VOLCANO: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision

Seongyun Lee[1] **Sue Hyun Park**[1] **Yongrae Jo**[2] **Minjoon Seo**[1]

[1]KAIST AI [2]LG AI Research

{seongyun, suehyunpark, minjoon}@kaist.ac.kr yongrae.jo@lgresearch.ai

Abstract

Large multimodal models suffer from multimodal hallucination, where they provide incorrect responses misaligned with the given visual information. Recent works have conjectured that one of the reasons behind multimodal hallucination is due to the vision encoder failing to ground on the image properly. To mitigate this issue, we propose a novel approach that leverages self-feedback as visual cues. Building on this approach, we introduce **VOLCANO**, a multimodal self-feedback guided revision model. VOLCANO generates natural language feedback to its initial response based on the provided visual information and utilizes this feedback to self-revise its initial response. VOLCANO effectively reduces multimodal hallucination and achieves state-of-the-art on MMHal-Bench, POPE, and GAVIE. It also improves on general multimodal abilities and outperforms previous models on MM-Vet and MMBench. Through qualitative analysis, we show that VOLCANO's feedback is properly grounded on the image than the initial response. This indicates that VOLCANO can provide itself with richer visual information through feedback generation, leading to self-correct hallucinations. We publicly release our model, data, and code at github.com/kaistAI/Volcano.

1 Introduction

Recent large multimodal models (LMMs) use substantial image-text or video-text pairs to help instruct-tuned large language models (LLMs) comprehend visual features produced by vision encoders (Alayrac et al., 2022; Liu et al., 2023b,c; Chen et al., 2023; Peng et al., 2023; Dai et al., 2023; Zhu et al., 2023; Ye et al., 2023a; Li et al., 2023a; Zhang et al., 2023b; Su et al., 2023; Maaz et al., 2023). With the introduction of fine-tuning methods such as visual instruction tuning (Liu et al., 2023b,c), LMMs are now evolving into assistants capable of understanding the world through multiple channels, akin to humans.

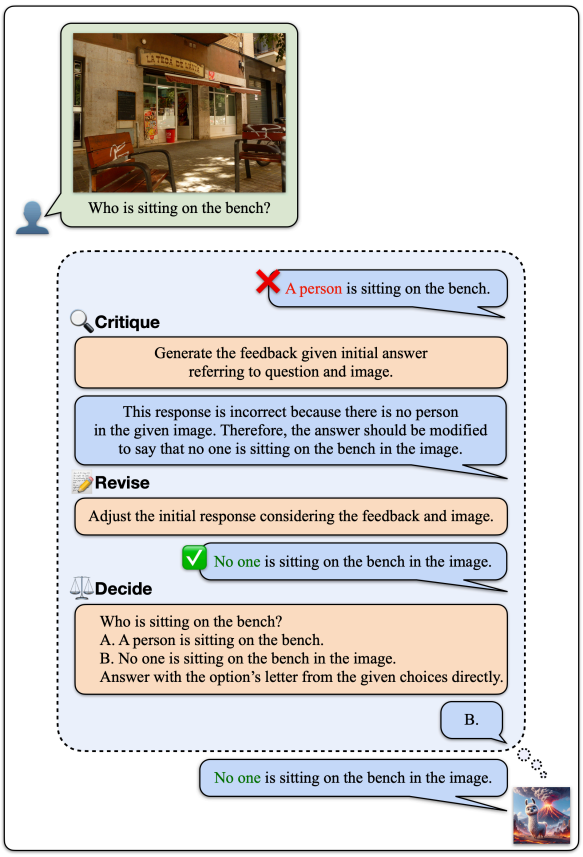

Figure 1: Overview of VOLCANO. This example illustrates the process undertaken by VOLCANO for a question in the MMHal-Bench dataset. Before giving the response, VOLCANO goes through a *critique-revise-decide* process. It critiques its initial response with natural language feedback, revises the response based on the feedback, and decides whether to accept the revised answer.

Despite the impressive performance observed on various benchmark tasks and qualitative outcomes, these models grapple with an issue called *multimodal hallucination*, where they produce responses that do not align with the visual information given in the question. Recent work (Zhai et al., 2023) demonstrates that multimodal hallucinations can occur when the vision encoder fails to ground im-

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 391–404
June 16-21, 2024

ages accurately. In other words, LMMs tend to rely more on their parametric knowledge than on provided visual features, causing them to guess and generate hallucinations. Wang et al. (2023b) empirically shows that models attend to the previous tokens more than image features as they generate tokens misaligned with the given image.

In this paper, we propose a novel method that utilizes natural language feedback to enable the model to correct hallucinated responses by providing detailed visual information. Building on this method, we introduce **VOLCANO**[1], a multimodal self-feedback guided revision model. VOLCANO is trained to first generate an initial response based on the given image and question, then sequentially revises the response until it determines that no more improvement is required. We collect our training data for multimodal feedback and revision using proprietary LLMs.

To verify the efficacy of VOLCANO in reducing multimodal hallucination, we evaluate its performance on multimodal hallucination benchmarks (Sun et al., 2023; Li et al., 2023d; Liu et al., 2023a). Results demonstrate consistent performance improvements across all benchmarks. Notably, when compared to previous methods specialized in mitigating multimodal hallucination (Zhou et al., 2023; Sun et al., 2023; Yin et al., 2023), VOLCANO showcases a 24.9% enhancement, underscoring its effectiveness in addressing the challenge. Further, on multimodal understanding benchmarks (Liu et al., 2023e; Yu et al., 2023), it is also shown effective in understanding and reasoning about visual concepts.

Through qualitative analysis, we find that the generated feedback attends to the image with higher intensity and higher coverage of features in the image. These findings explain that feedback carries fine-grained visual information. Even if the vision encoder fails to properly ground, the feedback can still guide the LLM to improve upon the hallucinated response, supporting the role of feedback in our proposed method.

