Prompt-aware Rubric (PAR)：任务拆考点，最大化原子能力

Case：将视频中城下的所有红色兵种全部替换为蓝色兵种，丰富古代战场的场景细节，添加悬崖峭壁元素，保持原有视频的镜头运动、画面节奏和时长不变

BEFORE

传统奖励模型：固定且通用 (fixed)



- 生成的视频是否满足了用户的指令？给出1~5分

AFTER

PAR = 出题手册+出题模型+答题模型

1. answer=0 score=0.000000 [画面编辑][替换元素] 生成视频是否把城下的红色兵种替换为蓝色兵种?

2. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在远处增加了山脉?

3. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在古代战场场景中增加了树木?

4. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在天空中增加了云朵?

5. answer=1 score=0.952519 [画面编辑][场景背景] 生成视频是否在城墙左侧和右侧都增加了悬崖峭壁?

6. answer=1 score=1.000000 [未编辑内容保真][背景道具] 生成视频是否在完成兵种替换和场景细节新增后保持原视频中的城墙和其他未指定修改的战场元素不变?

7. answer=1 score=1.000000 [未编辑内容保真][镜头动作连续性] 生成视频是否在完成上述编辑后保持原视频的镜头构图、运动节奏和原有动作关系不变?

8. answer=1 score=1.000000 [编辑质量][替换融合] 生成视频中城下蓝色兵种的替换结果是否自然稳定，且没有红色兵种残留、闪烁或明显变形?

9. answer=1 score=1.000000 [编辑质量][替换融合] 生成视频中新增的山脉、树木、云朵和悬崖峭壁是否与古代战场场景自然融合，且没有明显拼接感、边缘穿帮或闪烁?

- 视频理解任务 = 原子考点拆分任务+二分类预测
- 多rubric分数 [token prob p(1)] 聚集
- hint -> 更好的context
- rubric增加可解释性

PAR项目负责：技术选型->构造数据->优化模型 ->上线迭代

01 出题手册

业务文档 + 真实RL场景

从业务文档提炼编辑任务核心评估维度与原子能力点，定义 rubric 出题规则

02 出题手册

自动化 Agent 迭代

构建 Agent 自动搜索优化，持续迭代更新 System Prompt

03 答题模型

构造 PAR 训练数据

基于 System Prompt 生成 prompt-aware rubric 标注训练数据

04 答题模型

训练 PAR 答题模型

训练 prompt-aware rubric 奖励模型，精确评估编辑效果

05 PAR上线

数据筛选 & RL 训练

PAR 用于数据粗筛 + 后训练

06 上线后迭代

交付两个版本

Single PAR （上线）
Agentic PAR V1 （上线）
Agentic PAR V4 （探索中）

韩雨瑞 8895

PAR 三版本：从单点模型到动态奖励系统

版本迭代

01 PAR（已上线）：单一答题模型

出题手册 + 出题模型 + 答题模型

02 Agentic PAR V1（已上线）
单一模型 扩展 多工具池

单点答题模型扩展为多工具池，适应上线后的新数据问题

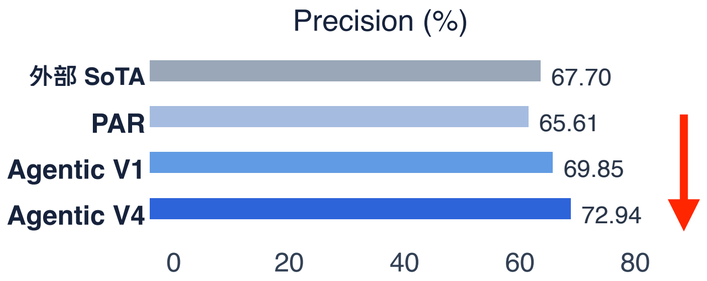
03 Agentic PAR V4（探索中）
多工具 + Harness 的奖励系统

Agent Brain + 多维动态 rubric Harness，形成动态奖励系统

Benchmark

Edit1500 评测集 Precision

外部 SoTA：出题模型 + Gemini-3-Pro-Preview + GPT-5.6-Sol



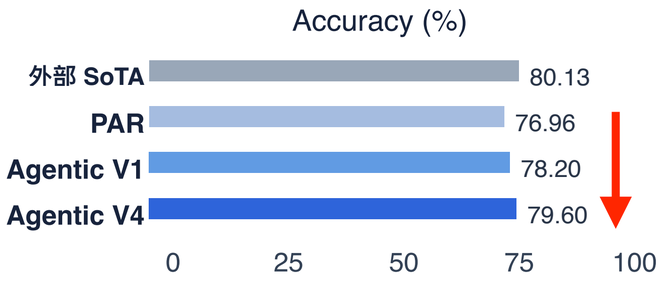
效果
V4 72.94% | 领先外部 SoTA +5.24pp

(预期)业务价值
进审量 ↓60%+ | 1000 → <400 条 | 人效翻倍

Benchmark

Edit1500 评测集 Accuracy

外部 SoTA：出题模型 + Gemini-3-Pro-Preview + GPT-5.6-Sol



效果
PAR → V4：+2.64pp | 76.96% → 79.60%

外部差距
V4 距外部 SoTA 仅 0.53pp

上线效果

百万数据真实业务场景（V1）



- 60.92% 严格可用
- 7.14% 可容
- 8.82% 不可用
- 23.12% 质量废弃

业务价值 | 数据可用率大幅提升 | 部分任务数据飞轮
可用 + 可容：31% → 68.06%

未来优化
质量废弃 23.12% 是主要卡点 (System 1 + System2)

口径：Edit1500 Benchmark；真实业务分布为严格可用 60.92%、可容 7.14%、可分 8.82%、质量废弃 23.12%

Prompt-aware Rubric (PAR)：任务拆考点，最大化原子能力

Case：将视频中城下的所有红色兵种全部替换为蓝色兵种，丰富古代战场的场景细节，添加悬崖峭壁元素，保持原有视频的镜头运动、画面节奏和时长不变

BEFORE

传统奖励模型：固定且通用 (fixed)



- 生成的视频是否满足了用户的指令？给出1~5分

AFTER

PAR = 出题手册+出题模型+答题模型

1. answer=0 score=0.000000 [画面编辑][替换元素] 生成视频是否把城下的红色兵种替换为蓝色兵种?

2. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在远处增加了山脉?

3. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在古代战场场景中增加了树木?

4. answer=1 score=1.000000 [画面编辑][场景背景] 生成视频是否在天空中增加了云朵?

5. answer=1 score=0.952519 [画面编辑][场景背景] 生成视频是否在城墙左侧和右侧都增加了悬崖峭壁?

6. answer=1 score=1.000000 [未编辑内容保真][背景道具] 生成视频是否在完成兵种替换和场景细节新增后保持原视频中的城墙和其他未指定修改的战场元素不变?

7. answer=1 score=1.000000 [未编辑内容保真][镜头动作连续性] 生成视频是否在完成上述编辑后保持原视频的镜头构图、运动节奏和原有动作关系不变?

8. answer=1 score=1.000000 [编辑质量][替换融合] 生成视频中城下蓝色兵种的替换结果是否自然稳定，且没有红色兵种残留、闪烁或明显变形?

