

Reason with Thumbnails, Answer with Focus

An Efficient and Effective Paradigm for Multimodal
Grounded Visual Reasoning

An-Lan Wang, Guozhi Tang, Lei Liao, Hanshen Zhu, Kai Huang,
Jingqun Tang, Jiaming Zhou, Kun-Yu Lin

ICML 2026

Motivation: grounded reasoning is interpretable, but costly

GVR trains the model to select relevant image regions before answering the question, but multi-round “ground-then-answer” reasoning imposes much more computational costs.

Why GVR?

- Conventional MLLMs typically rely on a “direct answer” approach.
- This suffers from low answer reliability and poor interpretability, and makes the model prone to hallucinations.
- GVR improves interpretability by selecting relevant image regions during the chain-of-thought process.

The bottleneck

- For every rollout round, the context includes the original image, textual query, and evidence-aligned image regions.
- The resulting visual-token cost is much higher than traditional non-GVR methods.

Key Idea: Thumbnails Often Preserve Enough Cues

In many cases, the key area required to answer questions can be inferred from low-resolution thumbnails, without the need for a full-resolution image.

Q: What's the title of this paper ?



Origin Resolution: 1062x1375



Resolution: 531x687



Resolution: 132x171

Thumbnail-first intuition

- When loading an image, a thumbnail is often displayed first; we can already grasp most of the content based on the context and the thumbnail.
- When the full image finishes loading, our attention focuses directly on the key regions.
- Therefore, thumbnails can support the grounded visual reasoning process.

Reason with Thumbnails



Answer with Focus

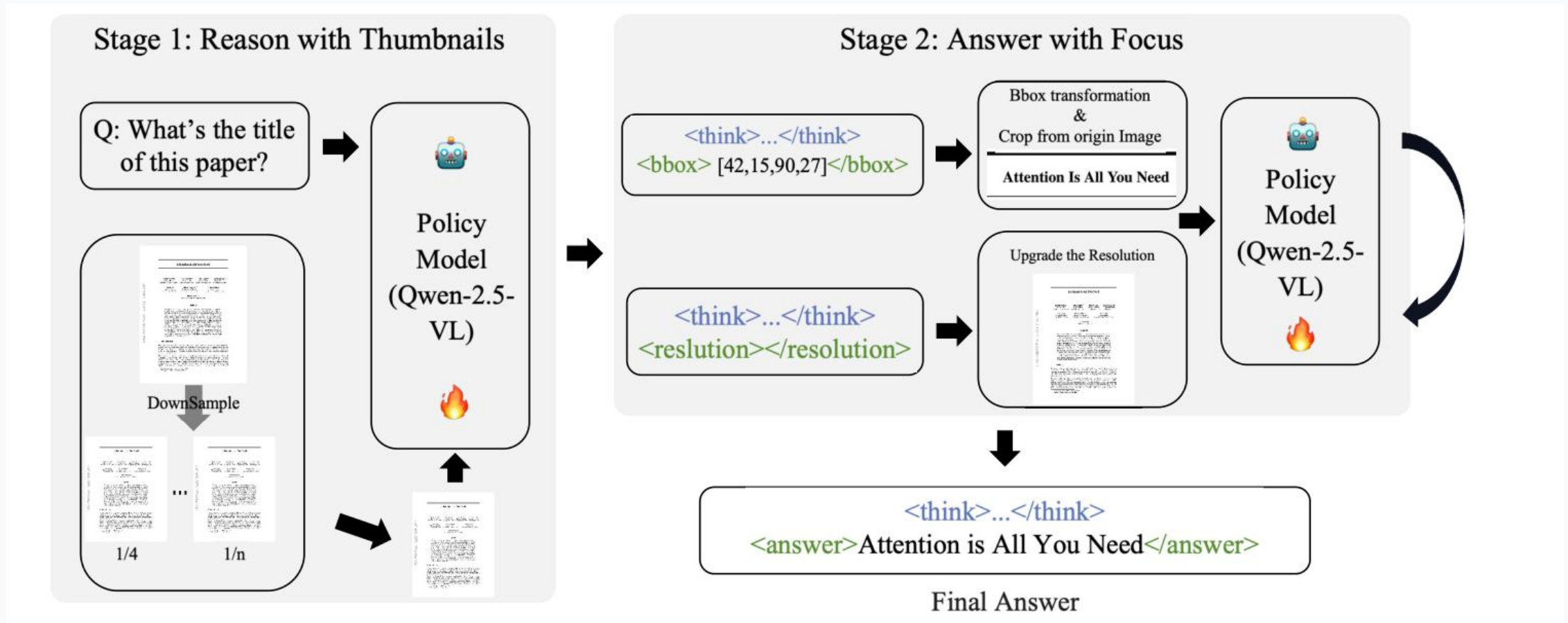
<think>The title “Attention is All You Need” is clearly visible at the top...</think>
<answer> Attention is All You Need </answer>

