豪泓科技
HAOHONG TECHNOLOGY

H2 AIDM · 技术总览 · 00

四层贯通，一体交付。

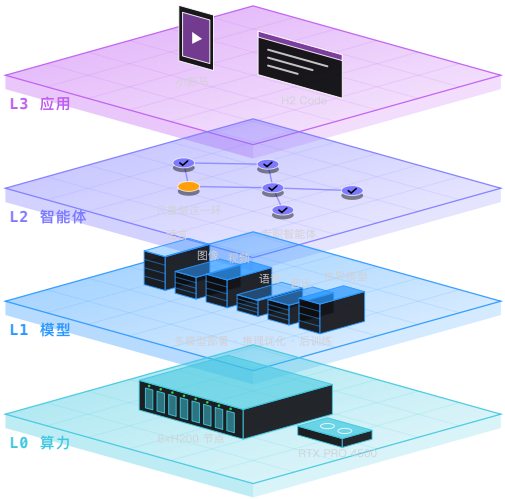
算力、模型、智能体、应用，四层都自己做。客户可以整套拿走，也可以只要其中一层。

上海豪泓信息技术有限公司 | v5.1 | 2026 年 11 月

数字口径：实测 / 生产实证 / 已上线 / 规划测算，见各页标注。规划值以实际运行数据为准。

四层贯通，一体交付。

算力、模型、智能体、应用，豪泓四层都自己做。
让每一家企业拥有自己的 AI，而不是租用别人的。



H2 AIDM · 一体智能

四层贯通，一体交付。

下层托起上层，上层打磨下层。客户可以整套拿走，也可以只要其中一层。

4 层 · 6 条产品线

每一层都能单独交付

↑ 向上支撑	算力 → 模型 → 智能体 → 应用，逐层托起。
↓ 向下反哺	应用的真实负载，逐层反哺到算力。

行业痛点 · 为什么要四层一起做

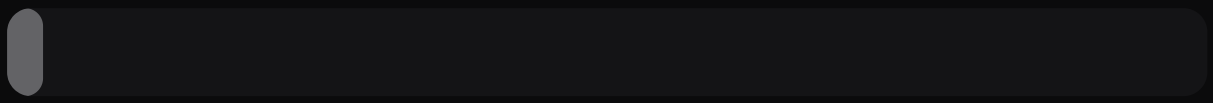
任务越长，越贵、越慢、越容易错。

聊天是一问一答。智能体要连续走几十步：边读、边调工具、边写，上下文越滚越长。看下面这个任务怎么一步步变重。

聊天 vs 智能体 示意

聊天

一问一答



上下文短，问完即止

智能体

一个任务 · 第 30 / 30 步



审查

上下文长度

100 万 词元

每走一步就长一截，可到几十万至百万。 行业观察

累计处理量

30 × 一次聊天

不做缓存，每一步都要把历史重读一遍：单任务是聊天的 5–30 倍。 行业观察

全部做对的概率

21 %

单步 95%，连续 30 步全对只有约 21%。 示意计算

示意动画：红色表示这一步要重读的历史，蓝色是这一步新增的内容。节奏也不同：智能体是「调一次模型、执行工具 30–60 秒、再调一次」，不是人打字的节奏。

越长越贵

历史被一遍遍重算，词元消耗成倍增加。

越长越慢

每一步都要先读完全部历史才能开口。

越长越容易错

一步出错，后面全跟着错，只能整段重来。

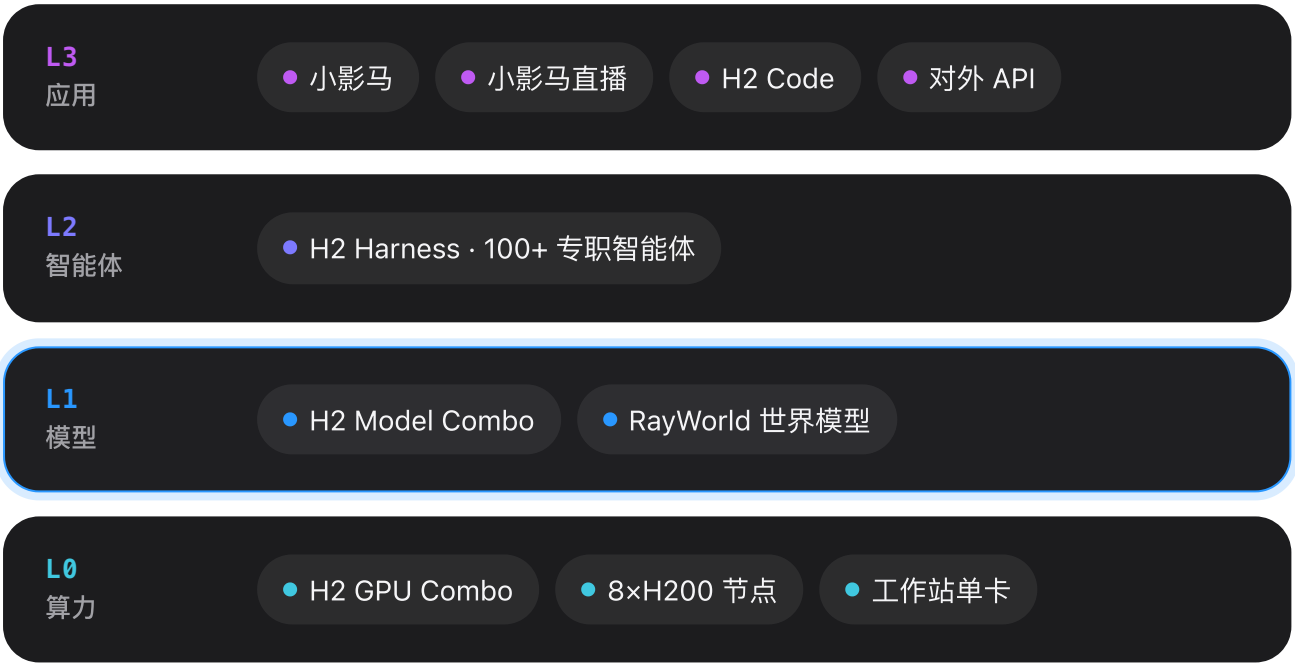
所以要四层一起做

算力让缓存留得住，模型只算新增，智能体只重做错的那一步，应用把它变成产品。

H2 AIDM 架构

四层清楚，
六条产品线彼此咬合。

名字借自半导体：从设计到制造都自己做的，叫 IDM。算力中心做算力，模型服务商做模型，应用公司做应用——AIDM 四层都做。



一条请求怎么走完四层

跟着一个任务，
看每一层在做什么。

用户说一句话，到拿到结果，中间要过七步。请求先一层层往下走，推理完成后再回到上面审查、结算。



① **下任务** · L3 应用
