

云端VLA方案：快速落地 vs 最优路线 v1.0

聚焦云端 VLA在“多源输入 → 统一 Observation → 理解 → 决策”的链路：

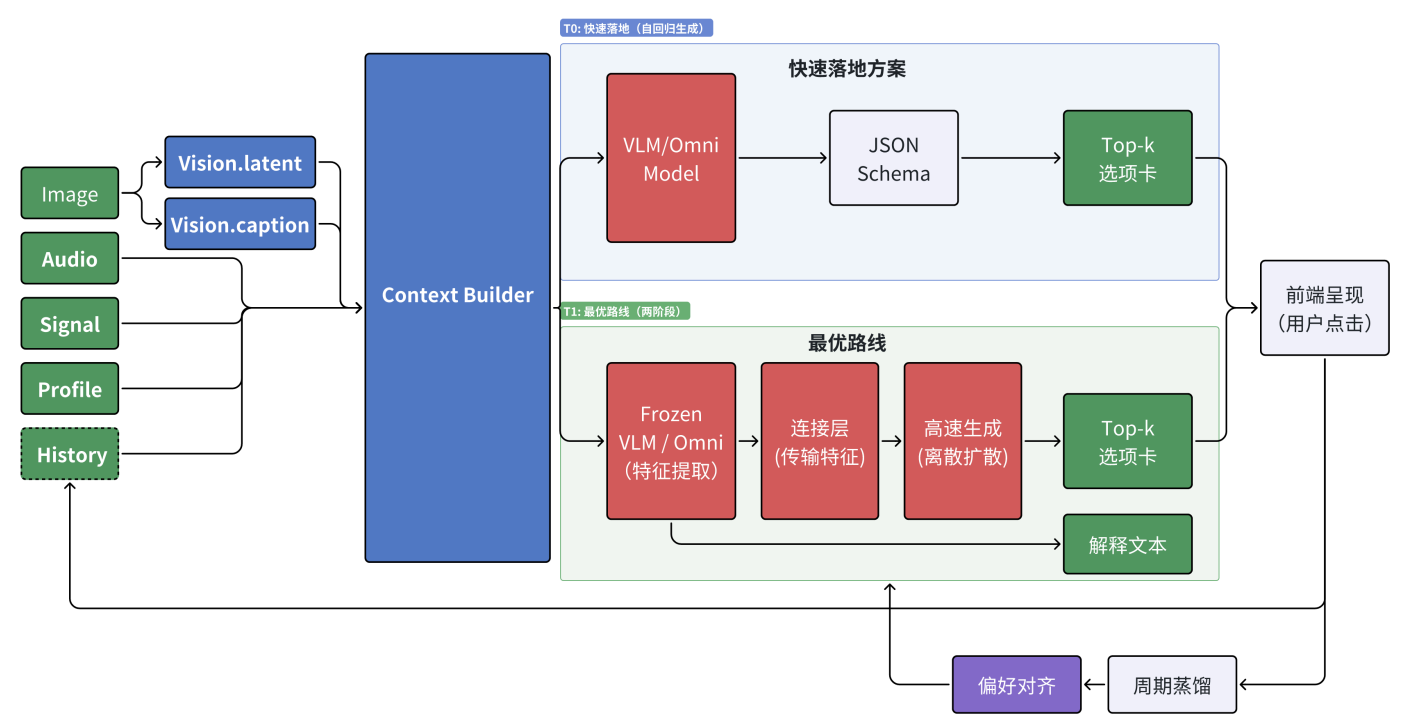
T0（快速落地）：复用开源全模态/多模态模型直接函数调用生成 Top-K OptionCards；

T1（最优路线）：以冻结 VLM/全模态模型为理解（Teacher），经可学习 Connector（如 MetaQuery）传递特征，驱动离散高速生成（Seed/MeanFlow/D2F 等）产生动作 DSL 序列，并支持解释异步补齐与执行绑定。视觉 caption 必填、latent 可选但推荐；TTFT（P95）对外承诺 $\leq 2s$ 。

