ESREAL: Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models

Minchan Kim*[1], **Minyeong Kim***[1], **Junik Bae***[1], Suhwan Choi[1], Sungkyung Kim[1], Buru Chang[2]

[1]Seoul National University, [2]Sogang University

DEEPEST

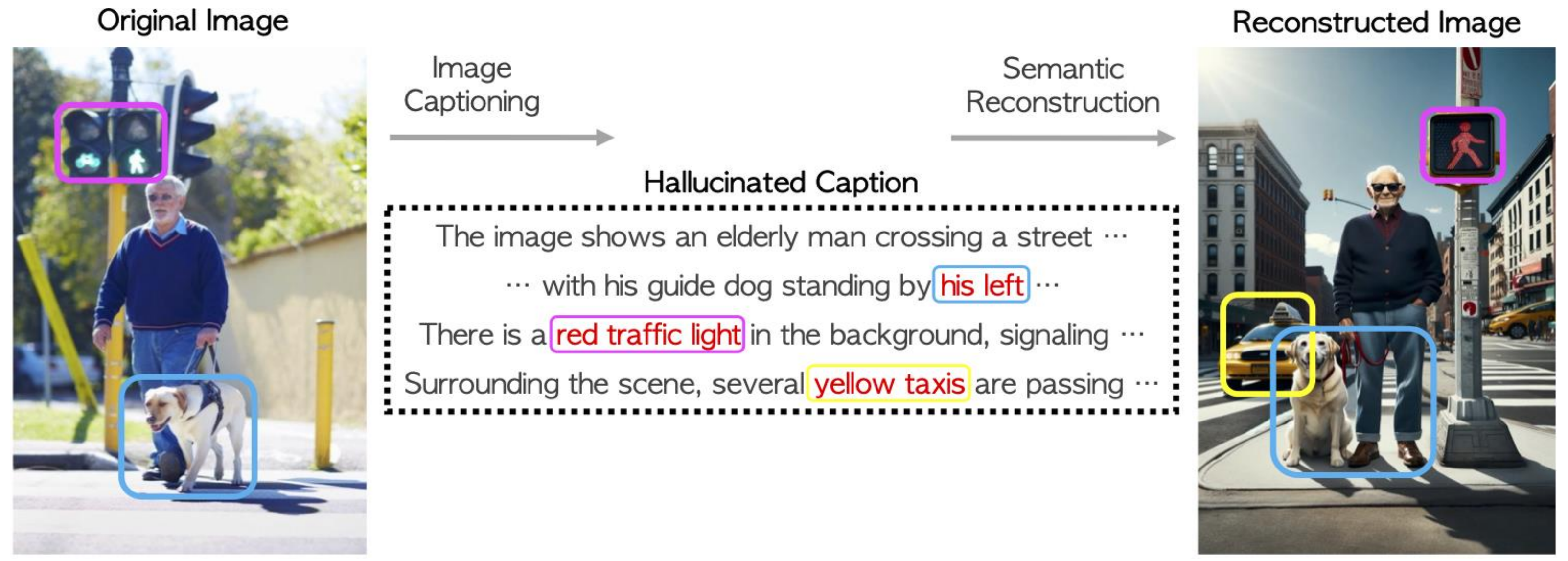
SUMMARY

(1) We propose ESREAL, a fully **unsupervised** hallucination mitigation framework. Our approach is scalable, eliminating the need for annotated data during training.
(2) We craft a hallucination detection pipeline, which facilitates **token-level identification** of hallucinations in generated captions in a reference-free manner via semantic reconstruction.
(3) We show that ESREAL can be applied across a variety of VLMs to effectively mitigate hallucinations.

Image	Generated Caption
	This image features a woman and a young boy sitting on a bed, surrounded by **three cats**. The cats are lying on the bed, with one on the left side, another in the middle, and the third on the right side. The woman and the boy are sitting on the bed, with the woman on the left side and the boy on the right side. There is also a **clock** on the wall, and a **book placed on the bed** possibly for the woman and the boy to read together.
	The image features a woman and a boy sitting on a bed, with **two cats** sitting on the bed as well. The woman is sitting on the left side of the bed, while the boy is on the right side. The cats are located in the middle of the bed, between the woman and the boy. There is a bed in the background, and a chair is visible in the corner of the room. There is also a lamp in the corner of the room and a **book placed on the table**. The scene appears to be a cozy and comfortable setting.
	The image features a yellow fire hydrant situated in a wooded area, surrounded by green bushes and trees. The fire hydrant is **located near the center** of the scene, with a tree on the left side and **another tree** on the right side. The area appears to be a mix of a forest and a park, with the fire hydrant serving as a focal point in the scene.
	The image features a fire hydrant located in a dark area near a large tree. The fire hydrant is situated on the **right side of the tree**, and the tree is positioned on the left side of the image. The scene appears to be dark, with the fire hydrant being the only visible object in the area.

