HalluSegBench: Counterfactual Visual Reasoning for Segmentation Hallucination Evaluation

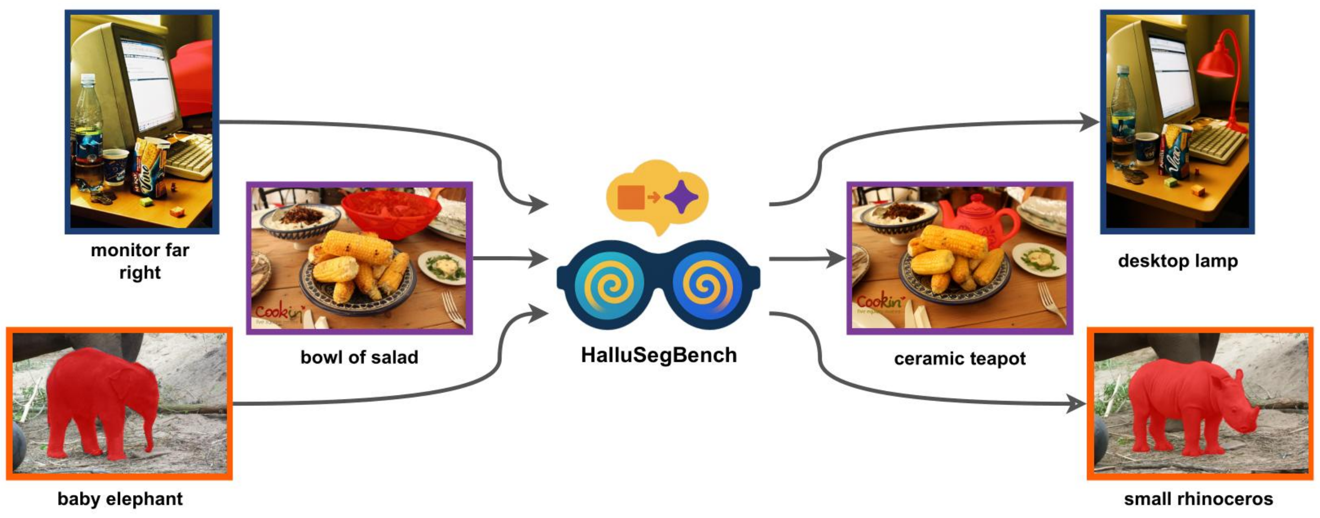
Xinzhuo Li*, Adheesh Sunil Juvekar*, Xingyou Liu, Muntasir Wahed,
Kiet A. Nguyen, Ismini Lourentzou

University of Illinois Urbana-Champaign

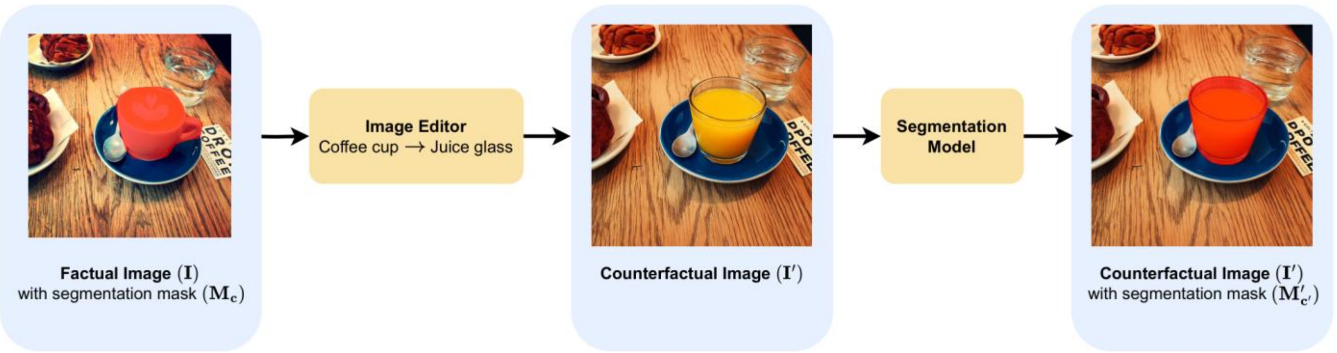
PERCEPTION & LANGUAGE
plan-lab.github.io
ILLINOIS

SaFeMM-AI
ICCV 2025

- ***New Segmentation Hallucination Benchmark.*** We introduce **HalluSegBench**, the first benchmark for pixel-grounding hallucination evaluation via **counterfactual interventions**.
- ***High-quality Dataset and Metrics.*** Our benchmark consists of a **dataset** of 1340 counterfactual pairs spanning 281 unique object classes, and 4 new **evaluation metrics** that quantify hallucination sensitivity under visually coherent scene edits.
- ***Experiments on SOTA Models.*** We conduct extensive experiments on state-of-the-art VLM-based segmentation models, highlighting the need for **counterfactual reasoning** to diagnose grounding fidelity.

