中国科学院大学
University of Chinese Academy of Sciences

Towards Trustworthy and Reliable Multimodal Foundation Model: Explainable Mechanisms with Enhancement Applications

Ruoyu Chen

Final year PH.D. Candidate

University of Chinese Academy of Sciences

https://ruoyuchen10.github.io/

2026.02.11

Outline – Three parts

- **Part 1** — Why Explainable AI?
  - Background
  - Evolution of Attribution Techniques
  - Challenges
- **Part 2** — Explainable Attribution Mechanisms
  - Subset Ranking-based Attribution
  - Explaining Autoregressive MLLM
- **Part 3** — Attribution-guided Learning
  - Prior-Aligned Training with Attribution Constraints
  - Counterfactual Data Augmentation

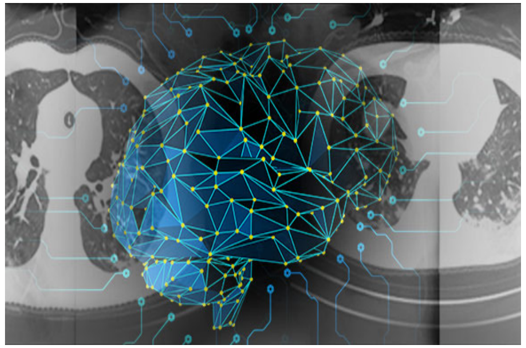
1 Why We Need Explainable AI?

The reliability and security of agents’ decisions are the core challenges in their practical applications, which directly determine whether they can be reliably deployed in the real world and win the trust of users.

