Carnegie Mellon University
School of Computer Science

ICLR
International Conference On Learning Representations

Self-Correcting Decoding with Generative Feedback for Mitigating Hallucinations in Large Vision-Language Models

Ce Zhang*[1] Zifu Wan*[1] Zhehan Kan[2] Martin Q. Ma[1] Simon Stepputtis[1]
Deva Ramanan[1] Russ Salakhutdinov[1] Louis-Philippe Morency[1] Katia Sycara[1] Yaqi Xie[1]

[1]Carnegie Mellon University [2]Tsinghua University

*Equal contribution.

{cezhang,zifuw,yaqix}@cs.cmu.edu

Poster Session 4 (April 25 Afternoon)

Project Page: https://zhangce01.github.io/DeGF/

Background

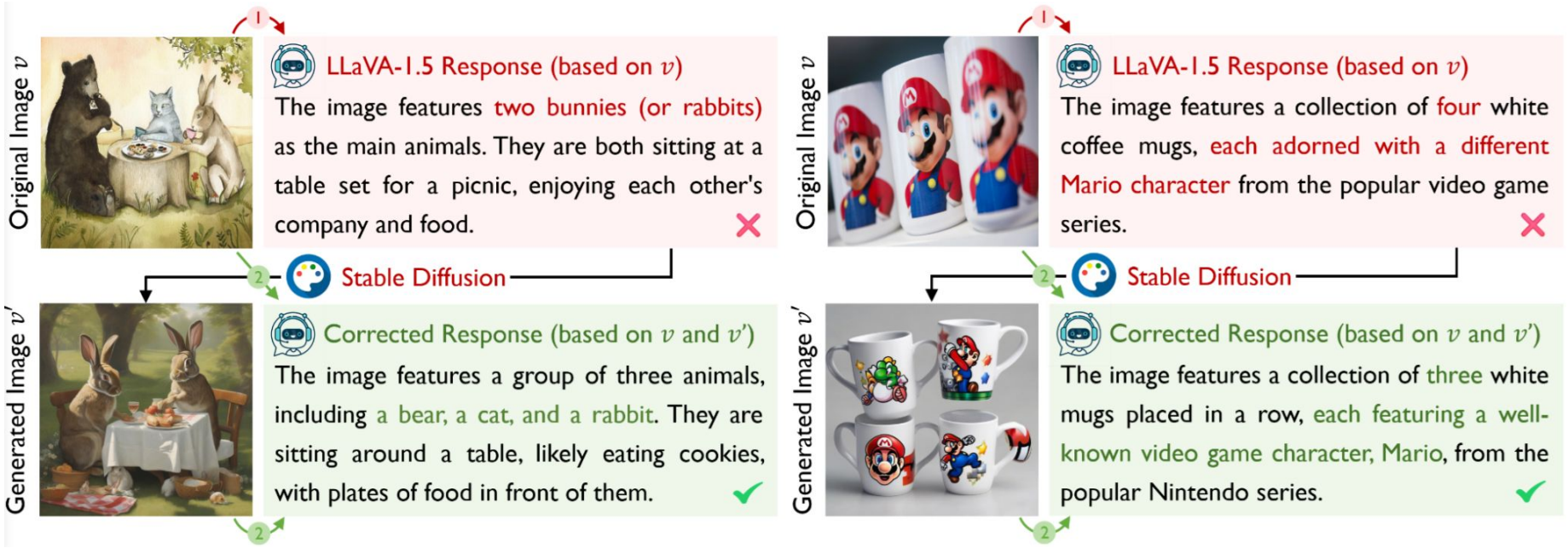
ICLR
International Conference On Learning Representations

- By extending the capabilities of powerful Large Language Models (LLMs) to incorporate visual inputs, recent Large Vision-Language Models (LVLMs) have demonstrated remarkable performance across various multi-modal tasks.
- Despite their proficiency in interpreting both visual and textual modalities, these models often suffer from hallucinations, where LVLMs erroneously produce responses that are inconsistent with the visual input.
- This potential for misinformation raises significant concerns, limiting the models' reliability and restricting their broader deployment in real-world scenarios.

