# CoFFT: Chain of Foresight-Focus Thought for Visual Language Models

Xinyu Zhang[1,2], Yuxuan Dong[1,2], Lingling Zhang[1,2 *], Chengyou Jia[1,2], ZhuoHang Dang[1,2], Besura Fernando[4,6], Jun Liu[1,3] , Mike Zheng Shou[5 *]

[1] School of Computer Science and Technology, Xi'an Jiaotong University
[2] Ministry of Education Key Laboratory of Intelligent Networks and Network Security, China
[3] Shaanxi Province Key Laboratory of Big Data Knowledge Engineering, China
[4] IHPC, Agency for Science, Technology and Research, Singapore
[5] Show Lab, National University of Singapore
[6] College of Computing and Data Science, Nanyang Technological University, Singapore
[*] Corresponding Author

# Background

## Challenges in VLM Performance

- Inherent complexity and redundancy in visual inputs significantly constrain VLM capabilities
- High sensitivity to visually salient yet semantically irrelevant elements

## Multimodal Chain-of-Thought Solution

- Enhances reasoning through prompt engineering
- Leverages different image information at various reasoning stages
- Unassessed regions may introduce irrelevant interference

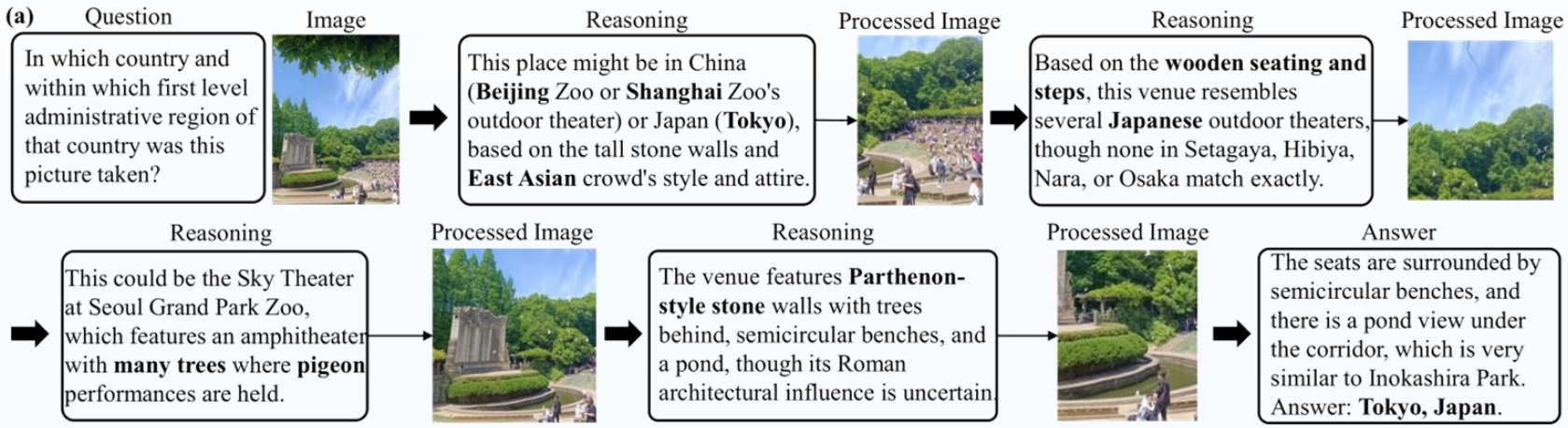
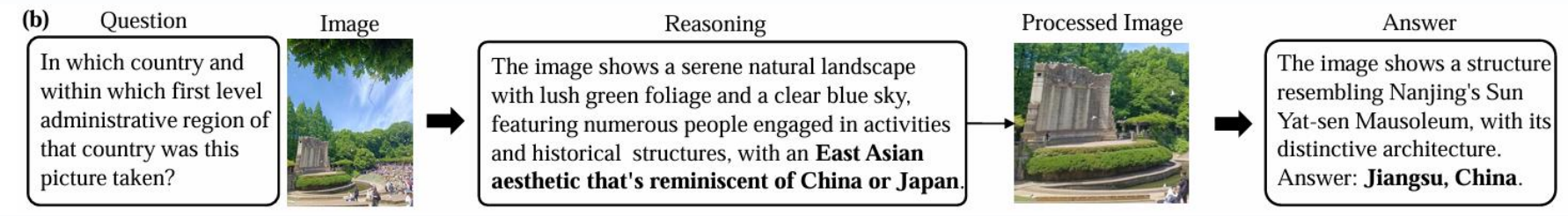



**Figure 1:** An example from the SeekWorld. (a) is the reasoning process of o3, and (b) is the reasoning process of o3 after human visual cognition. The correct answer is Jiangsu, China.

# Motivation

- **When humans analyze complex visual scenes, they evaluate which visual areas are most valuable for future reasoning**




Reduce attention to common elements (e.g., crowds, grids, trees); prioritize unique historical buildings.

