Relevance-aware Multi-context Contrastive Decoding for Retrieval-augmented Visual Question Answering

Supplementary Material

Method	E-VQA	InfoSeek
Exclude	36.3	20.2
Mean	37.2	20.2
$-\infty$ *(Ours)*	**39.5**	**21.0**

Table 8. **Analysis on choices for $s_{c_{n+1}}$ of an empty context.**

Method	E-VQA	InfoSeek
RMCD (No Constraint)	37.2	12.5
RMCD *(Ours)*	**39.5**	**21.0**

Table 9. **Ablation study on ensembled plausibility constraint.**

Method	n	Exact Match
IC-RALM [49] (LLaMA-7B)	5	24.1
+ RMCD *(Ours)*	5	**26.6**

Table 10. **Results on Natural Questions [28] benchmark.**

A. Ablation study on context score of an empty context

In Tab. 8, analysis of choices for a context score $s_{c_{n+1}}$ of an empty context c_{n+1} is provided. Note that a result with 'empty context' c_{n+1} denotes a case where no additional context is fed into an LVLM (*i.e.* unconditional generation), and $s_{c_{n+1}}$ denotes the retrieval score corresponding to c_{n+1}. We experimented with three design choices to determine $s_{c_{n+1}}$. First, 'Exclude' denotes that an empty context is excluded from calculating the final logit (Eq.8 and Eq.10 of the main paper). Second, 'Mean' denotes that $s_{c_{n+1}}$ is set as the mean of s_{c_j} of other contexts. Finally, '$-\infty$' is where $s_{c_{n+1}}$ is set as $-\infty$, which is the choice we adopted in the main experiments. As reported in the table, we empirically found that setting $s_{c_{n+1}}$ to $-\infty$ worked the best, while other choices have also shown decent performances.

B. Ablation study on ensembled plausibility constraint

In Tab. 9, ablation results without the ensembled plausibility constraint are reported. The ensembled plausibility constraint elaborated through Eqs. 9 to 13 of the main paper is an additional constraint that complements RMCD by restricting an LVLM from sampling highly unlikely tokens, thereby ensuring better generation results. As reported, RMCD without the proposed plausibility constraint shows sub-optimal performances, showing the importance of implementing the proposed constraint. Note that Contrastive Decoding-based methods performing worse without proper plausibility constraints is a common phenomenon, where the same result is also reported in the original CD paper [34].

C. Extending RMCD to text-based QA tasks

In Tab. 10, we report the performance of RMCD on Natural Questions [28] benchmark, an open-domain question-answering benchmark purely based on natural languages. In detail, we apply RMCD to IC-RALM [49], an RAG pipeline based on Large Language Models (LLMs). The results show that replacing the decoding method of IC-RALM with RMCD results in an additional gain of 2.5 points in accuracy, showing that RMCD can be also extended to pure NLP tasks to improve existing RAG pipelines.

D. Construction of KBs for InfoSeek

Since the Wikipedia Knowledge Base (KB) used as a retrieval source in the original InfoSeek [8] paper is not provided by the authors, we construct the KB using the Wikipedia dump that the authors have provided. Similar to the original paper, we first filter out entities in the Wikipedia dump without images, resulting in a filtered dump with a size of 2M. Then, we first construct the smallest KB by selecting 1.7K entities related to samples in InfoSeek validation set from the filtered dump based on annotations. Note that the 1.7K KB is a default for every experiment in the paper, except for experiments in Fig. 4. We additionally construct 100K and 2M KBs to test the robustness of decoding methods against retrieval quality. After constructing the 1.7K KB, we randomly sample non-relevant entities from the remaining entities, which are added to the 1.7K KB forming a KB with a size of 100K. Also, we simply define the largest 2M KB by using every entity in the filtered dump. As the ratio of entities irrelevant to QA pairs increases in larger KB, retrieval gets more difficult, as shown in the decline of the retrieval Recall@k metric in Fig. 3 of the main paper.

