A. Prompt Templates

Below, we provide the prompt templates of RISE-CoT and RISE-R1 stages for both classification (Emotion6) and detection (LISA) tasks, along with example inputs and expected outputs.

A.1. Dataset and Task Overview

The Emotion6[3] dataset is designed for emotion classification, where images are labeled with a probability distribution across six emotion categories: anger, disgust, fear, joy, sadness, and neutral. The classification task requires analyzing visual elements, such as objects, colors, and textures, to assign probabilities to each emotion category, reflecting the image's emotional atmosphere. Figure 1 provides an example image description and its Ground Truth classification annotation to illustrate the task.

The LISA dataset, originally a Visual Question Answering (VQA) dataset[2] with Mask annotations, was adapted for target detection. We converted the original questions into Target Descriptions and transformed the Mask annotations into Bounding Boxes (BBox), creating a Target Detection dataset. The detection task involves identifying and localizing the target object based on its description, focusing on visual cues like shape, color, texture, and context. Figure 2 provides an example image description, the original VQA question with its Mask answer, and the converted Target Description with its corresponding BBox used for RISE-CoT prompt design.

A.1.1. Reasoning Generation Prompts

The Reasoning Generation step prompts the VLM to produce a CoT (R_i) that justifies the annotation (A_i) based on visual and contextual cues, without leaking annotation specifics. The prompts are structured as follows:

- **Classification (Emotion6)**: *Analyze the image to explain how the visual elements and interactions support the emotional probability distribution **$prob_distribution$**, concisely in under 100 words. Describe key objects, actions, and expressions that contribute to the inferred emotional atmosphere, highlighting their connection to the emotional probabilities. Avoid directly stating the emotions or probabilities; instead, focus on the visual cues and their implications.*
- **Detection (LISA)**: *Please describe the reason that justify the region, with its bounding box **$[x1, y1, x2, y2]$**, contains **$target$**. You can focus on shape,color,texture or contextual cues. Note: Do not include specifc bounding box coordinates e.g. x1, y1, x2, y2 or directional terms (e.g., 'top-left') in the description.*

A.1.2. Annotation Reconstruction Prompts

The Annotation Reconstruction step prompts the VLM to predict the annotation ($\hat{A}_i$) based solely on the CoT (R_i) and the image (I_i), verifying the CoT's sufficiency. The prompts are structured as follows:

- **Classification (Emotion6):** *Analyze the provided image description and generate a probability distribution across emotion categories. Categories: ['anger', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'neutral']. Description: **$CoTs$**. Output the final answer in the following format: `<answer>{'anger': prob_0, 'disgust': prob_1, 'fear': prob_2, 'joy': prob_3, 'sadness': prob_4, 'surprise': prob_5, 'neutral': prob_6}</answer>`. Here,* prob_0, prob_1, ..., prob_6 *represent the probabilities for each emotion category, and they should sum to 1. Do not provide any additional explanations or reasoning. Only return the result in the specified format.*
- **Detection (LISA):** *According to the following reasoning description of a region corresponding to **$target$** in the image: **$CoTs$**, output its specific bounding box in the format: `<answer>[x1, y1, x2, y2]</answer>`. Please strictly follow the format.*

These prompts ensure that the CoTs are visually grounded and logically consistent, while the reconstruction step verifies their sufficiency for accurate annotation prediction.

Task	Template	Example
Classification	`prob_distribution`	`{'anger': 0.0, 'disgust': 0.1, 'fear': 0.2, 'joy': 0.389, 'sadness': 0.033, 'surprise': 0.122, 'neutral': 0.156}`
Detection	`[x1,y1, x2,y2]`	`[138, 182, 656, 428]`
Detection	`target`	`The object that could block the sun's glare and protect the eyes`

Table 1. RISE-CoT Template Variables and Examples. Examples are drawn from the first of the three targets in Figure 1 and Figure 2 respectively.

A.2. RISE-R1

The RISE-R1 stage designs prompts to initialize the reasoning process for both classification and detection tasks, emphasizing format compliance and reasoning.

- **Classification (Emotion6):**
  Analyze the visual elements and interactions in the image to explain the emotional atmosphere. Then,