Our contributions are summarized as follows:

1. We introduce VOLCANO, a self-feedback guided revision model that effectively mitigates multimodal hallucination. It achieves state-of-the-art performance on multimodal hallucination benchmarks and multimodal understanding benchmarks.

2. Our qualitative analysis shows that VOLCANO's feedback is rooted in the image, conveying rich visual details. This illustrates that feedback can offer guidance to reduce multimodal hallucination, even if the vision encoder imprecisely encodes the image and the model misinterprets the image initially.

3. We open-source VOLCANO (7B & 13B), along with data and code for training and inference.

2 Related work

2.1 Multimodal hallucination

Unlike language hallucination, where fabrication of unverifiable information is common (Ji et al., 2023; Zhang et al., 2023c; Li et al., 2023c), multimodal hallucination typically involves verifiable information misaligned with the input visual content. This phenomenon has been predominantly explored in the context of object hallucination, where generated content includes objects that are inconsistent with or absent from the target image (Rohrbach et al., 2018; Biten et al., 2022; Li et al., 2023d; Liu et al., 2023a; Zhai et al., 2023). More complex forms of multimodal hallucination, such as holistic misrepresentations involving entire scenes or environments, have only begun to be recognized and documented in recent studies (Sun et al., 2023).

To uncover the cause of failure in grounding, previous works analyze either the visual or language side. Zhai et al. (2023) pinpoints the lack of preciseness in visual features produced by the vision encoder. Other studies (Li et al., 2023d; Liu et al., 2023a; Wang et al., 2023b) focus on the tendency of LLMs to generate words more in line with common language patterns rather than the actual visual content. The error may be further exacerbated by autoregressive text generation (Rohrbach et al., 2018; Zhang et al., 2023a; Zhou et al., 2023).

2.2 Self-correcting from feedback

Learning from feedback can align LLMs to desired outcomes, to better follow instructions via human preference feedback (Ouyang et al., 2022), preference feedback generated by AI itself (Lee et al., 2023; Dubois et al., 2023), or even fine-grained feedback (Wu et al., 2023; Lightman et al., 2023). Compared to preference and fine-grained feedback which provide scalar values as training signals, natural language feedback provides more information (Scheurer et al., 2022; Ma et al., 2023) and has

[1]We call our model VOLCANO because it frequently erupts *LLaVA*

been effective for language models to correct outputs, especially for *self-correction* (Welleck et al., 2022; Pan et al., 2023). Inspired by successful iterative self-refining language models (Madaan et al., 2023; Ye et al., 2023b; Shinn et al., 2023; Gou et al., 2024), to the best of our knowledge, we are the first to achieve improvement in multimodal models through self-feedback guided refinement.

2.3 Mitigating multimodal hallucination

Previous methods for mitigating multimodal hallucinations have varied in their focus, including enhancing the quality of instruction tuning data, model training methodologies, and implementing post-hoc refinements. LRV-Instruction dataset (Liu et al., 2023a) ensures the balance of both negative and positive instructions and VIGC (Wang et al., 2023a) iteratively generates and corrects instructions to reduce hallucinated samples in training data. Adapting reinforcement learning from human feedback (RLHF) to train a single reward model as in LLaVA-RLHF (Sun et al., 2023) or training multiple or even without no reward models as in FDPO (Gunjal et al., 2023) has proven effective as well. LURE (Zhou et al., 2023) trains a revision model to detect and correct hallucinated objects in the base model's response. Woodpecker (Yin et al., 2023) breaks down the revision process into multiple subtasks where three pre-trained models apart from the base LMM are employed for the subtasks.

Unlike models using reinforcement learning, our approach does not require reward model training. Also, contrary to revision-only methods, our method trains a model to *self*-revise, eliminating the need for extra modules. Furthermore, we introduce natural language feedback before the revision process. This feedback serves a dual purpose: it revisits the visual features for enhanced clarity and specifically pinpoints the hallucinated elements that require correction, thereby enriching the information available for more effective revision.

3 VOLCANO

VOLCANO is a single LMM to generate initial responses, feedback, and revisions, as well as decisions to accept revisions. It follows a sequential procedure of an iterative critique-revision-decide loop. In Section 3.1, we introduce the process by which VOLCANO self-revises its responses iteratively. Section 3.2 describes the collection of multimodal feedback and revision data used to train VOLCANO. Finally, Section 3.3 provides detailed information about the models and data used in our study. The overall process is explained in Algorithm 1 and illustrated in Figure 2.

Algorithm 1 Feedback guided self-revision

Input: *model* M, *image* I, *question* Q
$R_{initial} = M(I, Q)$
$R_{best} = R_{initial}$
for up to 3 iterations **do**
  $F = M(I, Q, R_{best})$
  $R_{revised} = M(I, Q, R_{best}, F)$
  $R_{decided} = M(I, Q, R_{best}, R_{revised})$
  if $R_{decided} == R_{best}$ **then**
    break
  else
    $R_{best} = R_{revised}$
return R_{best}

3.1 Iterative self-revision

VOLCANO is a single model that generates improved responses through a sequential process of four stages. First, similar to other LMMs, it generates an initial response $R_{initial}$ for the image I and question Q and initializes the best response R_{best} with $R_{initial}$. This stage is performed only once in the process of creating the final response. Second, it generates feedback F based on the R_{best} (**stage 1**). Using this feedback, it self-revises the R_{best} (**stage 2**). Since there is no guarantee that the revised response $R_{revised}$ will be better than the existing R_{best}, there is a need to determine which response is better for the given Q and I. At this point, VOLCANO is given Q, I, and both responses, and it goes through the process of deciding which response is better (**stage 3**). The order of $R_{revised}$ and R_{best} in stage 3 is randomized to prevent the positions from affecting the results (Wang et al., 2023c). If the model decides that $R_{revised}$ is better than R_{best}, then R_{best} is updated with $R_{revised}$ and the procedure from stage 1 to stage 3 is repeated, with the predetermined maximum number of iterations. Otherwise, the loop is early-stopped, and R_{best} is selected as the final output. The prompts for inference at each stage are in Appendix B.1.