<think>...the bounding box around the title area to get a closer view... </think>
<bbox> [42,15,90,27] </bbox>

<think>... the second image provides clearly title...</think>
<answer>Attention is All You Need </answer>

Overall Paradigm: Reason with Thumbnails, Answer with Focus

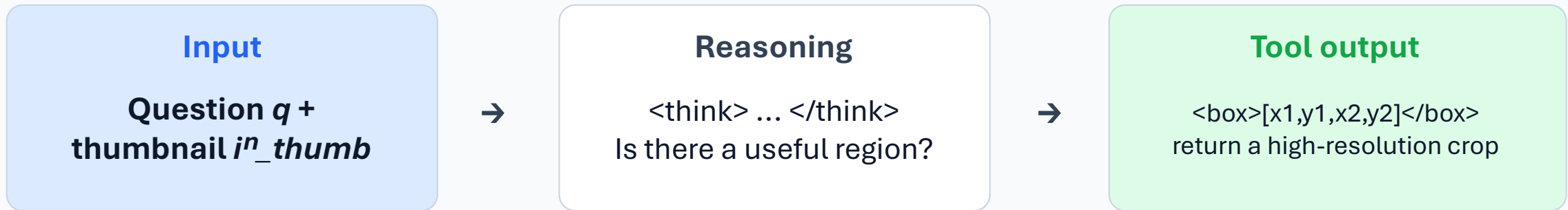
Unlike previous methods that use the full-resolution original image for reasoning, RTAF uses a low-resolution image to reason the key region and high-resolution crops to answer.



Low-resolution thumbnail → key region reasoning → coordinate transformation → high-resolution crop → final answer

Stage I: Reason with Thumbnails

The model reasons over a downsampled thumbnail and decides whether a specific region can better assist in answering the question.

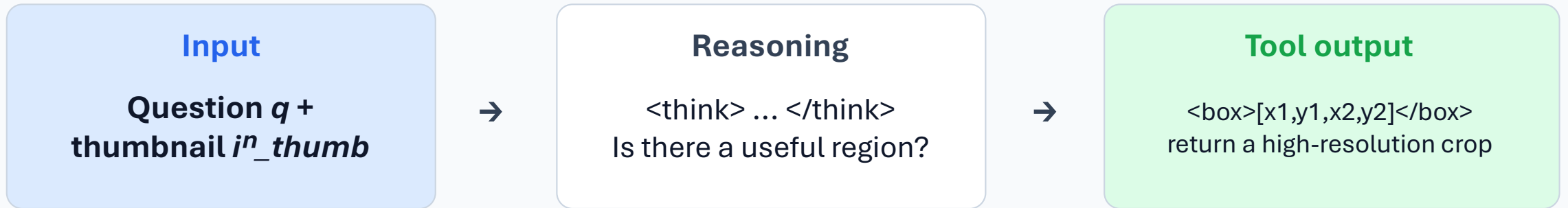


Extreme cases

- If the low-resolution image is insufficient, the model may perform unguided guessing of the key region.
- To avoid random guessing, RTAF equips the model with a tool to get images with higher resolution using a `<resolution></resolution>` tag.
- This process can be executed in multiple rounds until the model locates the correct key region.

Stage II: Answer with Focus

When no additional bounding boxes or higher-resolution images are needed, the model directly outputs the final answer within the <answer> tag.



Final answering

Based on all previous context e (e.g., the preceding reasoning trace), the model generates the final answer o .

$$o \sim \pi_{\theta_old}(\cdot | q, v, e)$$

Rewards

We design a suite of reward functions to supervise the model's behavior, including a reward designed specifically for reasoning efficiency .

$$R = R_{\text{format}} + R_{\text{acc}} + R_{\text{bbox}} + R_{\text{eff}}$$

Format

Ensure the model's prediction meets the required output format.

Accuracy

Exact answer match for numerical answers; ROUGE for free-form questions.

Bbox

Incentivize precise, question-relevant bounding boxes via IoU.