**Figure 2:** An example from the SeekWorld. (b) is the reasoning process of o3 after human visual cognition. The correct answer is Jiangsu, China.

- **Two core capabilities of human beings in analyzing complex visual scenes**
  - **Forward-looking: predict the most valuable visual area in future reasoning**
  - **Dynamic visual focus: precisely shift your attention to the areas most relevant to future reasoning**

# Our work

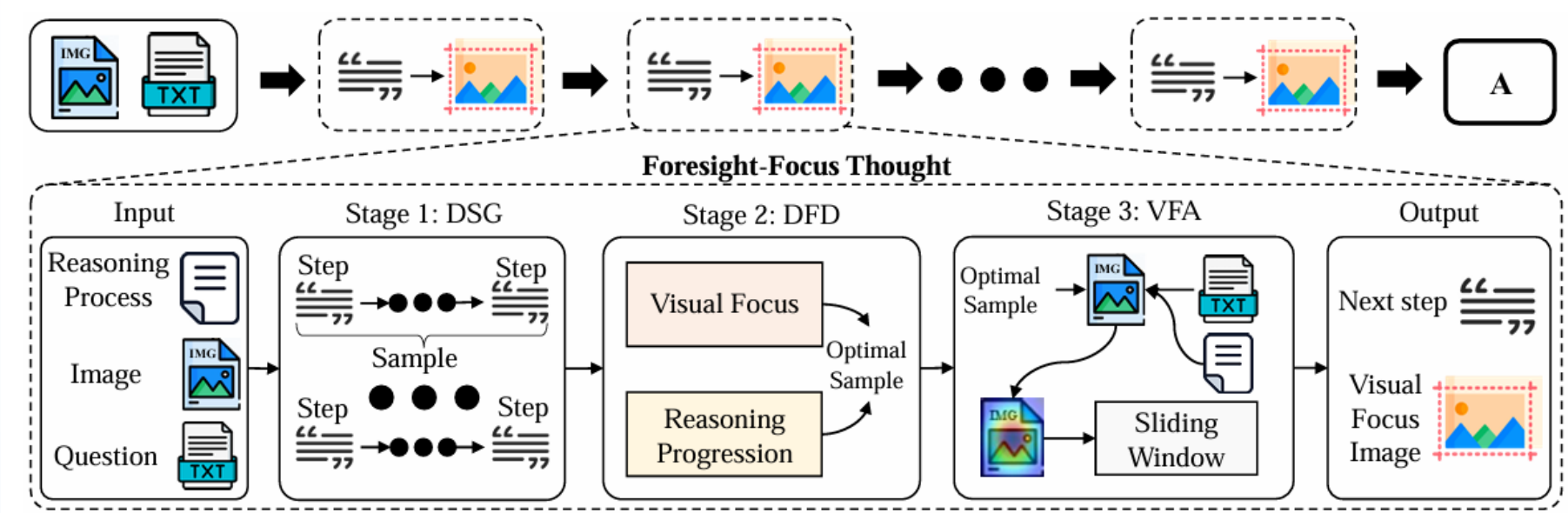



**Figure 3:** The overall approach of CoFFT, where Dual Foresight Decoding and Visual Focus Adjustment will be introduced in detail later.