3.2 Data collection

To train VOLCANO, we collect initial responses for visual questions from an open-source LMM and generate feedback and revisions using a proprietary LLM as shown in Figure 3 (Akyürek et al., 2023;

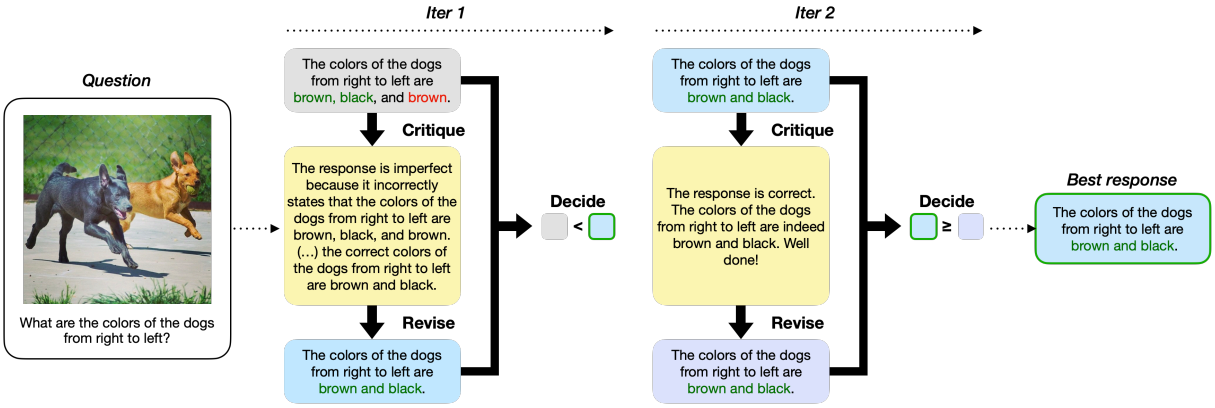

Figure 2: Overall process of VOLCANO. VOLCANO is a multimodal self-feedback guided revision model that takes an image and a question and then generates an improved response based on the self-feedback.

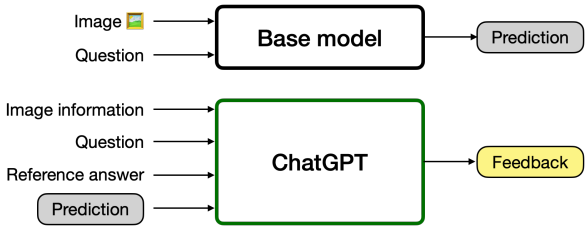

Figure 3: Data collection.

Madaan et al., 2023; Ye et al., 2023b; Wang et al., 2023d; Kim et al., 2023).

Since current proprietary LLMs cannot process images, we provide object details in text and image captions as a proxy for images. For each data instance, we feed the proprietary LLM image information consisting of object details and captions, question, initial response, and gold answer as reference answer, allowing the model to evaluate the given inputs and produce feedback.

The proprietary LLM might exploit the gold answer to generate the feedback, which can cause potential inaccuracies in feedback during inference time when it is not provided. To avoid this, we give the LLM clear prompts to focus on the text-formatted image details when generating feedback. When constructing the revision data, we set up a system to predict the existing gold answer as the output, using the feedback data, image, question, and initial response obtained from the previous steps as input, without involving any separate model generation process. The prompts for data collection are in Appendix B.2.

3.3 Implementation details

Data To construct multimodal feedback and revision data, we utilize the LLaVA-SFT-127k dataset (Sun et al., 2023). We only use the first turn of each instance in the dataset. When fine-tuning VOLCANO, we use the llava-1.5-mix665k as the visual instruction dataset (Liu et al., 2023b).

Model For the proprietary LLM, we employ OpenAI's gpt-3.5-turbo (OpenAI, 2022). We use the LLaVA-SFT+ 7B model to generate the initial response when creating feedback data and LLaVA-1.5 7B and 13B as backbone models of VOLCANO (Liu et al., 2023b,c). Details of computation and hyperparameters used are in Appendix C and Appendix D, respectively.

4 Experiments

4.1 Benchmarks

Multimodal hallucination benchmarks We use POPE (Li et al., 2023d), GAVIE (Liu et al., 2023a), and MMHal-Bench (Sun et al., 2023) as our benchmarks to test multimodal hallucination mitigation performance. POPE and GAVIE are benchmarks for assessing object-level hallucinations in images. POPE comprises 9k questions asking if a specific object is present or not in an image. GAVIE is composed of 1k questions evaluating how accurately the response describes the image (accuracy) and how well the response follows instructions (relevancy) using GPT-4. MMHal-Bench aims to evaluate the overall hallucination of LMMs, consisting of realistic open-ended questions. It comprises 96 image-question pairs across 8 question categories and 12 object topics. The overall score is computed by GPT-4, which compares the model's response to the correct answer based on the given object information. If the overall score is less than 3, the response is considered to contain hallucinations.