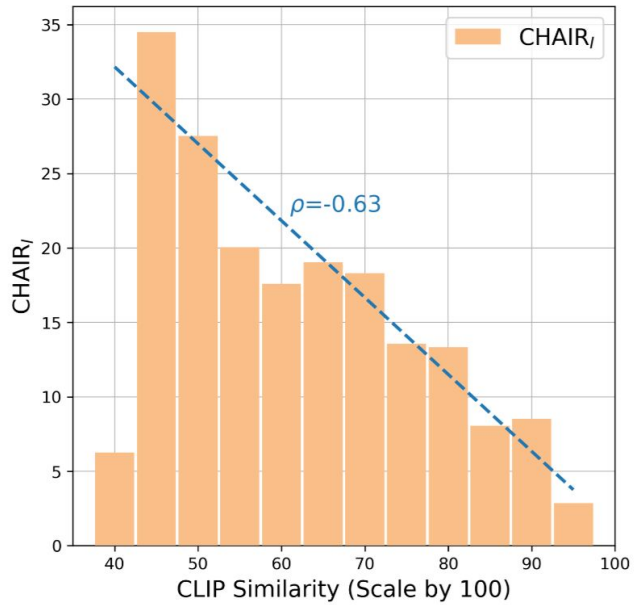
In this work, we explore the potential of leveraging powerful text-to-image generative models (e.g., Stable Diffusion) to mitigate various types of hallucinations in LVLMs.

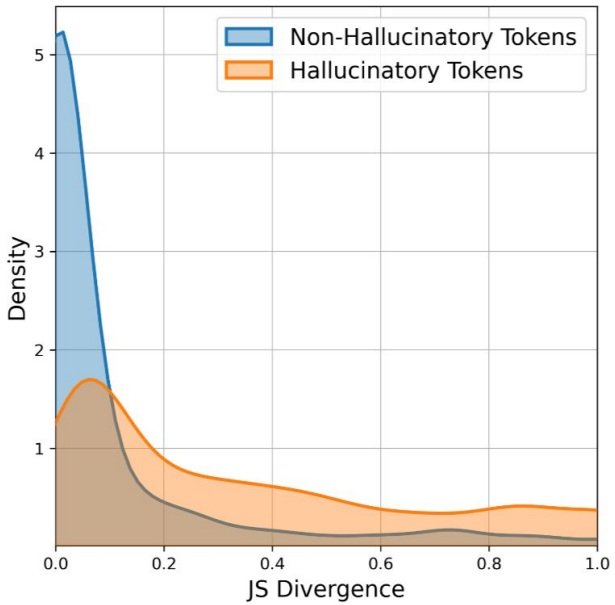
Motivation

- Our work is based on a simple yet intuitive hypothesis: Given a visual input and a textual prompt to an LVLM, if the generated response is accurate and non-hallucinatory, a text-to-image generative model should be able to reconstruct a similar image from that response.
- Alternatively, if there is a discrepancy between the original image and the generated image, this difference can serve as valuable self-feedback to correct potential hallucinations.



