DiffusionPID: Interpreting Diffusion via Partial Information Decomposition

NeurIPS 2024

Shaurya Dewan*, Rushikesh Zawar*, Prakanshul Saxena*
Yingshan Chang, Andrew Luo, Yonatan Bisk

Carnegie Mellon University

CMU

Rapid Advancements in Image Generation Models

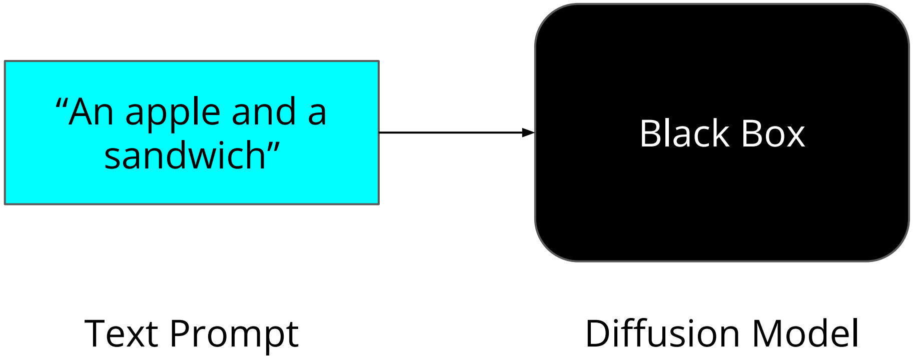
Gemini
Sure! Here's a punk rock flamingo in the clouds wearing a top hat
Generate more

Lack of Understanding

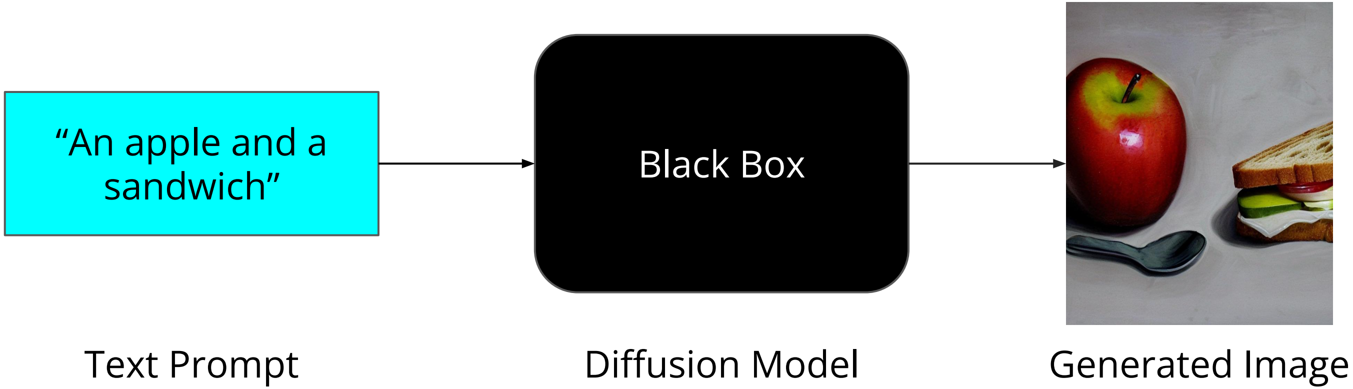
“An apple and a sandwich”

Text Prompt

Lack of Understanding



Lack of Understanding



“An apple and a sandwich”

Black Box

Text Prompt

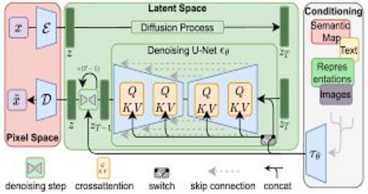
Diffusion Model

Generated Image

Introduction

“An apple and a sandwich”