Autonomous Drive

Education

Financial

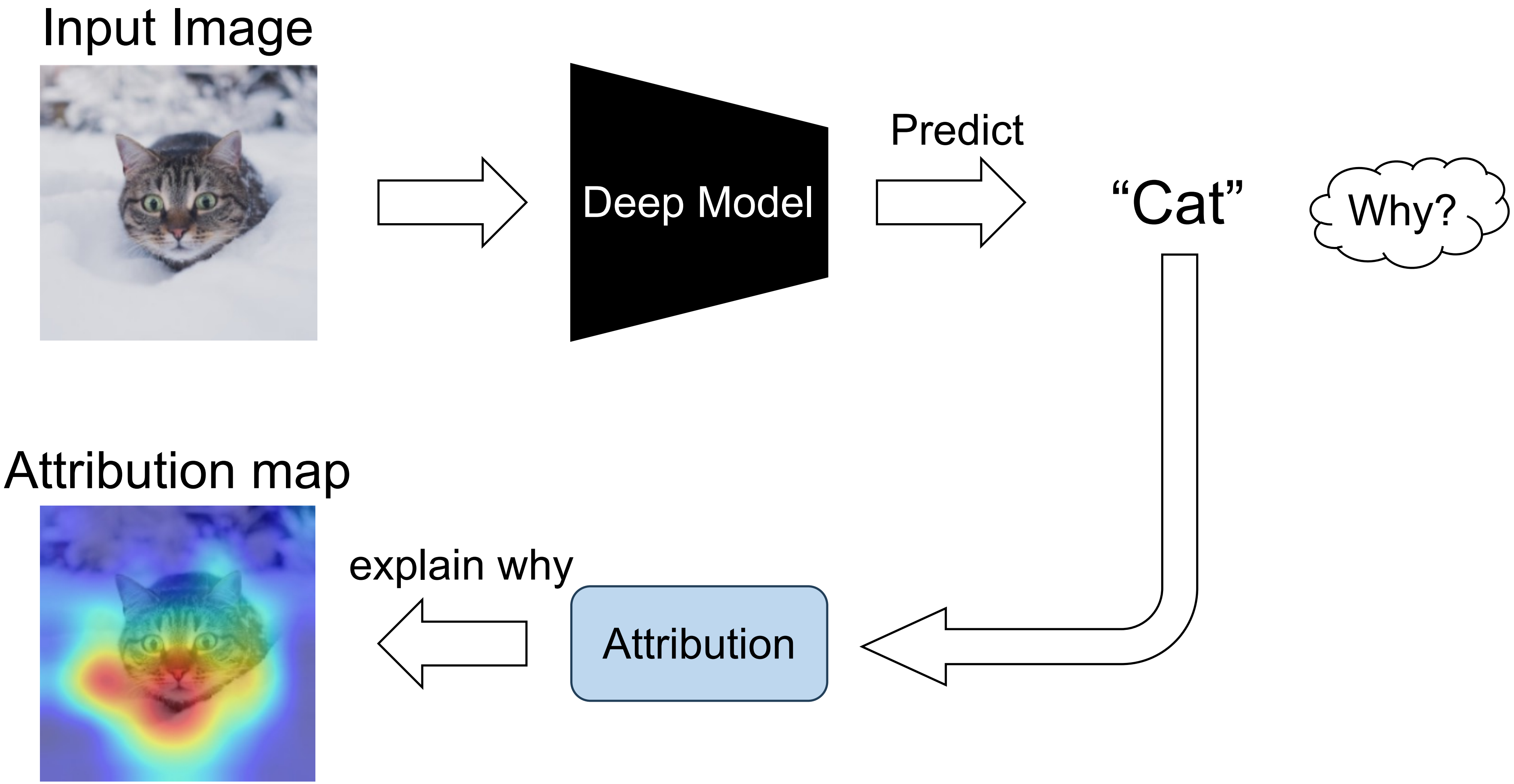


Healthcare

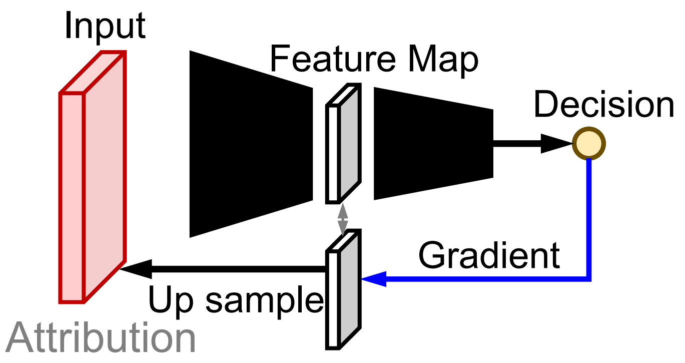
So we need explainable AI!

1 What's Attribution?

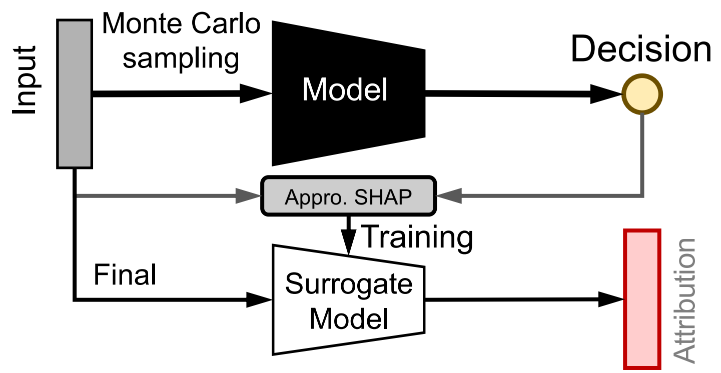
An Example of Image Attribution: The main objective in attribution techniques is to highlight the discriminating variables for decision-making.



