RAPPER: REINFORCED RATIONALE-PROMPTED PARADIGM FOR NATURAL LANGUAGE EXPLANATION IN VISUAL QUESTION ANSWERING

Kai-Po Chang[1] **Chi-Pin Huang**[1] **Wei-Yuan Cheng**[1] **Fu-En Yang**[1,2]
Chien-Yi Wang[2] **Yung-Hsuan Lai**[1] **Yu-Chiang Frank Wang**[1,2]
[1]Graduate Institute of Communication Engineering, National Taiwan University, Taiwan
[2]NVIDIA, Taiwan
{r11942093,r11942097,b08502072,r10942097}@ntu.edu.tw,
{fredy,frankwang,chienyiw}@nvidia.com

ABSTRACT

Natural Language Explanation (NLE) in vision and language tasks aims to provide human-understandable explanations for the associated decision-making process. In practice, one might encounter explanations which lack informativeness or contradict visual-grounded facts, known as *implausibility* and *hallucination* problems, respectively. To tackle these challenging issues, we consider the task of visual question answering (VQA) and introduce *Rapper*, a two-stage **R**einforced R**a**tionale-**P**rom**p**t**e**d Pa**r**adigm. By knowledge distillation, the former stage of *Rapper* infuses rationale-prompting via large language models (LLMs), encouraging the rationales supported by language-based facts. As for the latter stage, a unique Reinforcement Learning from NLE Feedback (RLNF) is introduced for injecting visual facts into NLE generation. Finally, quantitative and qualitative experiments on two VL-NLE benchmarks show that *Rapper* surpasses state-of-the-art VQA-NLE methods while providing plausible and faithful NLE.

1 INTRODUCTION

Deep learning has achieved remarkable success in vision-language (VL) tasks such as visual reasoning (Suhr et al., 2017), visual question answering (VQA, Goyal et al., 2017), and visual entailment (Xie et al., 2019). Take VQA as an example, while these models exhibit impressive ability in inferring answer descriptions from the given image-question pairs, its decision-making process remains an unsolved problem. As a result, such a black-box manner severely restricts their applicability in certain real-world scenarios (e.g., medical VQA, Lin et al., 2023), where the interpretability of the learning model is crucial for establishing trustworthy systems. To tackle this long-standing challenge, some approaches adopt attention mechanisms (Anderson et al., 2018) or gradient-based activations (Selvaraju et al., 2017), focusing on highlighting image regions which are relevant to the associated prediction. However, such visual explanations might not be desirable for VL tasks (e.g., those beyond classification) due to the lack of reasoning process (Kayser et al., 2021; Sammani et al., 2022). As a result, Natural Language Explanation (NLE) has emerged as a potential alternative, which aims to interpret the underlying reasoning process by natural language descriptions.

To extend NLE for vision-language tasks (i.e., VL-NLE), Park et al. (2018) and Kayser et al. (2021) introduced the benchmarks for explaining the decision-making process with NLEs for VQA and visual entailment tasks, respectively. Subsequent VL-NLE works have evolved into two research lines. The first research line (Park et al., 2018; Marasović et al., 2020) focuses on how to improve their pipeline from an architecture perspective for training NLE generators within a fully supervised learning manner. On the other hand, Sammani et al. (2022) and Suo et al. (2023) emphasize the utilization of unlabeled pre-training data to enhance the language models' NLE capability.

Despite significant advancements, most existing VL-NLE works require training in a full supervised manner. They might encounter problems where the explanations are irrelevant to the questions or contradictory to the established supporting facts (Majumder et al., 2021). The other potential concern

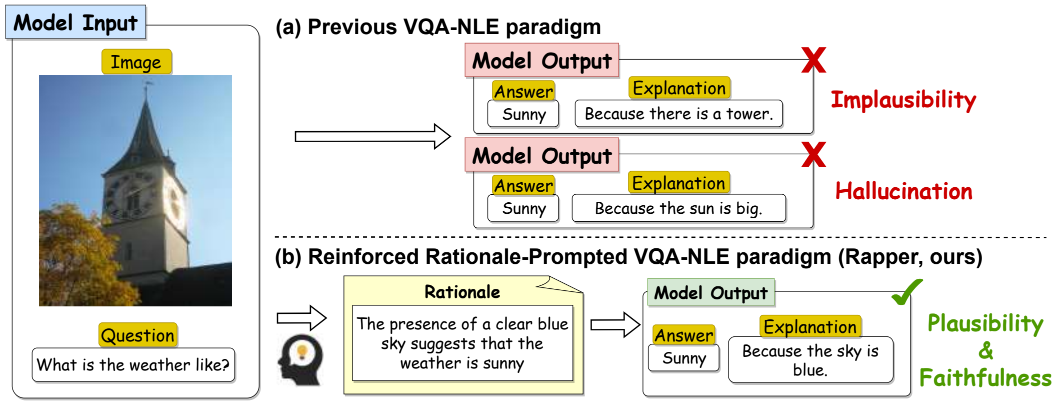

Figure 1: Comparison between (a) previous VQA-NLE paradigm and (b) our proposed reinforced rationale-prompted VQA-NLE paradigm of (*Rapper*). Instead of directly generating answer or explanation, *Rapper* learns plausible and faithful explanations which prompt the VQA model with improved performance.

is that the explanation is not related to the visual image (Ji et al., 2023). More specifically, the former problem is referred to as *implausibility*, while the latter is known as *hallucination*. Take visual input and question in Fig. 1 as an example, "Because there is a tower.' is an *implausible* explanation since it is irrelevant to question, and "Because the sun is big." is a *hallucinated* one since the sun is not visible in the image. Although these issues have been recently studied in the NLE community (Zhao et al., 2023; Turpin et al., 2023), they remain unexplored in the field of VL-NLE. As a result, generating *plausible* yet *faithful* NLEs for elucidating vision-language models continues to pose a crucial challenge.

