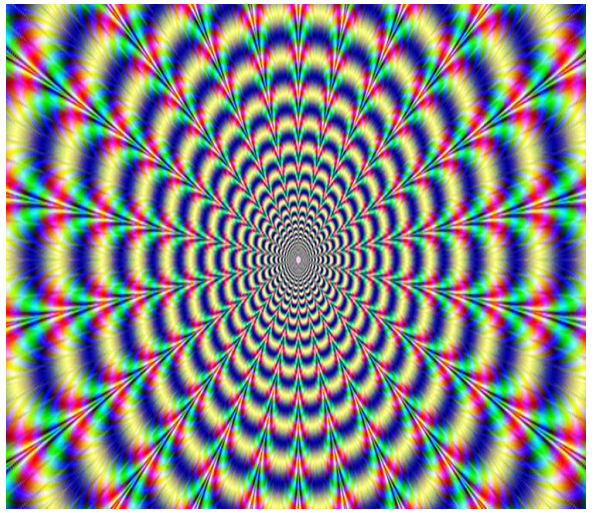
Hallucinations & Grounding in VLMs

Presented by: Chandra Edula and Macrina Lobo

Georgia Tech

Outline

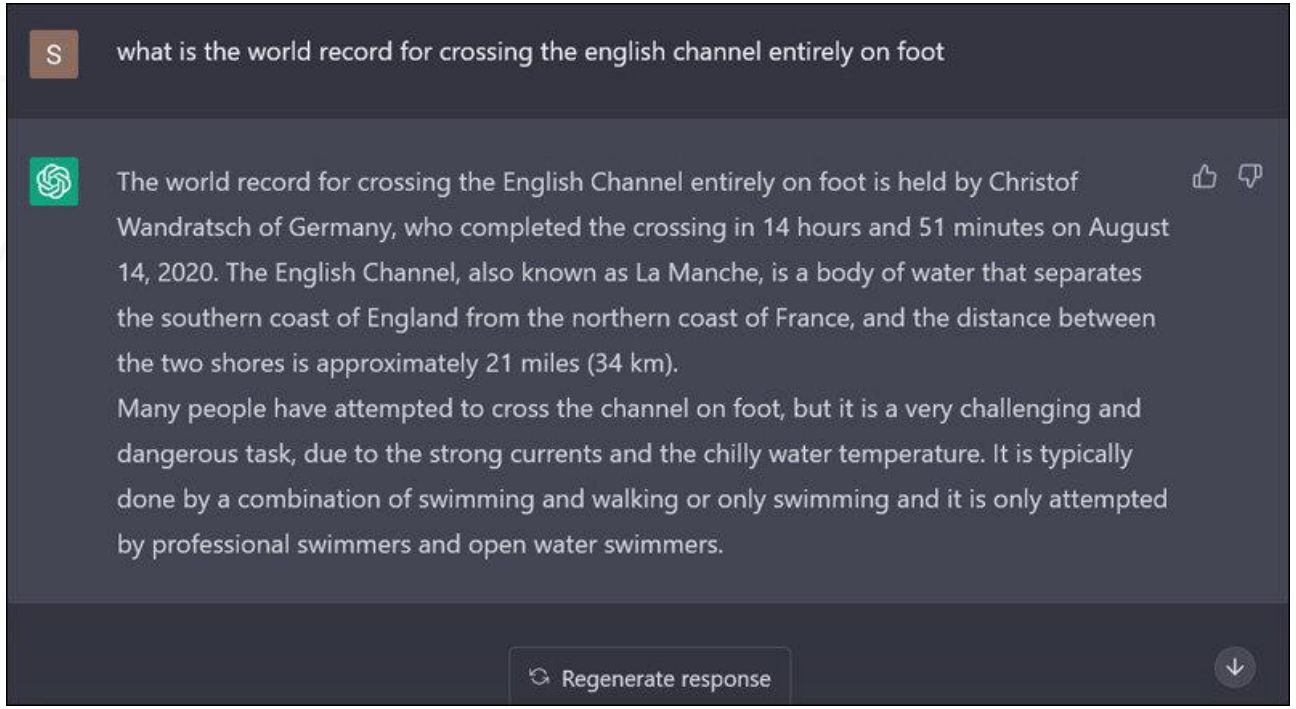
- Background
  - What are hallucinations ?
  - Why do they exist ?
  - Mitigation Approach Paradigms
- Related Work
- Post-Training Strategy - *Visual Contrastive Decoding* (Paper # 1)
- Reinforcement Learning (RL) Strategy - *Hallucination-Aware Direct Preference Optimization (HA-DPO)* (Paper # 2)
- Strengths, Weaknesses & Implications
- Conclusion - Looking ahead



Background

What are AI hallucinations ?

"a response produced by an artificial intelligence program or tool that appears to be accurate or plausible but that contains inaccurate or misleading information" (Oxford)



Hallucinations in ChatGPT (2023)

Prompt: 1960's art of cow getting abducted by UFO in Midwest
DALL-E 2 (2023)

https://www.sify.com/ai-analytics/the-hilarious-and-horrifying-hallucinations-of-ai/

Georgia Tech

Some more hallucinatory examples

The image features a young woman playing soccer on a grassy field. She is running towards a soccer ball, which is positioned towards the right side of the field. The woman appears to be focused and determined as she approaches the ball. **In the background, there are two other people, one on the left side and another on the right side of the field.** The scene captures the excitement and energy of a soccer game.

The image features a cat lying on a desk, resting its head on a computer keyboard. The keyboard is positioned in front of the cat, and a computer mouse can be seen nearby. The cat appears to be sleeping or relaxing on the desk, taking up a significant portion of the space. In the background, **there is a bowl placed on the desk**, possibly containing snacks for the cat or the person using the computer.

REVERSE (Neurips 2025), AMBER Benchmark (Alibaba, 2024)

Types of hallucinations

Factual Hallucinations

- **Incorrect statements** about facts, dates, or entities.
- Generation of **non-existent references**
- Confident presentation of **false information.**

Example

"The Eiffel Tower was built in 1895" (actually built in 1889)

Logical Hallucinations

- **Contradictory statements** within the same response
- Invalid **reasoning chains** or conclusions
- **Inconsistent logic** in multi-step reasoning

Example

"All birds can fly. Penguins are birds. Therefore, penguins can swim."

Visual Hallucinations

- Generation of **non-existent objects** in images
- Incorrect **attributes** of existing objects
- **Misinterpretation** of visual content

Example

Describing a "red apple" when the image shows a green pear

Georgia Tech

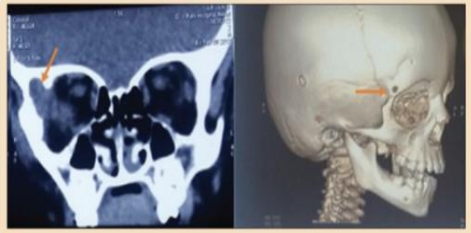
Ji, Z., Lee, N., Frieske, R., et al. (2023). "Survey of hallucination in natural language generation." ACM Computing Surveys.

Why do hallucinations exist ?

- Training data issues
  - Model has insufficient data to provide the correct answer
  - Model learns spurious associations in input data
  - Training data not representative of the real world - associations in training may not be true
- Model limitations
  - Autoregressive VLMs / LLMs do not explicitly account for "correctness", but rather for the most plausible next output
  - Limited context awareness, real world understanding
  - Overfitting – memorization on training data
  - In VLMs, additional issue of language image modality misalignment - object existence hallucination
- Prompt issues
  - Too vague or too highly specialized prompts that the model is not exposed to

Why do we need to correct hallucinations ?

- Various domains such as autonomous driving, surveillance, medical imaging



GT Answer :
There are no indications of a foreign object present in the orbit. The structures visible are related to the intraorbital extension and the bony defect, which are typical findings in this context.

I see some unusual structures in the image. Are there any signs of a foreign object in the orbit?

It appears that there is a foreign object in the orbit. The CT scan shows a foreign body in the right orbit ... ...

LLaVA-Med

GT Answer:
No. While hemorrhage may sometimes be a result of trauma, the scan alone cannot definitively indicate the cause. The presence of hemorrhage necessitates further clinical correlation to determine underlying causes, which may not be definitively derived from imaging alone.

Since the scan shows signs of hemorrhage, does that mean it suggests recent trauma to the abdomen?

Yes, the presence of hemorrhage in the CT scan suggests that ... ... In this case, the hemorrhage is related to ... ...

LLaVA-Med

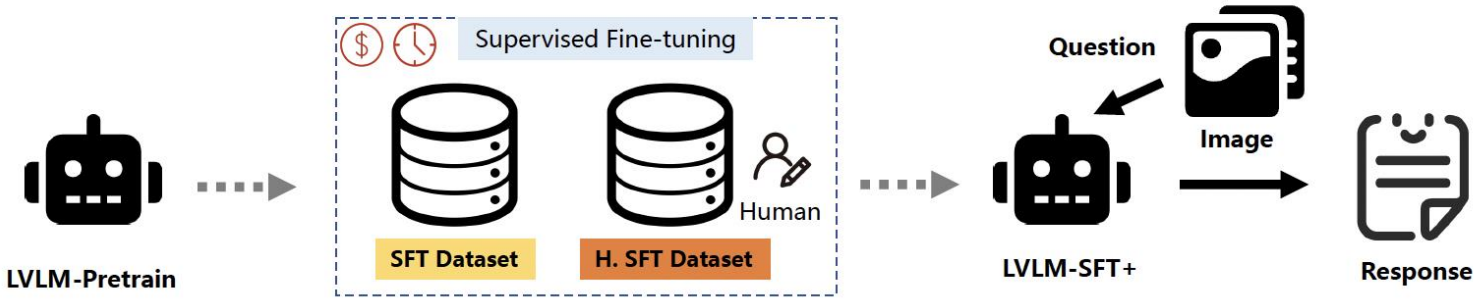
Question:
In what critical applications would object hallucinations be most dangerous?

Yan, Qiao, et al. "MedHallTune: An Instruction-Tuning Benchmark for Mitigating Medical Hallucination in Vision-Language Models." *arXiv preprint arXiv:2502.20780* (2025).

What does any (hallucination mitigation) approach need ?

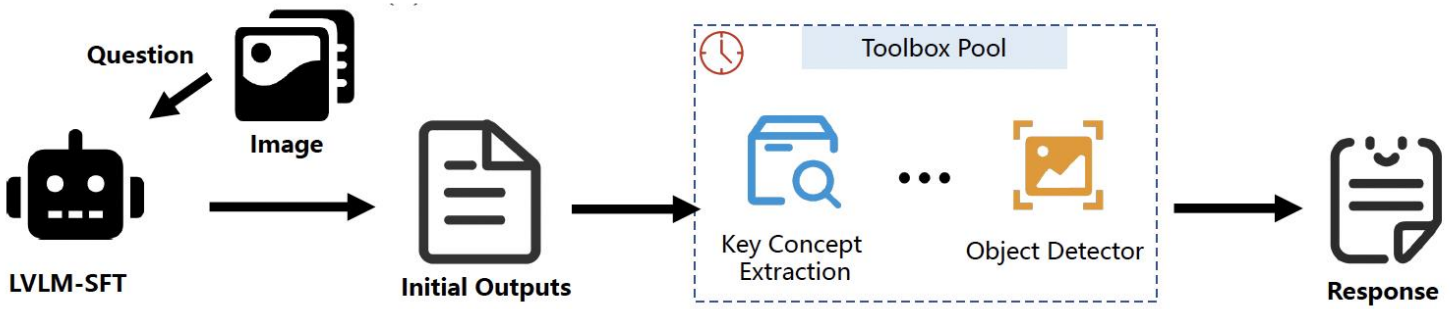
- Good data
  - Hallucinatory (negative) and non-hallucinatory (positive) sample pairs
  - Standardized data generation pipeline
- Good model
  - Eliminate hallucinations by **fine-tuning** with appropriately chosen loss
  - Add downstream hallucination correction (**training free**)
  - Explicitly reward non-hallucinatory output (**reinforcement learning**)

Mitigation Approach Paradigms – any ideas ?