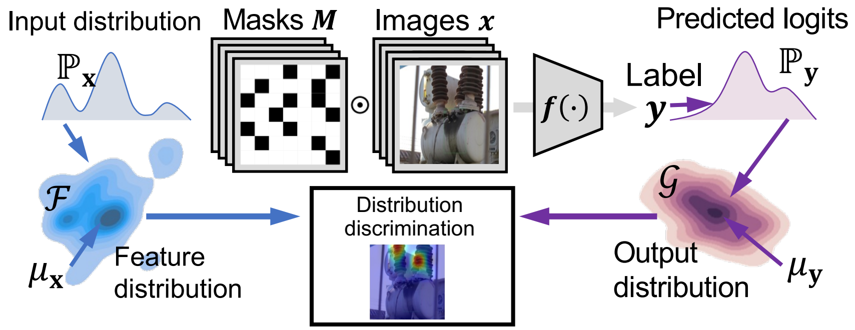
1 Evolution of Attribution Techniques



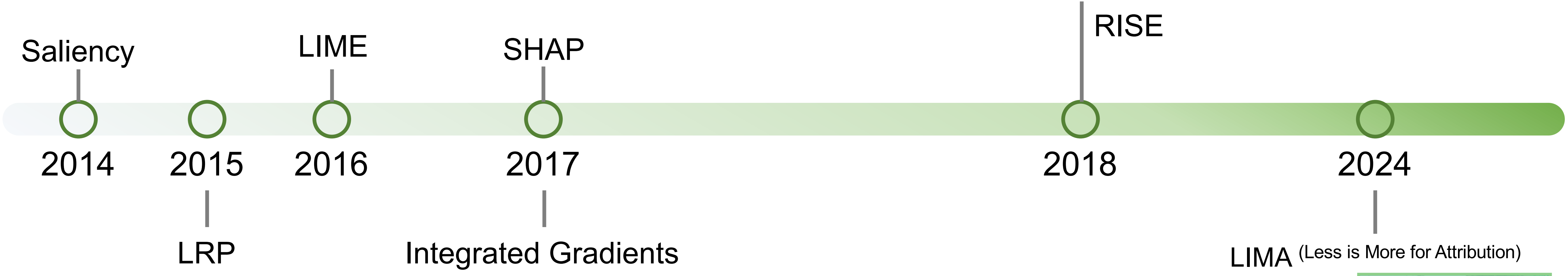
Gradient-based Attribution

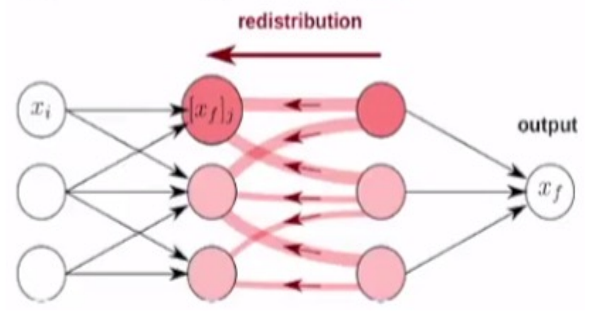


Shapely Value-based Attribution

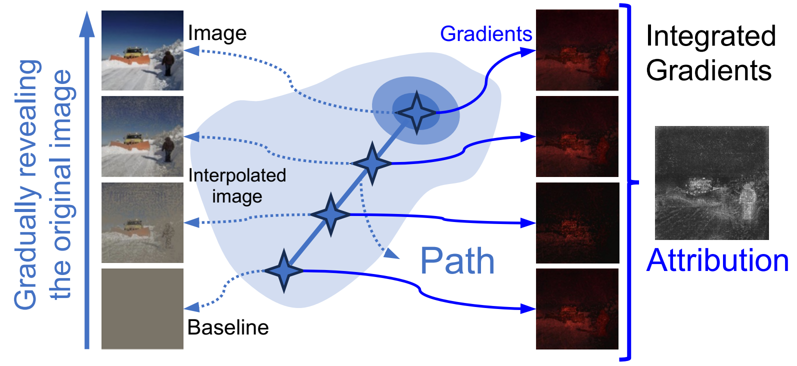


Perturbation-based Attribution