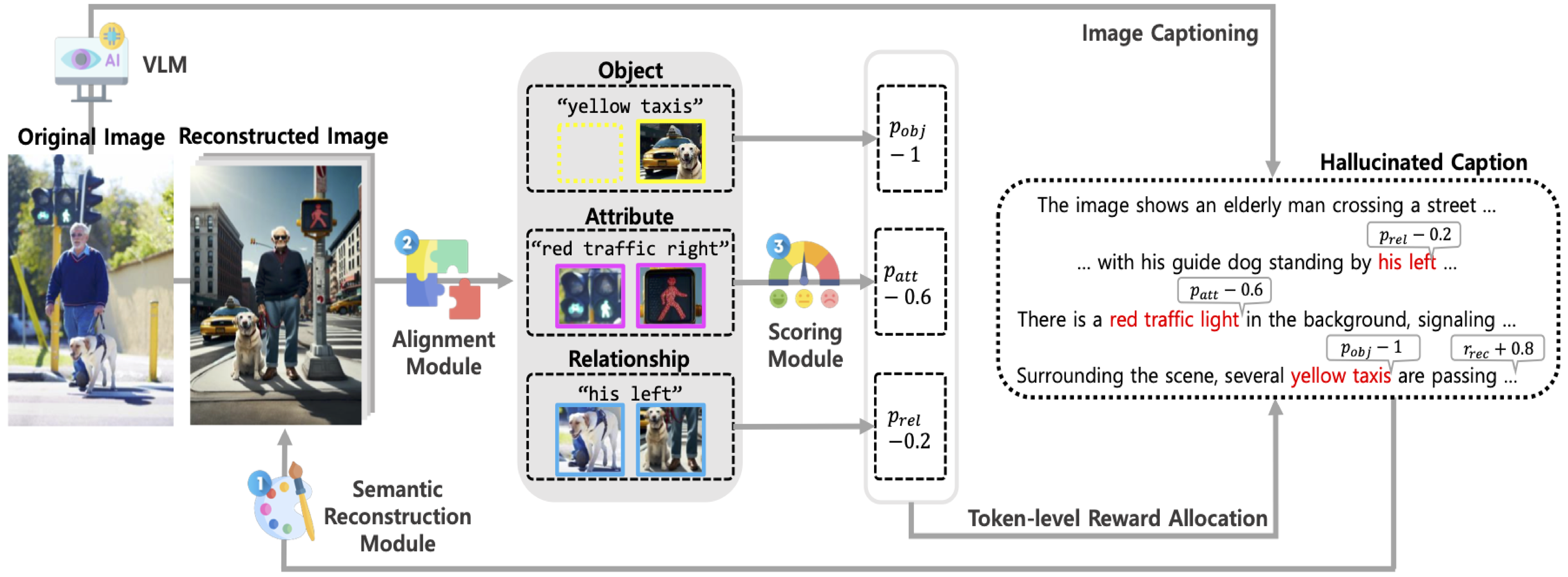
MOTIVATION

- Hallucination in the caption leads to **semantic misalignment** between the original and reconstructed images.
- By comparing the disparities among corresponding regions in the images, we can effectively **identify** and **penalize** the generation of **hallucinated tokens**.



