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

Shin'ya Yamaguchi
NTT

Kosuke Nishida
NTT

Daiki Chijiwa
NTT

Abstract

Large vision-language models (LVLMs) have demonstrated remarkable capabilities by integrating pre-trained vision encoders with large language models (LLMs). Similar to single-modal LLMs, chain-of-thought (CoT) prompting has been adapted for LVLMs to enhance multi-modal reasoning by generating intermediate rationales based on visual and textual inputs. While CoT is assumed to improve grounding and accuracy in LVLMs, our experiments reveal a key challenge: existing LVLMs often ignore the contents of generated rationales in CoT reasoning. To address this, we reformulate multi-modal CoT reasoning as a KL-constrained reward maximization focused on rationale-conditional log-likelihood. As the optimal solution, we propose rationale-enhanced decoding (RED), a novel plug-and-play inference-time decoding strategy. RED harmonizes visual and rationale information by multiplying distinct image-conditional and rationale-conditional next token distributions. Extensive experiments show that RED consistently and significantly improves reasoning over standard CoT and other decoding methods across multiple benchmarks and LVLMs. Our work offers a practical and effective approach to improve both the faithfulness and accuracy of CoT reasoning in LVLMs, paving the way for more reliable rationale-grounded multi-modal systems.

1. Introduction

Recent large language model (LLM) advancements exhibit impressive complex reasoning capabilities [1, 4, 52]. This progress now extends beyond single-modal text-to-text reasoning. The integration of pre-trained vision encoders, like CLIP [42], with LLMs has led to the development of large vision-language models (LVLMs) [3, 8, 10, 31, 74], enabling more complex multi-modal reasoning.

Chain-of-thought (CoT) prompting is a key factor in LLM reasoning abilities [23, 58]. CoT-prompted models first generate intermediate reasoning steps, termed *rationales*, then incorporate them into the context to produce the final query response. CoT facilitates models to understand queries deeply and promotes logical and coherent responses [44, 54, 57]. This principle has also been applied to multi-modal reasoning in LVLMs [16, 33, 36, 37, 69–72]. In CoT for LVLMs, models generate intermediate rationales from the input image and the text query. These rationales, with the original inputs, are then used for grounded image-text-to-text reasoning. Recent research also explores generating structured rationales (e.g., scene graphs) to enhance LVLMs' capabilities like spatial reasoning [36, 70]. Thus, CoT in LVLMs is widely assumed to be beneficial, offering multi-modally grounded understanding for more accurate responses.

However, our empirical results show that this assumption does not always hold. Our preliminary experiments, includ-

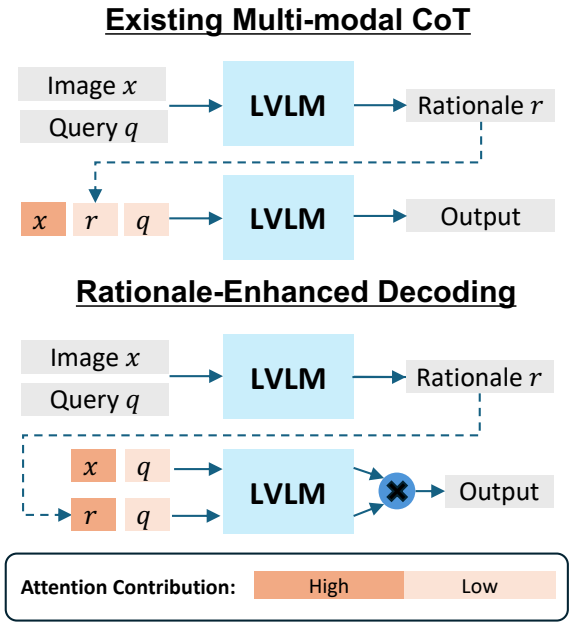

Figure 1. **Rationale-Enhanced Decoding (RED).** Existing multi-modal chain-of-thought (CoT) prompting by large vision language models (LVLMs) is a two-step generation of rationale and final output. It often focuses on input images and overlooks intermediate rationales in the final output generation. Our rationale-enhanced decoding (RED) addresses this issue by decoupling the image and rationale in decoding, and combining them at the logit level to provably ensure grounding outputs on the rationale.

ing (i) measuring input token contributions and (ii) swapping rationales, reveal a critical CoT limitation: existing LVLMs often ignore the contents of generated rationales. Indeed, CoT reasoning with rationales can even degrade performance compared to direct answering without CoT. Moreover, replacing a rationale with an irrelevant one often does not change model performance, implying LVLMs largely ignore rationale semantics in such cases. These findings suggest that the current CoT mechanism in LVLMs does not effectively ground the final prediction on the information captured by the intermediate rationale.

A straightforward solution to make models force rationale grounding would be to fine-tune LVLMs on datasets containing image-query-rationale-answer tuples, as in [69, 71]. However, this approach demands expensive annotated datasets and additional training costs. Therefore, we focus on developing a plug-and-play decoding strategy for pre-trained LVLMs. We address the following primary research question: *Can we enhance the LVLMs' grounding capability on the rationales in CoT and improve performance by solely modifying the decoding strategy without additional training?*

In this paper, we propose rationale-enhanced decoding (RED), a novel decoding strategy designed to harmonize information from visual and rationale tokens in LVLMs without additional training. Our core idea is decoupling the next token probability into distinct image-conditional $p(y|x,q)$ and rationale-conditional $p(y|r,q)$ and separately enhancing rationale grounding (x: image, r: rationale, q: query, and y: output). Based on this idea, we re-formulate multi-modal CoT prompting as a KL-constrained reward maximization [45, 46]. This aims to maximize policy regarding the rationale-conditional log-likelihood $\log p(y|r,q)$ as the reward while staying close to the image-conditional likelihood $p(y|x,q)$. Solving this problem makes LVLMs explicitly ground on both image and rationale information in the next token prediction. Without additional training, RED provably yields the optimal solution of this maximization by composing the next-token distribution as $p(y|x,q) \times p(y|r,q)^{\lambda}$. Practically, RED is implemented as a simple weighted sum of the log-softmax logits of $p(y|x,q)$ and $p(y|r,q)$, allowing easy, training-free, and plug-and-play integration with existing LVLMs without architecture modification.

We conduct extensive experiments comparing RED against standard CoT prompting and plug-and-play decoding methods for off-the-shelf LVLMs across multiple benchmark datasets and backbone LVLMs. Our results demonstrate that RED consistently and significantly improves reasoning performance. Furthermore, RED's advantages are amplified with high-quality rationales (e.g., intervening with GPT-4[1]). Our findings validate the effectiveness and practicality of RED for enhancing CoT reasoning faithfulness and accuracy in LVLMs, opening avenues for more reliable, interpretable, rationale-grounded multi-modal systems.

