Appendix of Rationale-Enhanced Decoding for Multi-modal Chain-of-Thought

A. Setting of Preliminary Experiments in Section 2.3

All experiments were done with greedy decoding on Gemma-3-12B. We used the test set of GQA for the evaluations.

For computing attention contribution scores, we used the outputs from the 15th transformer block, and averaged the scores across output tokens except for the last `EOS` token. We calculated the attention contribution scores for each token type and displayed the average percentage between test samples in Figure 1.

For intervention analysis with an irrelevant rationale r', we generated r' by the same prompting as CoT/CCoT with the original tuple (x, q). Irrelevant image and query pair (x', q') were uniform randomly selected from the other pairs in the test set of GQA.

B. Additional Experiments

B.1. Ablation Study of Eq. (8)

As alternatives for Eq. (8), we used (i) rationale-conditional probability $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$, (ii) mixture-of-experts $(1-\lambda)p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) + \lambda p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$, and (iii) reversed power-of-experts $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q) \times p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)^\lambda$. We see that our formulation by Eq. (8) achieves the best accuracy. While rationale-conditional probability $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$ is grounded on the rationale r in the next token prediction, it may cause the over-reliance on textual information, as reported in [3]. For mixture-of-experts, it produces the next tokens by the sum of $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ and $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$, which may fail to effectively collaborate image- and rationale-conditional prediction because it prioritizes tokens that have a high probability assigned to one of the distributions, similar to an OR operation. The result of reverse power-of-experts demonstrates the importance of maximizing the rationale-conditional probability as a reward in Eq. (7). Since the effects of the image x in $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ is already dominant as discussed in Section 2.3, predicting by $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q) \times p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)^\lambda$ further weaken the effects of the rationale r, leading a suboptimal performance similar to Baseline and CCoT. In contrast, our formulation of RED can appropriately focus on tokens important for both the image- and rationale-conditional probabilities by the multiplication, like an AND operation. More importantly, Eq. (8) has the theoretical guarantee that maximizes the rationale-conditional probability while maintaining the image-conditional prediction capability, achieving more stable and better accuracy gain for multi-modal CoT prompting.

Table 1. Ablation study varying next token probability formulation in Eq. (8) with CCoT on GQA.

	Formulation	GQA Accuracy (%)
Gemma-3-12B		
Baseline	$p_\theta(y_i\|\boldsymbol{y}_{<i}, x, q)$	45.34
CCoT	$p_\theta(y_i\|\boldsymbol{y}_{<i}, x, r, q)$	44.50 (-0.84)
Rationale-Conditional	$p_\theta(y_i\|\boldsymbol{y}_{<i}, r, q)$	37.87 (-7.47)
VCD + ICD + CCoT	$(1+\lambda)p_\theta(y_i\|\boldsymbol{y}_{<i}, x, r, q) - \lambda p_\theta(y_i\|\boldsymbol{y}_{<i}, x', r, q')$	44.09 (-1.25)
Mixture-of-Experts	$(1-\lambda)p_\theta(y_i\|\boldsymbol{y}_{<i}, x, q) + \lambda p_\theta(y_i\|\boldsymbol{y}_{<i}, r, q)$	44.67 (-0.67)
Rev. Power-of-Experts	$p_\theta(y_i\|\boldsymbol{y}_{<i}, r, q) \times p_\theta(y_i\|\boldsymbol{y}_{<i}, x, q)^\lambda$	44.01 (-1.33)
Power-of-Experts (Eq. (??))	$p_\theta(y_i\|\boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i\|\boldsymbol{y}_{<i}, r, q)^\lambda$	**47.50** (**+2.16**)

B.2. Effects on Object Hallucination

We evaluate RED on two object hallucination benchmarks for LVLMs: MMHal [2] and POPE [1]. Table 2 shows the results with multiple LVLM backbones. While RED is not particularly designed to mitigate object hallucination in LVLMs, it improved CoT/CCoT and achieved competitive performance with the specialized decoding methods for object hallucination, i.e., VCD and VCD + ICD. In contrast to these specialized methods, RED boosts not only anti-hallucination capability but also the quality of the response, as shown in MMHal's average scores. The improvements tend to increase when using high-performance LVLMs (e.g., Qwen2.5-VL-7B), implying the importance of rationale in RED as discussed in Section 4.4.

B.3. Effects of λ

We investigate the sensitivity of RED toward a selection of λ. Figure 1 shows the results on the validation set of GQA with Gemma-3-12B, where we vary $\lambda = \{0.0, 0.1, 0.3, 0.5, 1.0, 10.0\}$ and plot the baselines using image-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ and rationale-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$. We observe that the best value for λ depends on the type of rationale, i.e., text description for CoT and JSON-formatted scene graph. While the performance with larger λ gradually approaches the performance of rationale-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$, RED stably outperform the image-conditional baseline unless choosing extremely large values.

