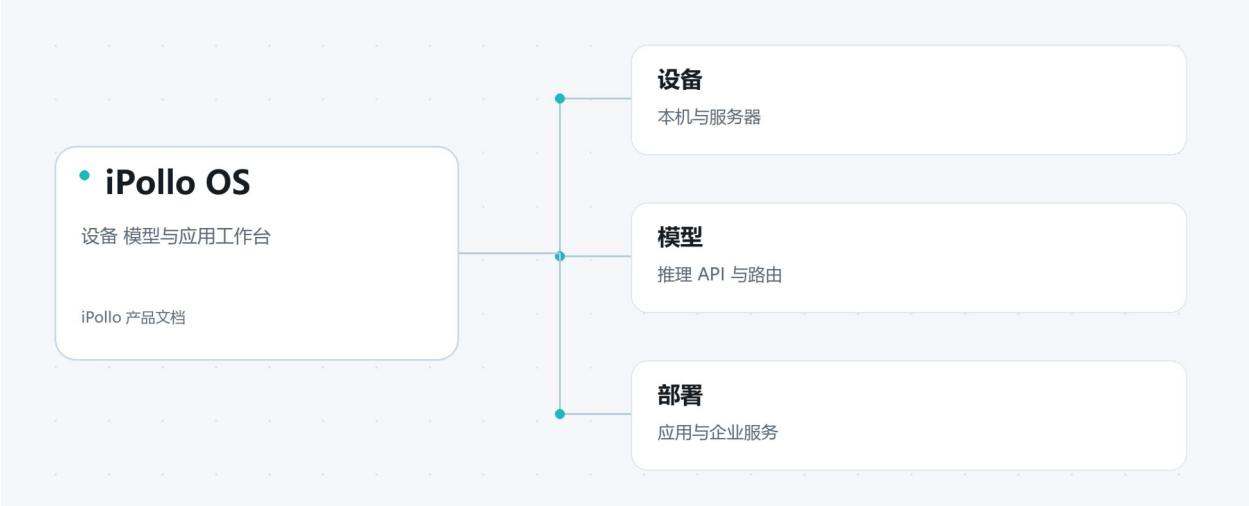
iPollo

iPollo 产品文档 / 05

iPolloOS
功能使用手册

使用指南与操作参考



iPollo

文档编号 IP-OS

2026 年 11 月 2 日

摘 要

本手册面向 iPolloOS 的设备所有者、模型管理员以及部署与运维人员，说明设备、模型、统一 API、服务器和应用服务的使用方法，帮助读者区分本机、远程设备、模型入口与企业管理服务，并按目标环境逐项完成配置。

全书按本机设备与模型 API、服务器路由与跨机功能、应用部署与持续运维、企业后台与权限总控四个主题组织各项功能。各项功能配有实际入口截图、准备条件、操作步骤和完成标志，附录进一步说明模型规格、应用部署、网络接口与当前使用边界。

使用前应核对运行位置、硬件条件、网络可达性和管理权限。模型下载、服务器连接、应用发布及恢复均需要在目标设备验证；截图中展示的配置表单和安装前状态用于帮助定位功能，不代表相应模型推理或远程部署已经完成。

关键词：iPolloOS；模型管理；统一 API；服务器；应用部署

适用读者　设备所有者、模型管理员、部署与运维人员

版本说明　操作界面采集于 2026 年 11 月 2 日。功能说明结合原项目指南与本地项目资料整理；实际能力以目标版本、权限和部署环境为准。

图示说明　截图聚焦操作入口与配置位置，图下注明关注点。模型调用、远程部署和外部服务的最终结果，需要在实际环境确认。

目 录

第 1 章 本机设备模型与统一 API 1
1.1 查看设备与选择运行位置 1
1.2 下载启用停止与模型配置 2
1.3 图片和视频模型任务 3
1.4 创建 API Key 与网络入口 4
1.5 向量媒体接口与计量 5
第 2 章 服务器路由与跨机功能 6
2.1 通过 SSH 接入服务器 6
2.2 模型路由与多个完整副本 7
2.3 CPU 双机与 GPU 跨机部署 8
2.4 网络诊断与服务器可达性 9
第 3 章 应用部署与持续运维 10
3.1 安装应用模板或公共镜像 10
3.2 从本地源码构建发布 11
3.3 环境变量健康检查与伸缩 12
3.4 日志预览启停与代码回退 13
第 4 章 企业后台与权限总控 14
4.1 部署与使用企业后台 14
4.2 部署总控并分配管理职责 15
4.3 成员模型授权与资源额度 16
4.4 持久服务备份与恢复 17
附录 A 产品能力与使用范围 18
A.1 设备检测与模型选择 18
A.2 本机模型生命周期 18
A.3 图片与视频模型 18
A.4 统一模型 API 与密钥 18
A.5 路由和完整模型副本 18
A.6 远程设备与跨机模型 18
A.7 iPolloOS 应用部署与运行边界 19
资料依据与使用边界 20

