

## Appendix

### A Algorithmic Details

---

**Algorithm 1** Chain of Foresight-Focus Thought (CoFFT)

---

**Require:** Original image  $V$ , question  $Q$ , VLM model  $M$ , number of samples  $k$ , foresight length  $l$ , balancing factor  $\lambda$ , exploration factor  $\alpha$ , temperature range  $[T_{\min}, T_{\max}]$

**Ensure:** Final reasoning result  $R_n$  with answer

```

1:  $V_0 \leftarrow V$  ▷ Initialize visual focus with original image
2:  $R_0 \leftarrow \emptyset$  ▷ Initialize reasoning chain as empty
3:  $t \leftarrow 0$  ▷ Initialize iteration counter
4: while not reached final answer do
5:    $t \leftarrow t + 1$  ▷ Stage 1: Diverse Sample Generation
6:    $S_t \leftarrow \emptyset$  ▷ Initialize empty sample set
7:   for  $i \leftarrow 1$  to  $k$  do
8:      $T_i \leftarrow \text{Random}(T_{\min}, T_{\max})$  ▷ Diverse temperature values
9:      $s_i \leftarrow \text{Generate}(M, V_{t-1}, Q, R_{t-1}, T_i, l)$  ▷ Generate sample with  $l$  steps
10:     $S_t \leftarrow S_t \cup \{s_i\}$ 
11:   end for ▷ Stage 2: Dual Foresight Decoding
12:   for each  $s \in S_t$  do
13:      $A^{rel}(V, s) \leftarrow \text{Softmax}(\frac{A(V,s)}{A(V,D)+\epsilon})$  ▷ Compute relative attention
14:      $E_{att}(s) \leftarrow 0.5 \cdot \cos(A^{rel}(V, Q), A^{rel}(V, s)) + 0.5 \cdot \text{IoU}^{30\%}(A^{rel}(V, Q), A^{rel}(V, s))$ 
15:      $p_0 \leftarrow \text{MeanLogProb}(R_{t-1})$  ▷ Base probability
16:      $E_{prob}(s) \leftarrow \frac{1}{l} \sum_{j=1}^l (\text{MeanLogProb}(R_{t-1} + s[1:j]) - p_0)$ 
17:   end for
18:    $E_{att\_norm} \leftarrow \text{Softmax}([E_{att}(s)] \forall s \in S_t)$ 
19:    $E_{prob\_norm} \leftarrow \text{Softmax}([E_{prob}(s)] \forall s \in S_t)$ 
20:    $E_t \leftarrow \lambda \cdot E_{att\_norm} + (1 - \lambda) \cdot E_{prob\_norm}$ 
21:    $s^* \leftarrow \arg \max_{s \in S_t} E_t(s)$  ▷ Select optimal sample
22:    $R_t \leftarrow R_{t-1} \cup \{\text{FirstStep}(s^*)\}$  ▷ Update reasoning with first step of optimal sample
▷ Stage 3: Visual Focus Adjustment
23:    $A^{rel}(V, R_t) \leftarrow \text{Softmax}(\frac{A(V,R_t)}{A(V,D)+\epsilon})$ 
24:    $C^{rel}(V, Q, R_t) \leftarrow \max(A^{rel}(V, Q) - \alpha \cdot A^{rel}(V, R_t), 0)$ 
25:    $A_{crop} \leftarrow 0.5 \cdot C^{rel}(V, Q, R_t) + 0.5 \cdot A^{rel}(V, s^*)$ 
26:    $\Omega \leftarrow \{\text{regions spanning 40\%-90\% of original dimensions}\}$ 
27:    $\mu_{V_0} \leftarrow \frac{1}{|V_0|} \sum_{(x,y) \in V_0} A_{crop}(x, y)$  ▷ Global mean score
28:    $\sigma_{V_0} \leftarrow \text{StandardDeviation}(A_{crop})$ 
29:    $\beta \leftarrow \sigma_{V_0} \cdot (1 - \cos(C^{rel}(V, Q, R_t), A^{rel}(V, s^*)))$ 
30:    $B^* \leftarrow \arg \max_{B \in \Omega} \frac{1}{|B|} \sum_{(x,y) \in B} A_{crop}(x, y)$  ▷ Find optimal region
31:    $\mu_{B^*} \leftarrow \frac{1}{|B^*|} \sum_{(x,y) \in B^*} A_{crop}(x, y)$  ▷ Maximum region score
32:   if  $\mu_{B^*} > \mu_{V_0} + \beta$  then
33:      $V_t \leftarrow \text{Crop}(V, B^*)$  ▷ Crop and scale region
34:   else
35:      $V_t \leftarrow V_0$  ▷ Revert to original image
36:   end if
37:   if  $R_t$  contains "REASONING_COMPLETE" then
38:     break
39:   end if
40: end while
41: return  $R_t$  ▷ Return final reasoning with answer

