Saliency-R1: Incentivizing Unified Saliency Reasoning Capability in MLLM with Confidence-Guided Reinforcement Learning

Long Li[1] Shuichen Ji[1] Ziyang Luo[1] Zhihui Li[2]
Dingwen Zhang[1] Junwei Han[1] Nian Liu[1†]
[1] Northwestern Polytechnical University [2] University of Science and Technology of China

Figure 1. Existing multimodal large language models (MLLMs) exhibit limitations in saliency reasoning. This paper proposes Saliency-R1 to incentivize unified saliency reasoning of MLLM across three representative tasks, *i.e.*, Salient Object Detection (SOD), Salient Instance Segmentation (SIS), and Co-salient Object Detection (CoSOD).

Abstract

Although multimodal large language models (MLLMs) excel in high-level vision-language reasoning, they lack inherent awareness of visual saliency, making it difficult to identify key visual elements. To bridge this gap, we propose Saliency-R1, the first unified MLLM framework that jointly tackles three representative and heterogeneous saliency tasks: Salient Object Detection (SOD), Salient Instance Segmentation (SIS), and Co-salient Object Detection (CoSOD), enhancing the model's capacity for saliency reasoning. We introduce a textual interface with structured tags (<rg>, <ins>) to encode region- and instance-level referring expressions, enabling a single referring segmenter to produce task-appropriate masks. To train the MLLM efficiently, we propose Confidence-Guided Policy Optimization (CGPO), a novel single-sample reinforcement learning algorithm. CGPO improves on GRPO by replacing group-normalized advantages with a per-sample signal based on reward–confidence discrepancy, thereby reducing computational waste, mitigating signal dilution, and lowering training overhead. Our model exceeds or matches the performance of robust open/closed-source MLLMs and specialized state-of-the-art methods across all three tasks, demonstrating the efficacy of our framework in saliency reasoning.

1. Introduction

Multimodal large language models (MLLMs) [10, 21, 26, 45, 56, 59, 72] integrate visual perception with linguistic reasoning, demonstrating remarkable capabilities in complex reasoning tasks. They excel in generalizing across various vision-language tasks, *e.g.*, visual question answering [18, 54], image captioning [1, 41], and referring expression comprehension [6, 57], which *rely on extensive world knowledge*. However, in the realm of saliency detection, tasks that *rely on scene-specific visual context* to identify visually prominent elements, MLLMs often underperform [11], as illustrated in Figure 1. This shortcoming indicates that current MLLMs lack inherent *saliency awareness*: they can comprehend *"what is visually described"* but struggle to identify *"what is visually prominent"* in a scene.

To bridge this gap, we aim to enhance the reasoning capabilities of MLLM in the context of saliency detection. This field encompasses various tasks, with the most representative being Salient Object Detection (SOD), Salient Instance Segmentation (SIS), and Co-salient Object Detection (CoSOD). SOD [33–35, 53, 73] segments all visually promi-

†Corresponding author: liunian228@gmail.com.

nent regions in a single image (*e.g.*, "a kitten" and "a stuffed animal"); SIS [17, 40, 55] distinguishes individual salient instances within a single image (*e.g.*, "a red cup on the left" and "a blue cup on the right"); CoSOD [29, 30, 47, 71] identifies common salient objects that belong to the same semantic category across a group of images (*e.g.*, "rubber duck"); The collection of these three tasks represents the core paradigms of modern saliency detection, covering single-image and image-group inputs, intra-image and inter-image modeling, and output granularity that ranges from region-level segmentation (unions) to instance-level disambiguation. Thus, to enable MLLMs to perform comprehensive saliency detection, we integrate these three representative tasks into a unified MLLM-based framework.

Since all three saliency tasks require segmentation outputs, we need to convert the MLLM's reasoning output into segmentation masks. Recent MLLM-based segmentation methods [8, 25, 32, 50, 52] address this by creating structured prompt representations that connect language reasoning with mask generation, typically inputting these into a promptable segmentation model (*e.g.*, SAM [24]). These prompt representations can be categorized into two paradigms. The first uses learnable special tokens (*e.g.*, `<SEG>` [25, 50]) to activate the SAM mask decoder. The second relies on geometric primitives, *e.g.*, bounding boxes, points [8], or attention-derived keypoints [32], as prompts for the SAM-style model to generate masks. Unfortunately, these designs are originally devised for single object segmentation and fail to capture the distinct output granularities needed for the three tasks: SOD (multi-category region unions), SIS (disambiguated instances), and CoSOD (single-category regions across a group of images).

To overcome this limitation, we propose a unified textual referring interface that employs text-based region-level and instance-level referring expressions, combined with a referring segmentation model to generate task-specific segmentation masks. This approach enables a single MLLM to effectively tackle all three tasks without requiring architectural modifications. Furthermore, our approach not only unifies these tasks but also maximizes the MLLM's text reasoning capabilities, facilitating easier training and application across diverse scenarios.

As for the training of saliency reasoning in the MLLM, we employ a two-stage training pipeline: Supervised Fine-Tuning (SFT) followed by Reinforcement Learning (RL) [19]. Although GRPO [42] is widely adopted in RL training, it suffers from three significant limitations: (1) it lacks confidence-aware learning, leading to wasted computation on uninformative responses; (2) reward and penalty signals are diluted due to group-mean baselining advantage calculations; and (3) it incurs excessive overhead from multi-sample generation and KL regularization. To address these issues, we propose Confidence-Guided Policy Optimization (CGPO), a novel RL algorithm that incorporates model reasoning confidence based on Bayesian calibration theory [12]. CGPO leverages the reward and confidence discrepancy as a per-sample advantage calculation, which not only minimizes wasted computation on uninformative responses but also prevents signal dilution by bypassing group-mean calculation. Moreover, this group-independent advantage calculation permits streamlining the algorithm process to single-sampling, substantially reducing overhead.

In summary, our main contributions are as follows:

- We are the first to incentivize the unified saliency reasoning capability for MLLM, using a unified textual interface to comprehensively tackle three representative and heterogeneous tasks *i.e.*, SOD, SIS, and CoSOD.
- We present CGPO, a lightweight single-sample RL algorithm that innovatively utilizes reward-confidence discrepancy as a per-sample advantage signal, reducing computational waste, mitigating signal dilution, and minimizing overhead.
- Our model matches or exceeds the performance of closed- and open-source MLLMs and most specialized SOTA methods across all three tasks.

2. Related Work

2.1. Saliency Detection

SOD. SOD aims to segment visually salient regions in an image. Early CNN-based methods [33] integrated global and local contexts, and label decoupling [53] refined boundaries. Later, part–whole modeling [35, 73] improved object completeness, while Transformer-based models [34] enabled unified global reasoning and detail refinement [63].

SIS. SIS extends SOD to category-agnostic, instance-level segmentation [17, 39]. Early multi-stage methods [27] suffered from error accumulation; end-to-end models [17, 55] improved robustness. Recent Transformer-based approaches [40] use query-driven, anchor-free detection, eliminating hand-crafted steps like NMS [49].

CoSOD. CoSOD identifies common salient objects across multiple related images [16, 44], relying on cross-image consensus. Recent methods achieve this via strategies including cross-image attention [66], democratic prototyping [62], group consistency [71], region mining [30], and foundation model adaptation with consensus prompts [47].

Despite progress in task-specific models, saliency reasoning remains fragmented. We unify SOD, CoSOD, and SIS by incentivizing unified saliency reasoning in MLLM with confidence-guided reinforcement learning.

2.2. MLLM-based Segmentation

Recent works [8, 25, 32, 50, 52] adapt MLLMs for segmentation by linking language reasoning with mask generation via structured interfaces. LISA [25] and SegLLM [50] introduce a `<SEG>` token whose hidden state prompts a

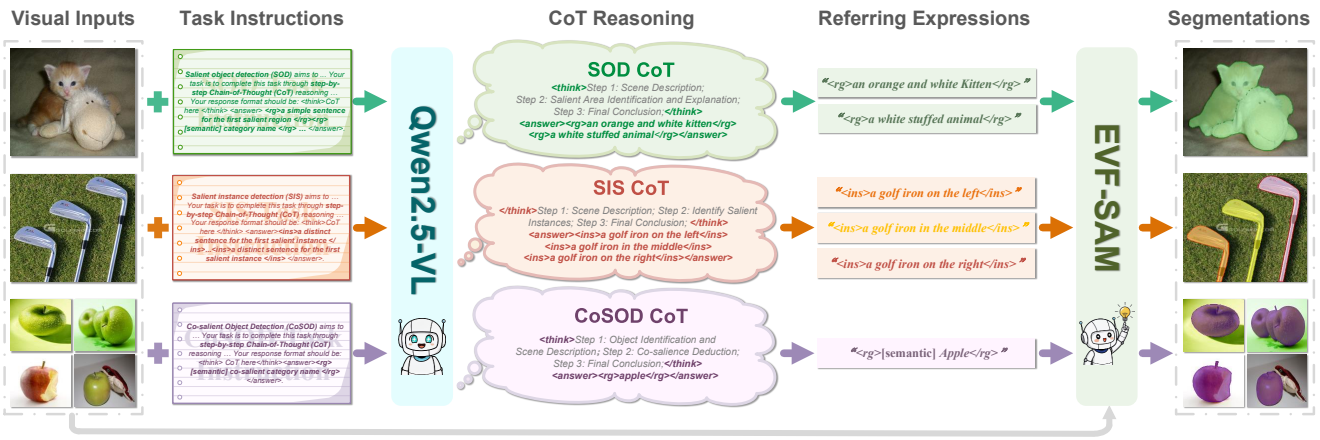

Figure 2. **Overview of the Saliency-R1 framework.** Given visual inputs and task-specific instructions, the MLLM (QWen2.5-vl) generates structured CoT reasoning, where referring expressions can be parsed and processed by a referring segmentation model (EVF-SAM) to produce task-specific masks for SOD, SIS, and CoSOD.

frozen mask decoder (*e.g.*, SAM [24]). HyperSeg [52] and SAM4MLLM [8] guide SAM in segmentation by transforming semantic information into geometric prompts, while LENS [32] enhances segmentation by utilizing keypoints extracted from attention maps.

However, these methods overlook the varied output granularities across SOD, SIS, and CoSOD. We introduce a structured textual interface that encodes task-aware granularity, enabling one segmenter to seamlessly handle all three tasks.

2.3. Advancements in GRPO

Recent advances have refined GRPO [42] to enhance the stability and scalability of reinforcement learning. DAPO [61] decouples clipping bounds and applies dynamic sampling to enhance exploration and filter uninformative batches. Dr-GRPO [36] provides an unbiased objective by removing response-length and group-variance normalization, improving token efficiency. GSPO [70] instead defines sequence-level importance ratios for more stable training.

Unlike them, our CGPO uses model confidence for per-sample advantage estimation, enabling the model to optimize informative responses. This also reduces multi-sample generation, minimizes response interference, and significantly lowers computational burden.

3. Proposed Method

3.1. Framework Construction

Overview. As shown in Figure 2, given a visual input $\boldsymbol{I}$ and a task instruction $\tau \in \{\tau^{SOD}, \tau^{SIS}, \tau^{CoSOD}\}$, the MLLM (Qwen2.5-VL [5]) generates a structured Chain-of-Thought (CoT) response, from which region- or instance-level referring expressions are parsed to guide a referring segmenter $\mathcal{S}$ (EVF-SAM [67]) in producing segmentation masks with task-specific granularity across all three tasks.