**Stage 1: Diverse Sample Generation (DSG)**

**Stage 2: Dual Foresight Decoding (DFD)**

**Stage 3: Visual Focus Adjustment (VFA)**

# Stage 1: Diverse Sample Generation

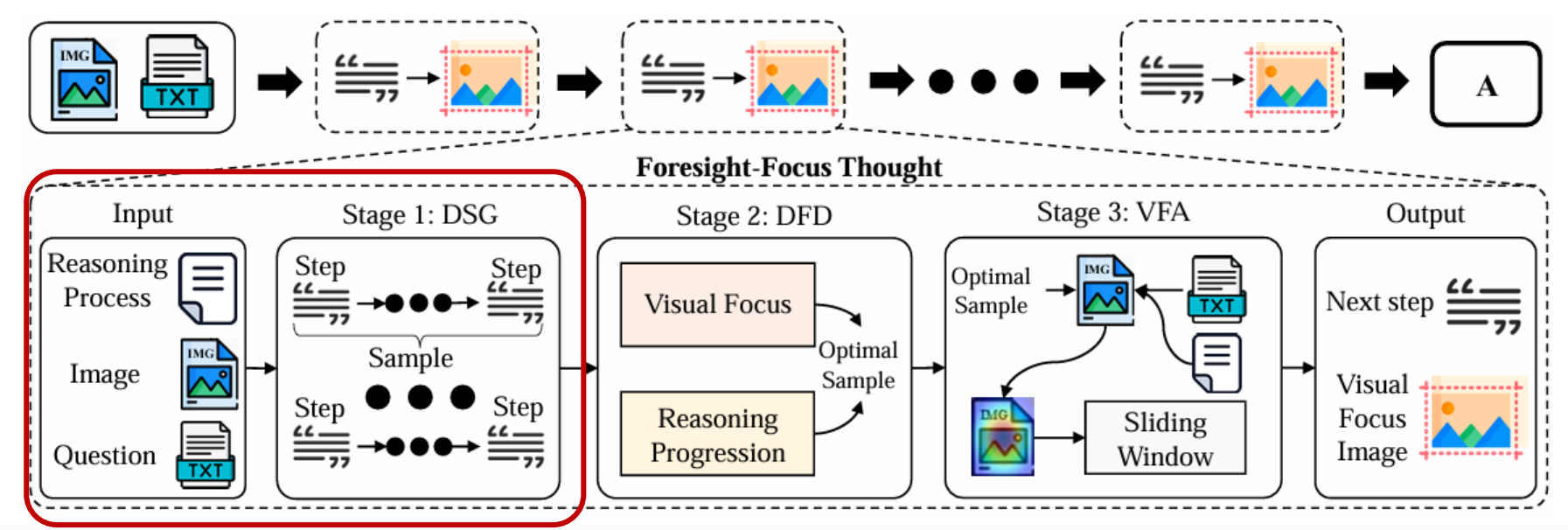



**Figure 3:** The overall approach of CoFFT, where Dual Foresight Decoding and Visual Focus Adjustment will be introduced in detail later.

- **The visual language model generates multiple candidate reasoning samples based on the current reasoning process, visually focused images, and the original problem.**

1. **Different temperature** coefficients are adopted to ensure **sample diversity**.
2. Each sample can retain a maximum of the **specified number of steps** in the reasoning process.

# Stage 2: Dual Foresight Decoding

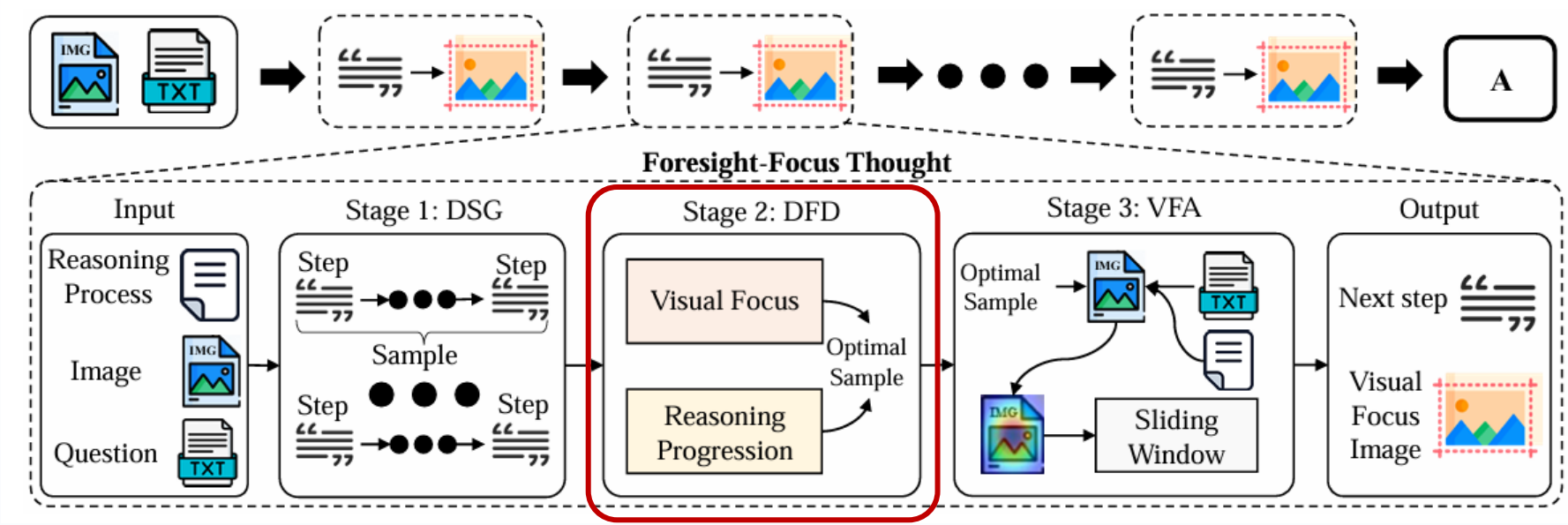



**Figure 3:** The overall approach of CoFFT, where Dual Foresight Decoding and Visual Focus Adjustment will be introduced in detail later.

- **The candidate samples are evaluated using the visual focus score and the reasoning progress score. The first step of selecting the best sample is integrated into the next reasoning process for iterative reasoning.**

1. **Visual Focus Score**: Measures the correlation between the reasoning process and the image.
2. **Reasoning Progress Score**: Quantifies the probability increase across reasoning steps.

