豪泓科技
HAOHONG TECHNOLOGY

H2 AIDM · L1 模型层 · 白皮书 02

# H2 Model Combo
# 为智能体重新做推理。

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。

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 / 示意，见各处标注。规划值以实际运行数据为准。

## H2 Model Combo （H2MC）

# 同样的模型，为智能体重新做推理。

为智能体负载重写推理路径，把五类模型部署成标准服务；再用架构级后训练，把业务约束写进模型本身。




当前步 **8 / 8**　　本步需要计算 **只算新增 6 单位**　　首字等待 **第 1 步冷启动，之后 ≈ 1 秒**

长上下文智能体任务，首字约 1 秒 实测；红色是同一套硬件上关闭缓存、每步完整重算的自测对照（约 34 秒）。第 1 步没有缓存，两种方式都要完整预填充（冷启动）；之后有状态缓存只算新增。时间按比例缩放，每步上下文长度为示意。

**约 1 秒**

长上下文智能体任务的首字时间（H200 实测）。

实测

**200–300**

单路出字速度（词元/秒），优化前约 90；H200 单路实测。

实测

**96 路**

整栈结果：4 卡 H200 实测 96 路并发推理。

实测

**约 100 万**

上下文最高约 100 万词元，能力不为速度让步。

与官方同档

问题

## 智能体负载，不是聊天负载。

按聊天设计的推理服务，每一步都要重算全部历史：首字要等半分钟，十几路并发就互相拖慢。

| 特征 | 聊天 | 智能体 |
| --- | --- | --- |
| **上下文** | 短 | **持续增长，可到几十万至百万词元** |
| **每步新增内容** | 整段都是新的 | **只有一小段是新的** |
| **请求节奏** | 人打字的节奏 | **请求一次模型，执行工具 30–60 秒，再请求一次** |
| **单任务消耗** | 基准 | **聊天的 5–30 倍** |

三项推理机制

# 缓存、锁定、加速，三项必须同时成立。

能力没有为速度让步：百万词元上下文、工具调用、看图、思考开关全部保留。

机制 01 · 有状态缓存

## 读过的，不再重读。

读过的历史以 KV 缓存驻留在算力侧（显存与主机内存两级），下一步只为新增的那一小段做预填充。第一步仍是冷启动；之后首字几乎不再随历史长度增长。长上下文智能体任务首字约 1 秒 实测 。