目录（快速定位）

- 0. 系统框图
- 1. 云端 VLA 建设目标
- 2. 云端 VLA 的输入与输出
 - 2.1 输入（Observation）
 - 2.2 输出（OptionCards + 执行绑定）
- 3. T0 快速落地
 - 3.1 基线模型与路线
 - 3.2 函数调用 Schema（示例）
 - 3.3 系统提示词模板（摘录）
 - 3.4 数据与轻量微调
 - 3.5 工程与 SLA（首 token 时延约束）
- 4. T1 最优方案
 - 4.0 两阶段总体架构
 - 4.1 多模态信息编码（From Qwen2.5 Omni）
 - 4.2 特征传输（Connector）设计与训练
 - 4.3 离散高速生成
 - 4.4 数据与蒸馏
 - 4.5 推理与 SLA

0. 系统框图（云端主流程，按 T0 / T1 分支）



1. 云端 VLA 建设目标

总体目标

在端侧已具备就地感知与简单指令能力的前提下，云端 VLA 作为“多源融合—规划与解释—持续学习”的中枢，面向规模化落地与可运营优化，达成以下目标：

- 1 | 多源融合：统一整合车况信号、用户画像、对话历史、端侧视觉摘要/潜特征与知识库，构建标准化 Observation 表示。
- 2 | 多步规划：将自然语言与多源上下文转为 动作 DSL 的 Top-K 可执行方案，附 可撤销（Undo）与解释说明，确保可回放与可审计（工程层面）。
- 3 | 高效生成：采用 Teacher（SOTA VLM/全模态）× Student（离散扩散/并行掩码）的非自回归动作生成路径，显著降低时延并提升一致性。
- 4 | 持续学习：将“曝光→点击/拒绝→撤销”沉淀为偏好/奖励信号，进行 RLAIIF/DPO 与蒸馏，周期性反哺端侧与云侧。
- 5 | 平台化治理：提供 A/B、模型路由、成本/延迟看板与分布漂移监控，支撑持续运营与稳定交付。

2. 云端 VLA 的输入与输出

2.1 输入（Observation）

- 视觉：vision.caption；vision.latent 可选但推荐（最佳效果，低清/潜特征均可）。
- 语音/文本：intent.text（ASR/文本）、intent.conf $\in [0,1]$ 。

- 车况信号：signals（如 cabin_temp, volume, media_state）。
- 用户画像摘要：profile（温度/音量偏好、勿扰时段等）。
- 历史：history（近 K 轮摘要与最近执行动作）。
- 元信息：scene（如 driving, time_of_day, child_present）、ts, vehicle_id, schema_ver, model_ver, trace_id。
标准结构（示例，最小化）：

```
{
  "ts": 1735912345123,
  "trace_id": "1689a-...",
  "scene": {"driving": true, "time_of_day": "night"},
  "signals": {"speed": 60, "cabin_temp": 28.5, "volume": 10},
  "intent": {"text": "有点热, 调舒服点", "conf": 0.86},
  "vision": {"caption": "后排小孩睡着", "latent": [1024]},
  "profile": {"temp_pref": 24, "volume_pref": 8},
  "history": [{"op": "set_temp", "args": {"zone": "all", "val": 25}}]
}
```

2.2 输出（OptionCards + 执行绑定）

- 默认返回 **K=3** 张候选卡片。每张卡片包含：
 - title**：方案标题/摘要。
 - action_plan**：动作 DSL 序列（受控词表、强类型参数）。
 - explanation**（可选）：方案说明与理由。
 - undo_plan**（可选）：回滚/补偿动作集合。
 - rank_score / confidence**：排序分数与置信度。
 - ttl**：有效期（秒）。

```
{
  "cards": [
    {
      "title": "安静降温（推荐）",
      "action_plan": [
        {"op": "set_temp", "args": {"zone": "all", "val": 24}},
        {"op": "set_fan", "args": {"zone": "front", "val": 2}},
        {"op": "media", "args": {"action": "volume_delta", "val": -2}}
      ],
      "undo_plan": [{"op": "restore_prev", "args": {"scope": ["ac", "audio"]}}],
      "explanation": "夜间行车且儿童入睡，优先安静与舒适。",
      "rank_score": 0.81, "ttl": 10
    },
    {"title": "仅降温", "action_plan": [{"op": "set_temp", "args": {"zone": "all", "val": 24}}]},
    {"title": "保持不变", "action_plan": []}
  ],
  "exec_binding": {"dsl_version": "1.1", "mode": "seq", "api_namespace": "HVAC/Media"} # 后端接口
}
```