Recently, rationale-based prompting techniques have been manifested to improve the capability of Large Language Models (LLMs) on complex reasoning tasks (Wei et al., 2022; Liu et al., 2022b). Such techniques involve elicitation of rationales from LLMs, producing knowledge-riched or fact-based intermediate to facilitate the reasoning capability of language model. Thus, these prompting manners are emerging as promising solutions for NLE (Zhao et al., 2023; Krishna et al., 2023). These rationale-prompting paradigms have been further extended to multi-modal regimes such as mm-CoT (Zhang et al., 2023) and mm-ReAct (Yang et al., 2023). However, mm-CoT (Zhang et al., 2023) relies on the ground-truth rationales for training, while mm-ReAct (Yang et al., 2023) have potential hallucinated outputs due to the information loss when converting visual signals into text for ChatGPT API call understanding.

In this paper, we propose ***R**einforced **R**ationale-**P**rom**p**t**e**d Pa**r**adigm* (*Rapper*) for providing accurate answers for VQA with sufficient NLE, which are plausible and faithful. As depicted in Fig. 1(b), our *Rapper* learns to exploit knowledge learned from LLM and incorporate the corresponding visual content from input images into *rationales* through two stages. Without observing any ground truth rationale during training, the first stage utilizes a knowledge distillation process to introduce LLM for enriching the rationales with supporting facts, encouraging NLE to be factual and plausible. The subsequent stage of *Reinforcement Learning from NLE Feedback* (RLNF) further exploits the answer-explanation feedback to enforce the produced rationales associated with both question and visual inputs, allowing faithful NLE.

We now summarize the contributions of this work below:

- A reinforced rationale-prompted paradigm, *Rapper*, is proposed for plausible and faithful NLE generation in VQA. This is achieved through two proposed stages: knowledge distillation process from LLM and *Reinforcement Learning from NLE Feedback* (*RLNF*).
- In *Rapper*, we first advance LLM and perform knowledge distillation. This results in predicted rationales being based on language-based facts, which prompt the VQA model for plausible NLE.
- To align NLE with the visual input, we introduce *Reinforcement Learning from NLE Feedback* (RLNF) to *Rapper*, which utilizes the answer-explanation feedback as rewards and prompts the VQA model with predicted rationales for faithful NLE.

- Our *Rapper* achieves new state-of-the-art performance for both VQA-X (Park et al., 2018) and e-SNLI-VE (Kayser et al., 2021) on NLE generation. We also demonstrate that *Rapper* outperforms existing VQA-NLE works with reduced implausibility and hallucination.

2 Related work

Plausible and Faithful Natural Language Explanation Research on plausibility and faithfulness in NLE (Majumder et al., 2021; King et al., 2022; Gou et al., 2023; Stacey et al., 2023) has garnered wide attention, particularly due to the evolution of Large Language Models (LLMs) and chain-of-thought (CoT) prompting techniques (Wei et al., 2022). Notably, the method of integrating external knowledge databases for fact generation or retrieval has been proven effective in enhancing the plausibility and faithfulness of NLEs (Majumder et al., 2021; Stacey et al., 2023). Based on this advancement, some recent approaches, such as the *verify-then-correct* pipeline by Gou et al. (2023) and novel decoding strategies proposed by Lan et al. (2023) and King et al. (2022), aim to mitigate hallucination in textual outputs. However, these works typically focus on isolated single text modality or rely on static external knowledge databases, limiting its scalability to multimodal data.

Natural Language Explanation for Vision-Language Tasks Most existing VL-NLE works (Wu & Mooney, 2018a; Park et al., 2018; Marasović et al., 2020; Kayser et al., 2021) generate explanations in a predict-then-explain fashion. Specifically, an answer is first predicted by a pre-trained VL model (e.g., UNITER (Chen et al., 2020) or Oscar (Li et al., 2020)), followed by the generation of the corresponding explanation via a separate language decoder (e.g., GPT2 (Radford et al., 2019)). As the answer and explanation are predicted separately, the explanation often contains irrelevant or contradictory descriptions of the given visual information, struggling to faithfully represent the underlying reasoning process. Recently, NLX-GPT (Sammani et al., 2022) proposes to jointly generate the answer and explanation by a unified sequence-to-sequence model, while S3C (Suo et al., 2023) further enforces the explanation to be consistent with the predicted answer. Although the above approaches have been shown to mitigate the hallucination issue, it is not clear how their NLE is established upon supporting facts or taking the visual input into consideration. Therefore, how to tackle the potential implausibile or hallucinated NLE remains a challenging task.

Reinforcement Learning for Language Models Several research works have explored RL and view it as the key component to enhance models across vision-language tasks such as image captioning (Rennie et al., 2017), novel object captioning (NOC) (Yang et al., 2022), and VQA (Lu et al., 2022a; Fan et al., 2018; Liu et al., 2018). There has been a concentrated effort to align LMs with natural language (NL) feedback (Akyürek et al., 2023; Yang et al., 2022; Liu et al., 2022a) as well as non-NL feedback (Bai et al., 2022; Lu et al., 2022b). For example, Liu et al. (2022a) utilizes the probability of the correct answer as a reward to stimulate an auxiliary module to produce beneficial knowledge, thereby enhancing QA-task performance. Similarly, Yang et al. (2022) employs a CIDEr optimization strategy to enhance the caption with sufficiently visual fidelity in the task of novel object captioning. Despite of their effectiveness, their RL framework or NL-feedback approaches cannot be easily applied for VL-NLE tasks.