9. answer=1 score=1.000000 [编辑质量][替换融合] 生成视频中新增的山脉、树木、云朵和悬崖峭壁是否与古代战场场景自然融合，且没有明显拼接感、边缘穿帮或闪烁?

- 视频理解任务 = 原子考点拆分任务+二分类预测
- 多rubric分数 [token prob p(1)] 聚集
- hint -> 更好的context
- rubric增加可解释性

PAR 优化：出题手册三步迭代构造策略

1

Step 1 · 评测业务经验搭建初步框架

固定要求 + rubric pool (anchor)，定义出题的 system prompt

2

Step 2 · RL 训练与推理获取反馈 [5大类 34细则]

用初步框架出题，通过 RL 训练和推理结果收集反馈

固定要求比如：要求出题模型出的Rubric

- **避免重复：**出题后需 self check，禁止把同一事实用多个类别重复提问、删除宽泛模糊定性的 rubrics
- **避免矛盾：**出题满足极性一致性约束
  - **1** = 生成视频更符合 要求
  - **0** = 不符合、未完成或证据不足
- **有限 Rubrics，重要优先：**不超过 12 条；超过则按“核心编辑目标 > 核心失败防护项 > …”重要性排序删减

韩雨琦 8895

PAR 优化：出题手册三步迭代构造策略

1 **Step 1 · 业务经验搭建初步框架**

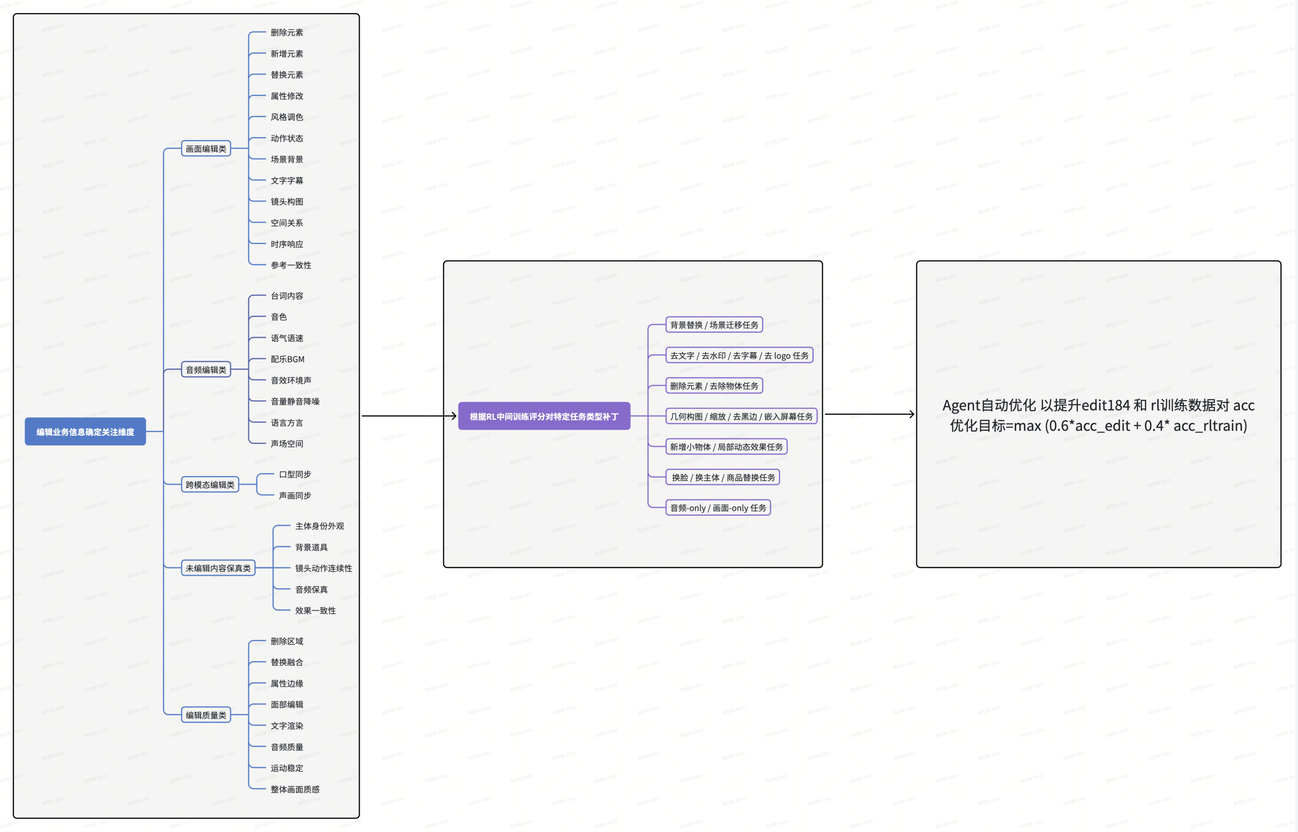
固定要求 + **rubric pool (anchor)**，定义出题的 system prompt

