VOLCANO: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision

Anonymous ACL submission

Abstract

Large multimodal models (LMMs) 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 might be 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 a 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, helping alleviate multimodal hallucination. We publicly release VOLCANO models of 7B and 13B sizes along with the data and code at http://www.omitted.link/.

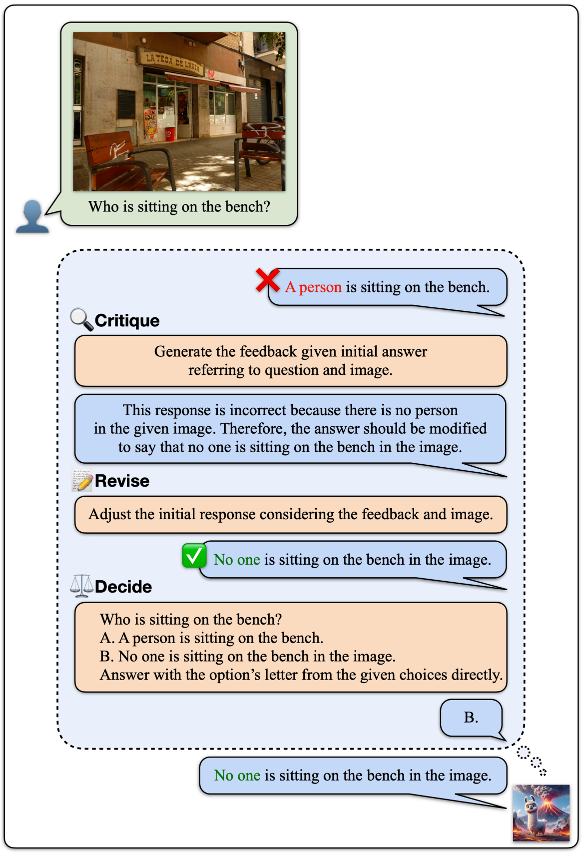

Figure 1: **Overview of VOLCANO.** This example illustrates the process undertaken by VOLCANO for a question in the MMHal-Bench dataset. Before giving the response, VOLCANO goes through a *critique-revise-decide* process. It critiques its initial response with natural language feedback, revises the response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

1 Introduction

Large multimodal models (LMMs) enable instruct-tuned large language models (LLMs) to comprehend the visual features conveyed by vision encoders with the help of substantial image-text or video-text pairs (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). Recently, with the introduction of fine-tuning methods such as visual instruction tuning, LMMs are evolving into assistants capable of understanding the world through multiple channels, akin to humans (Liu et al., 2023b,c).

Despite the impressive performance on various benchmark tasks and qualitative outcomes observed, 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 images accurately. In other words, LMMs tend to rely more on their own parametric knowledge than on provided visual features, causing them to respond with guesses and generate multimodal hallucinations. Wang et al. (2023b) empirically show that the model attends to the previous tokens more than image features when it generates hallucinated tokens.

In this paper, we propose a novel method that utilizes natural language feedback to enable the model to correct hallucinated responses by offering detailed visual information. Building on this hypothesis, 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 it until it determines that no more improvement is required. We collect our multimodal feedback and revision data for training 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). The results demonstrate consistent performance improvements across all benchmarks. Notably, when compared to previous works aiming at mitigating multimodal hallucination (Zhou et al., 2023; Sun et al., 2023; Yin et al., 2023), VOLCANO showcases an 24.9% enhancement, underscoring its effectiveness in addressing the challenge. Furthermore, on multimodal understanding benchmarks (Liu et al., 2023e; Yu et al., 2023), VOLCANO is also effective in understanding and reasoning about visual concepts.

Through qualitative analysis, we find that the generated feedback attends on the image with higher intensity and disperses the attention widely across the image. These findings explain that feedback carries fine-grained visual information and suggest that even if the vision encoder fails to properly ground, the feedback can still guide LLMs to improve upon a hallucinated response, supporting our claim.