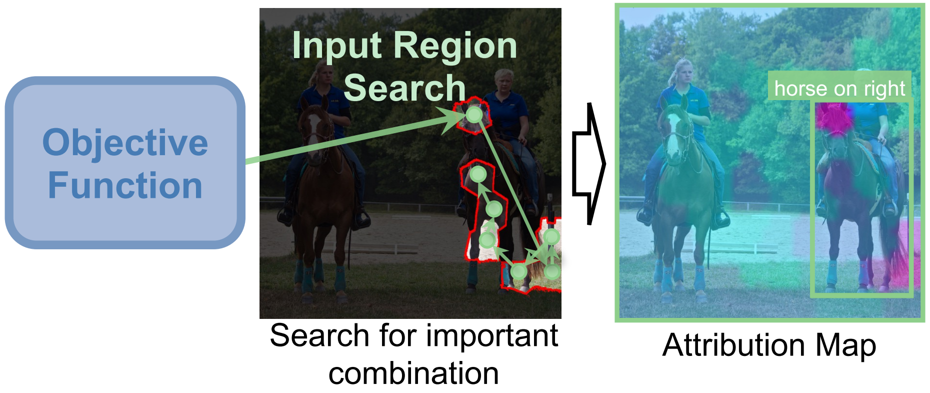




Propagation-based Attribution



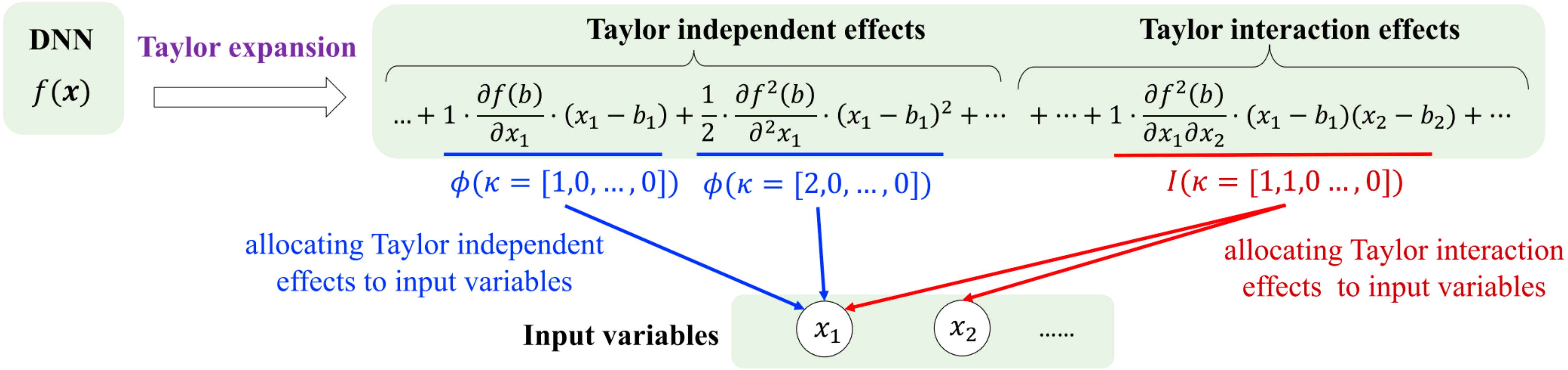
Path-based Attribution



Subset Ranking-based Attribution

1 Challenges of Attribution

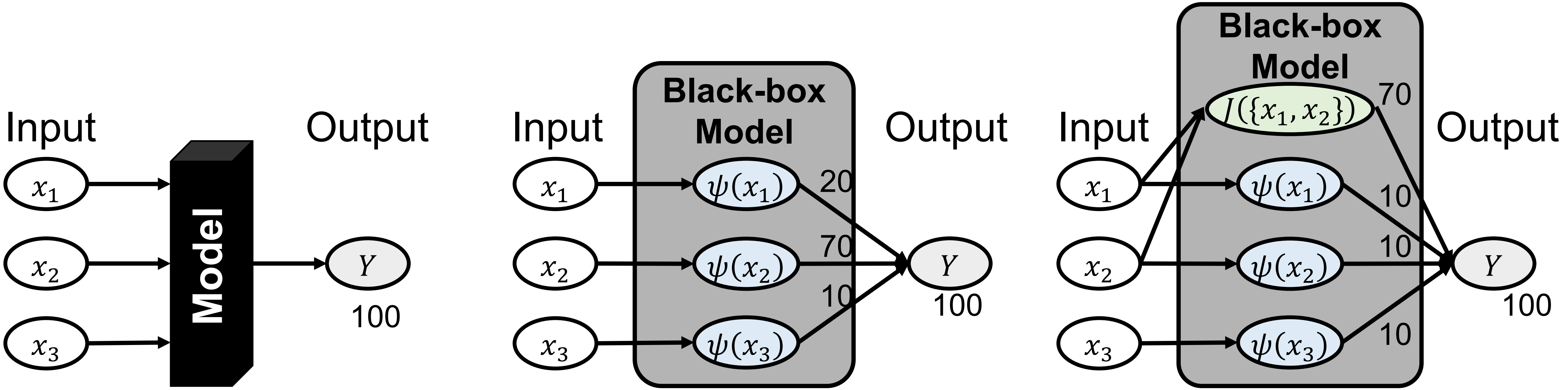
Deng *et al.* formulate the model's decision process using a Taylor expansion.


