HallusionBench → SAGE

Authors: Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, Tianyi Zhou

CVPR 2024, June 2024

Authors:
Tripti Shukla, Zsolt Kira

arXiv preprint, March 29, 2026

Yasmine Grane, Ghazal Mirzazadeh, Aryan Shah

CS8803-VLM Presentation

Georgia Tech

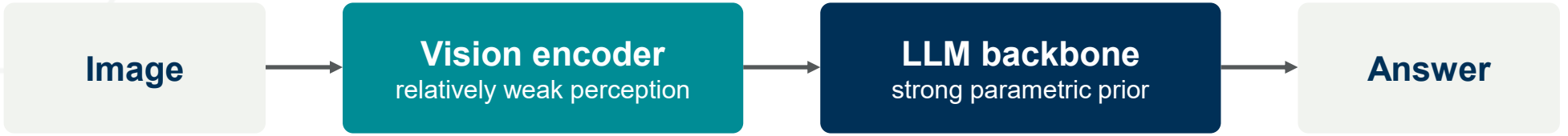
Outline

- Problem Statement
- Related Works
- Approach : Seminal and frontier
- Experiments & Results
- What persisted and what changed?
- Limitations and Implications
- Discussion Questions

Georgia Tech

Problem Statement

Do VLMs see what they think, or think what they see?



Strong LLMs bolted onto weaker vision encoders: a double-edged sword

Language hallucination

The LLM's prior knowledge overrides the image

- Answers what is usually true, not what is shown
  e.g., an edited map still gets the textbook answer

Visual illusion

The vision side misreads what is in the image

- Mis-perceived content becomes a confident wrong answer
  e.g., illusions, geometry, charts, frame order

Prior benchmarks report whether a model hallucinates. HallusionBench asks why, and separates the two causes, which are usually entangled.

Georgia Tech

Related Works

Work	What it measures	How	Blind spot
CHAIR (Rohrbach et al., 2018)	Object hallucination in captions	Share of mentioned objects absent from the image	Captions only; no cause
POPE (Li et al., 2023)	Object existence	Yes/no polling with random, popular, adversarial negatives	Natural images; objects only
GAVIE / LRV-Instruction (Liu et al., 2023)	Instruction-level hallucination	GPT-4-judged open answers; negative instructions for tuning	Scores errors, does not isolate why
MME, MMBench, MM-Vet, SEED-Bench	General multimodal ability	Broad perception and reasoning QA	Hallucination is not the focus
Woodpecker (Yin et al., 2023)	Mitigation	Training-free post-hoc correction of outputs	Fixes outputs; no diagnosis

Georgia Tech

HallusionBench at a Glance

346

images (165 original, 181 human-edited)

1,129

yes/no questions written by experts

455

visual-question control pairs

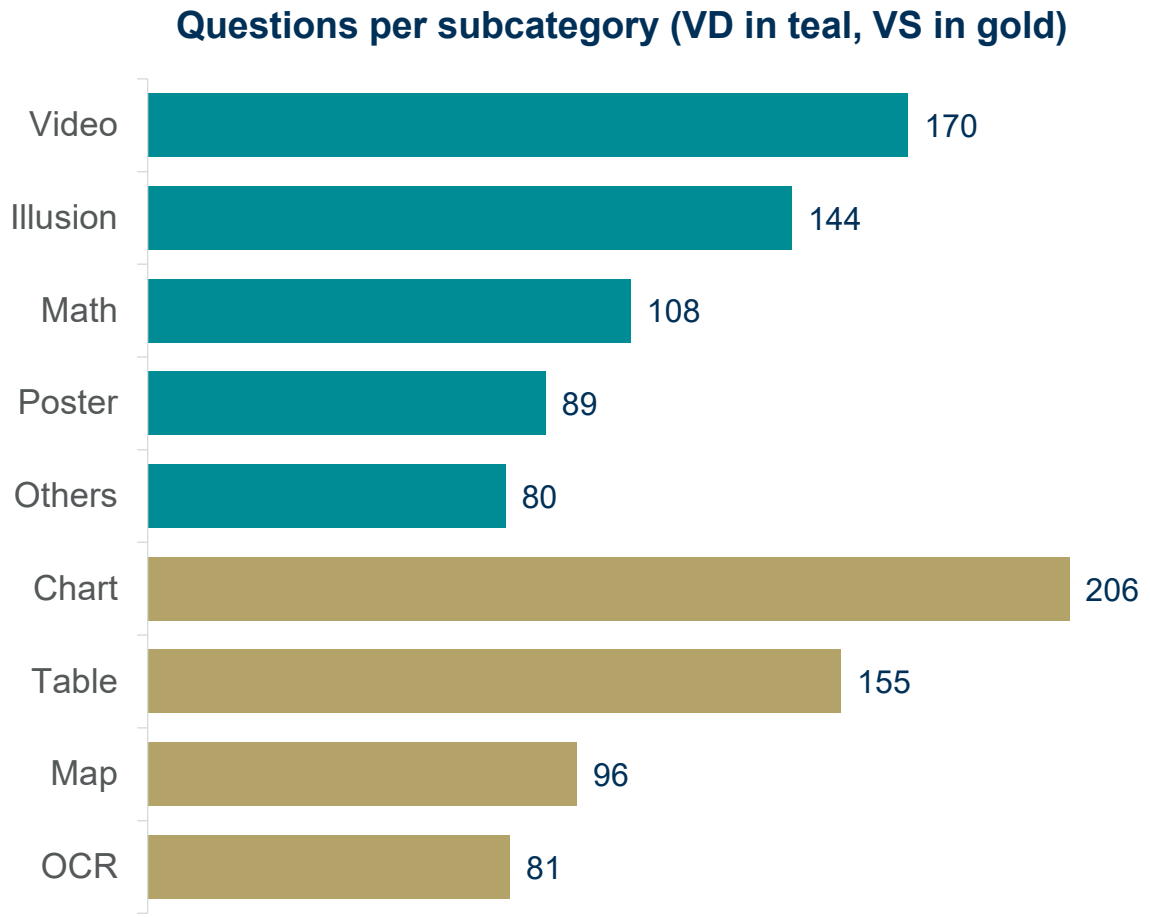
15

LVLMs evaluated, plus random chance

Visual Dependent (VD): no correct answer without the image

Visual Supplement (VS): answerable from world knowledge; the image adds or contradicts

Easy = original image; Hard = human-edited image



Data: Guan et al., HallusionBench (CVPR 2024), Table 2 (VD 591 Q, VS 538 Q)

Georgia Tech

HallusionBench: Control Pairs

Same question, two images
The original image and a human-edited copy where the correct answer is different

Isolates what the model actually sees
Only the image changes, so any change in the right answer comes from the visual content alone

Separates seeing from remembering
A model grounded in the image changes its answer
A model relying on prior knowledge gives the same answer and gets one of the two wrong

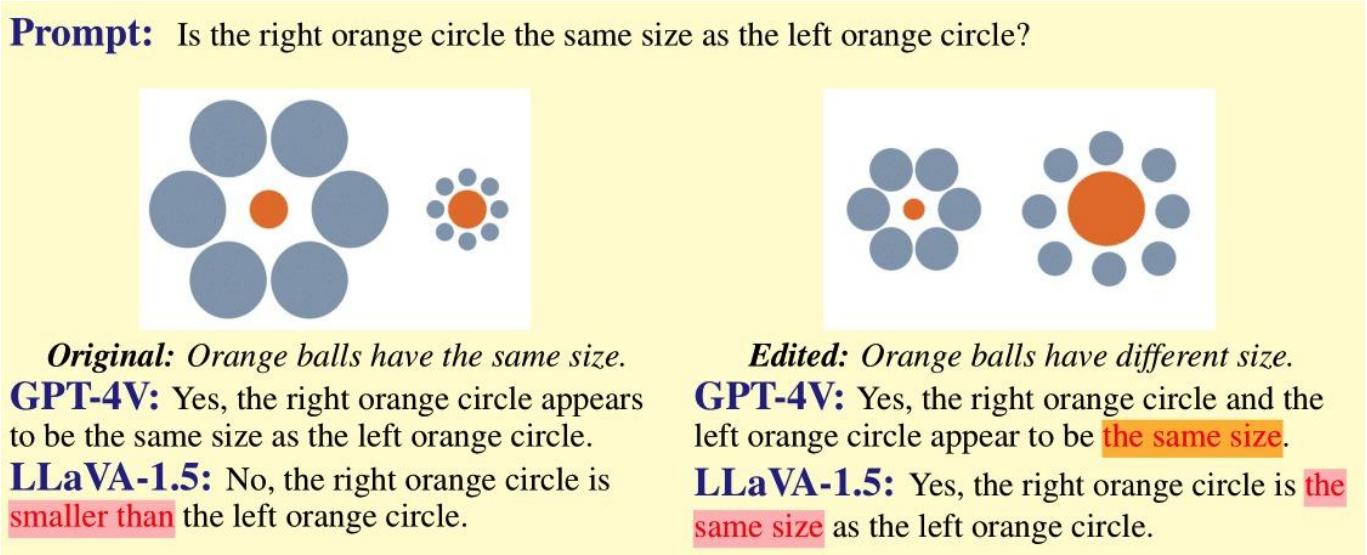


Figure from Guan et al., HallusionBench (CVPR 2024), Fig. 1 (excerpt). Red = visual illusion; orange = possibly mixed

On the edited image, both models still say "same size".