**Fix the image and select the best next reasoning step.**

# Stage 3: Visual Focus Adjustment

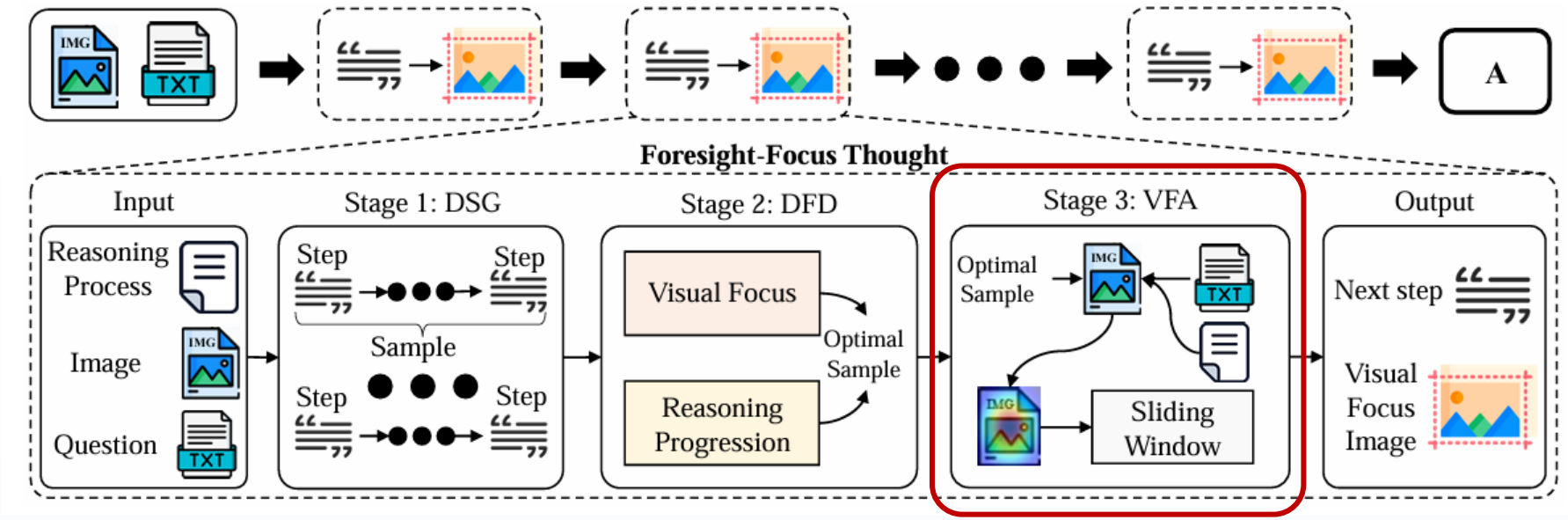



**Figure 3:** The overall approach of CoFFT, where Dual Foresight Decoding and Visual Focus Adjustment will be introduced in detail later.

- **Use the following scoring mechanism to evaluate the image:**
  1. **The relevance between the image and the problem.**
  2. **The relevance between the image and the future reasoning steps of the best sample.**
- **Slide the crop window, crop and enlarge the area with the highest average score as the visually focused image.**

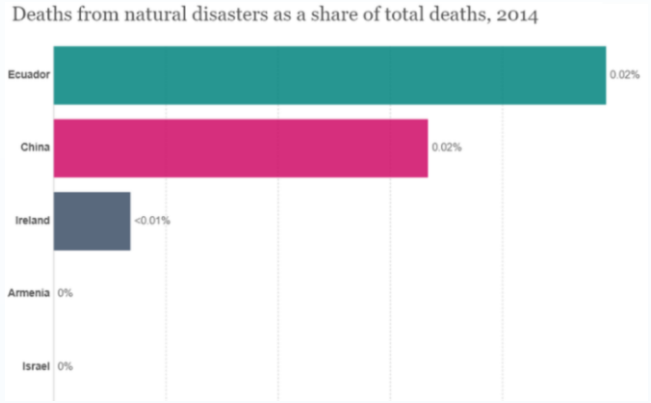
**Fix the reasoning steps and select the best visual aggregation image.**

