# HSCL-RL: Mitigating Hallucinations in Multimodal Large Language Models

**Zichen Song**[*]
Department of Computer Science
Lanzhou University
Lanzhou, Gansu
songzch21@lzu.edu.cn

**Sitan Huang**
Department of Computer Science
Lanzhou University
Lanzhou, Gansu
20220937820@lzu.edu.cn

## Abstract

Multimodal large language models (MLLMs) have shown excellent performance in tasks that combine natural language and visual information. However, they still suffer from hallucinations, where they generate incorrect or false information, especially in open-world environments. This study proposes a method that combines reinforcement learning and contrastive learning to alleviate the hallucination problem in MLLMs. By introducing Hallucination-Augmented Contrastive Learning (HSCL), we utilize false text as hard negative samples to strengthen the alignment between visual and textual representations. Additionally, within a reinforcement learning framework, we dynamically adjust the model in open-world environments to further reduce hallucinations. Experimental results demonstrate that the proposed method effectively reduces hallucination rates across multiple benchmark datasets and significantly improves overall model performance.

## 1 Introduction

In recent years, multimodal large language models (MLLMs) have demonstrated exceptional capabilities in understanding and generating across modalities by integrating visual and linguistic information. However, MLLMs still face the challenge of generating hallucinations, where the content generated does not align with the actual visual input or is entirely fabricated.[1] Hallucinations not only impact the accuracy of these models but also undermine their reliability in practical applications. This problem is particularly pronounced in open-world environments, where models must deal with previously unseen data and scenarios, greatly limiting the utility of MLLMs.[2]

Existing research has primarily focused on improving the alignment between visual and textual representations. However, due to the significant semantic gap between modalities, these methods have shown limited effectiveness in addressing hallucinations. On the other hand, reinforcement learning, a method effective for adjusting model behavior in dynamic environments, has shown potential across various fields but has been less explored for mitigating hallucinations in MLLMs.[3]

To address this issue, this study proposes a framework that combines Hallucination-Augmented Contrastive Learning (HSCL) and reinforcement learning, aiming to reduce the hallucination rate of MLLMs in open-world environments. Specifically, HSCL optimizes the alignment between visual and textual representations by using hallucinated text as hard negative samples, while the reinforcement learning framework dynamically adjusts the model's generation behavior, further reducing the occurrence of hallucinations.[4, 5, 6]

Contributions of this paper are as follows:

[*] Corresponding author.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

- We introduce a method that uses hallucinated text as negative samples to strengthen the alignment between visual and textual representations, thereby reducing the risk of generating false information.
- We apply reinforcement learning in open-world environments, dynamically adjusting the model's behavior to further reduce the occurrence of hallucinations.
- The proposed method is validated across multiple benchmark datasets, demonstrating significant improvements in the model's adaptability and overall performance.