Our work's contributions can be summarized as follows:

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

2. Our qualitative analysis shows that VOLCANO's feedback is effectively rooted on the image, conveying rich visual details. This underscores that feedback can offer guidance and reduce multimodal hallucination, even when a vision encoder inadequately grounds the image

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

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), the majority of multimodal hallucination occurs within verifiable information given the input visual content. Multimodal hallucination is mostly studied as a form of object hallucination where a generation contains objects 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), with misrepresentations of a scene or environment being documented until recently (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 Learning from feedback

Learning from feedback can align LLMs to desired outcomes, for instance 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.,

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

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 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), to the best of our knowledge, we are the first to achieve improvement in multimodal modals through iterative 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 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 of extra modules. Furthermore, we introduce natural language feedback prior to 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 employs 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

```
1:  Input: model M, image I, question Q
2:  R_initial = M(I, Q)
3:  R_best = R_initial
4:  for up to 3 iterations do
5:      F = M(I, Q, R_best)
6:      R_revised = M(I, Q, R_best, F)
7:      R_decided = M(I, Q, R_best, R_revised)
8:      if R_decided == R_best then
9:          break
10:     else
11:         R_best = R_revised
12: return R_best
```

3.1 Iterative self-revision

VOLCANO employs a single model to generate 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 the 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.

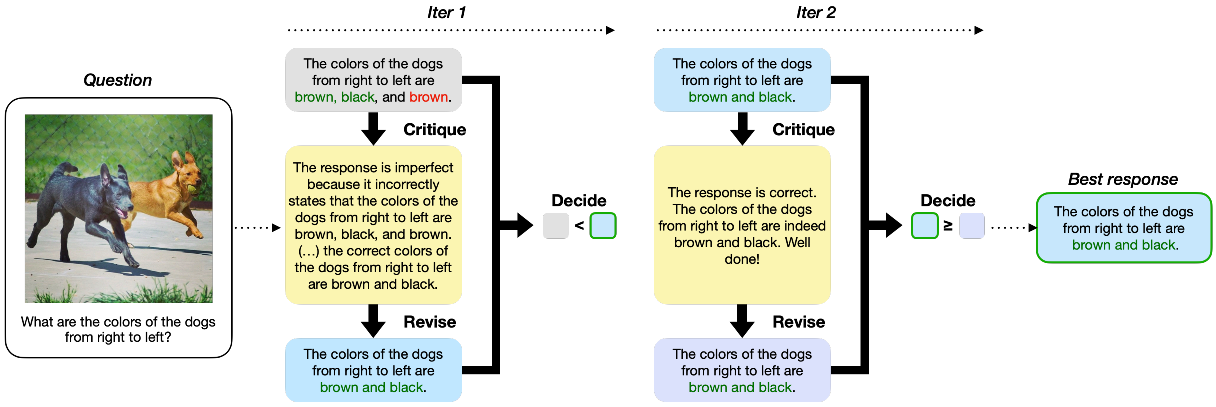

Figure 2: **Overall process of VOLCANO.** VOLCANO is a multimodal self-feedback guided revision model that takes an image and a question and then generates an improved response based on the self-feedback.

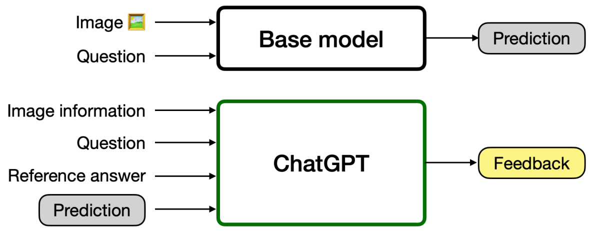

Figure 3: **Data collection.**

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; 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 image. 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 use text-formatted image details when generating feedback. When constructing the revision data, we set up the 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.

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 finetuning VOLCANO, we use the llava-1.5-mix665k as the visual instruction dataset (Liu et al., 2023b).

Model For proprietary LLM, we employ OpenAI's gpt-3.5-turbo[2]. We use the LLaVA-SFT+ 7B model[3] to generate the initial response when creating feedback data and LLaVA-1.5 7B[4] and 13B[5] as backbone models of VOLCANO (Liu et al., 2023b,c).

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 multimodal hallucination benchmarks. 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. GPT-4 evaluates an overall score by comparing

[2] `gpt-3.5-turbo`
[3] `LLaVA-RLHF-7b-v1.5-224`
[4] `llava-v1.5-7b`
[5] `llava-v1.5-13b`

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 of 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.

the model's response to the correct answer based on the given object information. If the overall score is less than 3, it is considered to have hallucinations.