提示缓存已是行业的通行做法。我们做的，是在自有算力、高并发下把它做稳：缓存分级驻留，会话亲和让同一会话始终落在存着它历史的通道上。

解码时注意力仍要读取完整的 KV 缓存，所以首字之后的出字速度交给机制 03。

关闭缓存 · 自测对照

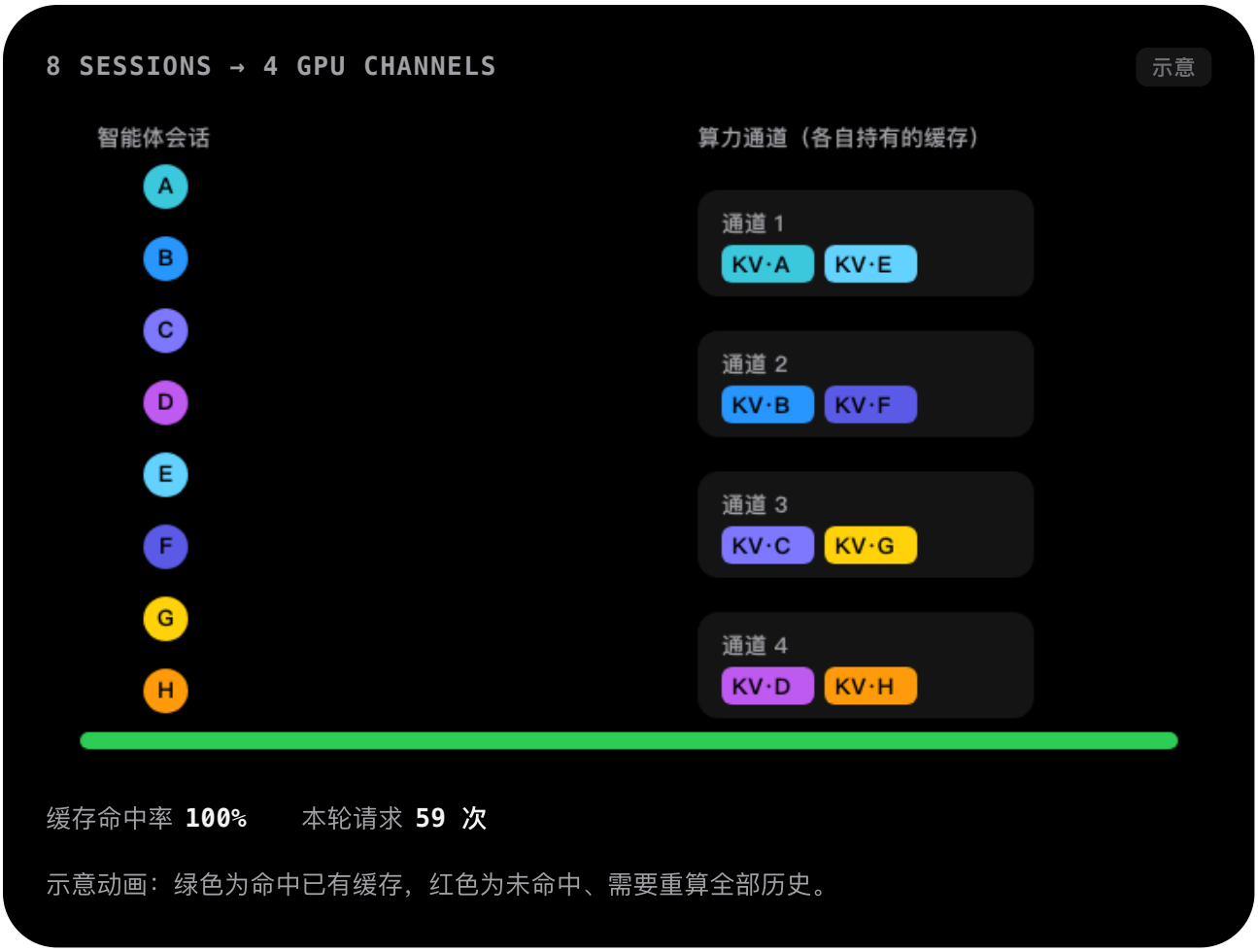
**约 34 秒**

同一套硬件上，每一步把全部历史重新算一遍。这是我们自己的对照基线，不代表其他厂商的水平。

有状态缓存 + 会话亲和

**约 1 秒**

只算新增内容，首字时间几乎不再随历史长度增长。




机制 02 · 通道锁定

## 同一个会话，<br>始终走同一条通道。

会话亲和（通道锁定）由 L0 调度实现、平台网关执行：同一个智能体会话始终落在持有它缓存的那条算力通道上。H2MC 把它列为三项推理机制之一，与 H2GC 协同。缓存命中是并发做上去的前提；整栈结果是 4 卡实测 96 路并发 实测 。

> 只开缓存不锁通道，缓存等于没开。

机制 03 · 推测解码

# 首字之后，<br>出字也要快。

在解码阶段用推测解码技术 DSpark 加速出字：单路从约 90 提到 200–300 词元/秒（H200 单路实测） 实测 。输出与原模型一致，工具调用、看图、思考开关全部保留。

DECODE · 同一段输出 示意

优化前 · 约 90 词元/秒 **41 字**

收到。我先把三份季度报表按科目对齐，再核对应收账款与合同台账的差异：第二季度有两笔

H2MC · 约 250 词元/秒 **114 字 · 完成**

收到。我先把三份季度报表按科目对齐，再核对应收账款与合同台账的差异：第二季度有两笔回款的入账日期晚于合同约定，合计影响现金流约四十万元。接下来我会调用表格工具生成对比图，并把需要财务确认的三处差异列在文末，等你批复后再写入系统。