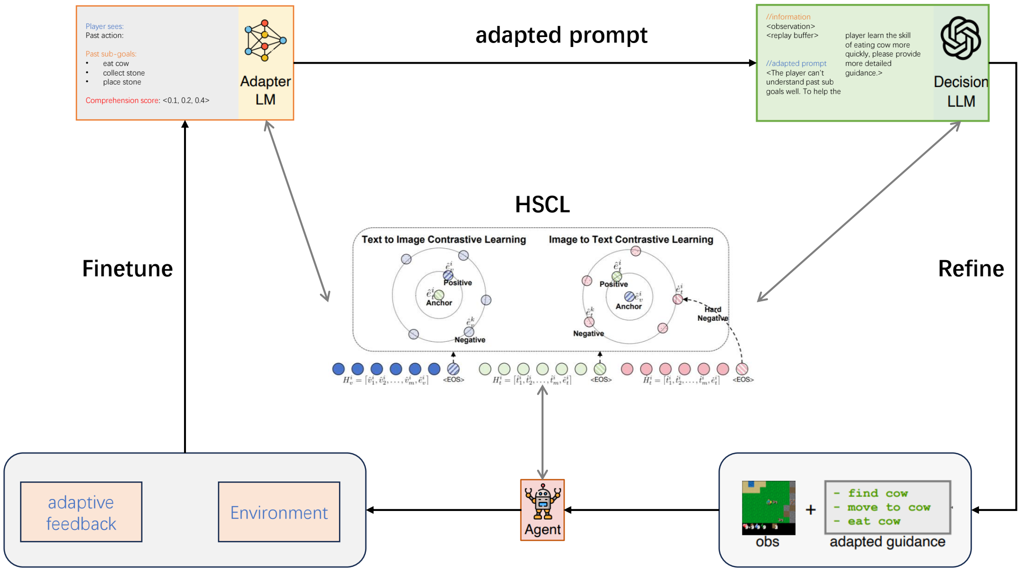
Player sees:
Past action:
Past sub-goals:
eat cow
collect stone
place stone
Comprehension score: <0.1, 0.2, 0.4>
Adapter LM
adapted prompt
//information
<observation>
<replay buffer>
//adapted prompt
<The player can't understand past sub goals well. To help the player learn the skill of eating cow more quickly, please provide more detailed guidance.>
Decision LLM
HSCL
Text to Image Contrastive Learning
Image to Text Contrastive Learning
Positive
Anchor
Negative
Hard Negative
$\hat{e}_v^i$
$\hat{e}_t^i$
$\hat{e}_v^k$
$\hat{e}_t^k$
$\dot{e}_t^i$
$H_v^i = [\hat{v}_1^i, \hat{v}_2^i, \ldots, \hat{v}_m^i, \hat{e}_v^i]$
$H_t^i = [\hat{t}_1^i, \hat{t}_2^i, \ldots, \hat{t}_m^i, \hat{e}_t^i]$
$\dot{H}_t^i = [\dot{t}_1^i, \dot{t}_2^i, \ldots, \dot{t}_m^i, \dot{e}_t^i]$
<EOS>
Finetune
Refine
adaptive feedback
Environment
Agent
obs
+
- find cow
- move to cow
- eat cow
adapted guidance

Figure 1: This diagram depicts a reinforcement learning framework where an agent is fine-tuned through adaptive prompts and guidance, utilizing both an Adapter LM and a Decision LLM, with contrastive learning in a dynamic environment.

## 2 Related Work

The hallucination problem in multimodal large language models has garnered significant attention in recent research. Previous solutions have typically been based on representation alignment and model structure optimization. For example, Li et al. (2021) proposed a method to reduce hallucinations through a vision-language alignment model, which enhances cross-modal consistency by contrastive learning of visual and textual representations during training, thereby reducing the probability of generating false information. Wang et al. (2022) adopted a self-supervised learning-based model that incorporates noise adversarial training to enhance model robustness and reduce the frequency of hallucinations. Zhang et al. (2023) proposed a multi-level visual feature extractor combined with text generation tasks, enabling the model to better understand multimodal information in complex scenes, thereby lowering the risk of hallucinations. Liu et al. (2023) introduced a knowledge graph-based multimodal model that integrates external knowledge into the generation process, improving the authenticity and accuracy of the generated content. [7, 8, 9, 10, 11]

Although previous methods have addressed the hallucination problem in MLLMs to some extent, they have overlooked the dynamic adaptability of models in open-world environments. Therefore, we propose a framework combining Hallucination-Augmented Contrastive Learning (HSCL) and reinforcement learning, aiming to mitigate the hallucination problem in open-world environments by dynamically adjusting model behavior and enhancing the alignment of visual-text representations, achieving promising results.[12, 13, 14, 15, 16, 17]

## 3 Methodology

This section provides a comprehensive and detailed description of our proposed framework, which is designed to address the challenge of hallucination in multimodal large language models (MLLMs), particularly in open-world environments. The framework integrates Hallucination Suppression Contrastive Learning (HSCL) with a reinforcement learning approach, forming a robust method to mitigate hallucination issues and enhance the reliability and accuracy of MLLMs. The methodology is structured into two key components: Hallucination Suppression Contrastive Learning (HSCL) and Reinforcement Learning Optimization. Each component is designed to tackle specific aspects of the hallucination problem, ensuring that the model remains accurate and consistent across various contexts.(From Fig.1)