3 Proposed method

3.1 Problem Formulation

Given a VQA input $X = (V, Q)$ consisting of an input image V and a textual input Q (i.e., question), our goal is to predict the answer $\hat{A}$ and the corresponding explanation $\hat{E}$ (denoted as $\hat{Y} = (\hat{A}, \hat{E})$) via a reasoning module M. In order to encourage M to provide correct answer with plausible and faithful explanation, we propose a ***R**einforced **Ra**tionale-**P**rom**p**t**e**d Pa**r**adigm* (*Rapper*) scheme, which learns an additional rationale generator G to jointly exploit the supporting facts from LLMs and the visual content observed from the conditioned image into rationales. Note that only the ground truth A and E are available during training, not the rationales. As depicted in Fig. 2, the learning of *Rapper* is decomposed into: (A) *Knowledge Distillation from LLM* (Sec. 3.2), and (B) *Reinforcement learning from NLE Feedback (RLNF)* (Sec. 3.3), which trains rationale generator G for providing auxiliary intermediates when predicting $\hat{Y} = (\hat{A}, \hat{E})$.

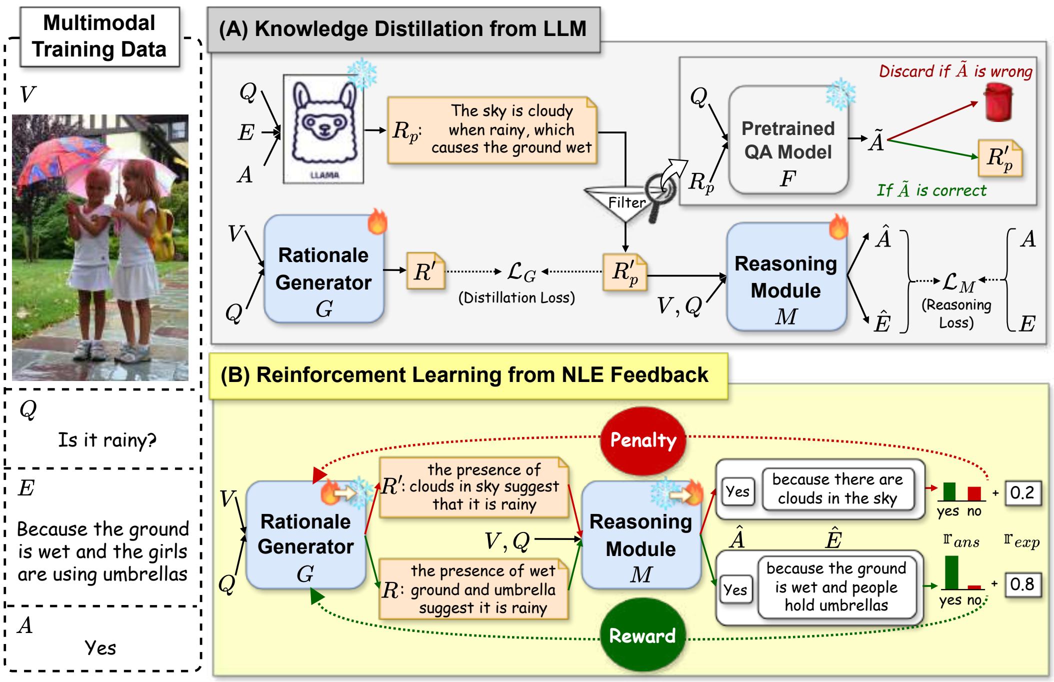

Figure 2: Overview of *Rapper*. *Rapper* involves two training stages: (A) *Knowledge distillation* introduce the rationales R'_p from LLM by offering established facts, facilitating the generation of plausible NLEs from the reasoning module M. (B) *Reinforcement learning from NLE feedback* (*RLNF*) further refines the rationales from R' to R by incorporating visual information, encouraging generation of faithful NLEs from M.

3.2 Plausible NLE Generation

Since VQA-NLE models typically rely on ground truth answers and explanations for training, it is not clear whether the underlying visual and language knowledge are exploited to support the predicted outputs. In the first stage of *Rapper*, we propose to leverage powerful *reasoning* capability inherent in LLM for plausible NLE generation. As depicted in Fig. 2(A), we propose to learn a rationale generator G by utilizing knowledge distillation from LLM (e.g., LLaMA-65B (Touvron et al., 2023)). This would have the reasoning module M elaborate the conditioned rationales before answering and explaining and encourage plausible NLE. We now detail this learning stage.

3.2.1 Knowledge Distillation for Facted-based Rationale Generation

With the recent success of LLMs showing great capability for generating rationale prompts as intermediate reasoning steps and knowledge (Wei et al., 2022; Kojima et al., 2022; Liu et al., 2022b) for reasoning task, we propose to advance the guidance of pre-trained LLMs to acquire such knowledge, so that supporting facts or knowledge can be exploited and serve as rationales for VL-NLE. Since no ground-truth rationales are available, we leverage the LLM to produce rationales as pseudo ground truth for training our rationale generator G. Inspired by Liu et al. (2022a;b) and Min et al. (2022), we elicit pseudo rationale r_p from LLM with a task-specific set of few-shot demonstrations (see Sec. A.5 for details) as follows:

$$R_p = \{r_p \mid r_p \sim P_{LLM}(y, q)\}, \tag{1}$$

where y is the ground-truth answer-explanation pair, q is question, P_{LLM} denotes the LLM in an autoregressive manner, r_p is the sampled pseudo rationale from P_{LLM}, and thus R_p is the set of all r_p.