第 1 章 本机设备模型与统一 API

1.1 查看设备与选择运行位置

入口与准备：顶部本机或服务器选择器；硬件详情。

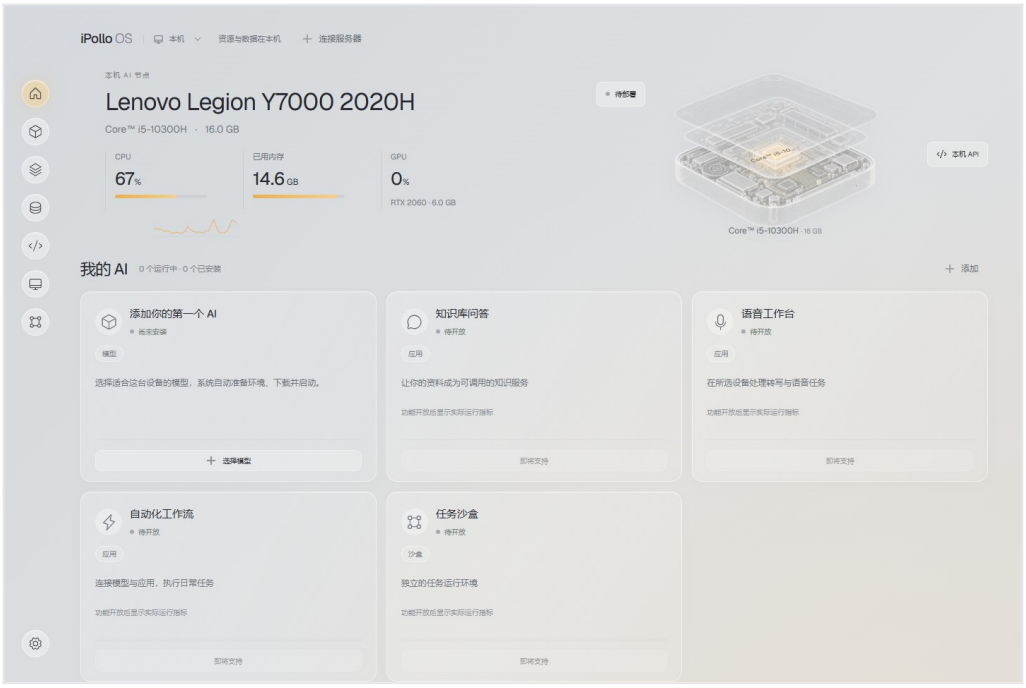


图 01 确认运行位置与设备硬件

先确认本机或目标服务器，再根据实际硬件选模型。

1. 先选择本机，查看 CPU、GPU、内存与磁盘；切到服务器时重新核对设备名称。
2. 打开硬件详情检查实际资源与驱动信息，按可用内存和显存选择模型规格。
3. 后续下载、启停与 Key 管理都先核对当前设备；切换设备不会自动迁移模型或数据。

完成标志：当前操作位置明确，所选模型符合实际设备条件。

1.2 下载启用停止与模型配置

入口与准备：模型市场与模型详情；确保足够空间和网络。



图 02 筛选模型能力与规格

核对能力 内存和存储要求，再安装和启用模型。

1. 选择模型家族和规格并安装，等待运行环境、下载与校验完成，勿把已下载当成已就绪。
2. 启用模型后进入详情试用；按需要设置上下文、驻留与随应用启动策略。
3. 不用时停止或释放内存；卸载前确认不再依赖本地模型文件。

完成标志：模型真实返回内容，状态与资源占用相符；失败先检查下载、内存、驱动和运行日志。

1.3 图片和视频模型任务

入口与准备：图片或视频模型详情；需受支持硬件与组件。



图 03 图片模型的组件与资源要求

图片和视频模型均需完整组件与匹配硬件；此处显示安装前状态。

1. 安装对应完整模型组件，检查磁盘和内存条件后启用。
2. 输入提示词及界面允许的参数，提交后保留任务记录并查看状态。
3. 成功后点击缩略图或结果预览并保存；取消操作后确认任务停止，勿重复提交未知状态任务。

