

面向多模态大模型幻觉缓解的模糊语义自反馈学习

RLSF-V: Mitigating Hallucinations in MLLMs via Fuzzy Semantic Self-Feedback



何长浩

四川大学

2026.08.13

HUAWEI分享·黄大年茶思屋

从一个典型幻觉开始

图像输入



模型输出

A dog is **catching a red frisbee** in the park.

视觉事实核对

对象: dog	✓ 有证据
场景: grass / park	✓ 有证据
动作: catching	✗ 无证据
物体: red frisbee	✗ 无证据

表述自然流畅 **≠** 与视觉事实一致



什么是多模态幻觉？



生成文本与图像证据、输入上下文或常识事实相冲突



1 存在性

例：图中有一只猫。✗



2 属性

例：狗的毛是黑色的。✗



3 状态

例：天气是下雨的。✗

视觉事实（证据）



晴天



公园



草地



狗



4 数量

例：图中有两只狗。✗



5 动作

例：狗在吃东西。✗



6 关系

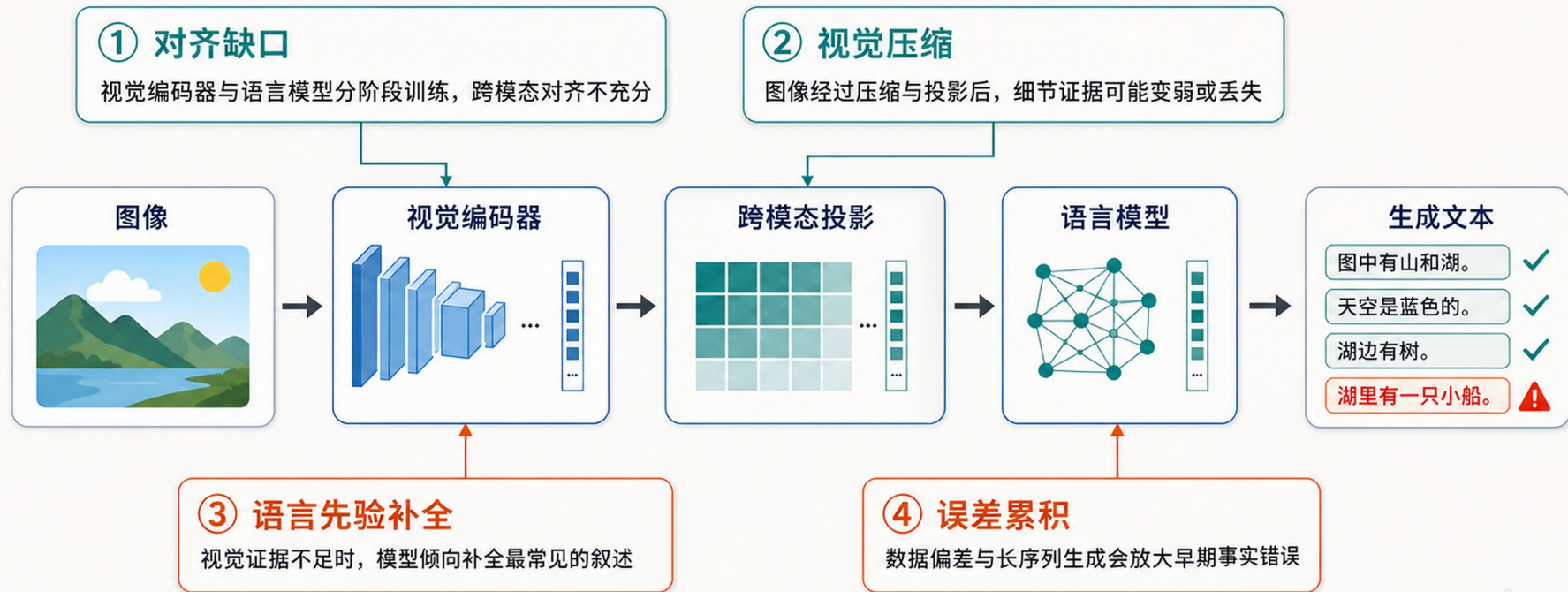
例：狗被人牵着。✗



回答可以很流畅、很合理，却没有视觉依据

流畅 $\neq$ 真实

为什么 MLLM 更容易“看着图说错话”？



流畅 ≠ 视觉忠实

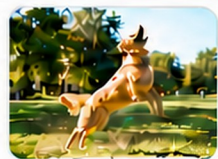
结构性缺口让合理语言覆盖真实图像证据

路线一：Training-free 的推理时抑制

背景 · 技术路线

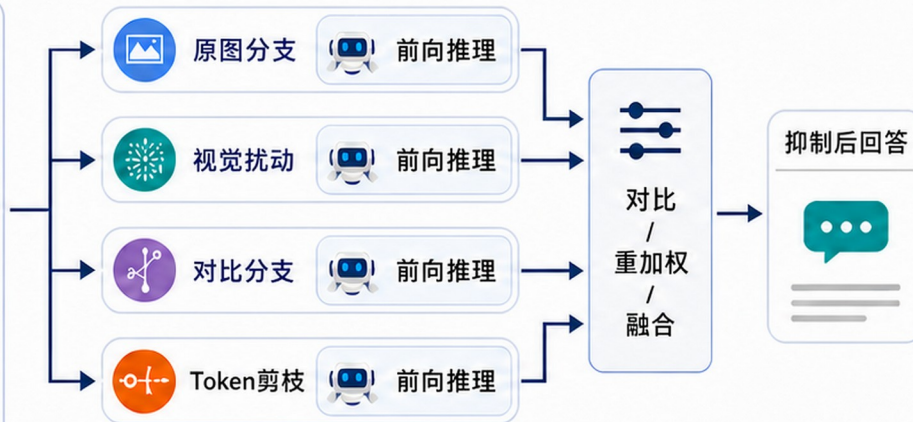
对比解码 | 输入扰动 | 视觉Token剪枝 | 注意力/激活控制

单次请求



Prompt

"Describe the dog's action in this image."



多次前向 + 额外分支 → 在线时延、显存↑



为什么有吸引力

无需重新训练 · 可直接插入现有模型



每次查询都要付费

多分支计算 · 延迟与显存开销重复发生



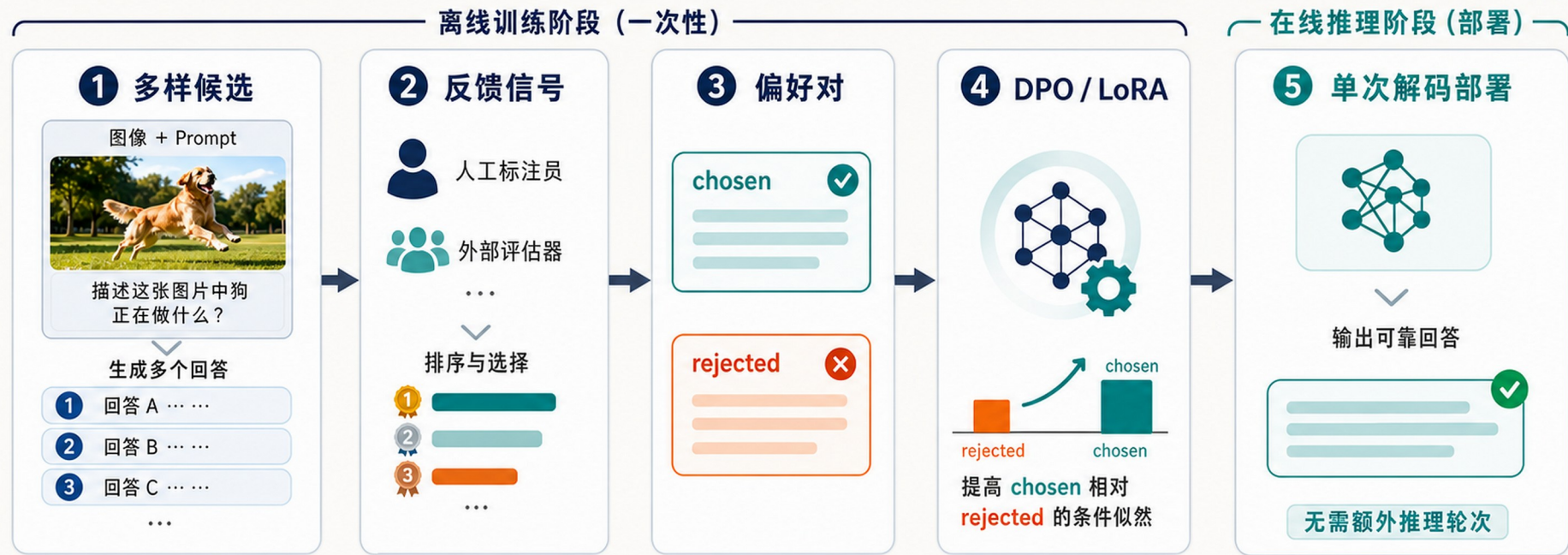
根本局限

超参数与提示敏感，稳定性依赖场景
只修正当前解码，基础策略没有被永久校准



问题被留在推理阶段：成本无法在训练阶段摊销

路线二：训练对齐把可靠性写进模型



训练成本离线支付，可靠行为写入模型



真正的瓶颈：谁来可靠、低成本地构建偏好对？

高质量偏好数据为什么昂贵？

三类外部监督来源各有局限，导致获取成本高、可扩展性差

✓ 人类反馈：可靠，但成本高、扩展慢

✓ 闭源模型改写：API成本、隐私与可复现性受限

✓ 更强MLLM/VLM评估：系统依赖转移给外部教师

✓ 新backbone往往要重做评估流水线

1 人类反馈 (Human)



- 标注 / 排序
- 编辑 / 改写
- 质量判定

成本



非常高

依赖



人力密集

隐私



风险可控

2 闭源模型改写 (Proprietary Model)



- 生成改写 / 排序
- 幻觉评分
- 偏好判断

成本



API 费用高

依赖



供给不稳定

隐私



数据外流风险

3 更强 MLLM / VLM 评估 (Stronger MLLM/VLM)



- 作为教师评分
- 作为参考排序
- 幻觉检测

成本



部署+算力高

依赖



性能上限转移

隐私



数据共享需求

! 常见问题

- ✓ 成本高昂，难以规模化
- ✓ 隐私与安全风险
- ✓ 跨领域 / 跨任务难以泛化

偏好数据池

(chosen / rejected 对)



