# Decoupled Clip and Dynamic Sampling Policy Optimization for Food Reasoning Segmentation

ACM multimedia MMFood'25 Dublin, Ireland 27-31.10.2025

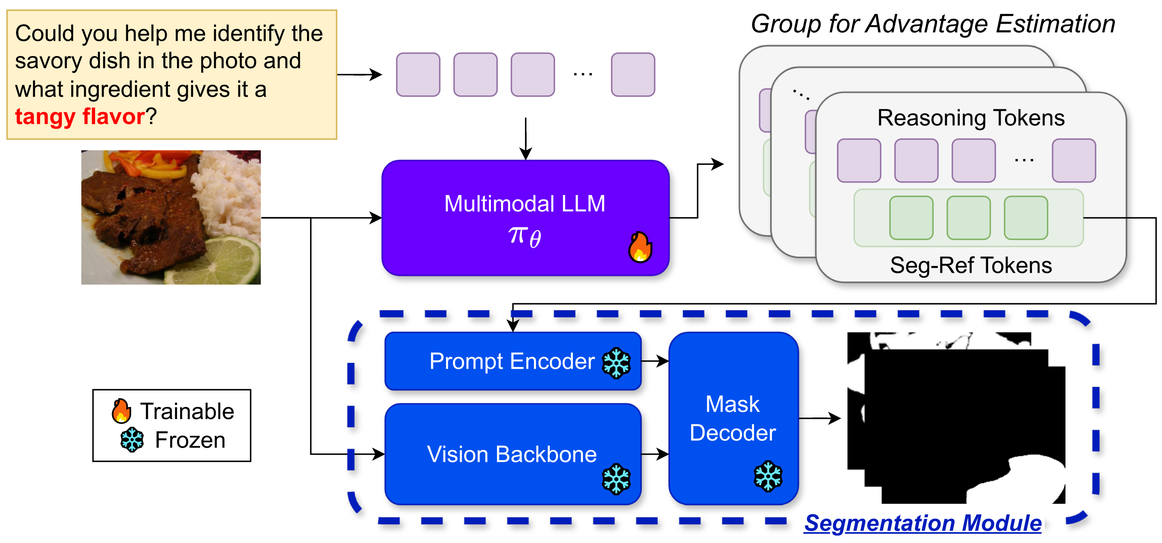
Hikaru Tanabe, Keiji Yanai
The University of Electro-Communications, Japan

## Background

- **Food Reasoning Segmentation** requires models to identify pixel-level target regions in food images based on **implicit textual queries**.
- Conventional **SFT-based models [1, 2] demonstrate limited reasoning capability**, while RFT has been shown to enhance reasoning performance of LLMs [3].
- We validate **GRPO-based RFT [4]** with a task-specific reward design for food reasoning segmentation and further **refine the optimization based on DAPO [5]**.

## Method

### Model Overview




① **Multimodal LLM (MLLM)** generates *segmentation references*, consisting of bounding boxes and points to the implied object.
② **Segmentation module** then produces a region mask for each *segmentation reference*.

### RL Objectives

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left( \min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon\Big)\hat{A}_{i,t}\right) - \beta D_{\text{KL}}(\pi_\theta||\pi_{\text{ref}})\right)\right]$$

$$\mathcal{J}_{\text{DAPO}}(\theta) = \mathbb{E}_{(q,a)\sim\mathcal{D},\{o_i\}_{i=1}^{G}\sim\pi_{\theta_{\text{old}}}(\cdot|q)} \left[ \frac{1}{\sum_{i=1}^{G}|o_i|}\sum_{i=1}^{G}\sum_{t=1}^{|o_i|}\min\left(r_{i,t}(\theta)\hat{A}_{i,t},\ \text{clip}\Big(r_{i,t}(\theta), 1-\varepsilon_{\text{low}}, 1+\varepsilon_{\text{high}}\Big)\hat{A}_{i,t}\right)\right]$$

$$\text{s.t.}\quad 0 < \Big|\{o_i \mid \texttt{is\_equivalent}(a, o_i)\}\Big| < G$$

$$\text{where}\quad \hat{A}_{i,t} = \frac{R_i - \text{mean}(\{R_j\}_{j=1}^{G})}{\text{std}(\{R_j\}_{j=1}^{G})}, r_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t}\mid q, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t}\mid q, o_{i,<t})}.$$

**Key modifications to GRPO [4] based on DAPO [5]**

① **Clip-Higher:** increases the upper clipping threshold to amplify gradients for low-probability tokens.
② **Dynamic Sampling:** resamples model outputs when the rewards are identical, preventing gradient vanishing.
③ **Token-level Policy Gradient:** averages the policy gradient loss across tokens, enhancing training stability in longer CoT scenarios.
④ **Removing KL term:** eliminates the KL regularization to reduce unnecessary constraints and expand exploration space.