完成标志：有实际生成文件且可打开。未完成硬件验收的模型按试验能力使用。

1.4 创建 API Key 与网络入口

入口与准备：API → 模型 API 网络设置及密钥管理。

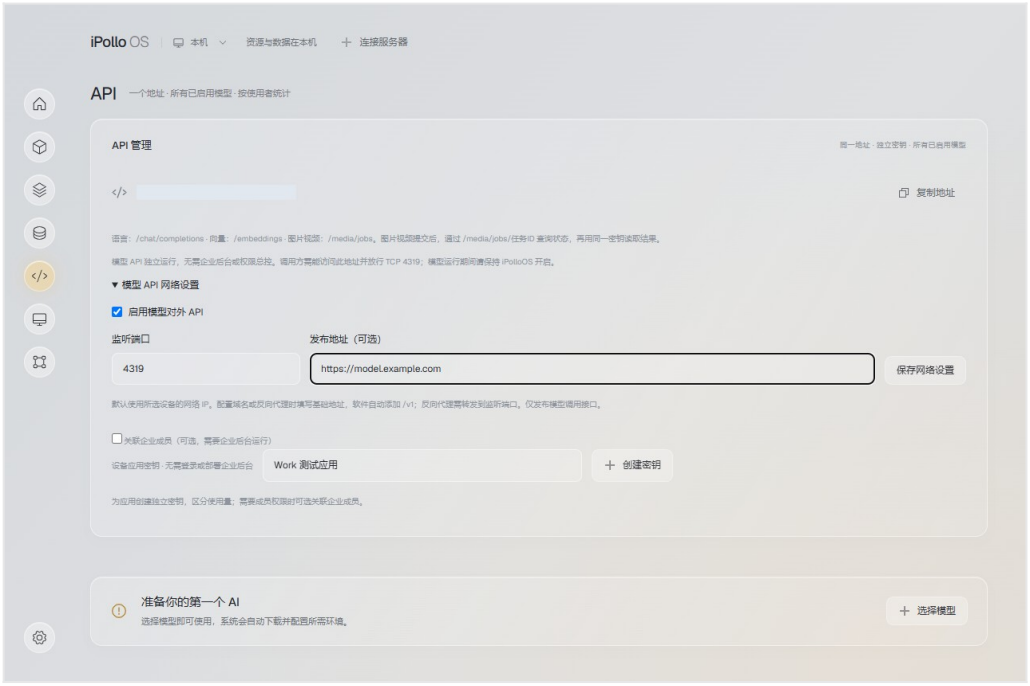


图 04 模型入口 网络设置与应用 Key

先核对发布地址和端口，再创建独立应用 Key；图中地址已遮挡。

1. 启用模型入口并检查端口和发布地址；默认通常是设备 IP 的 4319 端口，复制页面显示的 /v1 基址。
2. 创建设备应用 Key 并保存完整值；需要企业约束时再选择关联成员。
3. 从调用电脑请求模型列表，再使用列表内的 ID 发送聊天；停用 Key 后核对新请求被拒绝。

