HELPD: Mitigating Hallucination of LVLMs by Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding

Fan Yuan[1,2], Chi Qin[1,2], Xiaogang Xu[3], Piji Li[1,2*]

[1] College of Artificial Intelligence,
Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[3] The Chinese University of Hong Kong, Hong Kong, China
{fanyuan, qinchi, pjli}@nuaa.edu.cn, xiaogangxu00@gmail.com

Abstract

Large Vision-Language Models (LVLMs) have shown remarkable performance on many visual-language tasks. However, these models still suffer from *multimodal hallucination*, which means the generation of objects or content that violates the images. Many existing work detects hallucination by directly judging whether an object exists in an image, overlooking the association between the object and semantics. To address this issue, we propose Hierarchical Feedback Learning with Vision-enhanced Penalty Decoding (HELPD). This framework incorporates hallucination feedback at both object and sentence semantic levels. Remarkably, even with a marginal degree of training, this approach can alleviate over 15% of hallucination. Simultaneously, HELPD penalizes the output logits according to the image attention window to avoid being overly affected by generated text. HELPD can be seamlessly integrated with any LVLMs. Our experiments demonstrate that the proposed framework yields favorable results across multiple hallucination benchmarks. It effectively mitigates hallucination for different LVLMs and concurrently improves their text generation quality. [1]

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b), guided by human instruction, have demonstrated impressive performance in numerous Natural Language Processing (NLP) tasks (Qin et al., 2023). In light of the success of LLMs, researchers aspire to integrate the powerful capabilities of LLMs into multimodal domains, consequently introducing Large Vision-Language Models (LVLMs) (Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023; Ye et al., 2023; Liu et al., 2023b). Particularly,

---

[*]Corresponding author.
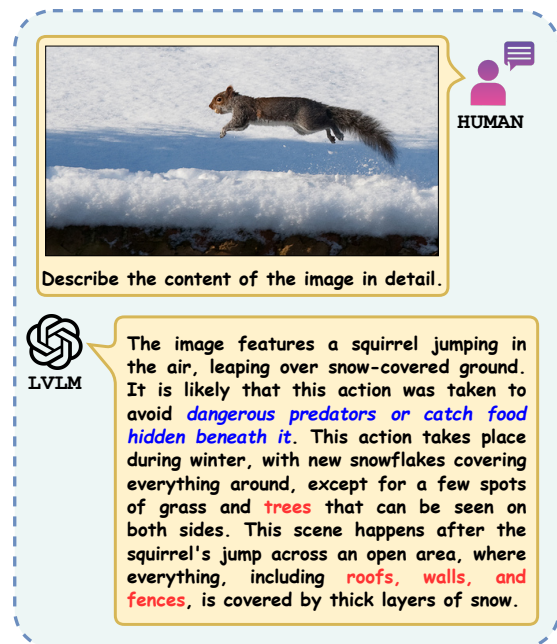[1]Code is available at https://github.com/F-Yuan303/HELPD

Figure 1: **A case of LVLM hallucination.** The parts marked in red are, in fact, hallucinations. The parts marked in blue would be mistaken for hallucinations by detection methods that focus only on objects.

GPT-4 (OpenAI, 2023) has been endowed with the capability to engage in complex, image-based dialogues with humans, while also being proficient in resolving a series of visual-language tasks.

Despite the fact that LVLMs have achieved quite considerable results on various tasks, problems with these models have gradually emerged. Within these problems, the hallucination (Rohrbach et al., 2018; Li et al., 2023c) has attracted significant attention. This is a phenomenon that LVLMs tend to generate content contradictory to the image, such as non-existent objects. In order to alleviate this phenomenon, many explorations have been carried out in recent work (Li et al., 2023c; Wang et al., 2023; Liu et al., 2023a; Zhou et al., 2023; Zhai et al., 2023; Lee et al., 2023; Huang et al., 2023; Wang et al., 2024). CoVe (Dhuliawala et al.,

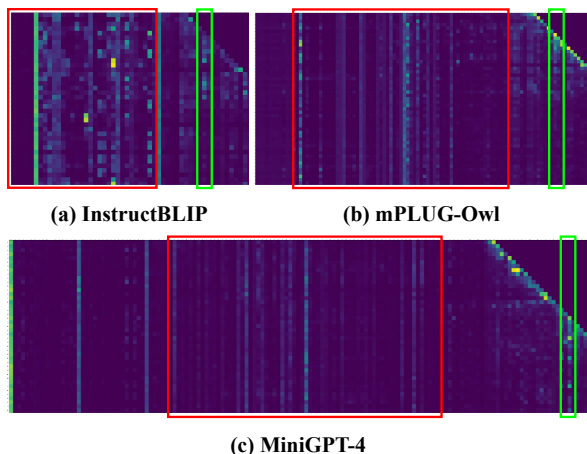**(a) InstructBLIP**   **(b) mPLUG-Owl**

(c) MiniGPT-4

Figure 2: **Attention visualization of LVLMs.** For the same input, each image represents the attention matrix of a specific LVLM generation instance. *Red* indicates the attention of the image, while *green* represents the phenomenon of "Over-trust" in the generated text.

2023) proposes a Chain-of-Verification method, it first generates verification questions, then executes them to check for hallucination, and finally gets a revised response. Additionally, some approaches aim to alleviate the hallucination from the decoding strategy (O'Brien and Lewis, 2023; Chuang et al., 2023; Huang et al., 2023). Dola decoding (Chuang et al., 2023) is a strategy of contrasting the mature layer and the immature layer of the model, followed by the determination of the next token based on the differences in logits. Opera (Huang et al., 2023) is a recently proposed decoding method that employs an Over-trust Penalty to determine the occurrence of hallucination, and utilizes a Retrospection-Allocation rollback mechanism for decoding.

However, most of the existing work focuses on alleviating object-level hallucinations. During this process, some methods excessively concentrate on whether the generated objects exist in the image, neglecting the association between these objects and the semantics of the whole sentence. As illustrated in Figure 1, words marked in red, such as "trees", are real object hallucinations. Solely considering the presence of objects, the parts marked in blue, such as "predators" and "food", would also be defined as hallucinations. Nevertheless, combined with context and semantics, such a definition is deemed inappropriate. Meanwhile, as indicated by the green box in Figure 2, "Over-trust" (Huang et al., 2023) does exist in LVLMs, which means certain generated tokens receive excessive attention, leading to a subsequent generation that deviates from the image. We initially assumed that insuffi-

cient focus on the visual part of the input might be one cause of this phenomenon. However, further observation of the attention matrix of these models reveals strong focus on the visual input (see red boxes in Figure 2). This indicates that considering the over-trust penalty only accounts for the impact of the text, additional focus on the image is therefore required to balance it.

