# 汇报 2

## 基于规则的 OCR 清洗策略与 VLM 描述能力对比

颜熙辰

2025 年 11 月 23 日

# 目录

# 任务一：核心问题与解决方案

**背景**：网络爬取的图片（如微博、小红书）底部常附带包含无关信息（UID、来源标识、广告）的水印条，严重影响后续数据训练质量。

**难点**：

- 水印透明度高，背景复杂，OCR 难以直接识别。
- 纯规则裁剪容易误伤正常图片内容。
- 纯 LLM 识别成本过高，且对坐标判断不精确。

**解决方案**：采用 “**图像增强 + OCR 预检 + 规则/LLM 混合决策**”

# 关键技术点 1：针对性图像增强

为解决水印“隐形”问题，代码实现了 enhance_watermark 函数。

**原理分析：**

1. **CLAHE (自适应直方图均衡化):**
   - clipLimit=3.0, tileGridSize=(8,8)
   - 作用：在局部小块内增强对比度。相比全局均衡化，它能把隐藏在复杂背景（如树叶、阴影）中的浅色水印“拉”出来。
2. **Gamma 校正 (Gamma=1.5):**
   - 作用：非线性提亮暗部细节。由于水印通常半透明且叠加在底部，Gamma 校正能显著提升文字边缘的清晰度。

```python
# 核心代码片段
clahe = cv2.createCLAHE(...)
enhanced = clahe.apply(gray)

gamma = 1.5
table = np.array([...])
enhanced = cv2.LUT(enhanced, table
    )
```

# 关键技术点 2：混合决策逻辑 (Hybrid Decision)

代码并未完全依赖 AI，而是设计了分级处理策略，兼顾**效率**与**准确率**。

## 处理流程 (Pipeline)

1. **物理裁剪预判**：仅提取图片底部 30% 区域进行分析，减少计算量。
2. **OCR 扫描**：使用 RapidOCR (CPU 版) 识别增强后的文字。
3. **一级决策（规则引擎 - 高效）**：
   - 匹配关键词黑名单（如" 微博", "uid", "©", " 摄影"）。
   - **为何有效**： 90% 的水印包含固定关键词。一旦命中，直接根据 OCR 坐标计算裁剪比例，**无需调用 LLM**，速度极快且裁剪位置像素级精准。
4. **二级决策（LLM 仲裁 - 兜底）**：
   - 仅当 OCR 识别出文字但不在黑名单中时（如路牌、衣服文字），调用 Doubao 模型进行语义判断。
   - 避免了“一刀切”导致的误裁。

# 任务二：评测概况

**评测目标**：对比不同大模型在生成“详细图片描述 (Caption)”时的准确性、细节捕捉能力及幻觉程度。

**测试集**： 5 张风格迥异的图片（风景、错位摄影、人文街拍、静物）。
**参测模型**： Gemini, GPT, Grok。
**Prompt**：统一要求生成详细的 JSON 格式描述，包含场景、光影、主体细节等。

# 模型表现对比：Gemini

**特点**：**结构化最强，分析像“理科生”**

- **优势**：
  - **字段拆分极细**：独有 `world_knowledge`（世界知识）和 `safety_and_constraints`（安全检测）字段。
  - **推理能力**：在“老鹰与人”的错位图中，明确指出了"Optical Illusion"（视觉错觉）和"Forced Perspective"（强迫透视），分析最为精准。
  - **OCR 结合**：对图片中的文字（如葡萄价格牌）有专门的 `visual_text` 字段分析。
- **不足**：语言风格偏向数据库条目，略显生硬，缺乏文学性。

典型案例（图 *2* 老鹰）：准确识别出是“人鹰合一”的错觉，并推测背景可能为驯鹰活动。

# 模型表现对比：GPT

**特点：叙事感最强，文笔像“摄影师”**

- **优势：**
  - **氛围营造：** `atmosphere_and_mood` 字段描写极佳（如 “Nostalgic” , "Cinematic"）。
  - **光影描述：** 对光线的描写（"Golden Hour", "Dappled effect"）非常细腻，适合用于指导文生图模型的 Prompt 优化。
  - **连贯性：** Overall Caption 是一段完整的、可读性极高的文章。
- **细节：** 在 “葡萄车” 图中，不仅识别了价格，还推测了生活气息（"Slice-of-life atmosphere"）。

# 模型表现对比：Grok

**特点：简洁直观，带有“极客”属性**

- **优势：**
  - **色彩量化**：唯一一个在描述中直接给出颜色 **Hex 代码**（如 #FFD700）的模型，这对设计参考非常有价值。
  - **风格识别**：倾向于识别图片的“网络属性”，例如将老鹰图描述为"Meme-like"（模因风格），非常接地气。
- **不足**：相比前两者，在部分细节的推演上略显保守，描述长度通常较短。

# 横向对比总结

| 维度 | Gemini | GPT | Grok |
|---|---|---|---|
| **结构化程度** | 高 (最佳) | 中 | 中 |
| **文学/叙事性** | 中 | 高 (最佳) | 中 |
| **错觉/逻辑推理** | 极强 | 强 | 较强 |
| **特殊功能** | 安全/隐私检测 | 氛围感描写 | **Hex 颜色代码** |
| **适用场景** | 数据清洗/打标 | 文案生成/Prompt 优化 | 快速检索/设计辅助 |

表: 三款模型在本次任务中的综合表现对比

# 收工

请哥指正