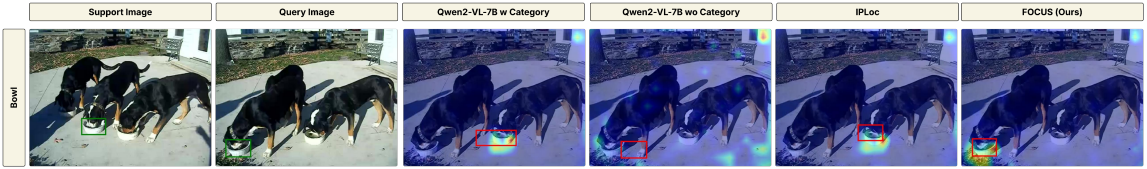
FOCUS: Forcing In Context Object Localization through Visual Support Constraints and Policy Optimization

Mohammed Asad Karim Vinay Kumar Verma

Problem: In Context Object Localization

Goal: Given a few support images with bounding boxes, localize the same object in a query image.



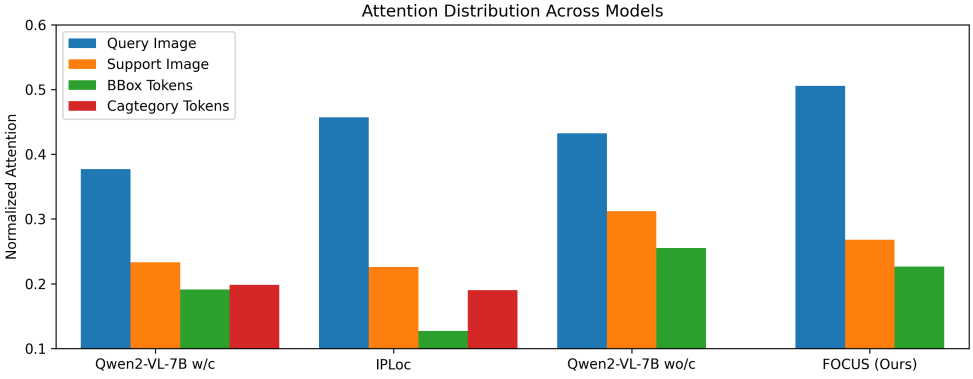
Why this matters:

- User-defined objects may not have clear category names.
- Applications include image editing, personalized visual search, retrieval, and interactive tracking.
- The model must infer the object from visual examples without updating its parameters.

Key Limitation of Existing VLMs

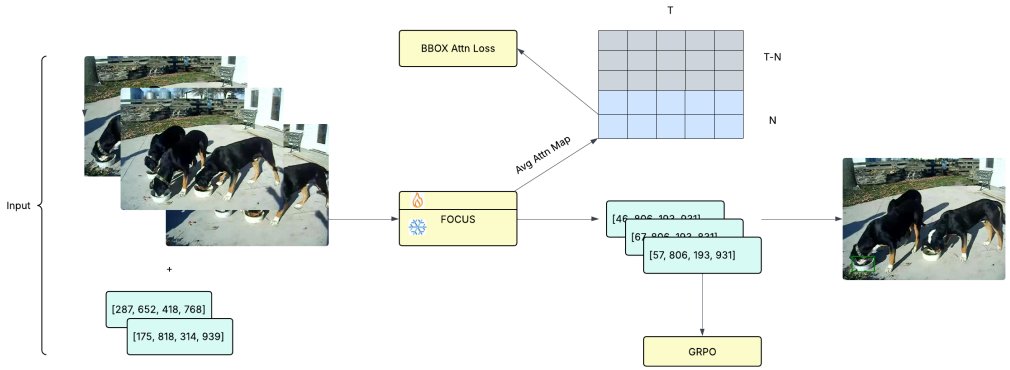
Existing approaches often rely on **category names** or semantic priors.

- Category labels bias the model toward prototypical or visually salient objects.
- When multiple objects share the same category, the model may localize the wrong instance.
- Removing category names alone is not sufficient, because the attention distribution may remain diffuse and weakly grounded.



FOCUS: Pure Visual In Context Localization

We propose a category-agnostic framework in which the model uses only visual support examples and bounding boxes.