! 可扩展性瓶颈

- ✗ 数据获取速度 << 模型迭代速度
- ✗ 每换一个 backbone 需重做流水线
- ✗ 长期维护与对齐成本持续上升

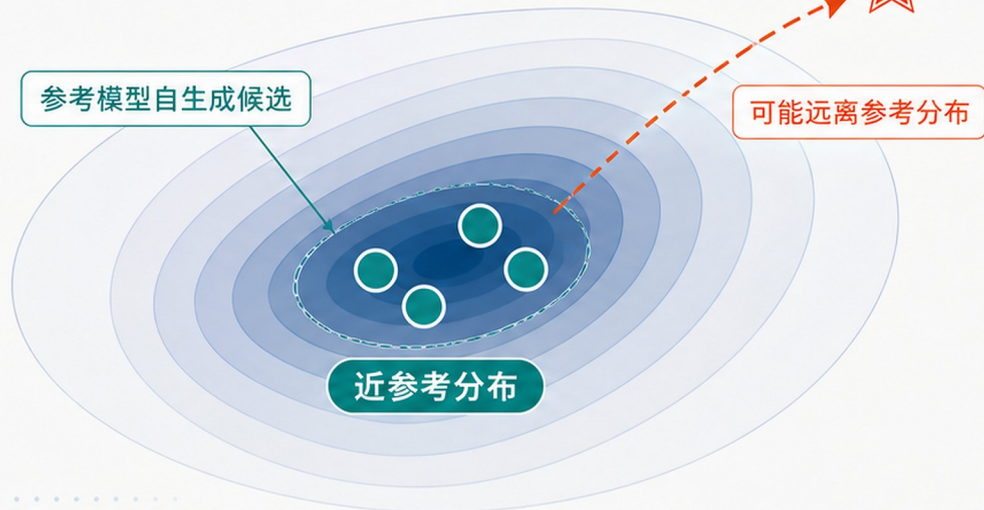


动机：能否利用模型自身的内部信号，低成本、近 on-policy 地自动构建高质量偏好数据？

更隐蔽的问题：外部纠正可能是 off-policy

— 参考策略分布 π_{ref}
(在潜在/样本空间中的示意)

外部大幅改写的 chosen



1 偏好样本的分布位置

DPO 的相对偏好学习，
通常更适配参考策略本身或其附近的样本

2 外部纠正带来的失配

人类或外部模型大幅改写的 chosen
可能远离参考分布，
使相对似然更新次优甚至偏移

3 RLSF-V 的近 on-policy 动机

由参考模型自生成多个候选，
再利用内部信号完成排序



理想路径：在同一参考策略的候选集合内 **自评价、自排序**

注：这是近参考分布的设计动机，并非声称 DPO 绝对要求严格 on-policy。

研究问题

1

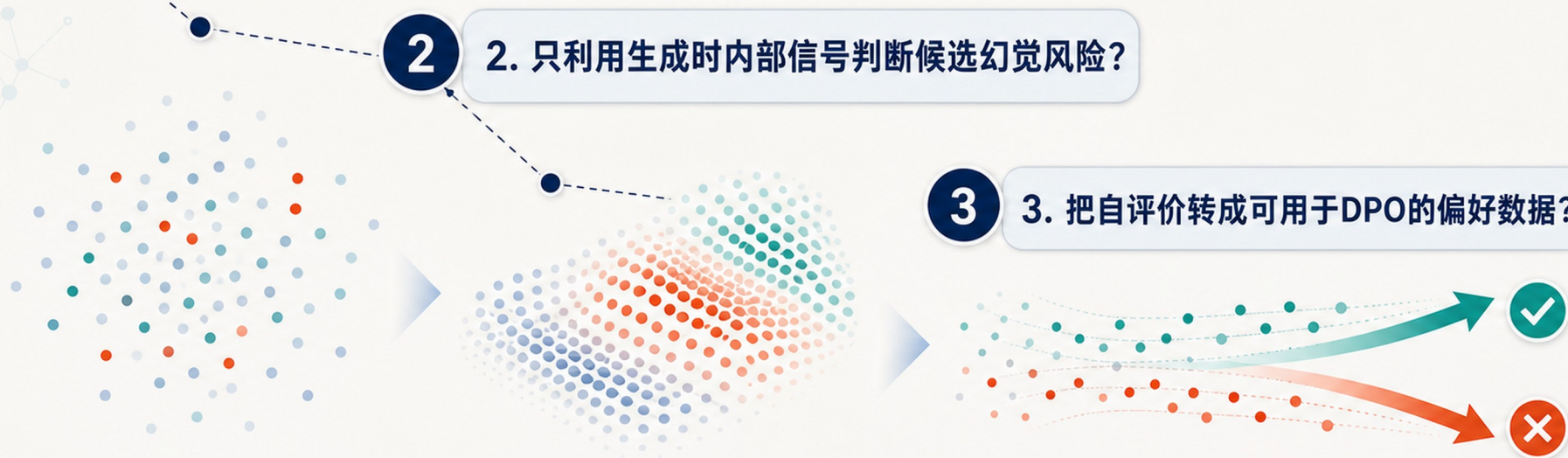
1. 不借助人工、闭源模型或更强MLLM?

2

2. 只利用生成时内部信号判断候选幻觉风险?

3

3. 把自评价转成可用于DPO的偏好数据?



logits



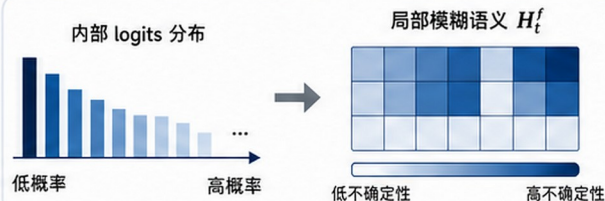
uncertainty



preference

RLSF-V 的一句话答案与贡献

1 内部 logits → 局部模糊语义不确定性



2 POS 内容词 + Top-R 聚焦事实承载位置

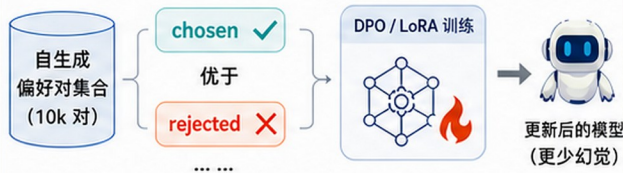
示例句子：我 喜欢 吃 香甜 的 苹果 和 新鲜 的 牛奶。

POS 标注： PRON VERB VERB ADJ ADJ NOUN CONJ ADJ ADJ NOUN PUNCT

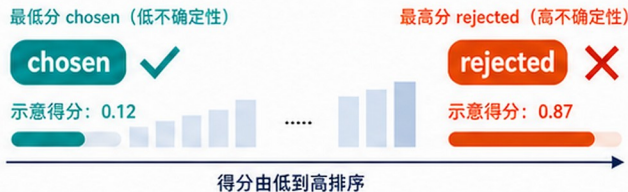
☐ NOUN (名词) ☐ ADJ (形容词) ☐ VERB (动词)

Top-R 聚焦事实承载位置 (内容词集合) 喜欢 吃 香甜 苹果 新鲜 牛奶 ...