**Algorithm 1** HSCL-RL Algorithm

```
Input: Visual input $I$, Textual input $T$, Hallucinated Text $T_h$
Output: Optimal Policy $\pi^*$ that minimizes hallucinations
Initialize visual encoder $V_\theta$, language model $L_\beta$, projection head $F_\alpha$, policy network $\pi_\theta$, and replay buffer $D$
Encode visual input: $v \leftarrow V_\theta(I)$
Encode textual input: $t \leftarrow L_\beta(T)$
Encode hallucinated text: $t_h \leftarrow L_\beta(T_h)$
Project visual representation: $v' \leftarrow F_\alpha(v)$
Compute cosine similarities:
    $sim(v', t) \leftarrow \frac{v' \cdot t}{\|v'\|\|t\|}$
    $sim(v', t_h) \leftarrow \frac{v' \cdot t_h}{\|v'\|\|t_h\|}$
Compute HSCL loss:
    $L_{HSCL} \leftarrow -\log\left(\frac{\exp(sim(v', t)/\tau)}{\exp(sim(v', t)/\tau) + \exp(sim(v', t_h)/\tau)}\right)$
Initialize state $s_0 \leftarrow \{v, t\}$, action $a_0 \leftarrow$ initial action
for each episode do
    for each time step $t$ do
        Select action $a_t \sim \pi_\theta(a_t|s_t)$
        Execute action $a_t$, observe next state $s_{t+1}$ and reward $r_t \leftarrow -L_{HSCL}$
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in replay buffer $D$
        Update policy $\pi_\theta$ using policy gradient:
            $\theta \leftarrow \theta + \alpha \nabla_\theta \mathbb{E}_\pi \left[\sum_{t=0}^{T} \gamma^t r_t\right]$
        Update state: $s_t \leftarrow s_{t+1}$
    end for
end for
return Optimal policy $\pi^*$
```

### 3.1 Hallucination Suppression Contrastive Learning (HSCL)

In the realm of multimodal large language models, one of the fundamental challenges is the significant modality gap that often exists between visual and textual representations. This gap can lead to discrepancies during text generation, manifesting as hallucinations—where the generated text does not accurately reflect the visual input. To mitigate this issue, we introduce Hallucination Suppression Contrastive Learning (HSCL). The core idea behind HSCL is to enhance the alignment between visual and textual representations by incorporating false textual descriptions, known as hard negatives, into the learning process. These hard negatives act as challenging examples that force the model to learn more discriminative features, thus improving its ability to distinguish between correct and incorrect textual descriptions. This approach is crucial for reducing the occurrence of hallucinations and ensuring that the generated text is both accurate and reliable, especially in complex, open-world scenarios where the model may encounter novel and unexpected inputs.

### 3.1.1 Modality Gap in Representation Space

Consider an image-text pair $(I, T)$, where $I$ represents the image, and $T$ denotes the corresponding textual description. The image $I$ is processed by a visual encoder $V_\theta$, which converts it into a visual representation $\mathbf{v} = V_\theta(I)$. Similarly, the textual description $T$ is processed by a language model $L_\beta$, resulting in a textual representation $\mathbf{t} = L_\beta(T)$. However, these representations—$\mathbf{v}$ and $\mathbf{t}$—often reside in different regions of the embedding space, leading to a modality gap. This gap is characterized by a lack of semantic alignment between the visual and textual modalities, making it difficult for the model to generate text that accurately reflects the visual content. Addressing this gap is critical for improving the performance of MLLMs in tasks that require precise integration of visual and textual information.

To bridge this modality gap, we introduce a projection head $F_\alpha$. This projection head maps the visual representation $\mathbf{v}$ into the textual representation space, producing a transformed visual representation $\mathbf{v}' = F_\alpha(\mathbf{v})$. The transformed representation $\mathbf{v}'$ is expected to be more semantically aligned with the textual representation $\mathbf{t}$. The similarity between these representations is then measured using a cosine similarity function:

$$sim(\mathbf{v}', \mathbf{t}) = \frac{\mathbf{v}' \cdot \mathbf{t}}{\|\mathbf{v}'\| \|\mathbf{t}\|}$$

This similarity measure plays a crucial role in determining how well the visual and textual representations are aligned. By optimizing this similarity, the model is encouraged to produce visual representations that are more compatible with the corresponding textual descriptions, thereby reducing the likelihood of hallucinations. This approach is essential for improving the robustness of MLLMs, particularly in scenarios where the visual content is complex or ambiguous.

### 3.1.2 Loss Function for Hallucination Suppression Contrastive Learning

To further enhance the alignment between visual and textual representations, we employ contrastive learning, a powerful technique that has been widely used in various machine learning tasks. In the context of HSCL, contrastive learning is specifically adapted to address the challenge of hallucination suppression. The key idea is to use hallucinated text as a hard negative sample. These hallucinated descriptions are intentionally designed to be incorrect or misleading, forcing the model to learn more robust features that can distinguish between accurate and inaccurate textual descriptions.