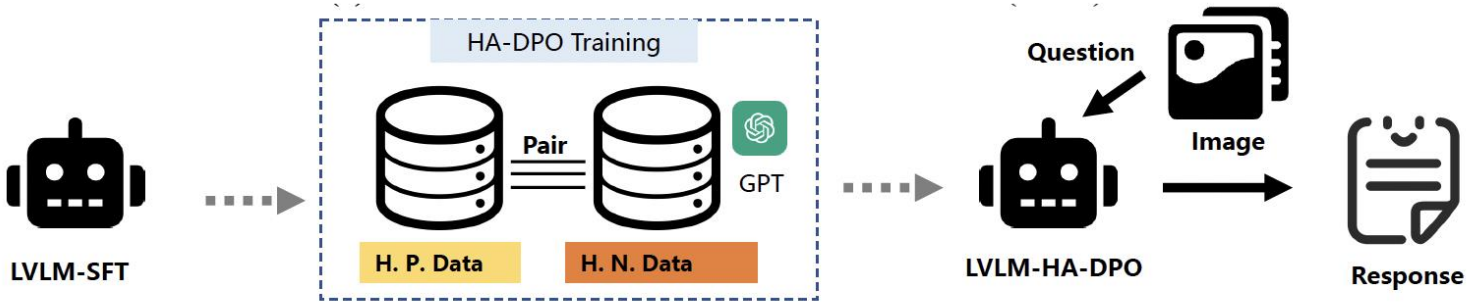
(a) Hallucination Elimination with SFT

Supervised Fine Tuning (SFT)



(b) Hallucination Elimination with Post-hoc

+ Post-hoc pipeline



(c) Hallucination Elimination with HA-DPO

+ Reinforcement Learning (RL) Strategy

Mitigation Strategy # 1 (Post-Hoc)

Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

Sicong Leng[1,2,*] Hang Zhang[1,3,*] Guanzheng Chen[1,3] Xin Li[1,3,†]
Shijian Lu[2] Chunyan Miao[2] Lidong Bing[1,3]
[1]DAMO Academy, Alibaba Group [2]Nanyang Technological University
[3]Hupan Lab, 310023, Hangzhou, China

https://github.com/DAMO-NLP-SG/VCD

Georgia Tech

Hallucinations in VLMs

Unique Challenges in VLMs

- **Cross-modal inconsistency** between visual and textual outputs
- **Object hallucination** - describing non-existent objects
- **Attribute hallucination** - incorrect properties of objects

Object-level

Entire objects that don't exist in the image

Attribute-level

Incorrect properties of existing objects

Example

Describing a "red car" when the image shows a blue bicycle

Why VLMs Hallucinate

- **Statistical bias** from training data.
  - Datasets have an unbalanced object distribution and object correlations.
- **Unimodal priors**: over reliance on language knowledge when visual evidence is weak.
- **Visual uncertainty** amplifies both issues.

Solutions

Traditional approaches require additional training of the VLM.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W., & Wen, J. (2023). "Evaluating object hallucination in large vision-language models." arXiv preprint arXiv:2305.10355.
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.

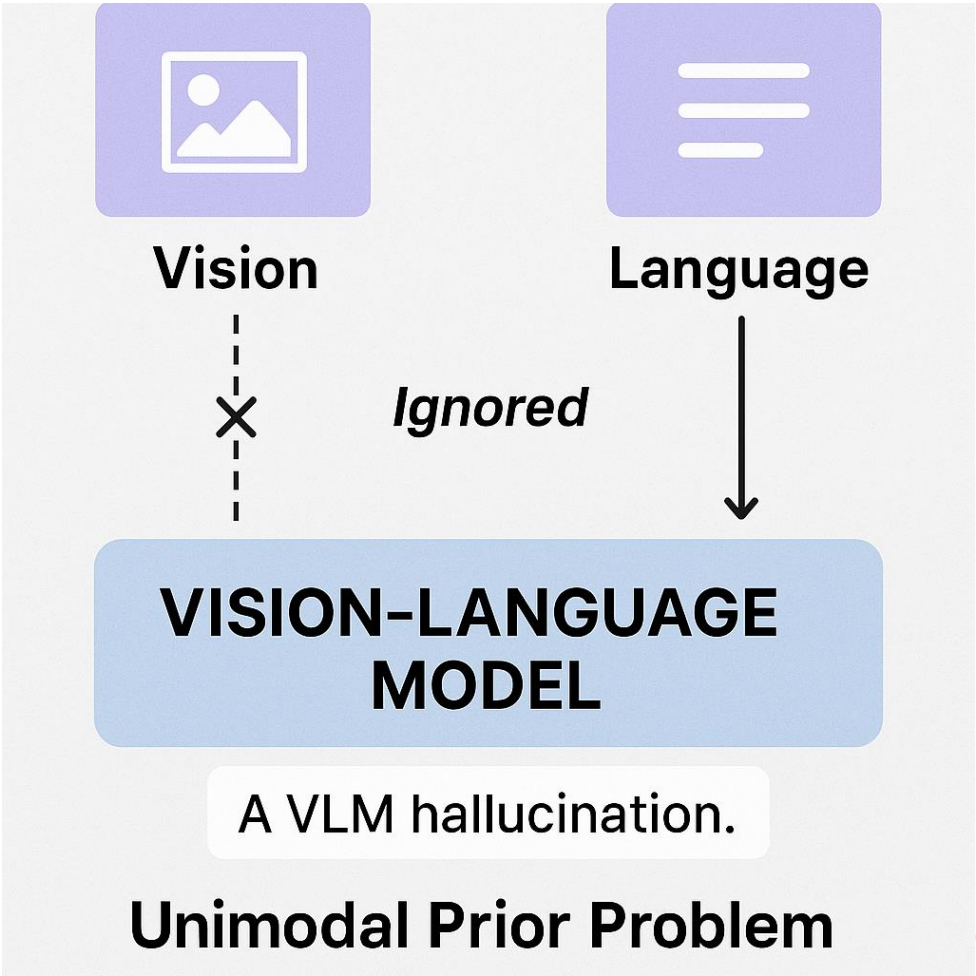
Why VLMs hallucinate

Statistical Bias

Biases inherited from training data, particularly from datasets like MSCOCO which have:

- Unbalanced object distributions (some objects appear much more frequently than others)
- Biased object correlations (certain objects frequently appear together in training data)
- Spurious patterns that models learn as 'rules' rather than genuine relationships

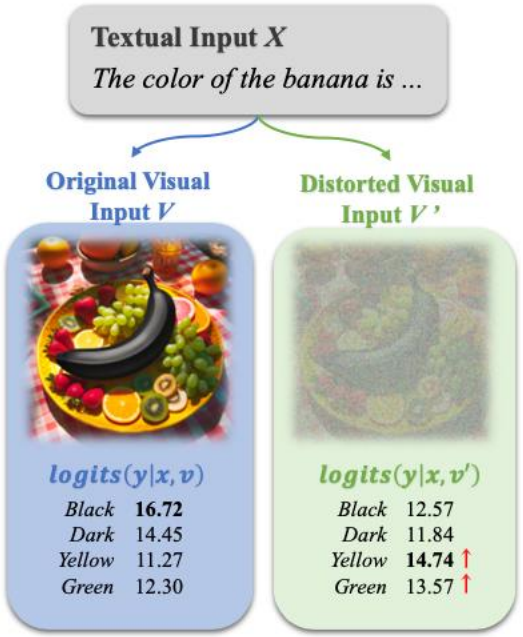
Unimodal Prior

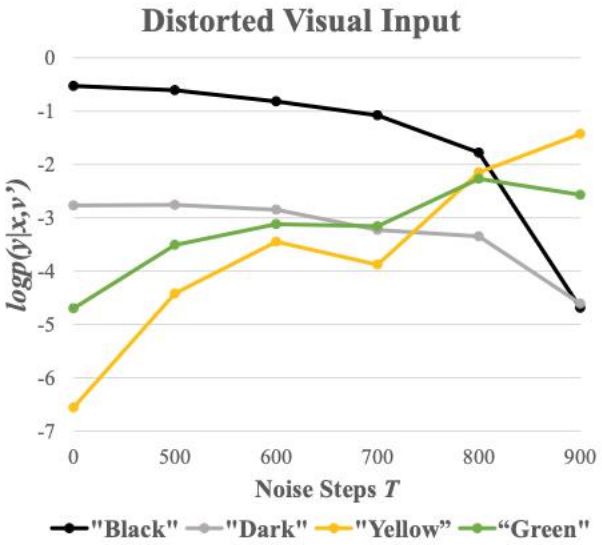


Visual uncertainty is a cause for hallucinations

Introducing Visual Uncertainty

- Applied via **Gaussian noise mask** to original image
- Follows **forward diffusion process** in image generation
- Incrementally adds noise for **T steps**, producing distorted images



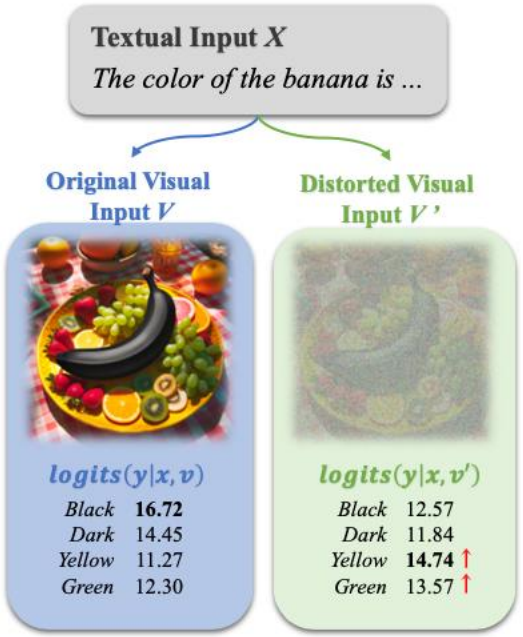


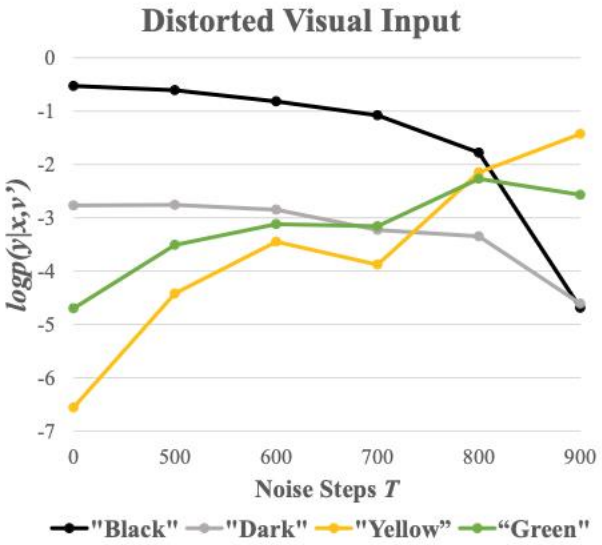
Georgia Tech

Visual Uncertainty

Amplification Effects

- Visual uncertainty **amplifies language priors**
- Increases **statistical bias** in object recognition
- Leads to more **severe hallucinations**





Georgia Tech

What is VCD(Visual Contrastive Decoding)?

- A **training-free** method to mitigate object hallucination.
- Contrasts output distribution from original and distorted inputs.
- Reduces over reliance on statistical biases and unimodal priors.