2. Related Work

Large Vision Language Models (LVLMs). Large vision language models (LVLMs) integrate visual encoders with LLMs, often via alignment training to represent images in the LLM input space [8, 10, 31, 74]. Despite remarkable multi-modal capabilities [25, 38, 40], LVLMs face challenges like poor visual recognition [67], object hallucination [14, 30, 60], and misalignment between image and text tokens [6, 55, 64]. Furthermore, as our preliminary experiments (Section 3.3) highlight, LVLMs also struggle to effectively leverage rationales in CoT reasoning. Previous solutions include preference/reward tuning [51, 73], improved visual instruction tuning [7, 9, 39, 47, 49], auxiliary model-enhanced decoding [50, 62, 65], plug-and-play decoding [12, 26, 35, 56], and CoT prompting [16, 36, 37, 69, 71, 72]. Our RED is categorized into a plug-and-play decoding for improving CoT prompting. While many CoT prompting methods focus on improving rationale generation via specific prompting or auxiliary model training (see Section 3.2), they often still rely on standard decoding with $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ that may not leverage rationales. Our work diverges with a novel plug-and-play decoding strategy designed to enhance rationale utilization in CoT reasoning, without additional training or auxiliary models.

Plug-and-play Decoding Strategies for LVLMs. Plug-and-play decoding strategies [12, 21, 26, 35, 56] are relevant as they operate at inference time without additional training like RED. These decoding strategies are mainly focused on mitigating object hallucination by contrastive decoding [29]. For instance, LCD [35] contrasts $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ with $p_\theta(y_i|\boldsymbol{y}_{<i}, q)$ to mitigate the language prior effects; VCD [26] subtracts hallucinated predictions by contrasting $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ with $p_\theta(y_i|\boldsymbol{y}_{<i}, x', q)$ (x' is the corrupted image). Thus, these decoding strategies basically aim to modulate image-conditional probability $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ to mitigate object hallucination. However, these methods do not ensure LVLMs faithfully use CoT rationales; they refine image-conditional output, not harmonize it for CoT reasoning. Therefore, RED is complementary to these decoding strategies because it grounds predictions by multiplying distinct image-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$, which is potentially pre-modulated by other methods, and rationale-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$. This allows combining RED with other plug-and-play methods for synergistic benefits of better hallucination mitigation and rationale grounding.

3. Preliminaries

This section introduces LVLM and CoT prompting principles, then presents preliminary experiments highlighting existing LVLMs' suboptimal rationale utilization in multi-modal CoT.

3.1. Next-token Prediction in LVLMs

Consider an auto-regressive LVLM, parameterized by θ, which is trained on large-scale image-text datasets to process images as input for its backbone LLM. Given an input image x and query q, an LVLM generates an output token sequence $\boldsymbol{y} = (y_1, \ldots, y_L) \in \mathcal{V}^L$ following:

$$p(\boldsymbol{y}|x, q) = \prod_{i=1}^{L} p_\theta(y_i|\boldsymbol{y}_{<i}, x, q), \tag{1}$$

where L is token length, $\mathcal{V}$ is token vocabulary, and $\boldsymbol{y}_{<i}$ are preceding output tokens. As in auto-regressive LLMs, $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ over $\mathcal{V}$ is the softmax of the model's output $\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$:

$$p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) = \text{softmax}(\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, x, q)) \tag{2}$$

$$= \frac{\exp(\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, x, q))}{\sum_{w\in\mathcal{V}} \exp(\text{logits}_\theta(y_i = w|\boldsymbol{y}_{<i}, x, q))}. \tag{3}$$

Each token y_i is generated from $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ via a decoding strategy such as greedy decoding, i.e., $y_i = \arg\max_{w\in\mathcal{V}} p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$. We define $\text{generate}(\cdot)$ as a utility function for an arbitrary decoding strategy:

$$\boldsymbol{y} = \text{generate}_\theta(x, q). \tag{4}$$

3.2. Multi-modal Chain-of-Thought Prompting

Inspired by single modal CoT prompting [58], multi-modal CoT aims to enhance LVLM reasoning by incorporating the intermediate rationale generation during decoding [36, 37, 69, 71]. In general, multi-modal CoT involves two reasoning steps: (i) rationale generation and (ii) output generation. First, LVLMs generate a rationale r from input image x, query prompt q with instruction prompt, e.g., "`Given the image, generate the rationale for answering the question`", as follows.

$$r = \text{generate}_\theta(x, q). \tag{5}$$

By using r, LVLMs generate the final output by the next-token prediction, similar to Eq. (4) :

$$\boldsymbol{y} = \text{generate}_\theta(x, r, q). \tag{6}$$

In this line of work, MM-CoT [69] and UnifiedQA [33] are the pioneering works introducing the concept of CoT prompting into the multi-modal reasoning of LVLMs. Successor works focus on improving the quality of r via auxiliary VQA models [71] and knowledge base retrieval [37]. Although these works successfully elicited LVLM reasoning, their reliance on additional training and/or auxiliary resources (e.g., VQA models and knowledge bases) limits broader applicability. To address this limitation, CCoT [36] enhanced rationale quality via structured scene graphs in JSON format generated from LVLMs' zero-shot reasoning without additional training. Most existing works focus on improving r in Eq. (5) and assume that accurate r ensures better next token prediction with $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$. In this regard, our work is orthogonal to them because we investigate the reliability of $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ as discussed in the next section. This work focuses on training-free CoT reasoning for maintaining simplicity and generalization of pre-trained LVLMs, but our findings can extend to any CoT methods in a plug-and-play manner.

3.3. Motivating Experiments

We show our motivation through preliminary experiments asking a simple question: *Do LVLMs perform CoT reasoning grounded on intermediate rationales?* Specifically, we assess how much r and its contents contribute to the output sequence decoded by $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ through two preliminary experiments: (i) measuring the contribution scores to output token predictions for input token groups corresponding to the image x, rationale r, and query q, and (ii) replacing rationale r with r' from another (x', q') pair. Experiment (i) aims to evaluate how much r contributes to output $\boldsymbol{y}$, and (ii) checks if LVLMs leverage the content of r associated with (x, q). We used Gemma-3-4B/12B [53] as the LVLM and GQA [18] as the evaluation dataset. For the rationale generation, we queried LVLMs with Eq. (5) to generate text description rationales and scene graphs with the prompts of CCoT [36]; hereinafter, we refer to the reasoning with text descriptions as CoT and that with scene graphs as CCoT. See Appendix for details.

Weakened influence of rationales. Here, we measure attention contribution scores [5, 20, 22] to quantify input token (x, r, q) contributions to output $\boldsymbol{y}$. Ideally, CoT prediction should derive substantial contributions from both image and rationale tokens. The contribution score from i-th token to j-th token at the l-th layer and h-th head is computed by $\|\alpha_{i,j}^{l,h} \boldsymbol{z}_j^{l-1} \boldsymbol{W}_{OV}^{l,h}\|$, where $\alpha_{i,j}^{l,h}$ is attention score, $\boldsymbol{z}_j^{l-1}$ is the output of the previous layer, and $\boldsymbol{W}_{OV}^{l,h}$ is the output projection matrix in transformer-based LVLMs. We describe more details of the evaluation protocol in Appendix. Figure 2 shows the attention contributions in the middle layer of Gemma-3-12B for each case of input for decoding, where the scores are computed for each token corresponding to x, r, and q, respectively, and displayed as percentages of the overall contribution score for each input type [1]. While image and rationales contribute largely when conditioned individually (i.e., (x, q), (r_{CoT}, q), and (r_{CCoT}, q)), in multi-modal CoT

[1]We analyzed a middle layer because prior work shows it is where LVLMs primarily perform multi-modal fusion to integrate vision and language [38]. This allows us to observe the model's attention balance between the image and rationale during this crucial integration process, before the representation becomes overly specialized for the final output task.