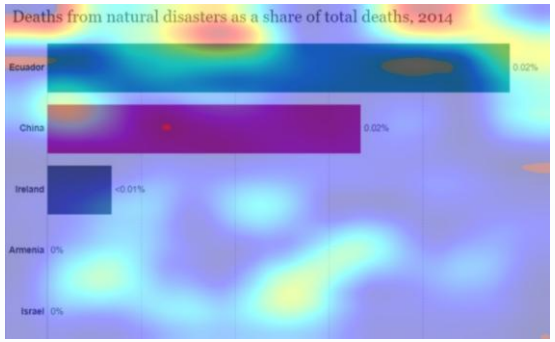
# Relative attention mechanism

**question:**

*Is the sum of two lowest bar is greater then the largest bar?*

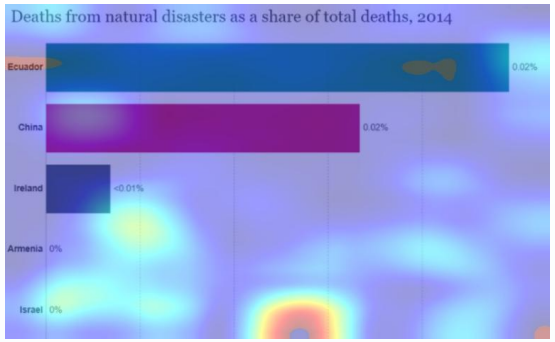
**image:**



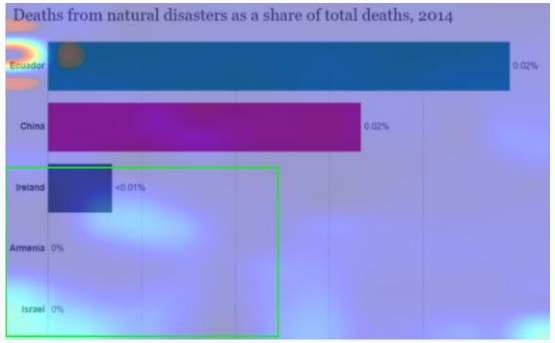

*Describe the image.*

$\boldsymbol{A(V, D)}$



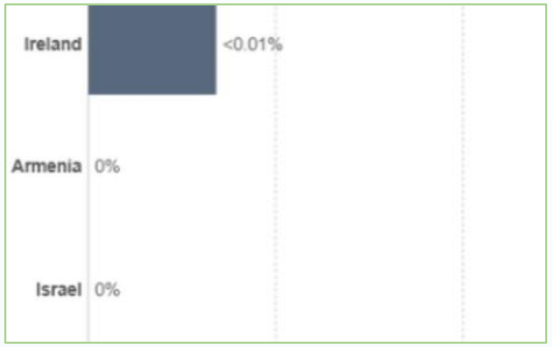

*1. Calculate the sum of the two smallest bars.*

$\boldsymbol{A(V, Q)}$




Relative attention

$\boldsymbol{A^{rel}(V, Q)}$




crop and enlarge

**Relative attention calculation formula:**

$$A^{rel}(V,Q) = Softmax\left(\frac{A(V,Q)}{A(V,D)+\varepsilon}\right)$$

# Dual Foresight Decoding

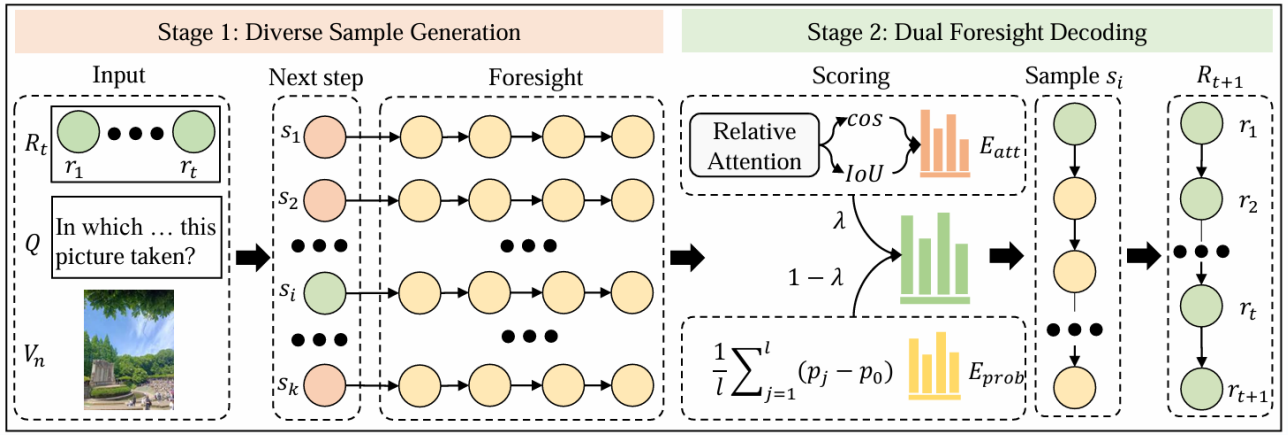



**Figure 4:** Dual Foresight Decoding, which evaluates different reasoning samples and selects the optimal sample based on visual focus and reasoning progression to enhance decision robustness.