Model	MMHal-Bench		POPE		GAVIE		
	Score ↑	Hal rate ↓	Acc ↑	F1 ↑	Acc score ↑	Rel score ↑	Avg score ↑
MiniGPT-4 7B	-	-	68.4	74.5	4.14	5.81	4.98
mPLUG-Owl 7B	-	-	51.3	67.2	4.84	6.35	5.6
InstructBLIP 7B	2.1	0.58	71.5	80.0	5.93	7.34	6.64
LLaVA-SFT+ 7B	1.76	0.67	81.6	82.7	5.95	8.16	7.06
LLaVA-RLHF 7B	2.05	0.68	81.8	81.5	6.01	8.11	7.06
LLaVA-SFT+ 13B	2.43	0.55	83.2	82.8	5.95	8.2	7.09
LLaVA-RLHF 13B	2.53	0.57	83.1	81.9	6.46	8.22	7.34
LLaVA-1.5 7B	2.42	0.55	86.1	85.1	6.42	8.2	7.31
LLaVA-1.5 13B	2.54	0.52	86.2	85.2	6.8	8.47	7.64
VOLCANO 7B	2.6	0.49	88.2	**87.7**	6.52	8.4	7.46
VOLCANO 13B	**2.64**	**0.48**	**88.3**	**87.7**	**6.94**	**8.72**	**7.83**

Table 1: Results on multimodal hallucination benchmarks. The MMHal-Bench score is measured on a 0-5 scale. Hallucination rate (Hal rate) is measured as the proportion of scores less than 3. Additionally, GAVIE's Acc score (Accuracy score) and Rel score (Relevancy score) are measured on a 0-10 scale, with Avg score representing the average of Acc and Rel scores. Detailed evaluation results for each benchmark by question type are in Table 6 and Table 7.

Multimodal understanding benchmarks We use MM-Vet (Yu et al., 2023) and MMBench (Liu et al., 2023e) as benchmarks to measure the general multimodal performance of LMMs. MM-Vet is a benchmark consisting of 16 tasks and 218 instances designed to evaluate LMM's ability in complex multimodal tasks. The score is measured by GPT-4, which compares the LMM's response to the gold answer. MMBench comprises 4,377 multiple-choice questions aimed at assessing visual perception and visual reasoning. We utilize the development set of MMBench in this study.

4.2 Baselines

We use Openflamingo (Awadalla et al., 2023), MiniGPT-4 (Zhu et al., 2023), mPLUG-Owl (Ye et al., 2023a), InstructBLIP (Dai et al., 2023), Otter (Li et al., 2023a), LLaVA-SFT+ (Sun et al., 2023), and LLaVA-RLHF (Sun et al., 2023) as baseline models. As multimodal hallucination corrector baselines, we employ LURE (Zhou et al., 2023) and Woodpecker (Yin et al., 2023). LURE utilizes MiniGPT-4 13B as its backbone model. Woodpecker uses gpt-3.5-turbo as its corrector, grounding DINO (Liu et al., 2023d) as its object detector and BLIP-2-FlanT5-XXL (Li et al., 2023b) for its VQA model.

4.3 Main results

VOLCANO achieves the best performance in multimodal hallucination benchmarks. As shown in Table 1, VOLCANO consistently outperforms the base model, LLaVA-1.5 and other existing LMMs in the multimodal hallucination benchmark. It shows strong performance in benchmarks that measure scores using proprietary LLMs (MMHal-Bench, GAVIE) and a benchmark using conventional metrics like accuracy and F1 score (POPE). Notably, results from GAVIE demonstrate that VOLCANO not only provides accurate answers for a given image but also enhances its ability to follow instructions. Full results are in Appendix A.1.

Model	MMHal-Bench	
	Score ↑	Hal rate ↓
LURE	1.9	0.58
Woodpecker	1.98	0.54
VOLCANO 7B	**2.6**	**0.49**
LLaVA-RLHF 7B	2.05	0.68
VOLCANO$^-$ 7B	**2.19**	**0.59**

Table 2: Performance comparison with recent methods focusing on reducing multimodal hallucination. VOLCANO$^-$ 7B is a model fine-tuned with our multimodal feedback and revision data on LLaVA-SFT+ 7B, which is the backbone model of LLAVA-RLHF 7B.

Natural language self-feedback is effective in revising responses. Table 2 shows VOLCANO's effectiveness by comparing it with previous models designed to tackle multimodal hallucination. Compared to LURE and Woodpecker, both of which revise responses without feedback, VOLCANO reduces hallucination better. This suggests that providing specific feedback is crucial for correcting multimodal hallucination. In addition, unlike the two methods that require a separate model specialized for revision, VOLCANO efficiently gives better responses using just one model. Another notable

Model	MMBench Acc ↑	MM-Vet Acc ↑
Openflamingo 9B	6.6	24.8
MiniGPT-4 13B	24.3	24.4
InstructBLIP 14B	36.0	25.6
Otter 9B	51.4	24.7
LLaVA-SFT+ 7B	52.7	30.4
LLaVA-RLHF 7B	52.7	29.8
LLaVA-SFT+ 13B	59.6	36.1
LLaVA-RLHF 13B	59.6	36.4
LLaVA-1.5 7B	59.9	31.2
LLaVA-1.5 13B	67.7	36.1
VOLCANO 7B	62.3	32.0
VOLCANO 13B	**69.4**	**38.0**

Table 3: Results on multimodal understanding benchmarks. The detailed evaluation results for each benchmark by question type are in Table 8 and Table 9.

Model	MMHal-Bench Score ↑	MMHal-Bench Hal rate ↓
Only prediction	2.45	0.52
No decision	2.33	0.56
VOLCANO 7B	**2.6**	**0.49**

Table 4: Module ablation results. The "Only prediction" is the result of performing only stage 1 for VOLCANO 7B. "No decision" is the outcome of completing stages 1 and 2.

Model	MMHal-Bench Score ↑	MMHal-Bench Hal rate ↓
Iter 1	2.54	0.51
Iter 2	2.58	0.5
Iter 3 (VOLCANO 7B)	**2.6**	**0.49**

Table 5: Iteration ablation results.

observation is that Woodpecker's improvement in hallucination is less significant compared to VOLCANO, despite converting visual information into text and feeding it to a proprietary LLM corrector. From this, we find that for reducing multimodal hallucination, conveying visual features directly to the corrector model is critical.

Compared to LLaVA-RLHF, which reduces LLM hallucination using RLHF, VOLCANO consistently performs better as well. For a fair comparison, we developed VOLCANO$^-$ 7B by fine-tuning the base model of LLaVA-RLHF 7B, LLaVA-SFT+ 7B, on our multimodal feedback and revision data. The results indicate that providing feedback in the form of natural language feedback, which the model can directly interpret, is more effective than providing feedback as scalar values.