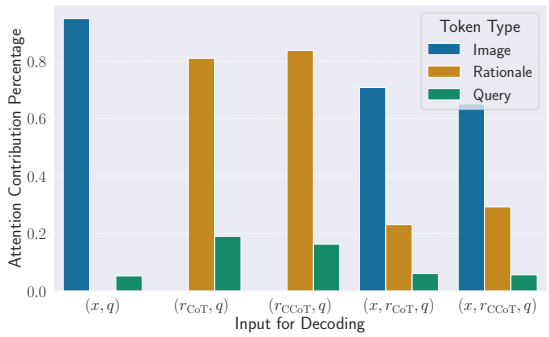

Figure 2. **Percentage of attention contributions by input token types for different decoding strategies (Gemma-3-12B)**. Rationale tokens contribute less to outputs than image tokens.

Input for Decoding	Gemma-3-4B	Gemma-3-12B
(x, q)	40.00	45.34
(x, r_{CoT}, q)	41.08 (+1.08)	41.76 (-3.58)
(x, r'_{CoT}, q)	41.88 (+1.88)	41.75 (-3.59)
(x, r_{CCoT}, q)	44.54 (+4.54)	44.50 (-0.84)
(x, r'_{CCoT}, q)	44.35 (+4.35)	44.30 (-1.04)
(r_{CoT}, q)	40.15 (+0.15)	37.87 (-7.47)
(r'_{CoT}, q)	7.40 (-32.60)	16.21 (-29.13)
(r_{CCoT}, q)	43.35 (+3.35)	44.23 (-1.11)
(r'_{CCoT}, q)	19.05 (-20.95)	10.25 (-35.09)

Table 1. **Test accuracy (%) on GQA**. Contrary to intuition, irrelevant rationale r' maintains the model performance in decoding with (x, r', q), indicating the ignorance of rationales in LVLMs.

(i.e., (x, r_{CoT}, q), and (x, r_{CCoT}, q)), image tokens dominate and rationale contribution is remarkably reduced.

Lack of grounding on rationale contents. Next, we analyze rationale content effects by replacing original rationale r with an irrelevant rationale r' generated from another random (x', q') pair. We expect that intervening by r' will largely degrade the performance. Table 1 shows the test accuracy on GQA for the combinations of decoding strategy and input for decoding. While CoT and CCoT with (x, r, q) improved the baseline performance with Gemma-3-4B, they degraded the performance with Gemma-3-12B. More importantly, in Table 1, CoT/CCoT with the irrelevant rationale r' via (x, r', q) performed similarity to using original r, indicating that LVLMs ignore the contents of r and the performance gain comes from other causes. Conversely, decoding with (r', q) largely dropped the scores from that with (r, q), implying that LVLMs use r when x is absent. This phenomenon could potentially be attributed to several factors, including position bias in the LLM backbones [66, 68], attention sink [13, 20, 61], and/or overfitting to visual instruction tuning [15, 30]. Since our focus is training-free CoT in off-the-shelf LVLMs, we leave a deeper analysis for future work.

In summary, the two experiments demonstrate that LVLMs are less likely to leverage rationale r mechanistically and semantically in the multi-modal CoT prompting. This challenge is important because CoT prompting is generally expected to enhance not only task performance but also faithfulness and interpretability [19, 24, 41]. This motivates our new decoding strategy to ground the inference on r.

4. Method

To address the previously identified challenges, we propose rationale-enhanced decoding (RED). Our approach begins by re-formulating multi-modal CoT prompting as a KL-constrained reward maximization focused on rationale-conditional probability. RED solves this without additional training by multiplying distinct image- and rationale-conditional token probabilities for predicting the next token. We prove that this formulation is equivalent to the optimal solution of the KL-constrained reward maximization, demonstrating its theoretically sound background. RED is implemented by a weighted sum of the log-softmax logits, allowing for its plug-and-play adaptation to any LVLMs.

4.1. Multi-modal CoT as a Reward Maximization on Rationale

As shown in Section 3.3, conventional multi-modal CoT prompting via $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ often fails to effectively leverage rationale r for output predictions. We also observed that the prediction with $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$ relies on r, but the performance is suboptimal. We aim for a next token distribution more effective than both $p(y_i|\boldsymbol{y}_{<i}, x, r, q)$ and $p(y_i|\boldsymbol{y}_{<i}, r, q)$. To this end, our core idea is decoupling the next token probability into distinct image-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ and rationale-conditional $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$ and separately enhancing rationale grounding. Enhancing rationale grounding indeed corresponds to maximizing the log-likelihood of $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$; we refer to this log-likelihood as *rationale-grounding reward*. Therefore, instead of $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$, we introduce a new next token distribution π maximizing the rationale-grounding reward based on a KL-constrained reward maximization [45, 46] as follows:

$$\max_{\pi} \mathbb{E}_\pi[R] - \beta \mathbb{D}_{\text{KL}}[\pi||\pi_{\text{ref}}], \tag{7}$$

where $R = \log p_\theta(y_i \sim \pi|\boldsymbol{y}_{<i}, r, q)$, β is a hyperparameter to balance the KL penalty term, and $\pi_{\text{ref}} = p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$. Intuitively, this maximizes expected rationale-grounding reward $\mathbb{E}_{y_i \sim \pi}[\log p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)]$, which strongly relies on r as shown in Section 3.3. The KL-constraint between π and $p(y_i|\boldsymbol{y}_{<i}, x, q)$ incorporates visual information from x. This formulation naturally makes the next token prediction ground on both x and r without using the problematic conditional probability $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$.

4.2. Rationale-Enhanced Decoding (RED)

We derive rationale-enhanced decoding (RED) to maximize Eq. (7) without additional training. RED's next token distri-

bution $\hat{p}_\theta(y_i)$ is formed by multiplying distinct distributions as follows:

$$\hat{p}_\theta(y_i) := \frac{1}{Z_\theta} p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)^\lambda, \quad (8)$$

where $Z_\theta = \sum_{w \in \mathcal{V}} p_\theta(y_i = w|\boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i = w|\boldsymbol{y}_{<i}, r, q)^\lambda$ is the normalization constant, and λ is a hyper-parameter for modulating the influence of $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$. Intuitively, Eq. (8) is a power-of-experts [17] emphasizing overlap between image- and rationale conditional probabilities. Indeed, Eq. (8) is the closed-form solution maximizing Eq. (7) as follows.