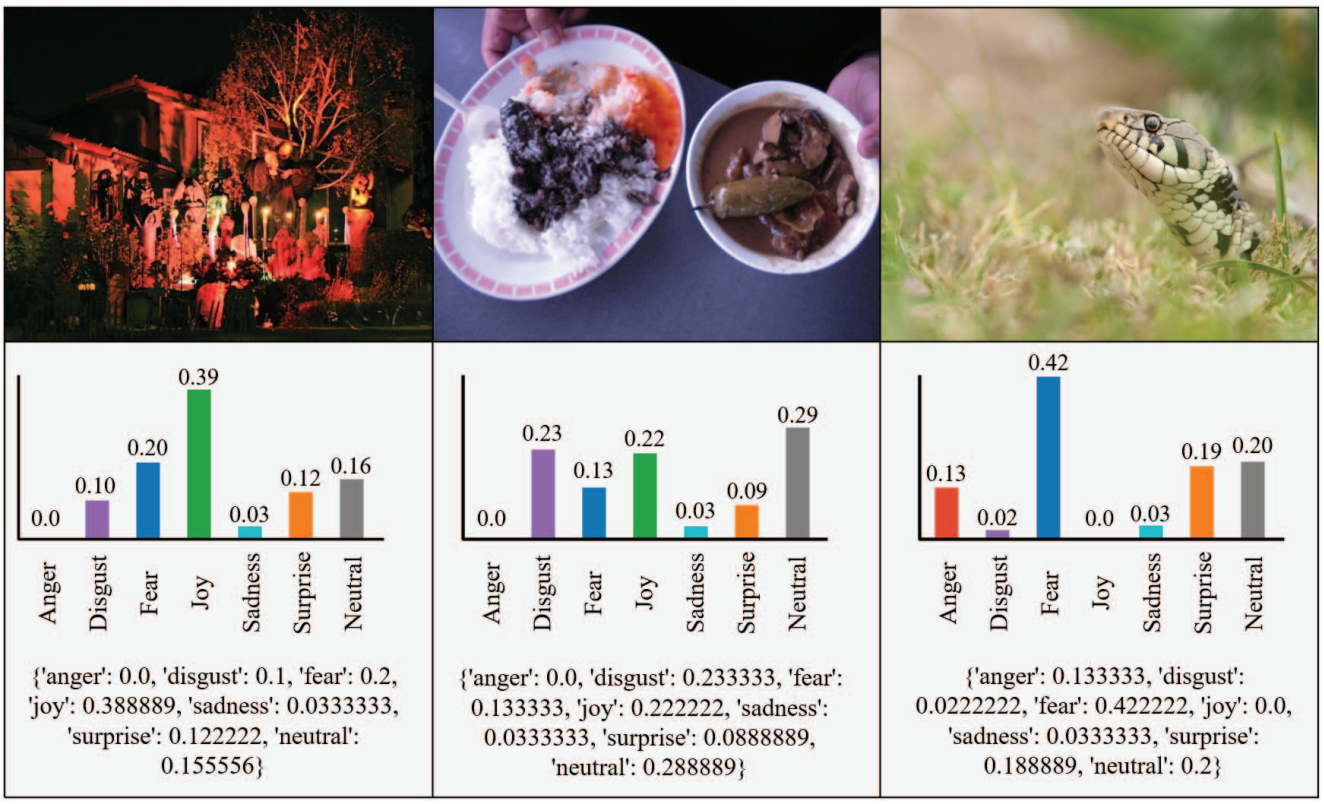


Figure 1: Example image from Emotion6 and Ground Truth classification annotation, showing a probability distribution over six emotion categories.

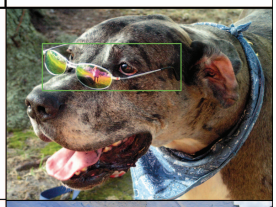
Image	Origin Question (Segmentation)	Mask Answer	Target Description (Detection)	Bounding Box
	From the given descriptions, what object in the picture could be used to block the sun's glare and protect the eyes?		The object in the picture that could be used to block the sun's glare and protect the eyes	
	If we want to enter the building shown in the picture, what object should we look for that serves as an entry point?		The object in the picture that we should look for as the building's entry point	
	When riding a bicycle, it is important to ensure the safety of the head. What object is the person in the picture wearing to protect their head?		The object that the person in the picture is wearing to protect their head when riding a bicycle	

Figure 2: Example image from LISA with its description, original VQA question and Mask answer, and converted Target Description with Bounding Box for target detection.

Task	Template	Example
Emo6	COTs *(Chain-of-Thought: Analyzes image elements like shape, color, texture, or contextual cues to support the given emotion distribution)*	The image depicts a house adorned with elaborate Halloween decorations at night, illuminated by red lights. Key objects include large, life-sized figurines of ghosts, ghouls, and other eerie figures, some of which appear to be holding candles. The decorations are meticulously arranged, creating a festive and somewhat eerie atmosphere. The shadows cast by the figures and the tree branches add depth to the scene, enhancing the sense of mystery and suspense. The overall emotional probability distribution suggests a mix of neutral and slightly negative emotions. The presence of the candles and the eerie figures might contribute to the feeling of unease, aligning with the neutral and slightly negative emotions inferred from the distribution.
LISA	target *(Unique per sample: Comprises the target object's name and its descriptive attributes)*	The object in the picture that could be used to block the sun's glare and protect the eyes
LISA	COTs *(Chain-of-Thought: Identifies the relationship between the target description and its location, forming evidence for selecting the bounding box)*	1. The object in question is a piece of fabric because it is enclosed within the given rectangular coordinates. 2. It is evident from the color and texture that the fabric is a type of object to block the sun's glare and protect the eyes. 3. The object has the appearance of being worn or hung around the neck, which is characteristic of a typical piece of fabric. 4. The surrounding context suggests that it is a dog, as evidenced by the reflection in the glasses and the overall appearance which is consistent with a dog. 5. Therefore, the logical conclusion is that this region corresponds to the target object, which is a piece of fabric used to protect the eyes. Hence, the object that corresponds to this region is a piece of fabric.