### Reward Design

$$R = R_{\text{fmt}} + R_{\text{acc}} + R_{\text{rep}}$$

① **Format Reward** $R_{\text{fmt}} = R_{\text{tf}} + R_{\text{sf}}$
- fixes the output format <think>…</think><answer>{…}</answer>

② **Accuracy Reward** $R_{\text{acc}} = R_{\text{IoU}}^{\text{bbox}} + R_{\text{L1}}^{\text{bbox}} + R_{\text{L2}}^{\text{pt}}$
- improves the *segmentation references*

③ **Non-Repeat Reward** $R_{\text{rep}}$
- penalizes sentence-level repetitions

**Accuracy Reward Criteria**

$$\text{IoU}(\hat{b}_i, b_j) > \tau_{\text{IoU}}$$
$$\|\hat{b}_i - b_j\|_1 < \epsilon_{\text{bbox}}$$
$$\|\hat{p}_i - p_j\|_2 < \epsilon_{\text{pt}} \ \wedge\ \hat{p}_i \in \hat{b}_i$$

[1] Xin Lai et al. LISA: Reasoning segmentation via large language model. In Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024.
[2] Yuehao Yin et al. FoodLMM: A versatile food assistant using large multi-modal model. arXiv preprint arXiv:2312.14991, 2024.
[3] Daya Guo, et al. DeepSeek-R1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948, 2025.
[4] Zhihong Shao et al. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300, 2024.
[5] Qiying Yu et al. DAPO: An open-source llm reinforcement learning system at scale. arXiv preprint arXiv:2503.14476, 2025.

## Experiments

### Implementation Details

- **MLLM**: Qwen2.5-VL 3B, **Segmentation Module**: SAM2
  - (SAM2: temporal modules were removed from the original model)
- $\varepsilon_{\text{low}} = 2.0, \varepsilon_{\text{high}} = 2.8, \tau_{\text{IoU}} = 0.5, \epsilon_{\text{bbox}} = 10, \epsilon_{\text{pt}} = 50, G = 8$
- **FoodReasonSeg-Single:** all conversations in FoodReasonSeg [2] are converted into single-turn question-answer pairs.
  - Boxes and points are extracted from the ground-truth masks. No ground-truth reasoning data is used for training.

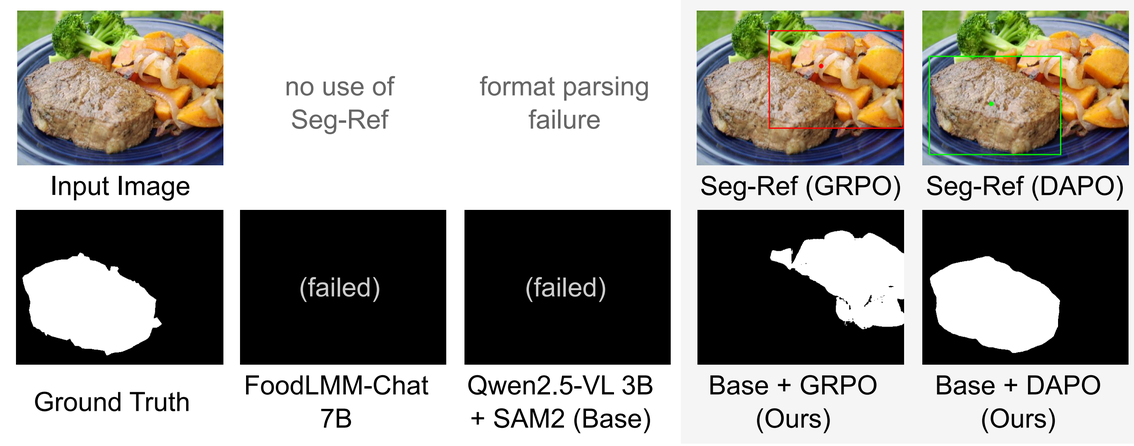
### Results on FoodReasonSeg-Single

| Method | gIoU | cIoU | |
|---|---|---|---|
| Qwen2.5-VL 3B + SAM2 | 0.07 | 0.08 | |
| LISA 7B [1] | 0.19 | 0.19 | SFT |
| FoodLMM-Chat 7B [2] | 0.25 | 0.20 | SFT |
| Ours (3B, GRPO) | 0.46 | 0.41 | RFT |
| Ours (3B, DAPO) | **0.51** | **0.45** | RFT |

- Our RFT-based 3B models performed more accurate segmentation than SFT-based 7B models.
- Especially the DAPO model outperformed the GRPO one.

### Segmentation Results

Query: Can you identify an ingredient from this image that **helps with muscle recovery**?




**FoodLMM-Chat 7B**
Yes, indeed. **Pumpkin** [SEG1] is a great source of antioxidants, particularly beta-carotene, which is known for aiding in muscle recovery.

**Qwen2.5-VL 3B + SAM2**
(incorrect format for parsing segmentation references)

**+ GRPO (Ours)**
<think>The image shows a plate with a piece of meat, ...(omitted)... Among the vegetables, **sweet potatoes** are a good source of carbohydrates, which can help with muscle recovery.</think><answer>...</answer>

**+ DAPO (Ours)**
<think>The ingredient that helps with muscle recovery is likely to be a protein source, such as beef or chicken. In this image, the main protein source is **the steak**.</think><answer>...</answer>

### Segmentation

- Our models successfully generate accurate segmentation masks.
- The DAPO-based model precisely estimates the ground-truth regions, demonstrating improved reasoning-based localization.

### Multimodal LLM

- Our models responses in a well-structured, parseable format.
- The DAPO-based model effectively reasons that the implied object is "steak" and focuses on the corresponding image region, showing enhanced reasoning ability.

## Conclusion

We introduced an RFT-based framework for food reasoning segmentation, validating the enhanced learning effectiveness of DAPO-based optimization and a task-specific reward design.

### Limitation and Future Work

- More comprehensive evaluation on out-of-domain scenarios
- Further exploration of the reward design, a tightly-coupled model architecture, and hybrid training strategy incorporating SFT stage.