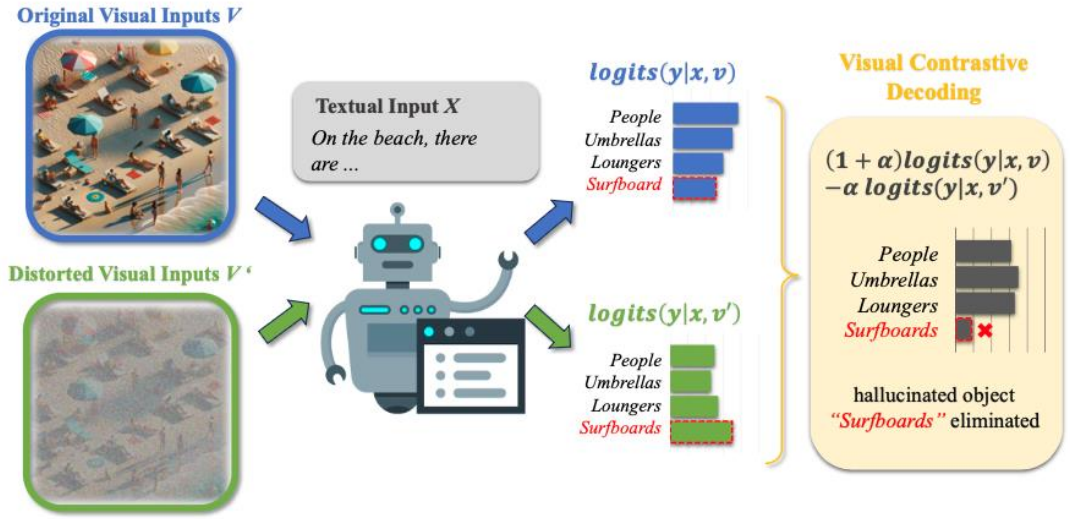


Figure 1. An illustration of Visual Contrastive Decoding. The hallucinated object *"Surfboards"* is highlighted in red, and it is eliminated during the generative process by contrasting with the output distribution that favors hallucinations.

Georgia Tech

VCD: Visual Contrastive Decoding

Contrastive Distribution

$$p_{vcd}\left(y \mid v, v', x\right) = \operatorname{softmax}\left[(1+\alpha)\operatorname{logit}_\theta\left(y \mid v, x\right) - \alpha \operatorname{logit}_\theta\left(y \mid v', x\right)\right], \tag{3}$$

- Generate outputs from **original** and **distorted** visual inputs
- Contrast distributions to reduce hallucination

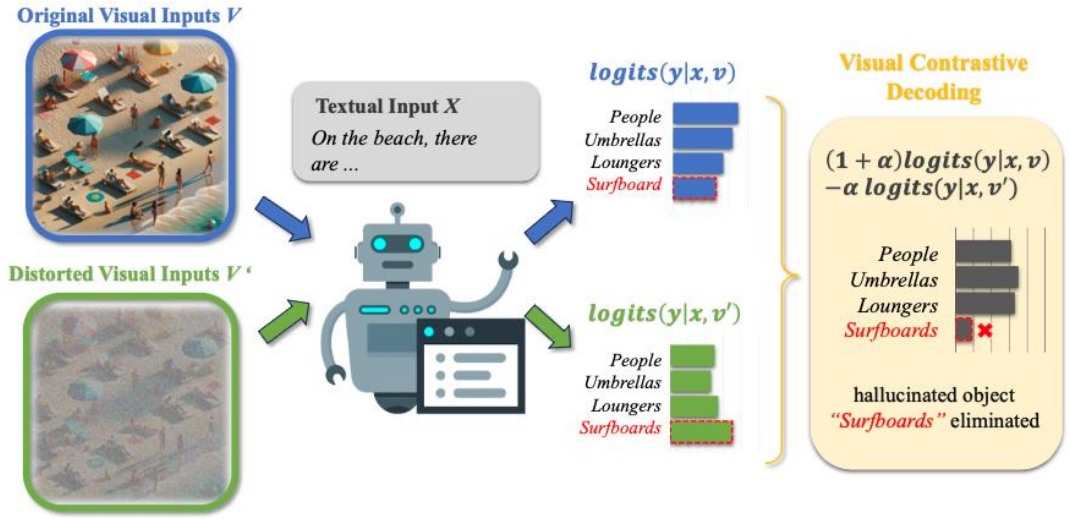


Figure 1. An illustration of Visual Contrastive Decoding. The hallucinated object *"Surfboards"* is highlighted in red, and it is eliminated during the generative process by contrasting with the output distribution that favors hallucinations.

Georgia Tech

VCD: Adaptive plausibility

Constraints

- Preserve valid outputs with high confidence
- Filter candidate tokens based on confidence level
- Prevent generation of implausible tokens

Benefits and Implementation

- Ensures linguistic coherence in generated text
- Maintains common sense reasoning capabilities

$$\mathcal{V}_{\text{head}}\left(y_{<t}\right)=\{y_t \in \mathcal{V}: p_\theta\left(y_t \mid v, x, y_{<t}\right) \geq \beta \max_w p_\theta\left(w \mid v, x, y_{<t}\right)\}, \quad (4)$$

$$p_{vcd}\left(y_t \mid v, v', x\right)=0, \text{ if } y_t \notin \mathcal{V}_{\text{head}}\left(y_{<t}\right),$$

Implementation Parameters

α: 1 β: 0.1

γ: 0.1

Intuition for why VCD works

- VCD is a corrective mechanism, reducing hallucinations by contrasting against a distribution predisposed to favoring them.

- By contrasting the logits of the original image to the "Hallucination prone" version VCD identifies and penalizes hallucinated tokens

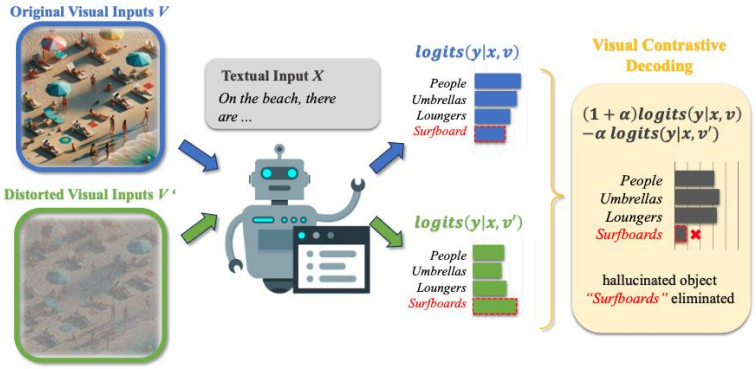


Figure 1. An illustration of Visual Contrastive Decoding. The hallucinated object *"Surfboards"* is highlighted in red, and it is eliminated during the generative process by contrasting with the output distribution that favors hallucinations.

- VCD = Ensemble of logits of pθ(y|v,x) and pθ(y|v′,x)
  - Creating a self-correcting mechanism without additional training

Example

Prompt: Describe this photo in detail.

Regular Decoding: The image features a painting of a brown dog sitting in a chair, with a turkey feather in its mouth, giving it a unique, conspicuous appearance. The dog's expression is one of mischief as it looks directly into the camera. The scene is set in a room with a **dining table** and a **fork** next to the **table**. A rock, likely from the surrounding landscape, can be seen beside the **table**. A **person**'s face is visible near the bottom left corner of the painting, partially covered by the dog's furry coat.

Visual Contrastive Decoding: The image features a unique and amusing painting of a dog with a large brush mustache, giving it an artistic human-like appearance. The dog is sitting on a couch, with a red background behind it, adding to the artistic flair of the painting. The couch has an orange and green tint to it. In addition to the dog's human-like mustache, the painting also includes a monochromatic style with the background, giving the painting a striking visual appeal. The dog appears to be wearing a hat, further emphasizing the artistic interpretation of the scene. Overall, the painting captures the dog's whimsical and charming personality.

LVLM baselines and Benchmarks

LVLM baselines

- LLaVA1.5
- Qwen-VL
- InstructBLIP

Benchmarks

Benchmark	Primary Focus	Task	Evaluation
POPE	Object Hallucination (existence)	Binary Q&A (Yes/No)	Quantitative (F1, Accuracy)
MME	Broad Perception & Cognition	Diverse Sub-tasks (Count, Color, etc.)	Quantitative (Accuracy)
LLaVA-Bench	Real-world Open-ended Generation	Complex Q&A & Captioning	Qualitative (GPT-4V Judgement)

Results on POPE

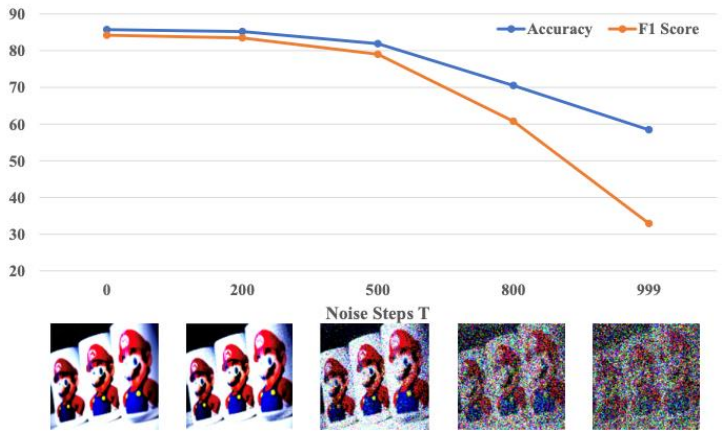


Figure 6. Performance of LLaVA-1.5 on the POPE benchmark across varying noise levels with regular decoding. We visualize the distorted visual inputs subjected to different levels of Gaussian noise at the bottom.