VOLCANO showcases high general multimodal understanding capabilities. As the tendency of hallucination decreases, it is expected that the LMM can answer user questions about images more accurately. In this sense, we anticipate that VOLCANO would score high in benchmarks measuring general LMM's performance. To demonstrate this, we evaluate VOLCANO on benchmarks assessing complicated visual reasoning and perception capabilities of LLMs (Table 3). VOLCANO achieves superior performance compared to existing LMMs. Notably, as shown in Table 8, when measuring the math score related to a model's arithmetic capability, VOLCANO 13B impressively scored about twice as high as LLaVA-1.5 13B. Full results are in Appendix A.2.

4.4 Ablation studies

Module ablation We test the influence of each stage in the critique-revise-decide process (Section 3.1) in reducing multimodal hallucination. As shown in Table 4, when we only use the initial response as the final response and skip iterative self-revision, it scores lower than going through both processes. Surprisingly, even after just completing stage 1 and without self-revision, it still scores higher than the base model LLaVA-1.5 7B. This shows that merely fine-tuning with multimodal feedback and revision data can effectively reduce the hallucination rate. We observe a decrease in performance when the revised response is given as the final output without executing stage 3, compared to when a decision is made. This highlights the role of stage 3 in decreasing hallucination as it can prevent unnecessary revisions. This also suggests that while it is hard for the model to produce the right answer initially, distinguishing between right and wrong answers is relatively easier.

Iteration ablation We test how changing the maximum number of iterations affects VOLCANO's performance. As shown in Table 5, as the maximum iteration count increased, the hallucination rate decreased. This indicates that answers are successfully refined through multiple revisions. However, there also exists a trade-off: as the iteration count goes up, the inference time also increases.

5 Qualitative analysis

We qualitatively analyze how feedback from VOLCANO is effective in reducing multimodal hallu-

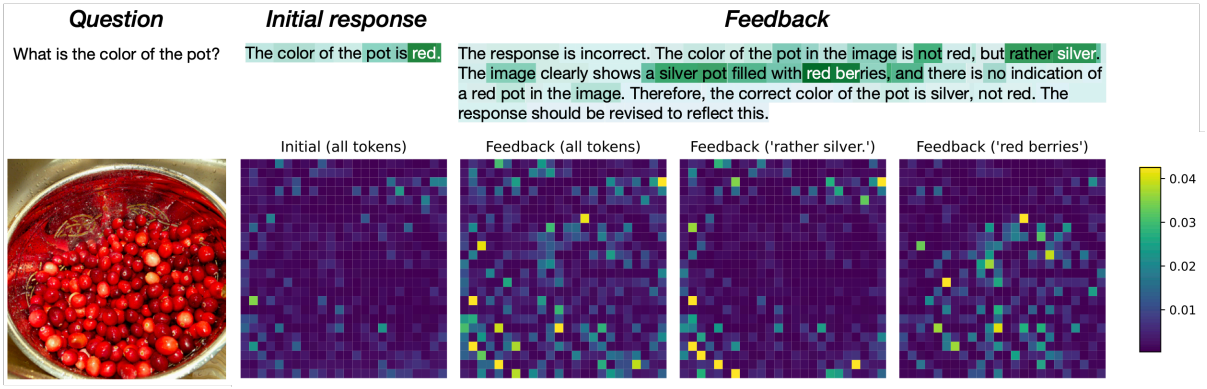

Figure 4: Coverage of image features attended during initial response and feedback generation on a single MMHal-Bench instance. The image attention heatmaps depict how the model's attention is distributed across image features, considering either all tokens or a subset of tokens in the output. In the text attention heatmaps above, the intensity of each token's background indicates the attention weight magnitude to image features, with darker highlights signifying higher weights. In the image attention heatmaps below, outliers at or above the 0.995th quantile are shown with the highest color intensity.

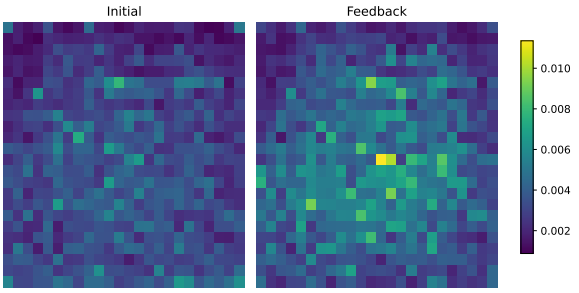

Figure 5: Average amount of attention to image features during the initial response (left) and feedback (right) generation. Attention weights are averaged across instances in MMHal-Bench where VOLCANO's revision enhances the initial response.

cination. In this section, we examine VOLCANO 7B results on MMHal-Bench in which the model's revised answer is chosen as the final answer. We compare the visual information content between the model's initial response and feedback, focusing on amount (5.1) and coverage (5.2).

5.1 Amount of visual information

Upon manual inspection of the instances, we observe that the initial response often correctly identifies object-level information but frequently misinterprets finer details such as object attributes or relationships between objects. On the contrary, we discover that the feedback text tends to describe the image contents more comprehensively.

To explore this phenomenon, we take inspiration from Wang et al. (2023b) and visualize how *much* do output tokens attend to input image features[2] while generating initial response and feedback tokens to the same image. For each instance, we perform top-k mean pooling to aggregate attention weights of initial response or feedback tokens on each image feature.[3] Specifically, we average top-3 attention weights across hidden layers, average top-3 weights across self-attention heads, and then average top-l weights across output tokens, where l is the length of the shorter output between initial response and feedback.

The results averaged across all instances are shown in Figure 5. Image features are more strongly attended by feedback than by initial response. Interestingly, even though attention to input would be more dispersed when generating feedback as its input includes an initial response in addition to the question, an increased concentration on wider areas of image features is visible. This suggests that visual information is largely reflected in the feedback text, supporting our manual inspection beforehand.