```

---

Algorithm 1 presents the complete Chain of Foresight-Focus Thought (CoFFT) approach, which systematically enhances visual reasoning through an iterative cognitive-inspired process. The algorithm leverages three interconnected stages that collaborate to produce accurate reasoning paths.

---

**Algorithm 2** Relative Attention Mechanism

---

**Require:** Image  $V$ , text input  $X$ , VLM model  $M$ , descriptive prompt  $D$  (e.g., "Describe the image in detail")

**Ensure:** Relative attention map  $A^{rel}(V, X)$

- 1:  $A(V, X) \leftarrow \text{ExtractAttentionMap}(M, V, X)$  ▷ Extract raw attention map
  - 2:  $A(V, D) \leftarrow \text{ExtractAttentionMap}(M, V, D)$  ▷ Extract descriptive attention map
  - 3:  $A^{rel}(V, X) \leftarrow \text{Softmax}(\frac{A(V, X)}{A(V, D) + \epsilon})$  ▷ Compute relative attention
  - 4: **return**  $A^{rel}(V, X)$
- 

In the Diverse Sample Generation stage (lines 4-11), the algorithm explores different reasoning trajectories by generating  $k$  diverse reasoning samples with varying temperature parameters. Each sample contains up to  $l$  future reasoning steps, providing a meaningful lookahead that helps evaluate potential reasoning directions before committing to the next step. This diversity is crucial for addressing complex visual reasoning tasks where multiple valid reasoning paths may exist.

The Dual Foresight Decoding stage (lines 12-22) represents the core evaluation mechanism, where samples are assessed through two complementary scores: a visual focus score ( $E_{att}$ ) that measures visual relevance, and a reasoning progression score ( $E_{prob}$ ) that evaluates logical coherence. These scores are normalized and combined using a balancing factor  $\lambda$ , allowing the algorithm to select reasoning steps that are both visually grounded and logically sound. The first step of the optimal sample is then integrated into the evolving reasoning process.

The Visual Focus Adjustment stage (lines 23-36) dynamically modifies visual attention based on current reasoning progress and future needs. It evaluates the significance of image regions using a dual-criteria scoring mechanism that considers both question relevance and future reasoning relevance. The algorithm then employs a sliding window approach with adaptive thresholding to select optimal regions for focus. This mechanism enables effective transitions between global context and local details throughout the reasoning process.

The iterative cycle continues until a conclusive answer is reached (signaled by "REASONING\_COMPLETE"), creating a synergistic loop where reasoning directs visual focus and optimized visual focus subsequently improves reasoning quality. This cognitively-inspired approach significantly enhances the model's ability to perform complex visual reasoning tasks that require attention to both global context and local details.

Algorithm 2 outlines the Relative Attention Mechanism, which serves as a foundational component for the visual focus calculations in Algorithm 1. This concise mechanism addresses the challenge of redundant information in images by normalizing text-image attention maps against a baseline descriptive attention distribution.

By performing element-wise division between task-specific attention and generic descriptive attention, followed by softmax normalization, this algorithm effectively highlights regions specifically relevant to the task while suppressing generally salient but task-irrelevant areas. This normalized attention map ( $A^{rel}$ ) is then utilized throughout Algorithm 1 to compute visual focus scores and guide the visual focus adjustment process, creating a coherent system that intelligently navigates visual information during complex reasoning tasks.

## B Implementation Details

### B.1 Parameter Configuration

The CoFFT algorithm requires several key parameters that influence its performance:

- **Number of samples ( $k$ ):** We set  $k = 4$  in our experiments, which balances computational efficiency with sufficient sample diversity.
- **Foresight length ( $l$ ):** We use  $l = 5$  as the maximum number of future reasoning steps to consider, providing sufficient lookahead while maintaining computational efficiency.