4 10k 自生成偏好对 + DPO, 无外部评估器



3 最低分 chosen、最高分 rejected



Self-Feedback
自反馈闭环

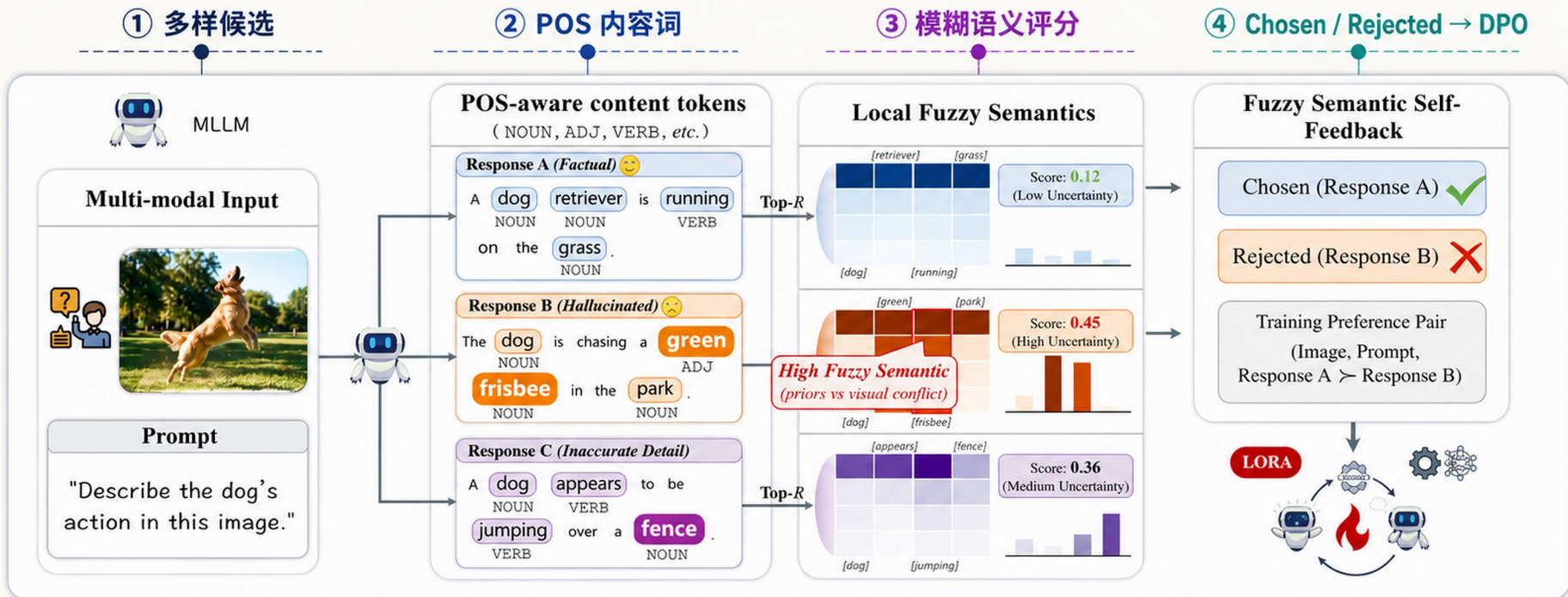


模型自身的局部 logit 竞争，就是低成本、近 on-policy 的幻觉自反馈信号

RLSF-V: 从内部 logits 到自生成偏好

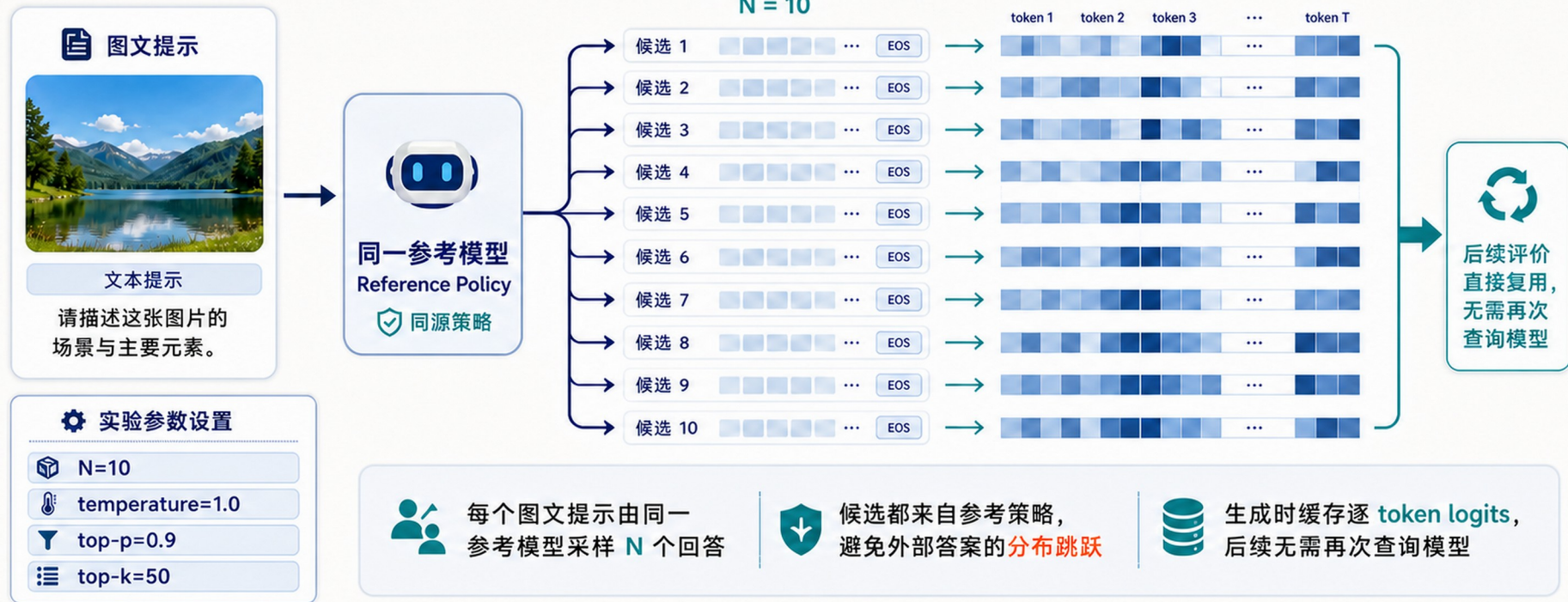
方法总览

同一参考模型完成候选生成、局部模糊语义评价与偏好学习



无需人工或外部评估模型：生成、评价与学习形成闭环

步骤一：生成多样候选并缓存 logits



同源采样 + 一次前向生成：同时获得候选与自反馈所需内部信号

内部 logits：模型“犹豫”的原始记录

每一步解码都在词表上做选择；局部竞争揭示模型的不确定与潜在幻觉

关键点

1 每个解码位置都有词表维度的logit向量



2 top候选的相对间隔反映局部竞争状态



3 幻觉常对应多个语义上合理但视觉证据不足的候选



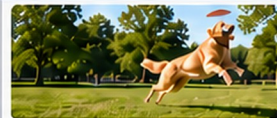
4 目标是定位最可能承载事实错误的局部犹豫



解码生成序列
(示例响应)

t_1 t_2 t_3 ... t_{k-2} t_{k-1} t_k t_{k+1}
The dog is ... in the park .

示例图像



放大解码位置： t_k (生成“park”)

Top 候选 (按 logit 排序)



示意 logits / 概率

间隔 (示意)
较小
(竞争激烈)

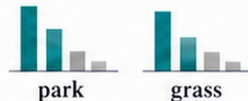
间隔 (示意)
较大
(显著落后)

Top1 与 Top2 间隔 (示意)：较小 (竞争激烈) → 不确定度较高

与视觉证据对照 (在该位置)

视觉上支持 (证据充分 / 不确定度更低)

“park”与“grass”与图像一致



视觉冲突 (证据不足 / 不确定度更高)

“fence”与图像冲突



局部 logit 竞争 = 模型“犹豫”的原始记录
定位最可能承载事实错误的局部位置



本质：logits 记录了模型在每一步对词表的偏好分布；间隔越小、竞争越激烈，局部不确定性越高，越可能隐藏幻觉风险。

经典基线：Shannon 预测熵

Shannon 预测熵 (token级)

$$H_t^{se} = - \sum_{v \in \mathcal{V}} p_{\theta}(v | y_{<t}, m, x) \log p_{\theta}(v | y_{<t}, m, x)$$

位置 t 的
Shannon 熵

在条件 $(y_{<t}, m, x)$ 下
词表项 v 的预测概率

完整词表
 $\mathcal{V}$ 上求和

自然对数

跨位置聚合得到回答级不确定性

均值聚合 (默认)

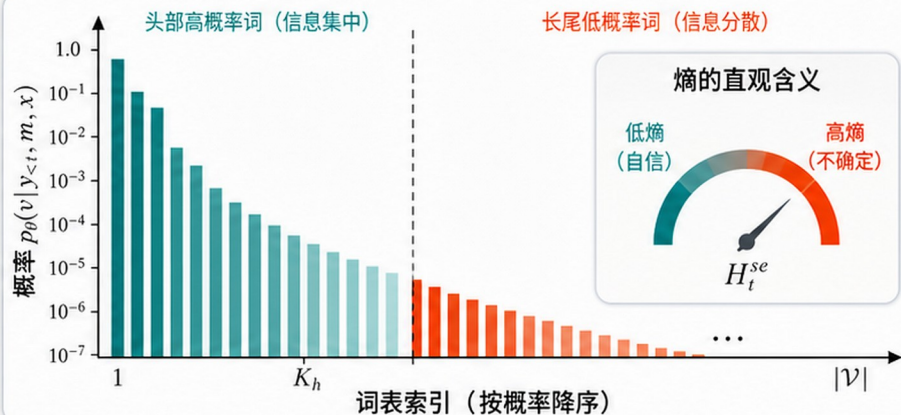
$$H_{\text{avg}}^{se} = \frac{1}{T} \sum_{t=1}^T H_t^{se}$$

位置加权聚合

$$H_w^{se} = \frac{\sum_{t=1}^T w_t \cdot H_t^{se}}{\sum_{t=1}^T w_t}$$

- T : 回答长度
- w_t : 位置权重 (可选)

完整词表概率分布 (位置 t)



核心直觉



高熵通常表示 **不确定**



低熵通常表示 **自信**

关键假设 (局限)



默认所有词表项同等重要



默认所有解码位置同等重要

公式符号说明

$y_{<t}$: 已生成的前缀文本; m : 多模态输入 (如图像); x : 文本指令/提示; $\mathcal{V}$: 完整词表; t : 当前解码位置

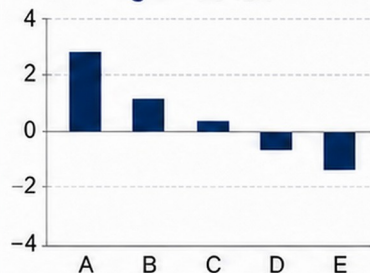
$p_{\theta}(v | y_{<t}, m, x)$: 模型在条件 $(y_{<t}, m, x)$ 下预测下一个 token 为 v 的概率 (softmax 后)

问题一：Shannon 熵对 logit 尺度过敏

缩放 logits 等价于改变 softmax 温度；排序不变，熵却快速饱和

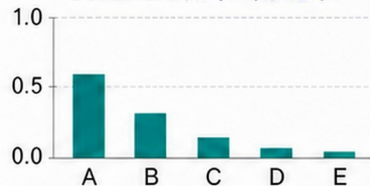
×1

logits (示意)



✓
候选排序不变
min-max
相对位置不变

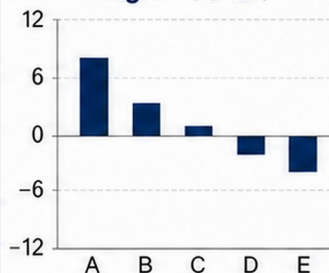
softmax 概率 (示意)



$H = 1.21$ (示意)

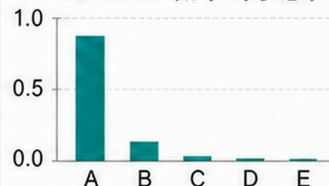
×3

logits (示意)



✓
候选排序不变
min-max
相对位置不变

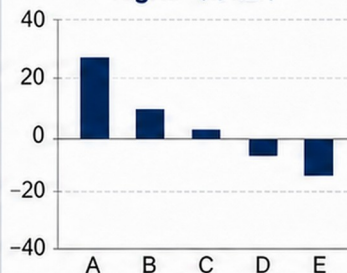
softmax 概率 (示意)



$H = 0.42$ (示意)

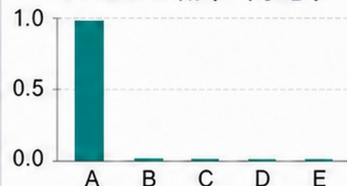
×10

logits (示意)



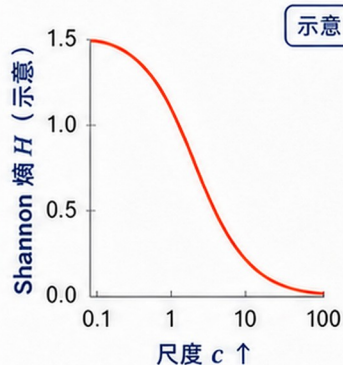
→
分布更尖锐
熵更低

softmax 概率 (示意)



$H \rightarrow 0$ (示意)

$$\text{softmax}(cz / T) \\ = \text{softmax}(z / (T/c))$$



低熵 ≠ 无语义冲突
它可能只是尺度 /
温度带来的过度自信



问题：Shannon 熵把 尺度变化 误读为 不确定性变化

问题二：全词表熵看不见局部语义竞争

反例说明：幻觉风险可集中于少数竞争语义词，而全词表归一化与跨句平均会稀释该信号

① 一个内容词位置



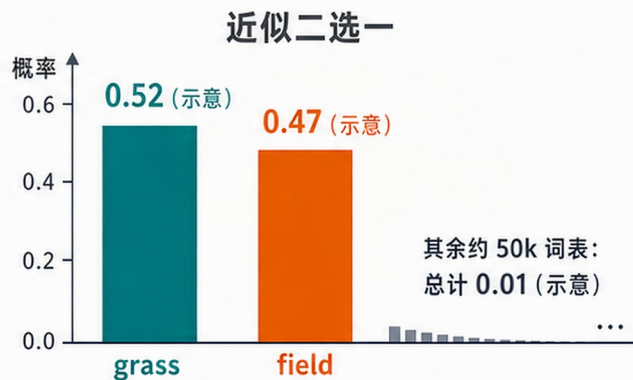
生成片段

A dog is standing on the
[grass / field].