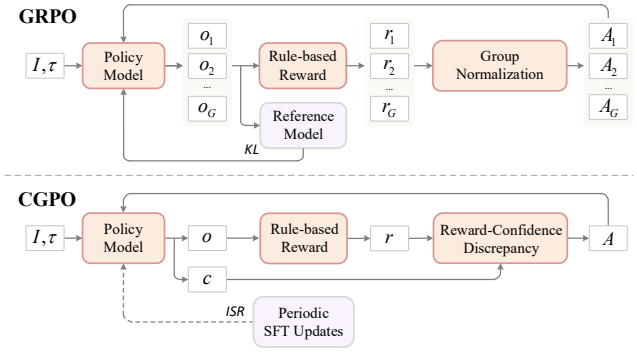

Figure 3. **Comparison between GRPO and our proposed CGPO.** GRPO uses multiple responses, group-normalized advantages, and KL regularization with a reference model. In contrast, CGPO employs a single response, calculates advantage with reward-confidence discrepancy, and replaces KL with ISR (Interleaved SFT Regularization).

Input structure. The visual input $\boldsymbol{I}$ and instruction τ are task-specific:

- **SOD/SIS**: $\boldsymbol{I} = \{x^{\text{SOD}}\}/\{x^{SIS}\}$ (single image); τ^{SOD} prompts step-by-step reasoning to identify all visually prominent regions, while τ^{SIS} requires enumerating all distinct salient instances with disambiguating attributes (*e.g.*, location, color).
- **CoSOD**: $\boldsymbol{I} = \{x_k^{\text{CoSOD}}\}_{k=1}^{K}$ (a group of K images); τ^{CoSOD} guides the model to deduce the single shared semantic category across the group.

Canonical CoT format. The MLLM output follows:

$$\texttt{<think>}\,\mathcal{T}\,\texttt{</think>}\,\texttt{<answer>}\,\mathcal{E}\,\texttt{</answer>}, \tag{1}$$

where $\mathcal{T}$ is a three-step reasoning process (scene/object identification → task analysis → conclusion), and $\mathcal{E}$ is the referring expression and can be parsed via string matching.

We design $\mathcal{E}$ using two structured textual tags to explic-

itly encode granularity heterogeneity: (1) <rg>···</rg>: denotes a salient region tied to one semantic category. For SOD, where multiple salient regions of the same or different semantic categories may coexist (*e.g.*, "an orange kitten" and "a white stuffed animal"), we use a combination of multiple <rg> blocks to refer to each region separately. For CoSOD, a single <rg> expresses the shared category across all images. (2) <ins>···</ins>: denotes an instance-level descriptor enriched with distinguishing attributes, used exclusively in SIS to disambiguate co-occurring salient instances. These tags avoid the limitations of implicit interfaces (e.g., <SEG>) that bind all salient units into a single hidden-state embedding, making multi-target segmentation infeasible without architectural redesign.

Finally, $\mathcal{E}$ for three tasks are formulated as:

$$\mathcal{E} := \begin{cases} \{\texttt{<rg>}d_i\texttt{</rg>}\}_{i=1}^{N} & \text{(SOD)}, \\ \{\texttt{<ins>}\phi_j\texttt{</ins>}\}_{j=1}^{M} & \text{(SIS)}, \\ \texttt{<rg>[semantic] } s\texttt{</rg>} & \text{(CoSOD)}. \end{cases} \tag{2}$$

where d_i is a single-category region description (*e.g.*, "an orange kitten" or "[semantic] person") for SOD, ϕ_j is an attribute-rich instance descriptor (*e.g.*, "a red cup on the left"), and s is the single shared semantic category for CoSOD (*e.g.*, "[semantic] rubber duck"). N is the number of salient regions in SOD, and M is the number of salient instances in SIS.

Referring segmentation. Given $\boldsymbol{I}$ and the parsed referring expressions $\mathcal{E} \in \{\mathcal{E}^{SOD}, \mathcal{E}^{SIS}, \mathcal{E}^{CoSOD}\}$, the referring segmenter $\mathcal{S}$ produces the final mask(s) $\boldsymbol{M}$ as:

$$\boldsymbol{M} = \begin{cases} \bigvee_{i=1}^{N} \mathcal{S}\big(x^{\text{SOD}}, \mathcal{E}_i^{\text{SOD}}\big) & \text{(SOD)}, \\ \big\{\mathcal{S}\big(x^{\text{SIS}}, \mathcal{E}_j^{\text{SIS}}\big)\big\}_{j=1}^{M} & \text{(SIS)}, \\ \big\{\mathcal{S}\big(x_k^{\text{CoSOD}}, \mathcal{E}^{\text{CoSOD}}\big)\big\}_{k=1}^{K} & \text{(CoSOD)}. \end{cases} \tag{3}$$

where $\bigvee$ denotes pixel-wise logical OR (*i.e.*, union) of binary masks, merging all <rg> regions into a single salient mask for SOD. This design enables a single segmentor $\mathcal{S}$ (EVF-SAM) to handle all three tasks without architectural modification.

3.2. CoT Alignment

We train the MLLM to generate high-quality, task-aligned CoT reasoning in two stages. First, Supervised Fine-Tuning (SFT) initializes the model using a curated dataset of ground-truth CoT sequences. Second, Reinforcement Learning (RL) refines the policy with a composite reward that jointly assesses reasoning format and segmentation accuracy (see supplementary material). This encourages structured, task-accurate CoT outputs.

3.2.1. Supervised Fine-Tuning

We curate the SFT data using an "output-to-reasoning" strategy. Specifically, we prompt a powerful closed-source MLLM (*i.e.*, Gemini [10]) with both the input image(s) and the corresponding ground-truth segmentation mask(s), and ask it to explain why the highlighted regions are salient. The resulting CoT explanations are treated as ground-truth reasoning traces for supervised training. This ensures logical consistency between the CoT and the ground-truth masks.

The MLLM is trained via the standard autoregressive cross-entropy loss to maximize the log-likelihood of the ground-truth CoT sequences. Formally, for visual input $\boldsymbol{I}$, task instruction τ, and the corresponding ground-truth CoT response $o^{\text{gt}} = (o_1^{\text{gt}}, \ldots, o_{T^{\text{gt}}}^{\text{gt}})$, the SFT objective is:

$$\mathcal{J}_{\text{SFT}}(\theta) = \mathbb{E}_{(\boldsymbol{I},\tau,o^{\text{gt}})}\left[\sum_{t=1}^{T^{\text{gt}}} \log \pi_\theta\big(o_t^{\text{gt}} \mid \boldsymbol{I}, \tau, o_{<t}^{\text{gt}}\big)\right]. \tag{4}$$

Here, $\pi_\theta(\cdot \mid \boldsymbol{I}, \tau, o_{<t})$ denotes the MLLM's autoregressive policy parameterized by θ, *i.e.*, the probability distribution over the next token given the visual input, task instruction, and previously generated tokens.

3.2.2. Reinforcement Learning

Preliminary. To contextualize our CGPO, we briefly review GRPO [42], a group-based RL algorithm for optimizing autoregressive language policies. As illustrated in Figure 3, for each input pair $(\boldsymbol{I}, \tau)$, GRPO samples G candidate responses $\{o_i\}_{i=1}^{G}$, computes a verifiable reward $r_i \in [0, 1]$ for each, and constructs a group-normalized advantage,

$$A_i = \frac{r_i - \mu_r}{\sigma_r}, \tag{5}$$

where μ_r and σ_r are the empirical mean and standard deviation of the G rewards. The policy update uses the importance sampling ratio:

$$\rho_{i,t}(\theta) = \frac{\pi_\theta(o_{i,t} \mid \boldsymbol{I}, \tau, o_{i,<t})}{\pi_{\theta_{\text{old}}}(o_{i,t} \mid \boldsymbol{I}, \tau, o_{i,<t})}, \tag{6}$$

combined with a PPO-style clipped objective:

$$\mathcal{C}_i = \frac{1}{T_i}\sum_{t=1}^{T_i} \min\left(\rho_{i,t}(\theta)A_i,\ \text{clip}(\rho_{i,t}(\theta), 1-\varepsilon, 1+\varepsilon)A_i\right), \tag{7}$$

and a KL penalty against a fixed reference policy π_{ref}:

$$\mathcal{J}_i^{\text{GRPO}} = \mathcal{C}_i - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}}). \tag{8}$$

The overall objective is:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}_{(\boldsymbol{I},\tau),\{o_i\}}\left[\frac{1}{G}\sum_{i=1}^{G} \mathcal{J}_i^{\text{GRPO}}\right]. \tag{9}$$

Confidence-Guided Policy Optimization. Despite its widespread use, GRPO faces three key limitations that impede learning efficiency and policy quality.

First, GRPO is confidence-agnostic: it computes advantages solely from group-level reward statistics and ignores the model's confidence. By discarding confidence, GRPO fails to prioritize informative responses, *i.e.*, HrLc (high-reward & low-confidence; correct but under-explored) and LrHc (low-reward & high-confidence; overconfident errors). As shown in Figure 4, these two types constitute only 45.8% of all responses, while the majority (54.2%) belong to non-informative categories: HrHc (high-reward & high-confidence; already well-learned) and LrLc (low-reward & low-confidence; unreliable). Crucially, these non-informative responses nonetheless receive non-negligible positive or negative advantages, leading to unnecessary parameter updates and wasted computation on samples that contribute little to policy improvement.

Second, the shared group-mean baseline causes signal dilution: advantages for informative responses are compressed. As shown in Figure 4, HrLc responses receive only modestly positive values, while LrHc responses incur insufficiently negative penalties. This occurs because dominant HrHc responses elevate the group mean, suppressing HrLc signals, and LrLc outliers depress it, softening penalties for LrHc, diluting learning signals and hindering policy improvement.

Third, GRPO incurs prohibitive computational overhead: it requires G forward passes, G reward evaluations, and storage of G KV caches per input due to group sampling, and further adds cost from the KL divergence penalty against a fixed reference model.

To this end, we propose Confidence-Guided Policy Optimization (CGPO), as shown in Figure 3, a lightweight single-sample RL algorithm that explicitly accounts for the model's confidence during policy updates to overcome GRPO's key limitations. Specifically, we reinterpret advantage estimation through a Bayesian lens [23]: the mean token-wise confidence

$$c = \frac{1}{T}\sum_{t=1}^{T} \pi_\theta(o_t \mid \boldsymbol{I}, \tau, o_{<t}) \tag{10}$$

serves as a proxy for the prior belief that the generated CoT is correct, while the task-specific reward $r \in [0, 1]$ acts as observational evidence of correctness. The ideal learning signal should be designed to align prior belief with observational evidence, a principle known as calibration [12, 37]. Since r is a soft estimation of correctness [43], the natural misalignment measure is the binary cross-entropy:

$$\mathcal{L}_{\text{NLL}} = -r \log c - (1 - r) \log(1 - c). \tag{11}$$

However, instead of minimizing $\mathcal{L}_{\text{NLL}}$ directly, we extract a policy update direction from its gradient. Since policy gradient methods maximize an objective, we follow the *negative*

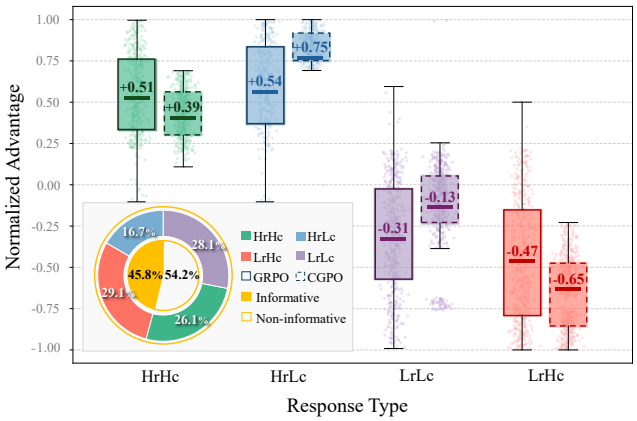

Figure 4. **Joint Response-Type Distribution and Per-Sample Advantage Analysis.** We analyze 8,000 responses generated by the SFT-initialized model, categorizing them into four types based on the bottom/top 20% empirical thresholds of reward and model confidence. Confidence is defined as the mean token-wise generation probability of the CoT output (Eq. 10). Advantages for GRPO and CGPO are rank-normalized [7] to $[-1, 1]$, respectively, to ensure fair comparison.

gradient of $\mathcal{L}_{\text{NLL}}$ with respect to c:

$$-\nabla_c \mathcal{L}_{\text{NLL}} = \frac{r - c}{c(1 - c)}. \tag{12}$$

The numerator $(r - c)$ captures the discrepancy between reward (evidence) and confidence (prior), *i.e.*, Bayesian surprise [22]. The denominator, however, causes numerical instability near $c \to 0$ or $c \to 1$. Discarding this unstable scaling factor, we adopt the signed error $(r - c)$ as our per-sample advantage:

$$A = r - c. \tag{13}$$

As shown in Figure 4, our design yields advantages for non-informative responses (HrHc and LrLc) that are weaker than GRPO, thereby suppressing unnecessary updates. Meanwhile, by leveraging the per-sample reward–confidence discrepancy as the advantage signal, CGPO alleviates signal dilution and ensures that informative responses (HrLc and LrHc) receive stronger (more positive or more negative) signals. Furthermore, requiring only a single reasoning trace, CGPO removes group sampling and substantially reduces the training overhead.