2 **Step 2 · RL 训练与推理获取反馈 [5大类 34细则]**

用初步框架出题，通过 RL 训练和推理结果收集反馈

Step 2 目的

同时考虑数据和训练场景下的标准，增加出题手册覆盖率



韩雨瑞 8895

PAR 优化：出题手册三步迭代构造策略

1 **Step 1 · 业务经验搭建初步框架**

固定要求 + rubric pool (anchor)，定义出题的 system prompt

2 **Step 2 · RL 训练与推理获取反馈 [5大类 34细则]**

用初步框架出题，通过 RL 训练和推理结果收集反馈

3 **Step 3 · AutoAgent 优化出题手册 像 SGD一样**

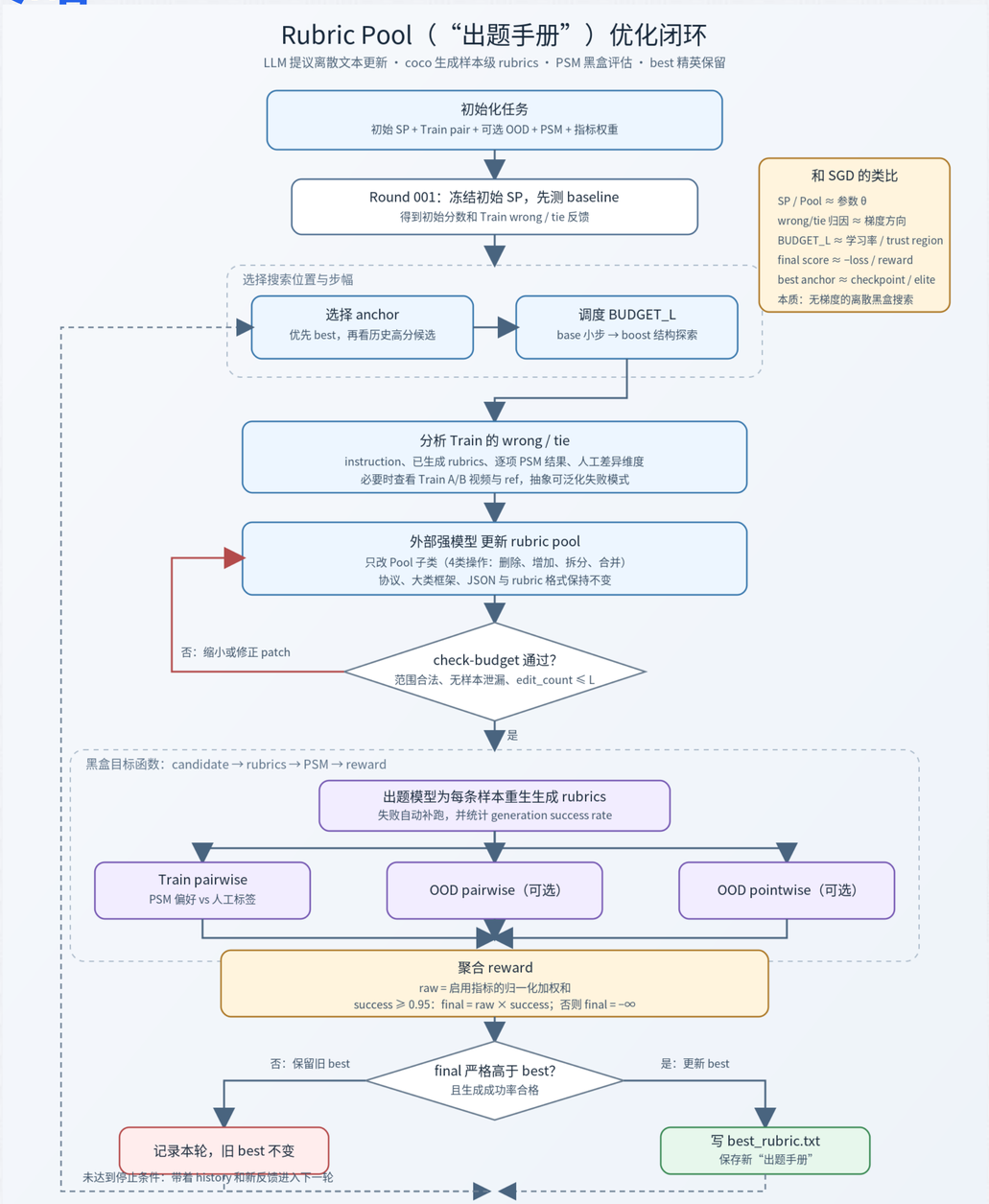
基于反馈自动优化 rubric generation prompt，形成闭环（**删除、增加、拆分、合并**）

Step 3 带来的 Rubric 变化 [5大类 35细则]

（1）每条 case 最大 rubric 数从 **12 → 14**

（2）属性/动作/文字/音频等 rubric **拆分更细**：五官、服装、道具、动作顺序与节奏、文字布局等

（3）新增 rubric：**[编辑质量][整体画面质感]**



PAR 优化：出题手册三步迭代构造策略

Step 1 · 业务经验搭建规则
固定要求 + rubric pool (anchor)，定义出题的 system prompt

Step 2 · RL 训练与推理获取反馈 [5大类 34细则]
用初步框架出题，通过 RL 训练和推理结果收集反馈

Step 3 · AutoAgent 优化出题手册 像 SGD一样
根据 eval benchmark 和 RL 训练结果，不断迭代 rubric generation sp （**anchor 4类操作：删除、增加、拆分、合并**）

效果对比（内部数据 Edit184 Accuracy）

Fixed Rubrics	0.636
Rubrics v0（业务文档）	0.712
Rubric v1 **Agent 搜索 + 训练优化，0/1 分**	**0.732**

无需训练： Accuracy 相对提升 **+10%**

增益来自范式： PAR 将 视频理解问题（困难） 拆分为 **带 Hints（rubric）的二分类问题（原子化）**

韩雨桐 8895

PAR优化：出题模型

94%

完全相关（5 分）

1000条训练样本 共 9728 条 rubric 人工标注

4.5%

部分相关（3 分）

1.3%

不相关（1 分）

结论：94% 的题目与 case 完全相关（精确率视角）

各大类 5 分占比（按 rubric 数量）



PAR优化：出题模型

训练模型	ChatGPT 5.5 出题	内部模型 (seed-2.0-lite) 出题
PAR ACC	**0.771**	**0.771**

评测集：内部数据 Edit 184，答题模型内部切换

Rubric 重合度分析

- 约**80%**样本 rubric 重合度超 60%
- 相同 rubric类别下，约**85%**问题语义基本一致

结论

出题策略是 Rubric 质量的决定性因素，足够详细易懂的出题策略使得，**出题模型的选择对最终 ACC 影响很小**。内部模型 可完全替代 ChatGPT 用于大规模出题。

韩雨桐 8895

PAR优化：答题模型

核心思路：答题模型本质为二分类模型。

GRPO （跳过SFT : 格式已掌握 + 数据量有限）: 每条样本问 N 条 rubric，模型输出长度为 N 的 0/1 字符串

三层 Reward 设计（format : accuracy = 1 : 9）

Format Reward 1

长度校验

输出位置数量必须等于 rubric 总数

少输出 / 多输出 → 整条 reward = -1

确保模型回答所有考点

Format Reward 2

格式校验

每个位置必须是 0 或 1

非数字字符 → 整条 reward = -1

其他数字 → 该位置 -1

确保输出可解析

Accuracy Reward

逐 rubric 平滑计分

预测正确 h_i = +1，错误 h_i = -1

p_i = 模型输出该位置的 token 概率

$r_i = 0.1 + 0.9 \times h_i \times p_i$

训练起点：内部模型

Rubric 维度：5 大维度 + 35 条细则，所有 rubric 等权

PAR优化：答题模型

exp1 · 基线 (无任何数据增强)

训练数据：约1.1w条数据

全 1 占比 4.99%，全 0 占比 0.4%

混合 0/1 占比 94%

分布严重不均衡

问题：模型学到 bias 捷径「多 rubric 通常含 0」，OOD acc 下降

OOD 最佳Step

77.1%

exp3 · augv1（增强方案1）

增强：删除 GT0 对应 rubric

保留 3~8 个 GT1 rubric，打乱顺序

总样本 1.1w → 1.5w

全 1 占比 4.99% →**31.85%**

效果：预测 1 增多，改善 FN（假阴性），关闭 thinking

OOD 最佳Step

80.7%

exp4 · augv2（增强方案2）

增强：进一步消除伪相关性

总样本 1.1w → 2.2w

全 1 / 混合 / 全 0 = 30% / 50% / 20%

平衡 rubric 长度与 01 分布

效果：预测 0 增多，改善 FP（假阳性），0 的准确率提升明显

OOD 最佳Step

81.4%

增益来自和评测专家对齐

数值为 ACC，评测集：内部数据 Edit 184

PAR优化：答题模型 GRPO算法优化

方法流程（per rubric）

① 正常 GRPO（8 次 rollout）

每个 prompt rollout 8 次，逐 rubric 计算 reward

有对有错 → 组内标准化得到 advantage

↓

② 触发检测：8 次全错?