用户一句话描述目标。

② **拆解分派** · L2 智能体
拆成环节，交给专职智能体。

③ **进网关** · 平台
鉴权、计量、冻结预算。

④ **路由** · L0 算力
把会话送到存着它历史的通道。

⑤ **推理** · L1 模型
命中读过的历史，只算新增。

⑥ **审查** · L2 智能体
不合格退回本环，只重做这一环。

⑦ **结算** · 平台
成功才扣费，结果交回用户。

为什么四层要一起设计

④ 路由错了，⑤ 的缓存就白存了；没有 ⑥，任务越长越做不完。任何一层单独做好，都接不住这条请求。

逐层证明

每一层，都有一个硬数字。

2 路 · 25fps

L0 算力 单张 RTX PRO 4500（32GB · 200W）并发两路实时直播，0 掉帧。

生产实证

约 1 秒

L1 推理 长上下文首字约 1 秒；整栈 4 卡 H200 实测 96 路并发。

实测

1/14

L1 世界模型 RayWorld R1 768p 约 0.07 元/秒，约为主流同类的 1/14。

已上线

100+

L2 智能体 H2 Harness 组织的专职智能体；1 天上线一部量产剧。

已上线

真实产出

不是概念图，
是线上生成的真实画面。

下面的片段都来自线上产品：RayWorld R1 样片取自 rayworld.ai，作品片段取自小影马。













六条产品线

每一层都能单独交付。

客户可以整套拿走 AIDM，也可以只要其中一层。

L0 · 算力

H2 GPU Combo

量体裁衣：按客户现有的算力设备定方案，有什么卡就把什么卡用好。

单卡两路 25fps 直播

了解 H2GC ›

L1 · 模型

H2 Model Combo

全模态部署、智能体推理优化与架构级后训练。

长上下文首字约 1 秒

了解 H2MC ›

L1 · 世界模型

RayWorld

R1 离线世界模型 × Live 实时直播模型，相机成了参数。

768p 约 0.07 元/秒

了解 RayWorld ›

L2 · 智能体

H2 Harness

服务端智能体运行时架构：分工 · 审查 · 单步重做 · 记忆与复用。

100+ 专职智能体

了解 H2 Harness ›

L3 · 应用

小影马

AIGC 内容创作平台：一句话进去，全自动出成片；短剧、影游、广告短片、直播都能做。

5 类内容形式 · 8 语种

了解小影马 ›

L3 · 应用

H2 Code

通用桌面 Agent：写材料、查资料、做表，能私有化部署。

长任务首字约 1 秒

了解 H2 Code ›

哪些是自研的

说清楚自己做了什么，也说清楚用了谁的开源。

层	豪泓自研	采用的开源技术