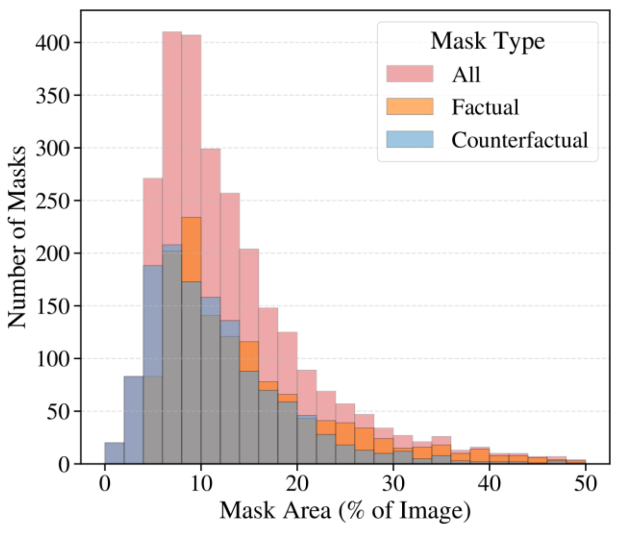


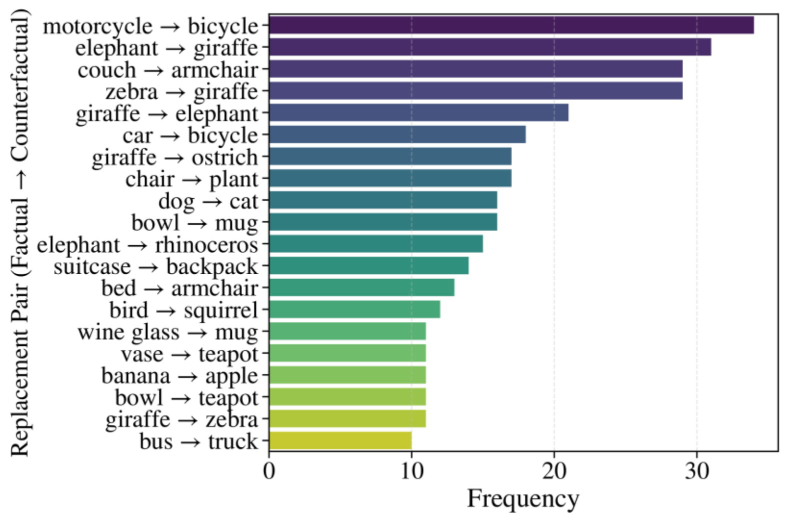
Data Generation

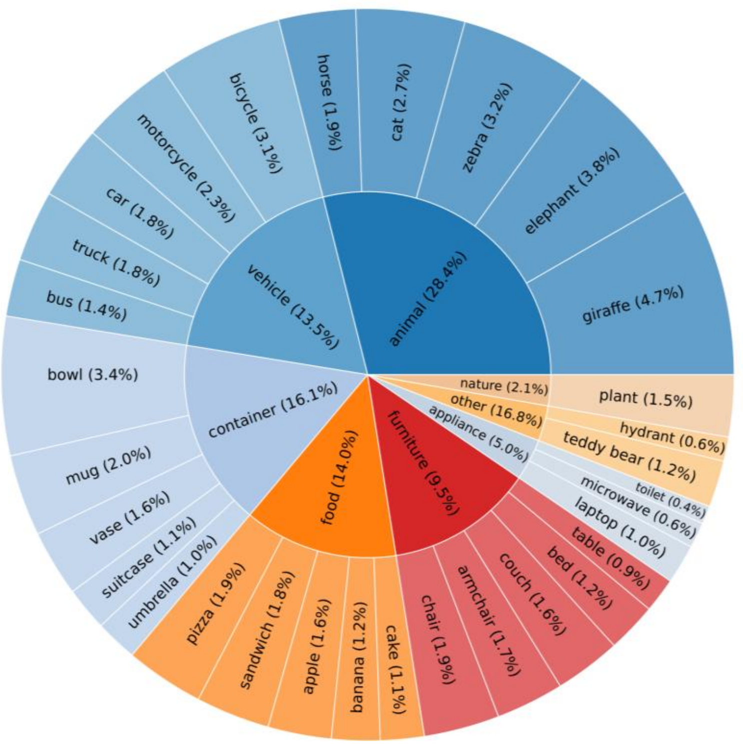

Factual Image (I) with segmentation mask (M_c)
Image Editor
Coffee cup → Juice glass
Counterfactual Image (I′)
Segmentation Model
Counterfactual Image (I′) with segmentation mask ($M'_{c'}$)

Dataset Creation Pipeline:
A factual image **I** is paired with a concise edit instruction to produce a counterfactual version **I'** where a target object is replaced with a visually similar alternative. Ground truth masks are obtained for both to support segmentation evaluation.