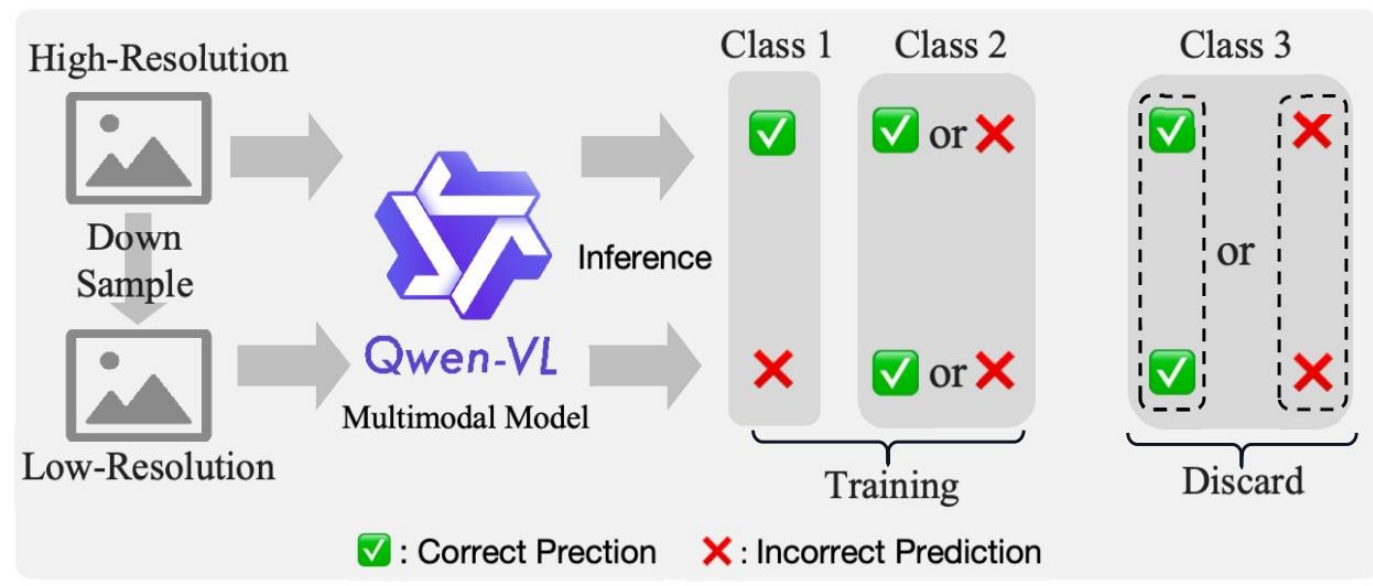
Efficiency

Encourage the model to answer directly when the answer is clearly visible.

Key insight: answer correctness remains the primary goal, while the model learns when to use tools and when to stop.

Resolution-aware Training Data Selection

Overly difficult or overly simple data can easily lead to training instability and inefficiency, such as reward collapse.



Selection pipeline

- Generate both high-resolution and downsampled low-resolution versions for each training sample.
- Perform multiple inferences using Qwen2.5-VL and categorize samples by resolution sensitivity.
- Use high-value samples to facilitate effective training.

Class 1: resolution sensitive

Class 2: resolution-neutral

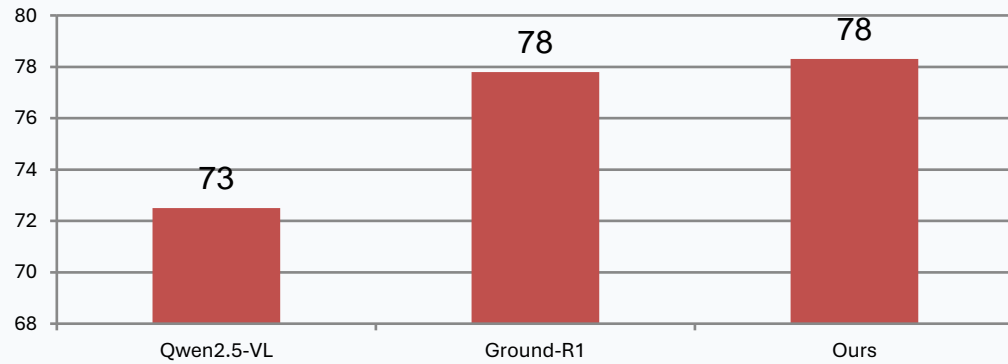
Class 3: discard samples

We use all Class 1 samples (2k) and 6k samples from Class 2; discard overly difficult/simple samples.