**Given the candidate sample set $S_{t+1}$ generated at reasoning step $t+1$, a composite score is calculated for each sample $s \in S_{t+1}$（with a maximum of $l$ steps）:**

$$E_{t+1} = \lambda \cdot Softmax([E_{att}(s)]\forall s \in S_{t+1}) + (1-\lambda) \cdot Softmax([E_{prob}(s)]\forall s \in S_{t+1})$$

**1. The visual focus score $E_{att}$ is utilized to measure the focus alignment of each sample with the image:**

$$E_{att}(s) = 0.5 \cdot cos(A^{rel}(V,Q), A^{rel}(V,s)) + 0.5 \cdot IoU^{30\%}(A^{rel}(V,Q), A^{rel}(V,s))$$

**2. The reasoning progress score $E_{prob}$ serves to quantify the degree of improvement in reasoning quality:**

$$E_{prob}(s) = \frac{1}{l}\sum_{j=1}^{l}(p_j - p_0)$$

# Visual Focus Adjustment

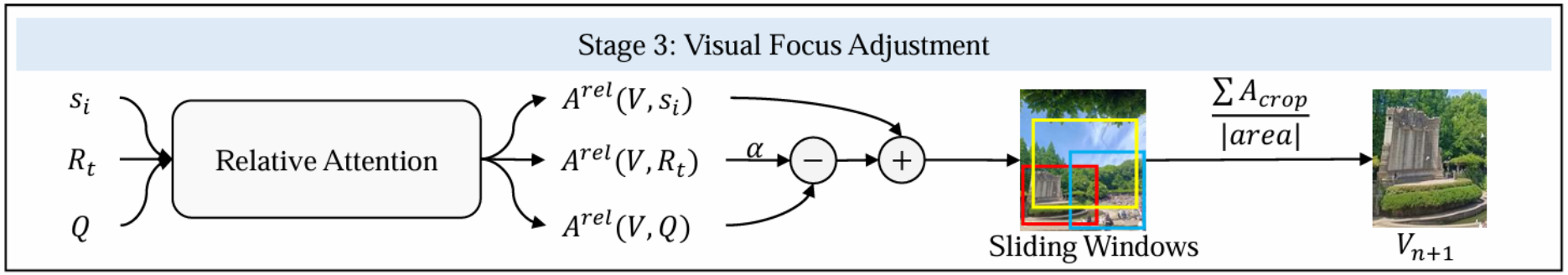



**Figure 5:** Visual Focus Adjustment, which adaptively modulates visual attention toward reasoning-relevant regions to optimize information understanding.

**Regarding the relevance between images and problems:**

$$C^{rel}(V, Q, R_t) = max(A^{rel}(V, Q) - \alpha \cdot A^{rel}(V, R_t), 0)$$

**Relevance of future reasoning steps between images and the best samples:**

$$A^{rel}(V, s_i)$$

**The final attention map assessment combines the two:**

$$A_{crop} = 0.5 \cdot C^{rel}(V, Q, R_t) + 0.5 \cdot A^{rel}(V, s_i)$$

**Select the best visual area using the dynamic sliding window $V_{t+1}$:**

$$V_{t+1} = \begin{cases} \arg max_{B \in \Omega} \sum_{(x,y) \in B} A_{crop}(x, y), if\, \mu_{B^*} > \mu_{V_0} + \beta \\ V_0, otherwise \end{cases}$$