However, the above pseudo rationales may be redundant, noisy or lengthy, which would not be desirable for subsequent NLE tasks (Li et al., 2023b). Thus, we apply a post-processing mechanism to filter pseudo rationales R_p to R'_p. To be specific, we apply a round-trip consistency by answering the input question on the pseudo rationales with a pre-trained question-answering (QA) model F [1]. The pseudo rationale is retained when the matching score between the ground-truth answer and the answer predicted by F exceeds a predetermined threshold τ. This matching score is quantified with the token-level F1 score (Wang et al., 2020). Thus, the process of collecting the filtered pseudo rationales R'_p is formulated as follows:

$$R'_p = \{r_p \mid \texttt{F1-score}(\tilde{a}, a) \geq \tau,\ \tilde{a} \sim P_F(Q, r_p), r_p \in R_p\}, \tag{2}$$

where a is the ground truth answer, $\tilde{a}$ is the answer predicted by F based on the pseudo rationale, and P_F denotes the pre-trained QA model F in an autoregressive fashion.

With the above R'_p serving as psuedo ground truth, we are able to train the rationale generator G with the distillation loss $\mathcal{L}_G$ described below:

$$\mathcal{L}_G = -\sum_{t=1}^{T} \log(p_G(r'_{p,t} | r'_{p,0:t-1}, x)), \tag{3}$$

where $r'_p \in R'_p$, $T = |r'_p|$, and $x = \{v, q\} \in X$.

3.2.2 Prompting by fact-based Rationale for Plausible NLE

With rationales R'_p better aligned with the facts, we can proceed to the training of the reasoning module M for NLE generation. We note that, since rationales R'_p are in the form of natural language, our the reasoning module M (which is also based on visual-language model) would be able to interpret them. Thus, in addition to the image-question pair X as the inputs to the reasoning module M, the derived pseudo rationales R'_p are further viewed as input prompts, which provide fact-supporting conditions when training M to perform VQA-NLE. As a result, we train M by calculating the reasoning loss L_M as follows:

$$\mathcal{L}_M = -\sum_{t=1}^{T} \log(p_M(y_t | y_{0:t-1}, r'_p, x)). \tag{4}$$

In the above cross-entropy loss, $y = [a; e] \in Y$ is the concatenation of the ground-truth answer a and explanation e.

3.3 Faithful NLE Generation

Although the above knowledge distillation process based on LLM introduces plausibility into our rationale generation, the predicted rationales might not be related to the visual input and thus encounter the hallucination problem. To tackle this issue, we introduce a novel technique of *Reinforcement Learning from NLE Feedback* (RLNF). This learning strategy is to encourage the rationale generator G to fully exploit multimodal input data, so that the output rationales are not only plausible but also faithful. Once G produces faithful rationales, we can fine-tune the reasoning module M for plausible yet faithful NLE.

3.3.1 RLNF for Injecting Visual Facts

To address the potential hallucination issue, we propose *Reinforcement Learning from NLE Feedback* (*RLNF*) by enforcing rationale generator G to derive the visual facts from the input image into rationales. To achieve this, we define a reward function via RL that penalizes the fact-based but hallucinated rationales R', while rewarding the rationales R that contain both established facts and visual content, as depicted in Fig. 2(B). To achieve this, we design our reward $\mathbb{r}_{total}$ to be the addition of answer scores $\mathbb{r}_{ans}$ and the explanation score $\mathbb{r}_{exp}$, which are the average predicted probability of the ground-truth answer and CIDEr score (Vedantam et al., 2015), respectively. For the answer score, inspired by and following Kadavath et al. (2022), we maximize the answer score to assess the faithfulness of the predicted explanation. This maximization enforces the rationale

[1] In the implementation, we follow (Changpinyo et al., 2022) and use UnifiedQA (Khashabi et al., 2022) as the pre-trained QA model.

Algorithm 1 Training RAPPER

Input: Rationale generator G, reasoning module M, LLM P_{LLM} and pre-trained QA model P_F
Data: Image-question pairs $X = \{x^i\}_{i=1}^N$, and answer-explanation pairs $Y = \{y^i\}_{i=1}^N$

/ Stage(A): KD for Plausible NLE Generation */*
$R_p \leftarrow$ Collect pseudo rationales (Eq. equation 1);
$R_p' \leftarrow$ Get filtered pseudo rationales from R_p (Eq. equation 2); ▷ Section 3.2.1
$G \leftarrow$ Update G with $\mathcal{L}_G$ (Eq. equation 3);
$M \leftarrow$ Update M with $\mathcal{L}_M$ (Eq. equation 4); ▷ Section 3.2.2

/ Stage(B): RLNF for Faithful NLE Generation */*
$G \leftarrow$ Update G with $\mathbb{R}_{total}$ (Eq. equation 8); ▷ Section 3.3.1
$M \leftarrow$ Update M with $\mathcal{L}_M$ (Eq. equation 10); ▷ Section 3.3.2

Output: G_θ, M_ϕ

---

generator G to inject more visual content into the rationale because the reasoning module M need more visual clues to correctly answer the question. Therefore, this process transform R' to R, and simultaneously provide the M with more visual fact-based rationale R to enable the explanation with sufficient faithfulness. On the other hand, the explanation score $\mathbb{r}_{exp}$ is (i.e., specifically CIDEr score) to maintain the plausibility of NLE after the first training stage. As a result, the reward $\mathbb{r}_{total}$ is formulated as follows:

$$\mathbb{r}_{total}(x, a, e, \hat{e}, r) = \mathbb{r}_{ans}(a, x, r) + \mathbb{r}_{exp}(e, \hat{e}), \tag{5}$$

$$\mathbb{r}_{ans}(a, x, r) = \mathcal{Z}(P_{M_\phi}(a \mid x, r)), \tag{6}$$

$$\mathbb{r}_{exp}(e, \hat{e}) = \mathcal{Z}(\texttt{CIDEr}(e, \hat{e})), \tag{7}$$

where $x = \{v, q\}$ is the input image-question pair, a denotes the ground-truth answer, e denotes the ground-truth explanation, $\hat{e}$ is the predicted explanation from M, and $r \in R$ is the sampled rationales from G.Notably, $\mathcal{Z}$ is an input-specific normalization function that follows Deng et al. (2022) to normalize reward for stabilizing the RL training process.