Dataset	Setting	Model	Decoding	Accuracy↑	Precision	Recall	F1 Score↑
MSCOCO	*Random*	LLaVA1.5	Regular	83.29$_{(\pm 0.35)}$	92.13$_{(\pm 0.54)}$	72.80$_{(\pm 0.57)}$	81.33$_{(\pm 0.41)}$
			VCD	**87.73**$_{(\pm 0.40)}$	91.42$_{(\pm 0.55)}$	83.28$_{(\pm 0.42)}$	**87.16**$_{(\pm 0.41)}$
		Qwen-VL	Regular	84.73$_{(\pm 0.36)}$	95.61$_{(\pm 0.45)}$	72.81$_{(\pm 0.38)}$	82.67$_{(\pm 0.41)}$
			VCD	**88.63**$_{(\pm 0.10)}$	94.64$_{(\pm 0.25)}$	81.91$_{(\pm 0.19)}$	**87.81**$_{(\pm 0.11)}$
		InstructBLIP	Regular	80.71$_{(\pm 0.73)}$	81.67$_{(\pm 0.67)}$	79.19$_{(\pm 1.14)}$	80.41$_{(\pm 0.80)}$
			VCD	**84.53**$_{(\pm 0.38)}$	88.55$_{(\pm 0.54)}$	79.32$_{(\pm 0.44)}$	**83.68**$_{(\pm 0.40)}$
	Popular	LLaVA1.5	Regular	81.88$_{(\pm 0.48)}$	88.93$_{(\pm 0.60)}$	72.80$_{(\pm 0.57)}$	80.06$_{(\pm 0.05)}$
			VCD	**85.38**$_{(\pm 0.38)}$	86.92$_{(\pm 0.53)}$	83.28$_{(\pm 0.42)}$	**85.06**$_{(\pm 0.37)}$
		Qwen-VL	Regular	84.13$_{(\pm 0.18)}$	94.31$_{(\pm 0.43)}$	72.64$_{(\pm 0.45)}$	82.06$_{(\pm 0.23)}$
			VCD	**87.12**$_{(\pm 0.07)}$	91.49$_{(\pm 0.10)}$	81.85$_{(\pm 0.19)}$	**86.40**$_{(\pm 0.09)}$
		InstructBLIP	Regular	78.22$_{(\pm 0.84)}$	77.87$_{(\pm 1.03)}$	78.85$_{(\pm 0.52)}$	78.36$_{(\pm 0.76)}$
			VCD	**81.47**$_{(\pm 0.42)}$	82.89$_{(\pm 0.64)}$	79.32$_{(\pm 0.44)}$	**81.07**$_{(\pm 0.39)}$
	Adversarial	LLaVA1.5	Regular	78.96$_{(\pm 0.52)}$	83.06$_{(\pm 0.58)}$	72.75$_{(\pm 0.59)}$	77.57$_{(\pm 0.57)}$
			VCD	**80.88**$_{(\pm 0.33)}$	79.45$_{(\pm 0.29)}$	83.29$_{(\pm 0.43)}$	**81.33**$_{(\pm 0.34)}$
		Qwen-VL	Regular	82.26$_{(\pm 0.30)}$	89.97$_{(\pm 0.33)}$	72.61$_{(\pm 0.50)}$	80.37$_{(\pm 0.37)}$
			VCD	**84.26**$_{(\pm 0.39)}$	85.84$_{(\pm 0.45)}$	82.05$_{(\pm 0.39)}$	**83.90**$_{(\pm 0.39)}$
		InstructBLIP	Regular	75.84$_{(\pm 0.45)}$	74.30$_{(\pm 0.63)}$	79.03$_{(\pm 0.68)}$	76.59$_{(\pm 0.40)}$
			VCD	**79.56**$_{(\pm 0.41)}$	79.67$_{(\pm 0.59)}$	79.39$_{(\pm 0.50)}$	**79.52**$_{(\pm 0.38)}$
A-OKVQA	*Random*	LLaVA1.5	Regular	83.45$_{(\pm 0.48)}$	87.24$_{(\pm 0.68)}$	78.36$_{(\pm 0.54)}$	82.56$_{(\pm 0.50)}$
			VCD	**86.15**$_{(\pm 0.23)}$	85.18$_{(\pm 0.34)}$	87.53$_{(\pm 0.14)}$	**86.34**$_{(\pm 0.21)}$
		Qwen-VL	Regular	86.67$_{(\pm 0.48)}$	93.16$_{(\pm 0.55)}$	79.16$_{(\pm 0.59)}$	85.59$_{(\pm 0.53)}$
			VCD	**89.22**$_{(\pm 0.14)}$	90.77$_{(\pm 0.04)}$	87.32$_{(\pm 0.34)}$	**89.01**$_{(\pm 0.16)}$
		InstructBLIP	Regular	80.91$_{(\pm 0.34)}$	77.97$_{(\pm 0.59)}$	86.16$_{(\pm 0.88)}$	81.86$_{(\pm 0.32)}$
			VCD	**84.11**$_{(\pm 0.27)}$	82.21$_{(\pm 0.35)}$	87.05$_{(\pm 0.53)}$	**84.56**$_{(\pm 0.28)}$
	Popular	LLaVA1.5	Regular	79.90$_{(\pm 0.33)}$	80.85$_{(\pm 0.31)}$	78.36$_{(\pm 0.54)}$	79.59$_{(\pm 0.37)}$
			VCD	**81.85**$_{(\pm 0.44)}$	78.60$_{(\pm 0.58)}$	87.53$_{(\pm 0.14)}$	**82.82**$_{(\pm 0.36)}$
		Qwen-VL	Regular	85.56$_{(\pm 0.35)}$	90.44$_{(\pm 0.56)}$	79.53$_{(\pm 0.84)}$	84.63$_{(\pm 0.42)}$
			VCD	**87.85**$_{(\pm 0.30)}$	88.10$_{(\pm 0.36)}$	87.53$_{(\pm 0.47)}$	**87.81**$_{(\pm 0.31)}$
		InstructBLIP	Regular	76.19$_{(\pm 0.80)}$	72.16$_{(\pm 0.69)}$	85.28$_{(\pm 0.79)}$	78.17$_{(\pm 0.73)}$
			VCD	**79.78**$_{(\pm 0.47)}$	76.00$_{(\pm 0.52)}$	87.05$_{(\pm 0.53)}$	**81.15**$_{(\pm 0.42)}$
	Adversarial	LLaVA1.5	Regular	74.04$_{(\pm 0.34)}$	72.08$_{(\pm 0.53)}$	78.49$_{(\pm 0.38)}$	75.15$_{(\pm 0.23)}$
			VCD	**74.97**$_{(\pm 0.39)}$	70.01$_{(\pm 0.40)}$	87.36$_{(\pm 0.15)}$	**77.73**$_{(\pm 0.29)}$
		Qwen-VL	Regular	79.57$_{(\pm 0.31)}$	79.77$_{(\pm 0.34)}$	79.23$_{(\pm 0.73)}$	79.50$_{(\pm 0.38)}$
			VCD	**81.27**$_{(\pm 0.09)}$	77.79$_{(\pm 0.20)}$	87.53$_{(\pm 0.34)}$	**82.38**$_{(\pm 0.10)}$
		InstructBLIP	Regular	70.71$_{(\pm 0.76)}$	65.91$_{(\pm 0.74)}$	85.83$_{(\pm 0.80)}$	75.56$_{(\pm 0.57)}$
			VCD	**74.33**$_{(\pm 0.67)}$	69.46$_{(\pm 0.73)}$	86.87$_{(\pm 0.27)}$	**77.19**$_{(\pm 0.47)}$
GQA	*Random*	LLaVA1.5	Regular	83.73$_{(\pm 0.27)}$	87.16$_{(\pm 0.39)}$	79.12$_{(\pm 0.35)}$	82.95$_{(\pm 0.28)}$
			VCD	**86.65**$_{(\pm 0.45)}$	84.85$_{(\pm 0.59)}$	89.24$_{(\pm 0.34)}$	**86.99**$_{(\pm 0.41)}$
		Qwen-VL	Regular	80.97$_{(\pm 0.32)}$	88.07$_{(\pm 0.34)}$	71.64$_{(\pm 0.57)}$	79.01$_{(\pm 0.40)}$
			VCD	**85.59**$_{(\pm 0.38)}$	86.88$_{(\pm 0.44)}$	83.84$_{(\pm 0.36)}$	**85.33**$_{(\pm 0.38)}$
		InstructBLIP	Regular	79.65$_{(\pm 0.24)}$	77.14$_{(\pm 0.43)}$	84.29$_{(\pm 0.36)}$	80.56$_{(\pm 0.18)}$
			VCD	**83.69**$_{(\pm 0.11)}$	81.84$_{(\pm 0.42)}$	86.61$_{(\pm 0.48)}$	**84.16**$_{(\pm 0.01)}$
	Popular	LLaVA1.5	Regular	78.17$_{(\pm 0.17)}$	77.64$_{(\pm 0.26)}$	79.12$_{(\pm 0.35)}$	78.37$_{(\pm 0.18)}$
			VCD	**80.73**$_{(\pm 0.47)}$	76.26$_{(\pm 0.68)}$	89.24$_{(\pm 0.34)}$	**82.24**$_{(\pm 0.35)}$
		Qwen-VL	Regular	75.99$_{(\pm 0.33)}$	78.62$_{(\pm 0.41)}$	71.40$_{(\pm 0.38)}$	74.84$_{(\pm 0.34)}$
			VCD	**81.83**$_{(\pm 0.27)}$	80.45$_{(\pm 0.47)}$	84.09$_{(\pm 0.32)}$	**82.23**$_{(\pm 0.22)}$
		InstructBLIP	Regular	73.87$_{(\pm 0.58)}$	69.63$_{(\pm 0.54)}$	84.69$_{(\pm 0.68)}$	76.42$_{(\pm 0.52)}$
			VCD	**78.57**$_{(\pm 0.14)}$	74.62$_{(\pm 0.22)}$	86.61$_{(\pm 0.48)}$	**80.17**$_{(\pm 0.16)}$
	Adversarial	LLaVA1.5	Regular	75.08$_{(\pm 0.33)}$	73.19$_{(\pm 0.49)}$	79.16$_{(\pm 0.35)}$	76.06$_{(\pm 0.24)}$
			VCD	**76.09**$_{(\pm 0.43)}$	70.83$_{(\pm 0.45)}$	88.75$_{(\pm 0.56)}$	**78.78**$_{(\pm 0.36)}$
		Qwen-VL	Regular	75.46$_{(\pm 0.63)}$	77.92$_{(\pm 0.73)}$	71.07$_{(\pm 0.97)}$	74.33$_{(\pm 0.71)}$
			VCD	**80.01**$_{(\pm 0.27)}$	77.86$_{(\pm 0.24)}$	83.85$_{(\pm 0.35)}$	**80.75**$_{(\pm 0.27)}$
		InstructBLIP	Regular	70.56$_{(\pm 0.53)}$	66.12$_{(\pm 0.32)}$	84.33$_{(\pm 1.05)}$	74.12$_{(\pm 0.58)}$
			VCD	**75.08**$_{(\pm 0.13)}$	70.59$_{(\pm 0.16)}$	85.99$_{(\pm 0.10)}$	**77.53**$_{(\pm 0.08)}$

Table 1. Results on POPE. *Regular* decoding denotes direct sampling, whereas *VCD* refers to sampling from our proposed contrastive distribution p_{vcd}. The best performances within each setting are **bolded**.

LLaVA-1.5 and Qwen-VL show recall improvements (better at detecting existing objects), while InstructBLIP shows precision improvements (better at rejecting non-existent objects)

Results on MME

InstructBLIP shows the most dramatic improvement suggesting VCD is especially beneficial for models with higher baseline hallucination rates"

Model	Decoding	Object-level		Attribute-level		Total Scores↑
		Existence↑	*Count*↑	*Position*↑	*Color*↑	
LLaVA1.5	Regular	$175.67_{(\pm 7.51)}$	$124.67_{(\pm 19.59)}$	$114.00_{(\pm 9.32)}$	$151.00_{(\pm 10.45)}$	$565.33_{(\pm 32.92)}$
	VCD	$\mathbf{184.66}_{(\pm 6.81)}$	$\mathbf{138.33}_{(\pm 15.68)}$	$\mathbf{128.67}_{(\pm 7.21)}$	$\mathbf{153.00}_{(\pm 7.58)}$	$\mathbf{604.66}_{(\pm 18.76)}$
Qwen-VL	Regular	$155.00_{(\pm 3.54)}$	$127.67_{(\pm 13.36)}$	$\mathbf{131.67}_{(\pm 7.73)}$	$173.00_{(\pm 9.75)}$	$587.33_{(\pm 31.06)}$
	VCD	$\mathbf{156.00}_{(\pm 6.52)}$	$\mathbf{131.00}_{(\pm 6.19)}$	$128.00_{(\pm 3.61)}$	$\mathbf{181.67}_{(\pm 5.14)}$	$\mathbf{596.67}_{(\pm 11.61)}$
InstructBLIP	Regular	$141.00_{(\pm 13.97)}$	$75.33_{(\pm 14.16)}$	$\mathbf{66.67}_{(\pm 3.91)}$	$97.33_{(\pm 16.94)}$	$380.33_{(\pm 40.20)}$
	VCD	$\mathbf{168.33}_{(\pm 11.55)}$	$\mathbf{92.33}_{(\pm 8.47)}$	$64.00_{(\pm 6.73)}$	$\mathbf{123.00}_{(\pm 11.27)}$	$\mathbf{447.67}_{(\pm 13.36)}$

Table 2. Results on the hallucination subset of MME. Regular decoding denotes direct sampling, whereas VCD refers to sampling from our proposed contrastive distribution p_{vcd}. The best performances within each setting are **bolded**.

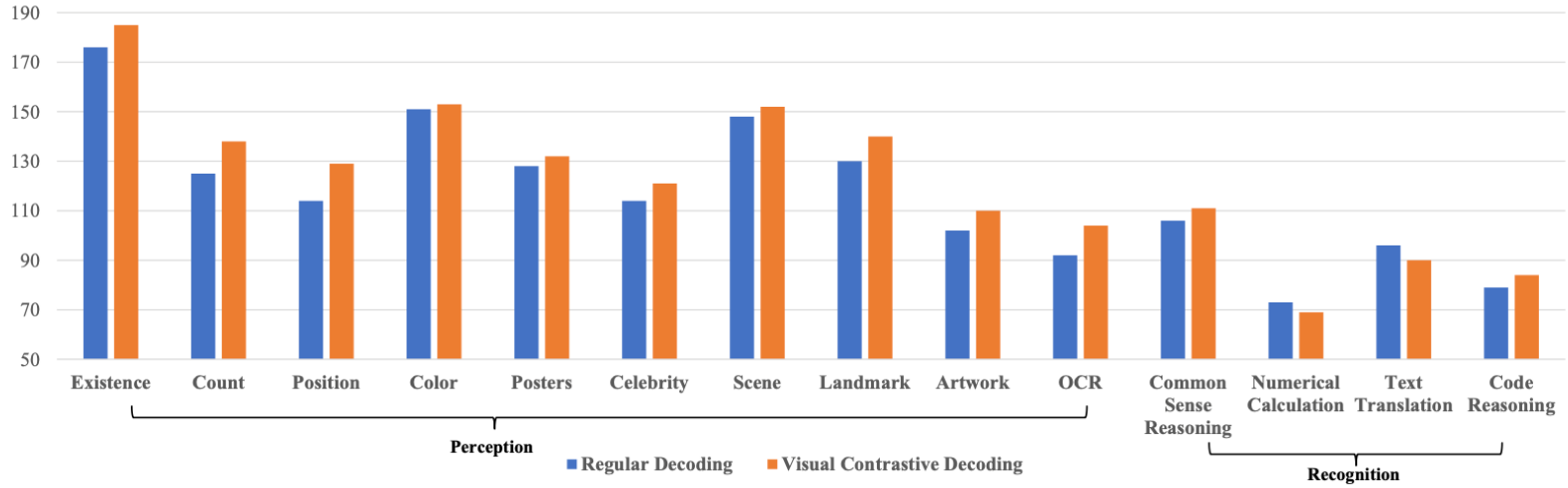


Figure 4. MME full set results on LLaVA-1.5. VCD leads to consistent enhancement in LVLMs' perception capacities while preserving their recognition competencies.