3.T0 快速落地（开源模型直连，自回归生成）

3.1 基线模型与路线

- 候选：
 - Qwen2.5-Omni（全模态）、InternVL3、GLM-4V（根据可得性/中文表现选择）。
- 方式：
 - i. 将 Observation 摘要化后作为 **系统Prompt + 工具Schema** 输入；
 - ii. 通过 **函数调用/JSON Schema** 直接产出 Top-K `action_plan[]` 与 `explanation` ；
 - iii. 规则层进行**安全校验与修正**；
 - iv. 返回 **OptionCards**。

3.2 函数调用 Schema（示例）

```
{
  "name": "propose_options",
  "description": "根据Observation生成3个可执行的候选方案",
  "parameters": {
    "type": "object",
    "properties": {
      "cards": {
        "type": "array", "minItems": 3, "maxItems": 3,
        "items": {
          "type": "object",
          "properties": {
            "title": {"type": "string"},
            "action_plan": {"type": "array", "items": {"type": "string"}},
            "expected_effects": {"type": "object"},
            "risk_tags": {"type": "array", "items": {"type": "string"}},
            "undo_plan": {"type": "array", "items": {"type": "string"}},
            "explanation": {"type": "string"}
          },
          "required": ["title","action_plan","explanation"]
        },
        "required": ["cards"]
      }
    }
  }
}
```

3.3 系统提示词模板（摘录）

用于全模态 Observation（ `vision.caption` + 可选 `vision.latent/裁剪图` + `signals` + `profile` + `history` ）的标准提示词，强调“最小变更、可撤销、可执行”，并对单位/范围/冲突做硬性约束。

你是云端 VLA 决策引擎。基于 Observation 的全模态信息（`vision.caption` + 可选 `vision.latent/裁剪图` + `signals` + `profile` + `history`），在“最小变更、可撤销、可执行”的原则下输出 3 个候选方案（JSON）。必须包含动作 DSL 序列（`action_plan`）；`explanation/undo_plan` 可按时延要求省略或延后补充。所有数值须符合单位与范围约束，避免相互冲突（如夜间音量上限、行车中限制）。

3.4 数据与轻量微调

- 快速数据集：构造 2–10 万条模板化样本（Observation → 3 候选），结合规则引擎自动标注约束/撤销。
- LoRA/SFT：小规模微调以稳定函数调用格式与 DSL 一致性；上线后以“曝光→点击/拒绝/撤销”日志开展 RLAI/DPO 周期蒸馏。
- latent 适配（可选）：训练 latent → embed 的小型投影层（MLP/线性投影），贴合所选全模态模型的视觉空间。

3.5 工程与 SLA（首 token 时延约束）

- 服务框架：vLLM；启用多模态 KV 缓存与提示裁剪。
- 时延目标：首 token (P95) ≤ 2.0 s。
 - 流式返回：先回传卡片骨架 (title + action_plan)，explanation/undo_plan 异步补齐；
 - 采用早触发函数调用 (early tool-call) 与模板压缩降低首 token 等待。
- 优点：改造最小、全模态即插即用；可从图像直传平滑切换到 latent 以降带宽与编解码开销。
- 局限：自回归生成对时延与稳定性敏感，性能受提示与模板设计影响。

4. T1 最优方案（冻结理解 × 可学习中间层传输 × 离散高速生成）

4.0 两阶段总体架构

通过冻结 VLM/全模态模型保障高质量理解与低成本稳定性，配合可学习特征传输驱动离散高速生成（4–8 步非自回归）在保持与自回归相当准确度的同时显著降低时延，并具备易扩展、易演进的工程优势。

- 阶段 A | 理解/规划 (Frozen Teacher)
采用冻结的 VLM 或冻结的全模态模型（如 Qwen2.5-Omni 等）完成任务识别、约束抽取与行动草案 y_draft 生成，同时输出多模态聚合表征 z_c。