Generative Self-feedback



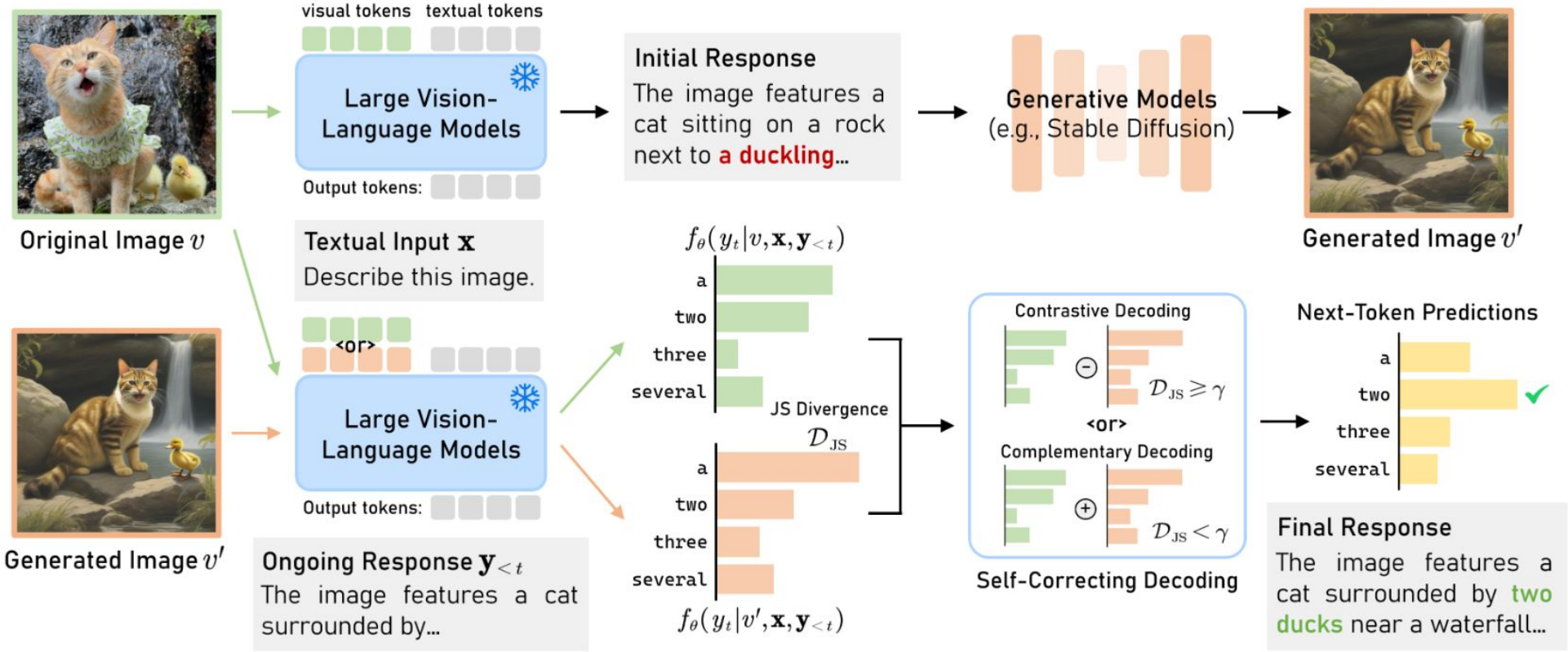


We validate that text-to-image generative models can provide valuable self-feedback for mitigating hallucinations at both the response and token levels:

- Lower similarity between the original image and generated image corresponds to higher rates of hallucinations at the response level.
- JS divergence between probabilities derived from the original and the generated image corresponds well to hallucinations at the token level.

Method

Building on this insight, we introduce self-correcting Decoding with Generative Feedback (DeGF), a novel training-free decoding algorithm that effectively incorporates feedback from text-to-image generative models to recursively enhance the accuracy of LVLM responses.



Method

ICLR
International Conference On Learning Representations

We generate two output distributions: one conditioned on the original image and the other conditioned on the synthesized visual reference. We then calculate the JS divergence on a token level.

$$p_\theta(y_t|v,\mathbf{x},\mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v,\mathbf{x},\mathbf{y}_{<t})],\quad p_\theta(y_t|v',\mathbf{x},\mathbf{y}_{<t}) = \texttt{Softmax}[f_\theta(y_t|v',\mathbf{x},\mathbf{y}_{<t})]$$

$$d_t(v,v') = \mathcal{D}_{\mathrm{JS}}\left(p_\theta\left(y_t|v,\mathbf{x},\mathbf{y}_{<t}\right) \parallel p_\theta\left(y_t|v',\mathbf{x},\mathbf{y}_{<t}\right)\right),$$

We consider two scenarios based on the token-level generative feedback:

- If the two predictions are aligned and both images agree on a specific token prediction, we confirm the original prediction as correct, and the auxiliary prediction from the generated image can be combined with the original prediction for enhancement.
- Conversely, if there is a significant discrepancy between the predictions, indicating that the original prediction is likely hallucinatory, we revise the original response by using the generated visual input as a contrasting reference to refine the initial next-token prediction.