L0 算力	**GPU 底层优化、异构协同、会话亲和路由与调度**	分页显存参考 PagedAttention 等开源推理框架思路
L1 文本推理	**有状态缓存、通道锁定等推理工程**	DeepSeek 等开源模型；推测解码技术 DSpark
L1 世界模型	**RayWorld R1 的结构改造、新增权重、验收标准**	开源视频基座模型
L2 智能体	**H2 Harness 编排框架**	对接开源与商用模型，按能力槽路由
L3 应用	**小影马、小影马直播、H2 Code**	—

对外是两门生意

同一套技术，两种交付方式。

业务一

AI 转型技术服务

给想转型为 AI 技术公司的企业，重点是上市公司：钱和卡都有，但卡的利用率低，没有自己的模型能力，也没有能写进财报的 AI 产品与收入。

- GPU 优化——验收看每张卡服务多少路、多少用户
- 模型优化——验收看首字、出字速度与能力完整度
- 模型部署——验收看可用性与容量
- 模型后训练——验收看评测集指标，权重归客户

业务二

企业 Agent 定制服务

给任何有重复性知识工作的企业。按客户的工作流定制一套多 Agent 系统，部署在客户本地，数据不出域。

- 工作流拆解 → Agent 架构设计 → 系统对接
- 本地化部署 → 知识接入与后训练 → 评测上线
- 权限、审计、人工确认点一并设计
- 证据：小影马把短剧行业的工作流交给了智能体

两样都要的客户，拿到的就是整套 AIDM。

一笔账 经营测算

同样 8 张卡，卖法不同，产值与估值都不同。

一台 8 卡 H200 服务器，每月产值。只出租，卖的是卡时；部署模型，卖的是推理；做成智能体，卖的是完成的工作。

只出租卡时

卖算力 · 按卡时计费

约 15 万元 1×

部署 DeepSeek

卖模型推理 · 按词元计费

约 50 万元 约 3.3×

做成智能体服务

卖完成的工作 · 按席位与任务

约 100 万元 约 6.7×

估值逻辑 A

算力租赁

重资产生意。收入跟着卡数与出租率走，市场按算力资产与租金回报给估值。

估值逻辑 B

AI 产品与服务

收入跟着用户与任务量增长，可以写进财报的是 AI 产品收入，市场按收入规模与增长给估值。

每月产值为经营测算，按一台 8 卡 H200 服务器估算，以实际运行数据为准。从 A 走到 B，差的不是卡，而是四层技术：L0 让卡跑得满，L1 让推理快且便宜，L2 让智能体把长任务做完，L3 把能力变成能卖的产品。

下一步

从一次演示开始。

接入方式与官方接口一致，最快 4–6 周上线第一个产品。

01

30 分钟演示

看平台与三款成品的实际使用。

02

3 天试用

结束后给一份延迟、用量、费用报告。

03

一页纸定方向

定下做哪几件事，按模块单买或整包。

联系我们 H2-AI@outlook.com 官网 H2-AI.tech