In addition to the single-sample advantage design, we simplify policy regularization by replacing GRPO's static KL divergence penalty with Interleaved SFT Regularization (ISR). Specifically, we alternate between RL updates and SFT steps based on a fixed period P.

The RL update uses the single-sample advantage and

follows a PPO-style clipped objective:

$$\mathcal{J}_{\text{RL}}(\theta) = \mathbb{E}_{(\boldsymbol{I},\tau),o}\Bigg[\frac{1}{T}\sum_{t=1}^{T}\min\Big(\rho_t(\theta)\,A, \\ \text{clip}\big(\rho_t(\theta), 1-\varepsilon, 1+\varepsilon\big)\,A\Big)\Bigg]. \tag{14}$$

The SFT update uses the objective defined in Eq. (4). We alternate between RL and SFT updates every P steps, yielding the final CGPO objective:

$$\mathcal{J}_{\text{CGPO}}(\theta) = \begin{cases} \mathcal{J}_{\text{RL}}(\theta), & \text{if } s \bmod P \neq 0, \\ \mathcal{J}_{\text{SFT}}(\theta), & \text{if } s \bmod P = 0, \end{cases} \tag{15}$$

where s is the current training step.

3.2.3. Reward Definition

During RL training, the total reward r combines a correctness term and a format term:

$$r = \lambda r_{\text{corr}} + (1-\lambda) r_{\text{fmt}}, \quad \lambda = 0.5. \tag{16}$$

r_{corr} measures segmentation quality, *i.e.*, S-measure [13] for SOD/CoSOD and our proposed Instance-Aligned S-Measure (IASM) for SIS, which simultaneously considers both instance detection and segmentation performance. r_{fmt} encourages well-structured, tag-consistent CoT outputs. The details of the IASM formulation and format scoring rules are provided in the supplementary material.

4. Experiments

4.1. Evaluation Datasets and Metrics

We evaluate our unified framework on standard benchmarks for all three tasks: SOD with DUT-O [60], ECSSD [58], and PASCAL-S [31]; SIS using ILSO [27], SIS10K [40], and SOC [14]; and CoSOD with CoSOD3k [16], CoCA [69], and CoSal2015 [65]. For SOD and CoSOD, we select the following commonly used evaluation metrics: Structure-measure (S_m) [13], max F-measure (F_{max}) [2], E-measure (E_m) [15], and Mean Absolute Error ($\mathcal{M}$) [9]. For SIS, we adopt the standard Average Precision (AP) metrics, specifically AP50 and AP70, which are commonly used for evaluating instance-level segmentation tasks.

4.2. Implementation Details

Training Data. For SFT data construction, we use Gemini-2.5-Pro [10] to generate CoT annotations conditioned on ground-truth masks, collecting: 2640 images from DUTS-TR [48] for SOD, 2467 images from ILSO [27], SIS10K [40], and SOC [14] for SIS, and 1262 image groups (8 images each) from CoCoSeg [46] for CoSOD. For CGPO training data, we use 3000 examples per task from the same sources, comprising 1800 held-out examples for RL updates and 1200 SFT examples for ISR.

Model Architecture. We employ Qwen2.5-VL-7B-Instruct [5] as our MLLM, freezing its vision encoder and fine-tuning only the language component via LoRA [20]. For the referring segmentation model, we adopt EVF-SAM-multitask [67] and further adapt it for saliency detection. Specifically, we fine-tune the mask decoder using structured referring expressions from our SFT data as prompts, supervised by binary cross-entropy against ground-truth masks.

Training Details. The training process uses AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$), with 3,000 SFT steps at a learning rate of $1\text{e}{-6}$, followed by 3,000 CGPO RL steps at a learning rate of $1\text{e}{-4}$. We implement ISR by alternating RL and SFT updates every 5 steps: for every 3 CGPO RL steps, we insert 2 SFT steps. All experiments are conducted on L20 GPUs using PyTorch [38].

4.3. Ablation Study

We validate our core contributions through ablation studies on representative datasets of three tasks: ECSSD (SOD), ILSO (SIS), and CoCA (CoSOD). Our analysis focuses on three aspects: (1) the effectiveness of our CGPO algorithm; (2) the necessity of explicit CoT reasoning; and (3) the benefits of multi-task unified training.

4.3.1. Effectiveness of CGPO

We assess the efficacy of our proposed CGPO by comparing it against standard GRPO and its recent representative variants, namely DrGRPO [36], DAPO [61], and GSPO [70], under identical experimental settings, which involve initializing with SFT followed by RL. As shown in Table 1, CGPO consistently outperforms GRPO and its variants across all three tasks, while avoiding performance degradation in metrics such as E_ξ and F_β, which is observed in GRPO and GSPO compared to the SFT baseline in SOD (ECSSD). This improvement is attributed to two key factors: (1) using per-sample reward–confidence discrepancy as the advantage signal to prioritize informative responses; (2) eliminating group-mean baseline to prevent signal dilution.

More importantly, as shown in Table 2, CGPO achieves superior performance with significantly lower training overhead, as it eliminates multi-sample generation. It requires only 38.7 GB peak GPU memory and 42.2 hours for RL training, in contrast to GRPO and its variants, which consume $\geq$ 65.0 GB memory and $\geq$ 127.0 training hours.

4.3.2. Effectiveness of CoT Reasoning

We evaluate the impact of explicit CoT reasoning using four variants (Table 1): (1) SFT with full CoT supervision; (2) $\text{SFT}_{\text{w/o CoT}}$ with direct instruction-to-answer mapping; (3) $\text{SFT}_{\text{w/o CoT}}$ + CGPO applying CGPO to the no-CoT model; and (4) SFT+CGPO, our full CoT-aware approach.

Results show that explicit CoT reasoning consistently boosts SOD and SIS performance at both SFT and CGPO

Table 1. **Ablation study on key components of our model and training strategy.** We conduct comprehensive ablation experiments on representative datasets from the CoSOD, SOD, and SIS tasks. We use bold to mark the best results.

Method	**SOD Task** (ECSSD [58])				**SIS Task** (ILSO [27])		**CoSOD Task** (CoSOD3k [16])			
	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$AP_{70} \uparrow$	$AP_{50} \uparrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Ablation Study on RL Algorithm										
○ SFT	0.890	0.929	0.906	0.046	0.683	0.711	0.853	0.897	0.861	0.057
○ SFT+GRPO[19]	0.895	0.928	0.889	0.049	0.895	**0.926**	0.899	0.944	0.900	0.035
○ SFT+DrGRPO [36]	0.915	0.948	0.931	0.035	0.802	0.830	0.901	0.941	0.903	0.033
○ SFT+DAPO [61]	0.929	0.961	0.938	0.026	0.716	0.738	0.901	0.942	0.899	0.031
○ SFT+GSPO [70]	0.894	0.923	0.899	0.044	0.755	0.780	0.897	0.937	0.894	0.319
• **SFT+CGPO**	**0.935**	**0.968**	**0.952**	**0.022**	**0.902**	0.922	**0.907**	**0.950**	**0.912**	**0.030**
Ablation Study for CoT Reasoning										
○ SFT	0.890	0.929	0.906	0.046	0.683	0.711	0.853	0.897	0.861	0.057
○ $SFT_{w/o\ CoT}$	0.870	0.908	0.883	0.062	0.525	0.542	0.887	0.933	0.903	0.039
○ $SFT_{w/o\ CoT}$ + CGPO	0.912	0.951	0.930	0.032	0.864	0.902	**0.911**	**0.956**	**0.919**	**0.027**
• **SFT+CGPO**	**0.935**	**0.968**	**0.952**	**0.022**	**0.902**	**0.922**	0.907	0.950	0.912	0.030
Ablation Study for Multi-task Unified Training										
○ Separate Training	0.925	0.958	0.938	0.027	0.898	0.921	0.903	0.947	0.907	0.031
• **Unified Training**	**0.935**	**0.968**	**0.952**	**0.022**	**0.902**	**0.922**	**0.907**	**0.950**	**0.912**	**0.030**

Table 2. **Training overhead comparison on a single H20 GPU: CGPO vs. GRPO and its other variants (DrGRPO, DAPO, GSPO). Traces**: reasoning traces per sample; **Mem**: peak GPU memory; **Step Time**: average training time per step; **Total Time**: wall-clock time for 3k RL steps.

Method	**Traces**	**Mem** (GB)	**Step Time** (s)	**Total Time** (h)
○ GRPO[19]	8	92.4	213.8	178.2
○ GRPO w/o KL	8	72.4	152.5	127.0
○ DrGRPO [36]	8	65.0	163.7	136.4
○ DAPO [61]	8	73.1	195.3	162.7
○ GSPO [70]	8	71.0	162.8	135.7
• **CGPO**	**1**	**38.7**	**50.6**	**42.2**

stages, as these tasks require context analysis and instance disambiguation, benefiting from stepwise, attribute-aware reasoning. CoSOD gains little from CoT since it only needs identifying a shared semantic category, making direct group-to-category mapping more stable for RL optimization. Thus, CoT proves crucial for fine-grained, compositional tasks, but less so for coarse, consensus-based ones.

4.3.3. Effectiveness of Unified Training

We compare separate training (three independently trained, task-specific models) against multi-task unified training (a single shared model trained jointly on all three tasks). As shown in Table 1, unified training achieves on-par or slightly better performance across SOD (ECSSD), SIS (ILSO), and CoSOD (CoSOD3k), despite using only one parameter set. This indicates strong task compatibility in the shared representation space, with no negative interference. Crucially, unified training matches or exceeds the performance of three separate models while offering clear gains in deployment simplicity, memory efficiency, and maintainability.

4.4. Comparison with MLLMs

We compare Saliency-R1 with several advanced MLLMs, including closed-source models such as GPT-4o [21], Gemini-2.5-Pro [10], Claude-Sonnet-4.5 [3], Qwen-VL-Max [4], and Grok-4 [56], as well as open-source models such as Qwen2.4-VL-7B-Instruct [5], Qwen3-VL-32B-Thinking [59], InternVL3-9B [72], LLaVA-OneVision-Qwen2-7B [26], and GLM-4.1V-9B-Thinking [45]. Although these MLLMs possess strong general reasoning capabilities, they show limited performance on visual saliency understanding, particularly in SOD and SIS tasks.

As shown in Table 3, our Saliency-R1 achieves consistent improvements across all tasks. Especially in SIS (ILSO), our model achieves an AP_{50} of 0.922, surpassing GPT-4o (0.558) by over 36% and Qwen3-VL-32B-Thinking (0.655) by 26.7%, which clearly demonstrates its superior instance-level reasoning capability.

These results indicate that Saliency-R1 effectively compensates for the deficiencies of current MLLMs in saliency reasoning. By integrating confidence-guided CoT reasoning with unified training, our model bridges the gap between general multimodal understanding and saliency reasoning.