视觉语义分岔集中在
一个内容词位置

② 局部竞争很强



简化后 $H_{\text{raw}} \approx 0.74$ nats (示意)

$H_{\text{raw}} \neq 0$

局部语义竞争：高风险

③ 为什么全局指标仍显得低？

① 全词表归一化

$$H_{\text{raw}} / \ln|V| \approx 0.07 \text{ (示意)}$$

$$|V| \approx 50k \text{ (示意)}$$



② 跨整句平均



单个高风险位置被
更多低风险 Token 稀释



全局信号：弱



幻觉风险来自少数语义分岔，而不是整张概率单纯形都均匀

模糊隶属度：只保留相对 logit 位置

Min-Max 归一化

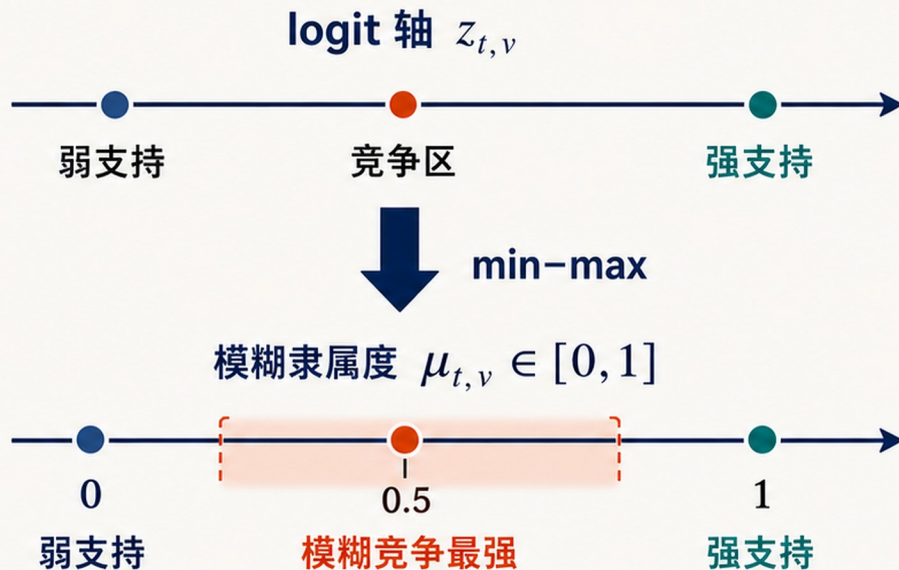
$$\mu_{t,v} = \frac{z_{t,v} - z_{t,\min}}{z_{t,\max} - z_{t,\min} + \varepsilon}$$

$z_{t,v}$: 候选词 v 的 logit

$z_{t,\min} / z_{t,\max}$: 集合 S_t 中的最小 / 最大 logit

ε : 避免分母为 0

只保留相对位置，不依赖绝对尺度



$\mu \rightarrow 1$ | 强支持



$\mu \approx 0.5$ | 模糊竞争

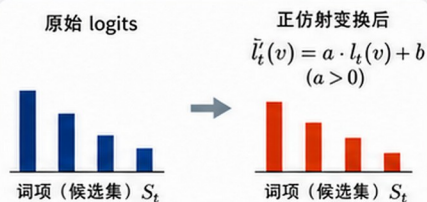


相对间隔 | 而非 softmax 绝对尖锐度

局部模糊语义：同时修正尺度与局部性失配

问题一 | 尺度敏感

正仿射变换改变绝对尺度，
但相对位置不变



解法：min-max 模糊隶属度

对正仿射 logit 变换保持不变

$$\mu_t(v) = \frac{l_t(v) - \min_{u \in S_t} l_t(u)}{\max_{u \in S_t} l_t(u) - \min_{u \in S_t} l_t(u)}$$

① μ_t 是“模糊隶属度/相对位置”，
不是不确定性。

✓ $\mu(az + b) = \mu(z)$, $a > 0$

核心：token级局部模糊语义

对候选集 S_t 上的 min-max 隶属度计算二元模糊熵并平均

$$H_t^f = -\frac{1}{|S_t| \ln 2} \sum_{v \in S_t} [\mu_t(v) \ln \mu_t(v) + (1 - \mu_t(v)) \ln(1 - \mu_t(v))]$$

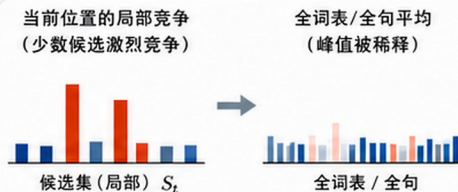


$H_t^f \in [0, 1]$ ，越高表示当前生成位置的局部语义越模糊

① 对 μ 与 $1 - \mu$ 计算二元模糊熵，并在候选集 S_t 上平均

问题二 | 全局稀释

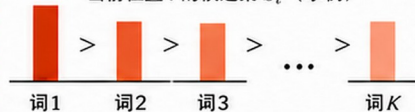
全词表/全句平均会稀释少数关键 token
的局部竞争峰值



解法：局部候选集 S_t + token级评分

- ✓ 只在当前生成位置的候选集合刻画相对间隔
- ✓ 避免全局平均掩盖局部峰值

当前位置 t 的候选集 S_t (示例)



两个失配，
两项修正

❗ 尺度敏感 → min-max / 正仿射不变

❗ 全局稀释 → 局部候选集 / token级评分



后续步骤 (未在本页完成)

后续：POS选择“看哪些位置” →
Top-R聚合“如何汇总”

关键性质：对正仿射 logit 变换保持不变

正仿射变换中的消去

① 变换: $\tilde{z}_i = az_i + b, \quad a > 0$

② 代入: $\tilde{\mu}_i = \frac{az_i + b - (az_{\min} + b)}{(az_{\max} + b) - (az_{\min} + b)}$

③ 消去: $\tilde{\mu}_i = \frac{z_i - z_{\min}}{z_{\max} - z_{\min}} = \mu_i$ ✓ 平移 b 、正尺度 a 均被消去

$$H_t^f(az + b1) = H_t^f(z)$$

同一 logit 形状，不同尺度

局部不确定性



min-max 关注相对 logit 位置;
softmax 熵会随温度/尺度饱和

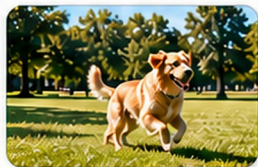


模糊语义更适合跨温度、跨模型比较局部歧义

同样“很自信”，局部语义可能完全不同

A | 视觉一致：低模糊

整体序列：高置信（定性示意）



A dog is running on the grass .

内容词位置 · 定性示意（grass 位置）



✓ 证据一致
↓
局部竞争弱
↓
低模糊

Global
confidence

≠

Local
semantic
certainty

B | 先验冲突：高模糊

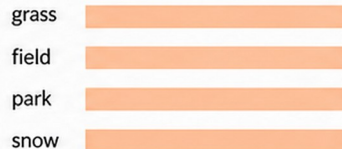
整体序列：高置信（定性示意）



A dog is running on the snow .



内容词位置 · 定性示意（snow 位置）



⚠ 证据冲突
↓
局部竞争强
↓
高模糊

语言先验 ↔ 视觉证据冲突

回答 token 热力条
（定性示意）

A | 视觉一致

B | 先验冲突



少数关键位置
会被全局序列
分数掩盖

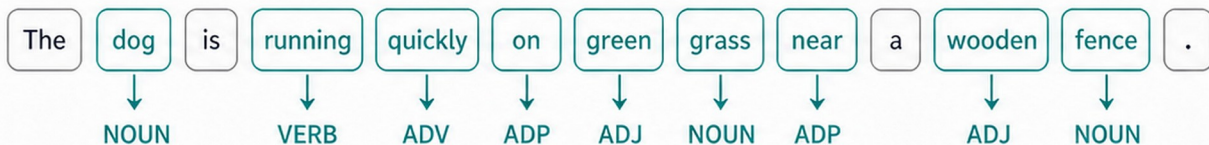


需要同时解决：看哪里（POS内容词） + 如何量化（局部模糊语义）

步骤二：POS-aware 内容 token 选择

语言学先验：只在承载视觉事实的位置聚合不确定性

英文回答



轻量 POS tagger



词级词性标签



映射回子词 token

running → [run] [ning]

同一词的字词沿用该词标签

保留的 POS 集合

NOUN

PROPN

NUM

ADJ

VERB

ADV

ADP

对象 · 专名 · 数量 · 属性 ·
动作/状态 · 修饰 · 关系

聚焦事实承载位置

对象



例：dog

数量

123

例：two, 3

属性



例：green

状态



例：standing

动作



例：chasing

关系



例：near



过滤功能词，
避免句子级平均稀释有效信号



更少噪声，更强事实敏感性

幻觉不是均匀分布的，而是局部爆发

全局平均会掩盖少数高风险内容词

示意

例句：A dog is running on the grass near a bench .

token序列 (示意)

每个 token 的
不确定性 (示意)