为避免大规模预训练，Teacher 默认不微调；仅允许极轻量的提示对齐/格式稳定（如少量规则或 LoRA <1% 参与度）。
- 阶段间 | 特征传输 (核心工作)
在 Teacher 与生成器之间引入可学习中间层 (Connector)，将 z_c（及 y_draft 的结构先验）映射为生成条件 h*。

Connector 可采用 MetaQuery / Adapter / Trans-Encoder 等，目标是低步数下的强条件对齐与稳定收敛。
- 阶段 B | 离散高速生成 (Student)
在条件 h* 下进行离散高速生成，采用离散扩散/并行掩码系列（如 Seed Diffusion / MeanFlow / D2F Diffusion），以 4–8 步非自回归生成动作 DSL 序列；解释由小型 LLM 解释头并行/异步补齐。

4.1 多模态信息编码（From Qwen2.5 Omni）

各模态处理

- 文本：Qwen 系 BPE 分词；直接作为 1D 序列供 LLM 消化。
- 音频：
 - 统一 16 kHz 重采样 → 128-mel 频谱（窗 25 ms / 步 10 ms）。

- Qwen2-Audio 编码器；时间轴分块注意力；约 40 ms/帧 的时间粒度。

• 图像/视频

- ViT 编码器 ($\approx 675\text{M}$)；patch=14。
- 2x2 token merge (MLP) 减少视觉 token。
- 视频使用动态帧率采样；静态图像视作两帧相同帧以与视频形态对齐。
- 支持分块/流式预填；内核加速（如 Flash-Attention）。

多模态对齐与编码（核心：TMRoPE）

- **TMRoPE (Time-aligned Multi-modal RoPE)**：将旋转位置编码拆为 Time / Height / Width 三路。
 - **时间对齐**：统一以 $40\text{ ms} = 1$ 个时间 ID；视频按真实时间推进；图像同帧内时间 ID 固定。
 - **空间编码**：视觉 token 按 H/W 网格编号；文本退化为 1D-RoPE。
- **序列拼接规则**：多模态 token 按块拼接，每种模态的起始位置 ID 为“上一个模态的最大 ID + 1”，避免冲突。
- **块内交错**：对“声画同源”的输入，每 2s 一块，块内按“视觉在前，音频在后”交错排列，保障 LLM 在同一注意力窗口内同时接收视听信息。
- **融合与推理**：LLM 直接在上述统一序列上做自回归推理，实现端到端跨模态融合与流式生成。

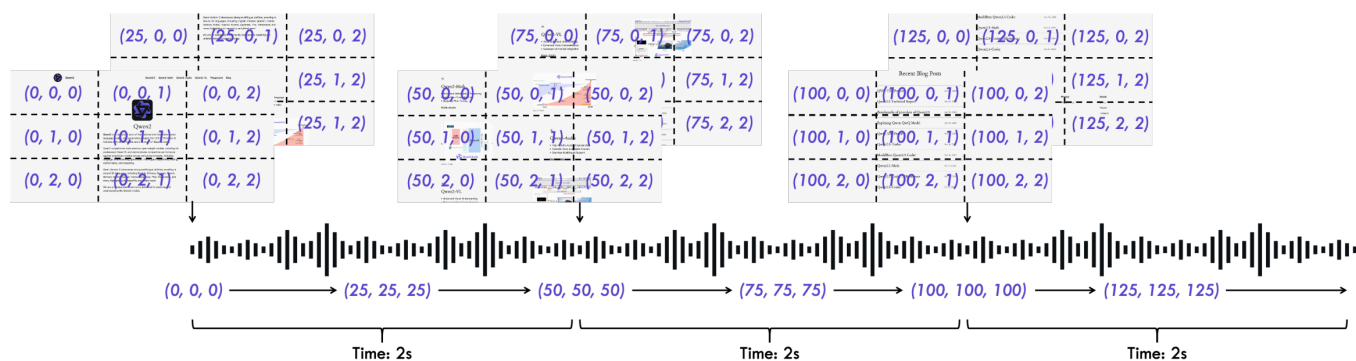


Figure 3: An illustration of Time-aligned Multimodal RoPE (TMRoPE).