4.5. Comparison With State-of-the-Art Methods

To further evaluate the effectiveness of Saliency-R1, we compare it with representative state-of-the-art (SOTA) methods in each task domain. For SOD, we choose VST [34], MENet [51], and SelFReformer [64]. For SIS, we benchmark against advanced category-agnostic instance segmentation methods including SCNet [39], S4Net [17], RDPNet [55], and OQTR [40]. Within CoSOD, the comparison involves competitive methods like GCoNet+ [71],

Table 3. **Comparison experiments with SOTA closed-source and opened-source MLLMs on the ECSSD, ILSO, and CoSOD3k datasets.** We use bold and underline to mark the best and second-best excellent results, respectively.

Method	SOD Task (ECSSD [58])				SIS Task (ILSO [27])		CoSOD Task (CoSOD3k [16])			
	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$AP_{70} \uparrow$	$AP_{50} \uparrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Closed-source MLLMs										
○ Gemini-2.5-Pro [10]	0.898	0.929	0.894	0.047	0.632	0.673	0.868	0.911	0.858	0.045
○ Claude-Sonet-4-5 [3]	0.868	0.893	0.855	0.072	0.644	0.678	0.904	0.947	0.907	**0.030**
○ GPT-4o [21]	0.707	0.727	0.680	0.203	0.529	0.558	0.904	0.948	0.909	0.031
○ Qwen-VL-Max [4]	0.823	0.846	0.808	0.108	0.608	0.635	0.878	0.910	0.860	0.038
○ Grok-4 [56]	0.790	0.808	0.753	0.116	0.561	0.597	0.900	0.941	0.900	0.032
Opened-source MLLMs										
○ Qwen2.5-VL-7B-Instruct [5]	0.518	0.531	0.510	0.372	0.558	0.590	0.808	0.850	0.785	0.086
○ Qwen3-VL-32B-Thinking [59]	0.830	0.853	0.814	0.099	0.625	0.655	0.776	0.770	0.676	0.075
○ InternVL3-9B [72]	0.576	0.580	0.559	0.325	0.501	0.545	0.866	0.909	0.856	0.048
○ LLaVA-OneVision-7B [26]	0.796	0.816	0.792	0.139	0.456	0.482	0.875	0.920	0.866	0.045
○ GLM-4.1V-9B-Thinking [45]	0.777	0.792	0.762	0.155	0.361	0.414	0.863	0.890	0.833	0.045
• **Saliency-R1(Ours)**	**0.935**	**0.968**	**0.952**	**0.022**	**0.902**	**0.922**	**0.907**	**0.950**	**0.912**	**0.030**

Table 4. **Quantitative comparisons with other SOTA methods in conventional SOD, SIS, and CoSOD tasks.** We conduct the comparison on nine benchmark datasets. We use bold and underline to mark the best and second-best excellent results, respectively.

Salient Object Detection (SOD)

Methods	Pub.	Training Images	DUT-O [60]				ECSSD [58]				PASCAL-S [31]			
			$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
○ VST [34]	ICCV2021	10553	0.850	0.888	0.800	0.058	0.932	0.964	0.944	0.033	0.873	0.920	0.871	0.060
○ MENet [51]	CVPR2023	10553	0.850	0.879	0.792	0.045	0.928	0.956	0.939	0.031	0.871	0.914	0.872	0.055
○ SelfReformer [64]	TMM2023	10553	0.861	0.894	0.807	**0.043**	**0.936**	0.965	0.948	0.027	0.881	0.926	0.883	0.052
• **Saliency-R1(Ours)**	-	**4440**	**0.864**	**0.907**	**0.833**	0.049	0.935	**0.968**	**0.952**	**0.022**	**0.887**	**0.935**	**0.891**	**0.042**

Salient Instance Segmentation (SIS)

Methods	Pub.	Training Images	ILSO [27]		SIS10K [40]		SOC [14]	
			$AP_{70} \uparrow$	$AP_{50} \uparrow$	$AP_{70} \uparrow$	$AP_{50} \uparrow$	$AP_{70} \uparrow$	$AP_{50} \uparrow$
○ S4Net [17]	CVPR2019	**500**	0.636	0.867	0.670	0.833	0.420	0.646
○ SCNet [39]	Neurocomputing2020	14207	0.674	0.846	0.692	0.692	0.414	0.609
○ RDPNet [55]	TIP2021	2900	0.885	0.734	0.694	0.820	0.482	0.606
○ OQTR [40]	TMM2022	9330	0.799	0.904	0.817	**0.881**	**0.725**	**0.895**
• **Saliency-R1(Ours)**	-	4267	**0.902**	**0.922**	**0.843**	0.870	0.714	0.745

Co-Salient Object Detection (CoSOD)

Methods	Pub.	Training Images	CoCA [68]				CoSal2015 [65]				CoSOD3k [16]			
			$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$S_m \uparrow$	$E_\xi \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
○ GCoNet+ [71]	TPAMI2023	208250	0.738	0.814	0.637	0.081	0.881	0.926	0.891	0.055	0.843	0.901	0.834	0.061
○ CONDA [29]	ECCV2024	208250	0.763	0.839	0.685	0.089	0.900	0.938	0.908	0.045	0.862	0.911	0.853	0.056
○ DMT+ [30]	TPAMI2025	208250	0.779	0.834	0.689	0.076	0.899	0.935	0.903	0.047	0.866	0.912	0.854	0.054
○ VCP [47]	CVPR2025	208250	0.819	0.871	0.752	0.054	0.927	0.962	0.941	0.030	0.895	0.938	0.893	0.043
• **Saliency-R1(Ours)**	-	**24496**	**0.899**	**0.946**	**0.871**	**0.023**	**0.928**	**0.965**	**0.942**	**0.024**	**0.907**	**0.950**	**0.912**	**0.030**

CONDA [29], DMT+ [28], and VCP [47].

As shown in Table 4, our model consistently outperforms most compared methods, even though it is trained with far less data than task-specific models. In SOD, it outperforms strong transformer- and CNN-based methods. For SIS, the model shows clear improvements, producing more coherent instance-level predictions. Its advantage is most pronounced in CoSOD, on the CoCA dataset, Saliency-R1 achieves an S_m of 0.899, surpassing the previous best method VCP (0.819) by about 8%, demonstrating outstanding capability in capturing cross-image consistency.

5. Conclusion

This paper proposes Saliency-R1 to incentivize unified saliency reasoning capability in MLLM through confidence-guided reinforcement learning. This design encourages the model to align its multimodal reasoning with spatial saliency cues, enabling coherent and interpretable visual reasoning. Experimental results on SOD, SIS, and CoSOD benchmarks demonstrate that Saliency-R1 effectively bridges multimodal reasoning and pixel-level visual saliency perception, outperforming both existing MLLMs and task-specific models.

Saliency-R1: Incentivizing Unified Saliency Reasoning Capability in MLLM with Confidence-Guided Reinforcement Learning

Supplementary Material

This supplementary material provides essential implementation details and analysis omitted from the main paper due to space limits, including: (1) full task instruction templates; (2) complete formulation of the reward function, especially the Instance-Aligned S-measure (IASM) for SIS; (3) the "output-to-reasoning" prompts used to generate SFT data; (4) parameter-scale comparison with other MLLMs; (5) qualitative comparisons with other MLLMs and specialize SOTA methods.

6. Task Instruction Design

Figure 5 illustrates the task instruction templates employed during the inference and confidence-guided RL training of Saliency-R1. Each instruction explicitly requires the MLLM to generate step-by-step CoT reasoning, adhering to strict formatting constraints: the reasoning trace must be enclosed in `<think> ... </think>`, and the final answer in `<answer> ... </answer>`. Crucially, the instructions are task-aware and prescribe distinct output semantics aligned with task granularity.

For SOD, the final answer may contain multiple `<rg>` blocks, each describing a salient region (*e.g.*, `<rg>a red cup</rg>`). When the description is a pure category name, it must be prefixed with `[semantic]` (*e.g.*, `<rg>[semantic] dog</rg>`).

For SIS, each salient instance must be expressed using an attribute-rich, disambiguating expression within an `<ins>` tag (*e.g.*, `<ins>a red cup on the left</ins>`).

For CoSOD, only a single `<rg>[semantic] s</rg>` expression is allowed, where s denotes the shared semantic category across the image group (*e.g.*, `<rg>[semantic] airplane</rg>`), ensuring cross-image consensus.

7. Reward Definition Details

To align the MLLM's textual reasoning with both task correctness and downstream parseability, our total reward is defined as a weighted sum:

$$r = \lambda \cdot r_{\text{corr}} + (1 - \lambda) \cdot r_{\text{fmt}}, \quad \lambda = 0.5, \tag{17}$$

where r_{corr} evaluates segmentation quality, and r_{fmt} enforces strict adherence to the structured CoT interface. The weighting $\lambda = 0.5$ is selected based on empirical validation to balance optimization of reasoning fidelity and output safety. Below, we detail each component.

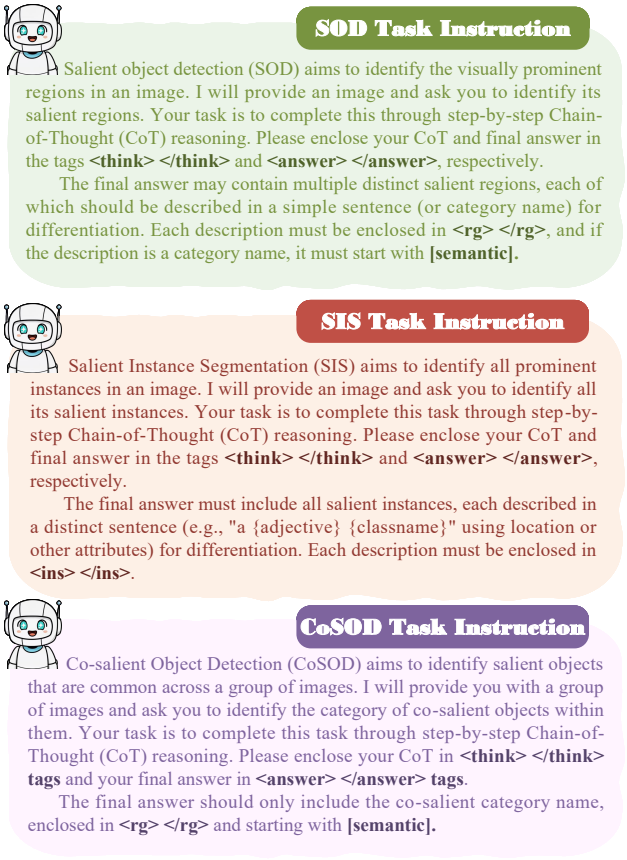

Figure 5. **Task instruction design for the three saliency tasks.**

7.1. Task-Adaptive Correctness Reward

The definition of r_{corr} is tailored to each saliency task to accommodate distinct input configurations and evaluation granularity.

