OVOD-Agent: A Markov-Bandit Framework for Proactive Visual Reasoning and Self-Evolving Detection

Supplementary Material

6. Appendix

This appendix provides additional technical details that complement the main paper. We first present the full set of visual-action operators used by **OVOD-Agent**. Next, we provide an expanded case study that demonstrates how the agent incrementally refines its textual hypotheses using both low-level and high-level visual cues. We further include a comparison between OVOD-Agent and other LLM-based COT methods, emphasizing the differences in inference latency across these approaches. Finally, we provide the exact GPT-5 evaluation prompts and scoring rubric used to assess trajectory coherence and groundedness, ensuring transparency and reproducibility of our analysis.

6.1. Visual Actions Space

This section presents the pseudocode for the seven interpretable visual actions (a_1–a_7) used by **OVOD-Agent**, as summarized in Algorithm 2. Each action extracts a specific visual cue from the ROI (e.g. color, texture, geometry, background, lighting, or spatial relation) and maps it to a short linguistic attribute that is later used to update the evolving caption in the main reasoning algorithm.

To ensure reproducibility, all visual cues are computed using standard computer vision toolkits. Specifically, we employ **OpenCV** for RGB–HSV color conversion, K-means clustering, edge detection, and basic shape analysis; **scikit-image** for LBP/GLCM texture extraction, brightness

Algorithm 2: Visual Actions a_1–a_7

Function a_1: Dictionary**:**
  obj ← PARSE_NOUN(c)
  syn ← WORDNET_SYN(obj)
  hyp ← WORDNET_HYPER(obj)
  cand ← syn ∪ hyp
  tokens ← FILTER_VISUAL_TERMS(cand)
  phrase ← FORMAT(“a object”, tokens)
end
Function a_2: Color**:**
  hsv ← TO_HSV(r)
  clst ← KMEANS(hsv, 3)
  dom ← LARGEST_CLUSTER(clst)
  col ← HSV_TO_COLOR(dom)
  phrase ← FORMAT(“ color”, col)
end
Function a_3: Texture**:**
  lbp ← LBP(r)
  glcm ← GLCM(r)
  feats ← {lbp, glcm}
  tex ← MATCH_TEXTURE(feats)
  phrase ← FORMAT(“ texture”, tex)
end
Function a_4: Background**:**
  fg ← FG_MASK(r)
  bg ← r - fg
  clt ← BG_CLUTTER(bg)
  tag ← IF(clt ¿ τ, “cluttered”, “clean background”)
  phrase ← FORMAT(“object against ”, tag)
end
Function a_5: Geometry**:**
  ar ← ASPECT_RATIO(r)
  sc ← BBOX_SCALE(r)
  geom ← CLASSIFY_GEOM(ar, sc)
  phrase ← FORMAT(“ shaped”, geom)
end
Function a_6: Lighting**:**
  hist ← V_CHANNEL_HIST(r)
  mean ← MEAN(hist)
  var ← VAR(hist)
  cond ← IF3($mean < \tau_{\text{dark}}$, “underexposed”,
    $mean > \tau_{\text{bright}}$, “overexposed”,
    $var > \tau_{\text{shadow}}$, “shadowed”,
    “well-lit”)
  phrase ← FORMAT(“ lighting”, cond)
end
Function a_7: Spatial**:**
  rel ← SPATIAL_REL(r, layout)
  phrase ← FORMAT(“the object”, rel)
end

Table 6. **Comparison with LLM-guided reasoning modules.** OVOD-Agent achieves competitive rare-category improvements while keeping inference in the *millisecond* regime, whereas LLM-based online reasoning methods (e.g., RALF) require *second-level* latency.

Method	LVIS AP_r	AP_c	AP_f	AP_{all}	COCO AP_{50}^N	Avg Latency	Worst-case	LLM Usage
GroundingDINO (baseline) [26]	35.4	51.3	55.7	52.1	30.8	**25 ms**	25 ms	**Free**
RALF (LLM-based RAG) [18]	38.6	52.0	56.1	52.9	33.2	**1.5 s**	3.0 s	**Online**
CoT-PL (Visual CoT) [5]	37.4	51.8	55.9	52.7	32.5	**1.2 s**	2.5 s	**Offline**
DVDet (VQA-refined descriptors) [17]	36.2	51.0	55.3	52.0	31.4	**30 ms**	45 ms	**Offline**
LLMDet [10]	40.8	43.1	54.3	48.3	55.6	35 ms	50 ms	**Offline**
OVOD-Agent (Ours)	**37.0**	**52.1**	**56.3**	**52.7**	**33.4**	**55 ms**	**175 ms**	**Free**

histogram estimation, and foreground/background masking; and **NumPy/SciPy** for region-level statistics, histogram aggregation, and geometric feature computation.