Georgia Tech

How the Edits Are Made

Six edit types, each chosen so that the correct answer changes while the question stays the same

1

Flip / rotate

Mirror an image or rotate a chart

Left/right and up/down answers flip

2

Order reversal

Reverse video frames or a sequence

e.g., the car now leaves the spot

3

Masking

Hide part of the image

The evidence the answer needs is gone

4

Text (OCR) edit

Change digits or words

e.g., π printed as 3.1615

5

Object edit

Resize, lengthen or move objects

e.g., enlarge one orange circle

6

Color edit

Recolor or swap regions

e.g., Delaware and Arizona swapped

Where the questions come from

45% on human-edited images

39% on original images found online

16% with no image at all (VS, knowledge only)

Hand-made: experts collected all 346 images and wrote every question and answer

Guan et al., HallusionBench (CVPR 2024), Sec. 3 (dataset construction)

Georgia Tech

Evaluation: Three Levels of Strictness

aAcc

All-question accuracy

Each yes/no question scored on its own

Most lenient

fAcc

Figure accuracy

Every question about one figure must be right: a consistency test

Stricter

qAcc

Question-pair accuracy

The same question must be right on every image in its control group

Strictest: headline metric

Scoring

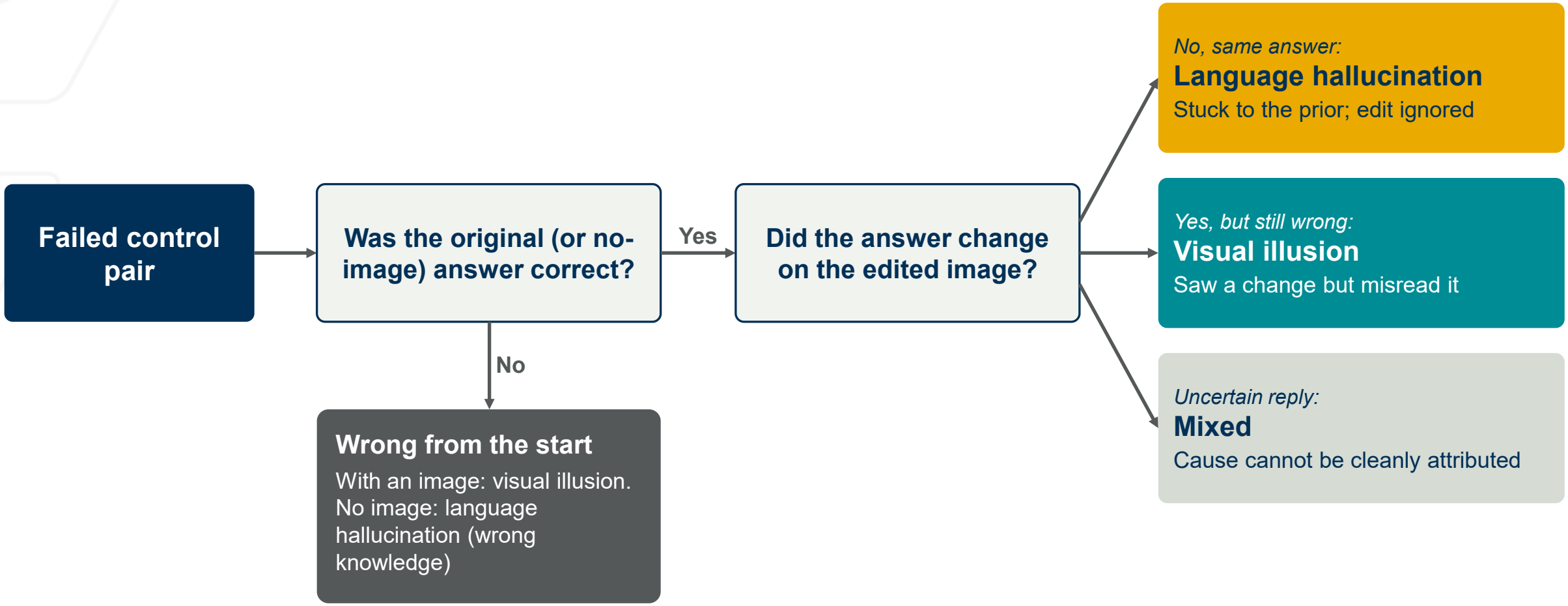
- GPT-4 judge labels each reply correct, incorrect, or uncertain; 3 runs at temperature 0, averaged
- Tracks human grading closely (GPT-4V qAcc: 31.4 human vs. 28.8 GPT-4)
- "Uncertain" counts as correct only for no-image VS questions

Analytical tests

- Yes/No bias: % difference in "yes" rate vs. ground truth (ideal 0) and false-positive ratio among errors (ideal 0.5)
- Consistency: a model is inconsistent on a figure if it gets only some of its questions right

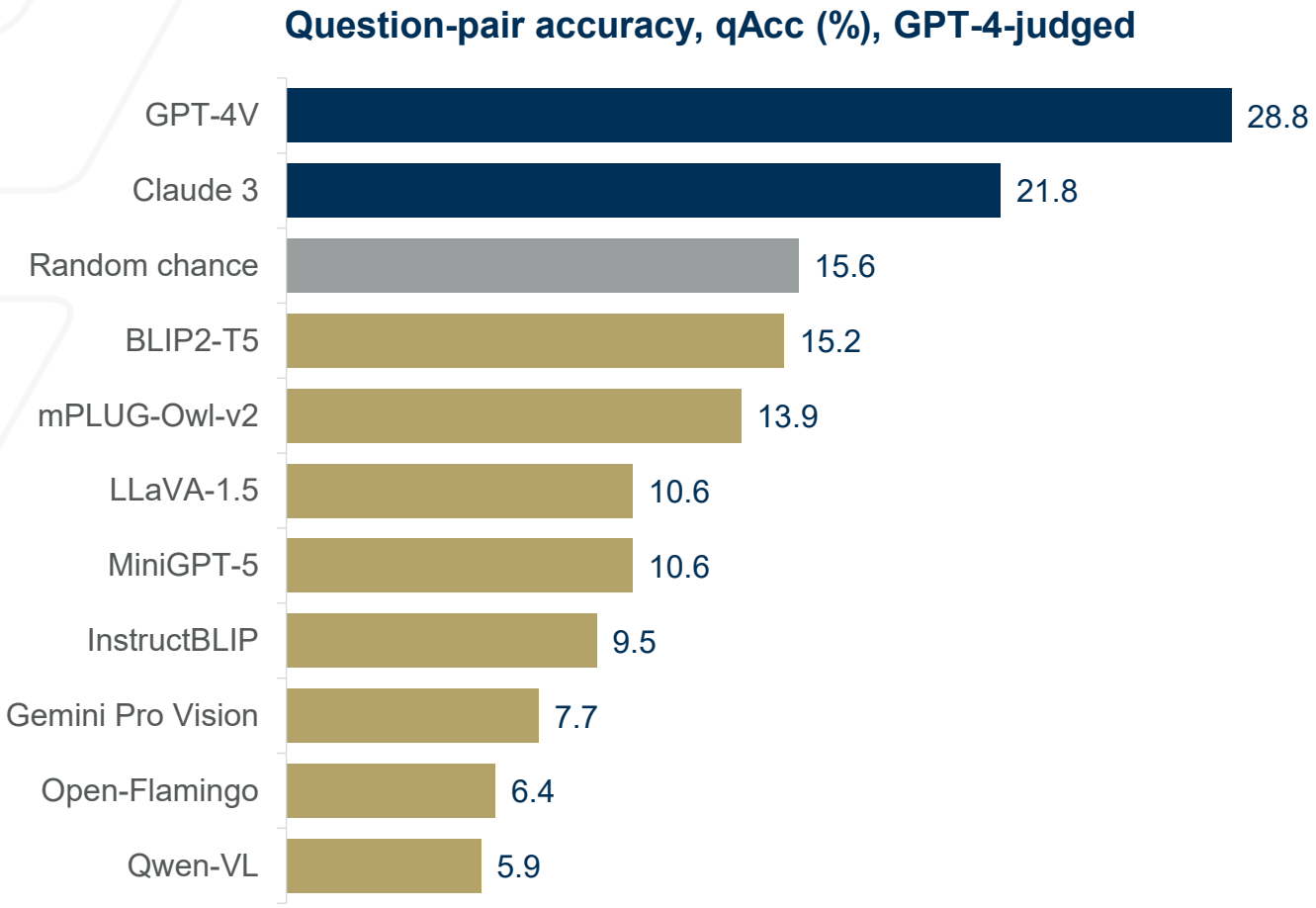
Diagnosing Why a Model Failed

Each failed control pair is attributed to one cause by comparing the two answers



Simplified from the decision tree in Guan et al., HallusionBench (CVPR 2024), Fig. 3