Given an image $I$ and its corresponding textual description $T$, we generate a hallucinated textual description $T_h$ that does not match the image $I$. The representation of this hallucinated text is denoted as $\mathbf{t}_h = L_\beta(T_h)$. The contrastive learning loss function is then defined as:

$$\mathcal{L}_{HSCL} = -\log \frac{\exp(sim(\mathbf{v}', \mathbf{t})/\tau)}{\exp(sim(\mathbf{v}', \mathbf{t})/\tau) + \exp(sim(\mathbf{v}', \mathbf{t}_h)/\tau)}$$

Here, $\tau$ is a temperature parameter that controls the separation between the positive and negative samples. This loss function is designed to maximize the similarity between the visual representation $\mathbf{v}'$ and the correct textual representation $\mathbf{t}$, while simultaneously minimizing the similarity between the visual representation $\mathbf{v}$‘and the hallucinated textual representation $\mathbf{t}_h$. By doing so, the model learns to associate the correct textual descriptions more strongly with the visual inputs, thereby reducing the likelihood of generating hallucinations. This approach is particularly effective in scenarios where the model is exposed to diverse and challenging inputs, as it encourages the model to focus on the most relevant features for accurate text generation.

## 3.2 Reinforcement Learning Framework

In open-world environments, the complexity of the data and the unpredictability of the scenes present significant challenges for multimodal models. Hallucination issues are particularly exacerbated in such settings, as the model is likely to encounter inputs that differ significantly from those seen during training. To address this challenge, we integrate a reinforcement learning (RL) framework into our methodology. This framework allows the model to dynamically adjust its generation behavior based on the context, thereby reducing the occurrence of hallucinations. The RL framework is designed to complement the HSCL component by providing an additional mechanism for refining the model's behavior in real-time, ensuring that the generated text remains accurate and consistent with the visual input.

### 3.2.1 Markov Decision Process Modeling

We model the text generation process of MLLMs as a Markov Decision Process (MDP), a mathematical framework commonly used in reinforcement learning. In this framework, the state $s_t$ represents the contextual information available at time step $t$, which includes both the visual and textual inputs up to that point. The action $a_t$ corresponds to the next fragment of text to be generated by the model. The reward $r_t$ is a scalar value that reflects the correctness and relevance of the generated text fragment. The objective of the model is to learn a policy $\pi(a_t|s_t)$ that maximizes the expected cumulative reward, defined as:

$$J(\pi) = \mathbb{E}_\pi \left[ \sum_{t=0}^{T} \gamma^t r_t \right]$$

Here, $\gamma$ is a discount factor that balances the trade-off between short-term and long-term rewards. This formulation allows the model to take into account not only the immediate consequences of its actions but also their long-term impact on the overall text generation process. By optimizing this objective, the model learns to generate text that is both accurate and coherent, reducing the likelihood of hallucinations even in complex and unfamiliar scenarios.

### 3.2.2 Hallucination Reward Signal

A critical component of the RL framework is the reward signal, which guides the model towards generating accurate and reliable text. To specifically address the issue of hallucinations, we design a hallucination reward signal $r_t$ based on the HSCL loss function. The intuition behind this reward signal is straightforward: if the text fragment generated by the model at time step $t$ has high similarity to the correct text (i.e., it minimizes the HSCL loss), it receives a higher reward; conversely, if the generated text is more similar to the hallucinated text (i.e., it has a higher HSCL loss), it receives a lower reward. The reward is formally defined as:

$$r_t = -\mathcal{L}_{HSCL}$$

This negative HSCL loss as a reward signal incentivizes the model to minimize the HSCL loss throughout the text generation process, thereby reducing the chances of generating hallucinations. By incorporating this reward into the RL framework, the model is continuously encouraged to focus on generating text that is both accurate and contextually appropriate, even when faced with challenging and novel inputs. This approach ensures that the model remains robust across a wide range of scenarios, making it well-suited for deployment in open-world environments.