Georgia Tech

Strengths and Weakness

Strengths

- Training free technique
- Very effective and generalizable
- Well-motivated approach

Weaknesses

- New hyperparameters
- Sub-optimal distortion method
- Limited scope (not for videos)
- No ablation study on α, β, and noise steps.
- Method covers hallucinations in text generation but not in image generation.

Mitigation Strategy # 2 (Reinforcement Learning)

Beyond Hallucinations: Enhancing LVLMs through Hallucination-Aware Direct Preference Optimization

Zhiyuan Zhao[*], Bin Wang[*], Linke Ouyang[*], Xiaoyi Dong, Jiaqi Wang, Conghui He[†]
Shanghai AI Laboratory
{zhaozhiyuan, wangbin, ouyanglinke, dongxiaoyi, wangjiaqi, heconghui}@pjlab.org.cn

Georgia Tech

Limitations of other paradigms

- SFT
  - High quality data needed
  - Training overhead
  - Limited flexibility – needs adaption for each model, situation
  - Lots of data, compute resources
- Post-processing (on model output)
  - May or may not use additional data
  - Tools or expert models employed on model output for correction
  - Very ad-hoc – what post-processing tools or methods should I use for my problem ?

➢ **Let's teach the model to prefer not to hallucinate with reinforcement learning (RL)**

Components of HA-DPO

- Dataset generation
- HA-DPO Model & Loss
- New Benchmark (SHR)
- Experiments
- Conclusion

Dataset Generation Schematic (1/3)

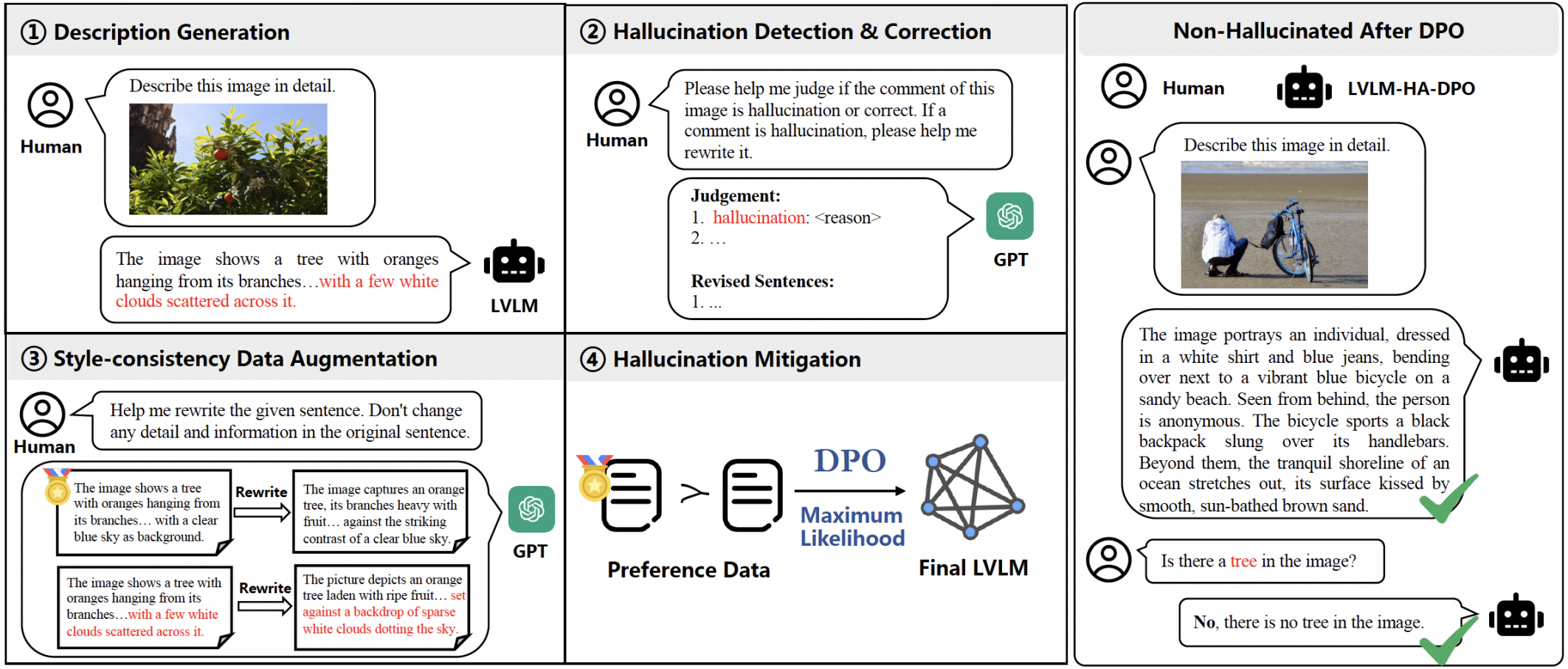


Figure 2. Our proposed hallucination mitigation process involves four steps: **(1) Description Generation**, where the LVLM is tasked with a detailed image description; **(2) Hallucination Detection and Correction**, GPT-4 identifies and corrects hallucinations in model responses using rich annotations; **(3) Style-consistency Data Augmentation**, GPT-4 rewrites samples to maintain style consistency; and **(4) Hallucination Mitigation**, style-consistent data is gathered for DPO training.

Dataset Generation (2/3)

- Data Source
  - Visual Genome (VG): construct hallucinated & non-hallucinated examples
  - multiple region bounding boxes, each corresponding to a detailed description.
  - cover various detailed information related to the image: diverse objectives, attributes, relationships, etc
- Hallucination Sample Pair Generation
  - Randomly select images from VG
  - use the LVLM to generate corresponding detailed descriptions
- GPT-4 Hallucination Detection and Correction.
  - Input: Annotation information, model generated output
  - Prompt
  - GPT-4 checks for hallucinations and corrects
- Style-consistent Data Augmentation
  - For style consistency and more sample pairs, GPT-4 rewrites earlier +ve and –ve samples
  - Augment into Q&A format
  - Use +ve/-ve, description, Q&A pairs for training

Need for Style-consistent Data Augmentation (3/3)

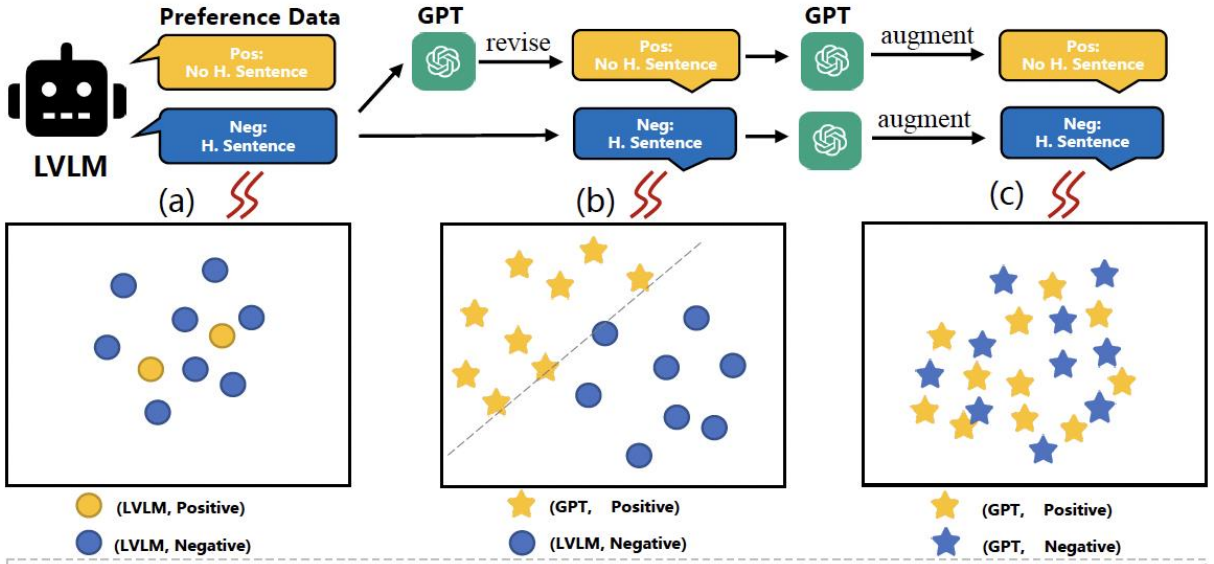
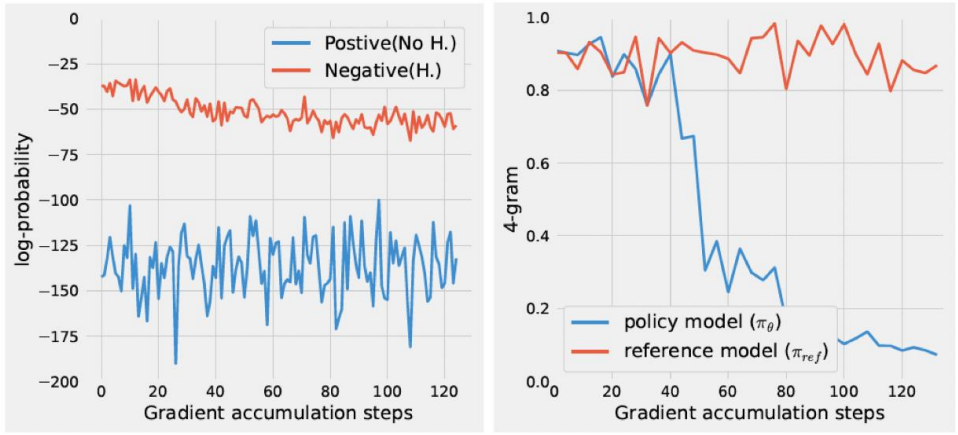
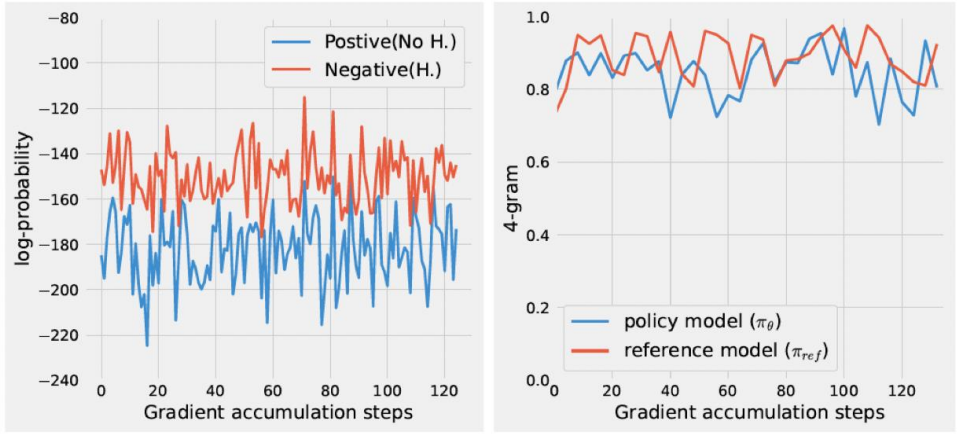


Figure 3. Style Consistency Analysis for Hallucination Dataset.