Two-stage training:

1. **Attention-optimized SFT:** encourages query visual tokens to attend to support bounding box information.
2. **GRPO refinement:** directly optimizes localization quality using IoU and formatting rewards.

Stage 1: Bounding Box Attention Optimization

Goal: encourage query image tokens to attend to support bounding box tokens.

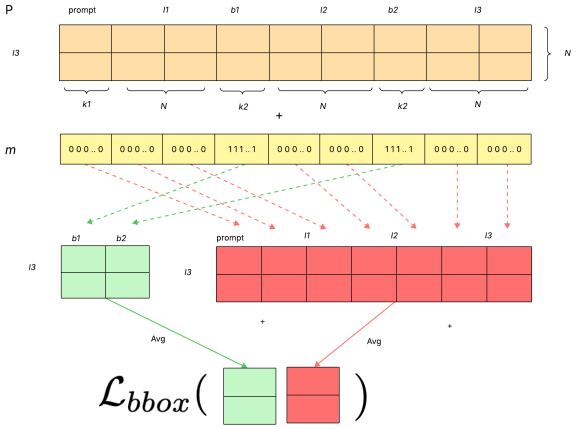
Let P_{ij} denote the attention from query image token i to input token j, and let $m_j = 1$ indicate support BBOX tokens.

$$p_i^+ = \frac{\sum_j P_{ij} m_j}{\sum_j m_j}, \qquad p_i^- = \frac{\sum_j P_{ij}(1 - m_j)}{\sum_j (1 - m_j)}$$

$$\Delta_i = p_i^+ - p_i^-$$

$$\mathcal{L}_{bbox} = \frac{1}{N} \sum_{i=1}^{N} \max(0, \mu - \Delta_i)^2$$

$$\mathcal{L}_{SFT} = \mathcal{L}_{LM} + \beta \mathcal{L}_{bbox}$$



This stage encourages the model to use support boxes as grounding evidence instead of relying only on semantic priors.

Stage 2: GRPO for Bounding Box Refinement

Goal: directly improve bounding box alignment after visual grounding.

For each query, the model samples multiple candidate bounding boxes. Each candidate receives a reward based on localization accuracy and output validity.

$$R = r_{iou} + r_{format}$$

$$r_{iou} = IoU(b_{pred}, b_{gt})$$

$$r_{format} = \begin{cases} 0.5, & \text{if the output follows the required BBOX format} \\ 0, & \text{otherwise} \end{cases}$$

Effect

GRPO assigns higher advantage to better localized boxes, improving precise bounding box alignment after attention based grounding.

Main Results

FOCUS outperforms stronger and larger VLM baselines.

Model	TAO 4 shot	GOT 4 shot	ICL LaSOT 4 shot
Qwen2 VL 7B	36.1	39.3	25.0
IPLoc 7B	56.1	68.7	59.4
Qwen2 VL 72B	55.6	59.9	55.4
InternVL2 76B	57.5	65.4	52.5
FOCUS	**68.5**	**82.6**	**65.6**

- FOCUS 7B surpasses 72B and 76B baselines.
- Gains are strongest in multi-object and unseen-category settings.
- This suggests that localization objectives matter more than scale alone.

Ablation: What Drives the Gain?

Model	1 shot	2 shot	4 shot
Qwen2 VL 7B	26.0	31.6	36.1
SFT	22.1	28.9	34.3
GRPO only	19.4	26.4	32.1
SFT plus attention loss	51.7	54.0	57.1
SFT plus attention loss plus GRPO	**55.8**	**63.0**	**68.5**

Takeaway:

- Naive SFT does not solve visual grounding.
- Attention loss provides the largest improvement.
- GRPO further improves precise bounding box alignment.

Conclusion

FOCUS reframes in context localization as visual correspondence rather than category recognition.

- Removes category labels from both support and query prompts.
- Forces the model to use support bounding boxes and visual context.
- Uses attention optimization for grounding and GRPO for precise localization.
- Achieves strong gains over much larger VLMs.

Main message

For in context object localization, the right training objective can matter more than simply scaling the model.

ICML International Conference On Machine Learning