Based on the observations above, we propose HELPD, a novel LVLM framework that utilizes **H**ierarchical Fe**E**dback **L**earning with Vision-enhanced **P**enalty **D**ecoding. We note the necessity of integrating both the inherent properties of an object and semantic meaning to determine the presence of hallucination. Thus, we propose the hierarchical feedback learning, which only requires a small amount of training, and we add this feedback mechanism at the end of the training period. On the one hand, the collection of objects is extracted from the sampled sentences and label sentences, and the object-level feedback is obtained through the comparison of the object sets. On the other hand, leveraging the powerful few-shot inference capabilities of GPT-4, we conduct a semantic comparison to obtain sentence-level feedback.

The manner of sampling constitutes a crucial component in the hierarchical feedback learning process. Opera decoding (Huang et al., 2023) predicts the next word by subtracting the over-trust penalty score from the logits, where the penalty score is computed based on the attention window of the generated text, disregarding the potent influence of visual attention. Consequently, we propose the Vision-Enhanced Penalty Decoding, which incorporates visual attention into the penalty score computation and makes the final logits place more emphasis on the image input. This approach effectively mitigates an over-reliance on the textual modality during the decoding process, and enhances the influence of visual modality, thereby alleviating the hallucination.

Our contributions can be summarized as follows:

- We propose a hierarchical feedback learning method, incorporating object-level and sentence-level hallucination feedback. It can mitigate the occurrence of hallucinations with only a minimal amount of training.

- With the analysis of the attention matrix during decoding, we introduce the vision-enhanced penalty decoding to enhance the influence of images on the generation process.

- Extensive experimental results indicate that our proposed framework shows better performance on multiple hallucination metrics and can effectively alleviate the hallucination of LVLMs.

2 Related Work

2.1 Large Vision-Language Models (LVLMs)

Owing to the success of Large Language Models (LLMs) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b; Chung et al., 2022; Zeng et al., 2023; Sun et al., 2021; Yang et al., 2023) in many Natural Language Processing (NLP) tasks (Qin et al., 2023), many researchers have endowed it with multimodal perception capabilities. Among these, Large Vision-Language Models (LVLMs) (Li et al., 2023a; Dai et al., 2023; Zhu et al., 2023; Ye et al., 2023; Liu et al., 2023b; OpenAI, 2023; Peng et al., 2023; Anil et al., 2023) have shown particularly notable performance.

LVLMs primarily consist of three components: a visual encoder, a modality alignment module, and a Large Language Model (LLM) (Yin et al., 2023; Zhang et al., 2024). Visual encoders include Vision Transformers (ViT) (Dosovitskiy et al., 2021), CLIP ViT (Radford et al., 2021), and others (Brock et al., 2021; Fang et al., 2023). Specifically, ViT splits images into patches, which are then input into Transformer blocks through linear mapping for feature learning. Since there exits a modality gap between the visual encoders and the LLMs, the modal alignment module is required as a bridge. Models such as Flamingo (Alayrac et al., 2022), BLIP-2 (Li et al., 2023a), and InstructBLIP (Dai et al., 2023) apply the Q-former, a method that extracts visual features in a query-based manner by employing a set of learnable vectors. Another more direct method involves using a linear interface for modality alignment. For instance, LLaVA (Liu et al., 2023b) employs a linear layer to map images to the textual embedding space.

2.2 Hallucination in LVLMs

Multimodal hallucination is a significant challenge faced by LVLMs, severely impairing the reliability and robustness of these models. It typically manifests as generating content that is inconsistent with the image or contradicts common sense. Generally, hallucinations can be divided into Intrinsic Hallucination and Extrinsic Hallucination. Intrinsic hallucination refers to the generation of content that conflicts with the input. On the other hand, extrinsic hallucination represents the generation of additional content that does not actually exist, such as objects not present in the image.

Recently, numerous efforts have been dedicated to the elimination of multimodal hallucination (Li et al., 2023c; Wang et al., 2023; Liu et al., 2023a; Zhou et al., 2023; Zhai et al., 2023; Lee et al., 2023; Huang et al., 2023; Wang et al., 2024; Gunjal et al., 2024; Zhao et al., 2023). CHAIR (Rohrbach et al., 2018) is an early proposed metric for evaluating object hallucinations in image captioning tasks. It assesses the degree of hallucination by calculating the proportion of objects that appear in the generated descriptions but not in the image itself. POPE (Li et al., 2023c) introduces a polling-based object probing evaluation method, which assesses the degree of hallucination based on the responses to questions like "Is there a <object> in the image?" that are posed based on the objects. CoVe (Dhuliawala et al., 2023) introduces a chain-of-verification that considers its own responses and self-corrects hallucination. Liu et al. (2023a) conducts visual instruction tuning on LVLMs with the newly proposed LRV-Instruction dataset to mitigate hallucinations.

3 Method

3.1 Hierarchical Feedback Learning

As we have illustrated in Section 1, to determine the occurrence of hallucination, it is necessary to consider not only whether the mentioned object appears in the image, but also to judge whether it is a reasonable association in combination with semantics. To address the aforementioned issue, we propose Hierarchical Feedback Learning, a learning method that enhances the model's intrinsic ability to avoid hallucination through different granularity hallucination detection feedback (see Figure 3).