E. Retrieval procedure

InfoSeek. For InfoSeek [8], we use the image corresponding to a QA pair as a query. A query feature is obtained with the image encoder of CLIP-ViT-B/32 [48]. Then, features

M	Encyclopedic-VQA
Non-RMCD Best	37.4
2.0	38.6
3.0	39.3
4.0	39.5
5.0	39.4

Table 11. **Sensitivity Analysis on M.**

m	Encyclopedic-VQA
Non-RMCD Best	37.4
-0.5	39.4
-1.0	39.5
-1.5	39.1
-2.0	38.9

Table 12. **Sensitivity Analysis on m.**

γ	Encyclopedic-VQA
Non-RMCD Best	37.4
0.05	39.3
0.1	39.2
0.3	39.5
0.5	39.0

Table 13. **Sensitivity Analysis on γ.**

β	Encyclopedic-VQA
Non-RMCD Best	37.4
0.1	39.4
0.15	39.3
0.2	39.5
0.3	39.1

Table 14. **Sensitivity Analysis on β.**

τ_1	Encyclopedic-VQA
Non-RMCD Best	37.4
1.0	37.7
2.0	39.0
3.0	39.5
4.0	39.3

Table 15. **Sensitivity Analysis on τ_1.**

τ_2	Encyclopedic-VQA
Non-RMCD Best	37.4
0.1	39.4
0.5	39.5
1.0	38.8
2.0	38.7

Table 16. **Sensitivity Analysis on τ_2.**

of every document in the KB are defined as a text feature of a summary corresponding to a document, obtained with a CLIP text encoder. Based on the obtained query and document features, the retrieval score between a query and a document is set as an inner product between the query feature and the document feature. Finally, we retrieve n documents with the highest retrieval scores, taking their summaries as retrieved contexts, where the retrieval scores s_{c_j} between a sample and a context c_j are set as retrieval scores used.

Encyclopedic-VQA. For Encyclopedic-VQA [45], we define the retrieval target documents using retrieval results with Google Lens [1] provided by authors, since CLIP-based retrieval is shown to be very challenging [45] on Encyclopedic-VQA. In detail, we adopt five documents with the highest Lens retrieval score as retrieval target documents. In the case where less than five Lens retrieval results exist, we additionally retrieve target documents with CLIP scores identical to the same process in InfoSeek until obtaining five target documents without duplicates. Documents retrieved with CLIP score are appended to Lens retrieval results. Since the document summary does not exist in KB for Encyclopedic-VQA, an additional step of determining passages in documents (*i.e.* paragraph) to retrieve is required. Among top-k documents having at least n passages in total, a retrieval score between a sample and a passage is defined with a BM25 similarity [50] between a question and the passage. Based on calculated BM25 similarities, we retrieve n passages with the highest similarity, where the retrieval score s_{c_j} between a sample and a context c_j is set as a BM25 similarity.

OK-VQA. For OK-VQA, we directly adopt the retrieval results used in the baselines [36, 37] where RMCD is implemented, without modification. Those retrieval results are obtained from Google Search Corpus [42] with retrievers proposed in each work.

F. Analysis on hyperparameters

Sensitivity analysis. From Tab. 11 to 16, we report performances on Encyclopedic-VQA under different hyperparameters. Overall, RMCD shows robust performance regardless of the choice of hyperparameter. Concretely, it consistently outperforms the second-best method ‘concat’, referred to as ‘Non-RMCD Best’ in tables, regardless of the hyperparameters. In terms of M and m controlling the maximum and minimum value of context weights, we found that larger values show better performance in general. Regarding τ_1 and τ_2, which control softmax temperature, results show that adopting $\tau_1 = 1$ or $\tau_2 = 1$ still outperforms the second-best performance. Such a result shows that adopting τ_1 or τ_2 is an optional choice for better performance, as setting a softmax temperature to 1 is equivalent to eliminating those hyperparameters. Finally, we observe that the choice of γ and β have minimal impact on the final performance.

Selection guideliens. Here, we provide guidelines for hyperparameter selection, although RMCD is generally robust to hyperparameters. For reflection strength M, smaller values are preferable under poor retrieval, thus we adopt relatively low $M = 3.0$ for OK-VQA with noisy web corpus. Setting m to overly small values (below -1.5) leads to suboptimal performance because it penalizes relevant contexts. For γ and β, performance remains stable across the ranges reported in Tab.13 and Tab.14. Softmax temperatures above 1.0 work well for γ_1, while values below 1.0 are better for γ_2.

G. Statistics of retrieval results

InfoSeek. In Tab. 18, retrieval results on InfoSeek validation set are reported. In detail, Recall@k, where $k = 1, 3, 5$ are reported in document-level. Since we construct the KB with three different sizes, results are reported on every KB.

Encyclopedic-VQA. We report the passage-level and document-level retrieval results on Encyclopedic-VQA in Recall@k (R@k), where $k = 1, 3, 5$. To be specific, only the retrieval results on the one-hop split are reported since two-hop questions require two different documents to answer the question, making it challenging to define the accurate Recall@k metric. Results are reported in Tab. 19.