These operations enable OVOD-Agent to extract stable and interpretable visual cues directly from the image, providing a consistent basis for subsequent textual refinement.

6.2. Detailed Step-by-Step Case Study

To illustrate how **OVOD-Agent** performs multi-step visual reasoning, we present a detailed case study (Fig. 4) that traces the complete prediction trajectory from an initial noun-only caption to a fully grounded, attribute-rich description. At each reasoning step, the agent executes one visual action, extracts a specific cue from the ROI (e.g. color via HSV analysis, texture via LBP/GLCM, or high-level cues from container/background geometry), converts the cue into a linguistic attribute, and updates the slot-based caption accordingly. For each step, we report: (1) the extracted visual evidence, (2) the updated caption, (3) the reward produced by the RM, and (4) the detector's grounding response (score and IoU). This case study demonstrates how progressive, attribute-aware refinement enables **OVOD-Agent** to stabilize open-vocabulary grounding even when initial predictions are incomplete or the detector temporarily fails to produce a bounding box.

6.3. Comparison with LLM-guided Methods

To demonstrate the efficiency of our **LLM-Free** paradigm, this section contrasts OVOD-Agent with representative LLM-guided modules, including RALF, CoT-PL, DVDet, and LLMDet. As summarized in Table 6, our approach eliminates the heavy dependencies that plague existing methods.

Inference Latency Bottleneck. Online reasoning methods like **RALF** are severely limited by their reliance on real-time LLM calls. Each "detection $\rightarrow$ LLM $\rightarrow$ re-detection" cycle drags the latency into the *second-level* regime (1.5 s), making them impractical for real-time deployment. While **CoT-PL**, **DVDet**, and **LLMDet** attempt to achieve faster inference (30–35 ms), they simply shift the burden to the training phase. These offline methods require massive computational resources and time to generate millions of pseudo-labels or descriptors using heavy LLMs (e.g., Qwen2-72B) before training can even begin.

The Superiority of LLM-Free Reasoning. In sharp contrast, **OVOD-Agent** is the only framework that remains entirely **LLM-Free** across both training and inference. It replaces expensive linguistic reasoning with lightweight visual actions (color, texture, geometry, spatial cues). Despite the lack of LLM intervention, OVOD-Agent achieves a competitive 37.0 AP_r on LVIS, outperforming several methods that rely on VQA-refined descriptors (e.g., DVDet at 36.2 AP_r). By formulating reasoning as a Markov-Bandit process, we achieve roughly $2.2\times$ faster inference than the base detector with reasoning, maintaining a strict millisecond latency (55 ms).

6.4. Blind GPT-5 Trajectory Scoring

For completeness, we include the prompt template used for the **blind GPT-5 evaluation**. GPT-5 does not participate in inference; it is used only to assign a continuous weak score to each sampled trajectory as an offline evaluator. To eliminate potential bias toward well-known algorithms, we implemented an **anonymized protocol** where all strategy names were replaced with generic identifiers (e.g., Strategy-A). As shown in Fig. 5, the evaluation consists of an instruction prompt (defining the evaluator's role and the **anonymization requirement**) and an input prompt containing the trajectory details. GPT-5 rates each trajectory according to the four criteria introduced in the main paper and outputs a final aggregated score in the range $[0, 5]$ in JSON format.

