模型与平台能力矩阵

对比国内外模型平台在视频理解与实时性上的真实边界。

先看结论

当前主流“先进模型 / 平台”可以分成三类：

1. 通用多模态实时模型 适合低延迟语音和视觉交互，不适合作为泳池防溺主检测引擎。
2. 通用视频理解模型 适合对视频做抽帧分析、摘要、事件定位、回放说明，但仍不是成品安全系统。
3. 专用计算机视觉栈 才是最适合做主告警引擎的路线。

能力矩阵

类别	平台 / 模型	官方支持边界	对防溺项目的价值	主要限制
国际	OpenAI Realtime API / `gpt-realtime`	官方文档强调低延迟、多模态应用；模型页显示支持文本、图像、音频，`Video: Not supported`	可做语音坐席、报警解释、事件摘要、后台问答	不能直接吃标准视频流做高帧率实时检测
国际	Google Gemini Live API	官方文档写明是低延迟实时语音与视觉交互；技术规格为音频、`JPEG <= 1FPS` 图像、文本	可做低频视觉流分析、交互式值守界面	1FPS 边界不适合高动态安全检测
国际	Google Gemini Video Understanding	官方文档支持视频理解；默认 1FPS 采样；并明确提示高速动作可能丢细节	适合回放分析、事件归因、非强实时理解	对溺水这种短时关键动作场景存在采样风险
国际	Google Cloud Video Intelligence Streaming	官方文档支持 RTSP/RTMP/HLS 等直播协议，提供对象检测与跟踪等实时能力	可作为原型、云侧分析或能力补充	页面标注 Beta / Pre-GA，且不是泳池专用产品
国内	Kimi K2.5	官方文档截至 2026-04-06 显示 `kimi-k2.5` 为最新多模态模型，支持视觉与文本输入，并给出视频理解 `video_url` 示例；支持常见视频格式，推荐视频分辨率不超过 2K	可做视频理解、录像复盘、事件解释、管理后台问答	文档展示的是文件 / 视频理解，不是 RTSP 直播安防接口；未见官方“监控级实时流分析”声明

类别	平台 / 模型	官方支持边界	对防溺项目的价值	主要限制
国内	阿里云百炼 / Qwen 视觉理解	官方文档支持图像与视频理解、直接传入 `video_url` 或图像帧列表，并通过 `fps` 参数控制抽帧，默认 `2.0`	可做视频摘要、事件理解、时间顺序分析、辅助分析	仍是抽帧分析，不是即插即用的泳池告警系统
国内	智谱 GLM 视觉模型	官方文档列出 `GLM-4V-Plus-0111` 支持视频、图像、文本输入，强调时间感知、动作理解、最长 2 小时视频理解；模型概览还列出 `GLM-Realtime` 具备跨文本、音频、视频的实时推理能力	在国内模型里，对视频理解与实时音视频的公开能力说明较强；适合辅助理解、复盘、交互式值守	公开文档仍未把它包装成“泳馆实时防溺产品”，落地仍需自建工程链路
国内	火山方舟 / 豆包视频理解	官方文档存在专门的“视频理解”页面，并提供多模态理解、Responses API、低延迟在线推理等文档入口	说明国内平台已有面向视频理解的正式产品能力	当前公开页面依赖 JS，能确认产品入口存在，但在当前环境中不便直接提取更细限制条件
国内	MiniMax 平台	官方模型总览显示覆盖文本、语音、视频、图像与音乐五大方向	说明国内平台具备多模态能力基础	官方公开材料里没有看到“泳池实时防溺”这种专用能力声明
工程主线	专用 CV 栈（检测 + 跟踪 + 姿态 + 规则）	不是单一平台，而是工程组合	当前最现实的主检测路线	需要算法、数据、现场标定和持续调优