局部高风险峰值

Global Average / 全局平均

Local Peaks / 局部峰值

高风险被多数正常token稀释

聚焦最不稳定的内容位置



1 长回答中通常只有少数 token 引入不存在对象或错误属性



2 对所有 token 平均，会让高风险峰值被正常 token 淹没



3 POS 过滤后，仍只有少数内容位置最不稳定



4 下一步：只聚合最危险的 Top-R token

Top-R pooling: 聚合最不稳定的内容位置



① 将 POS 过滤后的 H_t^f 按降序排列



只对最大的 R 个值求平均



内容 token 少于 R 时使用全部位置



轻量重要性聚合: 无需额外模型

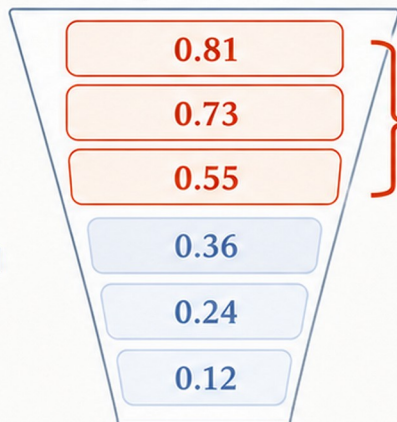
POS-filtered 内容 token 与不确定性 H_t^f

dog		0.12
running		0.36
grass		0.24
red		0.81
frisbee		0.73
park		0.55

示意值

按 H_t^f
降序排列

按 H_t^f 降序后的序列



Top-R
(示例: $R = 3$)

回答级不确定性 $U_{\text{pos}}(y)$



$$(0.81 + 0.73 + 0.55) / 3 = 0.70$$

示意

聚焦最不稳定的内容位置



$$U_{\text{pos}}(y^{(i)}) = \frac{1}{\min(R, |T^{(i)}|)} \cdot \sum_{r=1}^{\min(R, |T^{(i)}|)} H_{(r)}^f$$

若内容 token 数 $|T^{(i)}| < R$, 则对全部可用内容 token 求平均。

步骤三：从自评分到 chosen / rejected

按回答级不确定性排序，自动构建训练偏好对

① 对 N 个候选分别计算 U_{pos}

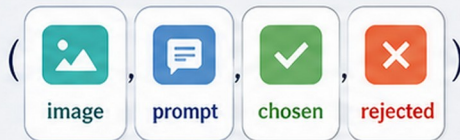
U_{pos} 递增



② 最低不确定性作为 y^{cho} ，
最高不确定性作为 y^{rej}

首尾配对

③ 自动生成偏好四元组



仅当分数有差异时加入

无外部 evaluator



自反馈闭环



① 参考模型既生成候选，也提供排序信号：
偏好数据无需人工或外部模型。

② 最低不确定性作为 y^{cho} ，
最高不确定性作为 y^{rej} 。

③ 形成 (image, prompt, chosen, rejected)
偏好四元组。

④ 分数有差异时加入数据集，
全程无外部 evaluator。

步骤四：用 DPO 把自反馈写回模型

1) 偏好对 (来自模型自反馈)



chosen:

低模糊、
视觉可证实



rejected:

高模糊、
存在视觉冲突

DPO: 提升 chosen 相对 rejected 的优势

$$\mathcal{L}_{\text{DPO}} = -\mathbb{E} \left[\log \sigma \left(\beta \left\{ \left[\log \pi_{\theta}(y^{+} | x) - \log \pi_{\theta}(y^{-} | x) \right] - \left[\log \pi_{\text{ref}}(y^{+} | x) - \log \pi_{\text{ref}}(y^{-} | x) \right] \right\} \right) \right]$$

可训练策略的
相对偏好

固定参考策略的
相对偏好 (锚点)



π_{θ} 可训练

同一基础模型初始化:

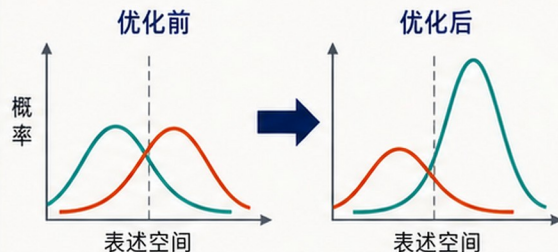


π_{ref} 固定。

3) 策略分布变化 (以 chosen/rejected 为例)

— chosen (低模糊、视觉可证实)

— rejected (高模糊、存在视觉冲突)



测试时: 单次生成更偏向
低模糊、视觉可证实的表述



π_{θ} 更新 (可训练)



π_{ref} 固定 (不更新)



$\beta = 0.1$



LoRA: 仅更新语言模型线性层



视觉塔不更新

完整算法：生成—评分—配对—优化

Algorithm 1 RLSF-V: Mitigating Hallucinations in MLLMs via Fuzzy Semantic Self-Feedback

Require: Base MLLM π_θ , multimodal training corpus $\mathcal{D} = \{(m, x)\}$ (image-text prompts), number of candidates N , top- R tokens R , temperature β .

Ensure: Fine-tuned policy π_τ for hallucination-mitigated generation.

```
1: // Diverse candidate generation
2: for each prompt  $(m, x) \in \mathcal{D}$  do
3:   Sample  $N$  candidate responses  $\{y^{(i)}\}_{i=1}^N \sim \pi_\theta(\cdot \mid m, x)$  using stochastic decoding.
4:   Cache token-level logits  $\{z_t^{(i)}\}$  for later local fuzzy semantics estimation.
5:   // POS-aware local fuzzy semantics
6:   for each candidate  $y^{(i)}$  do
7:     For each position  $t$ , compute the local fuzzy semantics  $H_t^{f,(i)}$  from logits  $z_t^{(i)}$  using Equation (4) and Equation (5).
8:     Employ the POS tagger on  $y^{(i)}$  and obtain the POS-aware content token index set  $\mathcal{T}^{(i)}$  as defined in Equation (11).
9:     Collect the POS-filtered content tokens  $\{H_t^{f,(i)} \mid t \in \mathcal{T}^{(i)}\}$  and apply the top- $R$  pooling in Equation (13) to obtain the local fuzzy semantics  $U_{pos}(y^{(i)})$ .
10:  end for
11:  // Fuzzy semantic self-feedback and preference construction
12:  Determine the chosen and rejected candidates  $y^{cho}$  and  $y^{rej}$  according to Equation (12), based on  $U_{pos}(\cdot)$ .
13:  if  $|U_{pos}(y^{cho}) - U_{pos}(y^{rej})| > 0$  then
14:    Add the ordered preference quadruple  $(m, x, y^{cho}, y^{rej})$  to the preference dataset  $\mathcal{D}_f$  as in Equation (15).
15:  end if
16: end for
17: // Fine-tuning with self-generated preferences
18: Initialize the policy  $\pi_\tau \leftarrow \pi_\theta$  and fix the reference policy  $\pi_{ref} \leftarrow \pi_\theta$ .
19: for each training step with mini-batch  $\mathcal{B} \subseteq \mathcal{D}_f$  do
20:   Compute the loss  $\mathcal{L}_{DPO}$  on  $\mathcal{B}$  following Equation (16) with temperature  $\beta$ .
21:   Update  $\pi_\tau$  by minimizing  $\mathcal{L}_{DPO}$ .
22: end for
```



① 多样候选生成

- 随机解码 N 个候选
- 缓存 token logits



② 模糊语义评分

- Local Fuzzy Semantics
- POS 过滤 + Top-R 聚合



③ 偏好配对

- chosen / rejected
- 构建 $\mathcal{D}_f$



④ DPO 优化

- 固定参考策略
- 微调目标策略

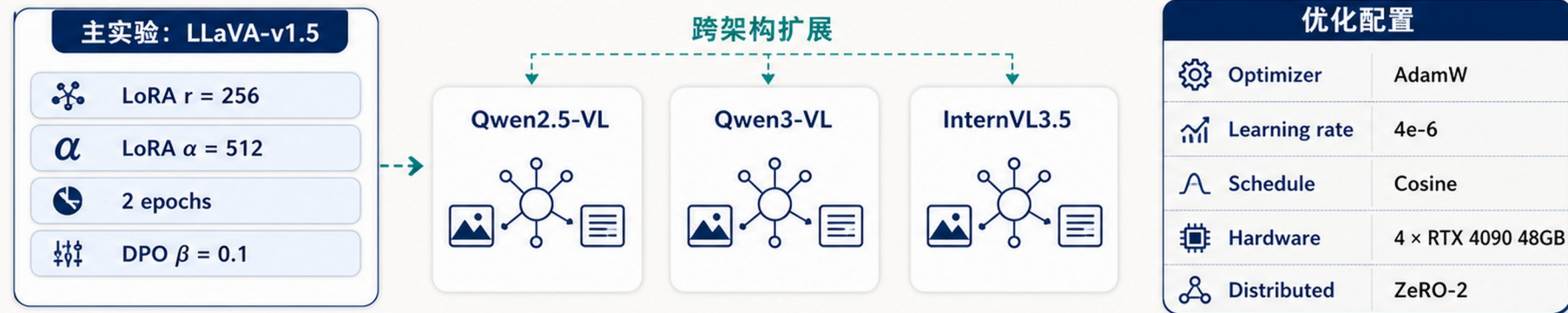
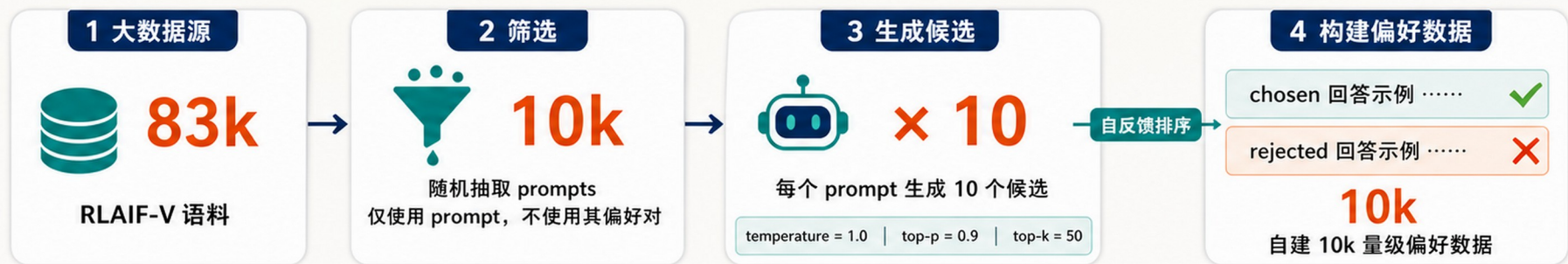


