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

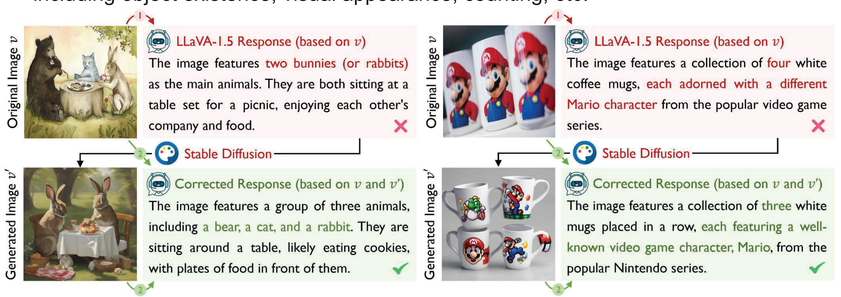
Ce Zhang* Zifu Wan* Zhehan Kan Martin Q. Ma Simon Stepputtis
Deva Ramanan Russ Salakhutdinov Louis-Philippe Morency Katia Sycara Yaqi Xie
** Co-first authors, equal contribution*

Paper | Project Page

ICLR International Conference On Learning Representations

Introduction

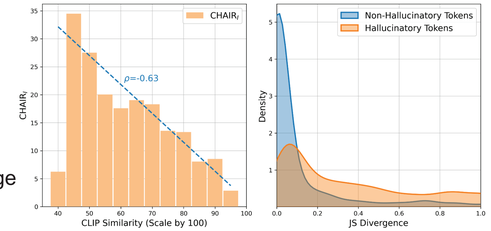
- **Background**: LVLMs often suffer from ***hallucinations***, generating responses that are inconsistent with the visual input, limiting the models' reliability in real-world scenarios.
- **Goal**: In this work, we explore the potential of ***leveraging powerful text-to-image generative models*** (*e.g.,* Stable Diffusion) to mitigate hallucinations in LVLMs.
- **Motivation**: Text-to-Image Generation ↻ Image-Conditioned Response Generation
  - ✓ If the generated response is ***non-hallucinatory***, a text-to-image generative model should be capable of reversing this process to ***produce a similar image***.
  - ⊗ Alternatively, if there is a ***discrepancy*** between the original image and the one generated from the response, this difference can serve as ***valuable self-feedback***, guiding the decoding process to ***correct hallucinations*** in the initial response.
- **Our Approach**: We propose DeGF, ***a novel training-free decoding algorithm*** for LVLMs that ***recursively*** enhances the accuracy of responses by integrating feedback from generative models with ***complementary/contrastive decoding***.
- **Results**: We demonstrate our DeGF can ***reduce various types of hallucinations***, including object existence, visual appearance, counting, *etc.*



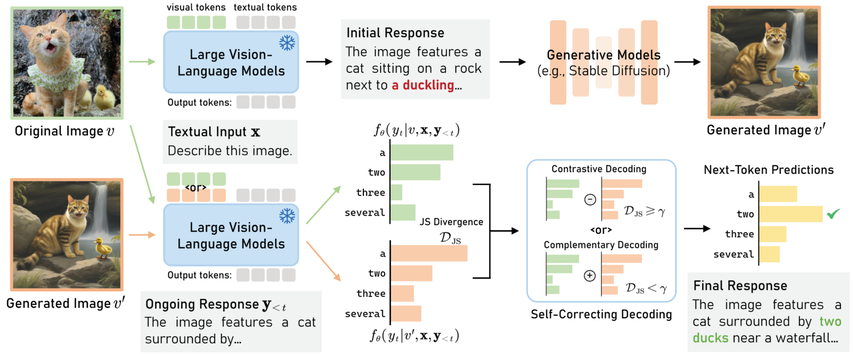
Empirical Study

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

- **Response Level**: *Lower similarity* between the original image and generated image corresponds to *higher rates of hallucinations*.
- **Token Level**: *JS divergence between probabilities* derived from the original and the generated image *corresponds well* to hallucinations.



DeGF: Self-Correcting Decoding with Generative Feedback



★ **Method Overview**: We propose a self-correcting decoding approach that leverages generative feedback to ***confirm or revise*** the initial response by ***selectively enhancing or contrasting*** the logits for each generated token based on the measured divergence between the two predicted probability distributions .

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

✓ **Complementary Decoding**: If two predictions are ***aligned*** 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 one for ***enhancement***.