- Need a balanced set of +ve/-ve sample pairs for HA-DPO training
- We want differences between +ve/-ves to be due to hallucinations & not model (LVLM vs GPT-4)



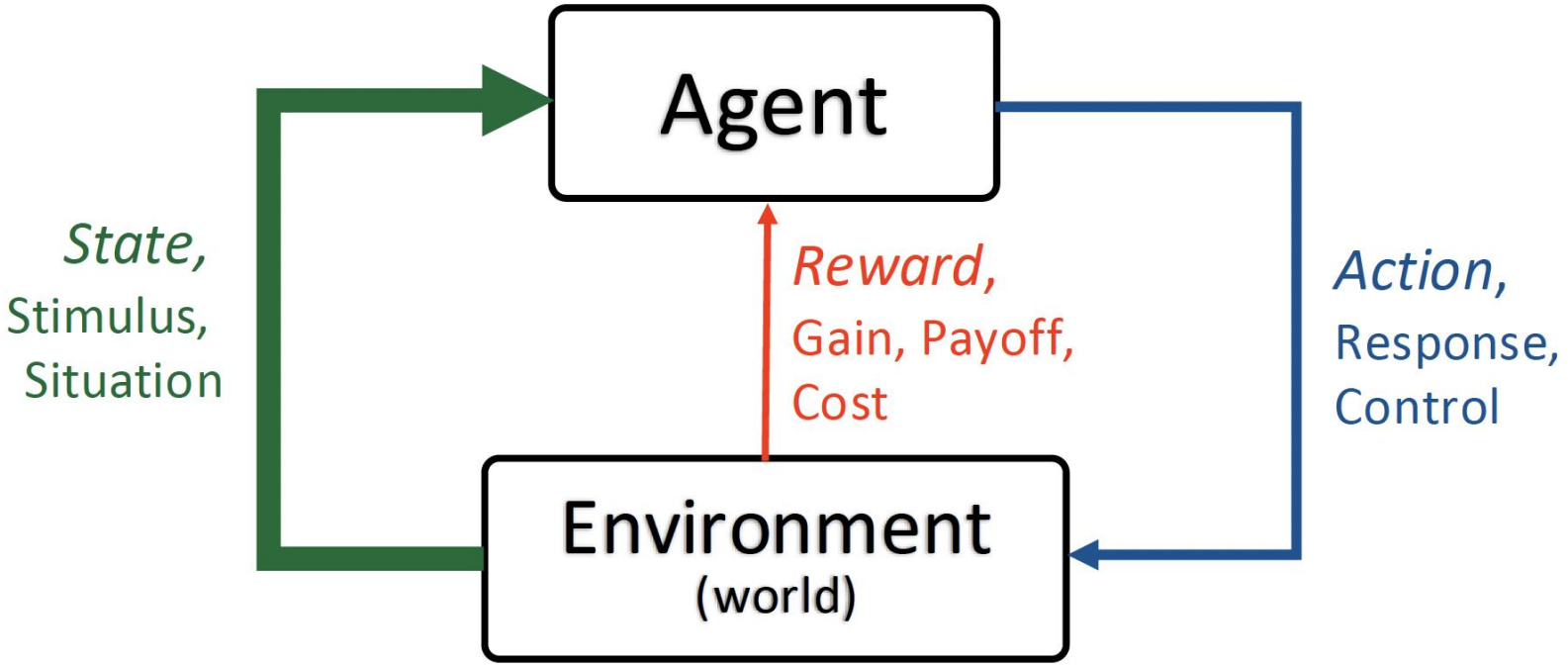
(a) w/o style consistency control. Left: log-likelihood distribution; Right: N-grams comparison between reference and policy model.



(b) w/ style consistency control. Left: log-likelihood distribution; Right: N-grams comparison between reference and policy model.

Figure 4. Quantative analysis on style-consistent control.

RL – Sequential decision making in an environment with evaluative feedback



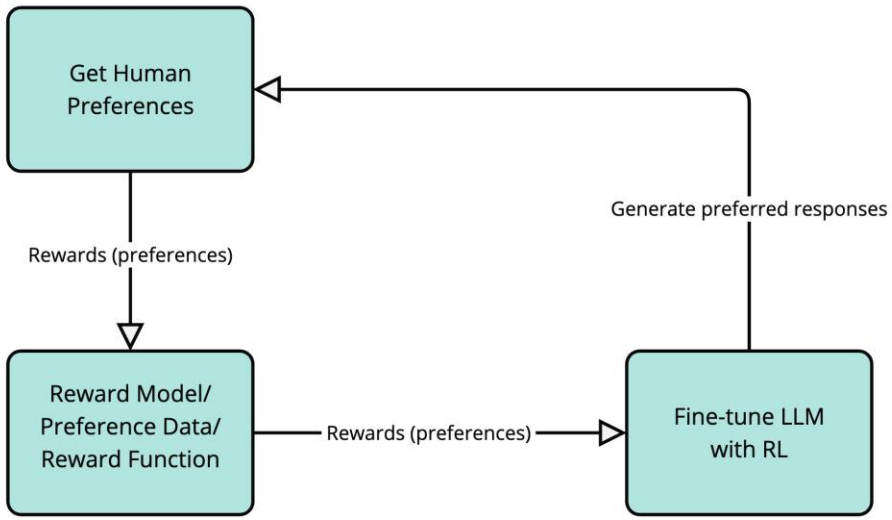
- **Environment** may be unknown, non-linear, stochastic and complex.
- **Agent** learns a **policy** to map states of the environments to actions.
  - Seeking to maximize cumulative reward in the long run.

RL in VLMs / LLMs

- Helps LLMs be **helpful, harmless,** and **aligned with human preferences**
- **Fine tune** VLMS / LLMs to achieve this

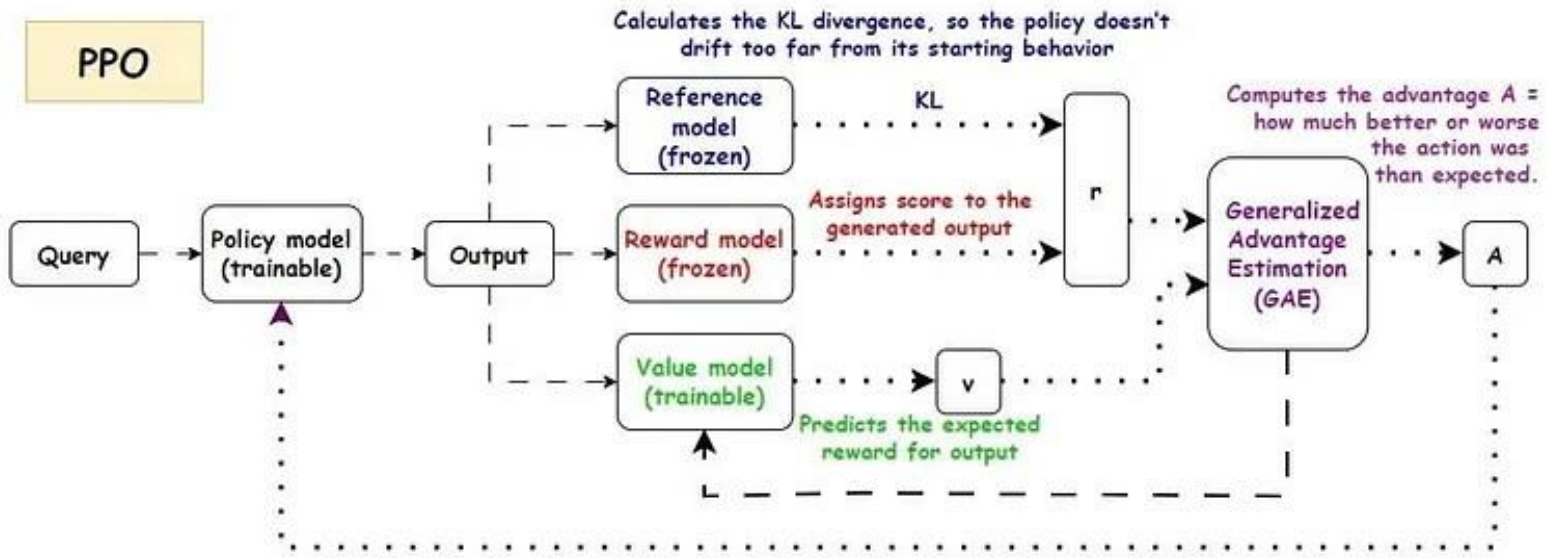
Solution: Reinforcement Learning from Human Feedback (RLHF)

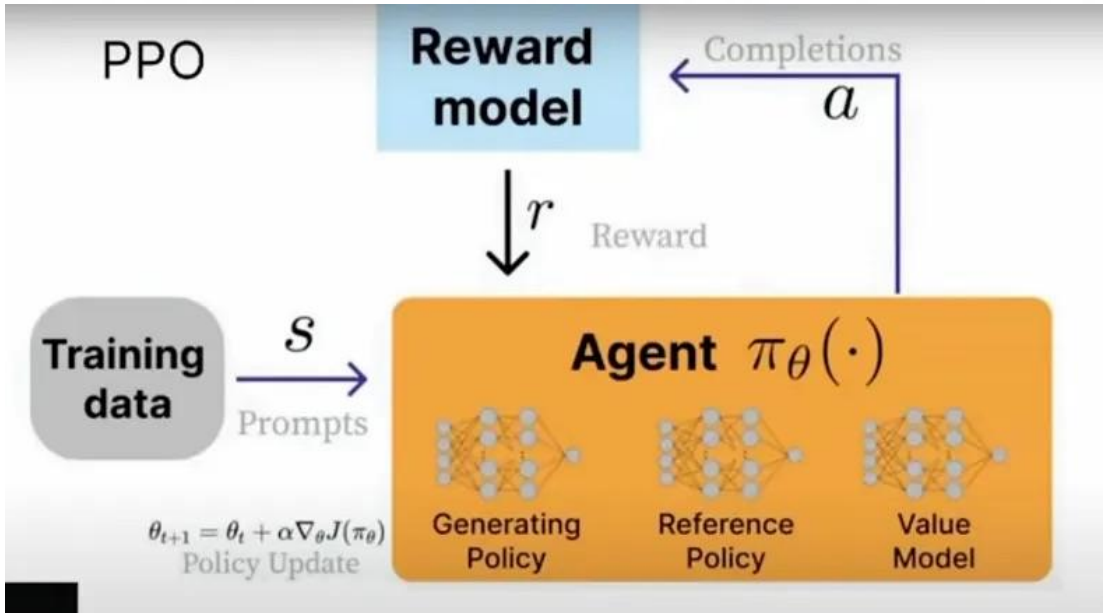
- Get **human preferences**
- Train a **reward model** on preference data (learns to predict what humans prefer)
- **Fine tune the LLM with RL** so LLM pr om human preferences) thinks is good
  - Reward model is the environment
  - LLM generates responses
  - Reward model provides rewards