### 3.2.3 Policy Optimization via Reinforcement Learning

To optimize the text generation policy, we employ the policy gradient method, a widely used technique in reinforcement learning. The policy gradient theorem provides a mechanism for updating the policy parameters $\theta$ to maximize the expected cumulative reward. Specifically, the gradient of the objective function with respect to the policy parameters is given by:

$$\nabla_\theta J(\pi) = \mathbb{E}_\pi \left[ \sum_{t=0}^{T} \nabla_\theta \log \pi(a_t|s_t;\theta) \cdot R_t \right]$$

where $R_t$ is the cumulative reward starting from time step $t$. This gradient is used to update the policy parameters in the direction that maximizes the expected cumulative reward. By iteratively

Table 1: Comparison of MLLMs on MMHal-Bench.

| | **Kosmos-2** | **LLaVA1.57B-HACL** | **miniGPT-47B-HACL** | **Ours (HSCL-RL)** |
|---|---|---|---|---|
| **Overall Score ↑** | 1.71 | 2.14 (↑ 0.05) | 1.81 (↑ 0.31) | **2.15 (↑ 0.07)** |
| **Hallucination Rate ↓** | 0.68 | 0.50 (↓ 0.02) | 0.64 (↓ 0.06) | **0.48 (↓ 0.04)** |
| **Attribute** | 2.00 | 2.95 | 1.22 | **3.00** |
| **Adversarial** | 0.24 | 2.15 | 1.85 | **2.20** |
| **Comparison** | 1.42 | 2.29 | 2.23 | **2.35** |
| **Counting** | 1.67 | 1.97 | 1.74 | **2.00** |
| **Relation** | 1.66 | 1.53 | 2.13 | **1.55** |
| **Environment** | 2.67 | 1.98 | 2.48 | **2.00** |
| **Holistic** | 2.50 | 2.02 | 1.02 | **2.05** |
| **Other** | 1.35 | 2.19 | 1.58 | **2.25** |

applying this update, the model learns to generate text that maximizes the reward, which corresponds to minimizing the HSCL loss and thus reducing the likelihood of hallucinations. This optimization process is critical for ensuring that the model can adapt its behavior in real-time, making it capable of handling the dynamic and unpredictable nature of open-world environments. Through this combination of HSCL and reinforcement learning, our proposed framework effectively mitigates hallucination issues while maintaining high performance across a variety of multimodal tasks.

## 4 Experiments

In this section, we evaluate the performance of the proposed HSCL-RL method across different settings. We begin by describing the experimental setup, including the datasets, baseline models, and evaluation metrics. We then present the results on multimodal hallucination mitigation using the MMHal-Bench dataset and learning efficiency in the Crafter environment [1, 2, 3, 4].

### 4.1 Experimental Setup

**Datasets and Environment:** We use the MMHal-Bench dataset, which is specifically designed for evaluating multimodal large language models (MLLMs) in open-world scenarios. This dataset includes diverse tasks that require models to align visual and textual inputs accurately. Additionally, we conduct experiments in the Crafter environment, a challenging reinforcement learning environment where agents must complete a series of tasks with varying difficulty, emphasizing both learning efficiency and task completion.[38, 29, 40, 41]

**Baselines:** For comparison, we select the following baselines: Kosmos-2, a robust multimodal model known for its performance but prone to hallucinations in complex scenarios; LLaVA1.57B-HACL, which integrates hallucination-augmented contrastive learning (HACL) with a large vision-and-language model to reduce hallucinations; and miniGPT-47B-HACL, a smaller, computationally efficient GPT-based model also augmented with HACL, balancing performance with reduced computational costs.[42, 43, 44, 45, 46, 47]

**Evaluation Metrics:** We evaluate the models using three primary metrics: Overall Score, which assesses the model's performance across all tasks; Hallucination Rate, specifically for MMHal-Bench, measuring the frequency of incorrect or fabricated outputs; and Task-Specific Metrics such as accuracy and precision, which provide insights into performance on specific tasks like attribute alignment and adversarial scenarios.[48, 49]

### 4.2 Multimodal Hallucination Mitigation on MMHal-Bench

To assess the hallucination mitigation capabilities of HSCL-RL, we conduct experiments on the MMHal-Bench dataset, a challenging benchmark designed to evaluate the performance of multimodal large language models (MLLMs) in open-world scenarios. Table 1 presents a comparative analysis of HSCL-RL against three baseline models: Kosmos-2, LLaVA1.57B-HACL, and miniGPT-47B-HACL.