复现要点

生成 → 评分 → 配对 → 优化
同一参考模型完成闭环

训练配置与数据构建

10k prompts × 10 candidates: 低成本构造自反馈偏好数据



同一参考模型完成候选生成与自反馈排序, 偏好数据 **近 on-policy**

评测设计：四类幻觉基准 + 通用能力

幻觉评测

1



生成式 | 对象事实
AMBER Generative

2



生成式 | 对象幻觉
Object HalBench

3



判别式 | 复杂推理
HallusionBench

4



判别式 | 对象事实
AMBER Discriminative

↓ 越低越好：CHAIR / HalRate / Cog / Resp / Ment

↑ 越高越好：All Acc / Acc / F1

通用能力（10 类基准）



MME



MMMU



MMStar



AI2D

等 10 类基准



统一官方 checkpoint；测试 temperature = 0



先看幻觉是否下降，再确认通用能力不退化

主结果：多数幻觉指标达到最好或次优

Model	Size	Feedback	AMBER GEN			HallusionBench	Object HalBench		AMBER DIS	
			CHAIR↓	HalRate↓	Cog↓		Resp↓	Ment↓	Acc↑	F1↑
LLaVA-v1.5-7B (Liu et al., 2023)	X	X	7.7	35.9	4.3	42.7	25.2	46.7	71.5	74.1
+HA-DPO (Zhao et al., 2023)	6k	GPT-4	6.7	30.5	3.2	43.7	20.3	36.7	74.2	78.0
+POVID (Zhou et al., 2024)	17k	GPT-4V	9.8	51.4	7.0	42.2	25.4	52.7	72.4	74.8
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	8.3	41.4	4.8	38.9	28.0	47.0	77.9	83.2
+TPO (He et al., 2024b)	21.4k	LLaVA-Next	3.6	20.5	1.6	31.6	4.4	8.0	78.8	86.2
+AMP-IG (Zhang et al., 2024)	11k	CLIP	17.6	73.8	9.1	38.2	26.7	43.7	74.2	80.8
+AMP-MEG (Zhang et al., 2024)	90k	LLaVA & CLIP	13.2	59.1	5.2	37.0	25.7	32.3	73.7	77.8
+HALVA (Sarkar et al., 2025)	21.5k	GPT-4V	7.0	33.6	3.6	44.2	23.8	43.0	77.9	83.4
+RLAIF-V (Yu et al., 2025b)	83k	LLaVA-Next	2.9	15.6	0.9	28.7	5.6	9.7	54.3	73.9
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	2.5	12.1	0.9	42.2	3.8	5.7	80.5	85.1
+SENTINEL (Peng et al., 2025)	8.7k	G-DINO & Yolo-W	3.0	14.8	1.3	42.9	3.0	5.0	76.1	79.3
+symMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	5.9	28.3	3.6	45.3	9.6	15.7	81.2	86.5
+RLSF-V	10k	self-feedback	2.4↓5.3	5.1↓30.8	0.3↓4.0	47.4↑4.7	2.9↓22.3	4.0↓42.7	78.8↑7.3	84.0↑9.9
LLaVA-v1.5-13B (Liu et al., 2023)	X	X	6.9	32.0	3.5	43.3	22.6	45.0	71.3	73.1
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	6.9	34.4	3.7	42.3	20.9	41.0	81.7	85.2
+AMP-IG (Zhang et al., 2024)	90k	CLIP	13.4	52.0	5.4	40.2	33.8	57.7	66.2	70.3
+AMP-MEG (Zhang et al., 2024)	11k	LLaVA & CLIP	11.6	51.1	5.4	39.9	24.9	38.7	62.8	63.4
+HALVA (Sarkar et al., 2025)	21.5k	GPT-4V	6.5	30.6	3.1	46.2	19.2	39.0	82.9	86.5
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	2.7	13.4	0.9	42.2	4.9	7.3	84.1	87.5
+SENTINEL (Peng et al., 2025)	7k	G-DINO & Yolo-W	2.7	12.0	0.9	43.8	2.4	4.0	76.2	78.7
+symMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	4.6	26.7	2.4	45.1	8.4	16.0	84.5	89.0
+RLSF-V	10k	self-feedback	2.0↓4.9	4.4↓27.6	0.3↓3.2	48.1↑4.8	1.2↓21.4	1.7↓43.3	84.8↑13.5	87.6↑14.5
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	X	X	8.0	49.1	5.2	53.1	9.7	15.7	88.0	91.0
+RLSF-V	10k	self-feedback	6.5↓1.5	36.5↓12.6	3.9↓1.3	53.9↑0.8	7.0↓2.7	9.3↓6.4	88.2↑0.2	91.1↑0.1
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	X	X	8.1	62.0	3.4	64.0	8.1	14.3	89.0	91.7
+RLSF-V	10k	self-feedback	6.7↓1.4	48.1↓13.9	2.3↓1.1	65.5↑1.5	5.5↓2.6	9.7↓4.6	89.5↑0.5	92.1↑0.4
InternVL3.5-8B-HF (Wang et al., 2025b)	X	X	7.6	63.8	5.3	57.6	5.5	9.4	86.7	89.4
+RLSF-V	10k	self-feedback	7.2↓0.4	58.7↓5.1	4.5↓0.8	58.1↑0.5	4.4↓1.1	6.7↓2.7	86.7↑0.0	89.4↑0.0



