Generate, but Verify: Reducing Hallucination in Vision-Language Models with Retrospective Resampling

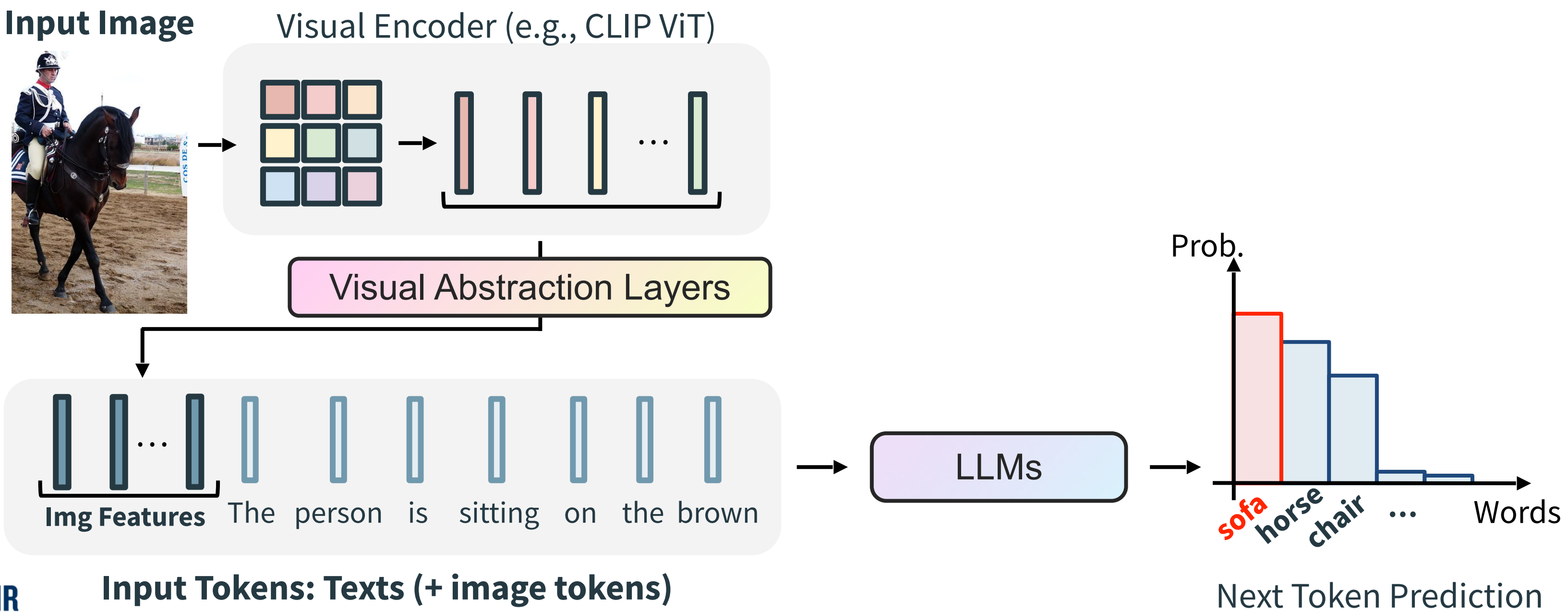
Tsung-Han (Patrick) Wu, Heekyung (Anne) Lee, Jiaxin Ge,
Joseph E. Gonzalez, Trevor Darrell, David M. Chan

UC Berkeley

NeurIPS 2025

VLMs Suffer from Hallucinations

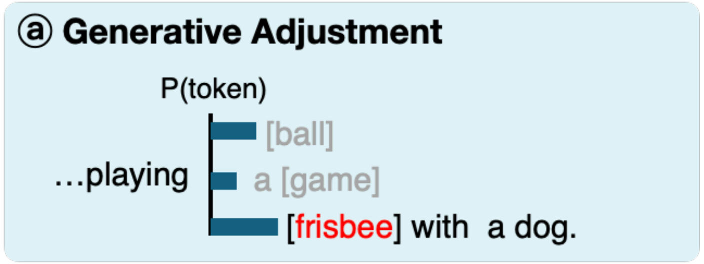
Visual Hallucinations: Describing nonexistent objects or concepts in the image, usually due to *training data biases* or *strong language priors*



Prior Methods

User “Describe this image.”