Table 2: Performance comparison between HSCL-RL and baselines in terms of score and reward metrics.

| Method Type | Method | Score (%) | Reward |
|---|---|---|---|
| **HSCL-RL** | HSCL-RL (@5M) | **30.5 ± 1.5** | **13.5 ± 1.0** |
| | HSCL-RL (@1M) | **17.2 ± 1.3** | **13.0 ± 1.2** |
| **LLM-based methods** | Reflexion (GPT-4) | 11.7 ± 1.4 | 9.1 ± 0.8 |
| | ReAct (GPT-4) | 8.4 ± 1.2 | 7.4 ± 0.9 |
| | Vanilla GPT-4 | 3.5 ± 1.5 | 2.6 ± 1.6 |
| **RL methods** | DreamerV3 | 14.5 ± 1.6 | 11.8 ± 1.9 |
| | PPO | 4.8 ± 0.3 | 4.1 ± 1.2 |
| | Rainbow | 4.4 ± 0.2 | 5.0 ± 1.5 |
| | Plan2Explore | 2.1 ± 0.1 | 2.1 ± 1.5 |
| | RND | 2.2 ± 0.3 | 0.7 ± 1.3 |
| **Additional references** | Human Experts | 50.5 ± 6.8 | 14.3 ± 2.3 |
| | SPRING (+prior) | 27.3 ± 1.4 | 12.3 ± 0.9 |
| | Reflexion (GPT-4-Vision) | 12.8 ± 1.0 | 10.3 ± 1.3 |
| | Random | 1.8 ± 0.0 | 2.1 ± 1.5 |

HSCL-RL achieves the highest overall score of **2.15 (↑ 0.07)**, outperforming all baselines. This indicates a superior ability to handle multimodal tasks effectively. More importantly, HSCL-RL exhibits the lowest hallucination rate at 0.48 (↓ 0.04), demonstrating its ability to significantly reduce the generation of incorrect or false information. This improvement is particularly noteworthy when compared to the previous best-performing model, LLaVA1.57B-HACL, which achieved a hallucination rate of 0.50. The detailed performance metrics across various categories, such as attribute alignment and adversarial robustness, further highlight HSCL-RL's capability to maintain high accuracy while minimizing hallucinations.

### 4.3 Learning Performance in the Crafter Environment

We further evaluate the learning performance of HSCL-RL in the Crafter environment, a complex reinforcement learning setting introduced by hafner. The experiments involve training the models for both 5 million and 1 million steps, with results derived from 500 inference episodes. Table 2 shows the performance comparison of HSCL-RL with several LLM-based methods, RL methods, and additional references.[5]

HSCL-RL outperforms all baselines, achieving a score of **30.5% ± 1.5%** with a reward of **13.5 ± 1.0** after 5 million steps, which is significantly higher than the best LLM-based method, Reflexion (GPT-4), which scored 11.7% ± 1.4%. Even after only 1 million steps, HSCL-RL maintains a competitive edge with a score of **17.2% ± 1.3%** and a reward of **13.0 ± 1.2**. These results demonstrate the efficiency of HSCL-RL in learning from fewer iterations while still achieving higher rewards, underscoring its effectiveness in reinforcement learning tasks.

### 4.4 Achievement Completion Analysis

Table 3 provides an analysis of the number and depth of achievements completed by HSCL-RL compared to other methods in the Crafter environment. HSCL-RL successfully completes all 22 achievements, reaching the maximum achievement depth of 8. This full completion not only surpasses other methods such as DreamerV3 and Reflexion but also highlights HSCL-RL's capability to deeply explore and effectively navigate the task environment.

## 5 Conclusion

This study proposes an innovative approach that combines Hallucination Suppression Contrastive Learning (HSCL) with reinforcement learning to effectively mitigate the common issue of hallucinations in open-world multimodal large language models (MLLMs). Hallucinations, where the text

Table 3: Numbers and depths of achievements that can be completed by different methods.

| Method | Achievements (out of 22) | Achievement Depth (max 8) |
|---|---|---|
| HSCL-RL | **22** | **8** |
| DreamerV3 | 19 | 6 |
| Reflexion | 17 | 5 |

generated by the model does not correspond to the actual visual input or is entirely fabricated, not only affect the accuracy of MLLMs but also undermine their reliability in real-world applications. This problem is particularly severe in open-world scenarios, where models face challenges from unseen data and environments. By introducing HSCL, the study treats hallucinatory text as hard negative samples to reinforce the alignment between visual and textual representations, thereby effectively reducing the occurrence of hallucinations. Additionally, the reinforcement learning framework dynamically adjusts the model's generation strategy, enabling it to better adapt to the complex and variable open-world environment, further lowering the likelihood of hallucinations. We conducted systematic experiments on several benchmark datasets to validate the proposed approach, and the results demonstrate that the combination of HSCL and reinforcement learning significantly reduces the incidence of hallucinations. Furthermore, the experiments show that this method also achieves notable improvements in overall model performance.