Theorem 4.1. *Let the reference policy π_{ref} be $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$, and the reward function $R(\cdot)$ be $\log p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$. Sampling by Eq. (8) is equivalent to sampling from the optimal policy π^* for Eq. (7).*

Proof. Consider the KL-constrained reward maximization objective [43, 45, 46]:

$$\max_\pi \mathbb{E}_\pi[R(s,a)] - \beta \mathbb{D}_{\text{KL}}[\pi(a|s)||\pi_{\text{ref}}(a|s)], \quad (9)$$

where R is a reward function, s is a state (input context), and a is an action (output). Obviously, Eq. (7) is a special case of Eq. (9). From Appendix A.1 of [43], the optimal policy $\pi^*(a|s)$ for this objective is given by

$$\pi^*(a|s) = \frac{1}{Z(s)} \pi_{\text{ref}}(a|s) \exp\left(\frac{1}{\beta} R(s,a)\right), \quad (10)$$

where $Z(s) = \sum_{a'} \pi_{\text{ref}}(a'|s) \exp(\frac{1}{\beta} R(s,a'))$ is the partition function. In the maximization of Eq. (7), given $a = y_i$ and $s = (\boldsymbol{y}_{i<}, x, r, q)$, we can set each component in Eq. (9) as $\pi_{\text{ref}}(a|s) = p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$, $R(s,a) = \log p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$, and $\beta = 1/\lambda$. Substituting them into the optimal policy of Eq. (10) yields

$$\begin{aligned}\pi^*(a|s) &= \frac{1}{Z_\theta} p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) \times \exp\left(\lambda \log p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)\right) \\ &= \frac{1}{Z_\theta} p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)^\lambda = \text{Eq. (8)}.\end{aligned}$$

Therefore, sampling by Eq. (8) is indeed equivalent to sampling from the optimal policy $\pi^*(a|s)$ for the KL-constrained reward maximization in Eq. (7). □

Theorem 4.1 shows that sampling by Eq. (8) yields the optimal distribution maximizing $p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$ at decoding time without any auxiliary reward models. This theoretical property supports the reliability of RED as a method for improving LVLMs in the multi-modal CoT. We also show a comparison to other possible alternatives to validate the efficacy of Eq. (8) in Appendix.

Algorithm 1 Rationale-Enhanced Decoding (RED)

Require: Input image x, query q, LVLM parameterized by θ, hyper-parameter λ
Ensure: Decoded output sequence $\boldsymbol{y}$

```
r ← generate_θ(x, q)   # Generate arbitrary rationales
y ← ∅
for |y| < L do
    logîts_θ(y) ← log softmax(logits_θ(y|y, x, q)) + λ log softmax(logits_θ(y|y, r, q))
    y ∼ softmax(logîts_θ(y))
    y .append(y)   # Add last token
end for
```

4.3. Algorithm

Practically, we generate the next token by combining $\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ and $\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, r, q)$:

$$\hat{p}_\theta(y_i) = \text{softmax}(\hat{\text{logits}}_\theta(y_i)), \quad (11)$$

$$\begin{aligned}\hat{\text{logits}}_\theta(y_i) := \log \text{softmax}(\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, x, q)) \\ + \lambda \log \text{softmax}(\text{logits}_\theta(y_i|\boldsymbol{y}_{<i}, r, q)). \quad (12)\end{aligned}$$

Eqs. (11) and (12) are derived from Eq. (8) by

$$\begin{aligned}&\hat{p}_\theta(y_i) = \exp \log(\hat{p}_\theta(y_i)) \\ &= \exp(\log(p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)) + \log(p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)^\lambda) - \log Z_\theta) \\ &\propto \text{softmax}(\log(p_\theta(y_i|\boldsymbol{y}_{<i}, x, q) \times p_\theta(y_i|\boldsymbol{y}_{<i}, r, q)^\lambda)). \quad (13)\end{aligned}$$

Thus, we can implement RED by computing the weighted sum of the log-softmax logits as the new logits for the next token prediction. To avoid the latency overhead, we simultaneously compute $\text{logits}_\theta(y|\boldsymbol{y}, x, q)$ and $\text{logits}_\theta(y|\boldsymbol{y}, r, q)$ by batch parallel inference. We show the overall procedures of RED in Algorithm 1.

5. Experiments

We validate the efficacy of RED via comprehensive evaluation on multiple multi-modal reasoning datasets with pre-trained LVLMs, comparing to existing multi-modal CoT methods and plug-and-play decoding baselines. We also analyze interventions on the rationale quality.

5.1. Settings

Baselines. The baselines include: Baseline (standard inference with $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$), CoT [58, 69] (prompting to generate text rationales), and CCoT [36] (a state-of-the-art training-free method generating JSON-formatted scene graph rationales). Other plug-and-play decoding baselines are: VCD [26], contrasting $p_\theta(y_i|\boldsymbol{y}_{<i}, x, q)$ with $p_\theta(y_i|\boldsymbol{y}_{<i}, x', q)$, where x' is a corrupted input image by adding Gaussian noise, and VCD + ICD [56], improving VCD using a variant query q' modified by adversarial

Table 2. **Performance comparison on general visual reasoning benchmarks across various LVLMs**. RED is applied to both CoT and CCoT. Values in parentheses indicate the relative delta from Baseline (direct decoding without CoT). Best scores for each LVLM-benchmark pair are bolded.