Table 2. RISE-CoT Reasoning and Target Examples.The examples are drawn from the first of the three targets in Figure 1 and Figure 2, respectively.

generate a probability distribution across the following emotion categories: {'anger', 'disgust', 'fear', 'joy', 'sadness', 'surprise', 'neutral'}. Please output the final answer in the following format: `<think>...</think><answer>{'anger': prob_0, 'disgust': prob_1, 'fear': prob_2, 'joy': prob_3, 'sadness': prob_4, 'surprise': prob_5, 'neutral': prob_6}</answer>`. *Here, prob_0, prob_1, ..., prob_6 represent the probabilities for each emotion category, and they should sum to 1.*

- **Detection (LISA):**
  Given an image, identify the region corresponding to {target}. Provide a structured output with a reasoning explanation followed by the predicted bounding box coordinates. Format the output as: `<think>Analyze the shape, color, texture, and surrounding objects that help localize the {target} in the image.</think><answer>[x1, y1, x2, y2]</answer>`

B. Reward Distributions and Threshold Analysis

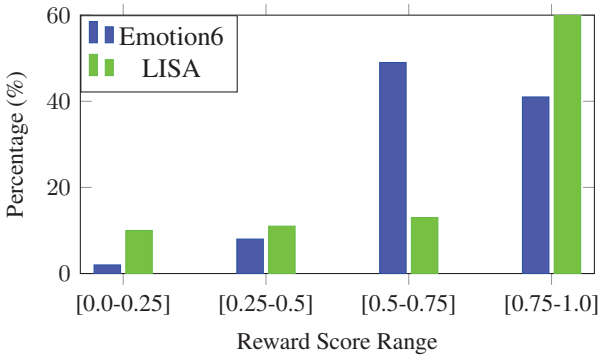


Figure 3: Reward score distribution for Emotion6 and LISA.

The RISE-CoT stage generates Chains of Thought (CoTs) (R^*), reconstructed annotations ($\hat{A}$), and reward scores ($r \in [0.0, 1.0]$) using the reward function $\mathcal{R}(A_i, \hat{A}_i, R_i)$. Figure 3 shows the reward score distribu-

tion for Emotion6 and LISA, with 41% (568/1,386) and 66% (231/350) of samples having $r \geq 0.75$, respectively, indicating that $\tau = 0.75$ captures a significant proportion of high-quality CoTs across tasks of varying reasoning difficulty.

To test robustness, we corrupted 30% of labels in each dataset. For Emotion6, we assigned a new random probability distribution, ensuring the highest-probability category differs from the original. For LISA, we selected a random bounding box region that does not overlap with the original. Figure 4 shows the reward score distribution for corrupted samples, with over 95% of corrupted samples having $r < 0.75$, demonstrating that $\tau = 0.75$ effectively filters noisy data.

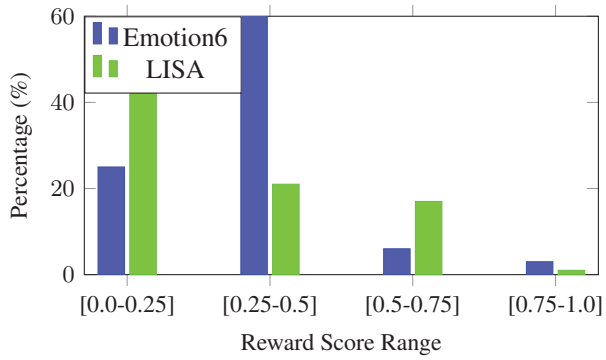


Figure 4: Reward score distribution of **corrupted samples** for Emotion6 and LISA.

Analysis: The choice of $\tau = 0.75$ as the threshold for filtering high-quality CoTs is justified by its ability to retain a substantial portion of accurate samples (41% for Emotion6, 66% for LISA) while excluding over 95% of noisy samples across both datasets. This balance ensures robust performance in diverse tasks. However, the effectiveness of $\tau = 0.75$ is closely tied to the design of the reward function $\mathcal{R}(A_i, \hat{A}_i, R_i)$, which evaluates the alignment between reconstructed and Ground Truth annotations. Changes to the reward function, such as altering the weighting of visual or contextual cues, could shift the optimal threshold, necessitating re-evaluation of τ to maintain filtering efficacy.