SOD. Given parsed region expressions $\mathcal{E}_i^{\text{SOD}}$ ($i = 1, \ldots, N$), the predicted binary mask $\boldsymbol{M} \in \{0,1\}^{H \times W}$ is obtained by pixel-wise logical OR:

$$\boldsymbol{M} = \bigvee_{i=1}^{N} \mathcal{S}\big(x^{\text{SOD}}, \mathcal{E}_i^{\text{SOD}}\big), \tag{18}$$

where $\mathcal{S}(\cdot)$ denotes the referring segmenter (EVF-SAM). Let $\boldsymbol{G} \in \{0,1\}^{H \times W}$ be the ground-truth mask. We adopt the Structure-measure $\mathcal{S}_m$ [13] as the correctness reward:

$$r_{\text{corr}} = \mathcal{S}_m(\boldsymbol{M}, \boldsymbol{G}). \tag{19}$$

CoSOD. In CoSOD, masks are generated for each image x_k^{CoSOD} with a shared expression $\mathcal{E}^{\text{CoSOD}}$:

$$\boldsymbol{M}_k = \mathcal{S}\big(x_k^{\text{CoSOD}}, \mathcal{E}^{\text{CoSOD}}\big), \quad k = 1, \ldots, K. \tag{20}$$

The ground-truth mask for the k-th image is $\boldsymbol{G}_k$, and the reward is the average S-measure across the group:

$$r_{\text{corr}} = \frac{1}{K} \sum_{k=1}^{K} \mathcal{S}_m(\boldsymbol{M}_k, \boldsymbol{G}_k). \tag{21}$$

SIS. Here, the standard S-measure is insufficient due to the need for instance correspondence. We introduce the Instance-Aligned S-measure (IASM):

Given M predicted instance masks $\{\boldsymbol{M}_j\}_{j=1}^{M}$ and I ground-truth masks $\{\boldsymbol{G}_i\}_{i=1}^{I}$, we construct the IoU matrix $\mathbf{U} \in \mathbb{R}^{I \times M}$ with entries

$$U_{ij} = \frac{\|\boldsymbol{G}_i \cap \boldsymbol{M}_j\|}{\|\boldsymbol{G}_i \cup \boldsymbol{M}_j\|}. \tag{22}$$

The Hungarian algorithm identifies the optimal one-to-one matching $\mathcal{M}$ to maximize total IoU, retaining pairs with $U_{ij} \geq \tau$ ($\tau = 0.1$). Unmatched instances are paired with zero masks, yielding aligned pairs $\{(\boldsymbol{G}'_\ell, \boldsymbol{M}'_\ell)\}_{\ell=1}^{L}$ (where $L = I + M - |\mathcal{M}|$). The correctness reward is defined as:

$$r_{\text{corr}} = \frac{1}{L} \sum_{\ell=1}^{L} \mathcal{S}_m(\boldsymbol{M}'_\ell, \boldsymbol{G}'_\ell). \tag{23}$$

This formulation penalizes both over- and under-segmentation, along with instance misassignment, which is crucial for instance-level reasoning.

7.2. Structured Format Reward

Our pipeline requires precise parsing of `<think>` and `<answer>` blocks, alongside task-specific tags (`<rg>`, `<ins>`, `[semantic]`). Any malformed response disrupts the system, thus r_{fmt} acts as a hard interface contract:

$$r_{\text{fmt}} = r_{\text{struct}} + r_{\text{tag}}, \tag{24}$$

where $r_{\text{struct}} = 0.5$ if and only if the response contains exactly one well-formed `<think>` block and exactly one well-formed `<answer>` block (without nesting, truncation, or duplication); otherwise $r_{\text{struct}} = 0$. For r_{tag}, it equals 0.5 if and only if the content inside `<answer>` adheres strictly to Eq. (2) of the main paper:

- **SOD**: $\mathcal{E} = \{\texttt{<rg>}d_i\texttt{</rg>}\}_{i=1}^{N}$, with $N \geq 1$ (at least one `<rg>` tag, each properly closed);
- **CoSOD**: $\mathcal{E} =$ `<rg>[semantic] s </rg>` (exactly one `<rg>` tag, starting with `[semantic]`);
- **SIS**: $\mathcal{E} = \{\texttt{<ins>}\phi_j\texttt{</ins>}\}_{j=1}^{M}$, with $M \geq 1$ (at least one `<ins>` tag, each properly closed).

Any deviation, such as missing tags, extraneous or missing closing brackets, incorrect tag types (*e.g.*, `<ins>` in SOD), or malformed `[semantic]` annotations, results in $r_{\text{tag}} = 0$.

This design ensures that only fully compliant outputs receive $r_{\text{fmt}} = 1.0$, thus providing a clear and strong reinforcement learning signal for interface safety.

8. Visualization of Model Predictions

Figure 6 illustrates representative outputs of Saliency-R1 across the three saliency tasks, *i.e.*, SOD, SIS, and CoSOD. This demonstrates how structured CoT reasoning facilitates precise segmentation.

In the SOD task, the model produces two region-level referring expressions, *i.e.*, "a military aircraft" and "a person", which are enclosed in `<rg>` tags. These expressions are processed independently by EVF-SAM, and their masks are unioned to form the final salient region map.

In the SIS task, the model differentiates two overlapping instances using attribute-rich descriptors, *i.e.*, "a green apple on the left" and "an orange on the right", within `<ins>` tags. EVF-SAM generates separate binary masks for each, allowing for accurate instance-level disambiguation without merging.

In the CoSOD task, the model analyzes four images, identifies "butterfly" as the only object category consistently present, and outputs a single semantic-tagged expression: `<rg>[semantic] butterfly</rg>`. This unified cue is applied across all images, yielding consistent co-salient masks.

Critically, all responses strictly adhere to the structured format (Eq. (1)–(2)), ensuring parseability. The resulting masks exhibit high fidelity: complete object coverage, sharp boundaries, and robustness to occlusion or clutter, thereby validating the effectiveness of our Saliency-R1 interface.

9. Prompts for SFT Data Generation via Output-to-Reasoning

We adopt the "output-to-reasoning" strategy, as introduced in Sec. 3.2.1, to generate high-quality CoT supervision for SFT. Instead of prompting the model to generate reasoning solely from instructions, we condition a powerful closed-source MLLM (Gemini-2.5-Pro [10]) on both the raw visual input *and* its corresponding ground-truth segmentation mask(s) as a visual cue. These ground-truth masks serve as precise saliency priors, allowing the model to anchor its stepwise reasoning in actual pixel-level evidence, thereby preserving naturalness and linguistic fluency.

As illustrated in Figure 7-9, each task employs a tailored prompt template that guides the model through a fixed three-step CoT structure (scene/object identification → task-specific analysis → conclusion), while ensuring strict adher-

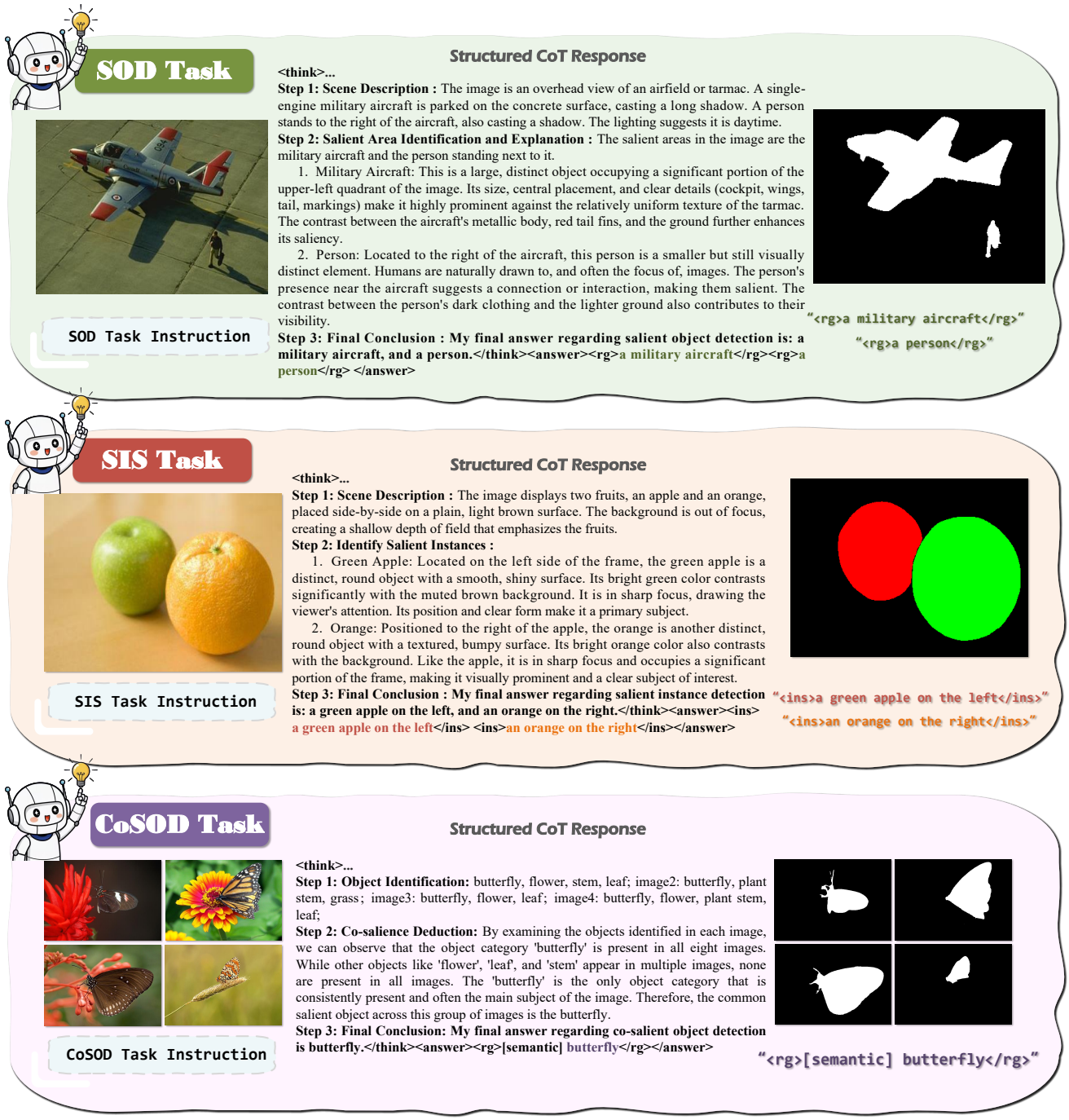

Figure 6. **Qualitative examples of Saliency-R1's reasoning and segmentation on SOD, SIS, and CoSOD.** Each shows multimodal inputs, structured CoT responses, parsed referring expressions, and final segmentation masks.

ence to the formatting outlined in Eq. (1)–(2). In the SOD task, where multiple salient regions may coexist, the model is instructed to enumerate all highlighted areas with concise and distinguishing descriptions within `<rg>` tags. For the SIS task, instance-wise mask cues prompt the model to produce attribute-rich, disambiguating expressions (*e.g.*, "a green apple on the left") within `<ins>` tags. For the CoSOD task, the model leverages cross-image mask consistency to identify the shared semantic category, outputting this as `<rg>[semantic] s</rg>`.

This conditioning ensures faithful alignment between reasoning and segmentation, reduces hallucination, and gen-

Prompt for SFT dataset construction **SOD Task**

You are a professional with a deep understanding of the mechanics of human visual attention. Co-salient object detection aims to identify salient objects that are common across a group of images. I will provide you with **a group of images** (they are named image1, image2, and so on), and **the correct category of co-salient object in the images (ground truth)**. It is known that the co-salient objects in this group of images is {}, which means that the group's co-salient objects you found are all present in the non-noise images (normal image). Your task is to provide a specific reason explaining the category of co-salient objects in this group of images with **step-by-step Chain-of-Thought (CoT) reasoning**.

In the CoT, **the first step is called 'Object Identification'**, which involves to identify objects in each image, (Notice that you don't care about object properties, just class), **the second step is called 'Co-salience Deduction'**, where you deduce and explain the co-salient object in this group of images. **The third step is called 'Final Conclusion'**, where you arrive at your final answer. Ensuring that your identified co-salient objects are consistent with the gt I provided. (However, please do not mention anything related to "ground truth" in any of your responses.) Please adhere strictly to the following step titles as specified: **"Step 1: Object Identification," "Step 2: Co-salience Deduction," and "Step 3: Final Conclusion."** Please note that there may be some noisy images that do not contain a common salient object, but you need to Perform comprehensive object recognition for EACH image in this group (include noise images and normal images) and provide MULTIPLE object categories per image include noise image.

When you come to your 'Final conclusion', please ensure that you begin with "My final answer regarding co-salient object detection is: **<rg>[semantic]"** and end with**"</rg>"**, start your task. Ensure that your answers are formatted uniformly each time. (eg: **"My final answer regarding co-salient object detection is: <rg>[semantic]airplane</rg>"**). Additionally, do not apply any font changes (such as bolding) to the text in the 'Final Conclusion'.