Multimodal understanding benchmarks We use MM-Vet (Yu et al., 2023) and MMBench (Liu et al., 2023e) as benchmarks to measure the general performance of LMMs. MM-Vet is a benchmark consisting of 16 tasks designed to evaluate LMM's ability in complex multimodal tasks. It has about 218 instances. GPT-4 measures the score by comparing 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 a dev split 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+, and LLaVA-RLHF (Sun et al., 2023) as baseline models. For the multimodal hallucination corrector baseline, we employ LURE (Zhou et al., 2023) and Woodpecker (Yin et al., 2023). LURE utilize MiniGPT-4 13B as its backbone model. Woodpecker use 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 Results

VOLCANO achieves the best performance in the 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 show strong performance in benchmarks that measures scores using proprietary LLMs (MMHal-Bench, GAVIE) and a benchmark evaluating with 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.

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: **Results of competitive test.** VOLCANO^{-} 7B is a model fine-tuned with multimodal feedback and revision data on LLaVA-SFT+ 7B.

Natural language self-feedback is effective in revising responses. Table 2 shows VOLCANO's effectiveness by comparing it with previous studies designed to tackle multimodal hallucination. It reduces hallucination more than LURE and Woodpecker, which try to revise responses without feedback. This suggests that specific feedback is crucial for correcting multimodal hallucination. Unlike the two methods that need a separate model to revise, VOLCANO efficiently gives better responses with just one model. In addition, Woodpecker converts visual information into text and feeds it to the proprietary LLM corrector. Its improvement in hallucination is less significant compared to VOLCANO.

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 of multimodal benchmarks.** The detailed evaluation results for each benchmark by question type are in Table 8 and Table 9.

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

Table 4: **Results of module ablation.** 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 ↑	Hal rate ↓
Iter 1	2.54	0.51
Iter 2	2.58	0.5
Iter 3 (VOLCANO 7B)	**2.6**	**0.49**

Table 5: **Results of iteration ablation.**

From this, we find that for reducing multimodal hallucination, it is effective to convey visual features directly to the corrector model. When compared to LLaVA-RLHF, which reduces LLM hallucination using RLHF, VOLCANO consistently performs better. LLaVA-RLHF 7B employs LLaVA-SFT+ 7B as its core architecture. To ensure a fair comparison, we fine-tune this model using multimodal feedback and revision data, resulting in the development of a VOLCANO^{-} 7B. The result shows that giving natural language feedback, which the model can directly understand, is more powerful than providing feedback in scalar value form.

VOLCANO is also effective for general multimodal understanding tasks. As multimodal 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 prove this, we evaluate VOLCANO on benchmarks assessing LMM's complicated visual reasoning and perception capabilities (Table 3). It 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.

4.4 Ablation studies

Module ablation We test the influence of each stage in reducing multimodal hallucination. As shown in Table 4, when we skip iterative self-revision and only use the initial response as the final response, 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 the number of max iterations affects the VOLCANO's performance. As shown in Table 5, as the max iteration count increased, the hallucination rate decreased. This demonstrates that multiple revisions can refine the answers. However, there also exists a trade-off: as the iteration count goes up, the inference time also increases.

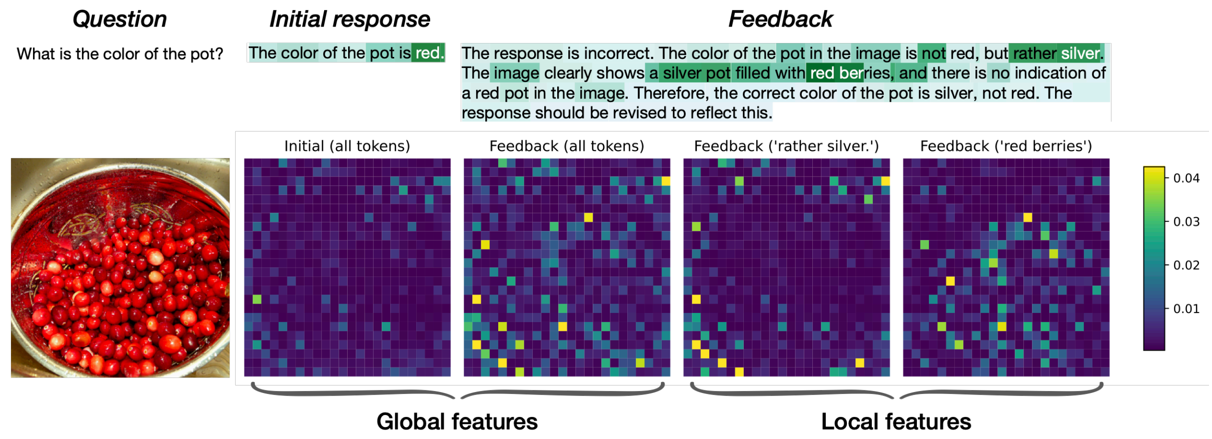

Figure 4: **Case study of image feature attention in initial response and feedback generation.** For the heatmaps above, the intensity of the highlight behind each token corresponds to the magnitude of attention weight from the token to image features, with darker highlights indicating higher attention weights. For the heatmaps below, values at or above the 0.995-th quantile are represented with the maximum color intensity on the colorbar.