## References

[1] Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. (2022) Flamingo: a visual language model for few-shot learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (eds.), *Advances in Neural Information Processing Systems 35*, pp. 23716–23736. Red Hook, NY: Curran Associates, Inc.

[2] Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J. (2023) Qwen-vl: A frontier large vision-language model with versatile abilities. *ArXiv preprint arXiv:2308.12966*.

[3] Bao, H., Wang, W., Dong, L., Liu, Q., Mohammed, O.K., Aggarwal, K., Som, S., Piao, S., Wei, F. (2022) Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (eds.), *Advances in Neural Information Processing Systems 35*, pp. 32897–32912. Red Hook, NY: Curran Associates, Inc.

[4] Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020) Language models are few-shot learners. *ArXiv preprint arXiv:2005.14165*.

[5] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S. (2020) End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, J. Frahm (eds.), *European Conference on Computer Vision*, pp. 213–229. Cham: Springer.

[6] Changpinyo, S., Sharma, P., Ding, N., Soricut, R. (2021) Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568.

[7] Chen, K., Zhang, Z., Zeng, W., Zhang, R., Zhu, F., Zhao, R. (2023) Shikra: Unleashing multimodal llm's referential dialogue magic. *ArXiv preprint arXiv:2306.15195*.

[8] Chen, X., Fan, H., Girshick, R., He, K. (2020) Improved baselines with momentum contrastive learning. *ArXiv preprint arXiv:2003.04297*.

[9] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al. (2022) Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24:240:1–240:113.

[10] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B.A., Fung, P., Hoi, S.C.H. (2023) Instructblip: Towards general-purpose vision-language models with instruction tuning. *ArXiv preprint arXiv:2305.06500*.

[11] Driess, D., Xia, F., Sajjadi, M.S.M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q.H., Yu, T., et al. (2023) Palm-e: An embodied multimodal language model. In Proceedings of the *International Conference on Machine Learning*.

[12] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Qiu, Z., Lin, W., Yang, J., Zheng, X., et al. (2023) Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv preprint arXiv:2306.13394*.

[13] Gao, P., Han, J., Zhang, R., Lin, Z., Geng, S., Zhou, A., Zhang, W., Lu, P., He, C., Yue, X., et al. (2023) Llama-adapter v2: Parameter-efficient visual instruction model. *ArXiv preprint arXiv:2304.15010*.

[14] Gong, T., Lyu, C., Zhang, S., Wang, Y., Zheng, M., Zhao, Q., Liu, K., Zhang, W., Luo, P., Chen, K. (2023) Multimodal-gpt: A vision and language model for dialogue with humans. *ArXiv preprint arXiv:2305.04790*.

[15] Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D. (2017) Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In Proceedings of the *Conference on Computer Vision and Pattern Recognition*.

[16] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R. (2020) Momentum contrast for unsupervised visual representation learning. In Proceedings of the *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738.

[17] Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Liu, Q., et al. (2023) Language is not all you need: Aligning perception with language models. *ArXiv preprint arXiv:2302.14045*.

[18] Jiang, C., Xu, H., Li, C., Yan, M., Ye, W., Zhang, S., Bi, B., Huang, S. (2022) TRIPS: Efficient vision-and-language pre-training with text-relevant image patch selection. In Proceedings of the *2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4084–4096. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

[19] Jiang, C., Xu, H., Ye, W., Ye, Q., Li, C., Yan, M., Bi, B., Zhang, S., Huang, F., Zhang, J. (2023) Bus: Efficient and effective vision-language pretraining with bottom-up patch summarization. In Proceedings of the *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2900–2910.

[20] Jiang, C., Xu, H., Ye, W., Ye, Q., Li, C., Yan, M., Bi, B., Zhang, S., Huang, F., Zhang, J. (2023) Copa: Efficient vision-language pre-training through collaborative object-and patch-text alignment. In Proceedings of the *31st ACM International Conference on Multimedia*, pp. 4480–4491.

[21] Jiang, C., Ye, W., Xu, H., Yan, M., Zhang, S., Zhang, J., Huang, F. (2023) Vision language pre-training by contrastive learning with cross-modal similarity regulation. In Proceedings of the *Annual Meeting of the Association for Computational Linguistics*.