Results: Most Models Fall Below Chance



31.4%

GPT-4V's best qAcc (human-graded): it fails over two-thirds of control pairs

2

models beat random chance (15.6%) on qAcc: GPT-4V and Claude 3

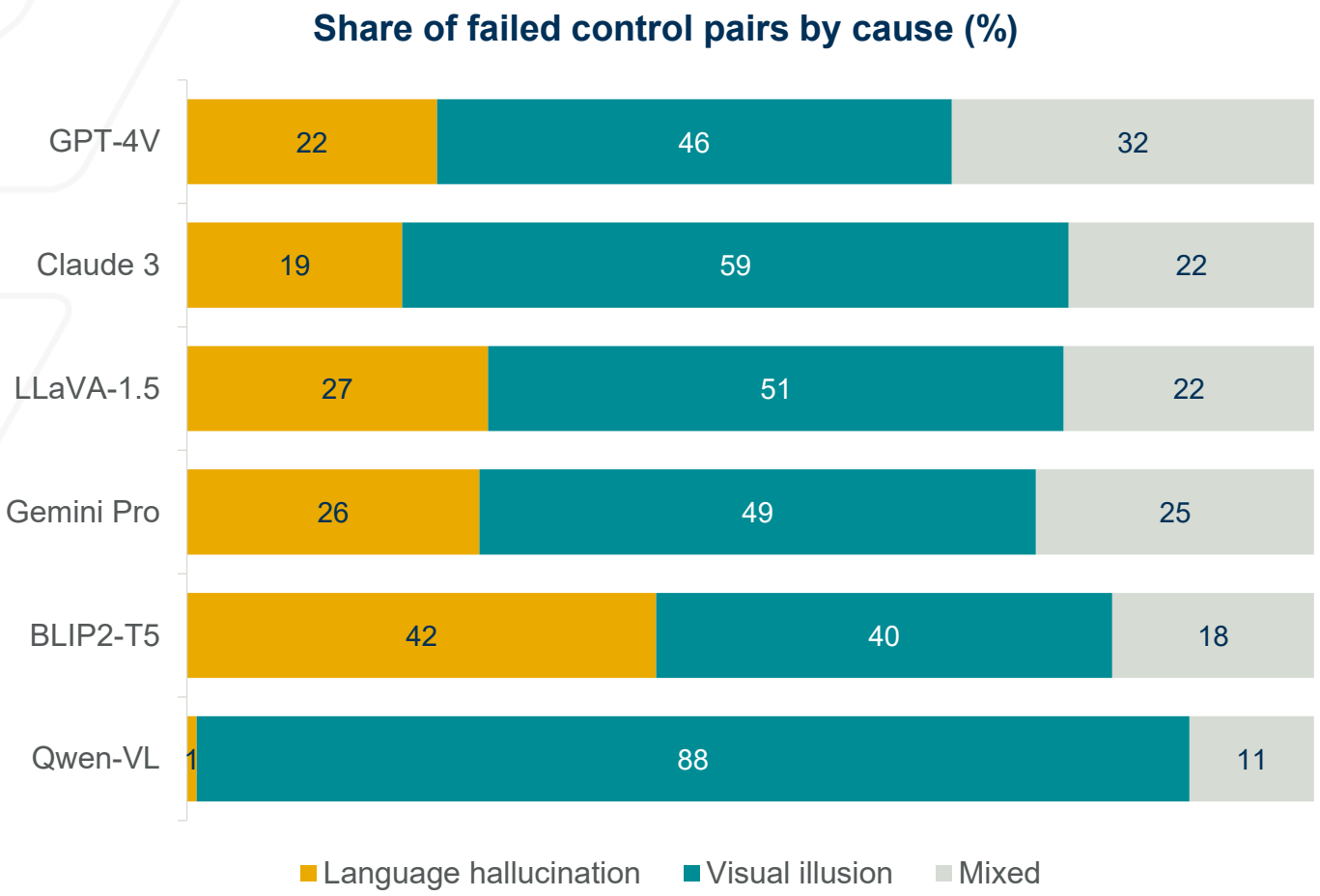
79.6 → 38.4

GPT-4V accuracy drops from easy to hard (edited) questions

Data: Guan et al., HallusionBench (CVPR 2024), Table 3 and project leaderboard (subset of models shown)

Georgia Tech

Failure Analysis: Which Mistake, How Often?



Data: Guan et al., HallusionBench (CVPR 2024), Table 3 (GPT-4-judged)

Visual illusion dominates

- Largest cause for nearly every model (46 to 88%); BLIP2-T5 is a near tie. Perception is the main bottleneck

Priors bite the capable models

- About 20 to 27% for GPT-4V, Claude 3, LLaVA-1.5, Gemini; 42% for BLIP2-T5
- Weak models (Qwen-VL: under 1%) fail before priors even matter

Bias and inconsistency

- LLaVA-1.5 leans "yes" (FP ratio 0.75; ideal 0.5)
- Inconsistent on 38% of figures (GPT-4V) to 60% (Gemini)

Georgia Tech

Bias and Consistency: Beyond Accuracy

Yes-rate difference

$d = (1/|V|) \Sigma [\ 1(\hat{y} = yes) - 1(y = yes)\ ]$

Range −1 to 1. Ideal 0. Positive = says "yes" too often

False-positive ratio

$r = (1/|W|) \Sigma$ over wrong answers $1(\hat{y} = yes)$

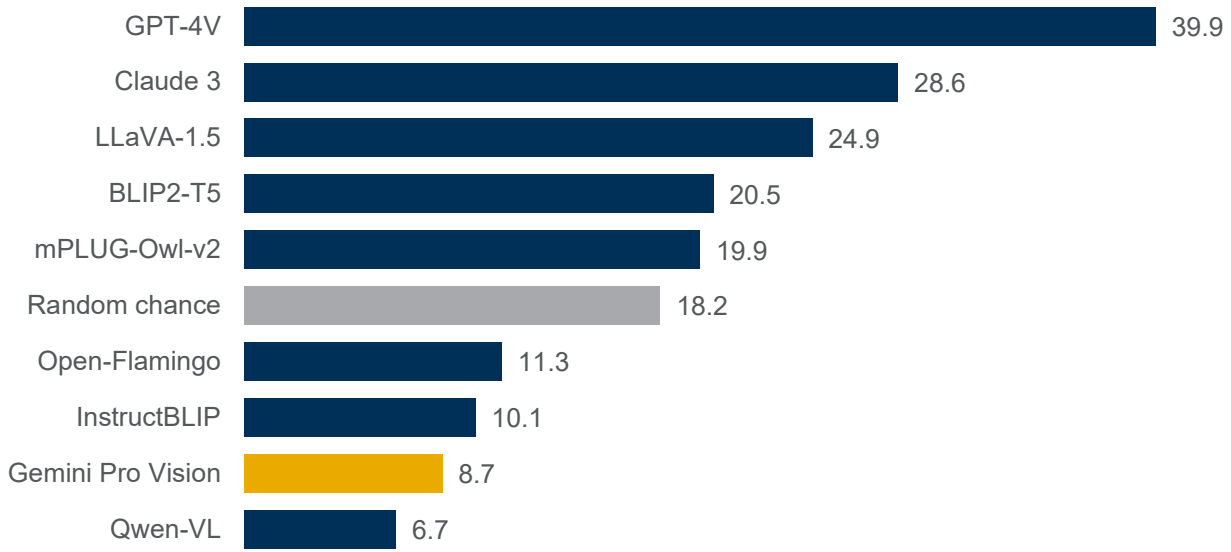
Share of errors that are a wrong "yes". Ideal 0.5

Figure consistency

Correct · Inconsistent · Wrong

All, some or none of a figure's questions right

Figures with every question right (%), GPT-4-judged



5 of 15

models are more consistent than random guessing (18.2%)

No bias ≠ grounded

Gemini Pro Vision is the least biased model (d = −0.02, r = 0.48), yet gets every question right on only 8.7% of figures

Data: Guan et al., HallusionBench (CVPR 2024), Table 3 (GPT-4-judged); Sec. 4 (metric definitions)

Georgia Tech

Case Studies: Memory Over Pixels

Visual Supplement – Map

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?

No Visual

GPT-4V: Yes, among the states you listed—Arizona, Oklahoma, North Carolina, and Delaware, Delaware is the farthest north.

LLaVA-1.5: No, Delaware is not the farthest north among Arizona, Oklahoma, North Carolina, and Delaware. North Carolina is the farthest north among these four states. Delaware is located to the north of Maryland and to the east of Pennsylvania, while Arizona is located to the southwest of Delaware, Oklahoma is located to the west of Delaware, and North Carolina is located to the north of Delaware.

Prompt: Among Arizona, Oklahoma, North Carolina, and Delaware on the map, Is Delaware the farthest north?



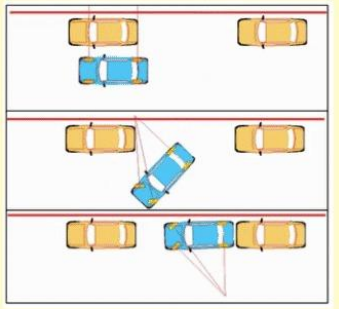
Delaware and Arizona are swapped in the following image.

GPT-4V: Yes, among Arizona, Oklahoma, North Carolina, and Delaware on the map, Delaware is the farthest north.

LLaVA-1.5: Yes, Delaware is the farthest north among the four states on the map.

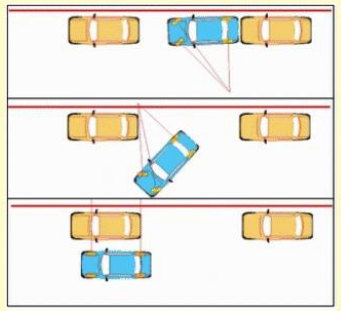
Figure from Guan et al., HallusionBench (CVPR 2024), Fig. 14

Prompt: According to the positive sequence of the images, does the blue car attempt to park?



Original sequence (From Top to Bottom)

GPT-4V: Yes, according to the sequence of the images, the blue car appears to be attempting to parallel park between the two yellow cars.



Reversed sequence (From Top to Bottom)

GPT-4V: Yes, according to the sequence of the images, the blue car appears to be attempting to parallel park between the two yellow cars.

Figure from Guan et al., HallusionBench (CVPR 2024), Fig. 8 (excerpt)

Map (VS): Delaware and Arizona swapped. Both models still name Delaware as farthest north: language hallucination

Sequence (VD): frames reversed. GPT-4V still says the car parks: no temporal reasoning

Georgia Tech

Strengths and Open Questions

What it gets right

A controlled experiment
Only the image changes, so a flipped answer traces back to the pixels

A diagnosis, not just a score
Splits prior override from misperception

Formats people actually ask about
Charts, tables, maps, OCR, math and video, not only objects

Hard to pass by guessing
Random scores 15.6% qAcc, not 50%

Low language hallucination is not a good sign
Qwen-VL (0.9%), mPLUG-Owl-v1 (4.0%) and LRV (4.5%) barely show it, because they misread the image before a prior can win. The cause mix tracks perception quality, so we expect the prior's share to grow as encoders improve.

Yes/no cannot reward the best answer
On the swapped map, the ideal reply is "On this map Arizona is farthest north, but in reality it is Delaware." A yes/no format can only score image vs. prior, never flagging the conflict.

Discussion:
When the image contradicts world knowledge, should a VLM always trust the image? Does that change for a medical scan vs. a doctored chart?

Our analysis; numbers from Guan et al., HallusionBench (CVPR 2024), Tables 2–3

HallusionBench Takeaways → SAGE

What HallusionBench showed (2023)

1 Hallucination has two separable causes: language priors and visual misperception

2 Edits expose it: even GPT-4V fails about 70% of control pairs; most models sit below chance

3 Models rarely check their answer against the image, so they ignore edits and contradict themselves

→

What it leaves open

A benchmark diagnoses says that and why a model fails, not when during generation grounding breaks

Frontier question (2026)

Can we detect a grounding failure while decoding and steer attention back to the image, with no retraining?

→ **SAGE: sink-aware grounded decoding**

Through-line: both papers ask whether the answer is grounded in the image. HallusionBench measures it from outside the model; SAGE checks it from inside, token by token.