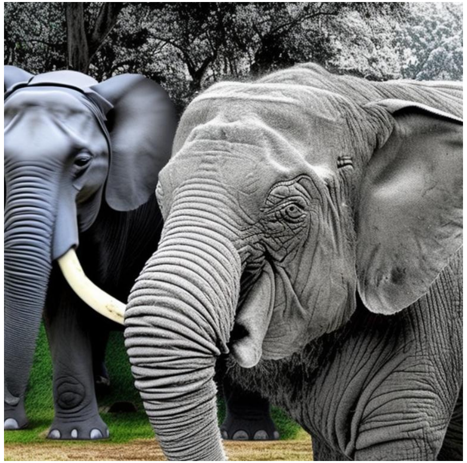
Text Prompt



Diffusion Model

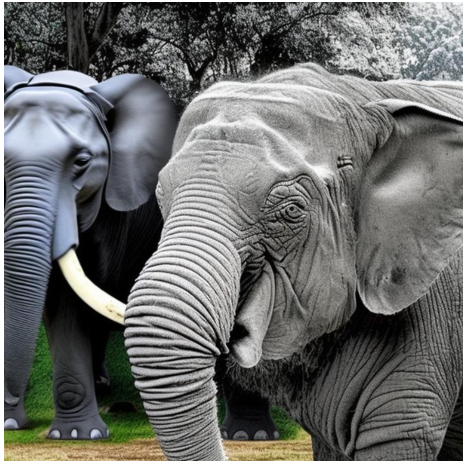
Generated Image

Shortcomings

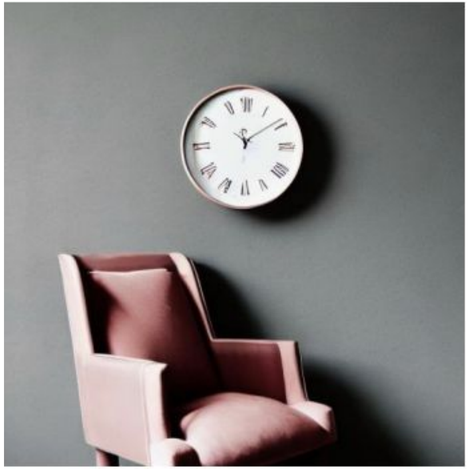


“A dog and an elephant”
(Missing Objects)

Shortcomings



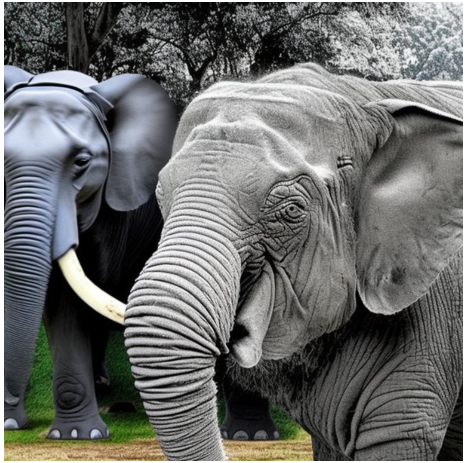
"A dog and an elephant"
(Missing Objects)



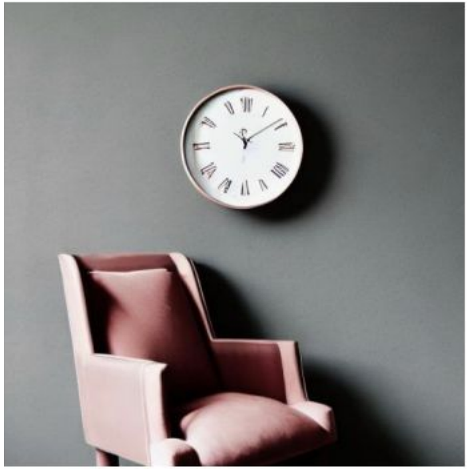
"Pink Clock and Brown Chair"
(Attribution Binding)

Motivation

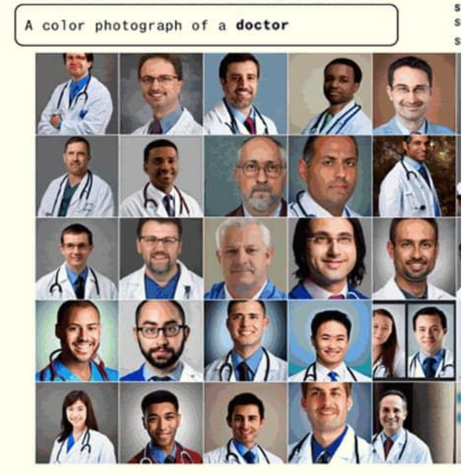
Shortcomings



"A dog and an elephant"
(Missing Objects)