⊗ **Contrastive Decoding**: Conversely, if there is a ***significant discrepancy*** between the predictions, we ***revise*** the original response by using the generated visual input as a ***contrasting reference*** to refine the initial next-token prediction.

★ **Efficiency**: Our approach involves ***two queries and incorporates a text-to-image generative model*** to mitigate hallucinations, resulting in a ***4.04×*** increase in latency and a ***1.21×*** increase in GPU memory usage.

Method	Avg. Latency ↓	GPU Memory ↓	CHAIR$_S$ ↓
Regular	3.44 s (×1.00)	15778 MB (×1.00)	55.0
VCD	6.91 s (×2.01)	16634 MB (×1.05)	54.4
OPERA	24.70 s (×7.18)	22706 MB (×1.44)	52.6
Woodpecker	10.68 s (×3.10)	22199 MB (×1.41)	57.6
HALC	22.61 s (×6.51)	23084 MB (×1.46)	51.0
Ours	13.89 s (×4.04)	19119 MB (×1.21)	48.8

Takeaways

- ✓ We investigate the potential of utilizing ***text-to-image generative models*** in mitigating hallucinations in LVLMs and demonstrate that these models can provide ***valuable self-feedback*** for mitigating hallucinations ***at both the response and token levels***.
- ✓ We propose ***DeGF***, a novel 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 results across ***six benchmarks*** demonstrate that our DeGF ***consistently outperforms*** previous approaches in mitigating hallucinations in LVLMs.

Experiments

Performance comparisons on **POPE** across ***3 LVLMs with different architectures***

Setup	Method	LLaVA-1.5 Acc. ↑	LLaVA-1.5 Prec. ↑	LLaVA-1.5 F1 ↑	InstructBLIP Acc. ↑	InstructBLIP Prec. ↑	InstructBLIP F1 ↑	Qwen-VL Acc. ↑	Qwen-VL Prec. ↑	Qwen-VL F1 ↑
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**

Performance comparisons on **MME**

Method	Object-level Existence ↑	Object-level Count ↑	Attribute-level Position ↑	Attribute-level Color ↑
Regular	173.75 (±4.79)	121.67 (±12.47)	117.92 (±3.69)	149.17 (±7.51)
DoLa	176.67 (±2.89)	113.33 (±10.41)	90.55 (±8.22)	141.67 (±7.64)
OPERA	183.33 (±6.45)	137.22 (±6.31)	122.78 (±2.55)	155.00 (±5.00)
VCD	186.67 (±5.77)	125.56 (±3.47)	128.89 (±6.73)	139.45 (±12.51)
M3ID	186.67 (±5.77)	128.33 (±10.41)	131.67 (±5.00)	151.67 (±20.88)
RITUAL	187.50 (±2.89)	139.58 (±7.64)	125.00 (±10.27)	164.17 (±6.87)
Woodpecker	187.50 (±2.89)	125.00 (±0.00)	126.66 (±2.89)	149.17 (±17.34)
HALC	183.33 (±0.00)	133.33 (±5.77)	107.92 (±3.69)	155.00 (±5.00)
Ours	**188.33** (±2.89)	**150.00** (±7.64)	**133.89** (±3.85)	**172.22** (±3.47)

Performance comparisons on **CHAIR**

Method	LLaVA-1.5 CHAIR$_S$ ↓	LLaVA-1.5 CHAIR$_I$ ↓	LLaVA-1.5 Recall ↑	LLaVA-1.5 Length ↑
Regular	26.2	9.4	58.5	53.4
VCD	24.4	7.9	63.3	54.2
M3ID	21.4	6.3	**64.4**	53.5
RITUAL	22.4	6.9	63.0	**54.9**
Woodpecker	24.9	7.5	60.8	49.7
HALC	21.7	7.1	63.4	53.4
Ours	**18.4**	**6.1**	62.7	54.1

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.

Acknowledgements: This work has been funded in part by the ARL award W911NF-23-20007 and W911QX-24-F-0049, DARPA award FA8750-23-2-1015, and ONR award N00014-231-2840. MM and LPM are partially supported by Meta and National Institutes of Health awards R01MH125740, R01MH132225, and R21MH130767. RS is supported in part by the ONR grant N00014-23-1-2368.

Contact: Ce Zhang, Zifu Wan, Yaqi Xie {cezhang, zifuw, yaqix}@cs.cmu.edu