	GQA	LLaVA-Bench	MME		MMVet	SEED-I	TextVQA	MathVista
			Perception	Cognition				
Gemma-3-4B								
Baseline	40.00	73.20	1211.34	370.00	44.00	65.39	63.95	41.40
VCD	38.74 (-1.26)	75.70 (+2.50)	1184.11 (-27.23)	338.57 (-31.43)	44.80 (+0.80)	65.19 (-0.20)	64.02 (+0.07)	39.30 (-2.10)
VCD + ICD	38.62 (-1.33)	73.30 (+0.10)	1182.29 (-29.05)	340.71 (-29.29)	46.70 (+2.70)	65.20 (-0.19)	64.03 (+0.08)	40.40 (-1.00)
CoT	41.08 (+1.08)	72.90 (-0.30)	1254.77 (+43.43)	341.07 (-28.93)	46.20 (+2.20)	65.90 (+0.51)	60.62 (-3.33)	40.10 (-1.30)
CCoT	44.54 (+4.54)	73.60 (+0.40)	1294.46 (+83.12)	396.43 (+26.43)	46.50 (+2.50)	66.63 (+1.24)	63.23 (-0.72)	41.10 (-0.30)
CoT + RED	42.19 (+2.19)	76.30 (+3.10)	1325.77 (+114.43)	**645.35 (+275.35)**	46.60 (+2.60)	**66.89 (+1.50)**	65.13 (+1.18)	**43.50 (+2.10)**
CCoT + RED	**45.87 (+5.87)**	**76.90 (+3.70)**	**1330.07 (+118.73)**	611.43 (+241.43)	**49.10 (+5.10)**	66.73 (+1.34)	**65.54 (+1.59)**	42.00 (+0.60)
Gemma-3-12B								
Baseline	45.34	79.00	1171.37	545.71	57.70	71.01	69.81	52.10
VCD	44.11 (-1.23)	78.40 (-0.60)	1152.02 (-19.34)	543.93 (-1.79)	58.30 (+0.60)	70.96 (-0.05)	69.40 (-0.51)	51.50 (-0.60)
VCD + ICD	44.16 (-1.18)	79.70 (+0.70)	1153.86 (-17.50)	651.79 (+106.07)	57.40 (-0.30)	71.01 (+0.00)	69.50 (-0.41)	51.80 (-0.30)
CoT	41.76 (-3.58)	78.90 (-0.10)	1507.67 (+336.30)	661.07 (+115.36)	57.50 (-0.20)	70.75 (-0.26)	66.23 (-3.58)	53.50 (+1.40)
CCoT	44.50 (-0.84)	79.00 (+0.00)	1289.67 (+118.30)	604.29 (+58.58)	53.00 (-4.70)	71.69 (+0.68)	69.70 (-0.11)	51.20 (-0.90)
CoT + RED	46.07 (+0.73)	**81.00 (+2.00)**	**1574.52 (+403.15)**	**695.36 (+149.65)**	**59.50 (+1.80)**	72.15 (+1.14)	70.73 (+0.92)	**54.80 (+2.70)**
CCoT + RED	**47.50 (+2.16)**	80.60 (+1.60)	1359.28 (+187.91)	642.50 (+96.79)	58.40 (+0.70)	**72.76 (+1.75)**	**71.09 (+1.28)**	53.50 (+1.40)
Qwen2.5-VL-7B								
Baseline	60.88	82.10	1665.22	621.07	56.70	58.13	77.76	64.70
VCD	59.34 (-1.54)	82.20 (+0.10)	1650.80 (-34.42)	631.07 (+10.00)	58.30 (-0.40)	60.32 (+2.19)	77.53 (-0.23)	64.50 (-0.20)
VCD + ICD	59.40 (-1.48)	82.40 (+0.30)	1584.17 (-101.04)	651.79 (+30.71)	57.50 (-1.20)	60.71 (+2.58)	75.45 (-2.31)	64.80(+0.10)
CoT	46.70 (-14.18)	81.80 (-0.30)	1555.52 (-109.70)	705.00 (+83.93)	56.80 (+0.10)	60.49 (+2.36)	70.32 (-7.44)	61.30 (-3.40)
CCoT	46.69 (-14.19)	81.00 (-1.10)	1559.71 (-105.51)	634.29 (+13.22)	52.90 (-3.80)	73.18 (+15.05)	74.60 (-3.16)	61.30 (-3.40)
CoT + RED	61.06 (+0.18)	82.20 (+0.10)	**1706.25 (+41.03)**	**706.79 (+86.72)**	**60.70 (+4.00)**	76.50 (+18.37)	77.98 (+0.22)	**70.60 (+5.90)**
CCoT + RED	**61.92 (+1.04)**	**84.60 (+2.50)**	1704.83 (+39.61)	648.21 (+27.14)	56.80 (+0.10)	**78.37 (+20.24)**	**78.61 (+0.85)**	68.10 (+3.40)
Llama3-LLaVA-Next-8B								
Baseline	65.22	65.60	**1583.67**	332.14	37.70	72.57	65.01	37.70
VCD	63.60 (-1.62)	67.60 (+2.00)	1512.10 (-71.57)	341.79 (+9.64)	40.50 (+2.80)	71.93 (-0.64)	63.51 (-1.50)	35.90 (-1.80)
VCD + ICD	63.88 (-1.34)	70.50 (+4.90)	1485.12 (-98.55)	317.50 (+14.64)	40.30 (+2.60)	71.71 (-0.86)	62.97 (-2.03)	36.20 (-1.50)
CoT	61.82 (-3.40)	65.60 (+0.00)	1462.63 (-121.04)	415.71 (+83.57)	**42.10 (+4.40)**	72.38 (-0.19)	64.05 (-0.96)	36.50 (-1.20)
CCoT	60.43 (-4.79)	63.50 (-2.10)	1424.86 (-158.81)	343.57 (+11.43)	35.80 (-1.90)	72.74 (+0.17)	64.18 (-0.83)	36.70 (-1.00)
CoT + RED	65.48 (+0.26)	74.10 (+8.50)	1583.56 (-0.11)	410.36 (+78.22)	40.60 (+2.90)	73.19 (+0.62)	**66.66 (+1.65)**	**39.10 (+1.40)**
CCoT + RED	**65.91 (+0.69)**	**76.50 (+10.90)**	1562.72 (-20.95)	**441.79 (+109.65)**	**42.10 (+4.40)**	**73.22 (+0.65)**	66.04 (+1.03)	38.80 (+1.10)

instruction. While they were proposed for object hallucination in LVLMs, comparing RED with them helps validate the practicality because they reportedly improved general performance [26, 56].

Benchmark Datasets. We used six diverse visual reasoning benchmark datasets: GQA [18], TextVQA [48], MME [11], SEED-I [27], LLaVA-Bench [32], MM-Vet [63], and MathVista [34]. These benchmarks are generally used for evaluating various types of multi-modal capabilities, including general question answering, visual recognition, OCR, mathematical reasoning, multi-modal comprehension, and spatial understanding.

Models. We used publicly available LVLMs on HuggingFace [59]: Gemma-3 (4B, 12B, 27B) [53], Qwen-2.5-VL (7B, 32B, 72B) [3], and Llama3-LLaVA-Next (8B) [28].

Evaluation Protocols. We used greedy decoding. λ for RED was chosen from $\{0.1, 0.3, 0.5, 1.0, 10.0\}$ using scores on validation or development sets; We discuss the effect of λ in Appendix. Tables show scores with parenthesized relative improvement or degradation from Baseline.

5.2. Evaluation on General Visual Reasoning Tasks

Table 2 shows RED's general visual reasoning improvements. While CoT/CCoT sometimes outperformed Baseline, they were inconsistent, with large drops on some LVLM-benchmark pairs, especially TextVQA, which requires understanding texts in images. In contrast, RED consistently improved CoT/CCoT and outperformed all baselines in nearly all cases, even on TextVQA. RED thus successfully addresses the issues of the existing multi-modal CoT using $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$ in terms of performance. The consistent improvements offered by RED highlight the potential value of intermediate rationales generated in multi-modal CoT and its practicality for leveraging in diverse domains.

5.3. Detailed Analysis for Multi-modal Capabilities

Table 3 summarizes the detailed performance analysis with respect to nine multi-modal capabilities of LVLMs provided

Table 3. **Detailed performance analysis on SEED-I for different CoT strategies**. RED's outputs evolve in accordance with the characteristics of specified CoT strategies. For example, CoT with natural text rationales enhances text understanding, while CCoT with structured scene graph rationales improves visual reasoning capability.