DNN
$f(\boldsymbol{x})$
Taylor expansion
Taylor independent effects
Taylor interaction effects
$\dots + 1 \cdot \frac{\partial f(b)}{\partial x_1} \cdot (x_1 - b_1) + \frac{1}{2} \cdot \frac{\partial f^2(b)}{\partial^2 x_1} \cdot (x_1 - b_1)^2 + \cdots + \cdots + 1 \cdot \frac{\partial f^2(b)}{\partial x_1 \partial x_2} \cdot (x_1 - b_1)(x_2 - b_2) + \cdots$
$\phi(\kappa = [1,0,\dots,0])$
$\phi(\kappa = [2,0,\dots,0])$
$I(\kappa = [1,1,0 \dots,0])$
allocating Taylor independent effects to input variables
allocating Taylor interaction effects to input variables
Input variables
x_1
x_2
......

Deng, Huiqi, et al. "Unifying fourteen post-hoc attribution methods with taylor interactions." *TPAMI* 46.7 (2024): 4625-4640.

1 Challenges of Attribution

Interaction：The nonlinear relationship among input elements. In general, the stronger the nonlinearity, the more complex the interaction is considered [1,2].



Input–output relationships

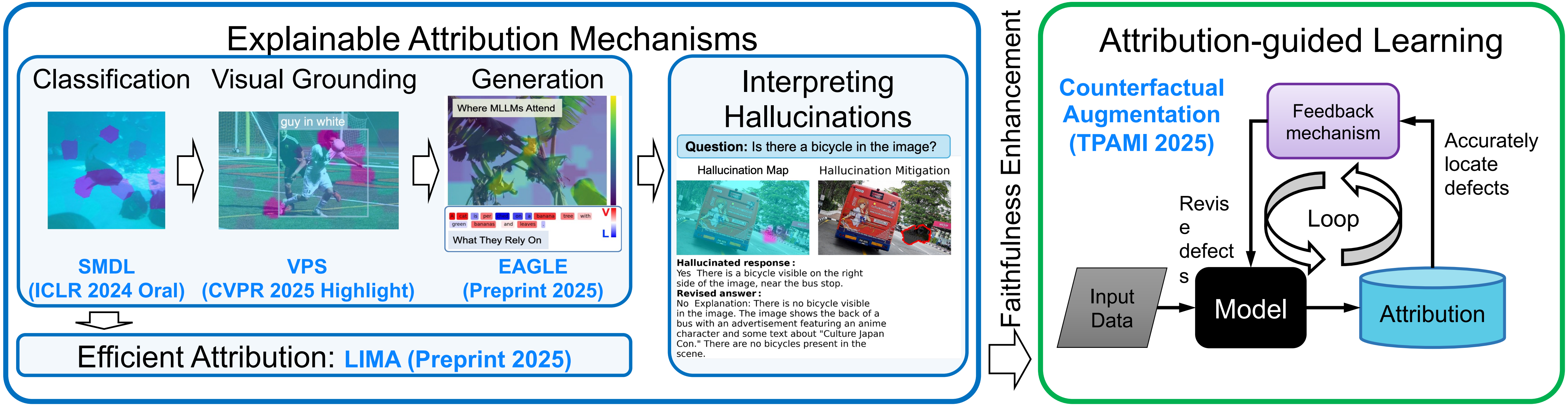
Independent effects within interactions

Combinational effects within interactions

The more combinational effects that exist (i.e., $J(S)$), the more complex the interactions are considered, and consequently, the more difficult attribution becomes.

[1] Chen, Lu, et al. “Can LLMs Reason Soundly in Law? Auditing Inference Patterns for Legal Judgment." *ICLR* 2026.
[2] Deng, Huiqi, et al. "Unifying fourteen post-hoc attribution methods with taylor interactions." *TPAMI* 46.7 (2024): 4625-4640.