完成标志：实际客户端能取得模型并返回内容；设备控制 4318、企业 4320 与模型 4319 分别使用。

1.5 向量媒体接口与计量

入口与准备：已启用相应模型；调用端持有 OS Key。

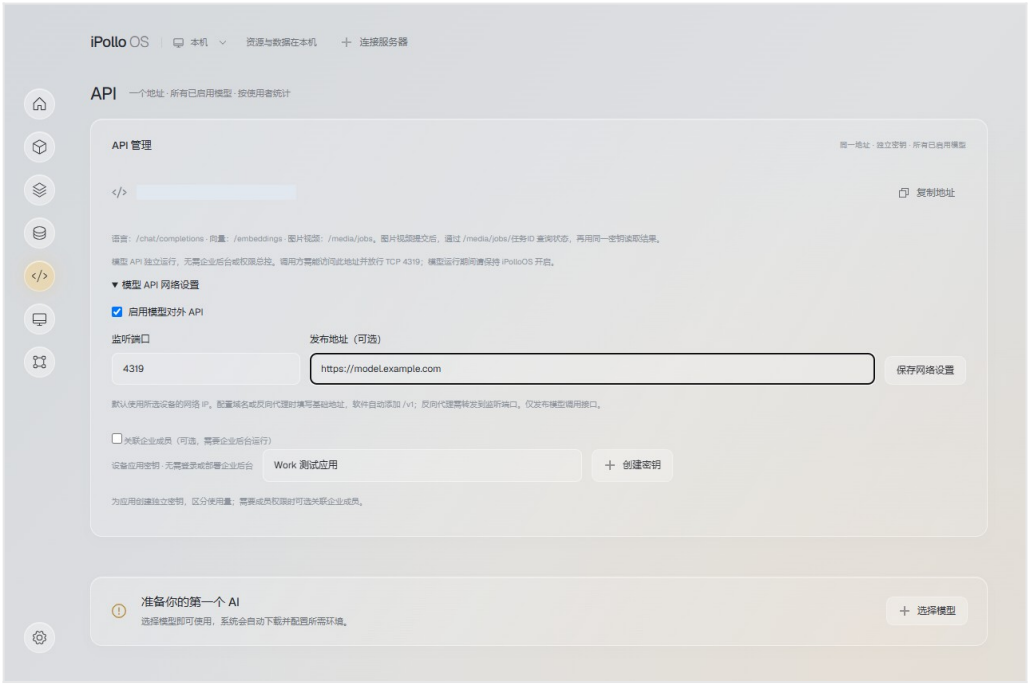


图 05 模型 API 的调用类型和计量入口

从页面核对语言 向量与媒体接口，按实际模型能力调用。

1. 文本生成调用 /chat/completions；向量模型调用 /embeddings，模型名使用实际发现结果。
2. 媒体先查 /media/models，再提交 /media/jobs，并按任务 ID 查询、取消或读取结果。
3. 在模型详情或 API 计量检查来源、请求和 Token，区分已接受、成功与失败任务。

完成标志：接口类型与模型能力匹配，计量可核对；未实现的其他兼容平台 API 不应直接套用。

第 2 章 服务器路由与跨机功能

2.1 通过 SSH 接入服务器

入口与准备：顶部 → 连接服务器；SSH 地址、用户、端口和私钥。



图 06 通过 SSH 连接服务器

填写目标地址和有效 SSH 身份，先检测再准备运行环境。

1. 填写信息并检测；已运行 OS 的设备直接接入，未安装的受支持 Linux 设备按检查结果准备后台。
2. 安装前确认 root 或免交互 sudo、Python、空间与网络等条件，等待健康检查就绪再进入服务器。
3. 在远程位置核对硬件、服务和连接诊断；需要受限管理时按当前权限登录流程认证。

完成标志：目标设备后台就绪且可管理。删除连接只清除本机连接记录，不自动卸载服务器数据。

2.2 模型路由与多个完整副本

入口与准备：API → 模型路由；已有可用模型或已连接设备。

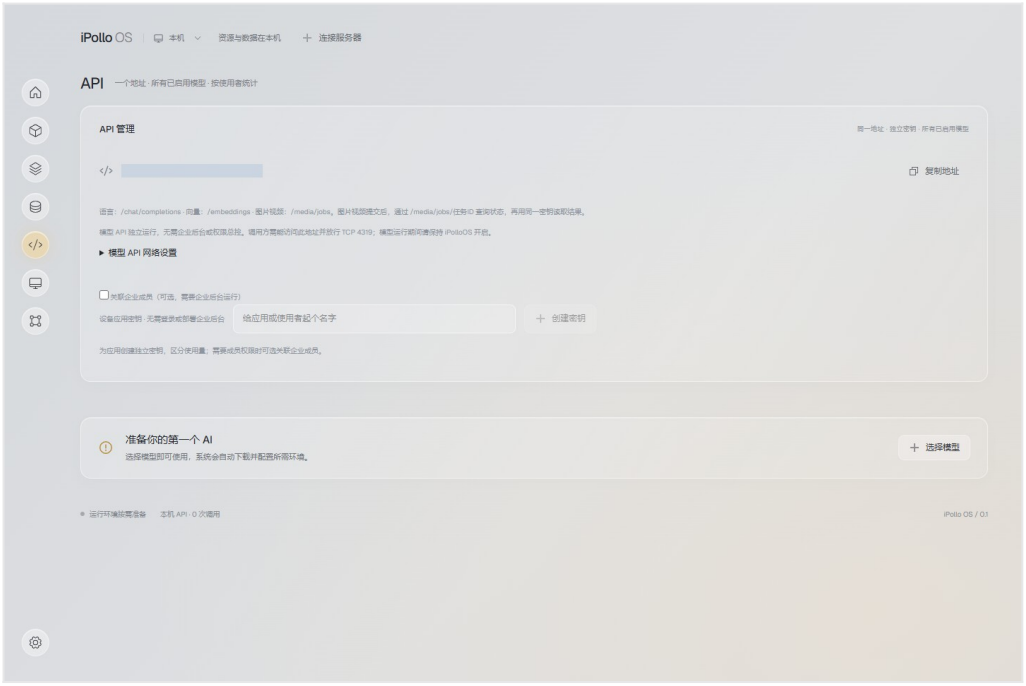


图 07 统一模型入口

设备和模型就绪后再创建路由；使用页面返回的模型 ID 验证调用。

1. 新建路由并选已登记的模型目标；同模型选择均衡方式，不同模型按允许的顺序中转。
2. 设置并发、超时和故障冷却，或选择多台设备批量安装并启用同一模型，分别等到模型就绪。
3. 复制 route 模型 ID 试用，查看各节点状态、调用与失败记录；修改前等待活动请求结束。