Σ 1[y_i,j = y_j*] = 0 → 该 rubric 进入 Rescue

有 ≥1 次正确 → 不触发，沿用正常 GRPO

↓

③ Gold-Rescue 注入

虚拟加入 1 条 Gold 正确答案 → 共 9 个 reward

[0,0,0,0,0,0,0,0, 1] 标准化后：

错误 rollout A = -1/3，Gold token A = +8/3

核心公式

正常 Rubric Reward

$$R_{i,j} = 0.1 + 0.9 \cdot h_{i,j} \cdot p_{old}(y_{i,j})$$

普通 Rollout PPO Clipped Loss

$$r_t(\theta) = \exp(\log \pi_\theta - \log \pi_{old})$$

$$L_{PPO} = \max(-A_t \cdot r_t, -A_t \cdot \text{clip}(r_t, 0.8, 1.28))$$

Gold-Rescue Loss（π_teacher = 1）

$$p_t = \pi_\theta(y_t^* \mid x, y_{<t}^*)$$

$$L_{gold,t} = -A_t^{gold} \cdot p_t / (p_t + \gamma),\ \gamma = 0.1$$

韩雨桐 8895

PAR 优化：答题模型 GRPO 算法优化

训练曲线



评测效果 · 相比 PAR augv2 的提升

内部数据 1 · Edit184（画面编辑评测）

PAR augv2	PAR augv2_rescue
81.4%	82.7%

ACC +1.3pp

GRPO 算法通过 rescue 机制优化答题模型，
在无需额外训练数据的情况下实现准确率提升。

PAR：编辑任务OOD benchmark EditResponse154 评测1

模型	总分	响应分	一致性	质量分
Baseline fixed sp	0.474	0.630	0.747	0.727
PAR 出题优化 + 答题训练	**0.747** +27.3pp	**0.857** +22.7pp	**0.805** +5.8pp	**0.734** +0.7pp
外部 SoTA Gemini-3-Pro-Preview	0.630	0.662	0.792	0.792

PAR 相较 Baseline：总分 +27.3pp · 响应分 +22.7pp · 一致性 +5.8pp

相较外部 SoTA：总分 +11.7pp · 响应分 +19.5pp · 一致性 +1.3pp（质量分 –5.8pp）

韩雨瑞 8895

PAR：OOD benchmark Edit1500 采样自真实的编辑数据场景

模型	Precision	Recall	后续环节数据进审量 (baseline 31%)
PAR 出题优化 + 答题训练	**67.6%** –0.1pp	**76.6%** –3.7pp	**下降 54.1%**
外部 SOTA 出题优化 + Gemini-3-Pro-Preview + GPT-5.6-Sol	67.7%	80.3%	下降 54.2%

PAR 与外部 SoTA 表现接近：Precision 仅差 0.1pp，Recall 差 3.7pp

内部训练方案已逼近外部多模型组合

韩雨瑞 8895

PAR数据侧评测

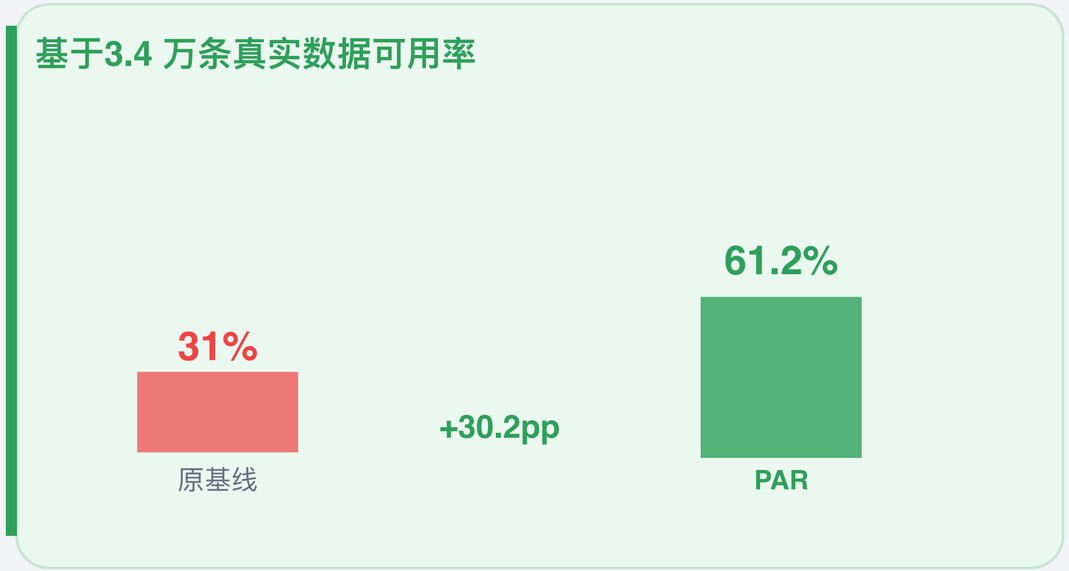
模型 （内部数据 Edit1500）	Precision	Recall
PAR augv2_rescue (出题模型 + 内部训练答题模型)	67.6% -0.1pp	76.6% -3.7pp
外部SOTA （出题模型 + Gemini-3-Pro-Preview + GPT-5.6-Sol）	67.7%	80.3%

黄雨琦 8895

PAR：OOD benchmark Edit1500 采样自真实的数据场景

模型	Precision	Recall	后续环节数据进审量
PAR 出题优化 + 答题训练	**67.6%** –0.1pp	**76.6%** –3.7pp	**下降 54.1%**
外部 SoTA Gemini-3-Pro-Preview + GPT-5.6-Sol	67.7%	80.3%	下降 54.2%

PAR：上线效果



PAR：可视化

编辑 Prompt

将从第9秒开始，女人的台词 `{这不是冬天到了吗？买一些菊花茶，泡脚包之类的。}` 改为 `{我给你买了件新外套，天冷了穿。}`，并把男人接下来的台词 `{花多少钱？}` 改为 `{是吗？真贴心。}`，同时确保口型同步。

Human Remark

台词改对了，口型对不上

答题模型准确归因，识别错误点：台词对 · 口型错

Main PSM 回答串：11001111111　平均分：0.835515　Rubric 通过：9 / 11

Reference video 1



打开视频

Target / edited video



打开视频

▼全部 Rubric 模型输出（11 条）

#	Rubric	回答	得分	P(0)	P(1)
1	[音频编辑][台词内容] 生成视频从约第9秒开始，女人的台词是否已从“这不是冬天到了吗？买一些菊花茶，泡脚包之类的。”替换为“我给你买了件新外套，天冷了穿。”?	1	0.892937	0.107059	0.892899
2	[音频编辑][台词内容] 生成视频中，在女人说完新台词后，男人的台词是否已从“花多少钱啊？”替换为“是吗？真贴心。”?	1	0.999820	0.000180	0.999275
3	[跨模态编辑][口型同步] 生成视频中，女人说出新台词“我给你买了件新外套，天冷了穿。”时，其口型是否与新语音整体同步，无明显提前或滞后？ 与人工一致	0	0.293404	0.706342	0.293299
4	[跨模态编辑][口型同步] 生成视频中，男人说出新台词“是吗？真贴心。”时，其口型是否与新语音整体同步，无明显提前或滞后？ 与人工一致	0	0.005248	0.994674	0.005248
5	[未编辑内容保真][音频保真] 生成视频在约00:09秒之前的全部原始对话内容是否与输入视频<视频1>保持一致?	1	0.999983	0.000017	0.999956