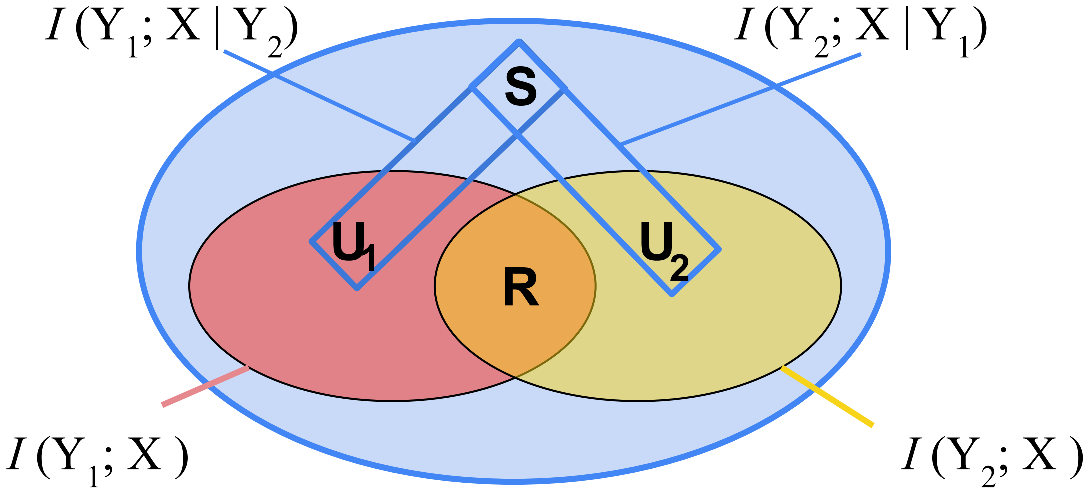


"Pink Clock and Brown Chair"
(Attribution Binding)



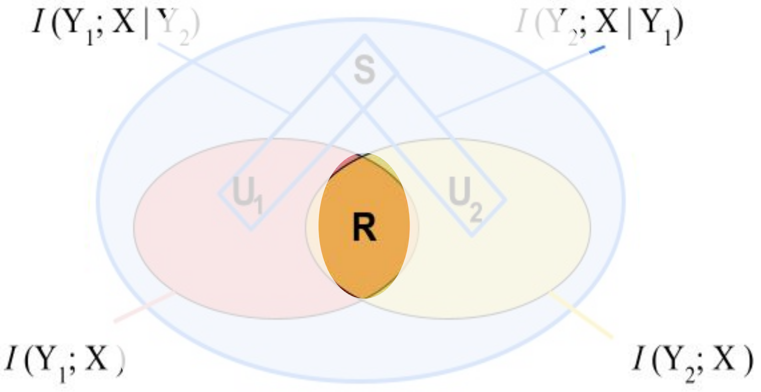
"Doctor"
(Gender Bias)

Partial Information Decomposition (PID)



$$I(Y_1, Y_2; X) = R(Y_1, Y_2; X) + U(Y_1 \backslash Y_2; X) + U(Y_2 \backslash Y_1; X) + S(Y_1, Y_2; X)$$

Partial Information Decomposition (PID)

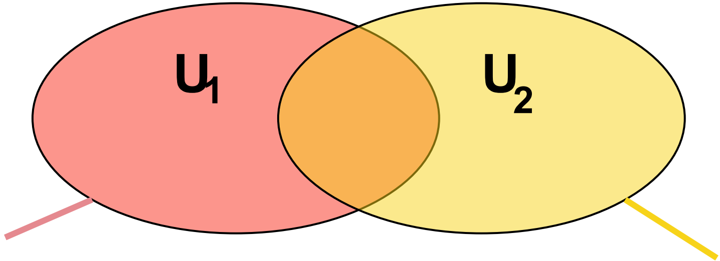


$$I(Y_1, Y_2; X) = \mathbf{R}(\mathbf{Y_1}, \mathbf{Y_2}; \mathbf{X}) + U(Y_1 \backslash Y_2; X) + U(Y_2 \backslash Y_1; X) + S(Y_1, Y_2; X)$$

Partial Information Decomposition (PID)