完成标志：统一 ID 可调用且请求分配可观察。完整副本模式不合并多台设备显存或内存。

2.3 CPU 双机与 GPU 跨机部署

入口与准备：模型服务的跨机部署入口；仅在版本展示该开发功能时使用。



图 08 准备跨机部署的目标设备

本图显示服务器接入前的配置；跨机分组需在目标设备连接后操作。

1. CPU 验证选择两台符合条件的 Linux 服务器及固定小模型；GPU 组选择受支持节点和规格，先跑资源与通信预检。
2. 创建组后查看分片、日志和整体就绪，完成实际推理；CPU 拆分需确认两台均持有非零模型缓冲区。
3. 使用模型组 ID 验证调用、停止和恢复；测试结束使用组清理功能并核对资源。

完成标志：完整组有真实输出。CPU 小模型验收不代表大模型容量，GPU 多机性能仍需目标硬件实测。

2.4 网络诊断与服务器可达性

入口与准备：服务器端口与连接；识别为阿里云 ECS 时可有授权及安全组入口。



图 09 服务器地址和连接检测入口

按目标网络检查 SSH 与服务端口；连接成功后再排查外部可达性。

1. 先检测 SSH、HTTP、HTTPS 和已配置服务端口，查看检测时间与所用通道。
2. 需要外部访问时核对防火墙、云安全组和反向代理；实际端口以该服务配置为准。
3. 若当前版本提供阿里云安全组管理，先配置有权限的云账号授权，预览目标规则后再应用。

完成标志：从使用者所在电脑可达实际服务。诊断本身不修改规则，其他云平台按各自运维方式处理。

第 3 章 应用部署与持续运维

3.1 安装应用模板或公共镜像

入口与准备：应用与集群 → 部署中心；目标容器环境已准备。



图 10 选择应用模板或公共镜像

核对部署目标和环境准备情况，再进入对应创建流程。

1. 先选择正确服务器和容器空间；从已支持模板创建，或填写公共镜像。
2. 填写应用名、端口、CPU、内存和实例范围；按需要配置访问域名及就绪检查。
3. 提交部署并查看任务，待实例就绪后从应用详情访问或创建私有预览。