Dataset Insights







Top Left: Distribution of mask area percentage.

Top Right: Top object replacement pairs.

Left: Distribution of object categories

Evaluation Metrics

Consistency-based Performance Metrics

$$\mathrm{IoU}_{\mathrm{fact}} = \frac{|\mathbf{M}_c \cap \hat{\mathbf{M}}_c|}{|\mathbf{M}_c \cup \hat{\mathbf{M}}_c|}, \quad \mathrm{IoU}_{\mathrm{textual}} = \frac{|\mathbf{M}_c \cap \hat{\mathbf{M}}_{c'}|}{|\mathbf{M}_c \cup \hat{\mathbf{M}}_{c'}|}, \quad \mathrm{IoU}_{\mathrm{visual}} = \frac{|\mathbf{M}'_{c'} \cap \hat{\mathbf{M}}'_c|}{|\mathbf{M}'_{c'} \cup \hat{\mathbf{M}}'_c|}$$

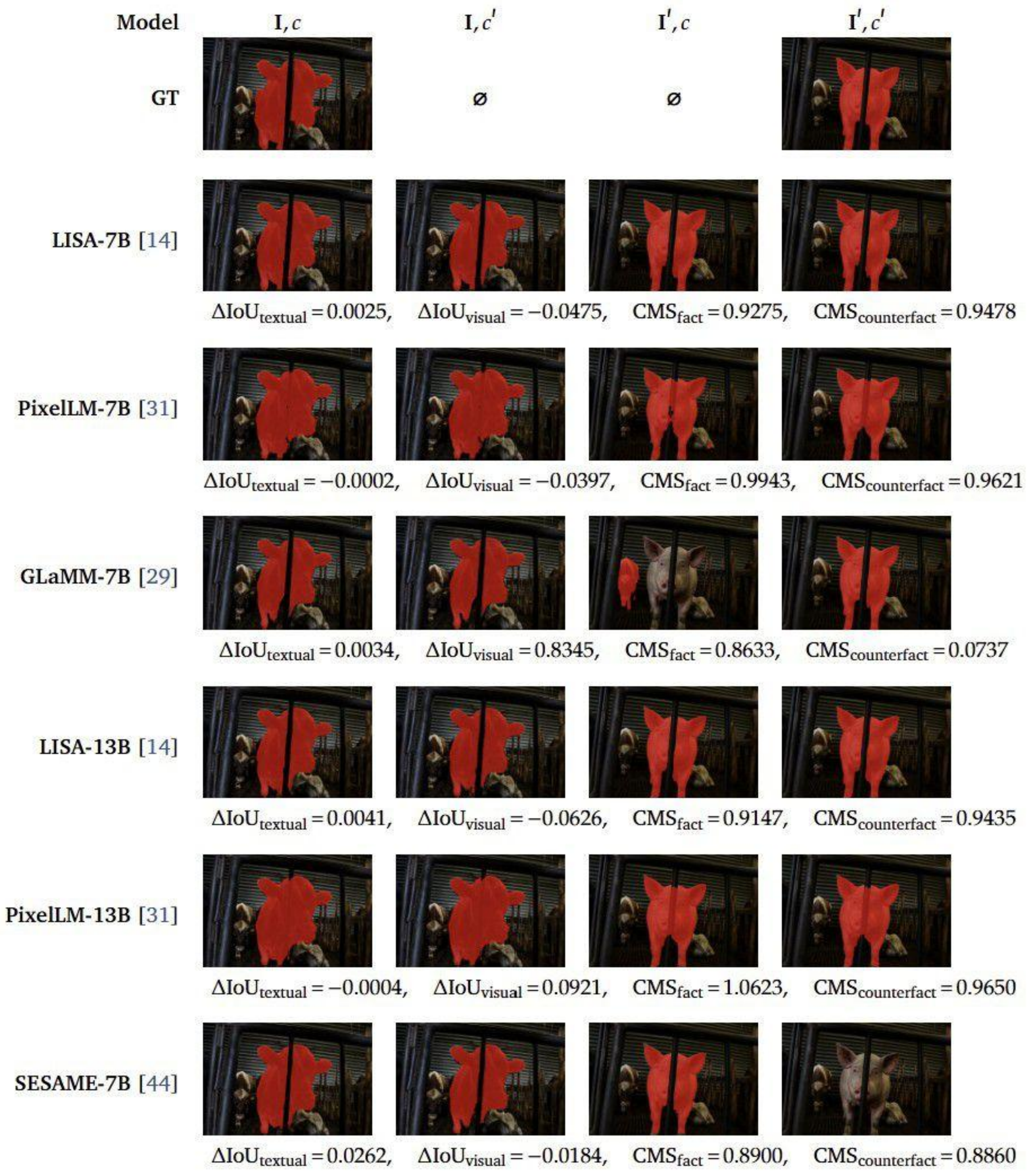
$$\Delta\mathrm{IoU}_{\mathrm{textual}} = \mathrm{IoU}_{\mathrm{fact}} - \mathrm{IoU}_{\mathrm{textual}} \qquad \Delta\mathrm{IoU}_{\mathrm{visual}} = \mathrm{IoU}_{\mathrm{fact}} - \mathrm{IoU}_{\mathrm{visual}}$$