RLNF Formulation Our RLNF employs Proximal Policy Optimization (PPO) (Schulman et al., 2017) as the RL algorithm. As the policy model updated, the rationale generator G is to maximize the following reward $\mathbb{R}_{total}$:

$$\max\{\mathbb{R}_{total}(x, a, e, \hat{e}, r)\}, r \sim \prod_{t=1}^{T} P_G(w_t | w_{<t}), \tag{8}$$

where $r = \{w_i\}_{i=0}^T$, $T = |r|$, and $x = \{v, q\}$. However, we need to ensure the generated rationales are understandable by humans and do not deviate too far from the distilled knowledge. To achieve this, we add a KL penalty term between the learned policy θ and the initial policy θ_{init} after the knowledge distillation phase. Therefore, the overall reward is defined as:

$$\mathbb{R}_{total}(x, a, e, \hat{e}, r) = \mathbb{r}_{total}(x, a, e, \hat{e}, r) - \alpha \log \frac{p_G(r|x;\theta)}{p_G(r|x;\theta_{\text{init}})}, \tag{9}$$

where $\mathbb{R}_{total}(x, a, e, \hat{e}, r)$ is the reward in Eq. 5.

3.3.2 PROMPTING BY VISUAL-FACT-BASED RATIONALE FOR FAITHFUL NLE

Once the rationale generator G is trained with the introduced RLNF, it is encouraged to produce visual fact-based rationales R that are encapsulated with established facts and visual content from visual input. Again, since R are natural language prompts, they are inherently interpretable by our reasoning module M. Therefore, for the given image-question pairs X, we utilize R as part of input prompts during the reasoning process of M. This ensures the NLEs from M retain plausibility because of the established supporting facts lies in R, together with the enhanced faithfulness because of the derived visual content embedded in R. We optimize M to achieve this with the reasoning loss $\mathcal{L}_M$ defined as follows:

$$\mathcal{L}_M = -\sum_{t=1}^{T} \log(p_M(y_t | y_{0:t-1}, r, x)), \tag{10}$$

where $r \in R$, $x = \{v, q\} \in X$, and $y = [a; e] \in Y$, which is the concatenated ground-truth answer a and explanation e sequence.

Therefore, through the complete *Rapper* training process as outlined in Algorithm 1, VQA-NLE tasks would be successfully enabled with adequate plausibility and faithfulness.

3.4 Inference

At inference time, for a given input image-question pair $x \in X$, we first generate rationale r on the fly from the rationale generator G:

$$r = \{w_i \mid w_i \sim P_G(w_{<i}, x); i = 0, \ldots, n\},$$

where $r = \{w_i\}_{i=0}^{n}$ is the sampled rationale, $n = |r|$, and $x = \{v, q\}$. Subsequently, we prompt the reasoning module M by concatenating the predicted rationale $\hat{r}$ with the image-question pair x for outputting the final answer and explanation sequence $\hat{y}$. This can be formulated as:

$$\hat{y} = [\hat{a}; \hat{e}] = \{z_i \mid z_i \sim P_M(z_{<i} \mid x, r); i = 0, \ldots, m\},$$

where $m = |\hat{y}|$, and $\hat{y} = \{z_i\}_{i=0}^{m}$ is the concatenated answer and explanation, denoted as $[\hat{a}; \hat{e}]$.

4 Experiments

4.1 Dataset and Setup

We follow (Kayser et al., 2021; Sammani et al., 2022; Suo et al., 2023) and consider two VL-NLE datasets. VQA-X (Park et al., 2018) builds upon VQAv2 dataset (Goyal et al., 2017). It is composed of 32.3K samples, divided into 29K for training, 1.4K for validation, and 1.9K for testing. e-SNLI-VE (Kayser et al., 2021) builds upon e-SNLI dataset (Camburu et al., 2018), consisting of 43K image-hypothesis pairs, divided into 40K for training, 1.4K for validation, and 1.6K for testing.

Rapper is consists of a rationale generator G and a reasoning module M, are both initialized from the pretrained image captioning model (Li et al., 2023a). The LLM for knowledge distillation during stage(A) is LLaMA-65B (Touvron et al., 2023). More implementation details are shown in Sec. A.1.

4.2 Evaluation Metrics

For NLE evaluation, we use BLEU@N (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE-L (Lin, 2004), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016) as the metrics, while using VQA accuracy to evaluate predicted answers. To evaluate the degree of plausibility and faithfulness of explanations, we measure them with CIDEr/SPICE and RefCLIPScore Hessel et al. (2021), respectively. In addition, we build human evaluation for explanation on plausibility and faithfulness since automatic metric measures not always reflect the correctness and logicality. Please refer to Appendix A.3 for the details of our human evaluation process.

Plausibility To quantitatively evaluate explanation plausibility, we employ CIDEr and SPICE scores. CIDEr measures the similarity between the generated explanation and human-written ground truth sentences, capturing human consensus by introducing tf-idf weight (Vedantam et al., 2015). On the other hand, SPICE converts sentences into semantic scene graphs, allowing evaluation to break grammatical constraints and thus closely resembling human judgment (Anderson et al., 2016).