H. Further implementation details

Encyclopedic-VQA preprocessing. We find that image URLs corresponding to some samples in the test set of Encyclopedic-VQA [45] are not accessible due to their expiration. Therefore, we excluded 38 samples from the test set whose image was not accessible. Every experimental result in this paper is reported with such samples excluded.

Details about baseline LVLMs. As baseline LVLMs for InfoSeek and Encyclopedic-VQA benchmarks, we adopt InternVL-2.5 [9], BLIP-2 [33] and LLaVA-1.5 [41]. We use the official pre-trained weights and implementations of InternVL-2.5 [9] by the by the Hugging Face transformers library [55], BLIP-2 by the LAVIS library [32], and LLaVA-1.5 by the Hugging Face transformers library [55]. As an image encoder of BLIP-2, ViT-g/14 variant of EVA-CLIP [16] is adopted. As a base Language Model (LM), OPT-6.7B [59], T5-XL, and T5-XXL [12] are used for BLIP-2. The input image size for BLIP-2 is set to 224×224 pixels, and the model is loaded in BF16 precision. As an image encoder of LLaVA-1.5, CLIP-ViT-L-336px [48] is adopted. As a base LM, Vicuna-13B [10] is utilized. The input image size for LLaVA-1.5 is set to 336×336 pixels. The model is loaded in BF16 precision, except for the Vicuna-13B language model, which uses 4-bit quantization [14]. As baseline LVLMs for OK-VQA, we use the pre-trained models provided by the authors [36, 37], where the model architectures are based on T5-XL and BLIP-2 with T5-XL as base LM, respectively. Nucleus sampling [19] with $p = 0.9$ and a temperature of 1.0 is applied for sampling.

Input prompt. In Tab. 17, prompt templates for unconditional decoding and retrieval-based methods (*i.e.* RAG, SCD, consistency, concat, max probability, and RMCD) are provided. For the unconditional generation, we adopt default prompts for question-answering tasks from the original implementations of BLIP-2 [33] and LLaVA-1.5 [41]. For retrieval-based methods, we simply append the retrieved context after the question. As an input prompt of RA-VQA [36] and FLMR [37] for the OK-VQA benchmark, we directly adopt the baselines' prompts for retrieval-based generation.

Evaluation metrics. In **InfoSeek**, the total accuracy is defined as the harmonic mean of accuracy on two validation splits, UNSEEN QUESTION and UNSEEN ENTITY. UNSEEN QUESTION consists of the QA pairs that do not exist in InfoSeek train set. UNSEEN ENTITY consists of QA pairs where the entity corresponding to the QA pair does not exist in the InfoSeek train set. Since RMCD is a training-free method, the two divisions are not significantly important. Still, we report performance on both splits for reference. Accuracy on each split is the average accuracy of three question types: STRING, NUMERICAL, and TIME. For STRING and TIME question types which contain multiple possible answers, evaluation is done following conventional VQA practices [17]. For NUMERICAL questions asking for detailed numbers, the Related Accuracy from [44] is utilized, which allows an error within a 10% tolerance range.

In **Encyclopedic-VQA** [45], questions are classified into three types: one-hop, multi-answer, and two-hop. While the original paper mainly reported performance on one-hop questions in the majority of experiments, we additionally reported the performance on every question as a reference. Each prediction in Encyclopedic-VQA except questions of the 'multi-answer' type is evaluated with BERT Matching (BEM) [5], which employs BERT [15] to classify whether the prediction is correct based on the given answer. In detail, if the BEM score between a prediction and the answer is over 0.5, the prediction is considered correct. For questions of the 'multi-answer' type, model predictions are converted into a set of strings, and the intersection-over-union (IoU) between the prediction set and answer set is calculated. If the IoU is over 0.5, the prediction is considered correct. Otherwise, the BEM score between concatenated predictions and concatenated answers is used to evaluate predictions.

In **OK-VQA** [43], we directly follow the evaluation protocol of baselines [36, 37] where the VQA score [43] is adopted as an accuracy metric.

I. Further qualitative examples

We provide further qualitative examples in Fig. 7. Examples demonstrate that RMCD effectively reflects relevant contexts while also deflecting irrelevant contexts, therefore generating better results compared to other decoding methods. For instance, RMCD successfully reflects evidence in relevant contexts ("Rocks from Moon & Mars", "located on both banks") which other decoding methods fail. Also, unlike other methods affected by wrong evidence existent in multiple irrelevant contexts thereby generating wrong answers ("apophyllite", "footbridge"), RMCD successfully deflects irrelevant contexts and thus opposes their effects.