1 Overview of This Talk



Explainable Attribution Mechanisms

Classification

Visual Grounding

Generation

guy in white
Where MLLMs Attend
A cat is per ched on a banana tree with green bananas and leaves .
V
L
What They Rely On

SMDL (ICLR 2024 Oral)

VPS (CVPR 2025 Highlight)

EAGLE (Preprint 2025)

Efficient Attribution: LIMA (Preprint 2025)

Interpreting Hallucinations

Question: Is there a bicycle in the image?

Hallucination Map

Hallucination Mitigation

Hallucinated response :
Yes There is a bicycle visible on the right side of the image, near the bus stop.

Revised answer :
No Explanation: There is no bicycle visible in the image. The image shows the back of a bus with an advertisement featuring an anime character and some text about "Culture Japan Con." There are no bicycles present in the scene.

Faithfulness Enhancement

Attribution-guided Learning

Counterfactual Augmentation (TPAMI 2025)

Feedback mechanism

Accurately locate defects

Loop

Revis e defect s

Input Data

Model

Attribution

This framework summarizes the main research pipeline from **explainable attribution mechanisms** to **attribution-guided learning** for reliable multimodal models.

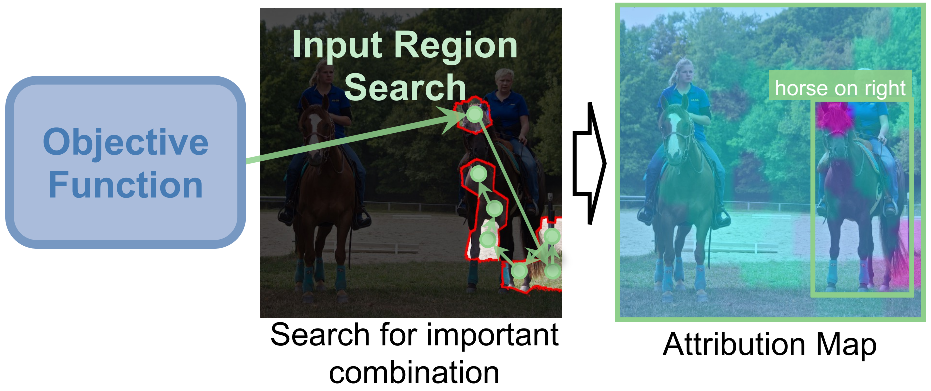
2 Subset Ranking-based Attribution

Divide the image into a set of small sub-regions and ranking the sub-regions according to their importance.

- Reformulate the attribution problem as a *submodular subset selection problem*;

$$\max_{S\subseteq V,|S|<k} \mathcal{F}(S)$$

- Employ regional *search* to expand the sub-region set to *alleviate the insufficient dense of the attribution region*;
- A novel *submodular mechanism* is constructed to *limit the search for regions with wrong class responses*.