LLaVA-v1.5-7B

AMBER HalRate 35.9 → 5.1

CHAIR 7.7 → 2.4



LLaVA-v1.5-13B

Object HalBench Resp 22.6 → 1.2

Object HalBench Ment 45.0 → 1.7



HallusionBench All Acc

提升到 47.4 / 48.1



训练成本

10k self-feedback

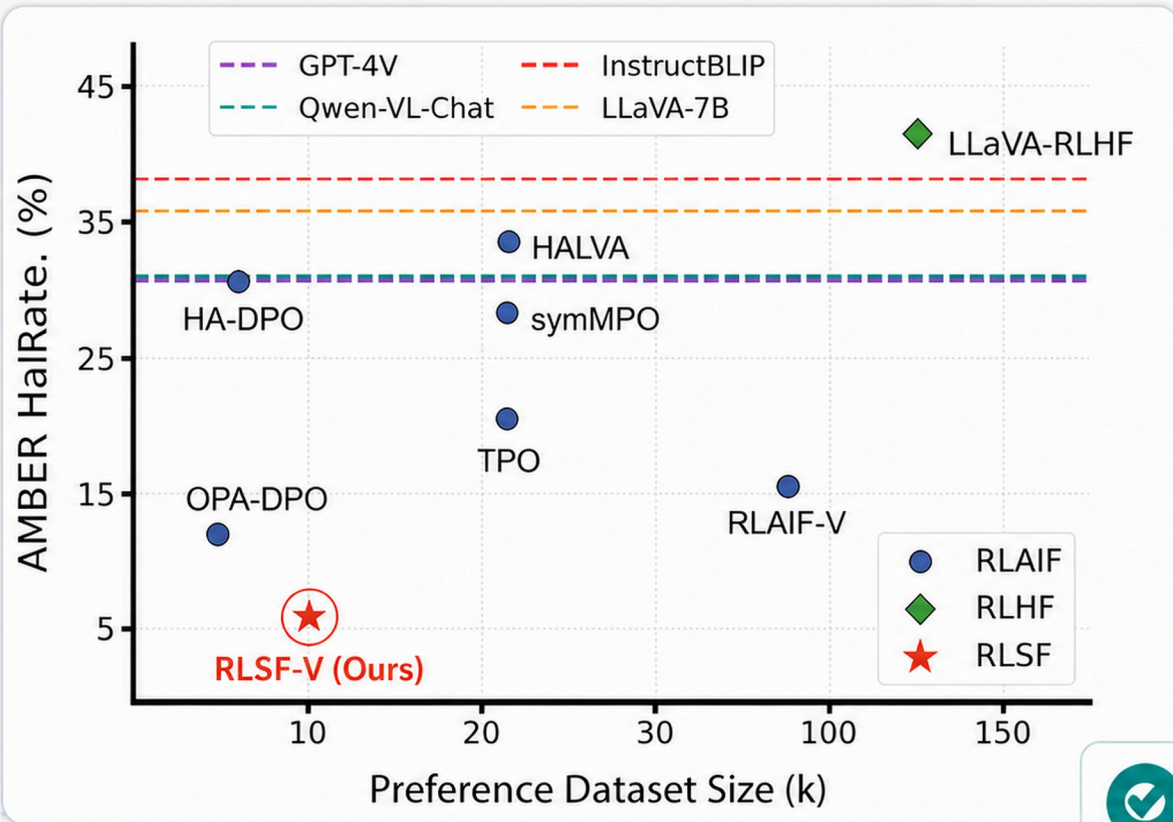
与人工 / 外部大模型方法竞争



只用 10k 自反馈数据，也能在多数幻觉指标上达到最好或次优。

数据效率：10k 自反馈位于性能—规模前沿

实验结果



横轴：偏好数据规模
纵轴：AMBER 幻觉率



仅 10k 自反馈数据
幻觉率低于大规模
RLHF / RLAIF 基线



无需外部教师
迁移新 backbone
成本更可控



更少偏好数据，更低幻觉率：
形成有利的性能—规模权衡

跨架构迁移：同一流水线适配四类 MLLM

论文 Table 1 | 跨架构结果（原始数据）

Model	Size	Feedback	AMBER GEN			HallusionBench	Object HalBench		AMBER DIS	
			CHAIR↓	HalRate.↓	Cog.↓	All Acc.↑	Resp.↓	Ment.↓	Acc.↑	F1.↑
LLaVA-v1.5-7B (Liu et al., 2023)	X	X	7.7	35.9	4.3	42.7	25.2	46.7	71.5	74.1
+HA-DPO (Zhao et al., 2023)	6k	GPT-4	6.7	30.5	3.2	43.7	20.3	36.7	74.2	78.0
+POVID (Zhou et al., 2024)	17k	GPT-4V	9.8	51.4	7.0	42.2	25.4	52.7	72.4	74.8
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	8.3	41.4	4.8	38.9	28.0	47.0	77.9	83.2
+TPO (He et al., 2024b)	21.4k	LLaVA-Next	3.6	20.5	1.6	31.6	4.4	8.0	78.8	86.2
+AMP-IG (Zhang et al., 2024)	11k	CLIP	17.6	73.8	9.1	38.2	26.7	43.7	74.2	80.8
+AMP-MEG (Zhang et al., 2024)	90k	LLaVA & CLIP	13.2	59.1	5.2	37.0	25.7	32.3	73.7	77.8
+HALVA (Sarkar et al., 2025)	21.5k	GPT-4V	7.0	33.6	3.6	44.2	23.8	43.0	77.9	83.4
+RLAIF-V (Yu et al., 2025b)	83k	LLaVA-Next	2.9	15.6	0.9	28.7	5.6	9.7	54.3	73.9
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	2.5	12.1	0.9	42.2	3.8	5.7	80.5	85.1
+SENTINEL (Peng et al., 2025)	8.7k	G-DINO & Yolo-W	3.0	14.8	1.3	42.9	3.0	5.0	76.1	79.3
+symMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	5.9	28.3	3.6	45.3	9.6	15.7	81.2	86.5
+RLSF-V	10k	self-feedback	2.4 _{1.5}	5.1 _{30.8}	0.3 _{1.0}	47.4 _{4.7}	2.9 _{12.3}	4.0 _{14.7}	78.8 _{77.3}	84.0 _{79.9}
LLaVA-v1.5-13B (Liu et al., 2023)	X	X	6.9	32.0	3.5	43.3	22.6	45.0	71.3	73.1
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	6.9	34.4	3.7	42.3	20.9	41.0	81.7	85.2
+AMP-IG (Zhang et al., 2024)	90k	CLIP	13.4	52.0	5.4	40.2	33.8	57.7	66.2	70.3
+AMP-MEG (Zhang et al., 2024)	11k	LLaVA & CLIP	11.6	51.1	5.4	39.9	24.9	38.7	62.8	63.4
+HALVA (Sarkar et al., 2025)	21.5k	GPT-4V	6.5	30.6	3.1	46.2	19.2	39.0	82.9	86.5
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	2.7	13.4	0.9	42.2	4.9	7.3	84.1	87.5
+SENTINEL (Peng et al., 2025)	7k	G-DINO & Yolo-W	2.7	12.0	0.9	43.8	2.4	4.0	76.2	78.7
+symMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	4.6	26.7	2.4	45.1	8.4	16.0	84.5	89.0
+RLSF-V	10k	self-feedback	2.0 _{1.9}	4.4 _{12.6}	0.3 _{1.2}	48.1 _{7.8}	1.2 _{12.1}	1.7 _{14.3}	84.8 _{73.5}	87.6 _{74.5}
Qwen2.5-VL-3B-Instruct (Bai et al., 2025b)	X	X	8.0	49.1	5.2	53.1	9.7	15.7	88.0	91.0
+RLSF-V	10k	self-feedback	6.5 _{1.5}	36.5 _{12.6}	3.9 _{1.3}	53.9 _{70.8}	7.0 _{12.7}	9.3 _{16.4}	88.2 _{70.2}	91.1 _{70.1}
Qwen3-VL-8B-Instruct (Bai et al., 2025a)	X	X	8.1	62.0	3.4	64.0	8.1	14.3	89.0	91.7
+RLSF-V	10k	self-feedback	6.7 _{1.4}	48.1 _{13.9}	2.3 _{1.1}	65.5 _{71.5}	5.5 _{12.6}	9.7 _{14.6}	89.5 _{70.5}	92.1 _{70.4}
InternVL3.5-8B-HF (Wang et al., 2025b)	X	X	7.6	63.8	5.3	57.6	5.5	9.4	86.7	89.4
+RLSF-V	10k	self-feedback	7.2 _{10.4}	58.7 _{15.1}	4.5 _{10.8}	58.1 _{70.5}	4.4 _{11.1}	6.7 _{12.7}	86.7 _{70.0}	89.4 _{70.0}

四类 MLLM Backbone 迁移链



Qwen2.5-VL-3B

AMBER HalRate

49.1 → 36.5



Qwen3-VL-8B

AMBER HalRate

62.0 → 48.1

HallusionBench

64.0 → 65.5



InternVL3.5-8B

多数幻觉生成指标改善

DIS 指标持平



同一 RLSF-V 流水线，
无需为每个 backbone 更换外部 evaluator

缓解幻觉是否牺牲通用能力？

通用能力证据

Benchmark	Qwen3-VL-8B-Instruct	+RLSF-V
MME-Cog. (Fu et al., 2026)	641.79	638.93 _{↓2.86}
MME-Per. (Fu et al., 2026)	1720.80	1715.30 _{↓5.50}
MMMU (Yue et al., 2024)	52.33	52.89 _{↑0.56}
MMStar (Chen et al., 2024c)	62.88	63.78 _{↑0.90}
AI2D (Kembhavi et al., 2016)	83.65	83.71 _{↑0.06}
RealWorldQA (xAI, 2024)	69.41	70.20 _{↑0.79}
ChartQA (Masry et al., 2022)	85.36	85.52 _{↑0.16}
OCRBench (Liu et al., 2024e)	83.00	82.40 _{↓0.60}
SeedBench (Li et al., 2023)	75.07	75.15 _{↑0.08}
DocVQA (Mathew et al., 2021)	95.61	95.95 _{↑0.34}

Table 2 · Qwen3-VL-8B 通用基准

Model	Size	Feedback	Subject			Context Modality			Grade		Avg
			NAT	SOC	LAN	TXT	IMG	NO	G1-6	G7-12	
LLaVA-v1.5-7B (Liu et al., 2023)	×	×	70.52	74.24	67.09	70.77	69.46	67.32	73.97	64.01	70.41
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	68.69	77.28	64.73	71.11	70.25	64.25	72.61	63.81	69.46
+TPO (He et al., 2024b)	21.4k	LLaVA-Next	69.85	70.42	64.82	70.04	67.13	65.09	71.44	63.68	68.66
+AMP-IG (Zhang et al., 2024)	11k	CLIP	58.17	62.43	58.27	58.11	60.04	57.63	62.81	52.41	59.09
+AMP-MEG (Zhang et al., 2024)	90k	LLaVA & CLIP	58.97	62.65	57.00	60.22	59.54	55.12	62.96	52.54	59.23
+RLAIF-V (Yu et al., 2025b)	83k	LLaVA-Next	69.49	71.54	62.00	69.60	68.12	63.14	70.89	62.76	67.98
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	70.07	73.57	67.18	70.09	68.67	67.46	73.68	63.55	70.05
+synMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	70.38	72.33	67.45	70.43	68.17	67.74	73.38	64.01	70.03
Ours	10k	self-feedback	68.87	72.78	66.82	67.99	67.38	67.53	72.61	62.95	69.16
LLaVA-v1.5-13B (Liu et al., 2023)	×	×	75.22	77.05	72.64	75.61	72.83	73.45	77.72	69.94	74.94
+LLaVA-RLHF (Sun et al., 2024)	122k	Human	73.45	78.63	70.64	74.19	72.73	72.06	76.98	68.09	73.80
+AMP-IG (Zhang et al., 2024)	11k	CLIP	69.09	75.03	64.73	69.79	69.16	67.04	72.32	63.61	69.21
+AMP-MEG (Zhang et al., 2024)	90k	LLaVA & CLIP	66.52	73.00	63.00	68.48	68.72	62.44	70.52	60.58	66.97
+OPA-DPO (Wu et al., 2025)	4.8k	GPT-4V	74.16	76.15	64.09	74.29	72.14	66.55	76.06	64.60	71.96
+synMPO (Liu et al., 2025)	21.4k	DeepSeek-V3	73.93	75.37	70.00	74.39	71.05	71.29	75.92	68.36	73.21
Ours	10k	self-feedback	73.91	75.70	70.36	72.78	70.40	71.71	76.25	66.71	72.84

Table 5 · LLaVA 7B/13B · ScienceQA

10项通用基准中 7项提升

下降集中于 MME-Cog、MME-Per、OCRBench，且幅度较小

① MME-Per / MME-Cog → 感知 / 认知

② MMMU / MMStar → 跨学科推理 / 综合多模态推理

③ AI2D / ChartQA → 科学图示 / 图表理解

④ OCRBench / DocVQA → 文字识别 / 文档理解

⑤ SeedBench / RealWorldQA → 综合多模态理解 / 真实场景推理

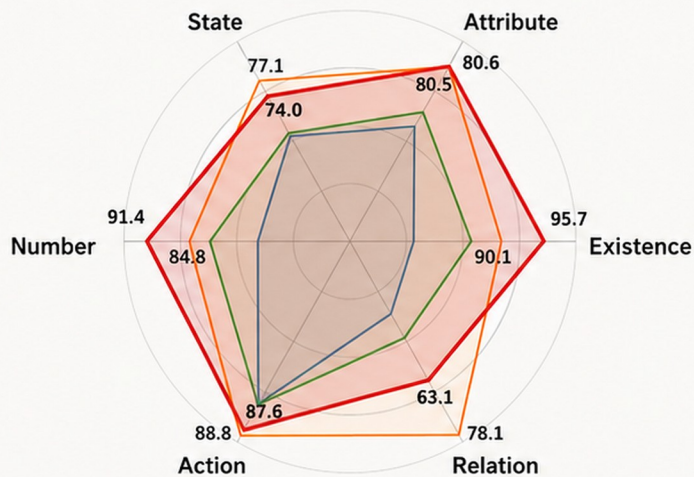
⑥ ScienceQA → 科学领域问答



显著抑制幻觉，没有带来普遍能力坍塌

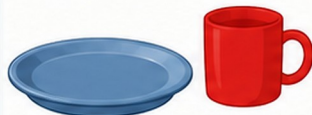
为什么关系幻觉更难抑制？

Relation 仍是相对短板



✓ 存在性、属性、状态、数量与动作上具有竞争力

定性示意



✓ 正确事实：红杯在蓝盘右侧

✗ 错误输出：

The red cup is to the ~~left~~ of the blue plate.