出题模型类别细致 原子化问题

PAR用于RL (Seedance2.5 & DiffusionNFT)

变好的 Case（编辑响应提升）

内部64 题 PAR=12 / fixed=7

提升主要集中在**编辑响应更准确**，模型对编辑指令的遵循度显著提高。

内部89 题 PAR=11 / fixed=3

运镜、音频编辑从不响应到完全响应；同时编辑响应更准确。

Case1: 细节编辑



Base fixed PAR

PAR用于RL (Seedance2.5 & DiffusionNFT)

变好的 Case（编辑响应提升）

内部64 题 **PAR=12 / fixed=7**

提升主要集中在**编辑响应更准确**，模型对编辑指令的遵循度显著提高。

内部89 题 **PAR=11 / fixed=3**

运镜、音频编辑从不响应到完全响应；同时编辑响应更准确。

Case2：风格编辑



Base fixed PAR

更好的ACC：PAR利用模型原子能力+偏好对齐，提升acc;

更高 ACC → 更好编辑响应

韩雨雨 895

PAR上线后迭代：从单一Verifier到Agentic Verifier

韩雨锦 8895

PAR上线迭代: 训练数据外的新问题

问题背景

真实场景下：新数据分布暴露single verifier中的答题模型在**背景声音一致性、编辑质量** 缺陷

FP

GT 0 · Pred 1

e3d9ed78-1f51-4754-a19c-506e34ace019_A-add_o2_A-rep_o1_s5_0003899

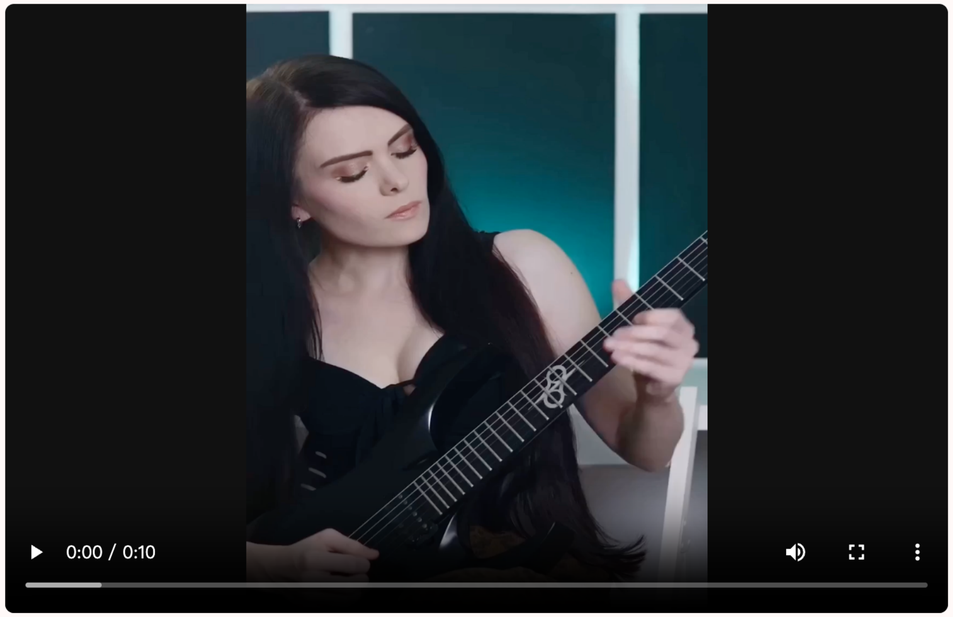
编辑前后音频识别失败 -> 弱能力定点补强

编辑 Prompt

把画面中弹吉他的黑发女子替换成一个穿着黑色T恤的长发男人，并且在他身后增加一个黑色的大音箱。

任务标签：新增元素、替换元素 | mean score：0.9987 | threshold：0.9965 | 响应 GT：1 | 一致性 GT：0 | 质量 GT：1

Source media



Target / model video



Agentic V1：工具引入优化PAR (已上线)

引入算子工具: 补强答题模型

单一答题模型 → 多工具池

Step 1
当前 Agentic Verifier 推理

Step 2
业务解释 + 错误 Case 分析

Step 3
调整路由策略、引入算子

Step 4
新 Agentic Verifier 推理

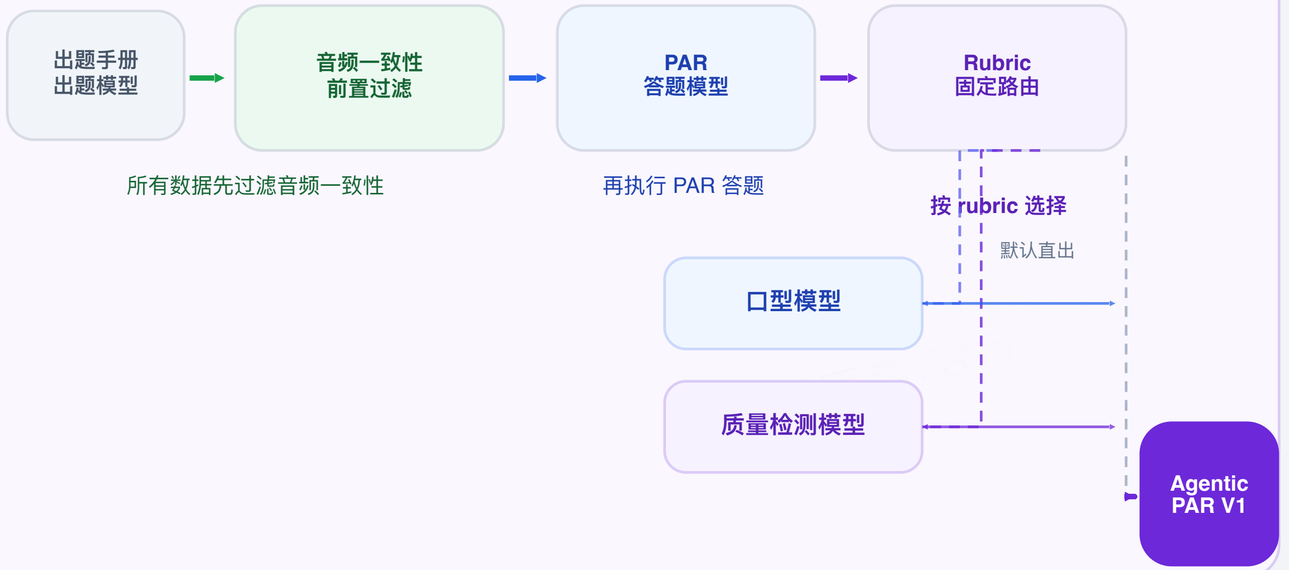
Step 3 展开 · 算子发现与验证

1. **构建 Benchmark** — 50 异常 + 50 正常
2. **强模型调用候选算子** — ChatGPT 5.6 Sol
3. **双指标验证** — 高识别率 · 低误杀率
4. **筛选上线算子** — 纳入固定路由