$B^*$ represents the optimal region
$V_0$ represents the initial image

# Experimental Results

**Mathematical and geometric reasoning**

**Cross-disciplinary visual reasoning**

**Chart understanding**

| Models | Math | | Multi-subjects | | Chart | Geography | | Avg. |
|---|---|---|---|---|---|---|---|---|
| | MathVista | MathVision | MMStar | M3CoT | Charxiv | S.W-China | S.W-Global | |
| **Closed source VLMs** | | | | | | | | |
| GPT-4o | 63.8 | 18.75 | 64.7 | 65.75 | 50.5 | 31.90 | 56.50 | 50.27 |
| Claude-3.5 | 65.4 | 26.21 | 65.1 | 66.05 | 60.2 | 29.22 | 52.50 | 52.10 |
| Gemini-2 | 73.1 | 31.83 | 69.4 | 67.73 | 53.2 | 30.83 | 55.31 | 54.49 |
| **Qwen2.5VL-7B-Instruct** | | | | | | | | |
| Baseline | 68.2 | 18.09 | 63.9 | 59.62 | 42.5 | 21.45 | 25.31 | 42.72 |
| MCTS | 69.6 | 18.75 | 66.2 | 60.87 | 44.5 | 26.27 | 26.56 | 44.68 |
| Pred. Dec. | 69.9 | 19.73 | 65.7 | 61.34 | 45.3 | 26.54 | 26.86 | 45.05 |
| ICoT | 47.5 | 10.53 | 42.1 | 61.17 | 27.5 | 29.22 | 26.37 | 34.91 |
| DyFo | 68.4 | 16.78 | 64.5 | 61.26 | 43.7 | 32.44 | 27.81 | 44.98 |
| CoFFT | **70.4** | **23.36** | **69.4** | **62.47** | **47.2** | **35.12** | **29.37** | **48.19** |
| **LLaVA-NeXT-7B** | | | | | | | | |
| Baseline | 34.6 | 9.87 | 34.2 | 38.27 | 13.9 | 10.72 | 15.31 | 22.41 |
| MCTS | 35.1 | 10.53 | 36.7 | 39.52 | 14.6 | 12.33 | 16.56 | 23.62 |
| Pred. Dec. | 34.8 | 11.36 | 35.1 | 40.16 | 15.3 | 11.80 | 17.19 | 23.67 |
| ICoT | 27.3 | 11.18 | 31.2 | 39.34 | 9.5 | 12.87 | 16.56 | 21.14 |
| DyFo | 34.8 | 8.22 | 36.1 | 39.86 | 15.7 | 13.67 | 17.50 | 23.69 |
| CoFFT | **35.6** | **12.17** | **38.3** | **40.68** | **16.8** | **15.55** | **19.69** | **25.54** |
| **InternVL2.5-8B-Instruct** | | | | | | | | |
| Baseline | 64.4 | 22.00 | 60.5 | 57.16 | 32.9 | 23.32 | 25.63 | 40.84 |
| MCTS | 65.0 | 24.67 | 62.0 | 58.24 | 34.3 | 25.20 | 26.56 | 42.28 |
| Pred. Dec. | 65.4 | 25.00 | 62.4 | 58.76 | 35.1 | 26.81 | 27.19 | 42.95 |
| ICoT | 42.9 | 16.12 | 39.4 | 58.50 | 16.4 | 27.35 | 26.88 | 32.51 |
| DyFo | 64.7 | 20.39 | 61.5 | 58.58 | 34.2 | 29.22 | 28.13 | 42.39 |
| CoFFT | **66.5** | **28.29** | **64.5** | **59.19** | **36.6** | **31.37** | **30.63** | **45.30** |
| **Qwen2.5VL-32B-Instruct** | | | | | | | | |
| Baseline | 74.7 | 25.33 | 69.5 | 62.81 | 44.5 | 24.13 | 28.41 | 47.05 |
| MCTS | 76.2 | 27.31 | 70.6 | 64.15 | 47.6 | 28.69 | 29.06 | 49.09 |
| Pred. Dec. | 76.6 | 27.96 | 71.1 | 64.62 | 48.2 | 29.22 | 30.31 | 49.72 |
| ICoT | 58.7 | 21.05 | 54.6 | 63.93 | 47.3 | 32.17 | 31.56 | 44.19 |
| DyFo | 75.6 | 24.67 | 70.1 | 64.32 | 47.7 | 35.38 | 32.19 | 49.99 |
| CoFFT | **77.5** | **29.93** | **72.7** | **66.08** | **50.9** | **38.61** | **34.38** | **52.96** |

**SeekWorld-Global:** utilizing panoramic images from Google Maps;
**SeekWorld-China:** integrating data from the Xiaohongshu app.

**Optimal results were achieved in all datasets**

# Experimental Results

## ➢ Computational Overhead

| Models | Baseline | MCTS | Predictive decoding | ICoT | DyFo | CoFFT |
|---|---|---|---|---|---|---|
| FLOPS | 8.35e+12 | 4.05e+14 | 1.85e+14 | 1.88e+13 | 1.98e+13 | 2.38e+14 |