5.2 Coverage of visual information

We further empirically investigate how attention from individual tokens contributes to the *coverage* of critical visual information. Building upon the visualization method described in Section 5.1, we

[2]For every image, the vision encoder of VOLCANO, CLIP ViT-L/14 336px, processes it into 336px×336px size and divides it into 14px×14px patches, creating an image feature vector of size 576.

[3]We experimented with min, max, mean, and top-k mean pooling. We chose the top-k mean configuration as it provided the clearest visualization for our analysis.

compare attention weights to image features from *all* tokens in the output with those from *a subset of* tokens in the output. For the latter procedure, tokens that most intensely attend to image features during generation are deemed salient and are selected.

We provide attention heatmaps of a representative instance in Figure 4. The task in this example is to identify the color of the pot in the image, and the initial response incorrectly answers ("red") and then the feedback corrects the answer ("silver").

Such a correction can be explained by the difference in the distribution of attention to image features during the generation of each token. Based on the heatmaps of all tokens attending to image features, when VOLCANO generates the initial response, it mostly focuses on features on the outer edges, corresponding to the rim of the pot; when generating feedback, it attends to the entire image including outer regions corresponding to the silver pot and inner regions with red berries in it. Heatmaps of specific tokens attending to image features show that in the process of improving the initial response, VOLCANO indeed focuses on the exact areas of the image corresponding to key color descriptors "silver" and "red" when generating these words.

The findings suggest that during the feedback generation phase, the model develops an enhanced focus on an increased coverage of salient features, leading to a more comprehensive understanding of the image. This capability is beneficial for addressing the fundamental cause of multimodal hallucination of LLMs, which is that a lack of clear visual features leads LLMs to base their responses on pre-existing knowledge (Zhai et al., 2023; Li et al., 2023d; Liu et al., 2023a; Wang et al., 2023b). We propose that VOLCANO, with its ability to extract fine-grained visual information through feedback, can effectively reduce multimodal hallucination.

6 Conclusion

In our work, we suggest a novel approach that utilizes feedback as visual signals to direct the model to refine responses that do not accurately reflect the image. Building on this approach, we present VOLCANO, a multimodal self-feedback guided revision model. VOLCANO has not only achieved state-of-the-art results on a multimodal hallucination benchmark but also demonstrated its effectiveness by improving performance compared to baseline models on multimodal understanding benchmarks. Through qualitative analysis, we demonstrate that the feedback produced by VOLCANO is well-grounded on the image, and providing the model with rich visual information helps reduce multimodal hallucination. We hope our model and data open new pathways for strategies to mitigate multimodal hallucination and uncover the fundamental cause of the issue.

Limitations

In our study, we successfully demonstrated that VOLCANO can mitigate multimodal hallucination, as evidenced by our evaluations and analyses across various benchmarks. However, one notable drawback is the increased execution time. VOLCANO necessitates multiple calls to the model, making it less time-efficient than directly generating a response. On average, VOLCANO tends to be around 2 to 3 times slower than the base model, requiring 5.8 seconds to generate a response for a given image and instruction compared to 2.7 seconds by LLaVA-1.5. A strategy we use to reduce the overall execution time is to limit the number of iterations to 3. We think future work could explore improving the efficiency of the self-feedback-guided revision process.

Acknowledgements

We thank Seungone Kim, Seonghyun Ye, Doyoung Kim and Miyoung Ko for helpful discussion and valuable feedback on our work. This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2022-0-00264, Comprehensive Video Understanding and Generation with Knowledge-based Deep Logic Neural Network, 80%; No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST), 20%).

References

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

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda

Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models.

Ali Furkan Biten, Lluís Gómez, and Dimosthenis Karatzas. 2022. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2473–2482.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning.

Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback.

Zhibin Gou, Zhihong Shao, Yeyun Gong, yelong shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2024. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *The Twelfth International Conference on Learning Representations*.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2023. Detecting and preventing hallucinations in large vision language models.

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

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2023. Prometheus: Inducing fine-grained evaluation capability in language models.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Lu, Thomas Mesnard, Colton Bishop, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023a. Otter: A multi-modal model with in-context instruction tuning.

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

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023c. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023d. Evaluating object hallucination in large vision-language models.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023c. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. 2023d. Grounding dino: Marrying dino with grounded pre-training for open-set object detection.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023e. Mmbench: Is your multi-modal model an all-around player?

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Eureka: Human-level reward design via coding large language models.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback.

OpenAI. 2022. Chatgpt: Optimizing language models for dialogue.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Liangming Pan, Michael Saxon, Wenda Xu, Deepak Nathani, Xinyi Wang, and William Yang Wang. 2023. Automatically correcting large language models: Surveying the landscape of diverse self-correction strategies.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Jérémy Scheurer, Jon Ander Campos, Jun Shern Chan, Angelica Chen, Kyunghyun Cho, and Ethan Perez. 2022. Training language models with language feedback.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf.

Bin Wang, Fan Wu, Xiao Han, Jiahui Peng, Huaping Zhong, Pan Zhang, Xiaoyi Dong, Weijia Li, Wei Li, Jiaqi Wang, and Conghui He. 2023a. Vigc: Visual instruction generation and correction.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, Jitao Sang, and Haoyu Tang. 2023b. Evaluation and analysis of hallucination in large vision-language models.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023c. Large language models are not fair evaluators.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023d. Shepherd: A critic for language model generation.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2023. Fine-grained human feedback gives better rewards for language model training.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023a. mplug-owl: Modularization empowers large language models with multimodality.

Seonghyeon Ye, Yongrae Jo, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, and Minjoon Seo. 2023b. Selfee: Iterative self-revising llm empowered by self-feedback generation. Blog post.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023. Halle-switch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2023a. How language model hallucinations can snowball.