C. ImageNet-Sub Dataset Details

For image classification tasks, we use a subset of ImageNet [1], termed ImageNet-Sub, comprising 20 classes: *analog clock, backpack, ballpoint, Band Aid, barbell, barber chair, beer bottle, beer glass, binoculars, bolo tie, bookcase, bottlecap, brassiere, broom, bucket, buckle, candle, can opener, carton, and cellular telephone*. These classes were selected to ensure diversity in object types, covering everyday items (e.g., *cellular telephone*, *carton*), tools (e.g., *can opener*, *broom*), and specialized objects (e.g., *barbell*, *bolo tie*). Overall, these objects rely primarily on visual features such as shape, texture, and color for identification, requiring minimal complex reasoning. This selection balances visual distinctiveness and contextual variability, providing a representative subset for evaluating RISE's image classification performance while maintaining computational efficiency. Each class contains 25 training and 10 testing images, totaling 500 training and 200 testing images, with one-hot probability distribution annotations.

D. Results and Comparisons

D.1. RISE-CoT Intermediate Results

To illustrate the self-supervised learning process in RISE-CoT, we present the average reward $\mathcal{R}(A_i, \hat{A}_i, R_i)$ across training steps for Emotion6 and LISA datasets, as shown in Figure 5. The curves demonstrate the improvement in CoT quality, with rewards converging to higher values as training progresses, reflecting the optimization of visually grounded and logically consistent CoTs.

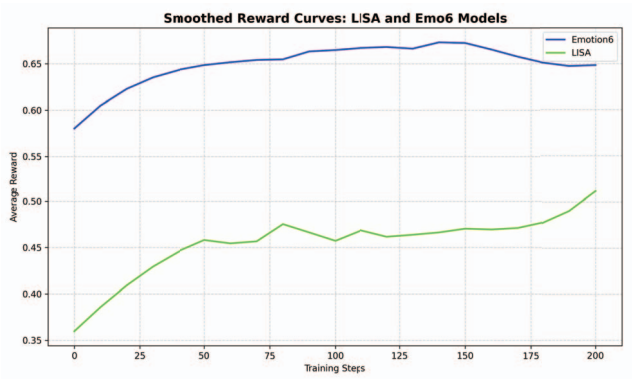


Figure 5: Average reward curves for RISE-CoT training on Emotion6 and LISA, showing improved CoT quality over 200 steps.

To further demonstrate the improvement in CoT quality, we present the CoT evolution for one sample from each dataset at training steps 0, 100, and 200, along with their corresponding reward scores, as shown in Figure 8 and Figure 9.

Discussion: The CoT evolution illustrates how RISE-CoT refines vague initial descriptions into detailed, visually grounded reasoning chains, leading to higher rewards due to improved reconstruction accuracy. The reward increase reflects the model's ability to generate CoTs that better align with the Ground Truth annotations.

D.1.1. Failure Case Analysis

To highlight the limitations of RISE-CoT, we analyze samples with low reward scores, indicating challenges in generating high-quality CoTs. Figure 10 and Figure 11 presents