## ➢ Ablation Study on Qwen2.5-VL-7B-Instruct

***Sample evaluation relies solely on the reasoning progress score.***

***The reasoning process employs only the original image.***

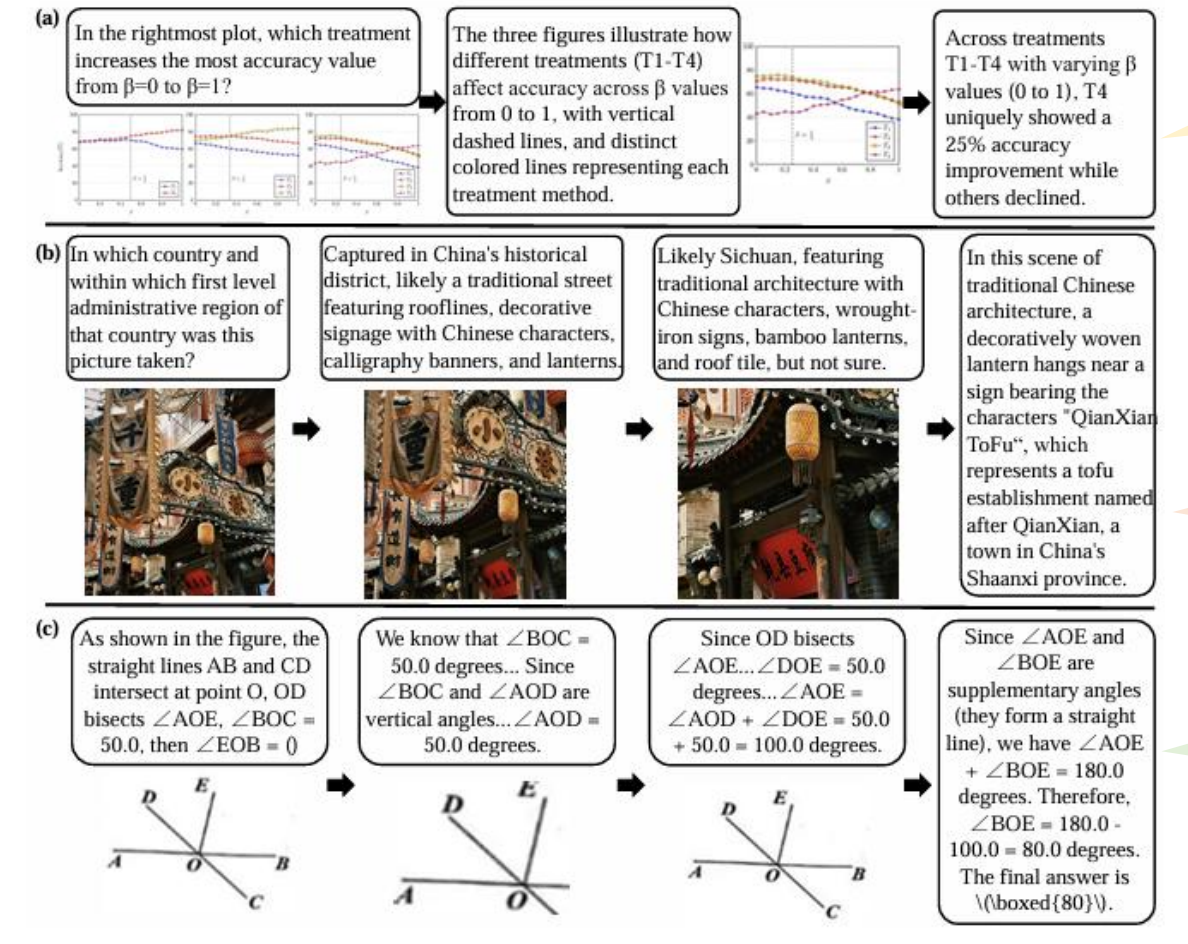
| Models | Math | | Multi-subjects | | Chart | Geography | |
|---|---|---|---|---|---|---|---|
| | MathVista | MathVision | MMStar | M3CoT | Charxiv | S.W-China | S.W-Global |
| Our | **70.4** | **23.36** | **69.4** | **62.47** | **47.2** | **35.12** | **29.37** |
| w/o DFD | 68.5 | 20.42 | 66.5 | 61.39 | 44.8 | 28.42 | 27.19 |
| w/o VFA | 69.3 | 21.71 | 67.4 | 61.09 | 44.7 | 27.08 | 26.25 |

## ➢ Impact of Foresight Param *l* & Sample Size *K*

***Study of parameter k with l held constant at 5.***
***Study of parameter l with k fixed at 4.***

| Foresight | MathVista | S.E-China | FLOPS | Sampling | MathVista | S.E-China | FLOPS |
|---|---|---|---|---|---|---|---|
| $l = 3$ | 68.8 | 33.51 | 1.41e+14 | $k = 2$ | 68.8 | 34.32 | 1.19e+14 |
| $l = 4$ | 69.3 | 34.58 | 1.91e+14 | $k = 4$ | 70.4 | 35.12 | 2.38e+14 |
| $l = 5$ | 70.4 | 35.12 | 2.38e+14 | $k = 6$ | 70.8 | 35.65 | 3.56e+14 |
| $l = 6$ | 70.5 | 35.65 | 2.86e+14 | $k = 8$ | 71.4 | 36.19 | 4.75e+14 |
| $l = 7$ | 70.7 | 35.92 | 3.33e+14 | $k = 10$ | 72.2 | 37.27 | 5.93e+14 |

# Case Study




**CoFFT successfully identified relevant diagrams to correctly answer questions about the full image.**

**CoFFT robustly identifies critical data and maintains accuracy amidst distracting information.**

**CoFFT reasons continuously by identifying key elements and dynamically refocusing on the source image.**

**Figure 6:** Illustrative cases demonstrating the reasoning process of CoFFT. Examples (a), (b) and (c) are respectively from Charxiv, SeekWorld-China and MathVista.

# Conclusion

- **CoFFT Introduction:** CoFFT, a training-free approach mimicking human visual cognition via three stages (Diverse Sample Generation, Dual-Foresight Decoding, Visual Focus Adjustment), bridges static visual processing and dynamic reasoning.
- **Experimental Gains:** Experiments show CoFFT boosts performance by 3.1% - 5.8% on challenging visual reasoning tasks, no specialized models or system changes needed.
- **Limitations & Future Work:** CoFFT solves new VLM problems but may cause errors in well-performing VLM cases, so robustness and systematic reasoning optimization need further research.

# Thanks

Arxiv Paper