$I(Y_1; X \mid Y_2)$

$I(Y_2; X \mid Y_1)$

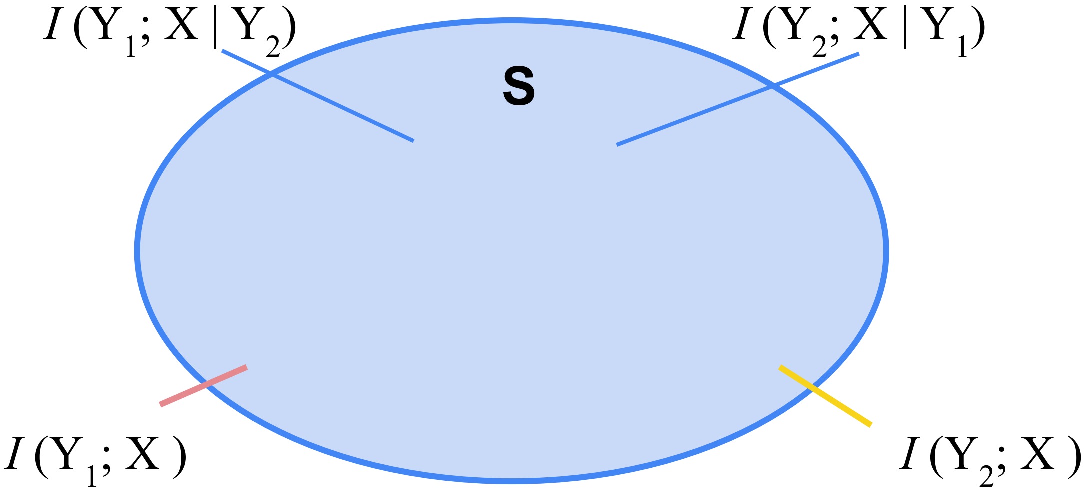


$I(Y_1; X)$

$I(Y_2; X)$

$$I(Y_1, Y_2; X) = R(Y_1, Y_2; X) + \mathbf{U}(\mathbf{Y_1} \backslash \mathbf{Y_2}; \mathbf{X}) + \mathbf{U}(\mathbf{Y_2} \backslash \mathbf{Y_1}; \mathbf{X}) + S(Y_1, Y_2; X)$$

Partial Information Decomposition (PID)



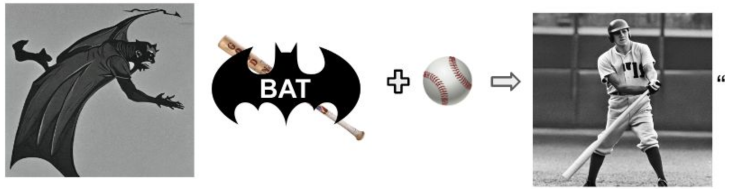
$$I(Y_1, Y_2; X) = R(Y_1, Y_2; X) + U(Y_1 \backslash Y_2; X) + U(Y_2 \backslash Y_1; X) + \mathbf{S}(\mathbf{Y_1}, \mathbf{Y_2}; \mathbf{X})$$

Our Method

He swung the bat

“Effect of Synergy”