VLM “The boy is sharing his umbrella…



Preventing hallucinations is hard, but detection isn’t — we need a **verifier** after the fact.

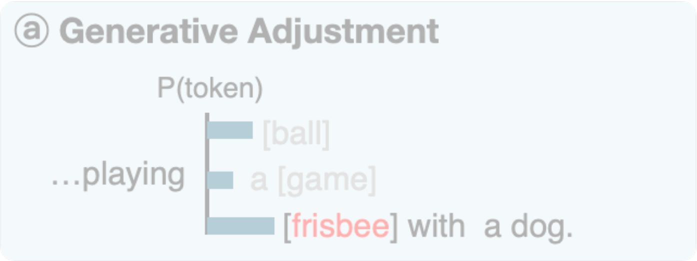
Prior Methods

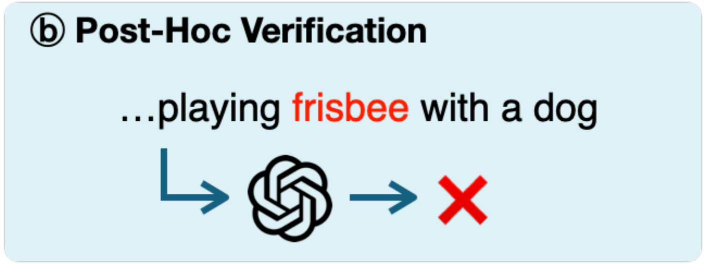
User "Describe this image."

VLM "The boy is sharing his umbrella…

Preventing hallucinations is hard, but detection isn't — we need a **verifier** after the fact.

Verifying after full outputs is slow and mostly ends in refusal — we need **correction** instead!





Prior Methods

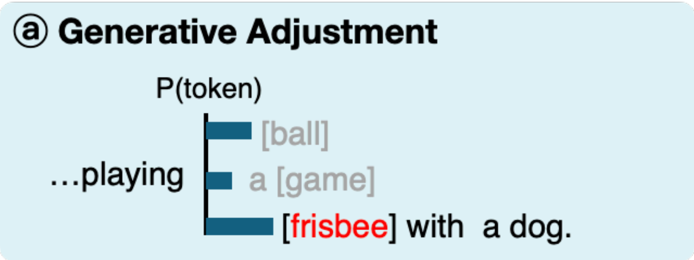
User “Describe this image.”

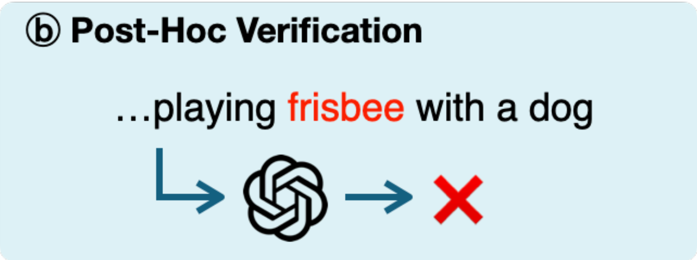
VLM “The boy is sharing his umbrella…

Preventing hallucinations is hard, but detection isn’t — we need a **verifier** after the fact.

Verifying after full outputs is slow and mostly ends in refusal — we need **correction** instead!

Why not both?





BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

REVERSE-VLM

REtrospective VERification and SElf-Correction

Generate, but Verify: The birth of REVERSE-VLM

A single VLM that can not only **generate**, but **verify and corrects** themselves on-the-fly

→ towards robust, controllable, and interpretable systems!

How to Enable This?

User “Describe this image.”

VLM “The boy is sharing his umbrella…

(1) **Train** a model that can do explicit confidence estimation

…with a girl </UN>

Unconfident! Backtrack and Correct!

…with a parent </UN>

Still bad, try again!