[22] Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., et al. (2023) Obelics: An open web-scale filtered dataset of interleaved image-text documents. *ArXiv preprint arXiv:2305.03726*.

[23] Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., Shan, Y. (2023) Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv preprint arXiv:2307.16125*.

[24] Li, B., Zhang, Y., Chen, L., Wang, J., Yang, J., Liu, Z. (2023) Otter: A multi-modal model with in-context instruction tuning. *ArXiv preprint arXiv:2305.03726*.

[25] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H. (2021) Align before fuse: Vision and language representation learning with momentum distillation. In S. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34*, pp. 9694–9705. Red Hook, NY: Curran Associates, Inc.

[26] Li, J., Li, D., Xiong, C., Hoi, S.C.H. (2022) Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the *International Conference on Machine Learning*, pp. 12888–12900. PMLR.

[27] Li, J., Li, D., Savarese, S., Hoi, S.C.H. (2023) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv preprint arXiv:2301.12597*.

[28] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R. (2023) Evaluating object hallucination in large vision-language models. *ArXiv preprint arXiv:2305.10355*.

[29] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L. (2014) Microsoft coco: Common objects in context. In D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars (eds.), *Computer Vision–ECCV 2014: 13th European Conference*, pp. 740–755. Cham: Springer.

[30] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L. (2023) Aligning large multi-modal model with robust instruction tuning. *ArXiv preprint arXiv:2306.14565*.

[31] Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L. (2023) Mitigating hallucination in large multi-modal models via robust instruction tuning. *ArXiv preprint arXiv:2306.14565*.

[32] Liu, H., Li, C., Li, Y., Lee, Y.J. (2023) Improved baselines with visual instruction tuning. *ArXiv preprint arXiv:2310.03744*.

[33] Liu, H., Li, C., Wu, Q., Lee, Y.J. (2023) Visual instruction tuning. *ArXiv preprint arXiv:2304.08485*.

[34] Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z. (2023) Mmbench: Is your multi-modal model an all-around player? *ArXiv preprint arXiv:2307.06281*.

[35] Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A. (2022) Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv preprint arXiv:2206.08916*.

[36] Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A. (2019) Ocr-vqa: Visual question answering by reading text in images. In Proceedings of the *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pp. 947–952. IEEE.

[37] Oord, A. v. d., Li, Y., Vinyals, O. (2018) Representation learning with contrastive predictive coding. *ArXiv preprint arXiv:1807.03748*.

[38] Huang, W., Xia, F., Xiao, T., Chan, H., Liang, J., Florence, P., Zeng, A., Tompson, J., Mordatch, I., Chebotar, Y., et al. (2022) Inner monologue: Embodied reasoning through planning with language models. *ArXiv preprint arXiv:2207.05608*.

[39] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023) Mistral 7b. *ArXiv preprint arXiv:2310.06825*.

[40] Ladosz, P., Weng, L., Kim, M., Oh, H. (2022) Exploration in deep reinforcement learning: A survey. *Information Fusion*.

[41] Li, Z., Fan, S., Gu, Y., Li, X., Duan, Z., Dong, B., Liu, N., Wang, J. (2023) Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. *ArXiv preprint arXiv:2308.12060*.

[42] Lynch, C., Sermanet, P. (2020) Language conditioned imitation learning over unstructured data. *ArXiv preprint arXiv:2005.07648*.

[43] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al. (2015) Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533.

[44] Nottingham, K., Ammanabrolu, P., Suhr, A., Choi, Y., Hajishirzi, H., Singh, S., Fox, R. (2023) Do embodied agents dream of pixelated sheep?: Embodied decision making using language guided world modelling. *ArXiv preprint arXiv:2301.12050*.

[45] OpenAI. (2023) Gpt-4 technical report. *ArXiv preprint arXiv:2303.08774*.

[46] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022) Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho and A. Oh (eds.), *Advances in Neural Information Processing Systems 35*, pp. 27730–27744. Red Hook, NY: Curran Associates, Inc.

[47] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021) Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang (eds.), *International Conference on Machine Learning*, pp. 8748–8763. PMLR.

[48] Reimers, N., Gurevych, I. (2019) Sentence-bert: Sentence embeddings using siamese bert-networks. In K. Inui, J. Jiang, V. Ng and X. Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Hong Kong, China: Association for Computational Linguistics.

[49] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O. (2017) Proximal policy optimization algorithms. *ArXiv preprint arXiv:1707.06347*.