For synonyms like dog and puppy to appear in the object recognition of non-noise images, you can only use one of the words to describe the object. So please use one word to describe the group's co-salient objects for the object recognition part of all non-noise images.

Please start with "The user is asking for the co-salient object in the provided images. Co-salient object means the object that is salient across all images. I need to look at all the images and identify the object that is present and noticeable in most or all of them." start your task.

Figure 7. **Prompt design for SFT dataset on the salient object detection (SOD) tasks.**

Prompt for SFT dataset construction **SIS Task**

You are a professional with a deep understanding of the mechanics of human visual attention. The task of salient instance detection is to identify all salient instances within an image. We will provide you with **an image** and **a mask image that highlights different salient instances in the original image using distinct colored regions**. Your task is to identify all salient instances in the image by employing **step-by-step Chain-of-Thought (CoT) reasoning**. Please ensure that your responses are consistent with the salient instances highlighted in the provided mask image. (However, please do not mention anything related to "mask image" or "masked areas" in any of your responses.)

In the CoT, **the first step is 'Scene Description'**, where you provide a brief description of the scene presented in the image. **The second step is 'Identify Salient Instances'**, which involves identifying the salient instances within the image and provide justifications for their saliency. This can be accomplished by identifying all instances within the masked areas of the provided image and describing the characteristics of these instances, as well as deducing and explaining why these instances are salient. Please note that when you explain the saliency, you should also highlight the relationship of this salient instance to the overall context, as well as its contrast relationship with other instances (whether salient or not). **The third step is 'Final Conclusion'**, where you generate a final response, summarizing each salient instance in a single sentence, please ensure that you begin with 'Final answer: '. Confirm that the salient instances correspond to the masked areas. Please adhere strictly to the following step titles as specified: **"Step 1: Scene Description", "Step 2: Identify Salient Instances", and "Step 3: Final Conclusion"**.

When you reach your conclusion, please follow these rules:

1.For each salient instance, use text to describe each one, ensuring that each description is distinct and clearly distinguishes each instance in the image (e.g., "**a {adjective} {classname}**" using location or other attributes).

2.Each description should begin with "**\n<ins>** ", such as "**\n<ins>** a black cat on the top". Please keep the string "**\n<ins>** " intact as a whole, without splitting it or adding spaces within it.

3.If the response after "**Final answer:** " includes multiple categories, each category should begin with "**\n<ins>**", such as " **Final answer: \n<ins> a black cat \n<ins> a red dog** ". Additionally, do not include any punctuation marks, such as periods or commas, in the response following 'Final answer:'

Please begin with 'The user wants me to identify the salient instances in the provided image. I need to analyze the input image, try to identify the salient instances within the image and describe it in a single sentence.' directly start your task.

Figure 8. **Prompt design for SFT dataset on the salient instance segment (SIS) tasks.**

erates high-granularity referring expressions, providing a robust foundation for subsequent CGPO refinement.

10. Parameter Scale Comparison

We also compare the parameter scales of our model with other MLLMs. As summarized in Table 5, our Saliency-R1 adopts Qwen2.5-VL-7B-Instruct as the MLLM backbone, with only about 7B dense parameters, substantially smaller than most closed-source counterparts (*e.g.*, GPT-4o [21]: ∼200B, Gemini-2.5-Pro [10]: ∼175B MoE) and even smaller than many open-source alternatives (*e.g.*, Qwen3-VL-32B-Thinking [59]: 32B). Despite this compact scale, Saliency-R1 significantly outperforms all listed models across SOD, SIS, and CoSOD (see Table 3 in the main text), indicating that its superiority stems not from model size, but from the synergistic design of our **confidence-guided RL training (CGPO)** in aligning the MLLM's reasoning with visual saliency priors, enabling more accurate,

Prompt for SFT dataset construction CoSOD Task

You are a professional with a deep understanding of the mechanics of human visual attention. Salient object detection (SOD) aims to identify the visually prominent (i.e., attention-grabbing) regions within an image. I will provide you with **an image** and **its ground truth mask map**, which highlights the salient areas of the image. Your task is to identify the salient areas in the image through **step-by-step chain-of-thought (CoT) reasoning**, ensuring that your identification of the salient areas is consistent with the salient areas highlighted in the provided ground truth mask.

In the CoT, **the first step is 'Scene Description',** which involves providing a brief description of the scene depicted in the image. **The second step is 'Salient Area Identification and Explanation'**, which involves identifying which areas are salient and explaining why; this can be accomplished by identifying all objects within the masked areas of the image I provide and describing their characteristics, as well as deducing and explaining why these objects are salient in the image. Please note that when you explain the salience, you should also highlight the contrast relationship of this object to the overall context, as well as its relationship with other objects. Additionally, please explain why some objects are not considered salient (why background objects are background) and analyze the relationship of non-salient objects with the overall context and other objects. **The third step is 'Final Conclusion'**, where you generate a final response that describes the salient areas using concise sentences, please ensure that you begin with "Final answer: ". Ensuring that the identified salient objects align with the mask image provided. (However, please do not mention anything related to "ground truth mask map" in any of your responses.) Please adhere strictly to the following step titles as specified: **"Step 1: Scene Description", "Step 2: Salient Area Identification and Explanation", and "Step 3: Final Conclusion"**. Note that the masked areas may consist of sub-regions belonging to different semantic categories. You need to locate each sub-region of the masks in the image using either a single word or a sentence, ensuring that each description is distinct.

When you reach your conclusion, please follow these rules:

1. Use concise text to describe each one, ensuring that each description is distinct and clearly distinguishes each object in the image (e.g., "**a {adjective} {classname}**" using location or other attributes). Keep responses brief, including category names.

2.For each sub-region, provide a description in one sentence or a single word (preferably a word). Each description should begins with "**\n<rg> **", such as "**Final answer: \n<rg> a black cat**". If you use a single word, please precede it with the " **[semantic]** " tag (e.g., "**Final answer: \n<rg> [semantic] person\n<rg> [semantic] dog**"), aiming to minimize the number of sub-region divisions. Please keep the string "**\n<rg> **" intact as a whole, without splitting it or adding spaces within it. Additionally, do not include any punctuation marks, such as periods or commas, in the response following 'Final answer: '.

3. If a sub-mask area or the entire mask area contains four or more objects of the same category, use terms like "group" or "several" instead of specifying the exact quantity.

4.Describe each sub-region in plain text.

Please start with 'The user wants me to identify the salient object in the provided image. I need to analyze the input image, try to identify the salient objects within the image and describe each one in a single sentence.' start your task.

Figure 9. **Prompt design for SFT dataset on the Co-salient object detection (CoSOD) tasks.**

Table 5. **Total parameter count of the MLLM backbone (including vision encoder) across models.** "Dense" denotes models where all parameters are activated in every forward pass; "MoE" (Mixture of Experts) denotes models where only a subset of parameters (experts) are activated per token. "Est." indicates community-derived parameter estimates for closed-source models (official values undisclosed). Open-source values follow official reports. All models use the same external EVF-SAM segmenter (not counted).

Model	Type	Params (B)
Closed-source Models		
○ Gemini-2.5-Pro [10]	MoE	175 (Est.)
○ Claude-Sonnet-4.5 [3]	Dense	175+ (Est.)
○ GPT-4o [21]	Dense	200 (Est.)
○ Qwen-VL-Max [4]	MoE	200+ (Est.)
○ Grok-4 [56]	MoE	300+ (Est.)
Open-source Models		
○ Qwen2.5-VL-7B-Instruct [5]	Dense	7
○ Qwen3-VL-32B-Thinking [59]	Dense	32
○ InternVL3-9B [72]	Dense	9
○ LLaVA-OneVision-Qwen2-7B [26]	Dense	7
○ GLM-4.1V-9B-Thinking [45]	Dense	9
• **Saliency-R1 (Ours)**	Dense	**7**

robust, and task-coherent CoT generation.

11. Qualitative Comparison

We conduct comprehensive qualitative analyses to demonstrate that Saliency-R1's reasoning quality drives superior mask fidelity across heterogeneous saliency tasks. As shown in Figure 10-15, our method matches or exceeds the segmentation performance of both closed/open-source MLLMs and specialized saliency models across diverse and challenging scenarios, *e.g.*, instance occlusion, scale variation, reflection interference, and cross-image consensus reasoning. These results highlight the effectiveness of our Saliency-R1 framework.

11.1. Comparison with MLLMs

Compared to closed- and open-source MLLMs (*e.g.*, GPT-4o [21], Gemini-2.5-Pro [10], Qwen3-VL-32B-Thinking [59], LLaVA-OneVision-7B [26], and so on), Saliency-R1 exhibits markedly superior visual grounding precision and segmentation coherence, as shown in Figure 10 (SOD), Figure 11 (SIS) and Figure 12 (CoSOD).

In SOD, as shown in Figure 10, common failure modes include false-positive activation on background regions, *e.g.*, the substrate beneath the chameleon (Column 1), the branches behind the bird (Column 2), the flowers surrounding the black dog (Column 3), and the water surface near the

deer's legs (Column 4). Our method preserves fine structure and enforces foreground–background contrast via explicit CoT reasoning.

In SIS, as shown in Figure 11, common failure modes of MLLMs include under-prediction of instances (*e.g.*, the buckets in Column 1, the car in Column 3, and the person in Column 4), over-prediction of non-instances (*e.g.*, spurious masks on the ground in Columns 1 and 3, and on the person in Column 2). Our method, guided by explicit CoT reasoning that generates distinct `<ins>`-tagged expressions (*e.g.*, "a person standing in front of the bus"), enables EVF-SAM to robustly isolate each instance.

In the CoSOD task, as shown in Figure 12, MLLMs often hallucinate or overlook small co-salient instances under scale variation (*e.g.*, distant boats in Column 11), while our model reliably identifies them. Additionally, in cluttered scenes (*e.g.*, penguin and band in Column 3), Gemini-2.5-Pro segments background distractors, while our CoT-induced consensus reasoning (Step 2: Co-salience Deduction) suppresses non-common elements, yielding masks that are strictly aligned with ground truth.

These results highlight a key advantage: our CoT is not just descriptive, but functionally coupled to segmentation, each referring expression acts as a discriminative prompt for the segmenter.

11.2. Comparison with Task-Specific Methods

We further compare our method against specialized SOTA methods, *i.e.*, VST [34], MENet [51], SelFReformer [64] (SOD); S4Net [17], RDPNet [55], OQTR [40] (SIS); GCoNet+ [71], CONDA [29], DMT+ [28], and VCP [47] (CoSOD). As shown in Figure 15–13, Saliency-R1 matches or exceeds their performance, even though we use far less task-specific training data.

In SOD, as shown in Figure 15, CNN/Transformer-based models (*e.g.*, MENet, SelfReformer) suffer from over-prediction of background regions (*e.g.*, the maintenance worker and scaffolding are not separated in Column 2, and the sign beneath the bench is incorrectly included in the bench's mask in Column 4) and under-prediction of object parts (*e.g.*, the legs of the bench are missed in Column 4). Our approach, guided by contrast-aware CoT, recovers sharp edges and complete object extent.