Georgia Tech

Approach: SAGE Overview

- **Training-free decoding strategy** that reduces VLM hallucination, with no retraining or architecture changes
- **Key idea:** treat attention sink tokens as checkpoints to check for hallucination rather than artifacts to suppress
- **Motivation:** 88.4% of hallucination errors occur within 5 decoding steps after a sink token (MSCOCO + HalluBench)
- **Contribution:** At each sink token, SAGE evaluates how grounded the previous segment is and sharpens or broadens attention based on the result

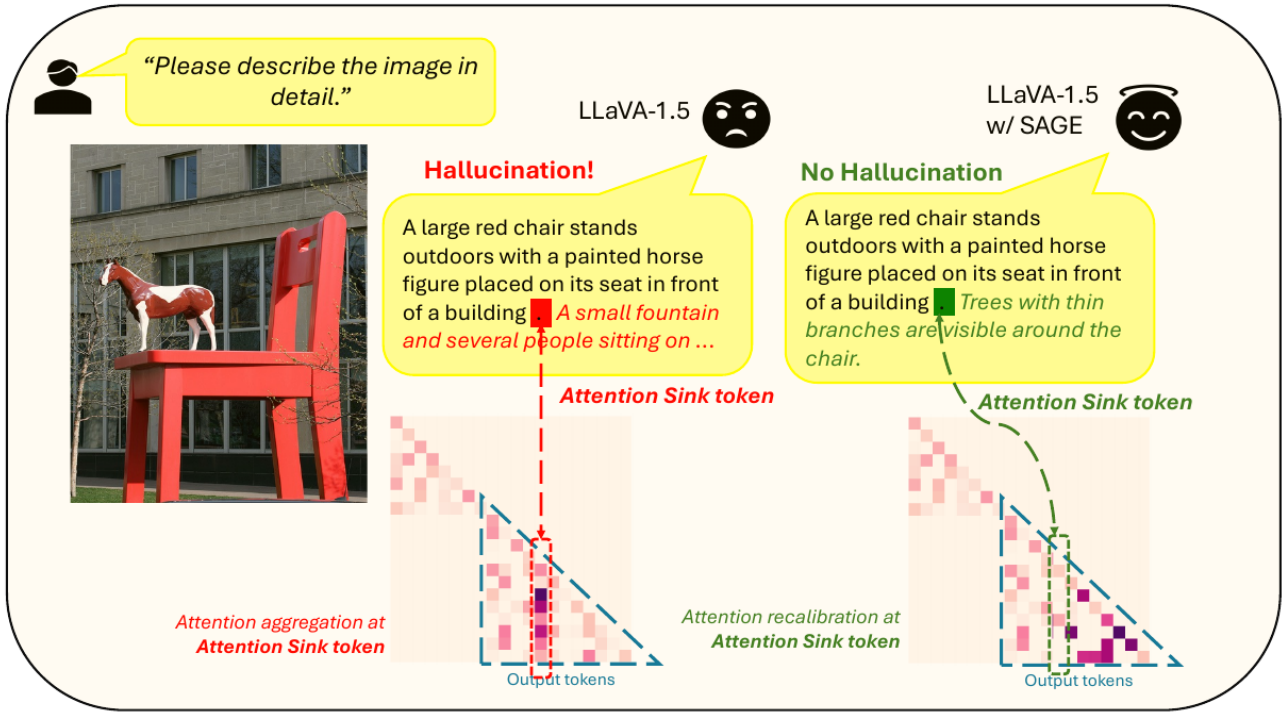


Figure from Shukla & Kira, SAGE (2026), Fig. 1

Georgia Tech

Background: Attention Sinks

Why sinks arise

- Sinks are tokens that receive disproportionately high attention despite little semantic content
- Softmax forces attention weights to sum to one; always-visible tokens absorb redundant attention
- In SAGE, punctuation and light conjunctions play this role

Why they matter for VLMs

- Text and image tokens compete in the same attention space
- Sinks leave less attention for image tokens, so visual evidence is underused
- Language model priors → hallucination!
- SAGE’s contribution is to use sink tokens as a marker to check grounding rather than simply penalize them

SAGE Pipeline

- **Stage 1: Detect and extract**
  - On a sink token, extract key concepts C from the segment since the previous sink
- **Stage 2: Measure grounding**
  - Compare overlap between attention map and GradCAM map using IoU
- **Stage 3: Modulate attention**
  - The overlap score is compared to a threshold, and if the overlap score is higher than the threshold, then attention is sharpened. Otherwise, attention is redistributed.

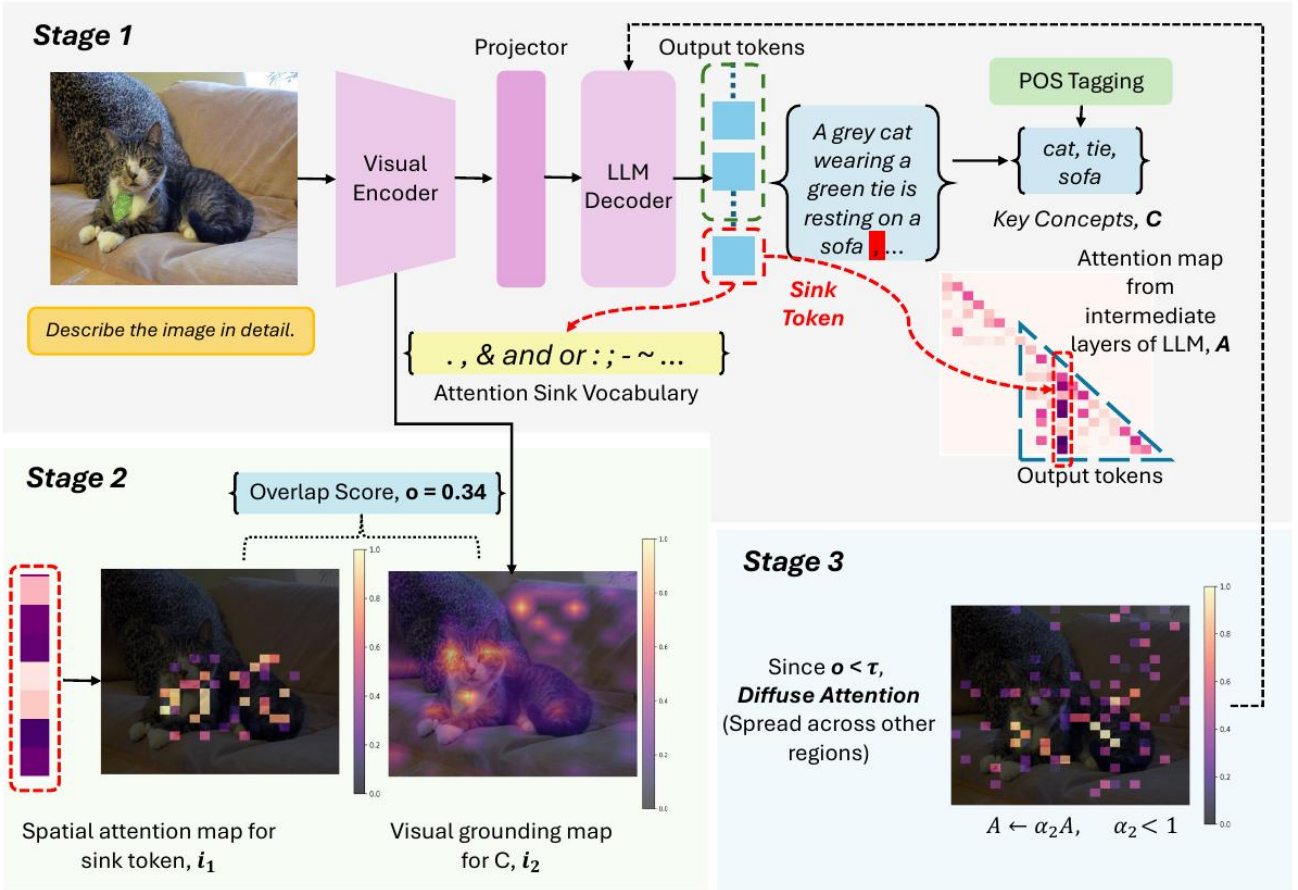


Figure from Shukla & Kira, SAGE (2026), Fig. 2

Georgia Tech

Stage 1: Detection & Concepts

Sink detection

- Monitor generated tokens continuously during autoregressive decoding
- A token is a sink if it is in a predefined set
  - Punctuation marks and simple conjunctions
- Each sink is a semantic boundary, meaning that the new tokens generated since the previous sink form a coherent descriptive segment

Key concept extraction

- Part-of-speech tagging on the new segment
- Keep tokens with high visual grounding potential
  - Nouns: objects and scene entities
  - Adjectives: color, texture, size
  - Compound noun phrases: "red sports car"
- Prepositions, auxiliary verbs, and other parts of speech that do not have a visual representation are thrown out
- Concepts are used to check grounding and guide future token generation

Georgia Tech

Stage 2: Attention Map (i_1)

- **Input:** the sink token's self-attention weights over image tokens only
- **Layers:** ablation study shown in figure on the right show that middle layers have the highest overlap with the GT boxes and lowest attention entropy
- **Head:** the head with the highest average attention mass on visual tokens at the sink position
- **Output:** project weights onto the image plane to get the attention map, the model's internal estimate of where it looks for this content

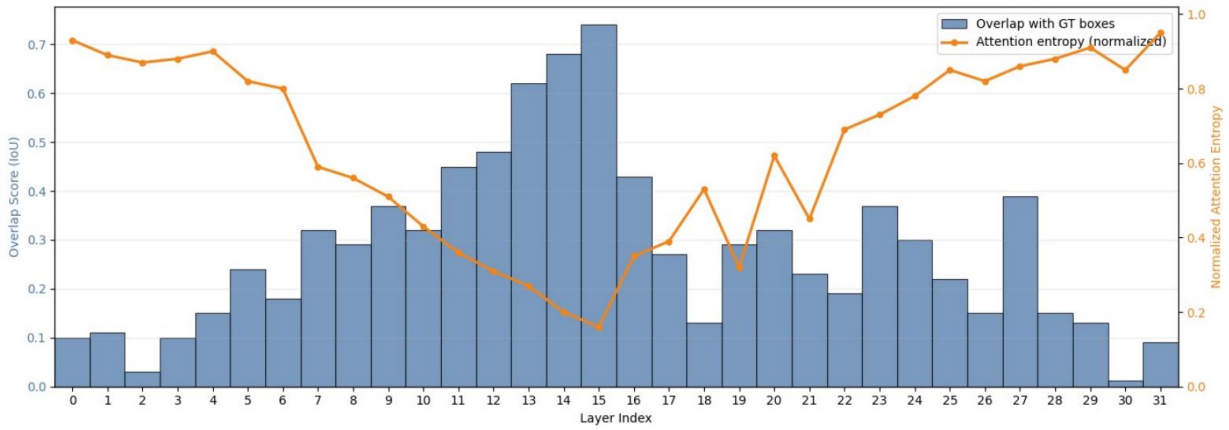


Figure from Shukla & Kira, SAGE (2026), Fig. 3a: layer-wise IoU with GT boxes and attention entropy (500 MSCOCO images)

Georgia Tech

Grad-CAM: How It Works

What it answers

- Which locations in a feature map pushed up the score y of a chosen target?

