实现方法

- 推理时开启 `output_attentions`，从 `outputs.attentions[step][layer]` 中切出 `<|vision_start|>` 之后的图像 tokens，按 `(llm_h, llm_w)` reshape 成网格，再叠加到原图。
- 渲染流程：min-max 归一化 → `gamma=0.7` 提亮 → 双三次插值 → `(25,25)` 高斯平滑 → `COLORMAP_JET` 上色并与原图 0.5:0.5 融合。
- 实用观察：Layer 12–24 往往目标对齐最好；`aggregation` 看整体，`keyword` 放大某词的瞬时关注（关键词尽量用不易被 BPE 切碎的词，如 `apple`）。

核心配置片段（脚本顶部）

```
PROMPT_TEXT = "Describe the child's action. What is the baby eating?"
# 仅在 'keyword' 模式下生效：你希望锁定的关键词
# INTEREST_KEYWORDS = ["dog", "cat", "man", "woman", "car", "sky", "tree"]
INTEREST_KEYWORDS = ["apple", "eating", "fruit", "mouth", "hand", "baby",
"child"]
```

示例与对比

示例使用图片 `data/R.jpg`，提示词：`Describe the child's action. What is the baby eating?`

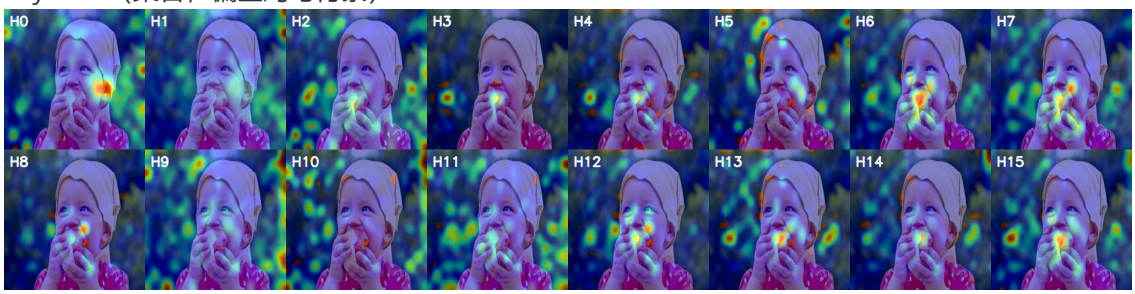
1) 输入示例

2) 同一图片，不同层的聚合策略对比

- Layer 12（聚合，主体对齐明显）

- Layer 15（聚合，主体与周边平衡）

- Layer 20（聚合，偏全局与背景）



说明：中层（12、15）更聚焦主体，深层（20）出现更多全局/背景权重，便于观察层间关注差异。

> 阅读提示：每张 `Lxx_HeadGrid` 覆盖该层所有 heads，左上角标注 `H0/H1/...`。颜色越暖（偏红黄）表示该 head 在对应图像 token 上的注意力越高。