…with a cat </CN> on a rainy day

(2) Do **retrospective resampling** inference for multi-round correction

1. SFT Dataset Construction

Source: 665K LLaVA-SFT Dataset

<SPAN>: Noun Phrase Opening   </CN>: Confident Token   </UN>: Unconfident Token

Human: What feature can be seen on the back of the bus?

GPT: The back of the bus features an advertisement.

Noun Phrase Extraction & Tagging

GPT: <SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>an advisement</CN> .

Positive Data

Negative Phrase Augmentation

GPT: <SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window</UN> .

Negative Data

1. SFT Dataset Construction

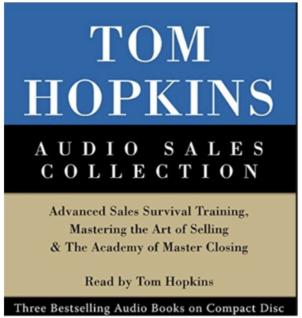
1.3M Open-Source Datasets on 🤗

Image

COCO/train2017/

Captioning Question

VQA task



	COCO/train2017/	Captioning Question	VQA task
Question	"How many total baseball players are shown in the image?"	"Describe this image in your own words."	"Who wrote this book?"
Pos Answer	"There are <SPAN>**three baseball players**</CN> shown <SPAN>in the image</CN> .",	"The image features <SPAN>**an old military aircraft**</CN> <SPAN>on display</CN> ..."	"<SPAN>**Tom Hopkins**</CN>"
Neg Answer	"There are <SPAN>**five soccer players**</UN>"	"The image features <SPAN>**a modern commercial airplane**</UN>"	"<SPAN>**John Steinbeck**</UN>"

Number | Attribute | Object

2. Hallucination-Aware Training

Model needs to:

1. **Do standard next token prediction**
2. **Avoid hallucination modeling**
3. **Learn to model confidence with </CN> and </UN>**

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>an advertisement</CN>

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window</UN>

Model trained to predict these tokens

Model **ignores** these tokens during training

3. Retrospective Resampling Inference

<SPAN>The back</CN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

3. Retrospective Resampling Inference

<SPAN>The back</CN> <SPAN>of the bus</CN> features

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

3. Retrospective Resampling Inference

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>a window</UN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.20 <UN>: 0.87

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

3. Retrospective Resampling Inference

<SPAN>The back</CN> <SPAN>of the bus</CN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

3. Retrospective Resampling Inference

<SPAN>The back</CN> <SPAN>of the bus</CN> features <SPAN>an advertisement</CN>

<UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0 <UN>: 0.0

Summary: Retrospective Resampling

① **User Query**

User

"Describe this image."

Experimental Results

SOTA on Captioning & Open-ended VQAs

Captioning Tasks

Base VLM	Method Type	Method	CHAIR-MSCOCO		AMBER-G			
			CHAIR_i (↓)	CHAIR_s (↓)	CHAIR (↓)	Cover (↑)	Hall (↓)	Cog (↓)
LLaVA-v1.5 7B [35]	None		15.4	50.0	7.8	51.0	36.4	4.2
	Gen-Adjust	VCD [28]	14.9	48.6	-	-	-	-
		OPERA‡ [23]	14.6	47.8	7.3	49.6	32.0	3.5
		DoLA† ‡ [16]	14.1	51.6	7.6	51.6	36.0	4.0
		AGLA [3]	14.1	43.0	-	-	-	-
		MEMVR [58]	13.0	46.6	-	-	-	-
	w/ Train	EOS [55]	12.3	40.2	5.1	49.1	22.7	2.0
		HALVA [41]	11.7	41.4	6.6	**53.0**	32.2	3.4
		HA-DPO [56]	11.0	38.2	6.7	49.8	30.9	3.3
	Post-hoc Refine	Woodpecker† [53]	14.8	45.8	6.9	48.9	30.4	3.6
	Combination	**REVERSE**$_{(\tau=0.003)}$	10.3	37.0	6.0	52.2	30.4	3.0
		REVERSE$_{(\tau=0.0003)}$	**6.1**	**13.6**	**4.0**	26.9	**10.2**	**0.9**