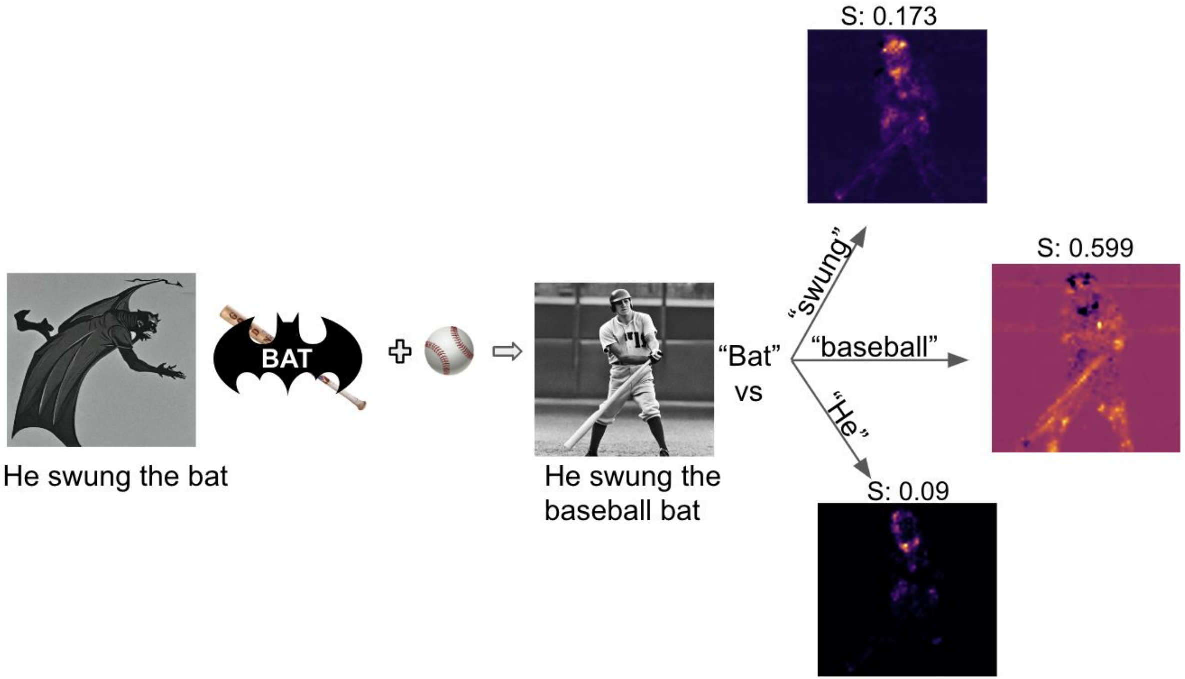
Our Method



He swung the bat

He swung the baseball bat

“Effect of Synergy”



"Effect of Synergy"

Our Method

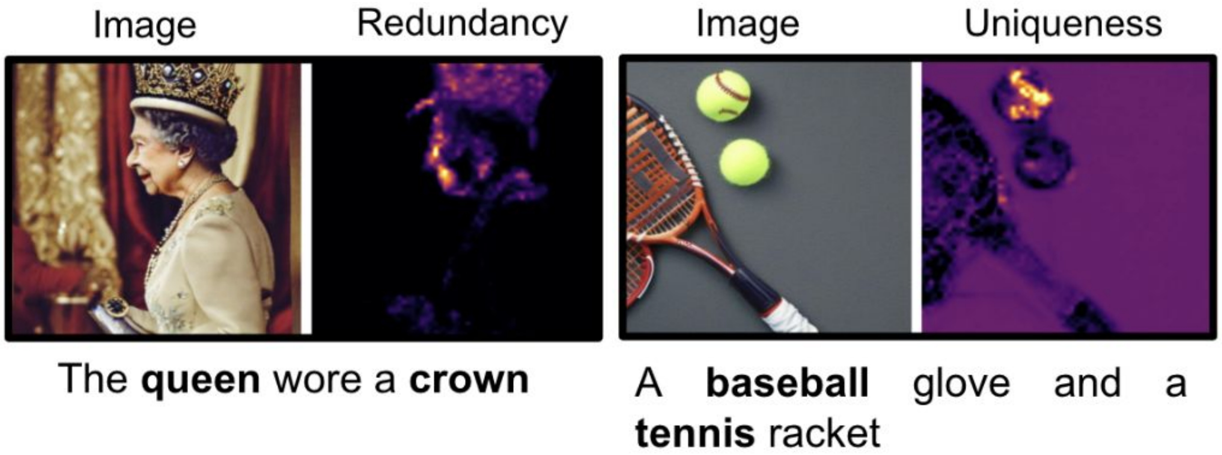
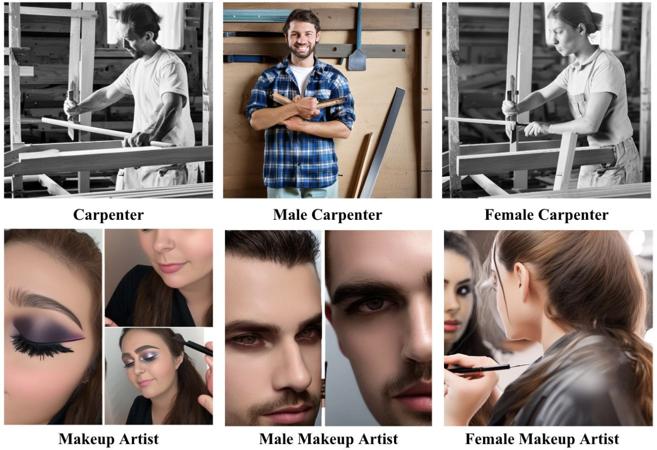


Image
Redundancy
Image
Uniqueness
The **queen** wore a **crown**
A **baseball** glove and a **tennis** racket