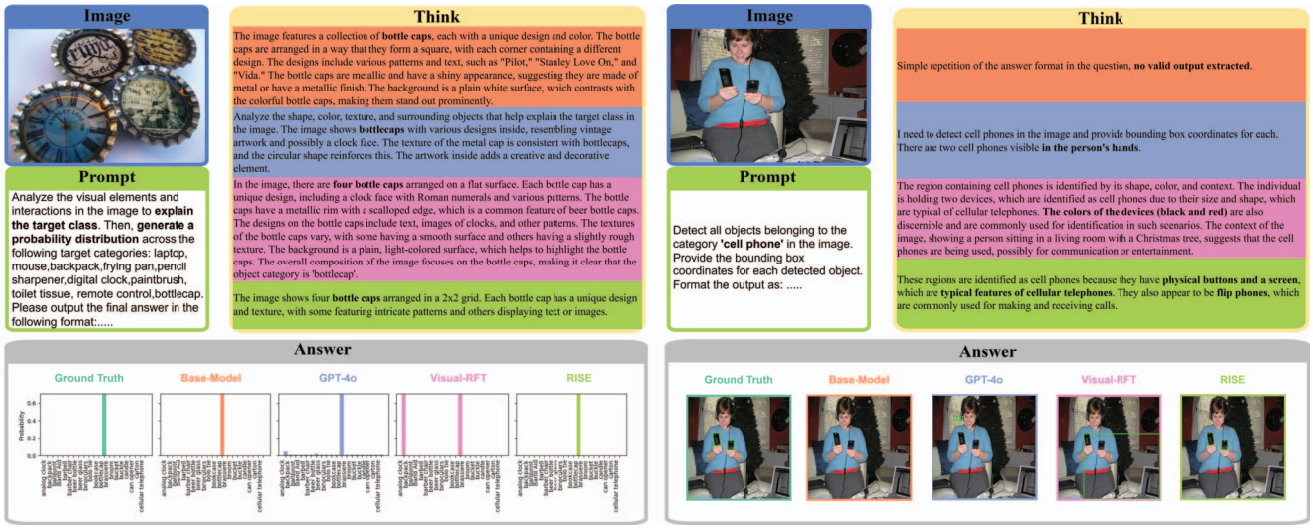


Figure 6: Extended "think-answer" comparisons for RISE-R1 on ImageNet-Sub, and COCO-Sub, showing outputs of RISE, Base-Model, SFT, Visual-RFT, and GPT-4o.

two failure cases from Emotion6 and LISA, including the original images, Ground Truth annotations, reconstructed annotations, and corresponding reward values. These cases typically involve images with low clarity (e.g., blurry visuals or ambiguous objects), which hinder the model's ability to extract distinct visual cues, resulting in less accurate CoTs and lower rewards. This analysis underscores RISE-CoT's dependency on clear visual information for effective reasoning and reconstruction, consistent with the limitations discussed in the paper.

D.2. RISE-R1 Training Process and Results

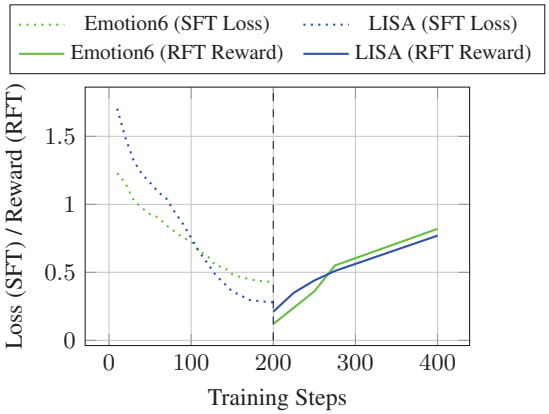


Figure 7: Training curves for RISE-R1 on Emotion6 and LISA, showing SFT loss (0–200 steps, dashed lines) and RFT reward (200–400 steps, solid lines).

The RISE-R1 stage consists of Supervised Fine-Tuning (SFT) on $\mathcal{D}_{\text{RISE}}^{\text{high}}$ (first 200 steps) followed by Reinforcement Fine-Tuning (RFT) on $\mathcal{D}$ (next 200 steps). Figure 7 shows the training dynamics: the SFT phase with the average loss (e.g., cross-entropy loss) for Emotion6 and LISA, and the RFT phase with the average reward $\mathcal{R}_{\text{R1}}(A_i, \hat{A}_i^{R1}, R_i^{R1})$. The decreasing loss in SFT reflects improved annotation accuracy using high-quality CoTs, while the increasing reward in RFT indicates enhanced CoT interpretability and alignment with Ground Truth annotations.

To showcase RISE-R1's final performance, we extend the qualitative comparisons in Figure 3 (main paper) with additional "think-answer" outputs for ImageNet-Sub, and COCO-Sub, comparing RISE with Base-Model, SFT, Visual-RFT, and GPT-4o, as shown in Figure 6. These results highlight RISE-R1's ability to generate detailed, visually grounded CoTs, leading to accurate annotations across diverse tasks.

D.3. Performance Comparison: RISE-R1 2B Model vs. Native Qwen2VL-72B Model

To validate the effectiveness of the RISE framework in **enhancing model reasoning capabilities** and evaluate our **dataset scale selection strategy** for tasks of varying complexity, we conduct a comparative analysis between our RISE-R1 2B parameter model and the largest available 72B parameter model from the same architecture family (Qwen2VL). This experiment aims to demonstrate that our approach not only enables smaller models to acquire reasoning capabilities surpassing their parameter scale, but more importantly, through performance differences across tasks of varying complexity, validates the rationality of our differentiated dataset scale selection strategy for different types of tasks.

Experimental results are shown in Table 3, highlighting

Table 3. Performance comparison between RISE-R1-2B model and Qwen2VL-72B model.