12% Gain

MM-Hal

Base VLM	Method	Score (↑)	Hall. Rate (↓)
LLaVA-MORE 8B	None†	2.50	0.53
	DoLA† [15]	2.54	0.51
	Woodpecker† [51]	2.28	0.58
	REVERSE$_{(\tau=0.003)}$	2.28	0.54
	REVERSE$_{(\tau=0.0003)}$	**2.93**	**0.40**

HaloQuest 34% Gain

Method	Avg. Acc. (↑)	FP Acc.	VC Acc.	IC Acc.
Qwen2.5-VL^{FT} 3B				
None†	33.5	25.4	**51.6**	26.4
DoLA† [15]	27.4	16.5	51.1	19.0
REVERSE$_{(\tau=0.01)}$	**45.1**	**42.9**	41.8	**55.5**

+ Multiple Models

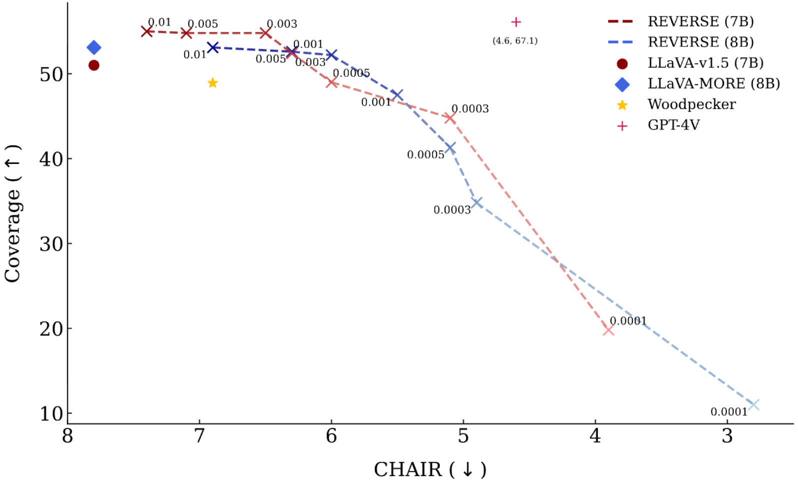
+ Multiple Tasks

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Towards Efficient Corrections & Controllable VLMs

Studies on AMBER-G Dataset

# Rounds (N)	0	5	10	20	50
CHAIR (↓)	7.8	7.1	6.8	6.7	6.0
#Tokens (%)	1.00×	1.75×	2.05×	2.63×	3.05×



15% gain from **50** round corrections but only **3.05x** more tokens

Tuning the **threshold (τ)** can control the trade-off between **expressiveness** & **hallucinations**

We can beat GPT-4V on the CHAIR metric with low threshold, making VLMs conservative!

BAIR
BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Conclusions

Takeaways

➔ Current VLMs are still prone to visual hallucinations.

BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

Takeaways

➔ Current VLMs are still prone to visual hallucinations.

➔ When verification is easier, we can insert retrospective reasoning tasks to encourage the model to “think twice” about its process.

Takeaways

- Current VLMs are still prone to visual hallucinations.
- When verification is easier, we can insert retrospective reasoning tasks to encourage the model to “think twice” about its process.
- We present REVERSE, the first and effective hallucination reduction method unifying the generation adjustment and post-hoc verification.

Takeaways

- Current VLMs are still prone to visual hallucinations.
- When verification is easier, we can insert retrospective reasoning tasks to encourage the model to “think twice” about its process.
- We present REVERSE, the first and effective hallucination reduction method unifying the generation adjustment and post-hoc verification.
- REVERSE is not endpoints but starting points that invite the community to explore the “generate-but-verify” paradigm.

Takeaways

Project Page

- ➔ Current VLMs are still prone to visual hallucinations.
- ➔ When verification is easier, we can insert retrospective reasoning tasks to encourage the model to "think twice" about its process.
- ➔ We present REVERSE, the first and effective hallucination reduction method unifying the generation adjustment and post-hoc verification.
- ➔ REVERSE is not endpoints but starting points that invite the community to explore the "generate-but-verify" paradigm.

Thanks for listening!