Uniqueness and Redundancy

Gender Bias



Carpenter

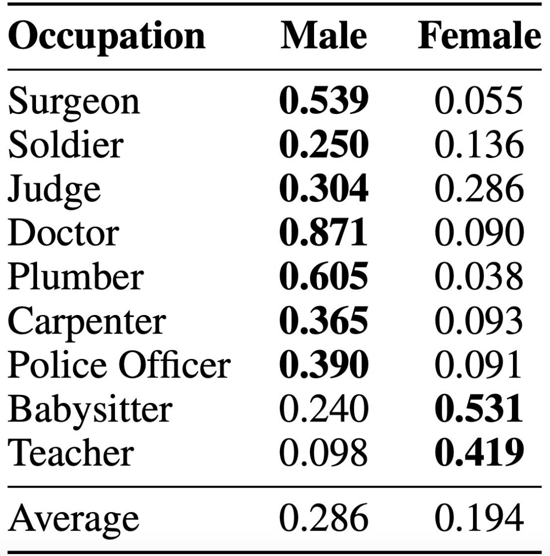
Male Carpenter

Female Carpenter

Makeup Artist

Male Makeup Artist

Female Makeup Artist



Occupation	Male	Female
Surgeon	**0.539**	0.055
Soldier	**0.250**	0.136
Judge	**0.304**	0.286
Doctor	**0.871**	0.090
Plumber	**0.605**	0.038
Carpenter	**0.365**	0.093
Police Officer	**0.390**	0.091
Babysitter	0.240	**0.531**
Teacher	0.098	**0.419**
Average	0.286	0.194

Results

Ethnic Bias

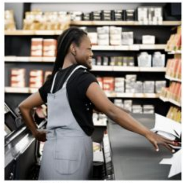
A hispanic environmentalist

A hispanic archaeologist

A caucasian farmer

A caucasian barista

A black data analyst

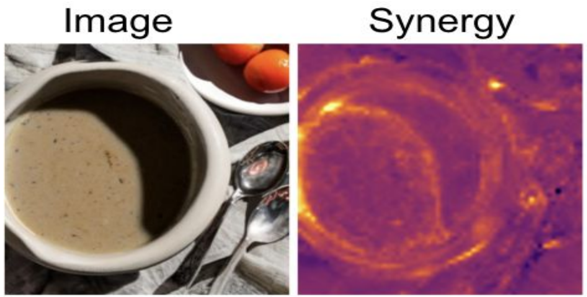

A black cashier

A asian athlete

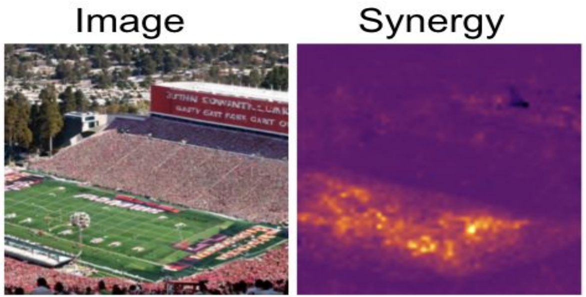
A asian airline pilot

Occupation	Black	Asian	Caucasian	Hispanic
Athlete	**0.321**	0.132	0.167	0.156
Artist	**0.106**	0.069	0.062	0.045
Engineer	0.126	**0.156**	0.080	0.097
Physicist	0.109	**0.209**	0.162	0.064
Butcher	0.110	0.179	**0.474**	0.396
Coach	0.118	0.107	**0.433**	0.128
Nurse	0.106	0.106	0.127	**0.219**
Agriculturist	0.046	0.117	0.337	**0.450**
Average	0.133	0.233	0.255	0.236

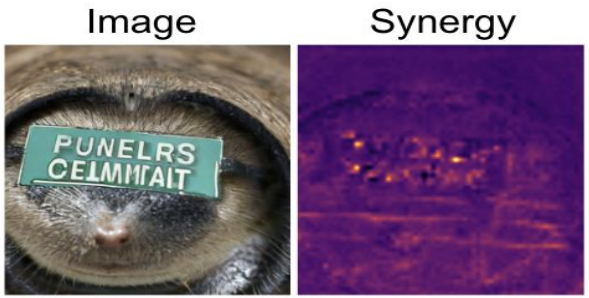
Homonyms



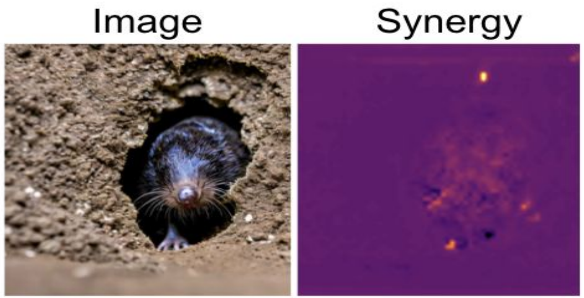
c = she served soup in a ceramic bowl
y = ‘bowl’ vs ‘ceramic



c = the football game was held at the rose bowl stadium
y = ‘bowl’ vs ‘game’