In practice, we conduct minimal further training on the LVLMs and incorporate this feedback-learning mechanism towards the end of the training process. More specifically, after every fixed number of training steps, we sample the logits of the model's output to obtain the actions (which means the sampled tokens) $A = \{a_{ij}\}_{j=1}^{t}, \quad i = 1, \ldots, b$, where t is the length of the sampled sentence and b is batch size. With the help of NLTK [2] and GPT-4, we extract objects from the sampled sentence and label sentence, respectively, obtaining the sampled
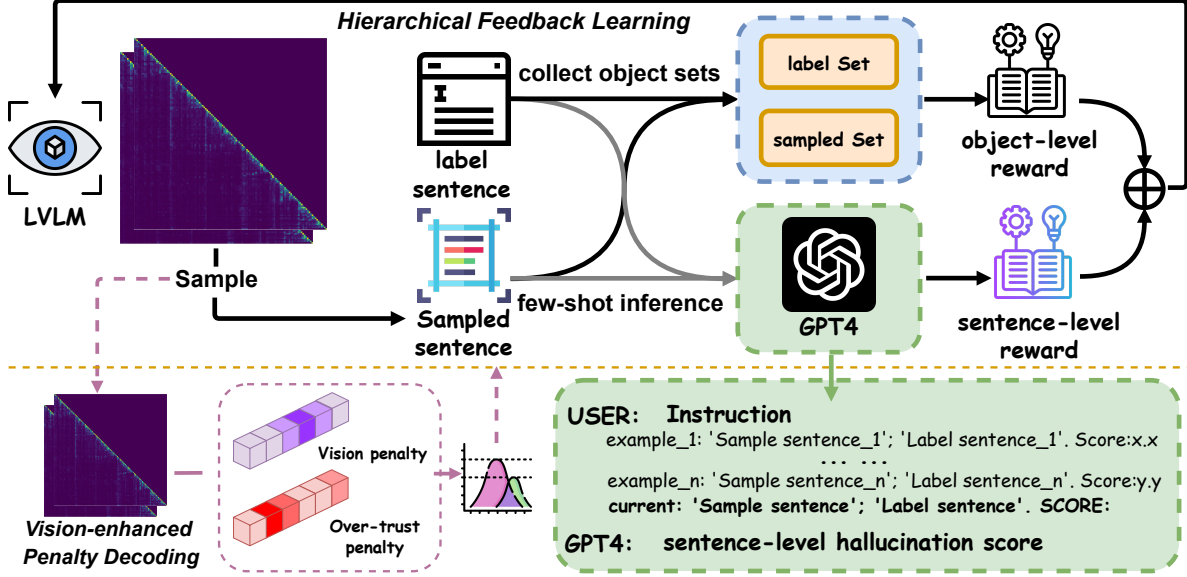
---

[2]https://www.nltk.org/install.html

Figure 3: **This diagram illustrates the framework of HELPD.** The Hierarchical Feedback Learning detects hallucination by obtaining object-level feedback from comparing object sets extracted from sampled and label sentences, and sentence-level feedback through semantic comparison using GPT-4's few-shot inference capabilities. To improve the effectiveness of sampling, the Vision Penalty Decoding augments the over-trust penalty score with a vision-enhanced penalty score, making the final logits closer to the image.

object set $S_{sam} = \{obj_1, obj_2, \ldots, obj_m\}$ and the label object set $S_{lab} = \{obj_1, obj_2, \ldots, obj_n\}$, where m and n represent the number of objects. Subsequently, we calculate the F1 score of these two sets as the object-level feedback scores R_{obj}:

$$\text{Precision} = \frac{|S_{sam} \cap S_{lab}|}{|S_{sam} \cap S_{lab}| + |S_{sam} \setminus S_{lab}|} \quad (1)$$

$$\text{Recall} = \frac{|S_{sam} \cap S_{lab}|}{|S_{sam} \cap S_{lab}| + |S_{lab} \setminus S_{sam}|} \quad (2)$$

$$R_{obj} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (3)$$

Sentence-level feedback is obtained through the few-shot inference capability of GPT-4. We provide a detailed evaluation method and pre-annotate several sentence pairs as context (see Appendix A for detail) to instruct it on discerning hallucination from semantics. The score ranging from 0 to 1, returned by GPT-4, is defined as R_{sen}.

Given that R_{sen} and R_{obj} are non-differentiable, they cannot be directly incorporated into training using the gradient method. Inspired by (Sutton et al., 1999), we introduce the reinforce algorithm to handle this problem. Specifically, based on the tokens sampled, we first retrieve their corresponding log probabilities from the original logits:

$$P_{i,j} = \log \left(\frac{e^{logits_{i,j,A_{i,j}}}}{\sum_{k=1}^{V} e^{logits_{i,j,k}}} \right), \quad (4)$$

where i is the index within the batch, j is the index within the sequence length, $A_{i,j}$ is the corresponding sampled action, and k is the index within the vocabulary of size V. A hyperparameter σ is set to determine the relative importance of the two types of feedback mentioned above (see Equation (5)). By summing the product of the feedback and the corresponding log probabilities of the actions, we obtain a negative weighted log-likelihood loss. To prevent the loss from infinitely increasing with the number of actions, the total loss is divided by the number of sampled actions to yield the loss function for reinforce algorithm, denoted as $\mathcal{L}_{RL}$:

$$R_i = \sigma R_{sen,i} + (1 - \sigma) R_{obj,i}, \quad (5)$$

$$\mathcal{L}_{RL} = -\frac{1}{N} \sum_{i=1}^{b} \sum_{j=1}^{t} P_{ij} \cdot R_i. \quad (6)$$

In the early stage of training, we employ cross-entropy loss $\mathcal{L}_{CE}$. When the training step reaches $c * \text{total steps}$, $\mathcal{L}_{RL}$ is added to the total loss. The total loss is defined as:

$$\mathcal{L}_{CE} = -\sum_{h}^{H} \log P(x_h \mid x_{<h}), \quad (7)$$

$$\mathcal{L} = \begin{cases} \mathcal{L}_{CE}, & \text{if steps} < c * \text{total steps,} \\ \frac{\mathcal{L}_{CE}}{\|\mathcal{L}_{CE}\|} + \frac{\mathcal{L}_{RL}}{\|\mathcal{L}_{RL}\|}, & \text{otherwise,} \end{cases}$$

$$\quad (8)$$
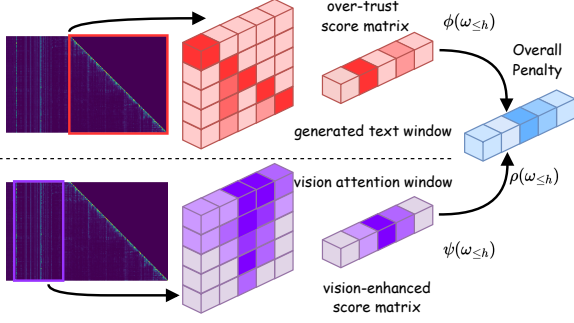
where x is the generated token.

Figure 4: **The illustration of Vision-enhanced Penalty Decoding.** The total penalty is composed of the vision penalty and the over-trust penalty. The over-trust penalty is computed based on the generated text (the upper region), while the vision penalty is computed from the vision attention window (the lower area).