Faithfulness We adopt RefCLIPScore, which computes the harmonic mean of CLIPScore (Hessel et al., 2021) and maximal reference cosine similarity, thereby encapsulating the correlation between the explanation and its reference. As noted by Hessel et al. (2021), RefCLIPScore surpasses prior metrics in correlating with human judgment for hallucination detection.

Method	VQA-X								
	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Accuracy
PJ-X (Park et al., 2018)	57.4	42.4	30.9	22.7	19.7	46.0	82.7	17.1	76.4
FME (Wu & Mooney, 2018b)	59.1	43.4	31.7	23.1	20.4	47.1	87.0	18.4	75.5
RVT (Marasović et al., 2020)	51.9	37.0	25.6	17.4	19.2	42.1	52.5	15.8	68.6
QA-only (Kayser et al., 2021)	51.0	36.4	25.3	17.3	18.6	41.9	49.9	14.9	-
e-UG (Kayser et al., 2021)	57.3	42.7	31.4	23.2	22.1	45.7	74.1	20.1	80.5
NLX-GPT (Sammani et al., 2022)	64.2	49.5	37.6	28.5	23.1	51.5	110.6	22.1	83.07
S3C (Suo et al., 2023)	64.7	50.5	38.8	30.7	23.9	52.1	116.7	23.0	85.6
Rapper (ours)	**65.5**	**51.6**	**40.5**	**31.8**	**24.3**	**52.9**	**124.0**	**24.5**	**87.25**

Method	e-SNLI-VE								
	B@1	B@2	B@3	B@4	METEOR	ROUGE-L	CIDEr	SPICE	Accuracy
PJ-X (Park et al., 2018)	29.4	18.0	11.3	7.3	14.7	28.6	72.5	24.3	69.2
FME (Wu & Mooney, 2018b)	30.6	19.2	12.4	8.2	15.6	29.9	83.6	26.9	73.7
RVT (Marasović et al., 2020)	29.9	19.8	13.6	9.6	18.8	27.3	81.7	32.5	72.0
QA-only (Kayser et al., 2021)	29.8	19.7	13.5	9.5	18.7	27.0	80.4	32.1	-
e-UG (Kayser et al., 2021)	30.1	19.9	13.7	9.6	19.6	27.8	85.9	34.5	**79.5**
NLX-GPT (Sammani et al., 2022)	37.0	25.3	17.9	12.9	18.8	34.2	117.4	33.6	73.91
Rapper (ours)	**40.5**	**28.1**	**20.2**	**14.7**	**20.8**	**35.9**	**128.6**	**34.9**	75.73

Table 1: Quantitative NLE comparisons of *filtered* results (i.e., NLE evaluation conditioned on correct answers) on VQA-X and e-SNLI-VE.

Method	Unfiltered					Filtered					Accuracy
	B@4	METEOR	ROUGE-L	CIDEr	SPICE	B@4	METEOR	ROUGE-L	CIDEr	SPICE	
Rapper	**30.0**	23.3	**51.3**	**116.0**	**23.2**	**31.8**	24.3	**52.9**	**124.0**	**24.5**	**87.3**
− RLNF	29.4	**23.6**	51.2	113.0	23.0	31.2	**24.5**	52.5	120.2	24.2	86.6
− RLNF − KD	27.1	21.8	49.7	103.2	20.7	29.3	23.0	51.6	112.1	22.3	85.0

Method	Unfiltered					Filtered					Accuracy
	B@4	METEOR	ROUGE-L	CIDEr	SPICE	B@4	METEOR	ROUGE-L	CIDEr	SPICE	
Rapper	**30.0**	**23.3**	**51.3**	**116.0**	**23.2**	**31.8**	**24.3**	**52.9**	**124.0**	**24.5**	**87.3**
Rapper w/o filtering	28.5	22.7	50.8	110.6	22.2	30.1	23.4	52.1	116.7	23.4	86.4

Table 2: Ablation studies of the proposed training schemes (up) and the filtering mechanism for knowledge distillation (bottom). We compare the performances in both filtered and unfiltered settings.

4.3 Quantitative Analysis

NLE evaluation. In Table 1, Table 5, and Table 6, we demonstrate that *Rapper* outperform previous state-of-the-art methods in NLE-related metrics on both VQA-X and e-SNLIV-VE datasets, with *filtered* and *unfiltered* settings. The *filtered* setting in Table 1 considers the explanations that are associated with correct answers. Conversely, the *unfiltered* setting in Table 5 and Table 6 in Appendix A.2 indicates evaluations of explanations without considering the correctness of the corresponding answers.

Method	RefCLIPScore(↑)
Much recent VL-NLE works	
NLX-GPT	64.06
S3C	65.09
Our stage-ablated approaches	
Rapper (w/o KD and w/o RLNF)	66.00
Rapper (w/o RLNF)	65.66
Rapper	**67.05**

Table 3: Faithfulness evaluation on the VQA-X dataset under filtered setting. Note that a higher RefCLIPScore indicates less hallucination.

Method	Plausibility (↑)	Faithfulness (↑)
NLX-GPT	0.771	0.795
S3C	0.797	0.811
Rapper	**0.845**	**0.859**

Table 4: Human evaluation on plausibility and faithfulness on VQA-X in the filtered setting.