	Scene Understanding	Instance Identity	Instance Attributes	Instance Location	Instances Counting	Spatial Relation	Instance Interaction	Visual Reasoning	Text Understanding
Gemma-3-4B									
Baseline	75.14	70.18	69.22	55.11	56.63	44.14	**72.16**	76.74	45.88
CoT	72.99 (-2.15)	70.02 (-0.16)	68.51 (-0.71)	57.41 (+2.30)	55.21 (-1.42)	49.92 (+5.78)	65.98 (-6.18)	72.21 (-4.53)	51.76 (+5.88)
CCoT	74.89 (-2.25)	**70.84 (+0.66)**	70.23 (+1.01)	**57.87 (+2.76)**	54.68 (-1.95)	44.81 (+0.67)	67.01 (-5.15)	75.23 (-1.51)	43.53 (-2.35)
CoT + RED	75.14 (+0.00)	**70.84 (+0.66)**	69.89 (+0.67)	57.46 (+2.35)	55.21 (-1.42)	**50.68 (+6.54)**	67.01 (-5.15)	74.62 (-2.12)	**52.94 (+7.06)**
CCoT + RED	**75.93 (+0.79)**	70.19 (+0.01)	**71.52 (+2.30)**	**57.87 (+2.76)**	**56.86 (+0.23)**	45.66 (+1.52)	67.01 (-5.15)	**77.34 (+0.60)**	43.53 (-2.35)
Gemma-4-12B									
Baseline	77.45	75.15	72.60	66.56	62.08	57.53	74.23	77.95	37.65
CoT	76.25 (-1.20)	74.11 (-1.04)	72.51 (-0.09)	63.80 (-2.76)	62.61 (+0.53)	61.04 (+3.51)	73.20 (-1.03)	76.74 (-1.21)	61.18 (+23.53)
CCoT	77.58 (+0.13)	75.42 (+0.27)	72.75 (+0.15)	65.85 (-0.71)	**64.98 (+2.90)**	56.62 (-0.91)	74.23 (+0.00)	**79.15 (+1.20)**	60.00 (+22.35)
CoT + RED	77.64 (+0.19)	75.20 (+0.05)	74.25 (+1.65)	64.93 (-1.63)	63.38 (+1.30)	**63.01 (+5.48)**	**76.29 (+2.06)**	78.55 (+0.60)	**63.53 (+25.88)**
CCoT + RED	**78.28 (+0.83)**	**75.97 (+0.82)**	**74.79 (+2.19)**	**68.10 (+1.54)**	64.53 (+2.45)	59.82 (+2.29)	75.26 (+1.03)	**79.15 (+1.20)**	60.59 (+22.94)
Qwen2.5-VL-7B									
Baseline	66.37	61.11	68.85	40.18	32.04	49.16	72.16	74.92	45.88
CoT	69.06 (+2.69)	62.37 (+1.26)	71.26 (+2.41)	45.19 (+5.01)	33.76 (+1.72)	51.29 (+2.13)	73.20 (+1.04)	77.04 (+2.12)	50.59 (+4.71)
CCoT	73.34 (+6.97)	65.21 (+4.10)	74.70 (+5.85)	54.29 (+14.11)	38.08 (+6.04)	54.49 (+5.33)	74.23 (+2.07)	77.95 (+3.03)	54.12 (+8.24)
CoT + RED	80.34 (+13.97)	81.21 (+20.10)	**81.33 (+12.48)**	72.70 (+32.52)	**73.80 (+41.76)**	**66.97 (+17.81)**	74.23 (+2.07)	80.66 (+5.74)	**84.71 (+38.83)**
CCoT + RED	**80.37 (+14.00)**	**81.70 (+20.59)**	81.05 (+12.20)	**73.11 (+32.93)**	73.64 (+41.60)	64.99 (+15.83)	**76.29 (+4.13)**	**81.27 (+6.35)**	77.65 (+31.77)
Llama3-LLaVA-Next-8B									
Baseline	77.77	76.52	76.23	65.24	64.36	52.21	**75.26**	76.13	55.29
CoT	77.23 (-0.54)	75.42 (-1.10)	77.16 (+0.93)	66.97 (+1.73)	61.10 (-3.26)	**55.71 (+3.50)**	72.16 (-3.10)	73.41 (-2.72)	56.65 (+1.36)
CCoT	76.88 (-0.89)	76.95 (+0.43)	77.35 (+1.12)	**67.08 (+1.84)**	63.10 (-1.26)	54.34 (+2.13)	70.10 (-5.16)	75.83 (-0.30)	51.76 (-3.53)
CoT + RED	**78.75 (+0.98)**	**77.06 (+0.54)**	**77.69 (+1.46)**	66.67 (+1.43)	64.12 (-0.24)	54.79 (+2.58)	73.20 (-2.06)	**78.25 (+2.12)**	**64.71 (+9.42)**
CCoT + RED	78.28 (+0.51)	**77.50 (+0.98)**	**78.55 (+2.32)**	**67.08 (+1.84)**	**64.94 (+0.58)**	53.27 (+1.06)	74.23 (-1.03)	77.04 (+0.91)	57.65 (+2.36)

by the SEED-I benchmark. Notably, RED amplified multi-modal capabilities depending on the nature of the given rationale, i.e., detailed textual descriptions in CoT and JSON-formatted scene graphs in CCoT. On the one hand, CoT + RED significantly enhanced text understanding and spatial relation capabilities, which require detailed information to be accurate. On the other hand, CCoT + RED remarkably improved instance attributes and instance localization, which can be clearly described using semi-structured data such as JSON. These observations indicate that RED can appropriately condition the rationale content to the final output and assign different capabilities to LVLMs according to the characteristics of a given rational format.

5.4. Intervention Analysis

We examine whether RED can overcome the issue of multi-modal CoT not being able to leverage rationale, as described in Section 3.3. We hypothesized that the performance should improve with higher-quality rationales and degrade with irrelevant ones. To evaluate this, we conduct intervention analysis using GPT-4 [1] generated rationales as the high-quality rationale, and the randomly swapped irrelevant rationales in the same way as Section 3.3. Table 4 shows the results of this intervention analysis on GQA with CCoT, where **Self** denotes using the rationales generated from the backbone LVLM itself, **GPT-4** denotes using high-quality rationales generated by GPT-4, and **Random** denotes using irrelevant rationales swapped from other instances. Note that we used the same λ tuned with the Self rationales across all cases. CCoT + RED with GPT-4 rationales greatly improved performance, Random rationales degraded it. The substantial performance gains observed with CCoT + RED using GPT-4 rationales, exceeding the CCoT counterparts, indicate that RED effectively leverages improved rationale content for multi-modal reasoning. These results indicate not only that RED can overcome the issue of conventional multi-modal CoT reasoning but also that RED offers some interpretability in the reasoning process, which is dependent on the rationale.

5.5. Scalability for Larger LVLMs

We test RED on LVLMs with a larger parameter size for evaluating the scalability. We additionally utilized Gemma-3-27B, Qwen2.5-VL-32B/-72B. These are increased language model sizes, and the vision encoders remain unchanged. Figure 3 shows the scalability on larger LVLMs. Contrary to intuition, neither Baseline nor CCoT consistently improved

	Self (↑)	GPT-4 (↑)	Random (↓)
Gemma-3-4B			
CCoT	44.54 (+4.54)	45.79 (+5.79)	44.35 (+4.35)
CCoT + RED	**45.87** (**+5.87**)	**48.93** (**+8.93**)	39.36 (-0.64)
Gemma-12-4B			
CCoT	44.50 (-0.84)	45.61 (+0.27)	44.30 (-1.04)
CCoT + RED	**47.50** (**+2.16**)	**50.04** (**+4.70**)	43.29 (-2.05)
Qwen2.5-VL-7B			
CCoT	46.69 (-14.19)	50.85 (-10.03)	51.76 (-9.12)
CCoT + RED	**61.92** (**+1.04**)	**63.11** (**+2.23**)	41.74 (-19.14)
Llama3-LLaVA-Next-8B			
CCoT	60.43 (-4.79)	57.84 (-7.38)	60.91 (-4.31)
CCoT + RED	**65.91** (**+0.69**)	**67.88** (**+2.66**)	58.70 (-6.52)

Table 4. **Intervention analysis on GQA with CCoT**. Performance is shown when using self-generated rationales (Self), rationales generated by GPT-4 (GPT-4), and randomly swapped irrelevant rationales (Random). RED can be improved by a higher-quality rationale and degraded by a lower-quality one, demonstrating its faithfulness in grounding on rationales.