示意：两边按约 90 与约 250（200–300 取中值）词元/秒的比例输出同一段文字，时间放慢 5 倍；一个汉字按一个词元近似。

全模态部署

# 一套硬件，<br>同时提供五类模型服务。

不同模态的瓶颈不同，做法也不同。五类模型对上归并为语言、图像、视频、音频四类能力槽（语音与音乐同属音频槽）。

| 模态 | 纳管的代表模型 | 主要瓶颈 | H2MC 的做法 | 对外服务 |
| --- | --- | --- | --- | --- |
| **语言与推理** | DeepSeek、GLM、Kimi 等开源模型 | 长上下文反复预填充；解码受显存带宽限制 | **有状态缓存 · 通道锁定 · 推测解码 · 融合 MoE 算子** | 智能体推理 API、包月、企业私有部署 |
| **流式语音合成** | 流式 TTS 模型（音色克隆、多语种） | 首包延迟敏感；高并发小请求调度碎片 | **分块流水线解码 · 动态批处理** | 配音、直播语音、多语种配音 |
| **AI 音乐** | 配乐与歌曲生成模型 | 长音频序列显存膨胀 | **分块解码 · 低精度算子融合 · 异步队列** | 短剧配乐、主题曲、音乐 API |
| **图像** | 多参考生图与局部编辑模型 | 多步去噪耗时；控制条件叠加占显存 | **步数蒸馏 · 注意力融合 · 权重共享** | 角色与场景母图、商拍、海报 |
| **视频与世界模型** | RayWorld R1 / Live、MiniMax H3、Wan | 时空注意力计算量大；视频解码易 OOM | **时空注意力稀疏化 · 步数蒸馏 · 分块解码 · 潜变量复用** | 按秒计费视频（RayWorld 768p 每秒 1 积分 = 0.07 元）、25fps 直播流 |

架构级后训练

## 把业务约束写进模型，而不是写进提示词。

方法由豪泓在 RayWorld R1 上首先落地，再沉淀为 H2MC 的后训练服务。给企业客户做后训练时，新增的权重归客户所有。

| 步骤 | 做什么 | 做错了会怎样 |
| --- | --- | --- |
| **① 拆解** | **找到最小改动面，只改必要的那一层** | 改大了，原有能力崩掉 |
| **② 接线** | **把结构化约束（如相机位姿、业务规则）接进注意力** | 接错层，约束不生效 |
| **③ 保能力** | **逐项回归，再用蒸馏和量化把成本压回单卡** | 只顾一致性，画质掉、价格涨 |

验收看评测集上的指标提升。RayWorld R1 的结果：相机位姿成了请求里的参数；在同一段生成内，镜头转开再转回，物体保持原位；单卡推理，768p 约 0.07 元/秒 已上线 。

平台与价格

## 能计费、能对账，可审计、可回滚。

01 · 迁移

### 只改地址和密钥

兼容 OpenAI 与 Anthropic 接口协议，客户现有代码与 Agent 框架只需改 base_url 和 api_key。

02 · 稳定

### 主备切换，自动熔断

主备自动切换；同一上游连续出错时自动熔断，避免故障扩散。

03 · 账务

### 先冻结，成功才结算

调用前冻结额度，成功才结算，失败原额退回，不重复扣费。

04 · 私有化

### 2 台普通服务器起

平台本身 2 台普通服务器起步，可部署在客户机房，数据不出域。

已上线

**69 / 199 / 369** 元/月

智能体推理包月三档。

已上线

**3 折**

按量计费，为官方价的 3 折。

已上线

**0.07** 元/秒

RayWorld R1 768p。在小影马上按积分售卖：1 积分 = 0.07 元，768p 每秒 1 积分。

定制

**一单一议**

专属节点与定制后训练项目。

在 H2 AIDM 中的位置

# 向下吃透算力，<br>向上供给能力槽。

↓ L0 算力 · 扎根

## H2 GPU Combo

运行在 H2GC 优化过的存量卡、RTX PRO 4500 与 8×H200 节点上；分页显存池与会话亲和（通道锁定）由 L0 调度实现，是缓存命中的前提。

了解 H2GC ›

↔ L1 世界模型 · 同层

## RayWorld

架构级后训练的方法由豪泓在 RayWorld R1 上首先落地，再沉淀为 H2MC 的后训练服务，开放给企业客户。

了解 RayWorld ›

↑ L2 智能体 · 供给

## H2 Harness

长任务动辄几十轮、几十万上下文；约 1 秒的首字让多角色编排跑得快、用得起。四类能力槽让底层模型可以随时切换。

了解 H2 Harness ›

↑ L3 应用 · 供给

## 小影马

小影马的文本、图像、配音、音乐、视频模型同时在线服务，全部经 H2MC 部署。

了解小影马 ›

↑ L3 应用 · 供给

## H2 Code

H2 Code 的长任务直接受益于有状态缓存与推测解码：首字约 1 秒，出字 200–300 词元/秒。

了解 H2 Code ›