Table 2. Hallucination Benchmark

	MMHal		POPE (All)
	Avg. Score (↑)	Hal. Ratio (↓)	
Gemma-3-4B			
Baseline	2.73	0.61	83.20
VCD	2.73 (+0.00)	0.56 (-0.05)	83.23 (+0.03)
VCD + ICD	2.75 (+0.02)	0.59 (-0.02)	83.27 (+0.07)
CoT	2.23 (-0.50)	0.65 (+0.04)	83.37 (+0.17)
CCoT	2.08 (-0.65)	0.66 (+0.05)	83.33 (+0.13)
CoT + RED	2.73 (+0.00)	0.58 (-0.03)	83.53 (+0.33)
CCoT + RED	2.76 (+0.03)	0.54 (-0.07)	83.57 (+0.37)
Gemma-3-12B			
Baseline	3.10	0.52	84.07
VCD	3.21 (+0.11)	0.52 (-0.00)	84.67 (+0.60)
VCD + ICD	3.29 (+0.19)	0.51 (-0.01)	84.73 (+0.66)
CoT	2.65 (-0.45)	0.52 (+0.00)	84.17 (+0.10)
CCoT	3.01 (-0.09)	0.48 (-0.04)	84.60 (+0.53)
CoT + RED	3.36 (+0.26)	0.46 (-0.06)	84.43 (+0.36)
CCoT + RED	3.23 (+0.13)	0.45 (-0.07)	85.20 (+1.13)

	MMHal		POPE (All)
	Avg. Score (↑)	Hal. Ratio (↓)	
Qwen2.5-VL-7B			
Baseline	3.85	0.36	87.90
VCD	3.86 (+0.01)	0.34 (-0.02)	89.40 (+1.50)
VCD + ICD	3.85 (+0.00)	0.33 (-0.03)	89.90 (+2.00)
CoT	3.84 (-0.01)	0.35 (-0.01)	84.20 (-3.70)
CCoT	3.77 (-0.08)	0.35 (-0.01)	86.33 (-1.57)
CoT + RED	3.99 (+0.14)	0.33 (-0.03)	88.63 (+0.73)
CCoT + RED	3.93 (+0.08)	0.33 (-0.03)	88.47 (+0.57)
Llama3-LLaVA-Next-8B			
Baseline	2.97	0.53	88.23
VCD	3.10 (+0.13)	0.47 (-0.06)	89.50 (+1.27)
VCD + ICD	3.12 (+0.15)	0.46 (-0.07)	89.43 (+1.20)
CoT	3.05 (+0.08)	0.48 (-0.05)	87.00 (-1.23)
CCoT	3.15 (+0.18)	0.49 (-0.04)	86.83 (-1.40)
CoT + RED	3.16 (+0.19)	0.46 (-0.07)	88.86 (+0.63)
CCoT + RED	3.14 (+0.17)	0.47 (-0.06)	88.63 (+0.40)

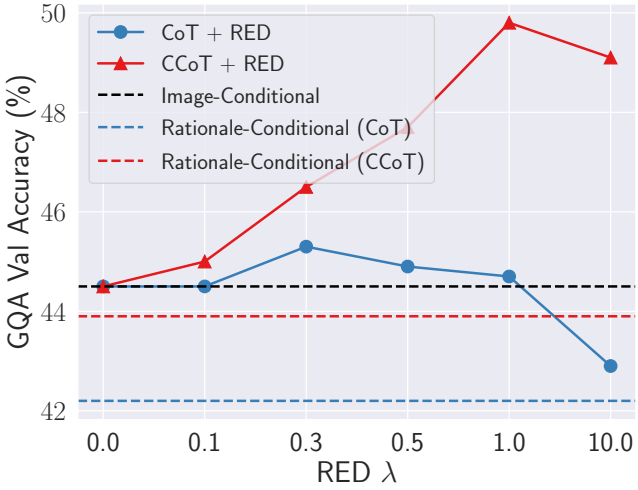

Figure 1. Effects of λ in RED on the GQA Validation Set.

Table 3. Evaluations on MMMU/MMMU-Pro with Qwen2.5-VL-7B

	MMMU		MMMU-Pro	
Model	Accuracy (%)	Speed (ms/token)	Accuracy (%)	Speed (ms/token)
Baseline	50.3	9.5	35.2	12.4
VCD + ICD	48.5	12.1	33.1	14.7
CoT	55.5	17.5	37.2	28.7
CoT + RED (Ours)	**61.6**	14.5	**40.5**	16.2

B.4. Evaluations on MMMU/MMMU-Pro

Table 3 shows the results on the MMMU benchmarks. RED largely outperformed baselines, indicating that it remains effective on more complex multi-modal reasoning tasks.

C. Broader Impacts

This work on rationale-enhanced decoding (RED) has the potential for several societal impacts, both positive and negative.

As a positive perspective, by making LVLMs' reasoning more grounded on rationales, RED can increase the transparency of these models. When users or developers can see that a model's output is a logical consequence of its intermediate reasoning steps (rationales), it can foster greater trust in AI systems. This is particularly important for critical applications where understanding why an AI made a certain decision is essential. Furthermore, when LVLMs are used to answer questions based on documents or to summarize information involving images and text, RED could ensure that the generated answers are more faithfully tied to the source material and the model's interpretation process, reducing the spread of misinformation stemming from model errors.

From a negative perspective, even if RED improves faithfulness, if the generated rationales themselves are flawed (due to biases in training data or limitations in the LVLM's rationale generation capability), users might still be unduly convinced by a seemingly logical but ultimately incorrect or biased output. This is a form of automation bias, where humans over-rely on AI-generated information, especially if it's presented with a "reasoning chain." To overcome these potential issues, it is important to develop robust methods for detecting and mitigating biases in both rationale generation and the final outputs, even when they appear logically derived.

References

[1] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 1

[2] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chun-

yuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In Findings of the Association for Computational Linguistics ACL 2024, pages 13088–13110, 2024. 1

[3] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In Findings of the Association for Computational Linguistics ACL 2024, 2024. 1