Direct Hallucination Metrics

$$\mathbf{C} = \hat{\mathbf{M}}_{c'} \cap \mathbf{M}_c \qquad \mathbf{N} = \hat{\mathbf{M}}_{c'} \setminus \mathbf{M}_c$$

$$\mathrm{CMS} = \frac{\alpha\,|\mathbf{C}| + |\mathbf{N}|}{\alpha\,|\mathbf{M}_c|} \qquad \mathrm{CCMS} = \frac{\mathrm{CMS}_{\mathrm{fact}}}{\mathrm{CMS}_{\mathrm{counterfact}}}$$

Results


Model
I, c
I, c′
I′, c
I′, c′
GT
∅
∅
LISA-7B [14]
ΔIoU_textual = 0.0025, ΔIoU_visual = −0.0475, CMS_fact = 0.9275, CMS_counterfact = 0.9478
PixelLM-7B [31]
ΔIoU_textual = −0.0002, ΔIoU_visual = −0.0397, CMS_fact = 0.9943, CMS_counterfact = 0.9621
GLaMM-7B [29]
ΔIoU_textual = 0.0034, ΔIoU_visual = 0.8345, CMS_fact = 0.8633, CMS_counterfact = 0.0737
LISA-13B [14]
ΔIoU_textual = 0.0041, ΔIoU_visual = −0.0626, CMS_fact = 0.9147, CMS_counterfact = 0.9435
PixelLM-13B [31]
ΔIoU_textual = −0.0004, ΔIoU_visual = 0.0921, CMS_fact = 1.0623, CMS_counterfact = 0.9650
SESAME-7B [44]
ΔIoU_textual = 0.0262, ΔIoU_visual = −0.0184, CMS_fact = 0.8900, CMS_counterfact = 0.8860

Qualitative Comparison - Reasoning Segmentation Predictions
Here, c = "front cow" and c′ = "front pig"

Model	HM	$\Delta\mathrm{IoU}_{\mathrm{textual}}$ ↑	$\Delta\mathrm{IoU}_{\mathrm{visual}}$ ↑	$\mathrm{CMS}_{\mathrm{fact}}$ ↓	$\mathrm{CMS}_{\mathrm{counterfact}}$ ↓	CCMS
LISA-7B [14]	✗ No	0.4534	0.2810	0.3080	0.7317	0.4209
PixelLM-7B [31]	✗ No	0.3952	0.4071	0.4748	0.7286	0.6517
GLaMM-7B [29]	✗ No	0.3273	0.3016	0.4196	0.6052	0.6933
LISA-13B [14]	✗ No	**0.4591**	0.3886	0.3194	0.6687	0.4776
PixelLM-13B [31]	✗ No	0.4285	**0.4273**	0.4306	0.7253	0.5937
SESAME-7B [44]	✓ Yes	0.4180	0.3605	**0.1983**	**0.4304**	0.4607

Comparison of SOTA Reasoning Segmentation Models on our Metrics

- ✓ **Vulnerability to Counterfactual Edits.** Current segmentation models are vulnerable to hallucination, particularly when object identity is subtly changed through counterfactual edits.
- ✓ **Failure of Mitigation Strategies.** Even methods explicitly designed to reduce hallucination remain susceptible to counterfactual visual manipulations, suggesting that prior mitigation strategies do not generalize well to visually grounded reasoning tasks.
- ✓ **Pixel-level Hallucination Elicitation.** HalluSegBench elicits pixel-grounded hallucinations more effectively than label-based methods.

🚀 Follow Our Work!

* Equal contribution

Acknowledgements: DARPA NCSA