Algorithm

1. Forward pass; keep the K feature maps A_k of one layer
2. Backprop the target score y to that layer: $\partial y/\partial A_k$
3. Average each channel's gradient over space to get its weight α_k
4. Weighted sum of the maps, then ReLU, then upsample to image size

$$\alpha_k = (1/Z)\ \Sigma_{i,j}\ \partial y\ /\ \partial A_k(i,j)$$

channel weight = gradient averaged over space

$$L = \mathrm{ReLU}(\ \Sigma_k\ \alpha_k \cdot A_k\ )$$

weighted sum of maps, positive evidence only

Why it works, and its limits

- Gradients say which channels matter; activations say where they fire
- ReLU keeps only regions that raise the score
- Coarse: resolution of the chosen layer. Gradients can saturate or be noisy
- Shows sensitivity of y, not proof that a region caused it

Georgia Tech

Stage 2: Grad-CAM Map (i_2)

- Attention shows where the model looks, not what drives a prediction, so a gradient-based reference is added
- Uses the decoder's output logit for each concept in the extracted concepts and backprops to the visual features in the final layer of the vision encoder
- Grad-CAM gives a concept-specific map of the regions most influential for predicting the concept
- Use a threshold to turn each concept map into a binary mask and then union all of the maps using OR operator over all masks to get final spatial map

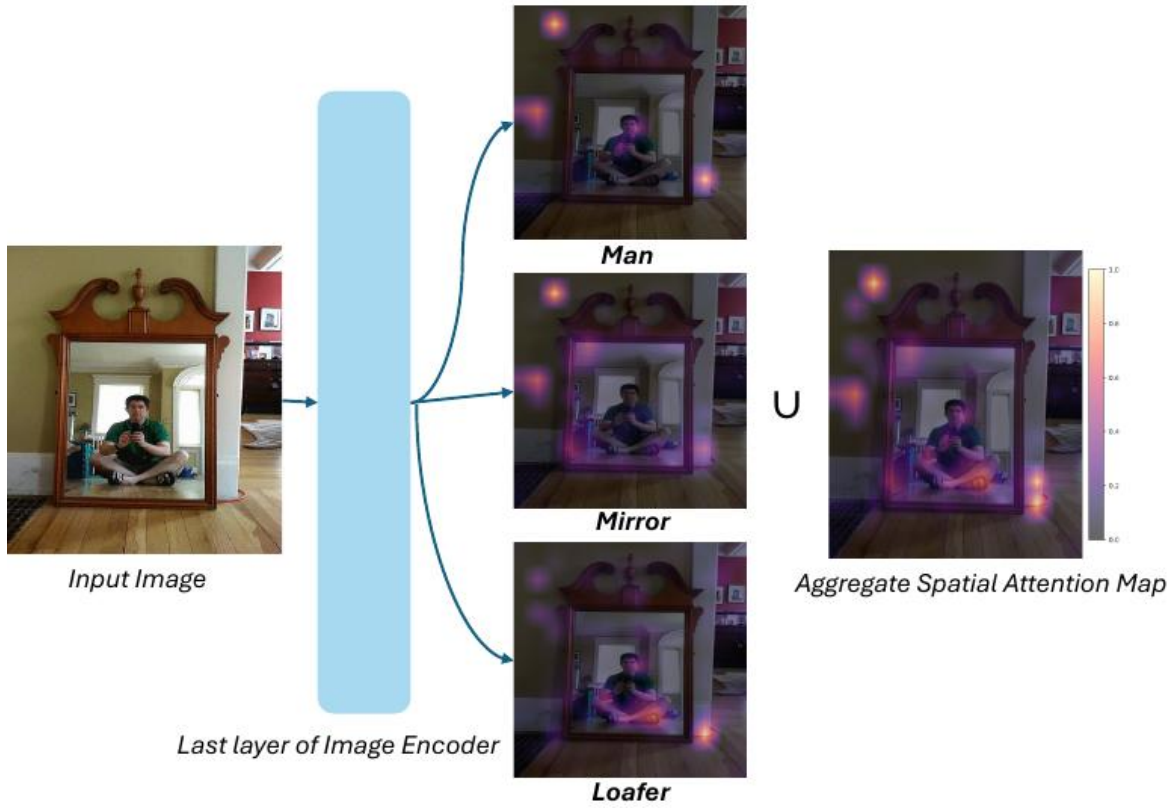
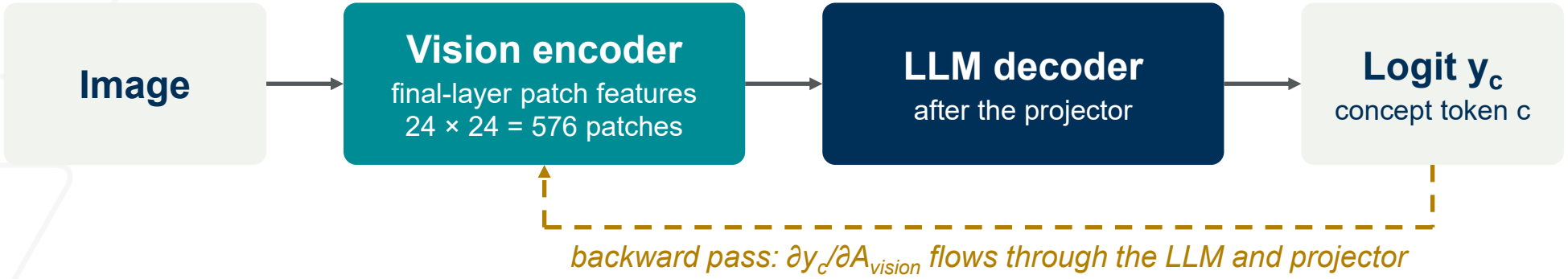


Figure from Shukla & Kira, SAGE (2026), Fig. 3b: per-concept Grad-CAM maps and their union

Georgia Tech

Grad-CAM in SAGE

Which image patches does a concept token's logit depend on?



1 Channel weights

$\alpha_{c,k}$ = average over patches of $\partial y_c/\partial A$ (hidden dimensions act as channels)

2 Heatmap

$M_c = \mathrm{ReLU}(\Sigma_k \alpha_{c,k} A_k)$, reshaped to 24 × 24 and upsampled

3 Mask and union

Threshold each M_c, then OR across all concepts in C to get i_2

4 Overlap

$o = \mathrm{IoU}(i_1, i_2)$, compared with threshold = 0.5 to choose reinforce or diffuse

Supervision comes from the language model, not the vision encoder!

Georgia Tech

Stage 3: Reliability & Modulation

Grounding reliability

- The overlap score between the attention map and Grad-CAM map asks whether the regions the model looks at match the regions that drive the output logits for a concept
  - High overlap means the attention map and spatial maps are correlated, and thus attention is sharpened
  - Low overlap implies misalignment, with attention either diffuse or on irrelevant regions, both of which lead to hallucinations
    - Attention is redistributed across the image tokens rather than being concentrated at the sink

Modulation rule (threshold = 0.5)

- **Overlap > threshold** → sharpen attention
  - $A \leftarrow \alpha_1 A, \alpha_1 > 1$
- **Overlap < threshold** → broaden attention
  - $A \leftarrow \alpha_2 A, \alpha_2 < 1$
- Applied only at sink tokens, so intervention lands on concept transitions where hallucinations are most likely
- Ablations: both modes are needed, and sink-triggered modulation beats a fixed 10-step schedule

Georgia Tech

Attention Modulation

- The attention modulation occurs on the raw attention scores before softmax
- The attention redistribution is similar to temperature by the way in which it affects the probability distribution
  - Sharpening inflates the differences between the logits, so softmax blows up those differences, making the model focus its attention on areas of highest importance
  - Diffusing makes the differences between the logits smaller, so the probability distribution is smoothed, and the attention broadens
- Sharpening attention → lower entropy and diffusing attention → higher entropy

Georgia Tech

SAGE in Action

- Sink token is highlighted in red; SAGE compares its attention map to the Grad-CAM map
- Top (o = 0.59 > threshold): reinforce
  - Attention is kept on grounded regions; the output stays faithful instead of a hallucinated watering can and gloves
- Bottom (o = 0.24 < threshold): diffuse
  - Attention is broadened; the hallucinated "colorful balloons" is replaced by a different continuation
- Right column shows the modulated attention maps