Experiments

ICLR
International Conference On Learning Representations

- Performance comparisons on POPE

	Setup	Method	LLaVA-1.5			InstructBLIP			Qwen-VL		
			Acc. ↑	Prec. ↑	F1 ↑	Acc. ↑	Prec. ↑	F1 ↑	Acc. ↑	Prec. ↑	F1 ↑
MS-COCO	Random	Regular	83.13	81.94	83.44	83.07	83.02	83.08	87.43	93.56	86.48
		VCD	87.00	86.13	87.15	86.23	88.14	85.88	88.80	93.89	88.11
		M3ID	87.50	87.38	87.52	86.67	88.09	86.41	**89.83**	95.44	89.17
		RITUAL	88.87	89.23	**88.81**	**88.83**	90.48	**88.60**	89.47	**96.32**	88.62
		Ours	**89.03**	**91.20**	88.74	**88.83**	**93.73**	87.71	89.73	93.19	**89.31**
	Popular	Regular	81.17	78.28	82.08	77.00	73.82	78.44	84.70	88.24	83.96
		VCD	83.10	79.96	83.94	80.07	77.67	80.89	85.13	87.27	84.69
		M3ID	84.30	81.58	84.95	80.97	77.93	81.85	86.27	89.19	**85.73**
		RITUAL	85.83	84.17	86.17	81.97	78.90	**82.87**	84.57	84.09	84.67
		Ours	**86.63**	**87.75**	**86.28**	**82.73**	**84.02**	82.10	**86.50**	**89.87**	85.71
	Adversarial	Regular	77.43	73.31	79.26	74.60	71.26	76.45	79.83	80.13	79.73
		VCD	77.17	72.18	79.47	77.20	74.29	78.49	81.33	80.60	81.55
		M3ID	78.23	73.51	80.22	77.47	73.68	79.14	82.03	81.47	82.19
		RITUAL	78.80	74.43	80.54	78.73	74.57	**80.39**	82.80	83.15	82.71
		Ours	**81.63**	**80.59**	**81.94**	**80.30**	**80.90**	80.11	**83.47**	**84.49**	**82.98**

Our method consistently outperforms other decoding methods on three LVLMs, achieving state-of-the-art accuracies across all settings.

Experiments

ICLR
International Conference On Learning Representations

- Performance comparisons on CHAIR

Method	LLaVA-1.5				InstructBLIP			
	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Length ↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Length ↑
Regular	26.2	9.4	58.5	53.4	31.2	11.1	59.0	53.6
VCD	24.4	7.9	63.3	54.2	30.0	10.1	61.8	54.2
M3ID	21.4	6.3	**64.4**	53.5	30.8	10.4	62.6	53.4
RITUAL	22.4	6.9	63.0	**54.9**	26.6	8.9	63.4	55.3
Woodpecker	24.9	7.5	60.8	49.7	31.2	10.8	62.3	51.3
HALC	21.7	7.1	63.4	53.4	24.5	8.0	63.8	55.1
Ours	**18.4**	**6.1**	62.7	54.1	**24.0**	**7.7**	**67.2**	**55.5**

- We also compare the performance of our methods and other state-of-the-art methods in the open-ended captioning task and report the CHAIR scores, recall, and the average length of response.
- Specifically, our method outperforms the second-best approach by 3.0% and 2.6% on the CHAIRS metric, while also enhancing the detailedness of generated responses compared to regular decoding, as indicated by the higher recall and increased response length.