3.2 Vision-enhanced Penalty Decoding

With the hierarchical feedback learning, we can effectively detect object-level and sentence-level hallucinations and correct them through back propagation. To obtain the sampled sentences, we need to sample each token from the model's logits. Based on our analysis of the attention matrix, we propose the Vision-enhanced Penalty Decoding based on Opera (Huang et al., 2023).

Over-Trust Logit Penalty. First, we provide a brief description of the original process. It sets a local window of length h for the attention matrix, where h represents the length of the current generated sequence. Upon obtaining such a lower triangular matrix, it pads the upper triangular part with zeros and scales up the values to avoid excessively small values. Subsequently, it conducts column-wise multiplication on this matrix and select the maximum value of the vector as the over-trust penalty $\phi(\omega_{\leq h})$. Finally, it subtracts this penalty from the original logits to predict the next token.

Vision-enhanced Penalty Decoding. As illustrated in Section 1, we note that the over-trust penalty establishes a local window whose size is limited to the length of the generated text. This approach can effectively extract over-trust patterns from past tokens, but it inadvertently amplifies the model's reliance on the text modality, thereby passively diminishing its focus on images.

To foster a greater focus on images during the sampling process, we set an additional local window $\mathbf{W}_l^h$ beyond the local window of the over-trust penalty, as shown in Figure 4, purposed for storing the image components within the attention matrix:

$$\mathbf{W}_l^h = \{\mathbf{w}^i\}_{i=1}^h, \quad \text{s.t. } \mathbf{w}^i = \{\omega_{i,j}\}_{j=1}^l, \quad (9)$$

where h is the length of the over-trust penalty window, l means the length of the visual input within the attention matrix, and $\omega_{i,j}$ represents the attention weight from i_{th} token to j_{th} token. Subsequently, we conduct the column-wise multiplication on the $\mathbf{W}_l^h$ to obtain a vector of column-wise scores, which represents the accumulated attention values of image:

$$\psi(\omega_{\leq h}) = \sum_{i=1}^h \omega_i, \quad \text{s.t. } \omega_i = \prod_{j=1}^l \omega_{i,j}, \quad (10)$$

where $\psi(\omega_{\leq h})$ means the vision-enhanced penalty.

Given the difference in numerical magnitudes, the initial step involves scaling $\psi(\omega_{\leq h})$ to match the order of magnitude of $\phi(\omega_{\leq h})$, then calculating the overall penalty weight $\rho(\omega_{\leq h})$, as:

$$\rho(\omega_{\leq h}) = \phi(\omega_{\leq h}) - \beta\psi(\omega_{\leq h}),$$
$$\text{s.t. } \beta = \frac{\overline{\sum_{j \leq h} \phi(\omega_j)}}{\sum_{j \leq h} \psi(\omega_j)}. \quad (11)$$

Then, this penalty weight is added to the original logits for the prediction of the next token $\hat{x}_h$, as:

$$\hat{x}_h = \arg\max_{x \in V}[p(x|x_{<h}) - \rho(\omega_{\leq h})], \quad (12)$$

where V is the size of vocabulary and x represents the predicted token.

4 Experimental Setups

4.1 Hallucination Benchmarks

CHAIR. Caption Hallucination Assessment with Image Relevance (CHAIR) (Rohrbach et al., 2018) is an evaluation metric employed for assessing hallucination in image captioning, and it is often used to evaluate LVLMs. CHAIR obtains scores for the degree of hallucination by calculating what proportion of objects generated are actually in the image according to the ground truth sentences and object segmentations. Specifically, it computes the hallucination at both instance level (defined as CHAIR_i) and sentence level (defined as CHAIR_s):

$$\text{CHAIR}_s = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}, \quad (13)$$

$$\text{CHAIR}_i = \frac{|\{\text{captions w/ hallucinated objects}\}|}{|\{\text{all captions}\}|}, \quad (14)$$

where CHAIR_s represents the proportion of hallucinated objects among all mentioned objects, and CHAIR_i denotes the proportion of captions with hallucinated objects among all captions.

POPE	Model	Accuracy	Precision	Recall	F1 Score	Yes (%)
Random	MiniGPT-4	53.06	51.58	99.60	67.97	96.53
	InstructBLIP	86.28	84.11	95.02	89.23	55.63
	mPLUG-Owl2	87.71	89.13	**85.92**	87.50	48.21
	mPLUG-Owl2 (w/ ours)	**88.02**	**89.90**	85.67	**87.73**	47.62
	LLaVA-1.5	89.44	87.21	90.11	88.63	53.09
	LLaVA-1.5 (w/ ours)	**89.65**	**87.84**	**91.98**	**89.86**	52.41
Popular	MiniGPT-4	50.53	50.26	99.60	66.81	89.06
	InstructBLIP	81.67	74.12	93.31	82.61	65.43
	mPLUG-Owl2	84.10	82.81	85.23	84.00	51.90
	mPLUG-Owl2 (w/ ours)	**85.67**	**85.81**	**85.27**	**85.53**	49.66
	LLaVA-1.5	84.91	81.02	90.71	85.59	55.51
	LLaVA-1.5 (w/ ours)	**85.79**	**81.85**	**91.97**	**86.62**	56.26
Adversarial	MiniGPT-4	50.46	50.23	99.61	66.78	99.13
	InstructBLIP	72.12	65.63	95.27	77.32	73.26
	mPLUG-Owl2	**81.70**	79.21	**85.93**	82.43	54.23
	mPLUG-Owl2 (w/ ours)	81.64	**80.28**	85.65	**82.87**	53.31
	LLaVA-1.5	77.61	71.72	92.55	80.81	63.83
	LLaVA-1.5 (w/ ours)	**78.15**	**72.04**	**92.91**	**81.15**	63.88

Table 1: **Results of LVLMs under three evaluation settings of POPE on the validation set of MSCOCO.** "Yes" denotes the proportion of answering "Yes" to the given question. "w/ ours" means the application of HELPD.

POPE. POPE (Li et al., 2023c) converts hallucination assessment into asking the model to answer a series of true or false questions about whether an object is present in the image. Specifically, given an image set and the object annotations contained in each image, POPE will construct a series of triples consisting of images, questions, and answers. It considers three polling strategies by sampling the objects randomly, from popular objects, and among those frequently co-occurring objects, respectively. Finally, POPE involves 3K questions for the captions of 500 images and uses the Accuracy, Precision, Recall, and F1 scores for evaluation.