音频一致性 · visual_only

波形相关	频谱余弦	平均相干	残差 SNR
≥0.15	≥0.65	≥0.15	≥0 dB

固定路由 · 单一 PAR → Agentic PAR V1



Agentic Verifier V1： 多专家工具引入优化PAR (已上线)

引入算子工具: 补强答题模型

单一答题模型 → 多工具池

Step 1
当前 Agentic Verifier 推理

Step 2
业务解释 + 错误 Case 分析

Step 3
定位rubric，引入算子

Step 4
新 Agentic Verifier 推理

PAR：OOD benchmark Edit1500 采样自真实的数据场景

模型 （内部数据 Edit1500）	Precision	Recall	后续环节数据进审量	Acc	F1
PAR 出题优化 + 答题训练	67.6%	76.6%	下降 54.1%	76.96%	71.76%
Agentic PAR V1 出题优化 + 答题训练 + 多专家工具	**70.7%** +3.1pp	**74.4%** −2.2pp	**下降 56.2%** +2.1pp	**78.20%** +1.24pp	**72.06%** +0.30pp
外部 SoTA Gemini-3-Pro-Preview + GPT-5.6-Sol	67.7%	80.3%	下降 54.2%	--	--

Agentic Verifier V4：PAR + 多工具池 + Agent Brain路由 （探索中）

第一步：工具池扩充: 重复Agentic V1版人类的分析流程



第二步：Agentic Verifier V4 脚手架框架（三阶段 Harness）

1 Rubric 回扫

Gap Review Brain 检查遗漏 rubric，补全 synthetic rubric（台词/背景音/响度保真）

2 多工具编排

工具和 PAR 绑定 · 无需收集 tool call 数据训练工具调用

2 训练模型 | Gemini API | 10 本地算子

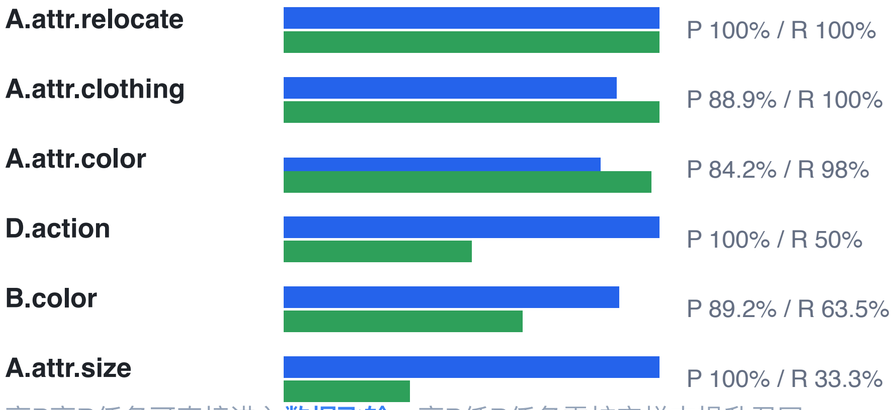
不同 rubric 绑定不同工具，逐 rubric 取证据

3 Brain 逐 rubric 判断

证据 bundle 综合决策；冲突时 Conflict Brain 独立裁决

效果验证：6 个任务 Precision ≥85%

高P高R占11% · 全部6项占17%



高P高R任务可直接进入**数据飞轮**；高P低R任务需扩充样本提升召回

PAR三版本效果

模型 （内部数据 Edit1500）	Precision	Recall	后续环节数据进审量	Acc	F1
Single PAR 出题优化 + 答题训练	67.6%	76.6%	下降 54.1%	76.96%	71.76%
Agentic PAR V1 出题优化 + 答题训练 + 多专家工具	70.7%	74.4%	下降 56.2%	78.20%	72.06%
Agentic PAR V4 出题优化 + 答题训练 + 多专家工具 + workflow	**72.94%** **相对 V1 +2.24pp**	**74.6%** **相对 V1 +0.2pp**	**下降 57.5%** **相对 V1 +1.3pp**	**79.60%** **相对 V1 +1.40pp**	**73.74%** **相对 V1 +1.68pp**
外部 SoTA **Gemini-3-Pro-Preview + GPT-5.6-Sol**	67.7%	80.3%	下降 54.2%	--	--

韩雨瑞 8895

PAR三版本效果

模型 （内部数据 Edit1500）	Precision	Recall	后续环节数据进审量	Acc	F1
Single PAR 出题优化 + 答题训练	67.6%	76.6%	下降 54.1%	76.96%	71.76%
Agentic PAR V1 出题优化 + 答题训练 + 多专家工具	70.7%	74.4%	下降 56.2%	78.20%	72.06%
Agentic PAR V4 出题优化 + 答题训练 + 多专家工具 + workflow	**72.94%** 相对 V1 +2.24pp	**74.6%** 相对 V1 +0.2pp	**下降 57.5%** 相对 V1 +1.3pp	**79.60%** 相对 V1 +1.40pp	**73.74%** 相对 V1 +1.68pp
外部 SoTA Gemini-3-Pro-Preview + GPT-5.6-Sol	67.7%	80.3%	下降 54.2%	--	--