Plausibility & faithfulness of NLE. We assess the plausibility and faithfulness in NLE through CIDEr/SPICE (in Table 1), RefCLIPScore (in Table 3), and human evaluation (in Table 4). In table 1, we demonstrate that *Rapper* outperforms previous state-of-the-art methods in NLG metrics on both VQA-X and e-SNLI-VE benchmarks, underscoring its superiority in generating plausible explanations. On the other hand, in table 3, *Rapper*'s superior RefCLIPScore indicates fewer hallucinations and increased faithfulness over other VQA-NLE works, although the RefCLIPScore of *Rapper* (w/o RLNF) is lower due to the hallucinations introduced by knowledge distillation from LLM. Nonetheless, *Rapper* still successfully reduce hallucination after the RLNF training. This demonstrates the effectiveness of our proposed RLNF to enable the model to generate faithful NLEs. Lastly, through human evaluation in Table 4, we provide further human-perceived evidence for the effectiveness of *Rapper* for improved NLE generation.

Multimodal Input / Methods		(a)	(b)	(c)
				
		(a) **Q**: Is the table cluttered? **GT A**: No **GT E**: There is only a single vase with flowers on it	**(b)** **Q**: Is this in an asian country? **GT A**: Yes **GT E**: there is an asian language used as text font in public	**(c)** **Q**: What kind of animal is this? **GT A**: Sheep **GT E**: The animal is covered in thick wool
NLX-GPT	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There are no objects in the table	There is a train on the tracks	It has a long face and long nose
S^3C	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There are only a few items on it	There is a train in the stations	It has a long snout and white fur
Rapper	R	The table is not cluttered because there is only one object on it	The presence of asian writing on the train suggests that it is in an asian country	A sheep is a type of animal that has wool on its body
	$\hat{A}$	No	Yes	Sheep
	$\hat{E}$	There is only one object on it	There is asian writing on the train	Its has wool on its body

Figure 3: Visualization of output answers and explanations predicted by different methods. Note that words in red denote hallucinated explanations, and those in orange denote implausible ones. Words in blue denote faithful and plausible explanations to the input image-question pair.

Ablation on the proposed stages. In top of Table 2, we evaluate our two-stage approach: (A) KD from LLM and (B) RLNF. Compared to the *Rapper* baseline without KD and RLNF, our method enhances explanation plausibility and faithfulness, highlighting the importance of both stages.

Ablation on the "Filter" mechanism. In bottom of Table 2, our filtering mechanism in knowledge distillation stage outperforms the baseline *Rapper* without filtering, by effectively removing overly redundant and noisy pseudo rationales that could impair model performance.

Ablation studies of derived rationales. In Fig. 4, we demonstrate that introducing two proposed stages improves the quality of derived rationales, benefiting the VQA performance of vision-language large model. Specifically, we test whether mPLUG-Owl (Ye et al., 2023) can answer accurately when given a pair (image, question, and $x \in (None, R', R)$), where $x = None$ indicates no rationales as input, $x = R'$ indicates the rationales are from *Rapper* with KD training, and $x = R$ indicates the rationales are from *Rapper* with both KD and RLNF training. Notably, we find that rationale quality improves progressively as we implement the stages we have proposed. This underscores the effectiveness of our designed stages in enhancing rationale quality.

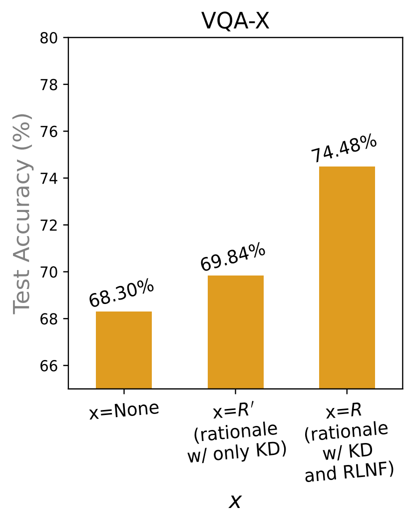

Figure 4: Ablation studies of derived rationales. Note the VQA accuracy on the VQA-X dataset is evaluated.

4.4 Qualitative Evaluation

In Fig.3, we compare NLX-GPT(Sammani et al., 2022), S3C (Suo et al., 2023), and our *Rapper* on the VQA-X dataset. *Rapper* consistently produces more plausible explanations. For example, Fig.3(a) highlights ability of *Rapper* to derive visual facts, such as identifying a single object on the table, surpassing previous methods that might produce hallucinated explanations. Similarly, in Fig.3(b), *Rapper* offers plausible explanations like recognizing Asian writing, contrasting with the implausible outputs of prior methods. Additional results and ablation studies are in Appendix A.4.

5 Conclusion

In this paper, we proposed *Rapper*, a two-stage Reinforced Rationale-Prompted Paradigm for enabling NLE with sufficient plausible and faithful properties. Our *Rapper* uniquely distills language-based knowledge from LLM and utilizes RL with natural language feedback from the VQA task, so that the designed rationale generator is able to produce rationales with the aforementioned desirable properties. By prompting such predicted rationales into the reasoning module, we demonstrated that satisfactory VQA performances can be achieved. Compared to SOTA VQA-NLE methods, possible implausible or hallucinated explanations can be mitigated by our *Rapper*.

Acknowledgments and Disclosure of Funding

This work is supported in part by the National Science and Technology Council via grant NSCT112-2634-F-002-007 and Center of Data Intelligence: Technologies, Applications, and Systems via grant NTU-113L900902. We also thank the National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

Afra Feyza Akyürek, Ekin Akyürek, Aman Madaan, Ashwin Kalyan, Peter Clark, Derry Wijaya, and Niket Tandon. Rl4f: Generating natural language feedback with reinforcement learning for repairing model outputs. *arXiv preprint arXiv:2305.08844*, 2023.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pp. 382–398. Springer, 2016.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31, 2018.