GAVIE. GPT4-Assisted Visual Instruction Evaluation (GAVIE) (Liu et al., 2023a) is an approach to measure the hallucination without the need for human-annotated ground-truth answers. GPT-4 takes the generated captions with bounding box coordinates as the image content and compares human instructions and model response. Then, ask GPT-4 to score the answers based on two criteria: (1) Accuracy: whether the response hallucinates with the image content. (2) Relevancy: whether the response directly follows the instruction. It is composed of 1k questions and uses accuracy and relevancy for evaluation.

MMHal-Bench. MMHal-Bench (Sun et al., 2023) has a focus on penalizing hallucinations

with 96 image-question pairs, ranging in 8 question categories and 12 object topics from OpenImages (Kuznetsova et al., 2018). It uses GPT-4 to compare the model's response to the correct answer based on the given object information. If the score is below 3, it is considered to have hallucinations.

4.2 Baselines

We use 4 recently released LVLMs as baselines: (1) MiniGPT4 (Zhu et al., 2023); (2) InstructBLIP (Dai et al., 2023); (3) LLaVA-1.5 (Liu et al., 2023b); (4) mPLUG-Owl2 (Ye et al., 2023). All models above have been tuned on their visual instruction data.

4.3 Implementation Details

We randomly select 5,000 images from the training sets of MSCOCO 2014 (Lin et al., 2014) and Flickr30k (Plummer et al., 2017). Given that each image corresponds to multiple short captions, we prompt GPT-4 to synthesize a longer caption for each image based on these short captions (see Appendix A). Then, we employ LoRA-tuning (Hu et al., 2022) and deepspeed zero stage 3 to conduct minimal training on LLaVA-1.5-7b and mPLUG-Owl2-7b for 1 epoch. We use the AdamW (Loshchilov and Hutter, 2019) optimizer for optimization purposes. The learning rate and weight decay are set to 0.0001 and 0.1, respectively. Dur-

Method	Model	$C_s\downarrow$	$C_i\downarrow$	Len
Beam_5	mPLUG-Owl2	46.6	14.5	**68.4**
	mPLUG-Owl2 (w/ ours)	**22.4**	**8.4**	67.1
	LLaVA-1.5	15.4	8.2	**63.2**
	LLaVA-1.5 (w/ ours)	**14.6**	**6.1**	57.8
Opera	mPLUG-Owl2	39.8	13.1	**64.5**
	mPLUG-Owl2 (w/ ours)	**21.4**	**8.2**	53.3
	LLaVA-1.5	14.1	6.1	58.9
	LLaVA-1.5 (w/ ours)	**13.7**	**5.0**	**59.9**
Vep	mPLUG-Owl2	36.2	13.0	**65.1**
	mPLUG-Owl2 (w/ ours)	**20.6**	**7.8**	54.0
	LLaVA-1.5	11.0	6.2	59.7
	LLaVA-1.5 (w/ ours)	**9.6**	**4.9**	**60.8**

Table 2: **CHAIR hallucination evaluation results.** "w/ ours" means the use of hierarchical feedback learning, and "*Vep*" is the vision-enhanced penalty decoding.

Model	Relevancy	Accuracy
MiniGPT-4	3.84	5.35
InstructBLIP	6.27	5.83
mPLUG-Owl2	8.29	5.68
mPLUG-Owl2 (w/ ours)	**8.88**	**6.12**
LLaVA-1.5	7.56	5.49
LLaVA-1.5 (w/ ours)	**7.98**	**6.01**

Table 3: **Evaluation results on GAVIE.** The metric scores of Relevancy and Accuracy are from 0 to 10. "w/ ours" means the application of HELPD.

ing the training process, we initiate a warm-up ratio of 0.03, after which we apply the cosine schedule to decay the learning rate. We set σ to 0.6. The values of c for LLaVA-1.5 and mPLUG-Owl2 are set to 0.7 and 0.8. Each model requires approximately 4 hours to train with 2 NVIDIA 3090 24Gb GPUs.

5 Results

5.1 Main Results

In general, it is noticeable that the application of HELPD with various LVLMs is able to enhance their performance across different evaluation metrics, compared to the original LVLMs.

Upon examining the results from the POPE benchmark, as detailed in Table 1, it is evident that the hierarchical feedback learning has led to enhancements in the accuracy, precision, and F1 score. This suggests that our proposed framework can provide effective hallucination detection and feedback during training, combining object entities and semantic information to guide the model in enhancing its ability to discern hallucinated objects.

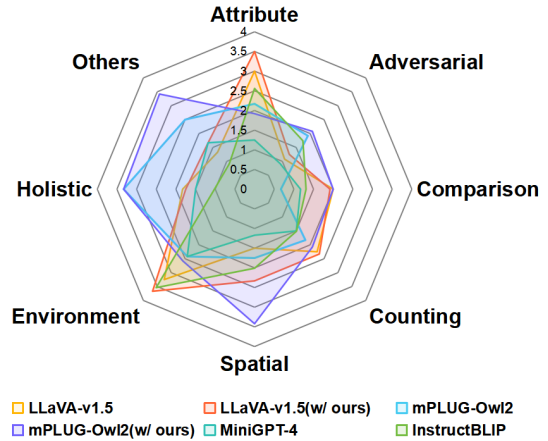As shown in Table 2, from the score of $CHAIR_s$ and $CHAIR_i$, it is evident that with the help of

Figure 5: **Detailed performance of LVLMs on the eight categories in MMHAL-Bench,** where "Overall" indicates the averaged performance across all categories. "w/ ours" means the application of HELPD.

HELPD, both mPLUG-Owl2 and LLaVA-1.5 have demonstrated varying degrees of hallucination reduction. Specifically, the trained mPLUG-Owl2, under various decoding methods, is able to reduce the $CHAIR_s$ by an average of 19.4 and the $CHAIR_i$ by an average of 5.4. This indicates that our proposed framework can effectively mitigate the generation of hallucinations, whether at the instance level or the sentence level. Moreover, it can be observed that the trained LVLMs do not exhibit significant fluctuations in the length of generated text. In most cases, LLaVA-1.5 can even increase the average generated length by 1.1 to 4.0. This illustrates that HELPD, while enhancing the anti-hallucination capabilities of LVLMs, does not excessively interfere with the generated length.