三版本提升来自哪里

Single → V1 提升来源

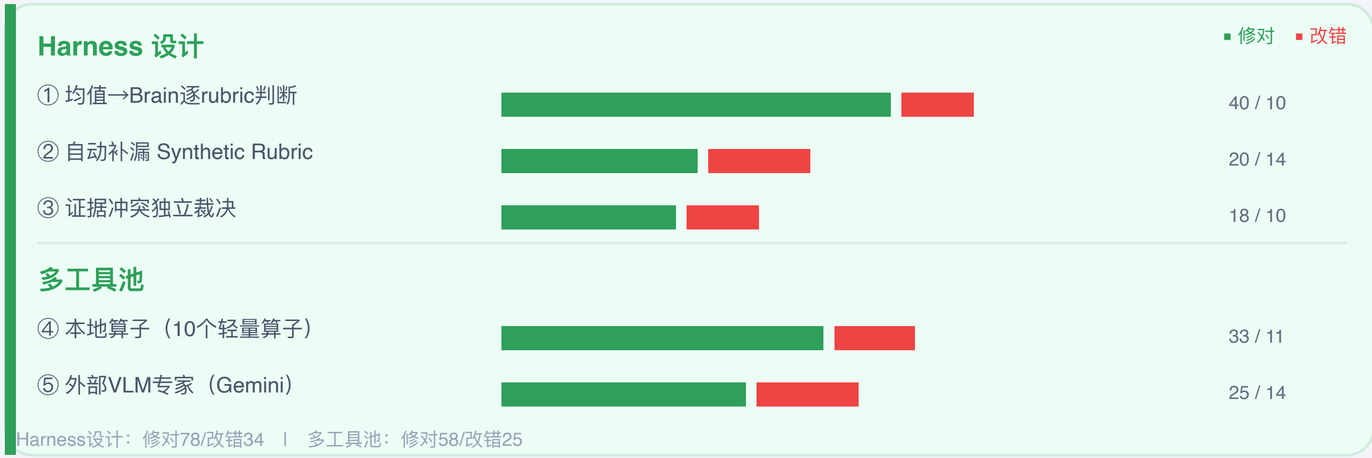
35 个

音频一致性算子识别不一致 case

24 个

质量专家模型识别低质量 case

V1 → V4 提升来源



PAR 提升来源 & 局限 & 未来优化

韩雨锦 8895

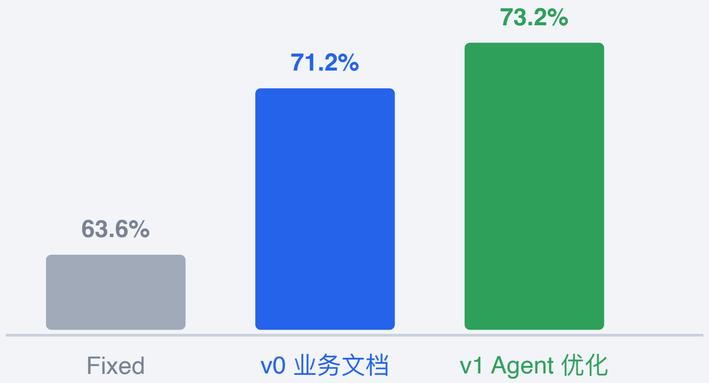
PAR 提升的两个来源

更好的 context 激发答题模型原子能力 × 专家偏好对齐答题模型训练

1 激发 VLM (答题) 原子能力

出题模型结合业务知识，用 rubrics 作为 hints；答题模型只需聚焦原子化问题（颜色更改、指定区域增删）；
更好的context更好激活基座能力。

实验证据：通过优化答题手册优化context (Edit184)



优势 · Scaling
PAR 上限随基座模型原子能力提升而提升

局限 · 受限于基座
细小崩坏识别、局部音画同步等原子能力不足

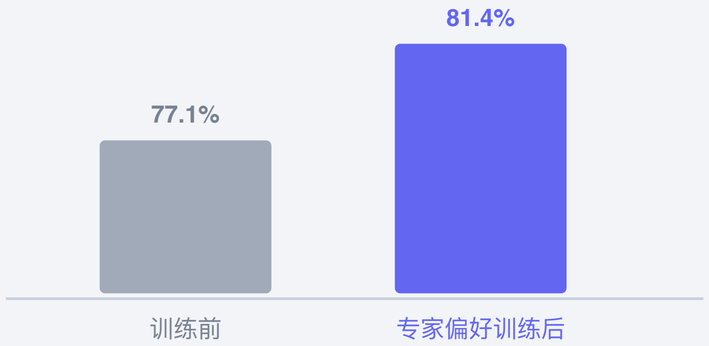
目前策略 · Agentic V1/V4
引入外部算子，rubric定位基座弱势的能力，定向补全

未来优化方向 · 激发 VLM 能力·答题模型不只是二分类
-针对细节识别问题使用带 Crop 工具的思维链（附录快速实验）
-训练策略优化（e.g., 附录Jev-RLCD实验）

2 与评测专家对齐

以专家偏好为监督训练答题模型，让模型判断口径对齐专家评测标准。

实验证据：答题模型 ACC（↑4.3pp）Edit184)



偏好对齐
答题模型学到专家评测口径，判断与专家一致；ACC 77.1% → 81.4%。

韩雨瑞 8895

基于PAR的思考：视频生成模型需要怎样的奖励环境?

一种可能：PAR-based Agentic Verifier —— PAR 是业务专家提供的 hints，也是 agentic verifier 的脚手架

01 定义完备、结果可验证

为什么 LLM 先在数学、代码突破?

定义清楚，且能获得明确、可验证的奖励信号——形式化证明、程序执行、数值仿真。

视频生成缺类似环境：

美学好不好、打2分还是3分、符合偏好吗——定义模糊，难稳定验证。

PAR 先把问题定义清楚：

- 出题手册：明确业务关注的任务维度
- rubric 作为判断标准和推理提示

PAR 把模糊质量问题转化为**可拆解、可解释、可验证**的**二分类**问题。

02 适应动态变化的环境（基座能力有限）

奖励环境不是静态的：

- **数据筛选**：数据分布不断变化
- **评测场景**：标注标准会变化
- **RL 场景**：不同训练阶段关注的问题不同 (响应->保真)

迭代机制：

- 根据PAR，结合标注解释与错误案例，定位薄弱rubric能力
- 训练二分类模型 + 引入专家算子
- 通过扩充工具池，迭代 verifier

不是**固定的 reward model**，而是一套随**数据、任务和模型阶段持续演化**的奖励环境。

03 可迁移可伴随基座Scaling

一套体系覆盖三大场景：

- **数据筛选**
- **RL 训练**
- **评测**

统一流程：出题手册 → 出题 → 答题

好的Context激发VLM原子能力

输入数据分布不同，但 PAR 体系可复用迁移。

区别：

- 答题模型输入数据的分布

核心结论：PAR-based Agentic Verifier = 可验证（定义完备）+ 可演化（动态适应）+ 可迁移（体系复用）

它把视频生成理解问题 拆分为 一个 文本理解（业务知识）+ 带暗示的视频理解问题（二分类），并且"可验证奖励 + 持续演化 + 跨场景复用"，为视频生成模型的 RL 训练提供稳定且可进化的奖励环境。

附录效果：PAR 三版本详细指标 Edit1500 Benchmark

PAR版本	架构	Precision	Recall	数据进审量下降	Acc	F1	备注
PAR	出题手册 + 出题模型 + 答题模型	65.61%	**79.17%**	54.1%	76.96%	71.76%	标准 PAR 范式 出题手册、出题模型、答题模型
Agentic PAR V1	出题手册 + 出题模型 + 答题模型 / 专家算子 + 固定路由	69.85%	74.40%	56.2%	78.20%	72.06%	Agentic PAR 多个专家算子补强答题模型，提供更多可信证据 路由规则固定
Agentic PAR V4 **（Gemini 大脑）**	出题手册 + 出题模型 + 多工具池 + Gemini 大脑路由	**72.94%**	74.56%	**57.5%**	**79.60%**	73.74%	Agentic PAR：Agent Brain 逐 rubric 综合证据 4 个 VLM 专家 + 10 个本地轻量算子； Harness 设计支持工具扩展与证据聚合
外部 SoTA	出题手册+出题模型+Gemini-3-Pro-Preview + GPT-5.6-Sol	67.7%	80.3%	54.2%	--	--	Gemini-3-Pro-Preview 将音频转为文字作为context输入给GPT5.6 sol

附录： PAR数据侧评测： 内部数据 1 & 内部数据 2

模型	内部数据 1 Edit184（ACC）	内部数据 2 Audi86（ACC）
fixed sp zeroshot	0.625	0.672
PAR zeroshot	0.781	0.683
PAR augv2_rescue（GRPO Rescue）	**0.827** +20.2pp	**0.732** +6.0pp