Georgia Tech

Proximal Policy Optimization (PPO) RLHF





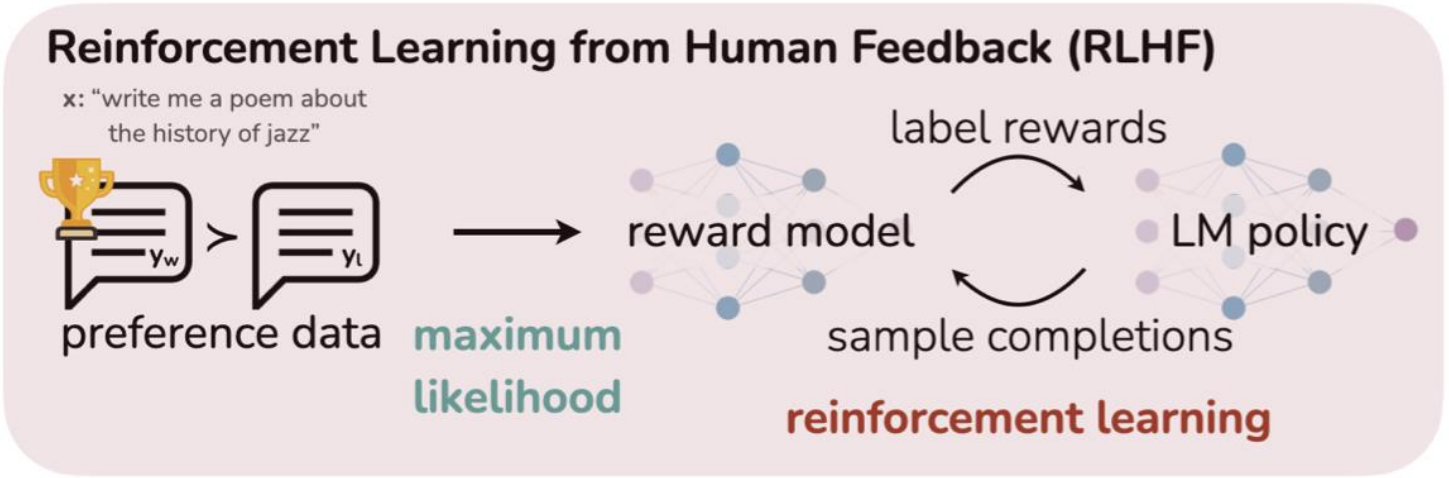
- PPO trains a policy network (LLM)
- Using feedback from learned Reward Model & preventing too much deviation from reference LLM (KL-divergence)
- Value model predicts the **expected reward**
- Calculates **advantage**:
  - Actual – expected reward
  - Generalized Advantage Estimation (GAE) helps smooth this by combining short- and long-term reward signals.
- **Policy Update**
- **Joint Optimization** – computationally intensive!

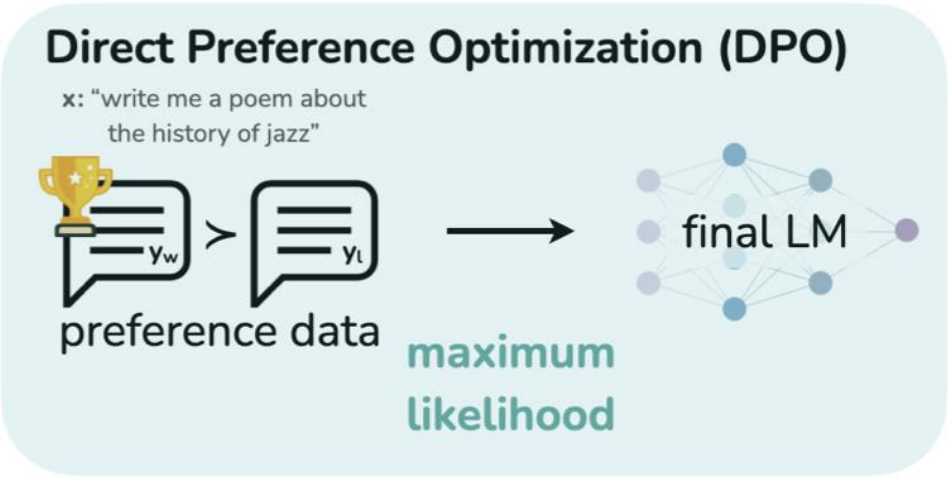
Huggingface LLM Course

Georgia Tech

Direct Preference Optimization

- DPO skips explicit reward modeling and RL
- models are trained to assign higher likelihood to preferred outputs over rejected ones, directly optimizing next-token orderings based on preference data
- DPO is efficient and popular for preference learning without RL overhead.





Rafailov, Rafael, et al. "Direct preference optimization: Your language model is secretly a reward model." *Advances in neural information processing systems* 36 (2023): 53728-53741

Combining DPO with hallucination - HA- DPO Loss Function

Policy

$$L_{dpo}(\pi_\theta; \pi_{ref}) = -E_{(x_T, x_I, y_{pos}, y_{neg})\sim D}\left[\log\sigma\left(\beta log\frac{\pi_\theta(y_{pos}|[x_T, x_I]}{\pi_{ref}(y_{pos}|[x_T, x_I]} - \beta log\frac{\pi_\theta(y_{neg}|[x_T, x_I]}{\pi_{ref}(y_{neg}|[x_T, x_I]}\right)\right]$$

Implicit reward

- Concurrently training reward and policy model skewing the reward model to prefer positive responses
- Reward learned implicitly

$$L = L_{dpo} + \lambda L_{aux}$$

- For stability, we have an additional auxiliary loss

$$L_{aux} = -\sum \log P(y|\pi_P; \pi_\theta), \{\pi_P, y\} \sim D_{sft}$$

- x_T - text prompt
- x_I - image prompt
- [ ] - feature concatenation
- π_{ref} - reference model
- π_{theta} - policy model
- D – style consistent dataset
- $\log\sigma$ - log sigmoid function

- x_P - text prompt

The L_{aux} loss is from InstructGPT

New Benchmark: Sentence Level Hallucination Ratio (SHR) (1/3)

- Limitations in current benchmark (eg POPE)
  - Few evaluation target categories
  - Need more categories, attributes, emotions, and other elements.
- SHR
  - Comprehensive, broad
  - All textual descriptions that do not match image
  - 200 images for validation

$$SHR = \frac{\sum_{i=1}^{N} h_i}{\sum_{i=1}^{N} s_i}$$

- N is the total number of images,
- h_i number of hallucinated sentences
  - 200 images from VG for validation
  - determined by GPT-4 based on model outputs & image annotations
- s_i all sentences in the response

Diversity covered in SHR (2/3)

Movement Hallucination

LVLM

One person is standing at the top of the slope, holding a ski pole, while the others are skiing down the slope.

GPT-4

Judgement: Hallucination. There is no description mentioning one person standing at the top of the slope while the others are skiing down.
Revise: A woman in ski gear is smiling at the camera, holding ski poles.

Emotion Hallucination

LVLM

The cat is looking up at the person with a curious expression on its face.

GPT-4

Judgement: Hallucination. There's no specific mention of the cat's expression in the region descriptions.
revise: The cat is actively engaging with the banana.

Spatial relation Hallucination

LVLM

The pole is made of metal and has a green light on top and a red light on the bottom.

GPT-4

Judgement: Hallucination. The green light is not on top and the red light is not on the bottom.
Revise: The pole is made of metal with a traffic signal hanging from it, showing a green light.

Nonexistent Object

LVLM

There are balloons and streamers hanging from the ceiling.

GPT-4

Judgement: Hallucination. There is no mention of balloons and streamers hanging from the ceiling in the region descriptions.
Revise: The room they are in has windows with bars and a railing visible outside.

Object Attribute Hallucination

LVLM

The table is made of wood and has a white tablecloth on it.

GPT-4

Judgement: Hallucination. The table is not described as made of wood in the region descriptions.
Revise: The image shows a table with a blue and white patterned tablecloth.

Physical State Hallucination

LVLM

The image shows a chocolate cake with orange slices on a white plate.

GPT-4

Judgement: Hallucination. There are no orange slices on the cake or plate, they are whole oranges.
Revise: The image shows a chocolate cake and whole oranges on a white plate.

Figure 5. SHR evaluation covers diverse types of hallucinations.

Advantages of SHR (3/3)

- Reliability
  - Includes manual annotations
  - 95% accuracy
- In addition to GPT-4 judging the content
  - Add extra factual info.
  - Compare MiniGPT4, GPT4 LLaVA-1.5, InstructBLIP with and without on 20 images (250 sentences) with manual checks

	P	R
H	88.81%	89.43%
C	86.99%	87.70%

(a) w/o factual information.

	P	R
H	97.76%	92.25%
C	92.60%	97.56%

(b) w factual information.

- Universality
  - Unlimited # object types (unlike COCO's 80) in VG images
- Comprehensiveness
  - wide spectrum of hallucinations tags any description that contradicts image content
  - Includes nonexistent objects, emotions, attributes, movements, etc

Experiments

- **Training Data**
  - Filtering on VG
  - Random 2k images
  - 3 GPT-4 rewrites – 2k images, 6k hallucinatory, 6k non-hallucinatory responses
  - Added Q&A format – 10k data pairs added
  - Total: 16k data pairs
- **POPE Evaluation**
  - 9,000 questions of 3 types.
  - POPE targets at object existence of fixed categories (80 COCO) in images
  - Yes/No responses.
  - Benchmarked against the ground truth answer.
- **SHR Evaluation**
  - 200 images from the VG dataset.
  - Output detailed descriptions for these 200 images.
  - SHR ratio measured
  - GPT-4 determines if a sentence is hallucinated by comparing the model output with VG annotations and human-annotated factual information.

Implementation Details

Mini-GPT-4

- Fine tune via LoRA
- Fixed except $q_{proj}, k_{proj}, v_{proj}$
- LoRA rank 64, $\alpha = 16$
- Cosine scheduler, LR 1e-4, warmup ratio 0.03, 100 rounds
- Batch size = 1
- HA-DPO β $0.1, \lambda = 0.5$
- 8 A100 GPUs 1-2 hours for 1k steps

InstructBLIP-13B

- Fine tune via LoRA
- Fixed except $q_{proj}, k_{proj}, v_{proj}$ in LM
- LoRA rank 64, $\alpha = 16$
- Cosine scheduler, LR 4e-6
- Batch size = 1
- HA-DPO β $0.1, \lambda = 0$
- 8 A100 GPUs, 1 epoch less 1 hour

LLaVA-1.5-7B

- Fine tune all linear layers with LoRA
- LoRA rank 256, $\alpha = 128$
- LR 2e-6, Cosine scheduler
- Batch size = 16
- HA-DPO $\beta = 0.1, \lambda = 0.0$
- 8 A100 GPUs, 1 epoch less 1 hour

Ablations - Hyperparameter β

- **Question: What do we expect ?**
  - Too small β ?
    - HA-DPO training is unstable
    - Model mostly learns noise rather than how to distinguish hallucinations
  - Too large β ?
    - loss more focused on constraining the consistency between the policy model and the reference model
    - Can't distinguish hallucination from non-hallucination

β	SHR ↓	1-gram	2-gram	3-gram	4-gram
0.3	57.2	56.7	82.3	86.4	87.9
0.4	55.8	57.8	84.3	88.4	90.0
0.5	52.3	59.0	85.9	90.3	91.8
0.6	51.4	60.1	87.4	91.7	93.1
0.8	52.3	59.0	85.9	90.3	91.8
1.0	56.7	61.0	88.8	93.1	94.6

Table 1. Ablation studies on β. Too low β can lead to unstable training, and too high β constraint model from learning knowledge about how to distinguish hallucinations.

Results – Hallucination Mitigation (POPE)

POPE	Model	HA-DPO	Accuracy	Precision	F1 Score	Yes Ratio (%)
Random	MiniGPT-4-LLama2-7B [36]	✗	51.13	50.57	67.13	98.66
		✓	86.13	92.81	84.96	42.20
	InstructBLIP-13B [5]	✗	88.70	85.03	89.26	55.23
		✓	89.83	**93.07**	89.43	46.23
	LLaVA-1.5-7B [13]	✗	89.60	88.77	89.70	**51.06**
		✓	**90.53**	92.99	**90.25**	47.13
Popular	MiniGPT-4-LLama2-7B [36]	✗	51.46	50.74	67.72	98.06
		✓	79.50	80.20	79.25	48.83
	InstructBLIP-13B [5]	✗	81.36	75.06	83.44	62.56
		✓	85.76	85.55	85.80	**50.03**
	LLaVA-1.5-7B [13]	✗	86.20	83.23	86.79	54.46
		✓	**87.90**	**88.07**	**87.81**	49.76
Adversarial	MiniGPT-4-LLama2-7B [36]	✗	51.26	50.64	67.16	98.40
		✓	75.66	74.36	76.29	**52.66**
	InstructBLIP-13B [5]	✗	74.50	67.64	78.64	69.43
		✓	80.70	77.72	81.68	55.36
	LLaVA-1.5-7B [13]	✗	79.76	74.43	81.75	60.90
		✓	**81.46**	**77.99**	**82.54**	56.20

- Hallucinations reduced on all datasets (random, popular and adversarial) for all methods
- Mini-GPT-4
  - Most improvement
- LLaVA
  - Least improvement since SFT was already pretty good due to data diversity

Table 2. Results on POPE Benchmark: HA-DPO significantly enhances the model's ability to discern hallucinatory objects in images.