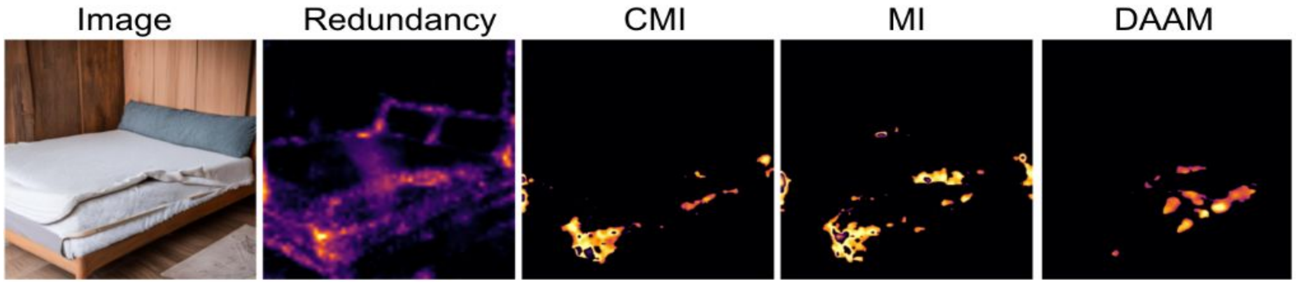


c = he suspected his coworker might be a mole leaking information to competitors
y = ‘mole’ vs ‘coworker’

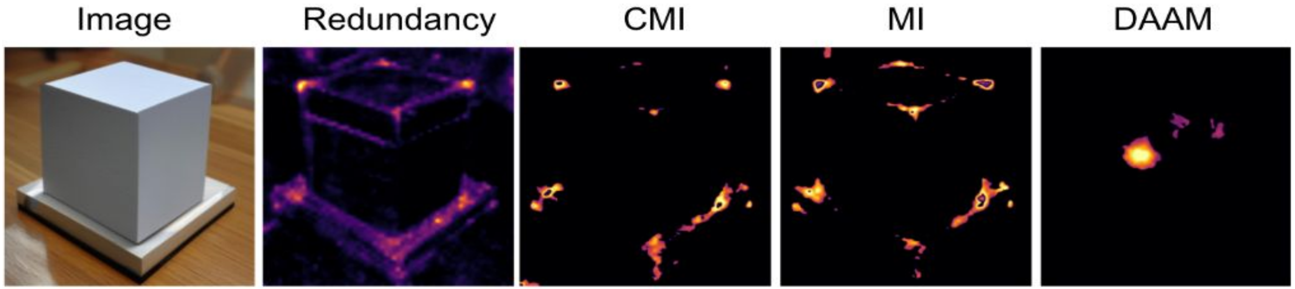


c = the mole burrowed underground searching for insects
y = ‘mole’ vs ‘searching”

Synonyms

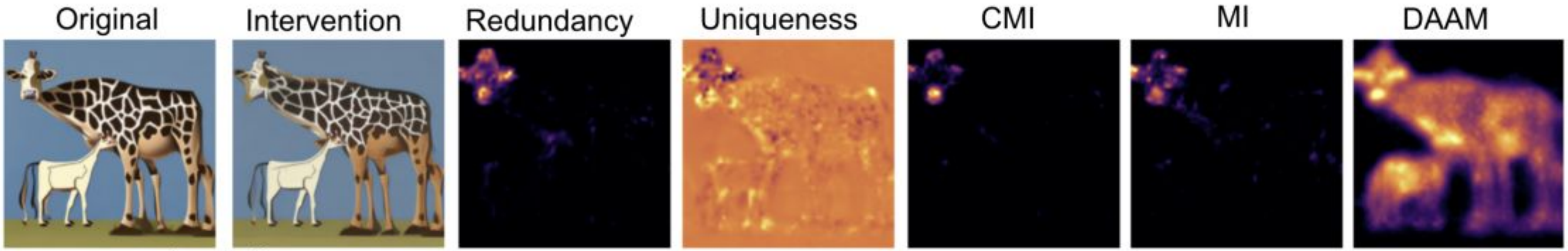


c = the bed was covered in a soft mattress with a plush blanket
y = ‘bed’ vs ‘mattress’