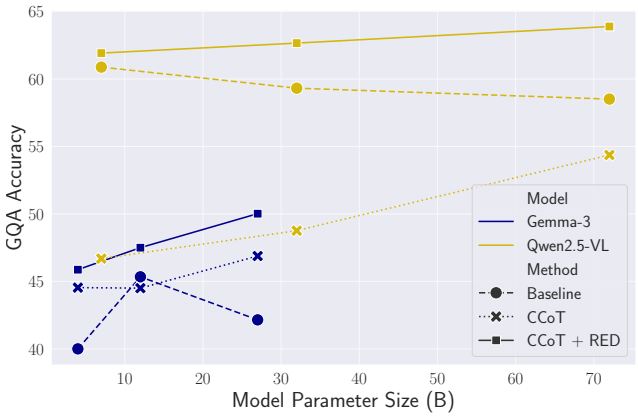

Figure 3. **Performance trends on GQA accuracy with increasing LVLM parameter sizes** (Baseline, CCoT, CCoT + RED for Gemma-3 and Qwen2.5-VL families). RED can consistently improve Baseline and CCoT in any models, and can unlock further performance scalability according to model sizes.

performance with model size. This is potentially because, as shown in Section 3.3, the decoding process often over-relies on visual input tokens at the expense of rationales; a similar observation is found in [2]. In contrast, our RED consistently improved performance in proportion to model size, indicating that RED can leverage sophisticated rationales from larger language models. This implies that RED could be a key factor in unlocking the full potential of larger LVLMs by enabling more effective multi-modal CoT.

5.6. Qualitative Evaluation

Figure 4 shows qualitative examples from GQA; we performed reasoning with CoT and CCoT at the top and bottom

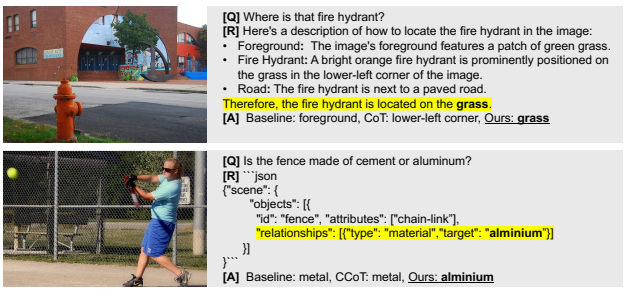

Figure 4. **Qualitative examples of CoT reasoning on GQA (Gemma-3-12B)**. For each example, the input image (left), the query [Q], the generated rationale [R] (text for CoT and JSON for CCoT), and the answers from Baseline, CoT/CCoT, and our method (RED) [A] are shown. RED successfully leverages the rationale to produce the correct answer.

of the figure, respectively. While generated rationales point to the correct answer, naive CoT/CCoT produce incorrect responses, highlighting their failure to leverage the rationales by $p_\theta(y_i|\boldsymbol{y}_{<i}, x, r, q)$. In contrast, RED extracts the conclusion and related attributes from the rationales in both cases.

5.7. Inference Efficiency

We evaluate the inference efficiency of RED. The inference steps are decomposed into (i) rationale generation by Eq. (5) and (ii) answer generation by Eq. (8). The former is the same as the existing CoT, and the latter requires a dual-forward pass, which we implemented using batch-parallel decoding. To assess the latency, we measured the average inference time for each query across all benchmarks with Gemma-3-12B on Baseline, CoT, and RED. The results were 3.01, 5.27, and 5.05 for Baseline, CoT, and RED, respectively. RED was faster than CoT because the contexts of (x, q) and (r, q) of the batch-parallel decoding were shorter than the full context (x, r, q) of CoT, emphasizing the efficiency of our method. Nonetheless, slightly increasing GPU memory consumption due to parallel RED is a limitation that should be resolved in future work.

6. Conclusion

This paper addressed the challenge of LVLMs ineffectively using rationales in CoT reasoning. We proposed Rationale-Enhanced Decoding (RED), an optimal, plug-and-play solution derived from KL-constrained reward maximization, which harmonizes visual and rationale inputs by multiplying their distinct conditional distributions. Extensive experiments demonstrated RED's consistent and significant reasoning improvements over existing methods.

A notable limitation is the increased inference overhead, a common trade-off for plug-and-play decoding strategies. Future efforts could focus on mitigating this cost. Despite this, RED offers a significant step towards more reliable, interpretable, and rationale-grounded multi-modal systems.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 1, 2, 7

[2] Haider Al-Tahan, Quentin Garrido, Randall Balestriero, Diane Bouchacourt, Caner Hazirbas, and Mark Ibrahim. Unibench: Visual reasoning requires rethinking vision-language beyond scaling. Advances in Neural Information Processing Systems, 2024. 8

[3] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923, 2025. 1, 6

[4] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022. 1

[5] Samyadeep Basu, Martin Grayson, Cecily Morrison, Besmira Nushi, Soheil Feizi, and Daniela Massiceti. Understanding information storage and transfer in multi-modal large language models. Advances in Neural Information Processing Systems, 2024. 3

[6] Declan Campbell, Sunayana Rane, Tyler Giallanza, Camillo Nicolò De Sabbata, Kia Ghods, Amogh Joshi, Alexander Ku, Steven Frankland, Tom Griffiths, Jonathan D Cohen, et al. Understanding the limits of vision language models through the lens of the binding problem. Advances in Neural Information Processing Systems, 2024. 2

[7] Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. PerturboLLaVA: Reducing multimodal hallucinations with perturbative visual training. In International Conference on Learning Representations, 2025. 2

[8] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024. 1, 2

[9] Wei Chow, Juncheng Li, Qifan Yu, Kaihang Pan, Hao Fei, Zhiqi Ge, Shuai Yang, Siliang Tang, Hanwang Zhang, and Qianru Sun. Unified generative and discriminative training for multi-modal large language models. Advances in Neural Information Processing Systems, 2024. 2

[10] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In Advances in neural information processing systems, 2023. 1, 2

[11] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. arXiv preprint arXiv:2306.13394, 2023. 6

[12] Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in LVLMs. In The Thirteenth International Conference on Learning Representations, 2025. 2

[13] Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In International Conference on Learning Representations, 2025. 4

[14] Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 18135–18143, 2024. 2

[15] Hulingxiao He, Geng Li, Zijun Geng, Jinglin Xu, and Yuxin Peng. Analyzing and boosting the power of fine-grained visual recognition for multi-modal large language models. In International Conference on Learning Representations, 2025. 4

[16] Liqi He, Zuchao Li, Xiantao Cai, and Ping Wang. Multi-modal latent space learning for chain-of-thought reasoning in language models. In Proceedings of the AAAI Conference on Artificial Intelligence, pages 18180–18187, 2024. 1, 2

[17] Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. Neural computation, 14(8): 1771–1800, 2002. 5