Dataset	Metric	RISE-R1-2B	Qwen2VL-72B
LISA (Reasoning Detection)	mAP	**0.4037**	0.3246
	Precision	**0.5714**	0.5326
	Recall	**0.5714**	0.5000
	F1 Score	**0.5714**	0.5158
Emotion6 (Emotion Analysis)	KL Divergence ↓	**0.3357**	0.7142
	JS Divergence ↓	**0.0712**	3.3287
	Accuracy	**0.9663**	0.9596
	Win-Rate	**10.27**	3.20
COCO (General Detection)	mAP	0.3015	**0.3164**
	Precision	0.5270	**0.6473**
	Recall	**0.5491**	0.4808
	F1 Score	0.5378	**0.5517**
ImageNet-Sub (General Classification)	KL Divergence ↓	5.3761	**0.5183**
	JS Divergence ↓	0.3015	**0.2425**
	Accuracy	0.5450	**0.9200**

performance on tasks requiring deep reasoning.

The experimental results indicate that the performance improvement brought by the RISE framework is closely related to task complexity, which aligns highly with our theoretical expectations.

On complex tasks requiring deep reasoning, RISE-R1-2B demonstrates significant advantages. In the LISA reasoning detection task, our model surpasses the 72B large model across four key metrics:mAP, Precision, Recall, and F1 Score. This advantage is even more pronounced in Emotion6 emotion analysis, where the substantial differences in JS Divergence and Win-Rate fully demonstrate that the Chain-of-Thought (CoT) mechanism can effectively enhance the deep comprehension capabilities of smaller models in complex reasoning tasks.

However, on relatively simple visual tasks, the performance gap between RISE-R1-2B and the large-scale model reflects the applicable boundaries of the CoT mechanism. In the COCO detection task, although our model holds a slight advantage in Recall, it slightly lags behind the 72B model in other metrics. This trend is more evident in the ImageNet-Sub classification task, where Qwen2VL-72B's significant lead in Accuracy indicates that for simple recognition tasks relying primarily on visual feature extraction, excessive reasoning processes may instead introduce unnecessary computational overhead.

This phenomenon validates our core viewpoint regarding the compatibility between CoT and task complexity: complex reasoning tasks can fully benefit from the deep thinking process enabled by CoT, whereas simple recognition tasks only require basic visual feature matching capabilities. Based on this understanding, we adopted a differentiated data strategy in our experimental design—focusing CoT training primarily on complex reasoning tasks, while using only an appropriate amount of data for model warm-up on simple tasks. This approach avoids potential interference of CoT with basic visual abilities while ensuring balanced model performance across multi-task scenarios.

In summary, by precisely matching the CoT mechanism with task characteristics, the RISE framework achieves breakthrough improvements for small models on complex reasoning tasks while maintaining their fundamental capabilities on simple tasks, providing an optimized solution for efficient model deployment in resource-constrained environments.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 4

[2] Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya 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. 1

[3] Kuan-Chuan Peng, Tsuhan Chen, Amir Sadovnik, and Andrew C Gallagher. A mixed bag of emotions: Model, predict, and transfer emotion distributions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 860–868, 2015. 1

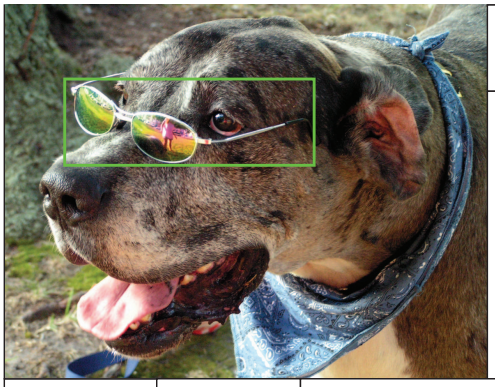
Prob Distribution

{’anger’: 0.0, ’disgust’: 0.1, ’fear’: 0.2, ’joy’: 0.388889, ’sadness’: 0.033333, ’surprise’: 0.122222, ’neutral’: 0.155556}

STEP	Reward	COT
0	0.643	**In the image, we observe a festive and eerie display of decorations surrounding a house at night. The decorations include various statues and figures that suggest themes of horror and mystery. Red lighting illuminates the scene, creating a stark contrast against the dark surroundings. The illuminated statues appear to be looking off into the distance, adding to the eerie ambiance. The presence of tall candles further emphasizes the somber mood. The overall interpretation of emotions based on this detailed description is likely:Fear: Given the presence of menacing-looking statues and the overall unsettling atmosphere created by the vivid lighting.These interpretations are derived from typical associations commonly associated with Halloween decorations, where horror stories and eerie themes are often emphasized.**
100	0.774	**The image depicts a house adorned with Halloween decorations, illuminated with red lights. The house features a large, tree-shaped display made up of numerous creepy and whimsical figures, including skeletons and creatures with intricate facial features. There is a mix of emotions depicted in the scene - the vivid red light casts an intense red hue over the entire structure, which could be perceived as both delightful and unsettling. The figure of an angry or menacing character among the displays could imply a negative emotional response, contrasting sharply with the lightness of some of the decorations. The overall atmosphere is one of mixed emotions, with a strong presence of horror and disapproval due to the aggressive or menacing figures, while the whimsical and colorful decorations provide a counterbalance of mirth and amusement.**
200	0.950	**The image depicts a Halloween house decorated with elaborate Halloween decorations illuminated by lights. The environment is dark with subtle hints of darkness surrounding the illuminated area around the center. Key elements include large, illuminated figures resembling ghosts or monsters, some holding candles. The setting appears festive but somber due to the overall darkness.Surprise: Low – There isn't anything specifically surprising about the image; it's more about the arrangement of objects than any immediate actions. It's more likely a festive activity rather than a situation with personal loss.**