5 Qualitative analysis

We qualitatively analyze how feedback from VOLCANO is effective in reducing multimodal hallucination. Using results on MMHal-Bench where VOLCANO 7B revision is selected as the final answer, we compare the visual information content between initial response and feedback, focusing on amount (5.1) and coverage (5.2).

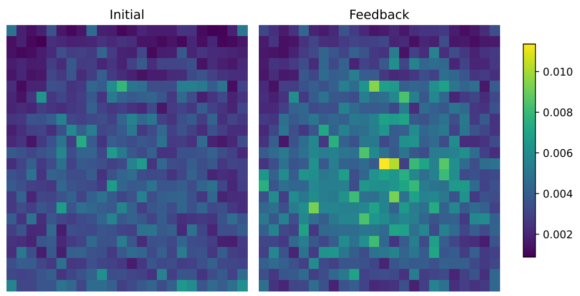

Figure 5: **Image feature attention in initial response and feedback generation.** Attention weights are averaged across instances in MMHal-Bench where VOLCANO's revision enhance the initial response.

5.1 Amount of visual information

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

To delve deeper into this phenomenon, we take inspiration from Wang et al. (2023b) by visualizing how attention weights connect output tokens to input image features during the generation of both initial responses and feedback. Specifically, we focus on the top-3 attention weights across hidden layers and attention heads. These weights are averaged to form a consolidated view. As there is a difference in the initial response length and feedback length, we choose the minimum k of the two and averaged top-k weights from the output.[6]

As shown in Figure 5, image features are more strongly attended by feedback compared to initial response. Interestingly, even though attention to input would be more dispersed when generating feedback due to the inclusion of the initial response as additional input, an increased concentration on large areas of image features is visible. This suggests that visual information is largely contained in the feedback text, supporting our manual observation beforehand.

5.2 Coverage of visual information

We further investigate the coverage of information to identify whether the visual information correctly aligns with both global and local image features. We perform a case study on an instance that asks the color of a pot (Figure 4). The initial response incorrectly answers "red" while the feedback makes it clear that the answer should be "silver".

The correction can be explained by the difference in distribution of attention to image features during

[6]This approach is chosen based on experiments with different aggregation methods—max, mean, and top-k-mean pooling. We find that the top-3 configuration provided the clearest visualization for our analysis.

each generation. Based on the global features visualization, when VOLCANO generates the initial response, it only focuses on features corresponding to the pot. When generating feedback, VOLCANO attends to the entire image including the areas corresponding to the pot and red berries in the it. Specifically, the local features visualization show that in the process of improving the initial response, it indeed focuses on the exact areas of the image corresponding to key color descriptors "red" and "silver" when generating these words. From these findings, we infer that VOLCANO can grasp a more holistic view of the image and distinguish information in local features at the same time.

In summary, existing LMMs may generate answers based on their prior knowledge if the visual features lack clarity, leading to multimodal hallucination. We suggest that VOLCANO can alleviate multimodal hallucination as it is capable of acquiring fine-grained visual information from its feedback. The feedback can effectively encompass a sufficient quantity of a broad spectrum of image features.

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, which means that it can provide the model with rich visual information. This helps reducing multimodal hallucination.

Limitations

In this study, we demonstrate through evaluation and analysis in benchmarks that VOLCANO can effectively alleviate multimodal hallucination. However, it requires more time to execute as it needs to call the model multiple times, compared to directly generating a response. To address this, we introduce stage 3, which allows for early stopping, thereby reducing the execution time.

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.

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.

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 Appendix

A.1 Detailed results

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. 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). 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).

A.2 Prompts

Prompt for generating multimodal feedback We introduce the prompt used in generating our multimodal feedback dataset. For a 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.

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).

A.3 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.

A.4 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 of 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 of 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 of 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 of 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**