4.2 特征传输（Connector）设计与训练

• 输入与输出范围：

支持两类视觉通路：① 场景摘要（caption）+ 车况信号；② 场景摘要 + 视觉潜特征（推荐）。Connector 将上述多源信息融合为生成条件表征，作为离散生成模块的标准输入。

• 结构设计建议：

- 采用 **MetaQuery** 等可学习查询机制（固定长度查询向量组），跨注意力聚合多源信息；
- 连接轻量投影/归一化层，产出定长条件表征，与离散生成接口无缝对齐；
- 兼容不同车型与版本的能力矩阵，提供可配置的词表与参数范围。

• 训练与优化目标：

- 一致性**：确保生成条件能稳定复现“被用户采纳的方案”（点击样本）；
- 规划对齐**：保留理解阶段给出的**行动草案**结构先验，减少无效探索；
- 平滑传输**：通过最优传输/流匹配（如 Sinkhorn-OT、Rectified-Flow、Noise-Free FM）约束，提升收敛稳定性与少步数可解性；
- 语义约束**：对越界参数、非法动作进行**硬屏蔽或惩罚**，确保与车型能力与规则一致。

- 训练流程：

先在冻结理解模块输出的离线数据上独立训练 **Connector**，打牢融合与对齐；随后与**离散高速生成模块**进行**联合微调**（理解模块继续冻结），逐步收敛至**低步数、高精度、强鲁棒**的在线推理路径。

4.3 离散高速生成

- 范式与步数：离散扩散/并行掩码（如 **Seed / MeanFlow / D2F**），**4–8 步**完成 ≤ 32 tokens 的动作 DSL 序列。
 - 目标：准确性对齐自回归，时延显著降低。
 - 推理：并行去噪/并行填充，结合置信度调度；解释由小型 LLM 解释头异步生成，不阻塞动作落地。
-

4.4 数据与蒸馏

- 数据来源：
 - i. 现有业务数据改造：将已有交互日志、车控操作轨迹和场景标签重新整理为 Observation → 候选方案的训练样本；
 - ii. 线上数据回流：采集用户真实交互数据（包含点击、拒绝、撤销），经统一清洗与脱敏后进入训练管线；
 - iii. 规则补充：利用规则引擎生成变体样本，并加入反事实样本作为负例，提升模型对约束和异常场景的鲁棒性。
 - 蒸馏路径：
 - 来自 Teacher 的规划与解释，蒸馏至 Connector 与小型解释头，提升一致性与生成稳定性；
 - 来自线上数据的偏好反馈（点击 / 拒绝 / 撤销），转化为奖励与排序信号，引导离散生成器优化输出质量与偏好对齐。
-

4.5 推理与 SLA

- 推理链路：

Observation → 冻结理解（VLM/Omni，产出场景表征与行动草案） → Connector（可学习中间层，生成条件表征） → 离散高速生成（产出 `action_plan`） → 规则/执行绑定校验 → （可选异步）解释返回。
- 性能目标（更贴近实际）：
 - 理解阶段（冻结 VLM/Omni）：P95 **400–800 ms**（含多模态编码与草案生成）；
 - **Connector**：P95 **10–30 ms**；
 - 离散生成（**4–8 步**）：P95 **80–180 ms**；
 - 后处理/校验/序列化：P95 **20–60 ms**；
 - 网络与排队开销（云侧内网 + 客户端RTT）：P95 **120–300 ms**；
 - 端到端（E2E）目标：P95 **0.9–1.3 s**。
- 时延与可用性约束：
 - 首 Token（TTFT）P95 ≤ 2.0 s；
 - 可用性：错误率 $\leq 1.5\%$ 。
- 达标手段：
 - 流式返回：优先返回 `title + action_plan`，`explanation` 异步补齐；

- 早触发工具调用 (**early tool-call**) 与 提示压缩/模板片段缓存;
- 多模态 **KV 缓存**、热点场景特征缓存 (caption/latent→embed);
- 小型解释头并行生成, 避免阻塞动作落地;
- 弹性并发与限速: Teacher 与生成器分层扩缩容, 峰值场景降级为保守单步方案。