Dimension	CoT 1 (Reward: 0.643)	CoT 2 (Reward: 0.774)	CoT 3 (Reward: 0.950)
Structure & Clarity	❌ Verbose and somewhat disjointed; conclusion feels bolted on at the end	✅ Clear flow of ideas with transitions from observation to emotion	✅ ✅ Very concise, step-by-step reasoning; flows logically and efficiently from setting to emotional interpretation
Visual Detail Usage	✅ Includes specific elements (statues, red lights, candles), but some are repetitive or generic	✅ ✅ Rich visual descriptions (e.g., “tree-shaped display,” “skeletons,” “creatures with facial features”)	✅ ✅ Selectively mentions key features (ghosts/monsters, candles) with a focus on what's emotionally relevant
Emotional Reasoning	❌ Limited to one emotion (“fear”); reasoning is generic and relies on stereotypical associations	✅ Highlights emotional complexity — blends “delightful” with “unsettling,” contrasts menace vs. whimsy	✅ ✅ Refines emotional judgment — interprets mood as “festive but somber,” and evaluates low *surprise* with logical justification
Contextual Awareness	✅ Refers to “Halloween” and common associations with horror	✅ Builds deeper context with contrasts (aggressive vs. whimsical), indicating nuanced understanding of the scene	✅ ✅ Applies functional context ("festive activity," "not personal loss") and links setting to emotion subtly
Conclusion Strength	❌ Weak – overly reliant on assumptions about Halloween rather than grounded evidence	✅ Balanced emotional conclusion grounded in visual cues and thematic contrast	✅ ✅ Strong, precise conclusion — avoids exaggeration, uses “surprise: low” as an analytical statement rather than a guess

Figure 8: CoT evolution for sample images in Emotion6 during RISE-CoT training, showing CoT descriptions and reward scores at steps 0, 100, and 200.



Target Description

The object in the picture that could be used to block the sun's glare and protect the eyes.

STEP	Reward	COT
0	0.406	1. The image shows a region with window features, including rectangular windows and a sloped design typically associated with roof windows.2. The window is in a room, not a building exterior.3. The environment suggests it was probably installed or being prepared for installation, as indicated by the ladder and other materials.4. The windows provide natural light, which is common in windows.Given these observations, the region corresponds to a window because it is rectangular, designed to admit light, and fits within a room setting and context.
100	0.536	1. Type of object: The image shows a window.2. Natural lighting feature: The window is designed to let in natural light, aligning with the region's role in providing illumination inside the room.3. Shape and configuration: The window has a typical rectangular shape and is set into framed beams, showing it is not a skylight.4. Contextual relevance: The region is a window because it has been mentioned as providing natural light and it is likely set in a room given the presence of other room elements.5. Conclusion: Based on these observations, the region corresponds to the target object, a window, due to its shape, lighting purpose, and contextual similarity to a window's appearance.Chain of Thought:1. Identifying the shape typical of windows for natural lighting.2. Identifying the natural-light-giving role of the window.3. Integrating the contextual elements such as other room components and roofline.4.Concluding the observation as a window based on its visual characteristics.
200	0.834	1. Observe the shape and size: The window has a square or rectangular shape and is mounted on the wall, which matches the characteristics of typical window shapes.2. Identify any indications of glass: There are no visible pieces of glass seen in this region, which further supports the identification that it is a window.3. Compare with known objects: Common windows are typically rectangular or square and can be opened or closed to allow light and air into a room.4. Context: The room has walls and flooring that are clean and unpainted, suggesting that this might be a new construction area.Based on these observations, the region corresponding to the target object (a window) is confirmed because of the square or rectangular shape typical of windows, the absence of glass, and the context of the room being under construction or refurbishment.