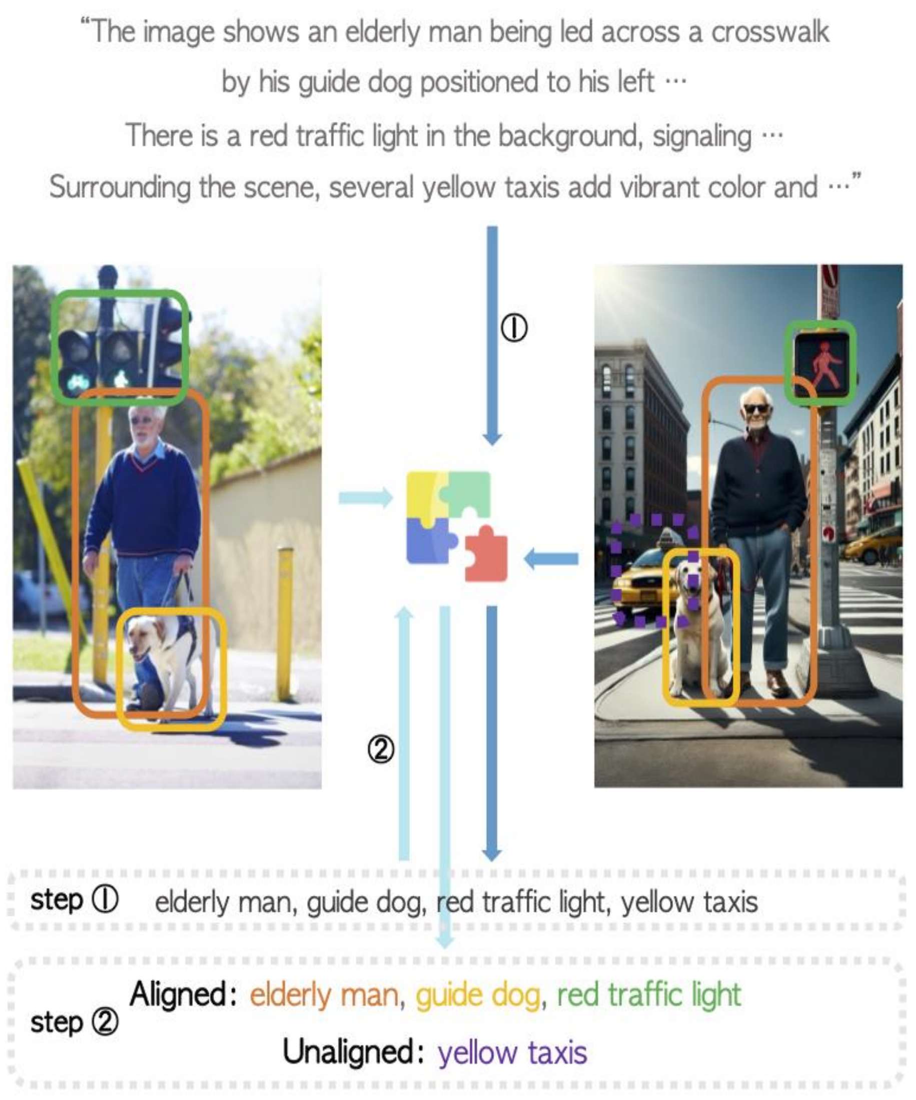
METHODOLOGY

ESREAL uses **token-level penalties** from a hallucination detection pipeline and a **fine-grained PPO approach** to selectively suppress hallucinatory content in generated text.



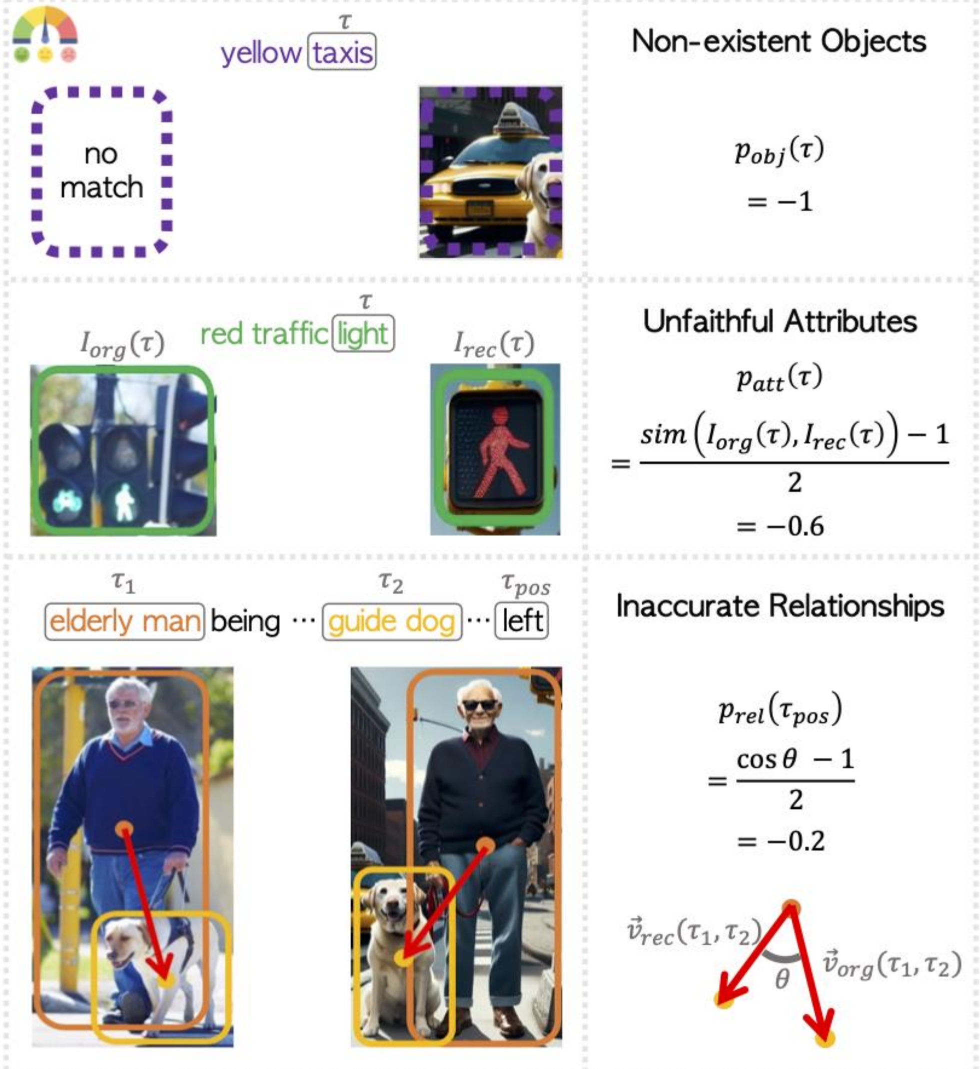
① Semantic Reconstruction Module

Reconstruct the image using the generated caption with the T2I model.