完成标志：目标实例真实就绪且能访问。域名需另有 DNS，数据库和持久卷按当前支持范围配置。

3.2 从本地源码构建发布

入口与准备：部署中心的源码入口；本机有完整项目目录。

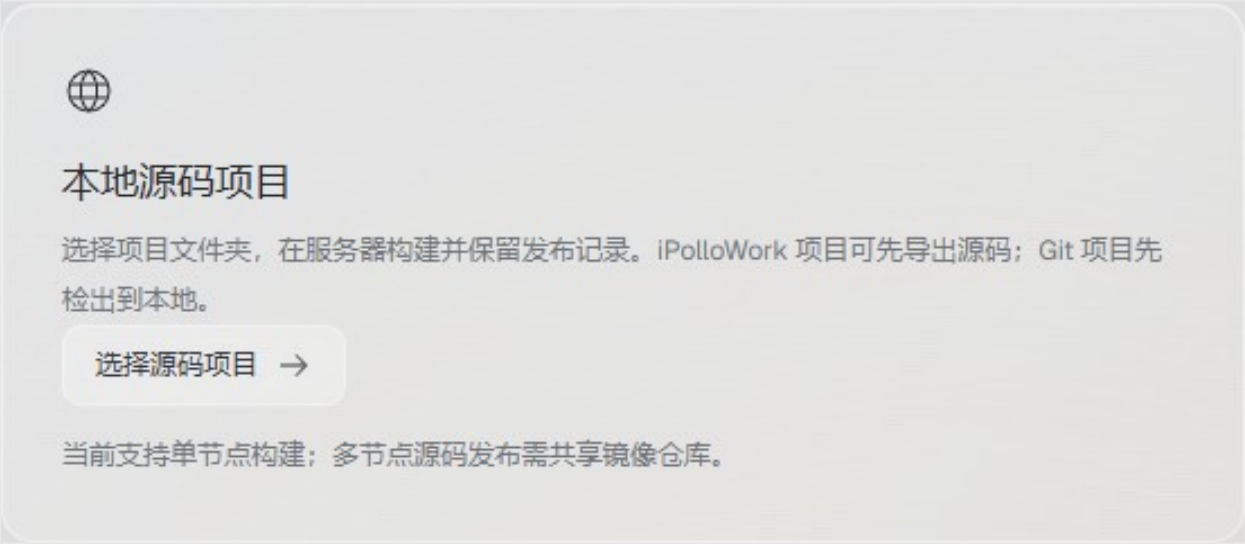


图 11 选择并检验源码的入口

选择包含完整源码的文件夹，再按检验结果配置目标空间与应用。

1. 选择源码目录，读取技术栈、端口和构建检测结果；用 .ipolloignore 排除额外无关文件。
2. 确认目标空间、资源和运行配置后上传当前快照，等待远程构建与部署阶段完成。
3. 修改后向同一应用再次更新，保留身份和地址；查看最新构建日志与发布版本。

完成标志：新代码真实运行。当前源码路径受单节点、80 MB 和 20000 文件等限制，不是保存即热更新。

3.3 环境变量健康检查与伸缩

入口与准备：应用详情 → 配置；具备应用管理权限。



图 12 应用与部署管理入口

进入具体应用后检查配置 健康状态和实例；需要已就绪的部署环境。

1. 按需调整镜像、资源、端口、命令、健康检查与实例上下限，核对预算和 HPA 目标。
2. 添加环境变量时填写名称与值；已有值按保留、替换或删除处理，源码 PORT/HOST 由系统管理。
3. 保存后查看任务和就绪实例，再验证业务请求；更新期间可能短暂中断。

完成标志：应用使用新配置并保持预期状态；HPA 调整应用副本，不自动购买云服务器。

3.4 日志预览启停与代码回退

入口与准备：应用详情的实例、日志、事件与版本入口。



图 13 从应用入口检查运行结果

在目标应用详情读取实例 日志与发布版本，并核对实际访问结果。

1. 选择真实实例读取日志，结合事件和资源定位失败；需要时查看上一次容器日志。
2. 预览后用本机浏览器检查；暂停应用用停止，再启动时等待就绪。
3. 在发布历史选择先前成功版本并回退，完成后验证实际镜像和业务；运行配置与数据恢复另行处理。

完成标志：问题可定位，预览或应用可访问，回退版本正确。私有预览只供本机，不作为公开分享链接。

第 4 章 企业后台与权限总控

4.1 部署与使用企业后台

入口与准备：应用与集群 → 部署中心 → 内置应用 → 企业后台管理。



图 14 部署企业后台的配置窗口

确认位置 发布地址和端口，服务就绪后再打开管理页面。

1. 核对顶部选定本机或云服务器，打开部署与配置，填写端口和真实发布地址后部署。
2. 服务就绪后在应用卡片进入管理网页；成员、部门、资源、审核和企业加入沿用 Enterprise 操作。
3. 复制当前企业地址与加入码给 Work；需停止服务时在卡片执行停止，勿仅关闭管理窗口。

完成标志：企业网页和客户端发现均可达。当前独立服务默认 4320，关闭桌面不等于停止该服务。

4.2 部署总控并分配管理职责

入口与准备：内置应用 → 权限总控 → 部署与配置；云端需管理登录与 SSH 权限。



图 15 部署权限总控的配置窗口

总控与企业后台分别管理；部署后按职责分配管理权限。

1. 部署总控，核对端口和发布地址；默认 4330，独立运行，不要求另建数据库。
2. 按当前页面提供的管理员流程登录，在用户表单创建账号，设置后台、运维、总控三类管理权限。
3. 用新用户登录验证：后台职责管理企业业务，运维职责管理设备与部署，总控职责管理用户；可按实际工作多选。