Experiments

ICLR
International Conference On Learning Representations

- Performance comparisons on MME and MMBench

Method	Object-level		Attribute-level		MME Score ↑	MMBench ↑
	Existence ↑	Count ↑	Position ↑	Color ↑		
Regular	173.75 (±4.79)	121.67 (±12.47)	117.92 (±3.69)	149.17 (±7.51)	562.50 (±3.96)	64.1
DoLa	176.67 (±2.89)	113.33 (±10.41)	90.55 (±8.22)	141.67 (±7.64)	522.22 (±16.78)	63.8
OPERA	183.33 (±6.45)	137.22 (±6.31)	122.78 (±2.55)	155.00 (±5.00)	598.33 (±10.41)	64.4
VCD	186.67 (±5.77)	125.56 (±3.47)	128.89 (±6.73)	139.45 (±12.51)	580.56 (±15.13)	64.6
M3ID	186.67 (±5.77)	128.33 (±10.41)	131.67 (±5.00)	151.67 (±20.88)	598.11 (±20.35)	64.4
RITUAL	187.50 (±2.89)	139.58 (±7.64)	125.00 (±10.27)	164.17 (±6.87)	616.25 (±20.38)	63.8
Woodpecker	187.50 (±2.89)	125.00 (±0.00)	126.66 (±2.89)	149.17 (±17.34)	588.33 (±10.00)	64.0
HALC	183.33 (±0.00)	133.33 (±5.77)	107.92 (±3.69)	155.00 (±5.00)	579.58 (±9.07)	64.2
Ours	**188.33** (±2.89)	**150.00** (±7.64)	**133.89** (±3.85)	**172.22** (±3.47)	**644.44** (±9.18)	**65.5**

- Beyond object hallucinations, we further compare the performance of our method with other approaches using the more comprehensive MME-Hallucination benchmark.
- This further illustrates the effectiveness of our approach in addressing a wide range of hallucinations.
- Our proposed DeGF enhances the general multi-modal understanding capabilities of LVLMs, as evidenced by its superior performance on the MMBench benchmark.

Qualitative Results

ICLR
International Conference On Learning Representations

- Case study on **LLaVA-Bench**

Visual Input:

Query: Describe this photo in detail.

[Regular Decoding] GPT-4V-Aided Evaluation: Accuracy: 4 Detailedness: 5

This image portrays a strategically placed tropical hill, punctured with a road branching out through it, giving the appearance of a mountain. The island below the mountain appears to have beautiful blue waters, making for a picturesque view. On the East coast of the bay, there are several houses and buildings, some situated on the beach near the water. The overall skyline captures a cohesive and captivating island landscape. Take a closer look at the cloudy sky, which creates a serene and natural atmosphere for the charming scene.

[Our Approach] GPT-4V-Aided Evaluation: Accuracy: 7 Detailedness: 7

The image displays a bird's eye view of an island with a large mountain in the center. The mountain is steep and appears to be a volcano, with greenery surrounding it. The island is located near the ocean, and it has a road running along the bottom of the mountain. The surrounding area is filled with houses, indicating that the island is inhabited. The island is surrounded by mountains, providing a picturesque and serene landscape. The combination of the ocean, the mountain, and the houses creates a unique and beautiful environment.

- Our response is more detailed, mentioning the volcano, the road, the surrounding greenery, and the inhabited areas, which gives a clearer understanding of the image's content.

Takeaways

ICLR
International Conference On Learning Representations

Our contributions can be summarized as follows:

- We investigate the potential of ***text-to-image generative models*** in mitigating hallucinations in LVLMs and demonstrate that text-to-image generative models can provide ***valuable self-feedback*** for mitigating hallucinations ***at both the response and token levels***.
- We propose self-correcting Decoding with Generative Feedback (DeGF), a training-free decoding algorithm for LVLMs that ***recursively enhances*** the accuracy of responses by integrating feedback from text-to-image generative models with ***complementary/contrastive decoding***.
- Extensive experimental evaluations ***across six benchmarks*** demonstrate that our DeGF consistently outperforms state-of-the-art approaches in effectively mitigating hallucinations in LVLMs.

Contact us: Ce Zhang (cezhang@cs.cmu.edu), Zifu Wan (zifuw@cs.cmu.edu), Yaqi Xie (yaqix@cs.cmu.edu)