Table 3 shows the performance of different LVLMs on the GAVIE benchmark, which asks GPT-4 to pretend to be a smart teacher and scores (0-10) the answers according to the image content and instructions. The trained models achieve improvements in both accuracy and relevancy. Specifically, the trained mPLUG-Owl2 attains scores of 8.88 and 6.12 in relevancy and accuracy respectively, surpassing the provided baseline models. This demonstrates that our proposed framework can aid LVLMs in more directly following instructions, and the generated responses are more accurate concerning the image content.

Detailed performance of LVLMs on the eight categories in MMHAL-Bench is shown in Figure 5. It is evident that the trained models surpass their corresponding baseline models in performance across all eight question categories, and achieve a score of over 3 in five categories, including Object attribute

Model	R_{obj}	R_{sen}	$C_s\downarrow$	$C_i\downarrow$	Rel	Acc
mPLUG-Owl2	✗	✗	46.6	14.5	8.2	5.6
	✓	✗	31.1	11.2	8.4	5.9
	✗	✓	25.9	9.9	8.7	5.9
	✓	✓	**22.4**	**8.4**	**8.8**	**6.1**
LLaVA-1.5	✗	✗	15.4	8.2	7.5	5.4
	✓	✗	14.9	7.1	7.6	5.7
	✗	✓	15.8	6.8	7.7	5.8
	✓	✓	**14.6**	**6.1**	**7.9**	**6.0**

Table 4: **Ablation results on different levels of feedback on CHAIR and GAVIE.** R_{obj} and R_{sen} represent object-level and sentence-level feedback, respectively.

and Spatial relation. This implies that their generated texts are somewhat informative and exhibit almost no hallucination.

5.2 Further Analysis

Break-down Study of Hierarchical Feedback Learning. As illustrated in Section 3.1. In order to detect hallucinations of different granularities during training and provide feedback for parameter updates, we introduce hallucination feedback at both the object and sentence levels. To verify whether these two types of feedback contribute to the mitigation of hallucination, we conduct ablation experiments, and the results are presented in Table 4. Both object-level and sentence-level feedback can aid in alleviating hallucination, making the generated text adhere more closely to instructions and rendering it more accurate concerning the image content. It can also be observed that, compared to object-level feedback, sentence-level feedback can more effectively enhance the model's ability to resist hallucination. We hypothesize that this is because object-level feedback is more uncontrollable, such as possible omissions in the process of object extraction, or score reductions due to the presence of synonyms. However, the sentence-level feedback generated by prompting GPT-4 can effectively compensate for the deficiencies of the object-level feedback, thereby enhancing the overall performance of hierarchical feedback learning.

The Timing of Incorporating Hierarchical Feedback Learning. To investigate at which stage of training the integration of hierarchical feedback learning can better enhance the model's anti-hallucination capabilities, we also conduct an ablation study on the hyperparameter c. The experimental results of LLaVA-V1.5 on the random set of POPE are shown in Table 5, with more details available in the Appendix B. It indicates that LLaVA-

Model	c	Precision	Recall	F1 Score
LLaVA-1.5	0.6	87.01	92.05	89.45
	0.7	**87.84**	91.98	**89.86**
	0.8	86.21	93.07	89.50
	0.9	86.09	**93.33**	89.56
	1.0	86.11	91.33	88.74

Table 5: **Ablation results on the timing of incorporating HELPD.** We only show the random set results on LLaVA-v1.5, more details can be seen in Appendix B.

Method	MiniGPT-4	InstructBLIP	mPLUG-Owl2	LLaVA-v1.5
Nucleus	58.6	78.9	82.9	82.3
Beam$_5$	69.2	82.1	84.7	84.7
Opera	73.3	84.7	85.1	85.4
Vep	**74.1**	**85.0**	**85.3**	**85.6**

Table 6: **Ablation results on the decoding strategy.** We exhibit the average F1-score computed on random, popular, and adversarial splits of POPE.

V1.5 exhibits fewer hallucinations when $c = 0.7$, while mPLUG-Owl2 performs better when $c = 0.8$. Therefore, we default to assigning $c = 0.7$ for LLaVA-V1.5 and $c = 0.8$ for mPLUG-Owl2.

Different Decoding Strategy. Based on the observations of the attention matrix, we propose the vision-enhanced penalty decoding based on opera. To validate its effectiveness, we conduct an ablation study on LVLMs. The experimental results are shown in Table 2 and Table 6. As can be observed, compared to the baseline decoding strategy, the vision-enhanced penalty decoding demonstrates superior performance on benchmarks such as POPE and CHAIR, and has a smaller impact on the length of the generated text. It should be noted that this decoding strategy pays more attention to the hallucination performance of long texts.

6 Conclusion

In this paper, we aim to alleviate hallucinations in Large Vision-Language Models (LVLMs), and propose the HELPD framework, which employs the hierarchical feedback learning for small amounts of training on the model. To enhance attention to the visual modality, we also propose a vision-enhanced penalty decoding strategy. To evaluate the effectiveness of our approach, we conduct evaluations on numerous benchmarks. Experimental results demonstrate that our proposed framework effectively mitigates hallucination for different LVLMs without impacting sentence length and concurrently improves their text generation quality. Future work could focus on a more comprehensive evaluation of hallucination at different granularities.

1775

Limitations

Although HELPD effectively mitigates the hallucination in VLVMs, it remains subject to certain limitations. Firstly, to further train LVLMs, even for minimal training, a rich corpus of modality-aligned data is required. Secondly, compared to traditional decoding strategies, our proposed vision-enhanced penalty decoding may slightly increase decoding time, thereby potentially limiting inference speed.

Ethics Statement

The data (Lin et al., 2014; Plummer et al., 2017) used in our work is all drawn from open-source datasets. The data and text involved in our research do not involve private information and social issues.

Acknowledgments

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022.*

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings,* *Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Slav Petrov, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy P. Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul Ronald Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Andy Brock, Soham De, Samuel L. Smith, and Karen Simonyan. 2021. High-performance large-scale image recognition without normalization. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 1059–1071. PMLR.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. *CoRR*, abs/2309.03883.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang,

Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *CoRR*, abs/2305.06500.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *CoRR*, abs/2309.11495.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. 2023. EVA: exploring the limits of masked visual representation learning at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 19358–19369. IEEE.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. *CoRR*, abs/2403.14003.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 18135–18143. AAAI Press.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2023. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *CoRR*, abs/2311.17911.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. 2018. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR*, abs/1811.00982.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. 2023. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. *CoRR*, abs/2311.07362.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023a. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023b. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 12286–12312. Association for Computational Linguistics.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 292–305. Association for Computational Linguistics.