完成标志：用户仅取得被授予的管理能力。软件管理登录不能替代 Linux 的 SSH 或 sudo 权限。

4.3 成员模型授权与资源额度

入口与准备：企业后台的成员、部门与资源设置；OS API Key 关联成员。

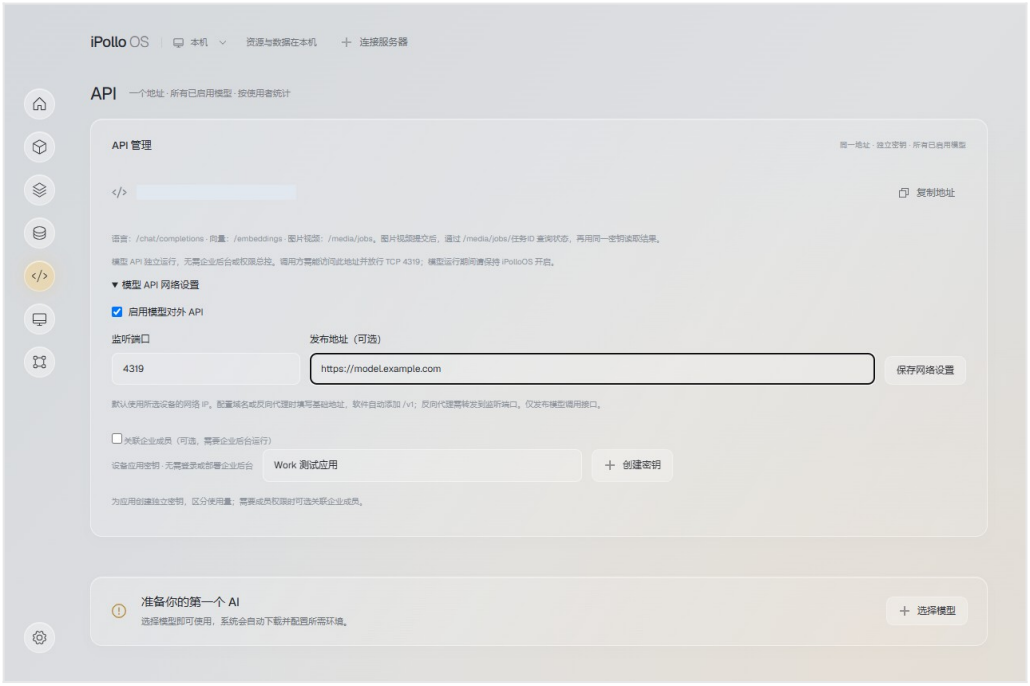


图 16 成员关联与模型 Key 的位置

需要成员约束时，先准备企业后台，再关联成员并验证实际调用权限。

1. 配置成员所属部门与资源使用范围；模型或应用分别选择使用和管理对象。
2. 需要成员审计时创建关联该成员的 API Key，再验证模型列表与实际调用权限。
3. 设置 CPU、GPU、显存、内存与存储额度后核对执行结果；共享推理无法保证硬额度时会拒绝受限调用。

完成标志：授权实时生效且可核对审计。设备应用 Key 与成员 Key 分开；可填额度不等于硬件独立隔离已完成。

4.4 持久服务备份与恢复

入口与准备：应用卡片、服务配置和部署数据目录；管理员操作。



图 17 区分企业后台和总控服务

分别记录服务地址 配置与数据位置，按本节顺序备份和恢复。

1. 分别记录设备服务、企业后台、总控和模型入口的地址与运行状态。
2. 备份企业数据与制品、服务配置、总控 JSON 权限资料，以及恢复所必需的加密密钥。
3. 恢复前停止实际数据持有服务，核对身份和目录；启动后验证成员、资源、加入码及模型 Key。

完成标志：数据和身份恢复一致。本机独立服务可能持续运行，当前 Windows 部署不自动等于开机自启。

企业员工加入和资源使用的七步完整流程见生态指南第三章；企业成员、部门及审核操作见 Enterprise 手册。

附录 A 产品能力与使用范围

本附录保留原项目指南中的产品能力说明，帮助核对功能范围与部署条件。引用编号统一对应生态指南第六章的资料索引。

iPolloOS 是桌面 AI 基础设施应用，把普通电脑和可连接的服务器组织成可管理、可调用的模型与应用运行环境。它提供设备检测、模型安装、推理服务、统一 API、设备连接及空间管理，企业治理复用 Enterprise。[5] [6]

A.1 设备检测与模型选择

读取操作系统、芯片、CPU、GPU、内存、磁盘及可获取的运行指标。模型市场展示规格、大小、建议内存、上下文和设备适配信息，帮助用户在当前硬件条件下选择模型。设备建议和模型目录信息应结合真实加载与输出验证。

A.2 本机模型生命周期

通过托管 Ollama 或已有本机 Ollama 安装与运行语言模型，集中处理下载、续传、校验、启用、停止、内存释放、启动策略和上下文配置。模型详情显示状态、调用次数、Token、速度与调用来源，减少为各模型单独维护启动方式。