- 三元绑定：**关系判断需要同时绑定 subject (red cup) — predicate (right of) — object (blue plate)，不是只识别单个实体。
- 多-token / 高频复用：**left of、right of、in front of 等是多-token 关系短语；of、in 等 token 又大量用于非空间句法，局部语义容易混淆。
- 语言先验自信：**即使两个实体都识别正确，常见关系表达仍可能流畅且高置信，单个关系 token 未必产生明显局部不确定性。

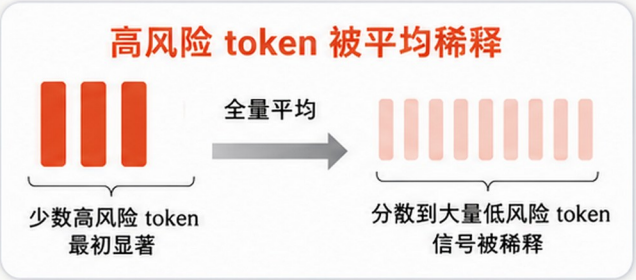
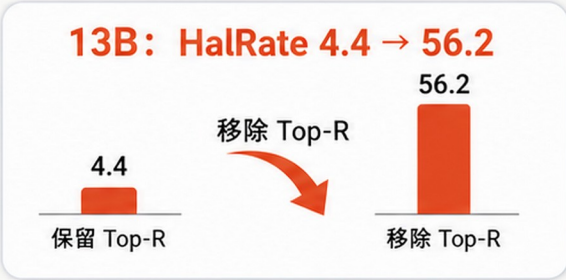
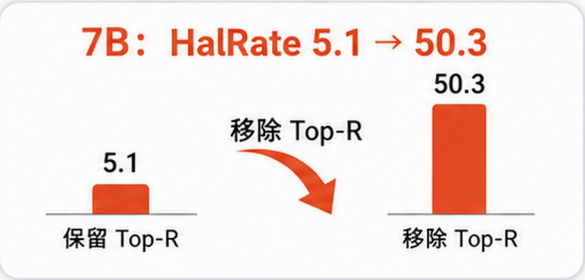


实体正确 ≠ 关系正确：关系错误可能藏在流畅、高置信的短语中

组件消融：Top-R pooling 是关键

Table 3 | 组件消融

Model	Type	AMBER GEN			HallusionBench	Object HalBench		AMBER DIS	
		CHAIR↓	HalRate.↓	Cog.↓	All Acc.↑	Resp.↓	Ment.↓	Acc.↑	F1.↑
LLaVA-v1.5-7B	RLSF-V	2.4	5.1	0.3	47.4	2.9	4.0	78.8	84.0
	W/O POS-aware Token Selection	2.4	5.9	0.4	44.5	3.1	4.7	76.3	79.2
	W/O Top-R token pooling	16.5	50.3	3.6	43.5	41.9	70.3	78.4	83.7
LLaVA-v1.5-13B	RLSF-V	2.0	4.4	0.3	48.1	1.2	1.7	84.8	87.6
	W/O POS-aware Token Selection	2.5	6.1	0.4	44.3	2.2	3.3	63.3	62.4
	W/O top-R token pooling	15.9	56.2	4.2	45.3	41.2	69.3	82.2	86.0



- 去除 POS 选择：一致但相对温和的下降 •



证据支持：幻觉往往集中在少数高风险 token，Top-R pooling 是关键聚合机制

为什么是模糊语义，而不是 Energy / Evidence / Entropy?

保持 POS 选择与 Top-R 不变，仅替换评分模块

Table 4. Comparison of different self-feedback within RLSF-V.

Model	Feedback	AMBER GEN			Object HalBench	
		CHAIR↓	HalRate.↓	Cog.↓	Resp.↓	Ment.↓
LLaVA-v1.5-7B	$\times$	7.7	35.9	4.3	25.2	46.7
	Energy	14.5	46.7	6.1	22.0	46.7
	Evidence	4.2	10.8	0.9	7.6	13.7
	Entropy	2.8	5.8	0.3	2.9	4.0
	Fuzzy	2.4	5.1	0.3	2.9	4.0
LLaVA-v1.5-13B	$\times$	6.9	32.0	3.5	22.6	45.0
	Energy	6.4	16.9	2.3	8.5	14.7
	Evidence	1.7	3.2	0.3	3.7	5.3
	Entropy	2.3	5.8	0.3	2.5	3.7
	Fuzzy	2.0	4.4	0.3	1.2	1.7

共同反例: $z = [4, 3, 1] \rightarrow z' = z - 3.5 = [0.5, -0.5, -2.5]$

softmax / 排序 / 相对间隔完全相同

ENERGY

$$U_{eng} = \log \sum \exp(z)$$
$$U_{eng}(z') = U_{eng}(z) - 3.5$$

预测不变，分数却随任意 **logit** 零点变化
且受尺度与词表大小影响

EVIDENCE

Top-K (K=50) $\rightarrow$ ReLU $\rightarrow$ S $\rightarrow$ EU / AU $\rightarrow U_{evi}$
 $S = \sum e$; $EU = K / (S + K)$; $U_{evi} = EU \cdot AU$ (AU为digamma离散度)

e: [4, 3, 1] $\rightarrow$ [0.5, 0, 0]; S: 8 $\rightarrow$ 0.5

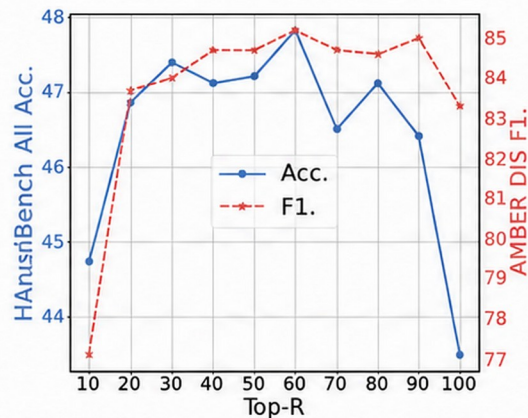
同一预测分布，**evidence** 强度却由无语义的零点决定
还依赖 K、尺度、ReLU 与 Dirichlet 映射

Entropy: 受全局分布、尺度与词表稀释影响，不直接聚焦局部候选竞争

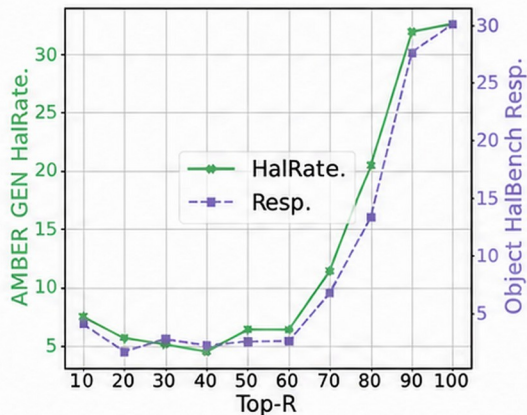
Fuzzy: $\mu(az + b) = \mu(z)$, $a > 0$

min-max 后只保留局部相对间隔；多数下游幻觉指标更优
并非每一项都绝对最优

Top-R 的本质：每个 backbone 都有自己的平衡点



(a) Discriminative benchmark



(b) Generative benchmark

LLaVA-v1.5-7B 示例：较优区间约 30-40

backbone-specific R

LLaVA → 30

Qwen2.5-VL → 20

Qwen3-VL → 5

InternVL3.5 → 40

跨 backbone 不存在统一最优值

Figure 4: LLaVA-v1.5-7B



自反馈真的能选对偏好对吗？



"chosen": "No, there is no lamp or chair in the image."

Fuzzy entropy: 0.18 ✓

"rejected": "Yes, there is a lamp in the image, which is located below the street sign."

Fuzzy entropy: 0.43 ✗



"Is there a lamp or a chair in this picture?"

街牌案例：低分回答正确否认灯/椅子；高分回答凭空添加灯



"chosen": "No, there is no person's head in the image. The image shows black curly hair and a pair of scissors."

Fuzzy entropy: 0.21 ✓

"rejected": "Yes, the head in the drawing is lighter than the rest of the body."

Fuzzy entropy: 0.39 ✗



"Is someone's head lighter?"

人头案例：低分回答依据剪刀与头发；高分回答作出无依据判断

低模糊熵 · chosen

高模糊熵 · rejected



高不确定性的 rejected 往往包含不存在对象、错误属性或无依据关系

自生成偏好与人类幻觉判断一致

训练后行为：从“过度补全”转向“有证据再说”

生成式回答：删去无视觉依据的细节



Prompt

"Describe this image."

"The image features a young boy sitting on the sand at the beach, enjoying his time playing in the sand. He is wearing a white shirt and red shorts, and he has a blue toothbrush in his hand."

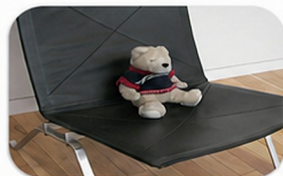


LLaVA-v1.5-7B

"The image features a young boy sitting on the sand at the beach. He is enjoying his time outdoors."



+RLSF-V



Prompt

"Describe this image."

"The image features a black leather chair with a teddy bear sitting on it. The chair is positioned in the center of the room, and the teddy bear appears to be the main focus of the scene."



LLaVA-v1.5-7B

"The image features a black chair in which a teddy bear is sitting."



+RLSF-V



Prompt

"Describe this image."

"The image features a black cell phone sitting on a wooden table. The phone is turned on and appears to be in good condition. The phone is positioned in the center of the table."



LLaVA-v1.5-7B

"The image features an old cell phone lying on a wooden surface."



+RLSF-V



Prompt

"Describe this image."

"The image features a large, fat cat standing on a toilet seat and looking into the toilet bowl. The cat appears to be curious about the water inside the toilet."



LLaVA-v1.5-7B

"The image features a cat standing on a toilet seat. The cat appears to be curious about its surroundings."



+RLSF-V

判别式回答：减少错误肯定



Prompt

"Are there two buckets in this image?"



"YES".

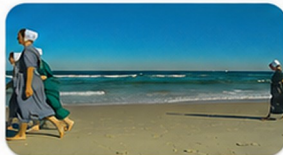


LLaVA-v1.5-7B

"NO".



+RLSF-V



Prompt

"Are there five women in this image?"



"YES".



LLaVA-v1.5-7B

"NO".



+RLSF-V



Prompt

"Is the beach dry in this image?"

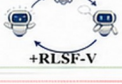


"NO".



LLaVA-v1.5-7B

"Yes".



+RLSF-V



Prompt

"Is the beach flat in this image?"



"NO".



LLaVA-v1.5-7B

"Yes".



+RLSF-V



Prompt

"Is the sea calm seas in this image?"



"NO".



LLaVA-v1.5-7B

"Yes".



+RLSF-V



有证据再说：校准描述细节与视觉证据，而非简单变短



去除无依据修饰



降低不存在对象的肯定

结论、边界与未来探索

核心结论

1 内部预测分布可以转化为有效的多模态自反馈

2 局部模糊语义 + POS选择 + Top-R + DPO形成闭环



3 10k数据 · 无外部评估器、适配多种backbone



边界

关系幻觉仍然更难；
Top-R仍需按backbone设置



未来探索

- 1 结构化实体—关系反馈
- 2 自适应Top-R
- 3 扩展纯文本LLM与其他反馈学习
- 4 内部与外部反馈协同



未来探索路线



结构化关系



自适应 Top-R



跨任务扩展



内外反馈协同