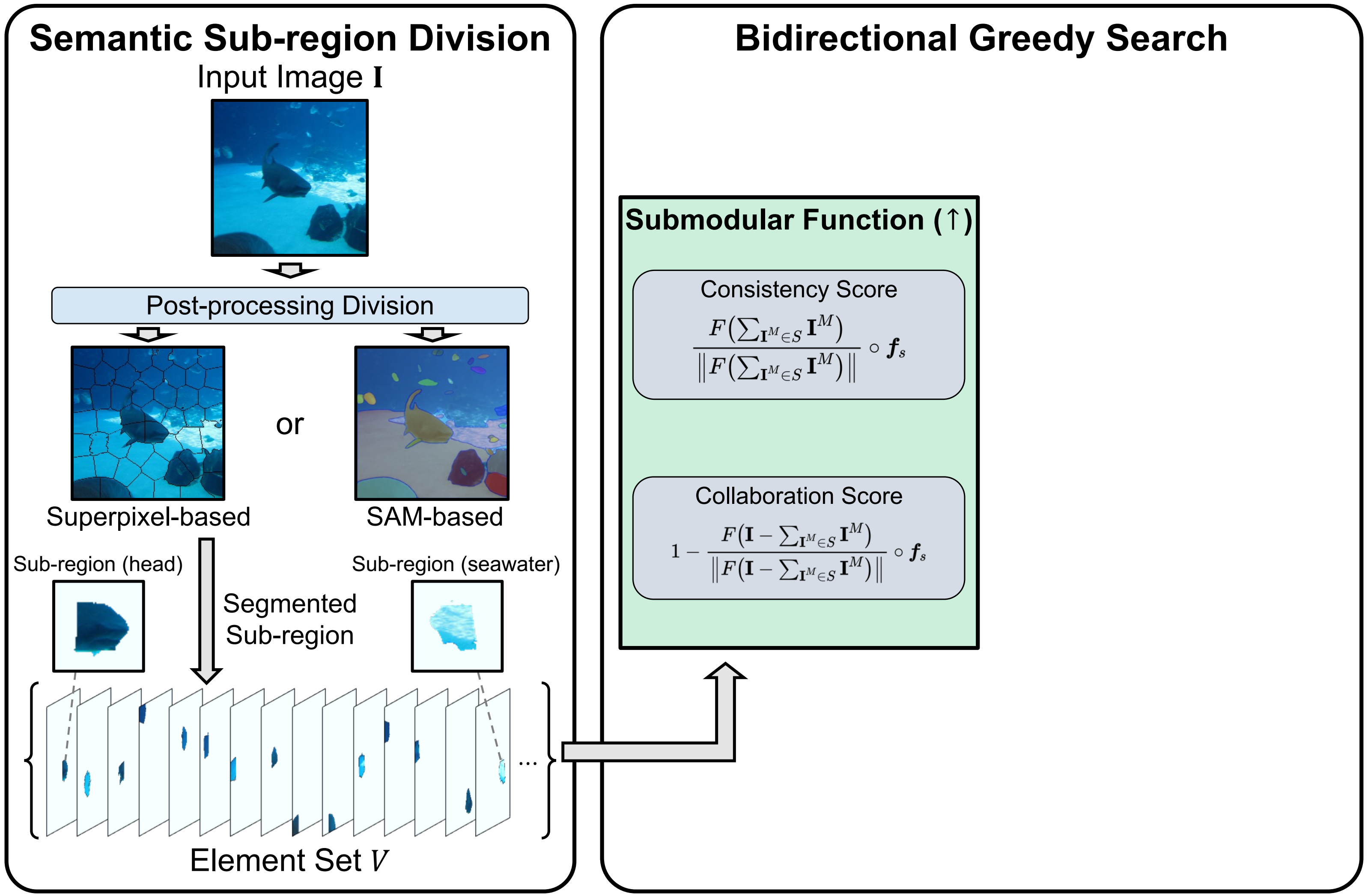


Search for important combination

Attribution Map

Subset Ranking-based Attribution

2 Subset Ranking-based Attribution — Method



Semantic Sub-region Division

Input Image $\mathbf{I}$

Post-processing Division

Superpixel-based

or

SAM-based

Sub-region (head)

Sub-region (seawater)

Segmented Sub-region

Element Set V

Bidirectional Greedy Search

Submodular Function (↑)

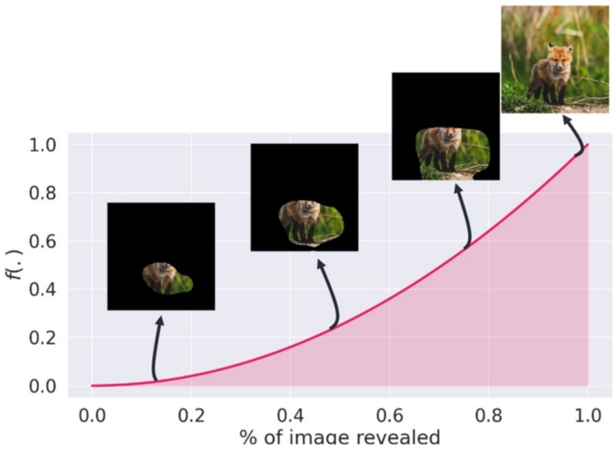
Consistency Score

$$\frac{F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$$

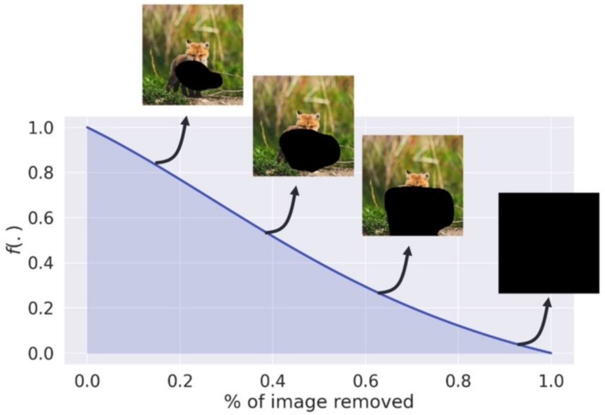
Collaboration Score

$$1-\frac{F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$$

Insertion* (high AUC = better faithfulness)