Model	Method	Template
BLIP-2 [33]	Unconditional	"Question: <question>, Short answer:"
	Retrieval-based	"Question: <question>, Context: <context> Short answer:"
LLaVA-1.5 [40]	Unconditional	"<question> Answer the question using a single word or phrase."
	Retrieval-based	"<question> Answer the question using a single word or phrase. Context: <context>"
RA-VQA [36]	Retrieval-based	"<question> <caption> <objects> <context>"
FLMR [37]	Retrieval-based	"Question: <question> Caption: <caption> Object: <objects> Knowledge: <context> Answer:"

Table 17. **Prompt templates.** <question> and <context> in templates are replaced with actual questions and contexts, respectively. <caption> and <object> are replaced with image captioning results and detected objects in [36, 37].

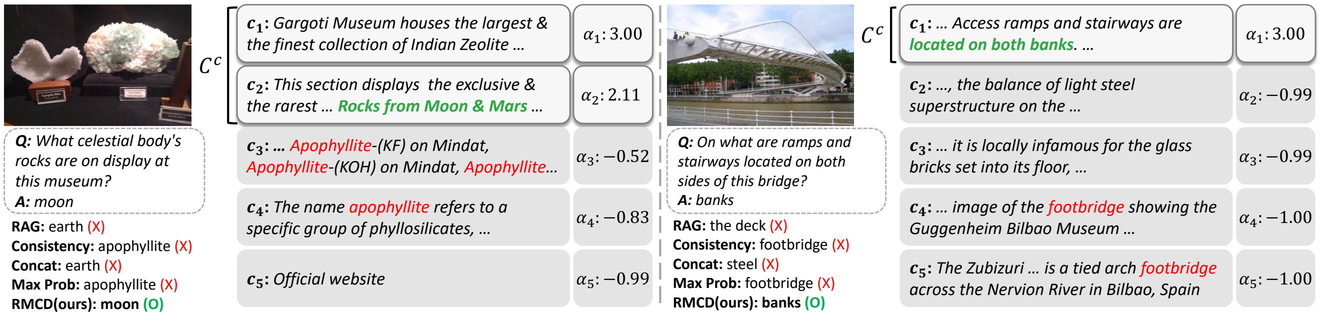
C^c
c_1: Gargoti Museum houses the largest & the finest collection of Indian Zeolite …
α_1: 3.00
c_2: This section displays the exclusive & the rarest … Rocks from Moon & Mars …
α_2: 2.11
c_3: … Apophyllite-(KF) on Mindat, Apophyllite-(KOH) on Mindat, Apophyllite…
α_3: −0.52
c_4: The name apophyllite refers to a specific group of phyllosilicates, …
α_4: −0.83
c_5: Official website
α_5: −0.99
Q: What celestial body's rocks are on display at this museum?
A: moon
RAG: earth (X)
Consistency: apophyllite (X)
Concat: earth (X)
Max Prob: apophyllite (X)
RMCD(ours): moon (O)
Q: On what are ramps and stairways located on both sides of this bridge?
A: banks
RAG: the deck (X)
Consistency: footbridge (X)
Concat: steel (X)
Max Prob: footbridge (X)
RMCD(ours): banks (O)
C^c
c_1: … Access ramps and stairways are located on both banks. …
α_1: 3.00
c_2: …, the balance of light steel superstructure on the …
α_2: −0.99
c_3: … it is locally infamous for the glass bricks set into its floor, …
α_3: −0.99
c_4: … image of the footbridge showing the Guggenheim Bilbao Museum …
α_4: −1.00
c_5: The Zubizuri … is a tied arch footbridge across the Nervion River in Bilbao, Spain
α_5: −1.00

Figure 7. **Further Qualitative Examples.** Predictions of each method are illustrated along with retrieved contexts $c_j \in \mathcal{C}$, their corresponding context weights α_j, and the constraint context set $\mathcal{C}^c$. Relevant and irrelevant information in contexts are marked green and red, respectively.

KB Size	Recall@1	Recall@3	Recall@5
1.7K	45.3	63.5	70.5
100K	19.5	30.6	36.1
2M	6.1	11.7	14.8

Table 18. **Retrieval results on InfoSeek validation set.**

Retrieval Target	Recall@1	Recall@3	Recall@5
Document	50.2	61.8	65.6
Passage	26.6	39.8	45.8

Table 19. **Document-level and passage-level retrieval results on Encyclopedic-VQA test set.**