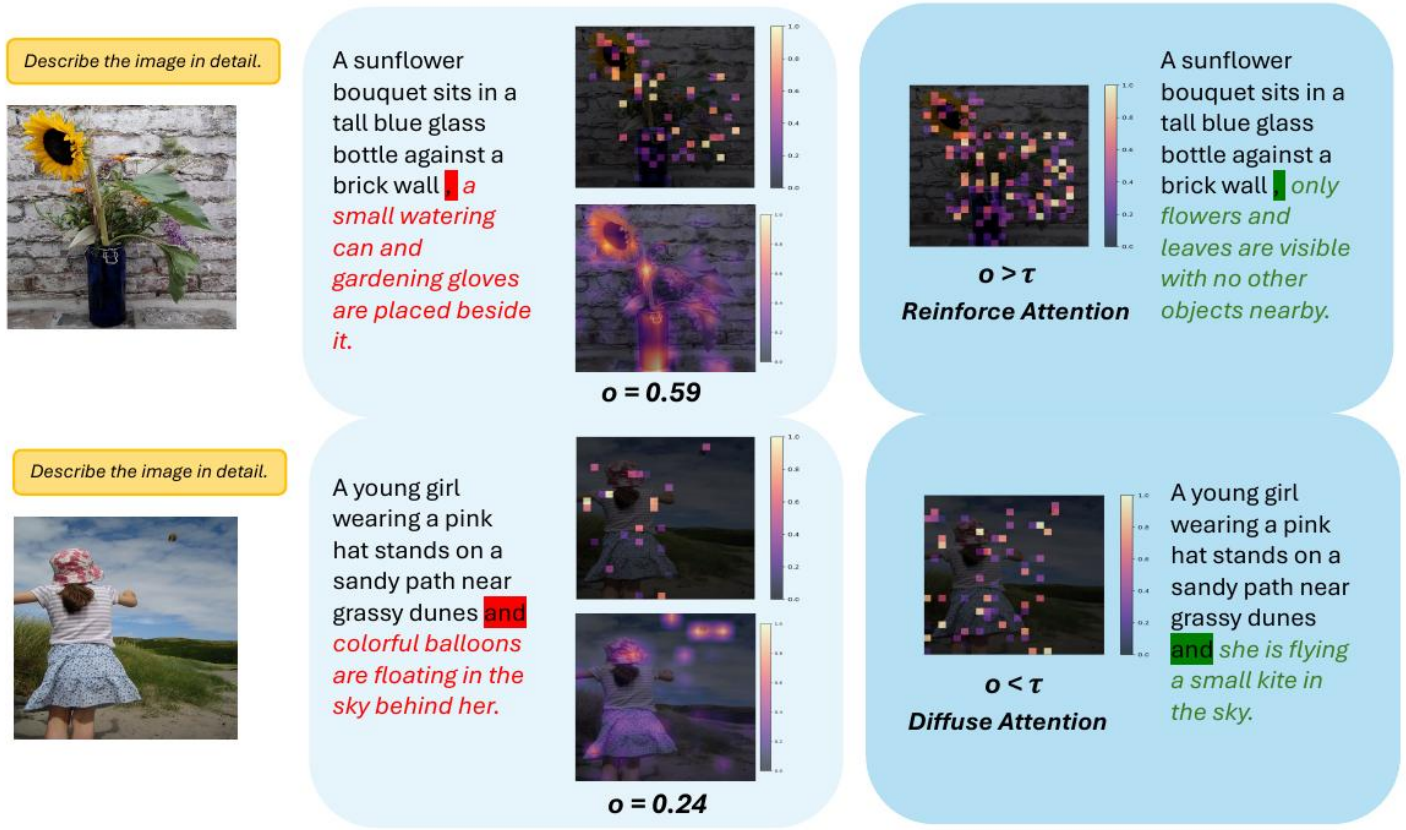


Figure from Shukla & Kira, SAGE (2026), Fig. 4

Georgia Tech

Discussion: Attention Maps vs. Grad-CAM Maps

	Attention map	Grad-CAM map
Question	Where does the sink token read from?	Where would changing the image most change this concept's logit?
Where	Middle decoder layers, one head	Final vision-encoder layer, with gradients from the decoder logit
Scope	One token: the sink standing in for the previous segment	One map per concept token, then a union (OR) across concepts
Can mislead when	Attention reads context from a neighboring object that does not drive the prediction; attention weights are not an explanation of the output	Gradients are noisy or saturate; ViT patches already mix global context, so maps are coarse; shows sensitivity, not proof of presence

- The two maps answer different questions, so low overlap is informative: attention may be using context that the concept logit does not depend on
- High overlap means two views of one forward pass agree: supporting evidence, not ground truth

Georgia Tech

Discussion: Scope, Assumptions, Failure Modes

Scope

- Object hallucination in long free-form captions (CHAIR, AMBER)
- Needs sink tokens, so short yes/no or VQA answers are out of scope
- Targets attention drifting from the image so the language prior takes over

Assumptions

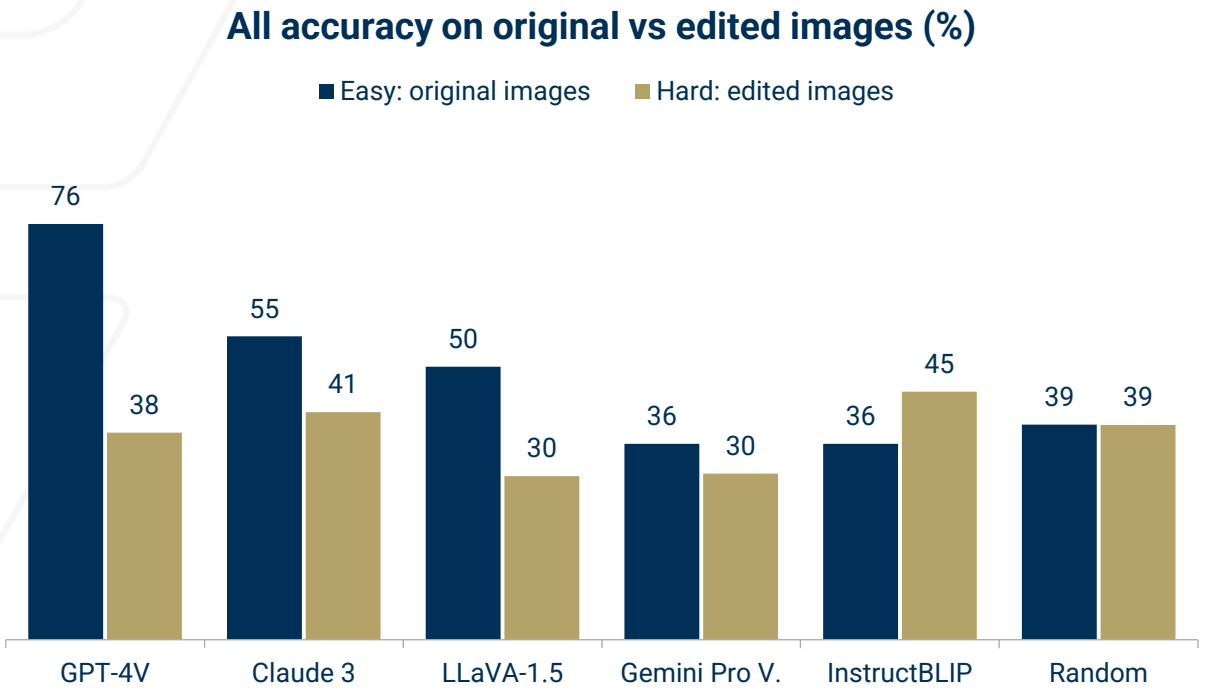
- Attention-attribution agreement means grounded: looking at the right place equals perceiving and reasoning correctly
- Concepts can be extracted from partial text with POS tags
- Grad-CAM is an independent reference, yet both signals come from the same model and forward pass

Failure Modes

- **Agree but wrong:** the logit is driven by a language prior or co-occurring context. Grad-CAM still returns a map, attention lands on the same region, o is high, and SAGE reinforces the hallucination
- **Disagree but right:** large or diffuse objects, or many concepts inflating the union, give low IoU for correct text and trigger needless broadening

Georgia Tech

HallusionBench: Original vs. Edited Images



Edits break almost every model. Accuracy falls from the original to the edited image, where answering from memory stops working.

Claude 3 drops less (55.2 → 41.4) and ends above GPT-4V on edited images.

InstructBLIP is the exception. It scores higher on edited images (45.1 vs 35.6), likely because it leans toward "no" (Pct. Diff −0.13).



***Question**:*
In 2017, was Tencent the company with the highest revenue from video games, with Sony as the second-highest earner?
In 2017, did Apple generate higher revenue from video games compared to Google?

HllusionBench Fig. 1: the edited chart (red) swaps Tencent and Google

Limitations: What HallusionBench Assumes

Recall (slides 6–8): control pairs + the strict qAcc metric are its strength. Below: what that design takes for granted.

What it assumes

- **Cause is read from outputs only:** an unchanged answer may mean the edit went unnoticed, which is a perception failure, not a prior. Nothing inside the model is inspected
- **Edits are unusual images:** pasted objects, recolored maps and reversed frames look odd; some failures may come from that rather than from strong priors
- **The judge is text-only:** GPT-4 grades answers without seeing the image

What it leaves out

- **Long answers:** captions, explanations, multi-turn chat
- **Reasoning chains:** errors that appear midway through a long answer
- **Fresh data:** 346 images, public since 2023, so newer models may have trained on them

Our take: HallusionBench measures whether the answer depends on the image, not whether the model perceived it correctly. With yes/no answers, a model can flip for the wrong reason and still be scored correct.

Based on Guan et al., HallusionBench (CVPR 2024), Sec. 3 (dataset design) and Sec. 4 (evaluation metrics); our analysis

SAGE: How the Evaluation Was Run

5 models, 3 connector designs

- **Linear projector**: LLaVA-1.5-7B, Shikra
- **Q-Former**: MiniGPT-4, InstructBLIP
- **Native integration**: Qwen2-VL-7B

6 decoding baselines

- OPERA, HALC, EAH, DML-LVLM, GIFT, LVLMs-Saliency

Protocol

- Prompt: "Please describe this image in detail."
- Max length 256 and 512 tokens, scores averaged
- α_1, α_2 and $\tau = 0.5$ tuned on 200-sample validation sets

CHAIR on MSCOCO

500 val2014 images. CHAIRs: % of captions with a hallucinated object. CHAIRi: % of mentioned objects not in the image. Lower is better.