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

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, volume 8693 of *Lecture Notes in Computer Science*, pages 740–755. Springer.

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. *arXiv preprint arXiv:2306.14565*, 1(2):9.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023b. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Sean O'Brien and Mike Lewis. 2023. Contrastive decoding improves reasoning in large language models. *CoRR*, abs/2309.09117.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

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. *CoRR*, abs/2306.14824.

Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2017. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *Int. J. Comput. Vis.*, 123(1):74–93.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 1339–1384. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

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, Brussels, Belgium, October 31 - November 4, 2018*, pages 4035–4045. Association for Computational Linguistics.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, Weixin Liu, Zhihua Wu, Weibao Gong, Jianzhong Liang, Zhizhou Shang, Peng Sun, Wei Liu, Xuan Ouyang, Dianhai Yu, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *CoRR*, abs/2107.02137.

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. *CoRR*, abs/2309.14525.

Richard S. Sutton, David A. McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12, [NIPS Conference, Denver, Colorado, USA, November 29 - December 4, 1999]*, pages 1057–1063. The MIT Press.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models. *CoRR*, abs/2307.09288.

Liam van der Poel, Ryan Cotterell, and Clara Meister. 2022. Mutual information alleviates hallucinations in abstractive summarization. In *Proceedings of*

the *2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5956–5965. Association for Computational Linguistics.

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. 2023. Evaluation and analysis of hallucination in large vision-language models. *CoRR*, abs/2308.15126.

Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *MultiMedia Modeling - 30th International Conference, MMM 2024, Amsterdam, The Netherlands, January 29 - February 2, 2024, Proceedings, Part IV*, volume 14557 of *Lecture Notes in Computer Science*, pages 32–45. Springer.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyu Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open large-scale language models. *CoRR*, abs/2309.10305.

Dingchen Yang, Bowen Cao, Guang Chen, and Changjun Jiang. 2024. Pensieve: Retrospect-then-compare mitigates visual hallucination. *CoRR*, abs/2403.14401.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. A survey on multimodal large language models. *CoRR*, abs/2306.13549.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, Weng Lam Tam, Zixuan Ma, Yufei Xue, Jidong Zhai, Wenguang Chen, Zhiyuan Liu, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. GLM-130B: an open bilingual pre-trained model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

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. *CoRR*, abs/2310.01779.

Duzhen Zhang, Yahan Yu, Chenxing Li, Jiahua Dong, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *CoRR*, abs/2401.13601.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. 2023. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *CoRR*, abs/2311.16839.

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. *CoRR*, abs/2310.00754.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Prompts

We show the prompt for generating sentence-level feedback score in Figure 6 and synthesizing longer captions for each image based on corresponding five short captions in Figure 7 with GPT-4.

B More Experimental Results

B.1 More Experimental Results Compared to Existing Methods.

For POPE benchmark in Table 9, compared to existing methods (Liu et al., 2023a; Favero et al., 2024; Yang et al., 2024), our approach performs well on both the random and popular test sets, but falls short in terms of accuracy on the adversarial test set. This is because the method by Zhou et al. (2023) uses multiple similar images for comparison, enabling multi-angle judgment of the test problem. In contrast, our proposed method requires less computational resources and is able to surpass the performance in terms of F1 score. While for CHAIR in Table 10, compared to existing work, we are able to demonstrate better performance.

B.2 Ablation Results About the Timing of Incorporating Hierarchical Feedback Learning.

In Section 5.2, to investigate at which stage of training the integration of Hierarchical Feedback

Learning can better enhance the model's anti-hallucination capabilities, we conduct an ablation study on the hyperparameter c. We show the additional results in Table 7.

B.3 Further Analysis About the Quality of the Generated Text.

Considering that interventions in the decoding process can impact the quality of the generated text, we also conduct experiments to measure the impact caused by the vision-enhanced penalty decoding. Evaluation was carried out on the generated text from CHAIR, using the BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), METEOR (Banerjee and Lavie, 2005), and SPICE (Anderson et al., 2016) metrics, with the experimental results presented in Table 8. It can be observed that our proposed vision-enhanced penalty decoding performs comparably to beam search in terms of text generation metrics, without demonstrating excessive decline. It even surpasses beam search on BLEU1, BLEU4, and ROUGE-L. As can also be seen from Table 2, compared to beam search and opera decoding, our proposed method is able to maintain the length of the generated sentences as well. This elucidates that our method can maintain the quality of the generated text while mitigating hallucinations. Additionally, to verify whether HELPD can mitigate hallucinations while preserving general capabilities, we conduct evaluations on VQA-v2 (Goyal et al., 2017) and MME (Fu et al., 2023). As shown in Table 11 and Table 12, LVLMs with HELPD can maintain relatively stable performance across various metrics, demonstrating that the framework does not significantly impair the model's foundational abilities.

C Cases

We show some generation cases in Figure 8, 9, and 10.

You will be presented with two pieces of text that describe the same image. First text comes from a dataset and can be considered as the label, and the other is generated by a model. Your task is to compare these two pieces of text and evaluate them from the perspective of hallucination. If the contents described by the two pieces of text are completely consistent and there is no hallucination, please give a score of 1. If hallucination is present in the model-generated text, i.e., it describes content not present in the label text, please give a score closer to 0. Note that the lower the score, the less obvious the hallucination. The score range is between 0 and 1. Please only provide a score and do not provide reason.

Output example 1:
Dataset text (Label): Label 1;
Model generated text: Text 1;
score: score 1

Output example 2:
Dataset text (Label): Label 2;
Model generated text: Text 2;
score: score 2

Output example 3:
Dataset text (Label): Label 3;
Model generated text: Text 3;
score: score 3

Output format:
Dataset text (Label):
Model generated text:
score:

Figure 6: **Prompt for generating sentence-level feedback score.**

Given the following five captions of the same image, please combine them into a single, comprehensive caption that includes all the information, especially objects, without any repetition:

Output example 1:
Caption 1: caption_1;
Caption 2: caption_2;
Caption 3: caption_3;
Caption 4: caption_4;
Caption 5: caption_5;
Combined caption 1: response_1

Output example 2:
Caption 1: caption_1;
Caption 2: caption_2;
Caption 3: caption_3;
Caption 4: caption_4;
Caption 5: caption_5;
Combined caption 2: response_2

Output example 3:
Caption 1: caption_1;
Caption 2: caption_2;
Caption 3: caption_3;
Caption 4: caption_4;
Caption 5: caption_5;
Combined caption 3: response_3