② Alignment Module

Match **object phrases** from the generated caption with regions in both the **reconstructed** image and the **original** image.

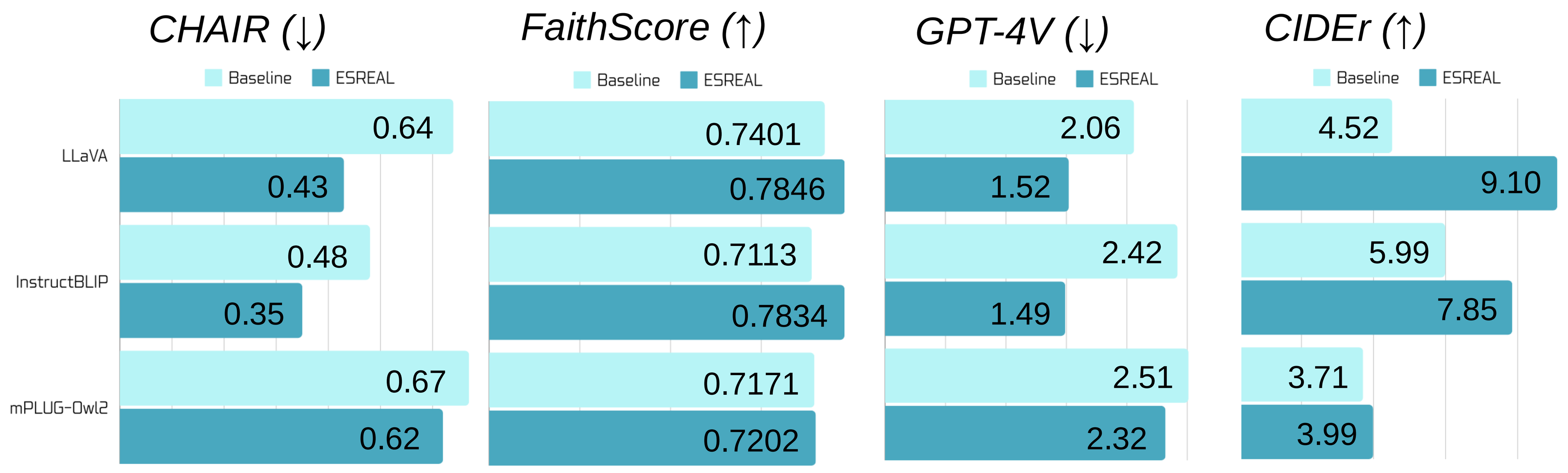


③ Scoring Module

Identify discrepancies between the aligned regions. Apply penalties for object, attribute, or relationship hallucinations.

EXPERIMENTS

① Does ESREAL reduce hallucinations while preserving VLMs' generative abilities?



② Does each penalty effectively target its type?

Model	Method	# Hallucinations per Caption (↓) Object	Attribute	Relationship	Total
InstructBLIP	Baseline	1.23	0.14	1.05	2.42
	ESREAL	0.80	0.06	0.64	1.49
	ESREAL w/o p_{obj}	1.18	0.14	0.92	2.24
	ESREAL w/o p_{att}	0.93	0.09	0.59	1.61
	ESREAL w/o p_{rel}	0.85	0.07	1.28	2.21
	ESREAL w/o r_{rec}	0.68	0.04	1.40	2.12
	ESREAL w/o p_{obj}, p_{att}, p_{rel}	1.15	0.10	1.01	2.26

③ How does T2I model stability impact ESREAL's performance?

T2I Models	Win Rate	Alignment	FaithScore (↑)	GPT-4V (↓)
SDXL-Turbo (1 Step)	0.76	1.86	0.7484	2.23
SDXL-Turbo (4 Steps)	0.79	2.30	0.7834	1.49
Hyper-SDXL (8 Steps)	0.80	2.58	0.8141	1.32
DALLE-3	0.82	2.71	-	-

CONTACT

Minchan Kim

Minyeong Kim

ACKNOWLEDGEMENT

We also thank great previous work including GroundingDINO, SDXL, SDXL-Turbo, Hyper-SDXL, DALLE-3, etc.