AMBER

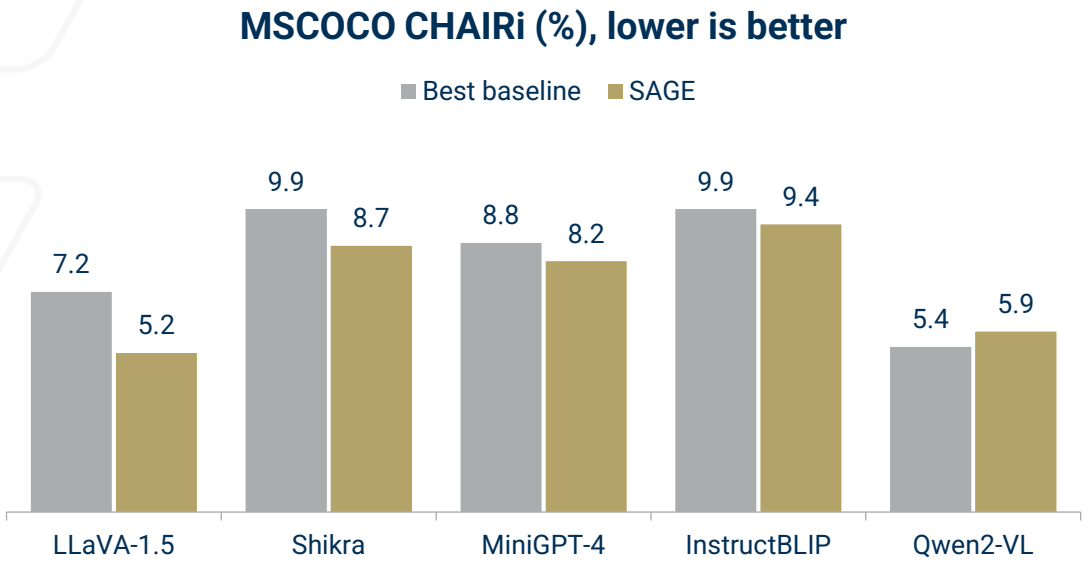
500 of 1,004 images. CHAIRi plus Cover: % of true objects the caption mentions. Higher Cover is better.

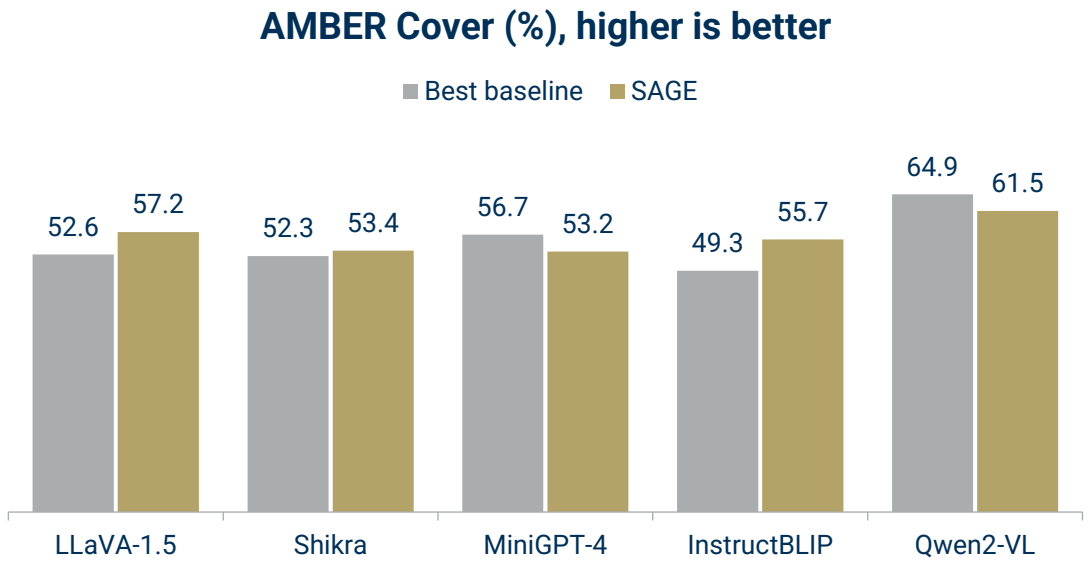
Gemini 3 Pro as judge

100 COCO images. Scores accuracy and detail 0–10 for SAGE, GIFT and LVLMs-Saliency side by side.

Georgia Tech

SAGE: Main Results





Wins in 15 of 20 model × metric cells

Exceptions: MiniGPT-4 (CHAIRs, Cover) and Qwen2-VL (CHAIRs, CHAIRi, Cover), where DML-LVLM or LVLMs-Saliency is better.

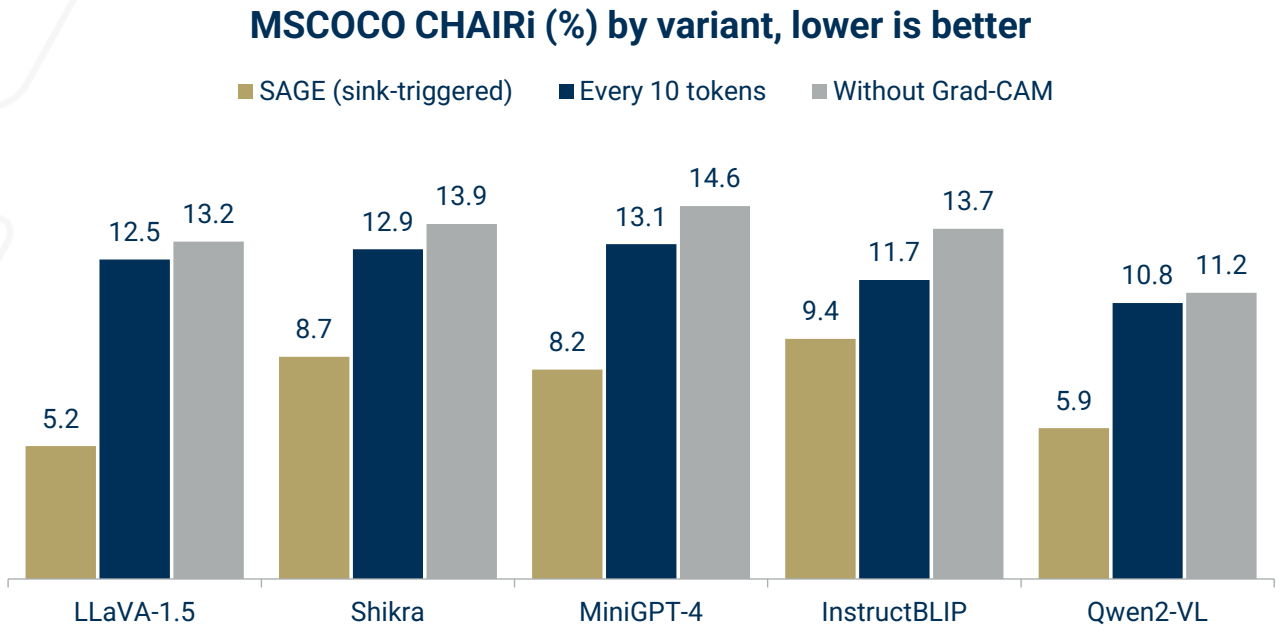
Gemini 3 Pro judge

SAGE scores highest on accuracy and detail on both models tested (LLaVA-1.5 accuracy 6.9 vs 5.6).

Data: SAGE (2026), Table 1 (CHAIR on MSCOCO), Table 2 (CHAIRi and Cover on AMBER), Table 3 (Gemini judge). Best baseline = strongest of the 6 methods per model and metric

Georgia Tech

SAGE: Ablations and Cost



Timing matters: checking at sink tokens beats a fixed 10-token schedule on every model.
Grad-CAM matters: without it, SAGE trails every baseline on LLaVA-1.5, MiniGPT-4 and Qwen2-VL.

+10.8%

extra decoding time (A100, batch 1)

Grad-CAM attribution	**7.6%**
Attention map + IoU	**1.7%**
Concept extraction (POS)	**1.2%**
Attention modulation	**0.3%**

Cost scales with the number of sink tokens, not sequence length

Data: SAGE (2026), Table 4 (with vs. without Grad-CAM), Table 6 (every 10 tokens vs. at sink tokens), Table 7 (runtime per stage)

Georgia Tech

Discussion: Does Agreement Mean Grounding?

	Maps agree IoU > 0.5 → reinforce	Maps disagree IoU < 0.5 → diffuse
Model perceives correctly	**Intended case** Reinforce keeps a good focus	**Side effect** Diffuse may pull attention off a good region
Model perceives wrongly	**Blind spot** Both maps share the model's prior, so they can agree on the wrong region, and SAGE reinforces the error	**Intended case** Diffuse lets the model look elsewhere

Why agreement ≠ correct perception

- **Attention is a weak explanation:** attention weights often fail to explain predictions (Jain & Wallace, 2019); saliency maps can survive randomized weights (Adebayo et al., 2018)
- **Much visual attention is noise:** LMMs park attention on "visual sink" tokens that can be removed without changing the answer (Kang et al., ICLR 2025)
- **Where ≠ what:** HallusionBench's illusion and chart failures happen while looking at the right region; SAGE only fixes where the model looks
- **Too late for the current segment:** SAGE checks text already written and can only change what comes next
- **The authors agree:** SAGE "may fail" when both signals are misaligned (their Limitations section, App. 8.9)

Shukla & Kira, SAGE (2026); Jain & Wallace (NAACL 2019); Adebayo et al. (NeurIPS 2018); Kang et al. (ICLR 2025); our analysis

Bridging the Evaluations (+ PAPO)

	HallusionBench (2024)	SAGE (2026)	PAPO (last class, ICLR 2026)
Goal	Diagnose why models hallucinate	Reduce hallucination at decoding time	Make RL training use the image
Output format	Short yes/no answers	Long free-form captions	Short answer after a reasoning chain
What counts as hallucination	Any answer that conflicts with the image: charts, maps, OCR, math, video	Objects named in a caption that are not in the image	Perception errors in math, geometry, logic and counting (67% of errors)
Ground truth	Human-written questions on human-edited images	COCO / AMBER object annotation lists	Verifiable answers from existing benchmarks
Who grades	Text-only GPT-4, checked against humans	CHAIR string matching, plus Gemini 3 Pro	Rule-based answer checker; no judge model
Models	15, including closed APIs (GPT-4V, Claude 3)	5 open 7B-scale models; needs attention and gradients	Qwen2.5-VL 3B and 7B; needs training

The Mitigation Landscape: Where SAGE and PAPO Fit

Family	What it changes	Representative work	Fixes well	Does not fix
Data	Instruction data	LRV-Instruction, HalluciDoctor, ShareGPT4V	Yes-bias, invented objects	Fine perception
Architecture	Vision encoder, resolution	MMVP / Mixture of Features, Cambrian-1, LLaVA-NeXT tiling	Misreading small details	Prior override
Training	Objective (RLHF, DPO, RL)	LLaVA-RLHF, RLHF-V, HA-DPO, POVID, mDPO, PAPO	Grounding across tasks	Reward hacking; needs training
Decoding	Token choice at inference	VCD, OPERA, HALC, PAI, VAR, SAGE	Priors in long captions	Short answers; closed models
Post-hoc	The output text	Woodpecker, LURE	Checkable object claims	Reasoning, illusions
Tools	The process: look again	V* search, DeepEyes, o3 zoom-in	Small details in large images	Strong priors

Leng et al. (CVPR 2024); Huang et al. (CVPR 2024); Tong et al. (CVPR 2024; NeurIPS 2024); Yu et al. (CVPR 2024; 2025); Wang et al. (EMNLP 2024); Yin et al. (2023); Zhou et al. (ICLR 2024)

Georgia Tech

Link to Last Class: One Question, Three Answers

HallusionBench
Measure · benchmark

Edit the image. Does the answer flip the right way?