Renrui Zhang, Jiaming Han, Chris Liu, Peng Gao, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, and Yu Qiao. 2023b. Llama-adapter: Efficient fine-tuning of language models with zero-init attention.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023c. Siren's song in the ai ocean: A survey on hallucination in large language models.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2023. Analyzing and mitigating object hallucination in large vision-language models.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models.

A Full results on benchmarks

In this section, we describe the detailed results from the benchmarks used in our work. The benchmarks are designed to evaluate the performance of LMMs from multiple perspectives, encompassing various sub-tasks and types of questions.

A.1 Multimodal hallucination benchmarks

For MMHal-Bench, the questions are categorized into 8 types: Attribute, Adversarial, Comparison, Counting, Relation, Environment, Holistic, and Other (Table 6). POPE evaluates three types of questions: random, popular, and adversarial (Table 7).

A.2 Multimodal understanding benchmarks

MM-Vet is composed of sub-tasks designed to measure 6 LMM capabilities: Recognition, OCR (Optical Character Recognition), Knowledge, Language generation, Spatial awareness, and Math (Table 8). MMBench is structured to evaluate across L-1, L-2, and L-3 dimensions. We followed previous works and conducted evaluations for the L-2 dimension. The L-2 dimension tasks include Coarse Perception (CP), Fine-grained Single-instance Perception (FP-S), Fine-grained Cross-instance Perception (FP-C), Attribute Reasoning (AR), Relation Reasoning (RR), and Logic Reasoning (LR) (Table 9).

B Prompts

B.1 Prompts for inference at each stage

For all prompts, we did not explicitly provide an image feature prompt. Instead, the image features are concatenated with the question during the tokenization process before being input to the model. Additionally, the prompt for the decision process is based on the work of (Liu et al., 2023b).

B.2 Prompt for generating multimodal feedback

We introduce the prompt used in generating our multimodal feedback dataset. For an LLM that cannot see images, we included the image contents in the form of text within the prompt, allowing it to provide feedback as if it had seen the image and initial response. We utilized object information and a gold caption as the image contents. In instances where no objects are present in the dataset, we didn't use a separate object detector to prevent the model's errors from propagating into the feedback. Instead, only the gold caption is provided in such cases. Additionally, to avoid erroneously generating feedback that suggests the presence of hallucination merely due to the use of different expressions, even when the initial response aligns sufficiently with the image information but uses different terms from the gold answer, we crafted the prompt to treat synonyms or paraphrases as correct answers. Drawing inspiration from previous research (Kim et al., 2023), we structured the prompt to ensure that it encapsulates these aspects well.

C Computation

For this research, we used an NVIDIA A100-SXM4-80GB GPU and an AMD EPYC 7513 32-Core Processor running at 2.0778 GHz. Training VOLCANO 7B required 8 GPUs and took a total of 15 hours, while training VOLCANO 13B took 30 hours. While the time taken to evaluate each dataset varies, VOLCANO takes about 2 to 3 times longer to complete the entire process compared to existing baselines that only generate responses.

D Hyperparameters

We used a batch size of 128, a learning rate of 2e-5, and trained for 1 epoch. The maximum length is set to 2048, with no weight decay. We employed a cosine scheduler for learning rate adjustments, with a warmup ratio of 0.03. Additionally, we incorporated gradient checkpointing and used DeepSpeed zero stage 3. The maximum number of iterations for self-revision is 3. When generating responses, we utilized greedy decoding following LLaVA-1.5.

Model	Attribute ↑	Adversarial ↑	Comparison ↑	Counting ↑	Relation ↑	Environment ↑	Holistic ↑	Other ↑	Score ↑	Hal rate ↓
Kosmos-2	2	0.25	1.42	1.67	1.67	2.67	2.5	1.33	1.69	0.68
IDEFIC 9B	1.58	0.75	2.75	1.83	1.83	2.5	2.17	1.67	1.89	0.64
IDEFIC 80B	2.33	1.25	2	2.5	1.5	3.33	2.33	1.17	2.05	0.61
InstructBLIP 7B	3.42	2.08	1.33	1.92	2.17	3.67	1.17	1.08	2.1	0.58
InstructBLIP 13B	2.75	1.75	1.25	2.08	2.5	**4.08**	1.5	1.17	2.14	0.58
LLaVA-SFT+ 7B	2.75	2.08	1.42	1.83	2.17	2.17	1.17	0.5	1.76	0.67
LLaVA-RLHF 7B	2.92	1.83	2.42	1.92	2.25	2.25	1.75	1.08	2.05	0.68
LLaVA-SFT+ 13B	3.08	1.75	2	**3.25**	2.25	3.83	1.5	1.75	2.43	0.55
LLaVA-RLHF 13B	3.33	**2.67**	1.75	2.25	2.33	3.25	2.25	**2.42**	2.53	0.57
LLaVA-1.5 7B	3.17	1.25	3.17	2.5	2.33	3.17	1.5	2.25	2.42	0.55
LLaVA-1.5 13B	**3.5**	2	2.67	2.33	1.67	3.33	2.58	2.25	2.54	0.52
VOLCANO 7B	3.42	2.42	3.08	1.75	**2.75**	3.75	1.33	2.33	2.6	0.49
VOLCANO 13B	3	1.75	**3.42**	1.67	2.33	3.75	**2.75**	**2.42**	**2.64**	**0.48**