Soravit Changpinyo, Doron Kukliansky, Idan Szpektor, Xi Chen, Nan Ding, and Radu Soricut. All you may need for vqa are image captions. *arXiv preprint arXiv:2205.01883*, 2022.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*, pp. 104–120. Springer, 2020.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3369–3391, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. URL `https://aclanthology.org/2022.emnlp-main.222`.

Zhihao Fan, Zhongyu Wei, Siyuan Wang, Yang Liu, and Xuan-Jing Huang. A reinforcement learning framework for natural question generation using bi-discriminators. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1763–1774, 2018.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. e-vil: A dataset and benchmark for natural language explanations in vision-language tasks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1244–1254, 2021.

Daniel Khashabi, Yeganeh Kordi, and Hannaneh Hajishirzi. Unifiedqa-v2: Stronger generalization via broader cross-format training. *arXiv preprint arXiv:2202.12359*, 2022.

Daniel King, Zejiang Shen, Nishant Subramani, Daniel S Weld, Iz Beltagy, and Doug Downey. Don't say what you don't know: Improving the consistency of abstractive summarization by constraining beam search. *arXiv preprint arXiv:2203.08436*, 2022.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*, 2022.

Satyapriya Krishna, Jiaqi Ma, Dylan Slack, Asma Ghandeharioun, Sameer Singh, and Himabindu Lakkaraju. Post hoc explanations of language models can improve language models. *arXiv preprint arXiv:2305.11426*, 2023.

Zhibin Lan, Wei Li, Jinsong Su, Xinyan Xiao, Jiachen Liu, Wenhao Wu, and Yajuan Lyu. Factgen: Faithful text generation by factuality-aware pre-training and contrastive ranking fine-tuning. *Journal of Artificial Intelligence Research*, 76:1281–1303, 2023.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pp. 121–137. Springer, 2020.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. Prompting large language models for zero-shot domain adaptation in speech recognition. *arXiv preprint arXiv:2306.16007*, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zhihong Lin, Donghao Zhang, Qingyi Tao, Danli Shi, Gholamreza Haffari, Qi Wu, Mingguang He, and Zongyuan Ge. Medical visual question answering: A survey. *Artificial Intelligence in Medicine*, pp. 102611, 2023.

Feng Liu, Tao Xiang, Timothy M Hospedales, Wankou Yang, and Changyin Sun. ivqa: Inverse visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8611–8619, 2018.

Jiacheng Liu, Skyler Hallinan, Ximing Lu, Pengfei He, Sean Welleck, Hannaneh Hajishirzi, and Yejin Choi. Rainier: Reinforced knowledge introspector for commonsense question answering. *arXiv preprint arXiv:2210.03078*, 2022a.

Jiacheng Liu, Alisa Liu, Ximing Lu, Sean Welleck, Peter West, Ronan Le Bras, Yejin Choi, and Hannaneh Hajishirzi. Generated knowledge prompting for commonsense reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3154–3169, Dublin, Ireland, May 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.225. URL https://aclanthology.org/2022.acl-long.225.

Jiaying Lu, Xin Ye, Yi Ren, and Yezhou Yang. Good, better, best: Textual distractors generation for multiple-choice visual question answering via reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4921–4930, 2022a.

Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, and Yejin Choi. Quark: Controllable text generation with reinforced unlearning. *Advances in neural information processing systems*, 35:27591–27609, 2022b.

Bodhisattwa Prasad Majumder, Oana-Maria Camburu, Thomas Lukasiewicz, and Julian McAuley. Knowledge-grounded self-rationalization via extractive and natural language explanations. *arXiv preprint arXiv:2106.13876*, 2021.

Ana Marasović, Chandra Bhagavatula, Jae sung Park, Ronan Le Bras, Noah A. Smith, and Yejin Choi. Natural language rationales with full-stack visual reasoning: From pixels to semantic frames to commonsense graphs. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8779–8788, 2018.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7008–7024, 2017.

Fawaz Sammani, Tanmoy Mukherjee, and Nikos Deligiannis. Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8322–8332, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Oana-Maria Camburu, and Marek Rei. Logical reasoning for natural language inference using generated facts as atoms. *arXiv preprint arXiv:2305.13214*, 2023.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 217–223, 2017.

Wei Suo, Mengyang Sun, Weisong Liu, Yiqi Gao, Peng Wang, Yanning Zhang, and Qi Wu. S3c: Semi-supervised vqa natural language explanation via self-critical learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2646–2656, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. Trl: Transformer reinforcement learning. `https://github.com/lvwerra/trl`, 2020.

Alex Wang, Kyunghyun Cho, and Mike Lewis. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5008–5020, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.450. URL `https://aclanthology.org/2020.acl-main.450`.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*, 2022.

Jialin Wu and Raymond J. Mooney. Faithful multimodal explanation for visual question answering. *ArXiv*, abs/1809.02805, 2018a.

Jialin Wu and Raymond J Mooney. Faithful multimodal explanation for visual question answering. *arXiv preprint arXiv:1809.02805*, 2018b.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

Cheng-Fu Yang, Yao-Hung Hubert Tsai, Wan-Cyuan Fan, Russ R Salakhutdinov, Louis-Philippe Morency, and Frank Wang. Paraphrasing is all you need for novel object captioning. *Advances in Neural Information Processing Systems*, 35:6492–6504, 2022.

Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Ehsan Azarnasab, Faisal Ahmed, Zicheng Liu, Ce Liu, Michael Zeng, and Lijuan Wang. Mm-react: Prompting chatgpt for multimodal reasoning and action. *arXiv preprint arXiv:2303.11381*, 2023.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*, 2023.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.