怎么理解这些边界

OpenAI

OpenAI 当前最适合放在这个项目里的位置是：

- 告警说明
- 语音对讲或操作员助手
- 事件总结
- 事后复盘

不适合做：

- 主实时检测
- 原始视频流高帧率视觉分析

Google Gemini

Gemini Live API 更接近“ 实时视觉交互助手 ”。 Gemini 视频理解更接近“ 抽样后的视频理解模型 ”。

这两者都能帮助理解视频，但都不应直接被等同为“ 可交付的防溺监控系统 ”。

国内多模态模型

Kimi

截至 2026 年 4 月 6 日，我在 Moonshot 官方文档里查到的最新多模态模型是 `kimi-k2.5`，官方还写明 `kimi-latest` 已于 2026 年 1 月 28 日 停止新用户使用。我没有在 Moonshot 官方文档里查到 `K2.7` 的正式页面，因此当前不建议把 `K2.7` 当作已确认可用前提。

Kimi 的官方价值在于：

- 提供视频理解示例
- 支持常见视频格式
- 适合用来做录像问答、复盘、报警解释

但它并没有在公开文档中给出 RTSP 直播安防级输入说明。

智谱 GLM

智谱在国内官方文档里，对视频与实时音视频的公开能力说明相对完整：

- `GLM-4V-Plus-0111` 支持视频、图像、文本输入
- 模型说明里写了时间感知和动作理解
- 模型概览还列出 `GLM-Realtime`，描述为跨文本、音频、视频进行实时推理

这说明如果优先考虑国内模型，智谱值得重点看。

阿里 Qwen / 百炼

阿里云的视觉理解能力已经很成熟，能做：

- 视频理解
- 高低 fps 抽帧控制
- 图像和视频混合理解

但这更适合：

- 录像审阅
- 规则验证
- 辅助分析

而不是直接替代专用安防检测栈。

火山方舟 / 豆包

火山方舟官方文档已经单列了“ 视频理解 ”和“ 低延迟在线推理 ”等入口，说明它在产品层面已支持视频理解工作流。

但在当前可抓取的公开页面里，详细限制条件不够透明，所以目前更适合把它列为：

- 候选国内平台
- 后续需要实际注册测试的能力

推荐使用方式

主检测层

采用专用 CV：

- 人体 / 头部检测
- 姿态估计
- 轨迹跟踪
- 区域规则
- 时间阈值

辅助理解层

采用通用多模态模型：

- 报警解释
- 对话式检索
- 事件回放摘要
- 场馆运营问答

数据闭环层

- 保留报警前后片段
- 建立误报 / 漏报标注机制
- 人工复核结果反哺规则和模型

对非技术人士最重要的一句提醒

“ 先进模型支持视觉或视频 ”这句话是真的；
但“ 因此我可以低成本做出稳定的泳池防溺系统 ”这句推论，大概率是假的。

关键来源

- OpenAI Realtime API: https://developers.openai.com/api/docs/guides/realtime

- OpenAI `gpt-realtime`: https://developers.openai.com/api/docs/models/gpt-realtime
- Gemini Live API: https://ai.google.dev/gemini-api/docs/live-api
- Gemini video understanding: https://ai.google.dev/gemini-api/docs/video-understanding
- Google Cloud Video Intelligence: https://docs.cloud.google.com/video-intelligence/docs/streaming/live-streaming
- Kimi 主要概念: https://platform.moonshot.cn/docs/introduction
- Kimi K2.5: https://platform.moonshot.cn/docs/guide/kimi-k2-5-quickstart
- 智谱模型概览: https://docs.bigmodel.cn/cn/guide/start/model-overview
- GLM-4V-Plus-0111: https://docs.bigmodel.cn/cn/guide/models/vlm/glm-4v-plus-0111
- 阿里云百炼视觉理解: https://help.aliyun.com/zh/model-studio/vision/
- 火山方舟视频理解: https://www.volcengine.com/docs/82379/1895586?lang=zh
- MiniMax 模型概览: https://platform.minimaxi.com/docs/guides/models-intro