- **Balancing factor ( $\lambda$ ):** Set to 0.3, controlling the weight between visual focus score and reasoning progression score. This value was determined through ablation studies showing optimal performance.
- **Exploration factor ( $\alpha$ ):** Set to 0.3, controlling how strongly the algorithm prioritizes unexplored regions over previously attended regions.
- **Temperature range:** We use  $T_{min} = 0.4$  and  $T_{max} = 1.0$  with an interval of 0.1, resulting in a set of temperature values  $\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$  to ensure diversity in the generated reasoning samples.

## B.2 Temperature Sampling Strategy

To ensure a comprehensive exploration of the reasoning space, we implemented an adaptive temperature sampling strategy:

- A temperature value is randomly selected from the set 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0 each time a sample is generated.
- To prevent repeated selections and promote diversity, the probability weight of each chosen parameter is reduced by half in subsequent sampling processes.
- The weights are reset to their initial values once all parameters have been selected, ensuring a balanced exploration of different temperature values throughout the reasoning process.

## B.3 Relative Attention Extraction

The relative attention mechanism requires extracting attention maps from the VLM. For this purpose, we extract attention weights from the last few layers of the visual encoder and language decoder cross-attention modules. The attention maps are averaged across attention heads and layers to obtain a single attention distribution per text-image pair. These raw attention maps are then normalized according to Algorithm 2.

## B.4 Stopping Criteria

The iterative reasoning process continues until one of the following conditions is met:

- The model outputs the token “REASONING\_COMPLETE” indicating a final answer has been reached
- The reasoning chain reaches a predefined maximum length (typically containing 3-7 reasoning steps, so we set this as 10)
- The reasoning converges (negligible changes in subsequent iterations)

# C Experimental Setup

## C.1 Benchmarks

Our evaluation encompassed multiple complementary benchmarks to assess various aspects of visual reasoning capabilities:

- **Mathematical and Geometric Reasoning:** MathVista [45] and MathVision [46]
- **Multidisciplinary Visual Reasoning:** M3CoT [47] and MMStar [48]
- **Chart Understanding:** Charxiv [14]
- **Geographical Reasoning:** SeekWorld-Global (Google Maps panoramic imagery) and SeekWorld-China (Xiaohongshu App data) from the SeekWorld dataset [16]

## C.2 Compared Methods

Our comparative analysis included the following state-of-the-art methods:

- **Search-based Method:** Monte Carlo Tree Search (MCTS) [49]
- **Foresight Reasoning Method:** Predictive Decoding [43]
- **Visual Search Methodology:** DyFo [37]
- **Multi-modal Chain-of-thought Prompting:** ICoT [41]

### C.3 VLM Models

Our experiments incorporated several state-of-the-art Vision Language Models:

- Qwen2.5-VL-Instruct (7B, 32B) [1]
- InternVL2.5-Instruct (8B) [2]
- Llava-Next (7B) [24]

These models were selected based on their architectural capabilities and demonstrated superior performance in visual reasoning tasks.

### C.4 Computational Infrastructure

All experiments were executed on four NVIDIA A100 GPUs with parallel processing capabilities. For efficiency, sample generation and evaluation were parallelized across the GPUs to accelerate the experimentation process.

### C.5 Performance Metrics

We adopted the following metrics for comprehensive evaluation:

- **Accuracy:** Pass@1 accuracy (Acc.) as the primary performance metric across all benchmarks
- **Computational Efficiency:** Floating point operations (FLOPS) calculated following the methodology in [50], where  $\text{FLOPS} \approx 6nP$  ( $P$  represents model parameters,  $n$  denotes generated tokens)

By computing the average number of tokens generated per example, we provided a standardized measure of computational cost across different methods based on Qwen2.5-VL-7B-Instruct to enable direct efficiency comparisons.

## D Practical Implementation Notes

### D.1 Region Selection Optimization

The sliding window implementation was optimized using stride-based approaches rather than exhaustively evaluating all possible regions. We implemented an efficient algorithm that:

- Uses a stride of 10% when scanning potential regions
- Prioritizes regions with high attention density
- Maintains aspect ratios within the original dimensions for subsequent reasoning iterations

### D.2 Inference Optimizations

To maximize computational efficiency, we implemented several optimizations:

- **Batch Processing:** Sample generation and evaluation were parallelized across multiple GPUs
- **Attention Caching:** Descriptive attention maps were cached to avoid redundant computation
- **Early Stopping:** Inference was halted when “REASONING\_COMPLETE” was detected

This implementation provided a complete blueprint for deploying CoFFT in practice, demonstrating how our approach systematically addresses the challenges of complex visual reasoning through a cognitively-inspired iterative process.