Signal: Human-made counterfactual pairs

Weak spot: Small, manual, yes/no only

SAGE
Steer · decoding

Do attention and Grad-CAM point to the same region?

Signal: The model's own internal maps

Weak spot: Needs open weights; captions only

PAPO
Train · RL objective

Does the answer become less likely when 60% of the image is masked?

Signal: KL between real-image and masked-image policies

Weak spot: Can be "hacked": collapse without an entropy loss

DeepSeek-R1: the reward only checks the answer

For a VLM, a correct yes/no answer cannot tell looking from guessing. Outcome-only RL can reinforce priors. PAPO finds 67% perception errors after GRPO; HallusionBench finds 40–88% visual illusion.

PAPO = the control pair as a training signal

Random masking replaces human edits. All three measure whether the output depends on the image; none checks that the model saw it correctly.

DeepSeek-R1 (2025), Sec. 2 (RL and rewards), Sec. 6 (limitations); PAPO (ICLR 2026), Sec. 2.2 (error analysis), 3.2 (method), 5.3 (KL hacking); Guan et al. (2024); Shukla & Kira (2026)

Georgia Tech

Frontier Models: Has the Problem Gone Away?

100% → 17%

Accuracy of o3, o4-mini, Gemini 2.5 Pro, Claude 3.7 Sonnet and GPT-4.1 on familiar vs. edited images (VLMBias)

76%

of their answers repeat the memorized answer, e.g. "3 stripes" for an edited 4-stripe logo

8 of 8

R1-style multimodal reasoning models hallucinate more than their base model (Liu et al., 2025)

What this means

- **Language hallucination persists:** HallusionBench's control-pair recipe still breaks 2025 frontier models, and asking them to double-check barely helps
- **More thinking is not more seeing:** longer reasoning chains attend less to the image, a direct echo of DeepSeek-R1-style training
- **The metric drifted:** reports now give one blended "HallBench avg" (GPT-4o 55.0, Claude 3.5 Sonnet 55.5, Qwen2.5-VL-72B 55.2), mixing the lenient and strict metrics, on a public 2023 test set
- **SAGE cannot reach closed models:** it needs attention and gradients, so GPT or Gemini can only be fixed through training (PAPO-style) or tools like zoom-in

Digging deeper into the same studies

- **Context triggers the prior:** removing the image background lifts accuracy by ~21 points (VLMBias)
- **Thinking has a ceiling:** accuracy peaks near 40%, then overthinking hurts (VLMBias)
- **Training recipe matters:** RL-only and larger models keep perception better (Liu et al.)

Vo et al., Vision Language Models are Biased (ICLR 2026); Liu et al., More Thinking, Less Seeing? (NeurIPS 2025); Qwen2.5-VL Technical Report (2025)

Testing SAGE on HallusionBench

Setup

1. **Make sink tokens appear**
   Prompt: "Describe the relevant part of the image, then answer yes or no"
2. **Use SAGE's own models**
   LLaVA-1.5-7B and Qwen2-VL-7B
3. **Compare four conditions**
   Short answer · reasoning first · reasoning + SAGE · reasoning + VCD
4. **Score with HallusionBench**
   qAcc, Hard accuracy, language-hallucination vs. visual-illusion split, yes/no bias

Our predictions

- **Reasoning alone may hurt**
  Longer outputs attend less to the image (Liu et al., 2025)
- **SAGE helps on charts, tables, maps**
  The answer sits in one region the model may miss
- **No gain on illusions and geometry**
  The model looks at the right place but misreads it
- **Prior-driven errors barely move**
  If the prior decides before the first sink token appears

What persisted and what changed?

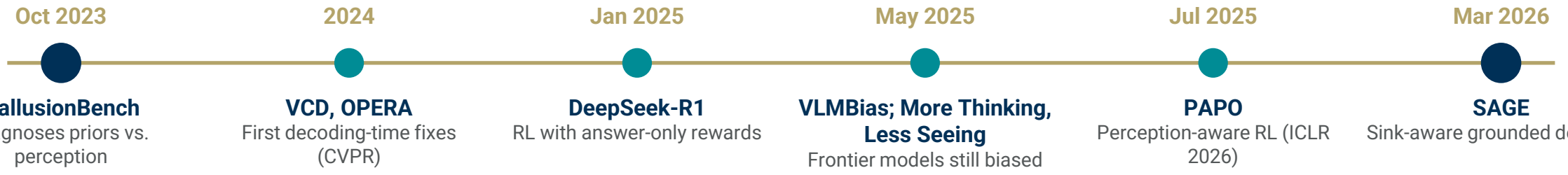
What persisted

- LLaVA-1.5 and InstructBLIP are still the standard test models
- Evaluation still relies on a strong model as judge (GPT-4, then Gemini 3 Pro)
- Not solved: with SAGE, LLaVA-1.5 still hallucinates in about 1 of 5 captions (CHAIRs 21.3)

What changed

- **From diagnosing to fixing**: a benchmark became a training-free decoding method
- **From black box to white box**: API answers gave way to reading attention and gradients
- **Narrower scope**: broad image-context reasoning became object hallucination in captions

How the field moved, 2023 → 2026



Georgia Tech

Recent Methods: Where the Field Is Heading

Method	Family	Key idea	Link
VAR (Kang et al., ICLR 2025)	Decoding	Gives attention wasted on visual sink tokens back to informative image tokens	Shows much visual attention is noise
EAH (Zhang et al., 2024)	Decoding	Finds vision sink tokens; strengthens image-focused attention heads in shallow layers	SAGE baseline
DML-LVLM (Jiang et al., 2025)	Decoding	Redistributes visual attention across heads in middle layers	SAGE baseline; same middle layers
GIFT (Qi et al., 2025)	Decoding	Builds a saliency map from simulated gaze shifts and amplifies salient regions	SAGE baseline
LVLMs-Saliency (Zhang et al., ICLR 2026)	Decoding	Filters candidate words when saliency on the image drops	SAGE baseline; beats SAGE on Qwen2-VL
PAPO, Visionary-R1, NoisyRollout (2025–26)	Training (RL)	Rewards or losses that make the model rely on the image while it learns to reason	PAPO
DeepEyes, Pixel Reasoner (2025)	Tools	RL teaches the model to zoom or crop during reasoning ("thinking with images", as in o3)	Same idea as o3; can reach closed models

Kang et al. (ICLR 2025); Zhang et al. (arXiv 2411.09968); Jiang et al. (arXiv 2411.16724); Qi et al. (arXiv 2510.22067); Zhang et al. (ICLR 2026); PAPO (ICLR 2026), Sec. 6; SAGE (2026), Sec. 4

Georgia Tech

Limitations and Implications

Limitations

HallusionBench

- Small (346 images), hand-made and public, so hard to refresh
- Failure cause inferred from outputs; text-only judge; yes/no only
- Some headline claims don't match its own tables (slide 23)

SAGE

- Object hallucination in captions only, on five open ~7B models
- Needs gradients (no closed models) and long outputs (no HallusionBench)

Implications

Hallucination is several problems
Prior override, misperception, invented objects and reasoning drift need different fixes

Check grounding against ground truth
Two internal signals agreeing is not evidence of correct perception

Benchmarks must keep moving
Static public test sets age as models train on the web

Fixes are moving into training
PAPO-style objectives and tool use, which still need perception-level benchmarks

Guan et al. (CVPR 2024); Shukla & Kira (2026); our analysis

Toward Better Hallucination Benchmarks

Keep the strict metric
Report qAcc and pair consistency, not blended averages

Generate, don't hand-edit
Procedural counterfactuals give fresh test sets at scale

Control for odd images
Add edits that do not contradict world knowledge

Score the evidence
Ask where the model looked; grade answer and evidence separately

Grade the reasoning
Find the first step that contradicts the image

Let judges see the image
And give credit for "I can't tell" on ambiguous images

Idea: the control pair as an RL reward
Reward the model only when it is right on both the original and the edited image (qAcc as the reward). A prior-driven model gets at most one of the two right, so it cannot collect the reward by guessing. This combines HallusionBench's test with DeepSeek-R1-style verifiable rewards.

Our proposals; VLMBias (Vo et al., 2025) for procedural counterfactuals; DeepSeek-R1 (2025) for verifiable rewards

Discussion Questions

1. **SAGE trusts the model when attention and Grad-CAM agree.**
   What would count as real evidence of correct perception: a detector, ground-truth boxes, or masking the attended region?

2. **HallusionBench infers "prior" vs. "misperception" from answers alone.**
   Could PAPO's masked-image test serve as a better diagnostic?

3. **Frontier models score 100% on familiar images and ~17% on edited ones.**
   Is that a perception failure or a prior problem, and will scale fix it?