In SIS, as shown in Figure 14, specialized methods exhibit diverse failure modes: under-prediction of small instances (*e.g.*, one water cup is missed in Column 1), failure to disambiguate overlapping instances (*e.g.*, adjacent golf clubs are merged by OQTR in Column 2), confusion or omission of multiple objects (*e.g.*, the three golden bars are incorrectly segmented as a single mask by OQTR in Column 3), and over-segmentation of a single instance into multiple parts (*e.g.*, the chair is split into separate masks by OQTR in Column 4). Our method, guided by explicit CoT reasoning

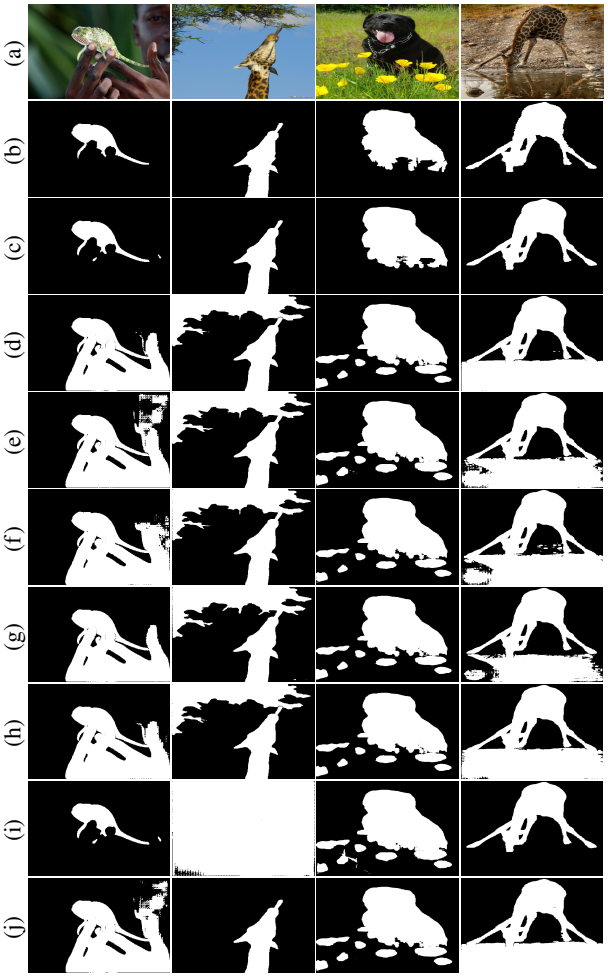

Figure 10. **Qualitative comparison with other MLLMs on SOD.** **(a)** denotes the input image, **(b)** the ground-truth mask, **(c)** our Saliency-R1, followed by closed- and open-source MLLMs: **(d)** Claude-Sonnet-4.5, **(e)** Gemini-2.5-Pro, **(f)** GPT-4o, **(g)** Grok-4, **(h)** Qwen3-VL-32B-Thinking, **(i)** Qwen2.5-VL-7B-Instruct, and **(j)** LLaVA-OneVision-Qwen2-7B.

that generates distinct `<ins>`-tagged descriptions (*e.g.*, "a golf iron on the left, slightly tilted"), enables EVF-SAM to robustly assign precise, instance-level masks.

In CoSOD, as shown in Figure 13, prior methods often segment irrelevant co-occurring objects (*e.g.*, the person beside the backpack in Column 1, or background items around the camera in Column 2). In contrast, our CoT's explicit consensus deduction enables EVF-SAM to suppress distractors and produce cleaner, more selective masks, evident in the precise segmentation of cameras and yellow ducks.

Importantly, all improvements stem from the same unified pipeline, no per-task architecture changes, demonstrating the transferability of CoT-guided reasoning.

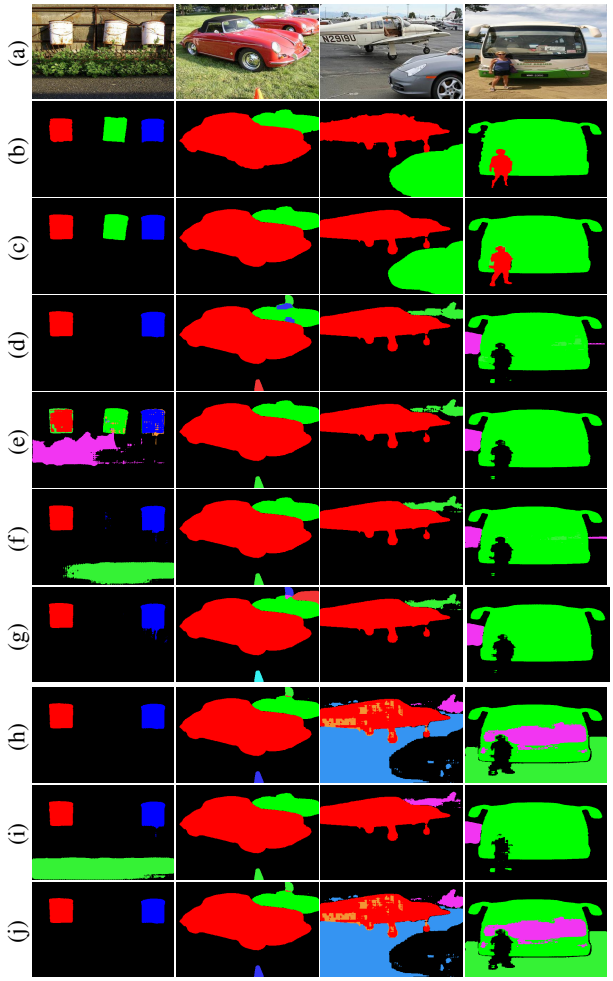

Figure 11. **Qualitative comparison with other MLLMs on SIS.** **(a)** denotes the input image, **(b)** the ground-truth instance masks, **(c)** our Saliency-R1, followed by: **(d)** Claude-Sonnet-4.5, **(e)** Gemini-2.5-Pro, **(f)** GPT-4o, **(g)** Grok-4, **(h)** Qwen3-VL-32B-Thinking, **(i)** Qwen2.5-VL-7B-Instruct, and **(j)** LLaVA-OneVision-Qwen2-7B.

References

[1] Huda Diab Abdulgalil and Otman A Basir. Next-generation image captioning: A survey of methodologies and emerging challenges from transformers to multimodal large language models. *Natural Language Processing Journal*, page 100159, 2025. 1

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 6

[3] Anthropic. System card: Claude sonnet 4.5. *Anthropic News*, 2025. 7, 8, 5

[4] Jinze Bai, Shuai Bai, Shijie Yang, et al. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 7, 8, 5

[5] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, J. Lin, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 7, 8, 5

[6] Jierun Chen, Fangyun Wei, Jinjing Zhao, Sizhe Song, Bohuai Wu, Zhuoxuan Peng, S-H Gary Chan, and Hongyang Zhang. Revisiting referring expression comprehension evaluation in the era of large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 513–524, 2025. 1

[7] Xiangyu Chen, Zhi Wang, Xiang Li, Chen Change Loy, and Dahua Lin. Rank normalization for stable and accelerated training of deep neural networks. In *International Conference on Learning Representations*, 2022. 5

[8] Yi-Chia Chen, Wei-Hua Li, Cheng Sun, Yu-Chiang Frank Wang, and Chu-Song Chen. Sam4mllm: Enhance multi-modal large language model for referring expression segmentation. In *European Conference on Computer Vision*, pages 323–340. Springer, 2024. 2, 3

[9] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. In *ICCV*, pages 1529–1536, 2013. 6

[10] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 4, 6, 7, 8, 2, 5

[11] Yasser Dahou, Ngoc Dung Huynh, Phuc H Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. Vision-language models can't see the obvious. *arXiv preprint arXiv:2507.04741*, 2025. 1

[12] A Philip Dawid. The well-calibrated bayesian. *Journal of the American statistical Association*, 77(379):605–610, 1982. 2, 5

[13] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, pages 4548–4557, 2017. 6, 1

[14] D. P. Fan, M. M. Cheng, J. J. Liu, S. H. Gao, Q. Hou, and A. Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *Proceedings of the European conference on computer vision (ECCV)*, pages 186–202, 2018. 6, 8

[15] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018. 6

[16] D. P. Fan, T. Li, Z. Lin, G. P. Ji, D. Zhang, M. M. Cheng, J. Shen, et al. Re-thinking co-salient object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44 (8):4339–4354, 2021. 2, 6, 7, 8

[17] R. Fan, M. M. Cheng, Q. Hou, T. J. Mu, J. Wang, and S. M. Hu. S4net: Single stage salient-instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6103–6112, 2019. 2, 7, 8, 6

[18] Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings*

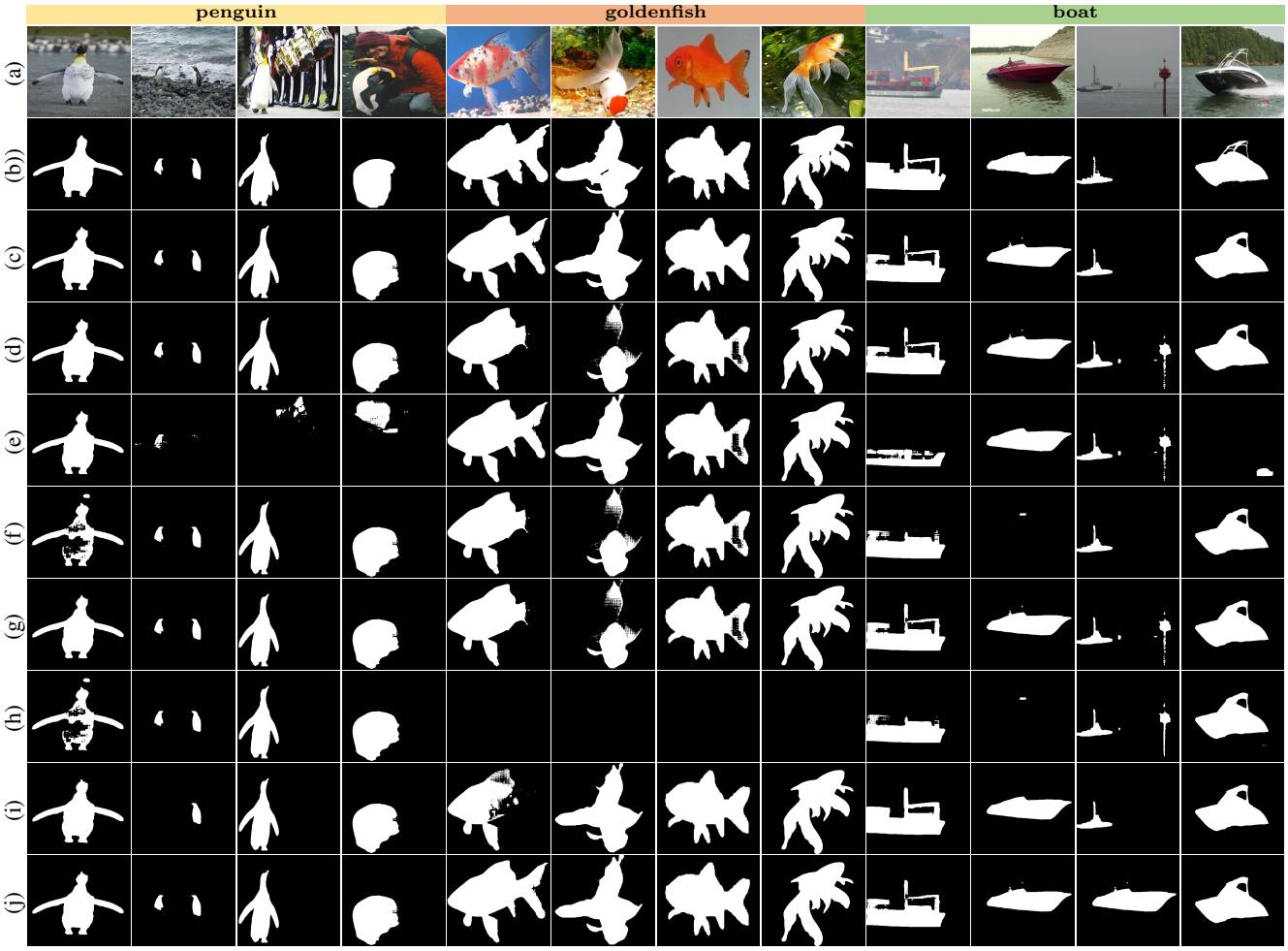

Figure 12. **Qualitative comparison with other MLLMs on CoSOD. (a)** denotes the input image group, **(b)** the ground-truth co-salient masks, **(c)** our Saliency-R1, followed by: **(d)** Claude-Sonnet-4.5, **(e)** Gemini-2.5-Pro, **(f)** GPT-4o, **(g)** Grok-4, **(h)** Qwen3-VL-32B-Thinking, **(i)** Qwen2.5-VL-7B-Instruct, and **(j)** LLaVA-OneVision-Qwen2-7B.

of the Computer Vision and Pattern Recognition Conference, pages 19597–19607, 2025. 1

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 2, 7

[20] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[21] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 1, 7, 8, 4, 5

[22] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306, 2009. 5

[23] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017. 5

[24] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, R. Girshick, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 2, 3

[25] X. Lai, Z. Tian, Y. Chen, Y. Li, Y. Yuan, S. Liu, and J. Jia. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589, 2024. 2

[26] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, others, and C. Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1, 7, 8, 5

[27] G. Li, Y. Xie, L. Lin, and Y. Yu. Instance-level salient object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2386–2395, 2017. 2, 6, 7, 8

[28] Long Li, Junwei Han, Ni Zhang, Nian Liu, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Fahad Shah-

Figure 13. **Qualitative comparison with specialized SOTA methods on CoSOD.** Here, **(a)** denotes the input image group, **(b)** the ground-truth co-salient masks, **(c)** our Saliency-R1, followed by: **(d)** VCP, **(e)** DMT+, **(f)** CONDA, and **(g)** GCoNet+.