Results – Hallucination Method Comparison

Method	Random		Popular		Adversarial	
	Accuracy	F1 score	Accuracy	F1 score	Accuracy	F1 score
LRV [12]	86.00	88.00	73.00	79.00	65.00	73.00
LLaVA-RLHF [26]	84.80	83.30	83.90	81.80	**82.30**	80.50
LLaVA-1.5-7B	89.60	89.70	<u>86.20</u>	<u>86.79</u>	79.76	<u>81.75</u>
InstructBLIP-13B	88.70	89.26	81.36	83.44	74.50	78.64
MiniGPT4-LLaMA2-7B	51.13	67.13	51.46	67.72	51.26	67.16
LLaVA-1.5-7B *w HA-DPO*	**90.53**	**90.25**	**87.90**	**87.81**	<u>81.46</u>	**82.54**
InstructBLIP-13B *w HA-DPO*	<u>89.83</u>	<u>89.43</u>	85.76	85.80	80.70	81.68
MiniGPT4-LLaMA2-7B *w HA-DPO*	86.13	84.96	79.50	79.25	75.66	76.29

Table 7. Results comparisons with other hallucination mitigation methods on POPE Benchmark. HA-DPO outperforms other competitive methods and achieves SOTA (state-of-the-art) in POPE accuracy and F1 score. Bolded denotes the best score and underline denotes the second best score.

- HA-DPO despite only 2k images, 16k +ve/-ve pairs achieves SOTA performance (2nd best on adversarial accuracy)
- LRV, LLaVA-RHLF used 400K, 160K training data

Results – Hallucination Mitigation (SHR)

Model	HA-DPO	SHR ↓
MiniGPT-4-LLama2-7B	✗	47.3
	✓	**44.4**
InstructBLIP-13B	✗	51.2
	✓	**49.1**
LLaVA-1.5-7B	✗	36.7
	✓	**34.0**

Table 3. Hallucination evaluation results on SHR benchmark.

- Hallucinations went down, but only marginally

➢ **We still have a long way to go to solve hallucination in LVLMs**

Results – General Performance Enhancement

Model	HA-DPO	Perception	Cognition
MiniGPT-4-LLama2-7B	✗	733.79	198.21
	✓	**1092.18**	**234.28**
InstructBLIP-13B	✗	1344.91	232.50
	✓	**1416.23**	**233.21**
LLaVA-1.5-7B	✗	**1510.74**	**355.71**
	✓	1502.58	313.93

Table 4. Results on MME Benchmark. Recognition and cognition each represent two major capability dimensions in MME, examining recognition and perceptual reasoning ability, respectively.

- Good general improvement on simpler models with limited SFT
- Performance drop on LLaVA-1.5-7B which already had good SFT due to data diversity. Does LoRA for hallucinations unlearn general capabilities ?

Some examples

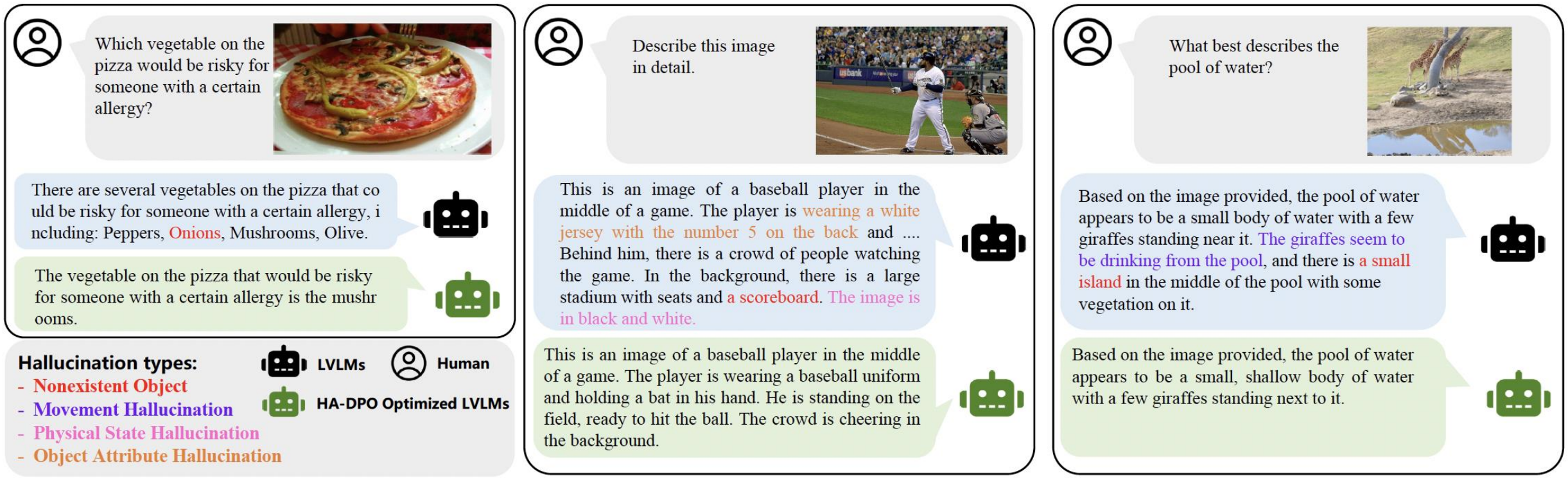


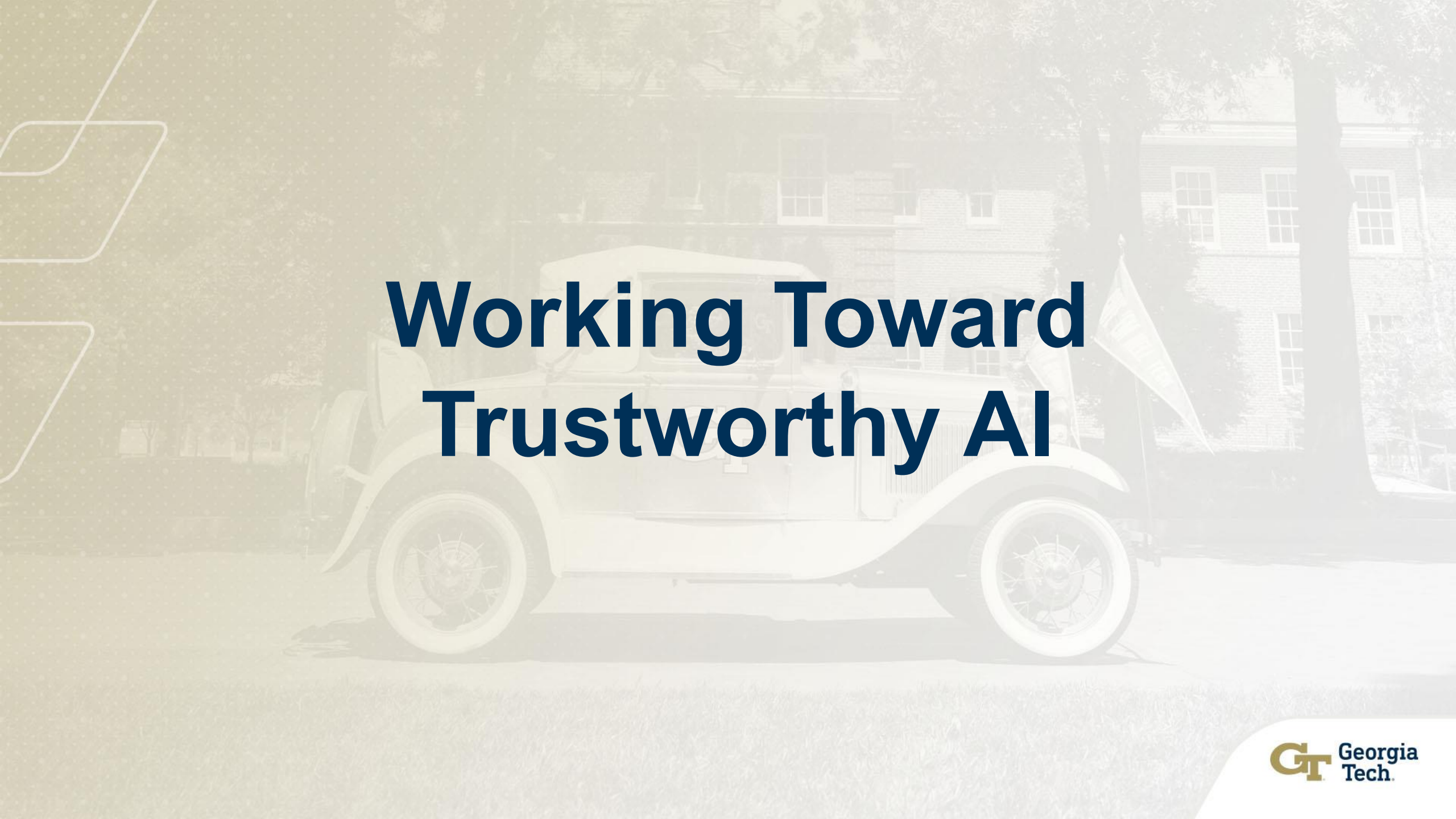
Figure 6. Comparison of model responses before and after our proposed hallucination elimination method.

- object existence hallucination,
- object attribute hallucination
- movement hallucination,
- physical hallucination

Conclusion & Discussion

- Limitations:
  - Tested on small VLMs (MiniGPT-4, InstructBLIP-7B, LLaVA-1.5-13B)
  - Needed to tune the HA-DPO parameters for each model
  - Is GPT-4 a good judge of hallucinations ?
  - Did not ablate λ to study effect of L_{aux}, in fact $\lambda = 0$ was often used
  - Unclear if increasing the size of HA-DPO training set improve performance (currently 2k images, 16k +ve/-ve pairs) ?
  - Did not benchmark other hallucination methods on SHR
  - Unclear how much it improves over **large** models with **good SFT.** Does general performance go down (eg LLaVA)?
- What we achieved with HA-DPO
  - Good dataset
    - Well matched +ve/-ve pairs
    - Q&A as well as descriptive
  - Good metric (SHR) – sentence level hallucinations irrespective of type of hallucination
- What's lacking in HA-DPO
  - Performance still not great (eg. SHR metric)
  - Only on text generation
  - Real-world situations
  - SHR is small - comprises only 200 images only from VG dataset

Working Toward Trustworthy AI



Georgia Tech

Have we solved the hallucination problem ?

- We have seen three paradigms – SFT, post-hoc model output processing and RL for hallucination mitigation
- While effective, **they do not completely solve the problem**
- Hallucinations are an inherent part of an LM or VLM

Why haven’t we solved the hallucination problem ? (Discussion)

- Models optimized to be good test-takers
- Guessing when uncertain improves test performance
- Models are penalized for IDK answers in benchmarks, leading the researchers to choose models which guess the answers.
- Binary grading rewards confident answers
- Misalignment between evaluation metrics and real-world needs
- Most work on hallucinations is done on the language output not on the generated image.

Student Analogy: When uncertain, students may guess on multiple-choice exams and bluff on written exams, submitting plausible answers with little confidence. Language models face similar incentives.

Kalai Adam Tauman, Nachum Ofir, Vempala Santosh S., Zhang Edwin "Why language models hallucinate." *arXiv preprint arXiv:2509.04664* (2025).

Proposed solution – a new evaluation strategy

Explicit Confidence Targets

- Add **confidence thresholds** to evaluation instructions.
- Similar to human exams with penalties for wrong answers.

Behavioral Calibration

- Alternative to **probabilistic confidence**
- Output IDK when correctness probability exceeds threshold

What else do you think we can do ?

Georgia Tech