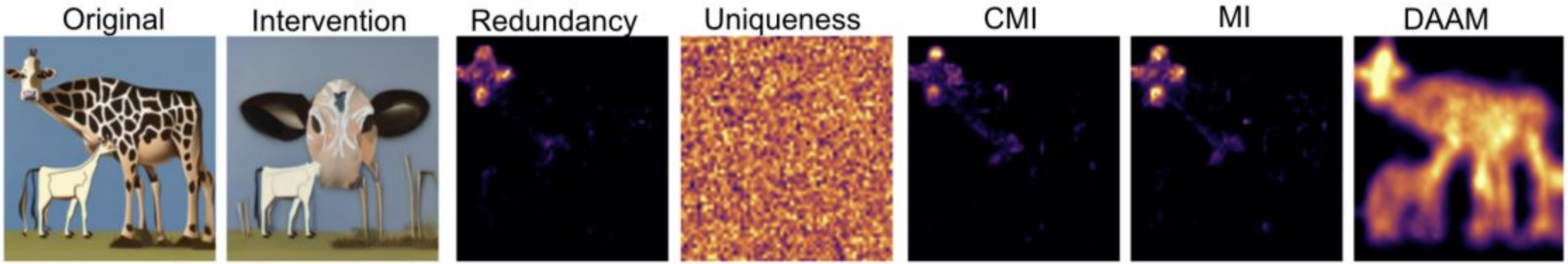


c = the cube was a perfect cuboid with equal sides and angles
y = ‘cube’ vs ‘cuboid’

Prompt Intervention

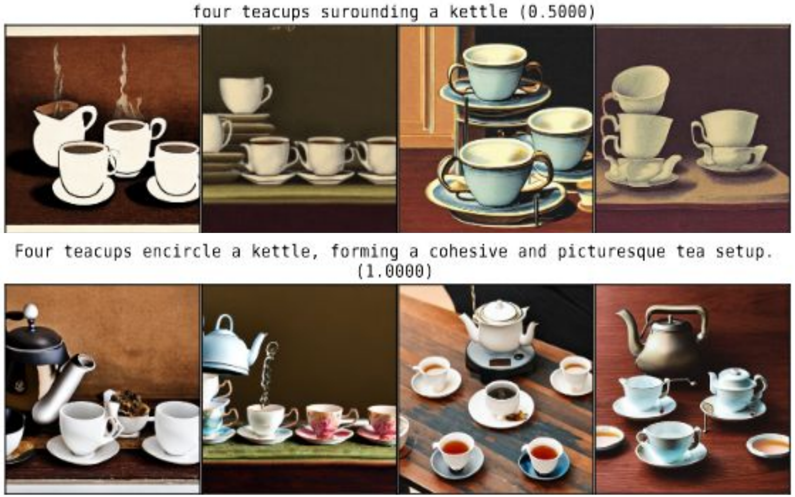


c = a cow and a giraffe
y = ‘cow’



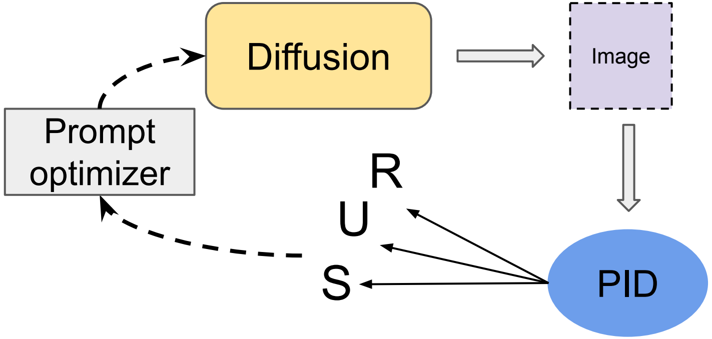
c = a cow and a giraffe
y = ‘giraffe’

Applications


four teacups surounding a kettle (0.5000)

Four teacups encircle a kettle, forming a cohesive and picturesque tea setup. (1.0000)

Prompt Engineering



Model Improvement

Future Works

1. PID gives a detailed breakdown about the Mutual Information between 2 concepts. This can be leveraged to better model uncertainty and can find applications in fields like Active learning.

2. PID can be extended to multiple modalities and other models.

Thank You