韩雨瑞 8895

附录：PAR用于RL (Seedance2.5 & DiffusionNFT)

变好的 Case（编辑响应提升）

内部64 题 **PAR=12 / fixed=7**

提升主要集中在**编辑响应更准确**，模型对编辑指令的遵循度显著提高。

内部89 题 **PAR=11 / fixed=3**

运镜、音频编辑从不响应到完全响应；同时编辑响应更准确。

思考与结论

PAR带来更好的ACC：PAR利用模型原子能力+偏好对齐，提升acc。

更高 ACC → 更好编辑响应：PAR 更高的 acc 体现为下游任务更好的编辑响应（与 diffusionNFT 这类梯度和 reward 成正比的方法相关）。

合训无奖励冲突：编辑奖励能上升，画面编辑维度融入音频编辑也能提升音频任务（更多任务能否通过 PAR 融合需进一步验证）。

Trade-off 仍存在：编辑响应增加的同时，编辑/参考一致性变差，可考虑rubric加权。

更多验证：基于Seedance 3.0

Case1: 发型一致性变差



Base fixed **PAR**

Case2: 字幕保持变差



Base fixed **PAR**

附录：PAR下一步优化方向 - 低编辑质量识别

做不好的归因分布

归因	数量	备注
Rubric 缺失	**15**	Rubric 没问到相关
含 Rubric 但工具判断错误	**162**	Rubric 存在，但工具 力弱为 V4 主要卡点

核心结论

- **87.1%**（162 条）：已含 Rubric，但工具判断错误
- **工具能力弱 = V4 主要卡点**
- 仅 **4.8%** 为 Harness 聚合失效
- Rubric 缺失占 8.1%

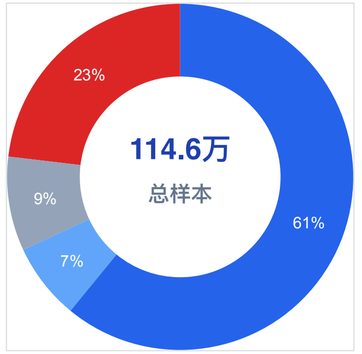
见 fail，但
聚合未生

工具池待优化：Accuracy 高、Recall 低

问题（占比）	对应工具表现（Acc / Reca
画面质量崩坏（24.6%）	主PSM 86.48% / **R11.01%** Quality PSM 69.61% / **R26** Ark Artifact 86.60% / **R7.22**
未编辑内容保真（12.83%）	主PSM **R5.77%** Ark Preservation **R5.00%**
背景/风格/光照/材质（13.37%）	主PSM **R29.07%** Ark Color Style **R20.83%** Quality PSM **R15%** Ark Artifact **R7.89%** rvation **R11.11%**
局部响应/时段覆盖（13.37%）	主PSM **R32.14%** Ark Completion **R30.77%**

Recall = 问题检出率：上述类别 Recall 多低于 30%，**工具普遍偏向 pass**、漏检明显

百万数据真实业务场景缺陷



- 1分　697,937（60.9%）
- 0.5分　81,810（7.1%）
- 0分　100,952（8.8%）
- 废弃　264,922（23.1%）

影响PAR precision的主因是因为低编辑质量导致的废弃

理想情况：识别编辑质量问题后，PAR precision 可到79.2%

附录：PAR进一步优化 — 质量崩坏问题

检测质量崩坏的两个核心问题：**崩坏人能否被模型发现?**

问题 1：崩坏是细节区域

全局画面看似正常
崩坏隐藏在局部细节 - 头发突起
竖版 720×1280 · 共 217 帧
截取连续帧 15–19



连续帧 15 / 16 / 17 / 18 / 19 —— 细节崩坏需引入 CoT + crop 工具定位

问题 2：均匀采样 FPS 能否覆盖崩坏?

崩坏仅闪现 1 帧即恢复
均匀采样极易漏掉
横版 1280×720 · 共 257 帧
崩坏发生在第 **192** 帧



第 191 帧（正常）



第 192 帧（崩坏）



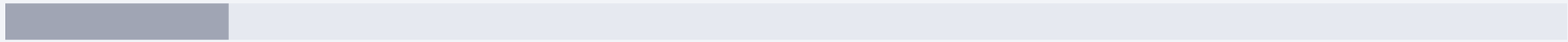
第 193 帧（正常）

附录：PAR进一步优化 — 质量崩坏问题

快速实验：**63 个明确被认定为编辑崩坏的 case**，识别率对比

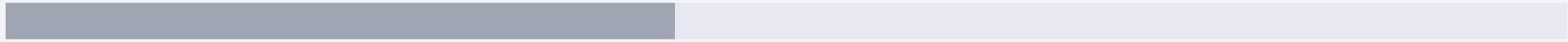
基线（无 CoT）

(1) 答题模型

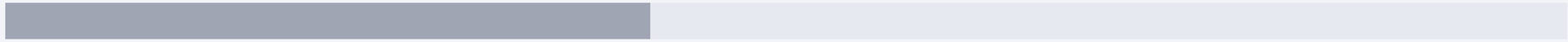


(2) 内部算子

(2A) 质量 rubric 模型



(2B) 基础维度崩坏检测



问题 1：崩坏是细节区域 → 引入 调用Crop工具的CoT

(3) 5 FPS · 无 CoT



(4) 5 FPS · CoT



对比：**57.1% → 63.5%（+6.4 pp）**

结论：CoT 引入可帮助细节区域的崩坏识别

问题 2：均匀采样 FPS 能否覆盖崩坏?

(4) 5 FPS · CoT



(5) 人工给帧 · CoT



对比：**63.5% vs 60.3%**

结论：5 FPS 已足够覆盖真实崩坏帧（人工给帧无增益）

总结论：质量崩坏优化方向为在高 FPS（5 FPS）下，提升 CoT 模型ACC

附录：Jev-RLCD 概率矫正方法调研与复现

Jev 官方未开源 RLCD 方案，基于 **Brier Score** 复现；在画面编辑 1500 benchmark 上验证

实验效果：标准 GRPO vs GRPO-RLCD

指标	标准 GRPO	**GRPO-RLCD**
Precision	68.91%	**71.56% (+2.7pp)**
Recall	75.04%	74.40% (-0.6pp)
F1	71.84%	**72.95% (+1.1pp)**
Accuracy	75.33%	**76.87% (+1.5pp)**

结论：RLCD 引入后 **Precision +3pp**，Recall 基本不变；
减少虚高的 rubric score = 0.999 预测

核心结论：抑制 GRPO 的过度自信

1. **Brier Score 矫正概率**：reward = (y - p)
   减弱对高置信正确预测的继续推高，更强惩罚高置信错误

2. **实验现象：rubric 输出分数整体变小**
   rubric 均值：0.803 → 0.775
   极端自信区间 [0.95, 1] 占比：75% → 67%

3. **理论解释：RLCD 稳定点 p = q**
   预测概率 p 朝真实条件概率 q=P(Y=1|x) 收敛，
   而不像标准 GRPO 把概率持续推向 0 或 1

意义：概率矫正让 rubric score 更可信，可直接用作奖励信号