要重定向到机器人，额外需要轨迹连续性、人体—机器人几何对应及接触信息。VLA 可直接以相机图像、任务语言和机器人本体状态生成动作，不要求所有场景都先跑 FoundationPose。

显式几何何时仍有用？对需要刀尖朝向、碰撞约束、可解释失败检查、物体掉落后重定位的任务，SAM/深度/6D 位姿可作为独立观测或安全检查接口。它与直接视觉策略是**互补路线**，不是非此即彼。

5. 实际使用：从模型输出到可靠系统

5.1 六个必须确认的接口条件

方面	必须确认的问题	常见误区
输入	是 RGB、RGB-D、双目、文本、CAD，还是参考视图？	把单张 RGB 直接当作 FoundationPose 完整输入。
尺度	深度是相对尺度还是可靠米制深度？物体网格是否已校准？	把生成网格尺寸用于毫米级装配。
坐标	Mask、深度、相机和物体坐标系是否一致？	把相机系 6D 位姿直接当机器人基座位姿。
时间	图像拍摄时刻和模型发布时刻是否区分？	用新时刻的机械臂姿态解释旧图像。
身份	目标是否仍是同一个实例？丢失后如何重定位？	Mask 看似连续，实际跟到了另一件物体。
资源	显存、分辨率、推理频率、权重许可是否匹配？	把论文某 GPU 的单模型 FPS 当作真机控制频率。

5.2 失败恢复比漂亮的单帧结果更重要

遮挡：允许短时跟踪，但不能把长期不可见仍标成“精确已观测”。**对称性：**某些旋转本来就无法由当前图像唯一确定。**细长、透明与反光物体：**深度和 Mask 可能同时受影响。**突然移动或掉落：**局部位姿跟踪可能跑偏，需重新检测、验证实例、重新初始化。**精密接触：**即使 6D 位姿准确，仍无法单凭视觉推定力、摩擦和微滑移。

机器人系统应把“是否看见目标”“几何是否可信”“观测是否新鲜”“目标 ID 是否连续”作为独立判断。必要时降低动作幅度、主动换视角或停止接触；这属于系统决策，而不是某个视觉模型会自动完成的事情。

5.3 何时运行大型模型？

事件触发的语义检索（任务开始、目标丢失）可以较低频；连续 Mask 或位姿跟踪可在相机更新时运行；控制器通常依赖更及时的本体观测和估计状态。吞吐量、单次模型延迟和“动作真正使用时信息有多旧”是不同指标。不要默认 SAM 3、FoundationStereo、SAM 3D Objects 和 FoundationPose 都需要每帧串行推理。

6. 常见问题与一句话总结

Q1：SAM 3 能替代 FoundationPose 吗？不能。前者主要给语义候选与 Mask，后者估计有对象先验的刚体 6D 位姿。

Q2：FoundationPose 的 model-free 是“完全零先验”吗？不是。它不要求现成 CAD，但需要参考视图等对象信息，具体公开工作流还有对象表示构建步骤。

Q3：有可靠 RGB-D，还要 FoundationStereo 吗？通常不需要；只有深度质量、视角和硬件条件确实受益时才增加双目路线。

Q4：SAM 3D Objects 输出的网格能直接给 FoundationPose 用吗？不能无条件直接使用。先验证尺度、坐标、几何和任务所需精度，再考虑作为对象先验。

Q5：只有单目 RGB 能获得精确的绝对深度吗？学习型单目模型能提供深度先验，特定 metric 权重有尺度能力；但不应未经校验就把它当成机器人精密操作的可信量测。

Q6：精确 6D 位姿就等于抓取成功吗？不等于。位姿是几何信息，接触力、抓持稳定性和滑移需要额外反馈。

最终记忆法：SAM 找目标、分割与追踪；FoundationStereo 用双眼看距离；DA3 从单目/多视图估计场景几何；SAM 3D 生成三维形状；FoundationPose 估计当前刚体姿态。选最小必要模块，再根据任务补充传感器、跟踪和失败恢复。

官方参考资料

1. SAM 2 — Segment Anything in Images and Videos (Meta, 官方仓库)；论文。
2. SAM 3 — Segment Anything with Concepts (Meta, 官方论文)；官方仓库与 SAM 3.1 更新；SAM 3.1 说明。
3. FoundationPose (NVIDIA, CVPR 2024, 官方仓库)；论文。
4. FoundationStereo (NVIDIA, CVPR 2025, 官方仓库)。
5. Fast-FoundationStereo (NVIDIA, CVPR 2026, 官方仓库)。
6. SAM 3D Objects (Meta, 官方仓库)；SAM 3D Body (Meta, 官方仓库)。
7. Depth Anything 3 (ByteDance Seed, 官方仓库与模型家族)。
8. Grounding DINO (IDEA-Research, 官方仓库)。

课程来源与署名：改写自《Foundation of Embodied AI》课程原有《视觉感知与状态估计》讲义（CC BY 4.0）。第三方模型、代码、权重及资料遵循各自许可证。2026 年 10 月修订；具体模型接口、许可与性能请以使用版本的官方资料为准。