[18] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 3, 6

[19] Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. How interpretable are reasoning explanations from prompting large language models? In Findings of the Association for Computational Linguistics: NAACL 2024, pages 2148–2164, 2024. 4

[20] Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In International Conference on Learning Representations, 2025. 3, 4

[21] Junho Kim, Hyunjun Kim, Yeonju Kim, and Yong Man Ro. Code: Contrasting self-generated description to combat hallucination in large multi-modal models. Advances in Neural Information Processing Systems, 2024. 2

[22] Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Attention is not only a weight: Analyzing transformers with vector norms. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 3

[23] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. Advances in neural information processing systems, 2022. 1

[24] Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, et al. Measuring faithfulness in chain-of-thought reasoning. arXiv preprint arXiv:2307.13702, 2023. 4

[25] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? Advances in Neural Information Processing Systems, 2024. 2

[26] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 2, 5, 6

[27] Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench: Benchmarking multimodal large language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 6

[28] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. arXiv preprint arXiv:2407.07895, 2024. 6

[29] Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023. 2

[30] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In The 2023 Conference on Empirical Methods in Natural Language Processing, 2023. 2, 4

[31] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In Advances in Neural Information Processing Systems, 2023. 1, 2

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 6

[33] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems, 2022. 1, 3

[34] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In International Conference on Learning Representations, 2024. 6

[35] Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). In Findings of the Association for Computational Linguistics ACL 2024, 2024. 2

[36] Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional chain-of-thought prompting for large multimodal models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024. 1, 2, 3, 5

[37] Debjyoti Mondal, Suraj Modi, Subhadarshi Panda, Rituraj Singh, and Godawari Sudhakar Rao. Kam-cot: Knowledge augmented multimodal chain-of-thoughts reasoning. In Proceedings of the AAAI conference on artificial intelligence, 2024. 1, 2, 3

[38] Clement Neo, Luke Ong, Philip Torr, Mor Geva, David Krueger, and Fazl Barez. Towards interpreting visual information processing in vision-language models. In International Conference on Learning Representations, 2025. 2, 3

[39] Timothy Ossowski and Junjie Hu. Olive: Object level in-context visual embeddings. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics, 2024. 2

[40] Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. A concept-based explainability framework for large multimodal models. Advances in Neural Information Processing Systems, 37:135783–135818, 2024. 2

[41] Debjit Paul, Robert West, Antoine Bosselut, and Boi Faltings. Making reasoning matter: Measuring and improving faithfulness of chain-of-thought reasoning. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15012–15032, 2024. 4

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In International conference on machine learning. PMLR, 2021. 1

[43] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. Advances in Neural Information Processing Systems, 2023. 5

[44] Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In International Conference on Learning Representations, 2023. 1

[45] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. Trust region policy optimization. In International conference on machine learning. PMLR, 2015. 2, 4, 5

[46] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347, 2017. 2, 4, 5

[47] Yucheng Shi, Quanzheng Li, Jin Sun, Xiang Li, and Ninghao Liu. Enhancing cognition and explainability of multimodal foundation models with self-synthesized data. In International Conference on Learning Representations, 2025. 2

[48] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019. 6

[49] Swetha Sirnam, Jinyu Yang, Tal Neiman, Mamshad Nayeem Rizve, Son Tran, Benjamin Yao, Trishul Chilimbi, and Mubarak Shah. X-former: Unifying contrastive and reconstruction learning for mllms. In European Conference on Computer Vision, 2024. 2

[50] Guangyan Sun, Mingyu Jin, Zhenting Wang, Cheng-Long Wang, Siqi Ma, Qifan Wang, Tong Geng, Ying Nian Wu, Yongfeng Zhang, and Dongfang Liu. Visual agents as fast and slow thinkers. In International Conference on Learning Representations, 2025. 2

[51] Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan 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, 2024. 2

[52] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805, 2023. 1

[53] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. arXiv preprint arXiv:2503.19786, 2025. 3, 6

[54] Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. Towards understanding chain-of-thought prompting: An empirical study of what matters. In Annual Meeting Of The Association For Computational Linguistics, 2023. 1

[55] Jiayu Wang, Yifei Ming, Zhenmei Shi, Vibhav Vineet, Xin Wang, Sharon Li, and Neel Joshi. Is a picture worth a thousand words? delving into spatial reasoning for vision language models. Advances in Neural Information Processing Systems, 37:75392–75421, 2024. 2

[56] 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. 2, 5, 6

[57] Zecheng Wang, Chunshan Li, Zhao Yang, Qingbin Liu, Yanchao Hao, Xi Chen, Dianhui Chu, and Dianbo Sui. Analyzing chain-of-thought prompting in black-box large language models via estimated v-information. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024. 1

[58] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35: 24824–24837, 2022. 1, 3, 5

[59] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771, 2019. 6

[60] Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. Evaluating and analyzing relationship hallucinations in large vision-language models. In International Conference on Machine Learning, 2024. 2

[61] Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In International Conference on Learning Representations, 2024. 4

[62] Runpeng Yu, Weihao Yu, and Xinchao Wang. Attention prompting on image for large vision-language models. In European Conference on Computer Vision, 2024. 2

[63] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In International Conference on Machine Learning, pages 57730–57754. PMLR, 2024. 6

[64] Yunan Zeng, Yan Huang, Jinjin Zhang, Zequn Jie, Zhenhua Chai, and Liang Wang. Investigating compositional challenges in vision-language models for visual grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14141–14151, 2024. 2

[65] Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q. Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia P. Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. In International Conference on Learning Representations, 2025. 2

[66] Meiru Zhang, Zaiqiao Meng, and Nigel Collier. Can we instruct llms to compensate for position bias? In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 12545–12556, 2024. 4

[67] Yuhui Zhang, Alyssa Unell, Xiaohan Wang, Dhruba Ghosh, Yuchang Su, Ludwig Schmidt, and Serena Yeung-Levy. Why are visually-grounded language models bad at image classification? Advances in Neural Information Processing Systems, 2024. 2

[68] Zhenyu Zhang, Runjin Chen, Shiwei Liu, Zhewei Yao, Olatunji Ruwase, Beidi Chen, Xiaoxia Wu, and Zhangyang Wang. Found in the middle: How language models use long contexts better via plug-and-play positional encoding. Advances in Neural Information Processing Systems, 2024. 4

[69] Zhuosheng Zhang, Aston Zhang, Mu Li, hai zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. Transactions on Machine Learning Research, 2024. 1, 2, 3, 5

[70] Changmeng Zheng, Dayong Liang, Wengyu Zhang, Xiao-Yong Wei, Tat-Seng Chua, and Qing Li. A picture is worth a graph: A blueprint debate paradigm for multimodal reasoning. In Proceedings of the 32nd ACM International Conference on Multimedia, pages 419–428, 2024. 1

[71] Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. Advances in Neural Information Processing Systems, 2023. 2, 3

[72] Guangmin Zheng, Jin Wang, Xiaobing Zhou, and Xuejie Zhang. Enhancing semantics in multimodal chain of thought via soft negative sampling. In Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), pages 6059–6076, 2024. 1, 2

[73] Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun

Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. Advances in Neural Information Processing Systems, 2024. 2

[74] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In International Conference on Learning Representations, 2024. 1, 2