Dimension	CoT 1 (Reward: 0.240)	CoT 2 (Reward: 0.626)	CoT 3 (Reward: 0.834)
Structure & Clarity	❌ Uses generic labels like "Cognitive Processing"; lacks a tight reasoning chain	✅ Step-by-step structure with clear visual and contextual progression	✅ ✅ Highly structured reasoning steps with logical flow and summary
Visual Feature Analysis	❌ Basic mention of shape and texture; lacks concrete visual grounding	✅ Analyzes shape, color, texture in more detail (e.g., rolled hem, shade of blue)	✅ ✅ Mentions specific visual cues (e.g., reflection in glasses, worn around the neck) that are directly relevant
Contextual Integration	❌ Minimal context use (e.g., says it's for comfort but doesn't support it well)	✅ Considers usage (blocking sun, worn as headband), better aligned with function	✅ ✅ Strong context use (mentions reflection in glasses, dog's appearance), reinforces object identity
Semantic Accuracy	❌ Overstates practicality ("designed for protective use") without evidence	✅ Reasonable assumptions about function (sun glare protection, headband)	✅ ✅ Function is well justified through visual + contextual synthesis
Conclusion Strength	❌ Vague phrasing: "most likely a bandana"	✅ Clearer and supported conclusion	✅ ✅ Strong and confident conclusion, tightly linked to all evidence presented

Figure 9: CoT evolution for sample images in LISA during RISE-CoT training, showing CoT descriptions and reward scores at steps 0, 100, and 200.

<table>
<tr><td>  </td><td>Prob Distribution</td></tr>
<tr><td></td><td>{'anger': 0.0333333, 'disgust': 0.0, 'fear': 0.266667, 'joy': 0.4, 'sadness': 0.0, 'surprise': 0.0333333, 'neutral': 0.266667}</td></tr>
<tr><td>Reward</td><td colspan="2">COT</td></tr>
<tr><td>0.077</td><td colspan="2">Here are the key objects, actions, and expressions observed from the image:1.2. Tree Lines: The line of trees flanking the river adds a sense of stability and calm, contrasting with the dynamic nature of the river.3.4. Cloudy Sky: The overcast sky can contribute to a sense of uncertainty or gloominess, which might contrast with the chaos of the river and reinforce the overall emotions.5. Water Depth: The varying depths of the river suggest movement and change, which can be interpreted as the unpredictable nature of emotions.Based on these observations:- Neutral (Neutral): The combination of serene forest setting, calm river flow, and cloudy sky creates a neutral atmosphere</td></tr>
<tr><td colspan="3">Failure Analysis of the Chain-of-Thought</td></tr>
<tr><td colspan="3">1.Confusing and incomplete structure: The numbering is unclear and some points are missing, disrupting the logical flow.
2.Vague descriptions: Key elements like “Tree Lines” and “Cloudy Sky” lack detailed visual observations.
3.Shallow reasoning: The conclusion of a “neutral” atmosphere is given without strong justification or clear links to the observations.
4.Repetitive language: Similar ideas are repeated without clear progression or deeper insight.
5.Lack of causal connections: Observations are fragmented and not well connected, leading to unsupported conclusions.</td></tr>
</table>

Figure 10: Failure cases in RISE-CoT for Emotion6, showing original image descriptions, Ground Truth annotations, reconstructed annotations, and low reward scores due to unclear visual cues.

Target Description
The object which medical professionals wear during surgery to maintain hygiene and prevent contamination

Reward	COT
0.260	1. The object in question is identified as a slice of bacon due to its characteristic shape and texture.2. The image shows the slice of bacon with a thin, elongated shape and a slightly curly edge, which resembles the common profile of bacon.3. The texture of the slice appears to be crispy and well-flavored, which is typical of bacon.4. The surrounding context, such as the beef steak and vegetables, is also consistent with traditional bacon pairing, where bacon is used to enhance the flavor of the steak.5. These visual cues and contextual information confirm that the object located at is a slice of bacon.Therefore, the final answer is that the region corresponds to a slice of bacon because it meets the criterion of a thin, elongated, crispy, well-flavored, often used meat to enhance the taste of other dishes.

Failure Analysis of the Chain-of-Thought

1.Lack of visual grounding: The CoT repeatedly uses subjective terms like “crispy” and “well-flavored”, which cannot be directly observed in the image. These are speculative descriptions rather than visual evidence.
2.Weak contextual integration: Although it mentions “beef steak and vegetables,” it relies on general food pairings rather than analyzing the actual positioning or interaction of objects in the image.
3.Redundant and repetitive reasoning: Multiple steps restate the same point—that the object is bacon due to its shape and texture—without adding new layers of reasoning, making the logic feel flat and circular.
4.Grammatical and structural issues: The sentence “the object located at is a slice of bacon” contains an obvious placeholder or formatting error, hurting clarity and professionalism.
5.Descriptive definition over analytical reasoning: The conclusion focuses on what bacon is (“a thin, elongated, well-flavored meat used to enhance dishes”) instead of why the object in the image specifically matches those features.

Figure 11: Failure cases in RISE-CoT for LISA, showing original image descriptions, Ground Truth annotations, reconstructed annotations, and low reward scores due to unclear visual cues.