Table 6: Results on MMHal-Bench

Model	Random			Popular			Adversarial			Overall	
	Acc ↑	F1 ↑	Yes (%)	Acc ↑	F1 ↑	Yes (%)	Acc ↑	F1 ↑	Yes (%)	Acc ↑	F1 ↑
Shikra	86.9	86.2	43.3	84	83.2	45.2	83.1	82.5	46.5	84.7	84.0
InstructBLIP	88.6	89.3	56.6	79.7	80.2	52.5	65.2	70.4	67.8	77.8	80.0
MiniGPT-4	79.7	80.2	52.5	69.7	73	62.2	65.2	70.4	67.8	71.5	74.5
mPLUG-Owl	54	68.4	95.6	50.9	66.9	98.6	50.7	66.8	98.7	51.9	67.2
LLaVA-SFT+ 7B	86.1	85.5	44.5	82.9	82.4	47.2	80.2	80.1	49.6	83.1	82.7
LLaVA-RLHF 7B	84.8	83.3	39.6	83.3	81.8	41.8	80.7	79.5	44	82.9	81.5
LLaVA-SFT+ 13B	86	84.8	40.5	84	82.6	41.6	82.3	81.1	43.5	84.1	82.8
LLaVA-RLHF 13B	85.2	83.5	38.4	83.9	81.8	38	82.3	80.5	40.5	83.8	81.9
LLaVA-1.5 7B	88.2	87.3	41.9	87.3	86.2	41.8	85.2	84.2	44	86.9	85.9
LLaVA-1.5 13B	88	87.1	41.7	87.4	86.2	41.3	85.5	84.5	43.3	87.0	85.9
VOLCANO 7B	89.9	89.4	43.9	**88.5**	**87.9**	45.1	86.2	85.7	46.6	88.2	**87.7**
VOLCANO 13B	**90.2**	**89.7**	44.3	88.1	87.4	44.5	**86.6**	**86.1**	46.7	**88.3**	**87.7**

Table 7: Results on Pope

Model	rec ↑	ocr ↑	know ↑	gen ↑	spat ↑	math ↑	total ↑
Transformers Agent (GPT-4)	18.2	3.9	2.2	3.2	12.4	4	13.4
MiniGPT-4-8B	27.4	15	12.8	13.9	20.3	7.7	22.1
BLIP-2-12B	27.5	11.1	11.8	7	16.2	5.8	22.4
MiniGPT-4-14B	29.9	16.1	20.4	22.1	22.2	3.8	24.4
Otter-9B	27.3	17.8	14.2	13.8	24.4	3.8	24.7
OpenFlamingo-9B	28.7	16.7	16.4	13.1	21	7.7	24.8
InstructBLIP-14B	30.8	16	9.8	9	21.1	10.5	25.6
InstructBLIP-8B	32.4	14.6	16.5	18.2	18.6	7.7	26.2
LLaMA-Adapter v2-7B 3	8.5	20.3	**31.4**	**33.4**	22.9	3.8	31.4
LLaVA-1.5 7B	37	21	17.6	20.4	24.9	7.7	31.2
LLaVA-1.5 13B	40.6	28	23.5	24.4	**34.7**	7.7	36.1
VOLCANO 7B	36.7	23.5	18.2	22	27.6	3.8	32
VOLCANO 13B	**42.9**	**30.4**	24.5	29.2	32.7	**15**	**38**

Table 8: Results on MM-Vet

Model	LR ↑	AR ↑	RR ↑	FP-S ↑	FP-C ↑	CP ↑	Overall ↑
OpenFlamingo	6.7	8	0	6.7	2.8	2	4.6
OpenFlamingo v2	4.2	15.4	0.9	8.1	1.4	5	6.6
MMGPT	2.5	26.4	13	14.1	3.4	20.8	15.3
VisualGLM	10.8	44.3	35.7	43.8	23.4	47.3	38.1
LLaMA-Adapter	11.7	35.3	29.6	47.5	38.6	56.4	41.2
μ-G2PT	13.3	38.8	40.9	46.5	38.6	58.1	43.2
mPLUG-Owl	16.7	53.2	47.8	50.2	40.7	64.1	49.4
Otter	32.5	56.7	53.9	46.8	38.6	65.4	51.4
Shikra	25.8	56.7	58.3	57.2	57.9	75.8	58.8
Kosmos-2	**46.7**	55.7	43.5	64.3	49	72.5	59.2
PandaGPT	10	38.8	23.5	27.9	35.2	48.3	33.5
MiniGPT-4	20.8	50.7	30.4	49.5	26.2	50.7	42.3
InstructBLIP	19.1	54.2	34.8	47.8	24.8	56.4	44
LLaVA-1.5 7B	30.8	**73.1**	53.9	67	57.2	77.2	59.9
LLaVA-1.5 13B	41.7	69.7	63.5	70	59.3	80.2	67.7
VOLCANO 7B	30.8	65.2	59.1	67.7	54.5	72.8	62.3
VOLCANO 13B	38.3	70.6	**67**	**72.4**	**62.8**	**82.2**	**69.4**

Table 9: Results on MMBench

System prompt

You are excellent multimodal feedback-generating assistant. You are given questions about the image contents, objects information, reference answers, image contents and the model's response to evaluate. Utilizing these informations, please give me some feedback on the model's response only if feedback is needed.

Rule
- Consider synonyms or paraphrases in response as a correct answer

---

User prompt

Your job is to generate multimodal feedback of the given response.

Object information:
{objs}

Image contents:
{Capts}

Question:
{question}

Response to Evaluate:
{prediction}

Reference Answer:
{answer}

* Feedback
- The feedback should each be an explanation of why the response is imperfect and how it could improve.
- The feedback should consider the image contents and object information.
- The feedback shouldn't just copy and paste the response, but it should also give very detailed feedback on the content of the response.

* Format
- DO NOT WRITE ANY GREETING MESSAGES, just write the feedback only.

Generated Feedback:

Figure 6: **Prompt for generating multimodal feedback**

Feedback prompt *(stage 1)*

Generate the feedback given initial answer referring to question and image.
Question: {question}
Initial answer: {initial response}

Revision prompt *(stage 2)*

Adjust the initial response considering the feedback and image.
Question: {question}
Initial answer: {initial response}
Feedback: {feedback}

Decision prompt *(stage 3)*

{question}
Answer with the option's letter from the given choices directly.
A. {initial response}
B. {revised response}

Figure 7: **Prompts for inference at each stage**