Output format:
Caption 1:
Caption 2:
Caption 3:
Caption 4:
Caption 5:
Combined caption:

Figure 7: **Prompt for synthesizing longer captions for each image based on corresponding five short captions.**

POPE	Model	c	Precision	Recall	F1 Score	Yes (%)
Random	mPLUG-Owl2	0.6	88.7	86.1	87.3	46.5
		0.7	88.9	**85.9**	87.3	46.4
		0.8	89.9	85.6	**87.7**	47.6
		0.9	**90.1**	85.2	87.6	47.3
	LLaVA-1.5	0.6	87.0	92.1	89.5	52.9
		0.7	**87.8**	91.9	**89.9**	52.4
		0.8	86.2	93.1	89.5	54.6
		0.9	86.1	**93.3**	89.7	54.3
Popular	mPLUG-Owl2	0.6	82.8	85.9	84.3	52.0
		0.7	84.2	**86.1**	85.1	52.0
		0.8	85.6	85.6	**85.6**	50.0
		0.9	**85.8**	85.2	85.5	49.6
	LLaVA-1.5	0.6	80.1	93.3	86.2	58.2
		0.7	**81.4**	92.0	**86.4**	56.5
		0.8	79.7	93.0	85.8	58.3
		0.9	79.5	**93.5**	85.9	58.8
Adversarial	mPLUG-Owl2	0.6	79.2	85.9	82.5	54.1
		0.7	79.8	85.9	82.7	54.2
		0.8	80.2	85.6	82.8	53.3
		0.9	80.7	85.2	82.9	52.8
	LLaVA-1.5	0.6	71.2	93.3	80.8	65.6
		0.7	**72.1**	92.0	**80.9**	63.8
		0.8	**72.1**	91.9	80.8	63.8
		0.9	70.8	**93.5**	80.6	66.0

Table 7: **Additional ablation results on the timing of incorporating HELPD.**

Model	Method	B1	B4	M	R-L	S
LLaVA-1.5	greedy	20.8	5.1	19.9	22.4	22.3
	beam$_5$	21.6	5.3	**21.2**	23.2	**23.4**
	vep	**22.3**	**5.5**	20.8	**23.3**	22.9
mPLUG-Owl2	greedy	18.9	4.8	16.7	17.1	16.9
	beam$_5$	19.8	5.1	**17.2**	**19.7**	**18.3**
	vep	**20.1**	**5.2**	**17.2**	18.9	18.1

Table 8: **Evaluation of the quality of the generated text from CHAIR.** B1, B4, M, R-L, and S are abbreviations for BLEU1, BLEU4, METEOR, ROUGE-L, and SPICE. "vep" represents the vision-enhanced penalty decoding.

Method	*Random*		*Popular*		*Adversarial*	
	Accuracy	F1 Score	Accuracy	F1 Score	Accuracy	F1 Score
Liu et al. (2023a)	87.3	87.3	73.4	80.1	65.0	73.9
Favero et al. (2024)	81.2	65.6	73.9	67.3	68.2	75.4
Yang et al. (2024)	87.5	86.1	85.1	84.8	**81.7**	80.7
HELPD (ours)	**89.6**	**89.8**	**85.7**	**86.6**	78.1	**81.1**

Table 9: **More experimental results on POPE benchmark.**

Method	CHAIR$_s$ ↓	CHAIR$_i$ ↓	Len
Zhou et al. (2023)	27.1	6.4	58.8
van der Poel et al. (2022)	16.2	6.7	59.6
Li et al. (2023b)	14.8	6.3	60.7
Liu et al. (2023a)	13.8	5.9	60.1
HELPD (ours)	**9.6**	**4.9**	**60.8**

Table 10: **More experimental results on CHAIR benchmark.**

Method	Yes/No	Number	Other	Overall
LLaVA-1.5	92.23	60.01	71.07	78.53
LLaVA-1.5 (w/ ours)	92.88	60.78	68.86	77.50
mPLUG-Owl2	91.96	63.24	70.51	79.05
mPLUG-Owl2 (w/ ours)	92.56	60.21	69.82	78.20

Table 11: **Experimental results on VQA-v2 benchmark.**

Method	Category		Total
	Perception	Cognition	
LLaVA-1.5	1508.5	355.7	1864.2
LLaVA-1.5 (w/ ours)	1505.6	342.8	1848.4
mPLUG-Owl2	1446.5	313.3	1759.8
mPLUG-Owl2 (w/ ours)	1434.5	312.7	1747.2

Table 12: **Experimental results on MME benchmark.**
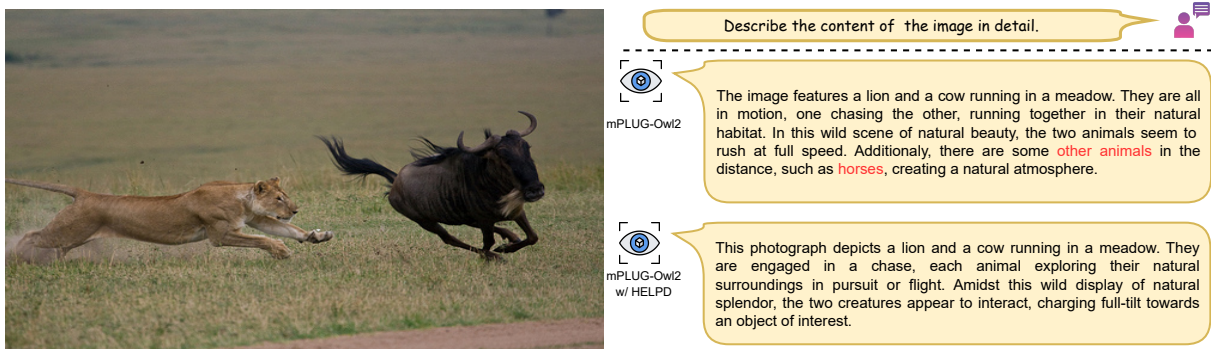
Figure 8: **An illustrative case is presented to compare the output of mPLUG-Owl2 and mPLUG-Owl2 with HELPD.**

Figure 9: **An illustrative case is presented to compare the output of LLaVA-v1.5 and LLaVA-v1.5 with HELPD.**

Figure 10: **An illustrative case is presented to compare the output of mPLUG-Owl2 and mPLUG-Owl2 with HELPD.**