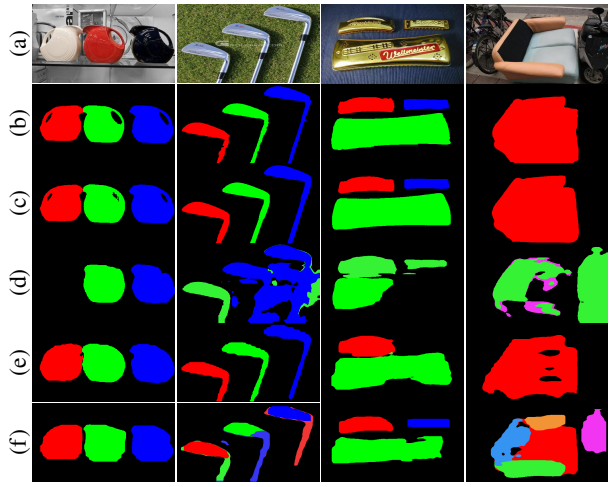

Figure 14. **Qualitative comparison with specialized SOTA methods on SIS.** **(a)** denotes the input image, **(b)** the ground-truth instance masks, **(c)** our Saliency-R1, followed by: **(d)** OQTR, **(e)** RDPNet, and **(f)** S4Net.

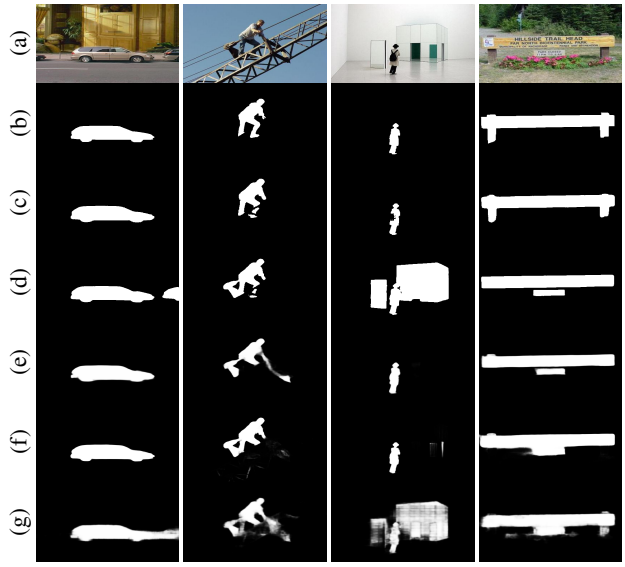

Figure 15. **Qualitative comparison with specialized SOTA methods on SOD.** **(a)** denotes the input image, **(b)** the ground-truth mask, **(c)** our Saliency-R1, followed by: **(d)** FOCUS, **(e)** DMT+, **(f)** SelfReformer, **(g)** MENet, and **(h)** VST.

baz Khan. Discriminative co-saliency and background mining transformer for co-salient object detection. In *CVPR*, pages 7247–7256, 2023. 8, 6

[29] Long Li, Nian Liu, Dingwen Zhang, Zhongyu Li, Salman Khan, Rao Anwer, Hisham Cholakkal, Junwei Han, and Fahad Shahbaz Khan. Conda: Condensed deep association learning for co-salient object detection. In *European Conference on Computer Vision*, pages 287–303. Springer, 2024. 2, 8, 6

[30] L. Li, H. Xie, N. Liu, D. Zhang, R. M. Anwer, H. Cholakkal, and J. Han. Advanced discriminative co-saliency and background mining transformer for co-salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 8

[31] Yin Li, Xiaodi Hou, Christof Koch, James M Rehg, and Alan L Yuille. The secrets of salient object segmentation. In *CVPR*, pages 280–287, 2014. 6, 8

[32] Jiazhen Liu and Long Chen. Segmentation as a plug-and-play capability for frozen multimodal llms. *arXiv preprint arXiv:2510.16785*, 2025. 2, 3

[33] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, pages 3089–3098, 2018. 1, 2

[34] Nian Liu, Ni Zhang, Kaiyuan Wan, Ling Shao, and Junwei Han. Visual saliency transformer. In *ICCV*, pages 4722–4732, 2021. 2, 7, 8, 6

[35] Yi Liu et al. Part-object relational visual saliency. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3688–3704, 2021. 1, 2

[36] Zhe Liu, Chen Chen, Wei Li, Peng Qi, Tian Pang, Chen Du, and Ming Lin. Understanding r1-zero-like training: A critical perspective. *arXiv preprint arXiv:2503.20783*, 2025. 3, 6, 7

[37] Allan H Murphy and Robert L Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 26(1):41–47, 1977. 5

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 32:8026–8037, 2019. 6

[39] J. Pei, H. Tang, C. Liu, et al. Salient instance segmentation via subitizing and clustering. *Neurocomputing*, 402:423–436, 2020. 2, 7, 8

[40] J. Pei, T. Cheng, H. Tang, and C. Chen. Transformer-based efficient salient instance segmentation networks with orientative query. *IEEE Transactions on Multimedia*, 25:1964–1978, 2022. 2, 6, 7, 8

[41] Sara Sarto, Marcella Cornia, and Rita Cucchiara. Image captioning evaluation in the age of multimodal llms: Challenges and future perspectives. *arXiv preprint arXiv:2503.14604*, 2025. 1

[42] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, D. Guo, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 2, 3, 4

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5

[44] L. Tang, B. Li, S. Kuang, M. Song, and S. Ding. Re-thinking the relations in co-saliency detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(8):5453–5466, 2022. 2

[45] V Team, Wenyi Hong, Wenmeng Yu, et al. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025. 1, 7, 8, 5

[46] Chong Wang, Zheng-Jun Zha, Dong Liu, and Hongtao Xie. Robust deep co-saliency detection with group semantic. In *AAAI*, pages 8917–8924, 2019. 6

[47] Jie Wang, Nana Yu, Zihao Zhang, and Yahong Han. Visual consensus prompting for co-salient object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 9591–9600, 2025. 2, 8, 6

[48] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, pages 136–145, 2017. 6

[49] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li. Solo: Segmenting objects by locations. In *European conference on computer vision*, pages 649–665. Springer, 2020. 2

[50] XuDong Wang, Shaolun Zhang, Shufan Li, Konstantinos Kallidromitis, Kehan Li, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Segllm: Multi-round reasoning segmentation. *arXiv preprint arXiv:2410.18923*, 2024. 2

[51] Y. Wang, R. Wang, X. Fan, et al. Pixels, regions, and objects: Multiple enhancement for salient object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10031–10040, 2023. 7, 8, 6

[52] Cong Wei, Yujie Zhong, Haoxian Tan, Yong Liu, Zheng Zhao, Jie Hu, and Yujiu Yang. Hyperseg: Towards universal visual segmentation with large language model. *arXiv preprint arXiv:2411.17606*, 2024. 2, 3

[53] Jun Wei, Shuhui Wang, Zhe Wu, Chi Su, Qingming Huang, and Qi Tian. Label decoupling framework for salient object detection. In *CVPR*, pages 13025–13034, 2020. 1, 2

[54] Weixi Weng, Rui Zhang, Xiaojun Meng, Jieming Zhu, Qun Liu, and Chun Yuan. Unsupervised domain adaptive visual question answering in the era of multi-modal large language models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6248–6258. IEEE, 2025. 1

[55] Yu-Huan Wu, Yun Liu, Le Zhang, Wang Gao, and Ming-Ming Cheng. Regularized densely-connected pyramid network for salient instance segmentation. *IEEE TIP*, 30:3897–3907, 2021. 2, 7, 8, 6

[56] x.ai. Grok 4 — xai. *x.ai News*, 2025. 1, 7, 8, 5

[57] Shiyu Xuan, Qingpei Guo, Ming Yang, and Shiliang Zhang. Pink: Unveiling the power of referential comprehension for multi-modal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13838–13848, 2024. 1

[58] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1155–1162, 2013. 6, 7, 8

[59] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, others, and Z. Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 7, 8, 4, 5

[60] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013. 6, 8

[61] Qi Yu, Zhe Zhang, Ruili Zhu, Yue Yuan, Xu Zuo, Yao Yue, and Mingshi Wang. Dapo: An open-source llm reinforcement learning system at scale. *arXiv preprint arXiv:2503.14476*, 2025. 3, 6, 7

[62] Siyue Yu, Jimin Xiao, Bingfeng Zhang, and Eng Gee Lim. Democracy does matter: Comprehensive feature mining for co-salient object detection. In *CVPR*, pages 979–988, 2022. 2

[63] Yi Ke Yun and Weisi Lin. Selfreformer: Self-refined network with transformer for salient object detection. *arXiv preprint arXiv:2205.11283*, 2022. 2

[64] Y. K. Yun and W. Lin. Towards a complete and detail-preserved salient object detection. *IEEE Transactions on Multimedia*, 26:4667–4680, 2023. 7, 8, 6

[65] Dingwen Zhang, Junwei Han, Chao Li, Jingdong Wang, and Xuelong Li. Detection of co-salient objects by looking deep and wide. *IJCV*, 120(2):215–232, 2016. 6, 8

[66] Ni Zhang, Junwei Han, Nian Liu, and Ling Shao. Summarize and search: Learning consensus-aware dynamic convolution for co-saliency detection. In *ICCV*, 2021. 2

[67] Yuxuan Zhang, Tianheng Cheng, Lianghui Zhu, Rui Hu, Lei Liu, Heng Liu, Longjin Ran, Xiaoxin Chen, Wenyu Liu, and Xinggang Wang. Evf-sam: Early vision-language fusion for text-prompted segment anything model. *arXiv preprint arXiv:2406.20076*, 2024. 3, 6

[68] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, pages 455–472, 2020. 8

[69] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In *ECCV*, pages 455–472, 2020. 6

[70] Cheng Zheng, Siyuan Liu, Minghui Li, Xin H. Chen, Bo Yu, Chen Gao, and Jiaxuan Lin. Group sequence policy optimization. *arXiv preprint arXiv:2507.18071*, 2025. 3, 6, 7

[71] Peng Zheng, Huazhu Fu, Deng-Ping Fan, Qi Fan, Jie Qin, Yu-Wing Tai, Chi-Keung Tang, and Luc Van Gool. Gconet+: A stronger group collaborative co-salient object detector. *IEEE TPAMI*, 2023. 2, 7, 8, 6

[72] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, others, and W. Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 1, 7, 8, 5

[73] Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient object detection via integrity learning. 2022. 1, 2