A.3 图片与视频模型

托管 ComfyUI 及固定媒体工作流，包含 FLUX.2 Klein、Z-Image Turbo、Wan 等组件路径。媒体生成使用异步任务，支持查询、取消和结果读取。FLUX.2 Klein 已有 Apple 设备真实生成记录；其他媒体模型和硬件组合仍需逐项验收。

A.4 统一模型 API 与密钥

提供 OpenAI 兼容的模型发现、聊天、流式输出和 Embeddings，以及专用媒体任务接口。调用方使用 OS 创建的 API Key，按模型 ID 路由至本机、节点池、集群或已接入的云供应商。Key 可停用，也可关联成员权限；兼容范围以已实现接口为限。[7]

A.5 路由和完整模型副本

可给模型配置 route 统一 ID，设置同模型副本分配、不同模型顺序中转、并发上限、超时与故障冷却。同模型副本也可在多台原生服务器批量安装、启用。每个副本独立运行完整模型，这类请求分配不合并设备内存。

A.6 远程设备与跨机模型

通过 SSH 检测和接入服务器，按目标环境准备后台或 K3s。另有独立的跨机拆分路径：固定小模型的 CPU 双机拆分已有实际调用与恢复记录；2 至 4 节点 GPU TP/PP 部署属于开发版实现，受同型号 GPU、驱动、网络和资源预检限制，尚需真实多机性能验收。[7]

A.7 iPolloOS 应用部署与运行边界

A.7.1 应用模板与容器部署

在准备好的 K3s 或符合条件的 Kubernetes 环境中部署当前支持的公共镜像和应用模板。模板市场展示已支持项目，用户配置资源、端口和实例范围后执行部署。是否能部署特定应用取决于镜像、认证、存储、网络和运行期配置要求。[5] [7]

A.7.2 源码构建与持续发布

检测本地源码项目，筛选和上传快照，云端构建镜像并发布到同一应用。技术栈识别覆盖 Dockerfile、Node.js、Next.js、Vite、Python、Go 和静态站点等路径。源码更新保持应用身份，发布历史可用于回退代码镜像；数据库和用户文件需要各自恢复机制。

A.7.3 应用运维与预览

查看实例状态、日志、事件、资源用量、配置和发布结果，执行启停、更新与代码回退。应用副本可配置范围并使用 HPA，私有预览便于开发者从本机访问云应用。实际环境变量、存储、HTTPS 和配套服务能力应按目标版本与部署模板核对。

A.7.4 企业后台与权限总控

应用与集群可分别部署企业后台管理和权限总控。企业后台维护成员、部门、资源、审核及成员资源授权；总控按后台、运维、总控三类职责管理软件操作者。模型 Key 可关联企业成员；共享推理无法保证硬额度时拒绝受限调用，不能把额度设置等同于独立硬件隔离。[6] [11]

A.7.5 网络接口的正确使用

内部控制服务使用 127.0.0.1:4318。当前模型对外入口默认独立使用 4319，企业后台通常使用 4320，权限总控通常使用 4330；地址和端口均以页面实际配置为准。模型基址从 API 页复制，常见形式为 http://设备 IP:4319/v1。模型 Key、企业会话和总控登录分别使用。[11]

A.7.6 验证状态与交付限制

能力	当前说明
本机语言与图片	Apple Silicon 和部分 Windows 设备已有真实运行记录；不能推广为所有设备均通过。
CPU 双机拆分	固定 SmolLM2 小模型已完成双机实际推理、故障、停止与恢复验证；不代表大模型性能。
GPU 跨机与云模型	GPU 部署链路和部分供应商适配已实现；多机 GPU 性能、真实厂商调用仍需目标环境验收。
应用与平台设施	完整云端复验依赖服务器可用；通用数据库市场、零副本唤醒、自动购机扩容等尚非完整能力。

资料依据与使用边界

本册整合《iPollo 项目清单与功能使用指南》和 iPolloOS 本地项目说明。企业接入相关内容同时依据《企业功能使用教程》；截图在 2026 年 11 月 2 日采集。

模型、容器镜像及子模块各有独立许可；OS 项目授权不自动涵盖模型权重的商业使用权。

涉及安装、部署、模型调用、支付、成员权限和外部发布时，应使用目标版本并核对实际结果。开发版、待开放及依赖外部账号的能力按页面提示使用；操作入口可见不自动代表整条业务链路已完成验收。

遇到问题时记录软件版本、所在页面、操作步骤、预期结果和实际提示，保留已产生的文件与任务记录；反馈前移除密钥、口令和客户资料。