Discover, Segment, and Select: A Progressive Mechanism for Zero-shot Camouflaged Object Segmentation

Supplementary Material

7. Boundary Contact Ratio

To quantitatively measure the extent to which a predicted mask touches the image boundary, we compute the boundary contact ratio $r_b \in [0, 1]$. Given a binary segmentation mask m of size $H \times W$, we define an outer margin of width n pixels along the four sides of the image. For each side, we count the number of foreground pixels that fall within this boundary strip, excluding overlapping corner regions to avoid double counting. Let C_t, C_b, C_l, C_r, denote the number of non-zero pixels in these four edge regions, and N_e be the total number of pixels within all edge strips. The boundary contact ratio is then defined as:

$$r_b = \frac{C_t + C_b + C_l + C_r}{N_e} \tag{11}$$

In our experiments, we set the margin width n to 10 pixels. A higher r_b indicates that the predicted object is more likely to be touching or extending beyond the image boundary, which is a common characteristic of background.

8. Extended Ablation on SMS Module

We conduct a quantitative study on MLLM selection accuracy, measuring how often different strategies choose the IoU-best mask. Also discuss efficiency and inference cost compared to simpler heuristics.

Tab. 7 compares several strategies for the final Semantic-driven Mask Selection (SMS) stage. For clarity, the methods are: (1) Heuristic Rules. Purely rule-based selection using scoring function defined in Eq. (7); (2) Single-pass Selection. Feed all candidate masks to the MLLM in a single prompt and let it select the best mask; (3) Top-K Single-pass. Pre-filter candidates by heuristic scores and present only the top-K masks to the MLLM in a single prompt; (4–6) Top-K pairwise variants—iterative pairwise comparisons among the top-K masks, differing in the order of comparisons (random, ascending/descending by heuristic score). Columns CAMO and CHAMELEON summarize the task performance under each selection strategy (higher is better). The Time column indicates the average inference time (s) per image for each strategy (lower is better).

From Tab. 7, we observe distinct trends in both effectiveness and efficiency. The Heuristic Rules baseline achieves reasonable accuracy with minimal cost (1.14 s/image), demonstrating that spatial and consistency cues with boundary contactness offer a fast but approximate proxy for semantic quality. However, its rule-based nature limits its adaptability in ambiguous cases where visual cues alone are insufficient. The Single-pass Selection strategy fails due to out-of-memory (OOM) errors, as feeding all masks at once results in extremely long multimodal prompts that exceed the context length of current MLLMs, highlighting a practical limitation. Although Top-K Single-pass mitigates this, it performs poorly because presenting multiple masks in one query often causes semantic confusion for the MLLM. In contrast, pairwise strategies substantially outperform single-pass variants, confirming that decomposing the reasoning process into binary comparisons helps the MLLM make more consistent and discriminative judgments. Among them, the Top-K pairwise (ascending) variant achieves the best accuracy across both datasets, as progressively eliminating low-quality masks guides the MLLM toward the most coherent result. However, this comes at the cost of increased inference time (30 s/image) due to multiple MLLM calls. Overall, reasoning-based progressive selection, particularly with ascending order, provides the most robust balance between accuracy and efficiency, and is therefore adopted as our default SMS configuration.

Table 7. Comparison of different Mask Selection strategies. CAMO and CHAMELEON columnes measure selection accuracy (%), while Time indicates average inference cost (seconds per image). "OOM" denotes out-of-memory errors.

Method	CAMO	CHAMELEON	Time (s)
Heuristic Rules	68.80	71.05	**1.14**
Single-pass Selection	OOM	OOM	OOM
Top-K Single-pass	40.24	36.84	14.25
Top-K pairwise (random)	54.00	59.21	31.14
Top-K pairwise (descending)	50.80	43.42	30.81
Top-K pairwise (ascending)	**74.40**	**73.68**	30.59

9. Prompt Design for MLLM

For reproducibility, we provide the exact prompt templates used in our experiments for both camouflaged object localization and mask selection tasks.

Prompt for camouflaged Object localization. Following [39], physical and dynamic description of objects are included in the prompt to help MLLM better perceive the camouflaged objects. The prompt used to guide the MLLM in locating camouflaged objects and generating bounding boxes is shown in Figure 7.

Prompt for camouflaged Object localization. The prompt used to instruct the MLLM to select the best mask from candidate masks is shown in Fig. 8.

Prompt Template for Camouflaged Object Localisation

```
{
    "role": "user",
    "content": [
        {"type": "image", "image": img_path},
        {"type": "text", "text": """The image may contain a few animal/insect or human whose
    shape, color, texture, pattern and movement closely resemble its surroundings. Please
    identify them and provide their locations in the format of coordinates, as precisely as
    possible. The output should be in JSON format, e.g.:
        {
            "bbox_2d": [[x1, y1, x2, y2],[x1, y1, x2, y2]],
            "label": "dog"
        }
        DO NOT ADD ANY EXTRA INFO, JUST JSON!"""}
    ]
}
```

Figure 7. Prompt template for camouflaged object localization using MLLM.

Prompt Template for Camouflaged Object Localisation

```
{
    "role": "user",
    "content": [
        {"type": "text", "text": "The image is this."},
        {"type": "image", "image": f"{img64}"},
        {"type": "text", "text": "Overlap from the original image through mask A is this."},
        {"type": "image", "image": f"{best_maskb64}"},
        {"type": "text", "text": "Overlap from the original image through mask B is this."},
        {"type": "image", "image": f"{maskb64}"},
        {"type": "text", "text": f"""
        CAMOUFLAGE MASK COMPARISON TASK
        IMAGE: The image may contain a few animal/insect or human whose shape, color, texture,
    pattern and movement closely resemble its surroundings.
        Overlap A: Current overlap mask
        Overlap B: New candidate overlap mask
        STEP-BY-STEP PROCESS:
        1. OBJECT IDENTIFICATION:
        - Carefully examine the image
        - Identify all hidden/concealed objects and their exact locations.
        2. SELECTION CRITERIA:
        - PRIMARY: Choose the mask that covers ALL identified objects completely
        - SECONDARY: Among masks that meet primary criterion, choose the one with least
    background
        - If no mask covers all objects, choose the one that covers the most objects
        OUTPUT JSON (DO NOT ADD ANY EXTRA INFO, JUST JSON!):
        {
            "better_mask": "Mask A" / "Mask B",
        }
        """},
    ],
}
```

Figure 8. Prompt template for mask selection using MLLM.