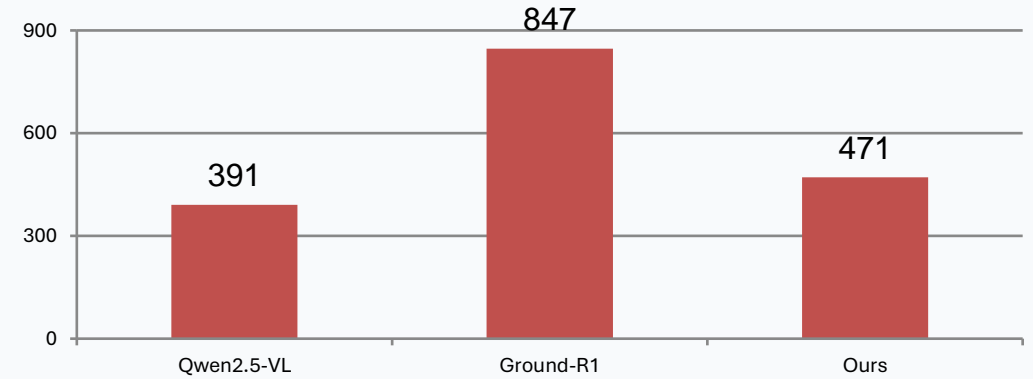
Main Results: Effective and Efficient GVR

RTAF achieves better or comparable performance with fewer visual tokens across VisCoT, grounding, and high-resolution benchmarks.

VisCoT average accuracy ↑



Average visual tokens ↓



+5.8%

performance gain over Qwen2.5-VL with comparable visual tokens

471 vs. 847

visual-token usage vs. Ground-R1

96.3

RefCOCO testA grounding score

69.5

HR-Bench-8K accuracy, tied best with fewer tokens

Ablations and Case Studies

Each additional component contributes to reducing token usage and, in most cases, improving accuracy.

Model	AVG ACC ↑	AVG V Token ↓
Baseline	77.8	847
+ RTAF	76.2	627
+ Reward	77.1	553
+ Data Selection	78.3	471

- RTAF lowers average visual tokens from 847 to 627, but low-resolution reasoning is more difficult.
- Reward functions improve accuracy while further reducing token usage.
- Resolution-aware data selection reaches the best balance: 78.3 accuracy with 471 tokens.

Ablations and Case Studies

Each additional component contributes to reducing token usage and, in most cases, improving accuracy.



Origin Image Resolution: 500*313 Thumbnail Resolution: 250*156

Question: *What activity is the man in the image engaged in?*

Answer:

Iteration1:

`<think>` The question asks about the activity of the man in the image. Since the man appears to be near a body of water, it's reasonable to infer he might be fishing. To confirm this, I need a closer view of his actions.`</think>`

`<box>`[179,149,230,168]`</box>`



Iteration2:

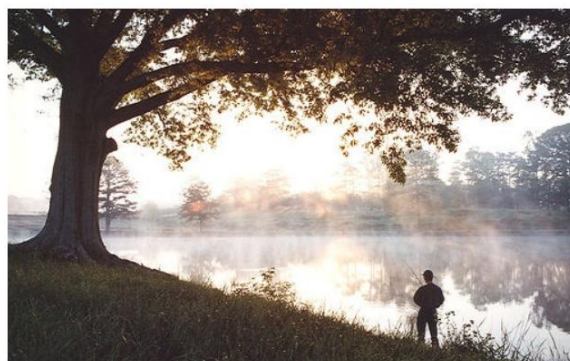
`<think>` The first image shows a man near a body of water, possibly fishing. In the second image, the man appears to be **standing in a field**, which suggests he might be engaged in an outdoor activity. However, the activity is not clearly visible due to the low resolution. To get a clearer view of what the man is doing, we need a higher-resolution image.`</think>`

`<resolution>` `</resolution>`

Iteration3:

`<think>` The man in the image appears to be fishing, as he is holding a fishing rod and standing by a body of water. There is no need for a higher-resolution image since the activity can be clearly identified from the image.`</think>`

`<answer>` The man is fishing. `</answer>`



Case studies show that RTAF uses low-resolution thumbnails for region reasoning and high-resolution crops for precise answering, balancing efficiency and effectiveness.

Thank You

Reason with Thumbnails, Answer with Focus