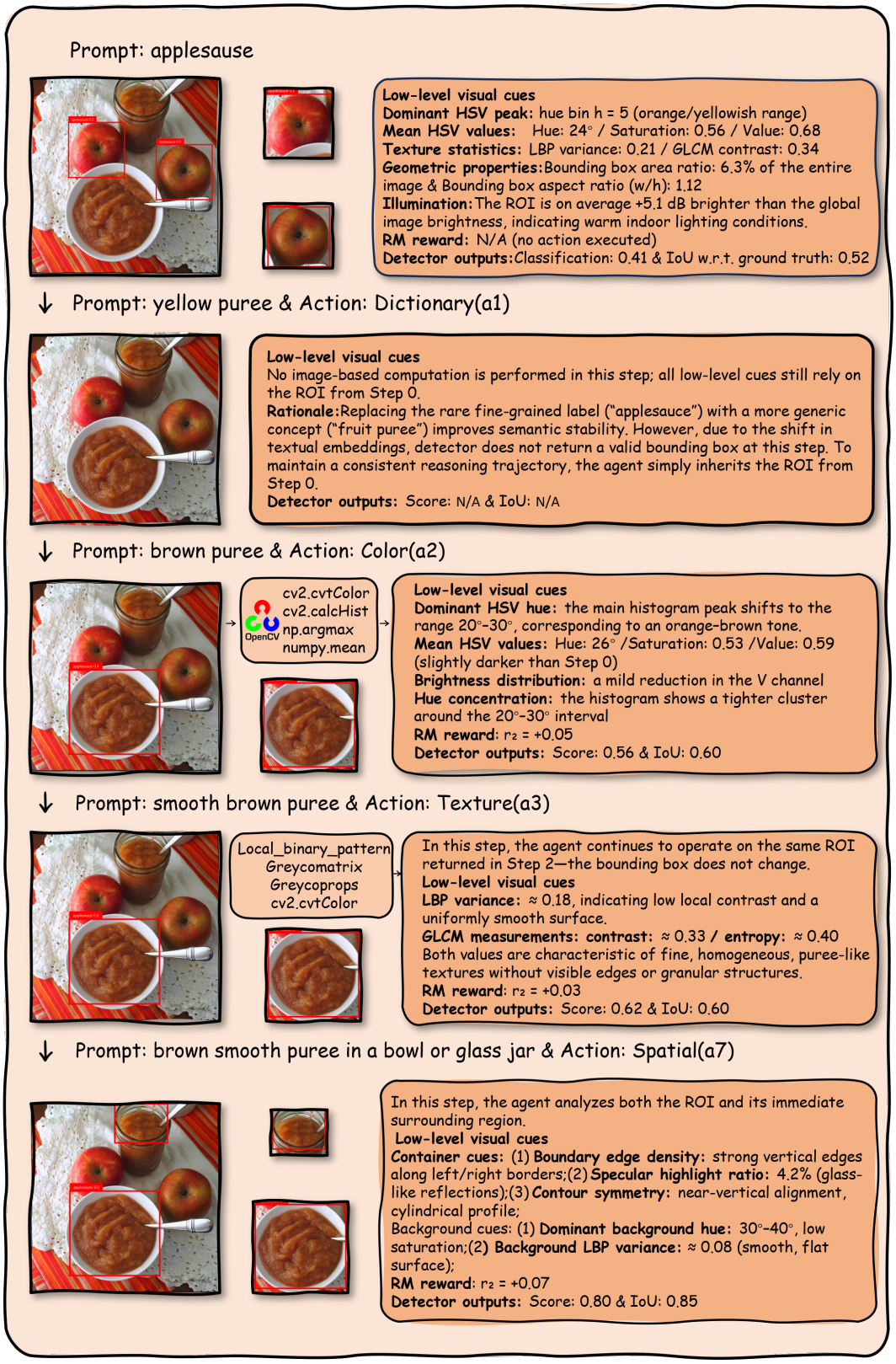

Figure 4. Step-by-step Case Study of **OVOD-Agent**, showing how visual actions (color, texture, container, background, spatial cues) progressively refine the caption and stabilize detector grounding.

---Role---
You are an expert evaluator specializing in multimodal reasoning and open-vocabulary visual grounding. Your responsibility is to assess the quality of visual reasoning trajectories. To ensure a fair and unbiased evaluation, the names of the sampling strategies have been anonymized.

---Goal---
You will rate a single trajectory sampled from an unnamed strategy. Evaluate the trajectory based solely on the provided reasoning steps and data, ignoring any potential assumptions about the underlying algorithm.

-**Trajectory Consistency:** Whether the sequence of actions forms a stable and contradiction-free reasoning process.

- **Visual Groundedness:** Whether inferred attributes (color, texture, geometry, spatial cues) match the ROI appearance.

- **Informational Gain:** Whether each step adds clearer, more specific, and more discriminative detail to the caption.

- **Detector Synergy:** Whether caption updates improve or stabilize detector confidence (score, IoU).

For each trajectory, provide a continuous score for all four criteria following the rubric above, and then compute and output the final total score in the range [0, 5].

Evaluation Instruction Prompt

---Trajectory Details---
- Strategy ID: {Anonymized_ID} (e.g., Strategy-A, Strategy-B)
- ROI description: {ROI_Text}
- Executed actions: {Action_List}
- Intermediate captions: {Caption_List}
- Detector feedback (score, IoU): {Detector_List}

Evaluate this trajectory using the four criteria listed above and provide a continuous score for each criterion.

Output your evaluation in the following JSON format:

```
{
  "Trajectory_Consistency": { "Score": "<float in [0,2.0]>", "Explanation": "<reasoning>" },
  "Visual_Groundedness":   { "Score": "<float in [0,1.5]>", "Explanation": "<reasoning>" },
  "Information_Gain":      { "Score": "<float in [0,1.0]>", "Explanation": "<reasoning>" },
  "Detector_Synergy":      { "Score": "<float in [0,0.5]>", "Explanation": "<reasoning>" },
  "Total_Score": "<float in [0,5]>"
}
```

Evaluation Input Prompt

Figure 5. Evaluation protocol for **blind** GPT-5 trajectory scoring, including the instruction prompt defining the evaluator's role and the **anonymized** input prompt to ensure unbiased assessment.