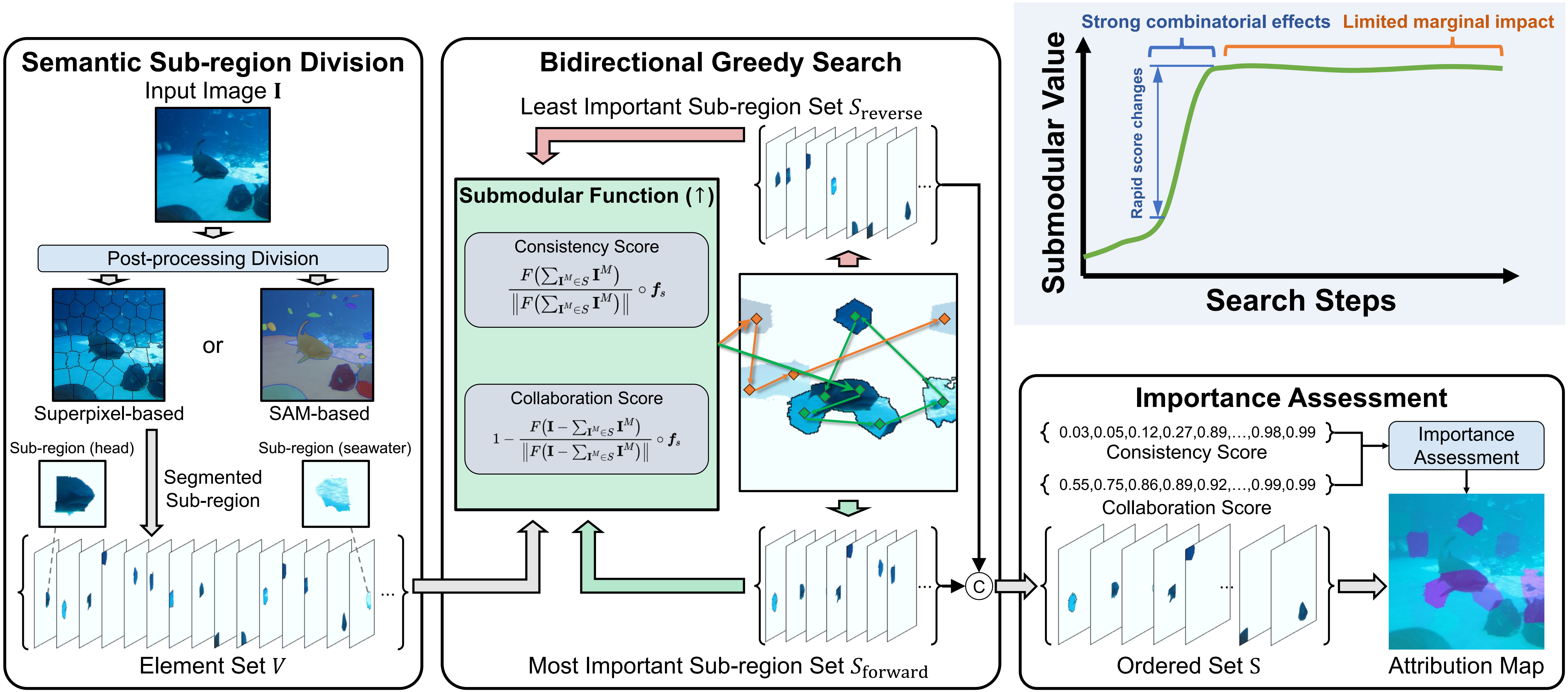


Deletion (low AUC = better faithfulness)



- **Ruoyu Chen**, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection." **ICLR 2024**. (**Oral Presentation**, **1.16%**)
- **Ruoyu Chen**, et al. "Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection." Preprint 2025.

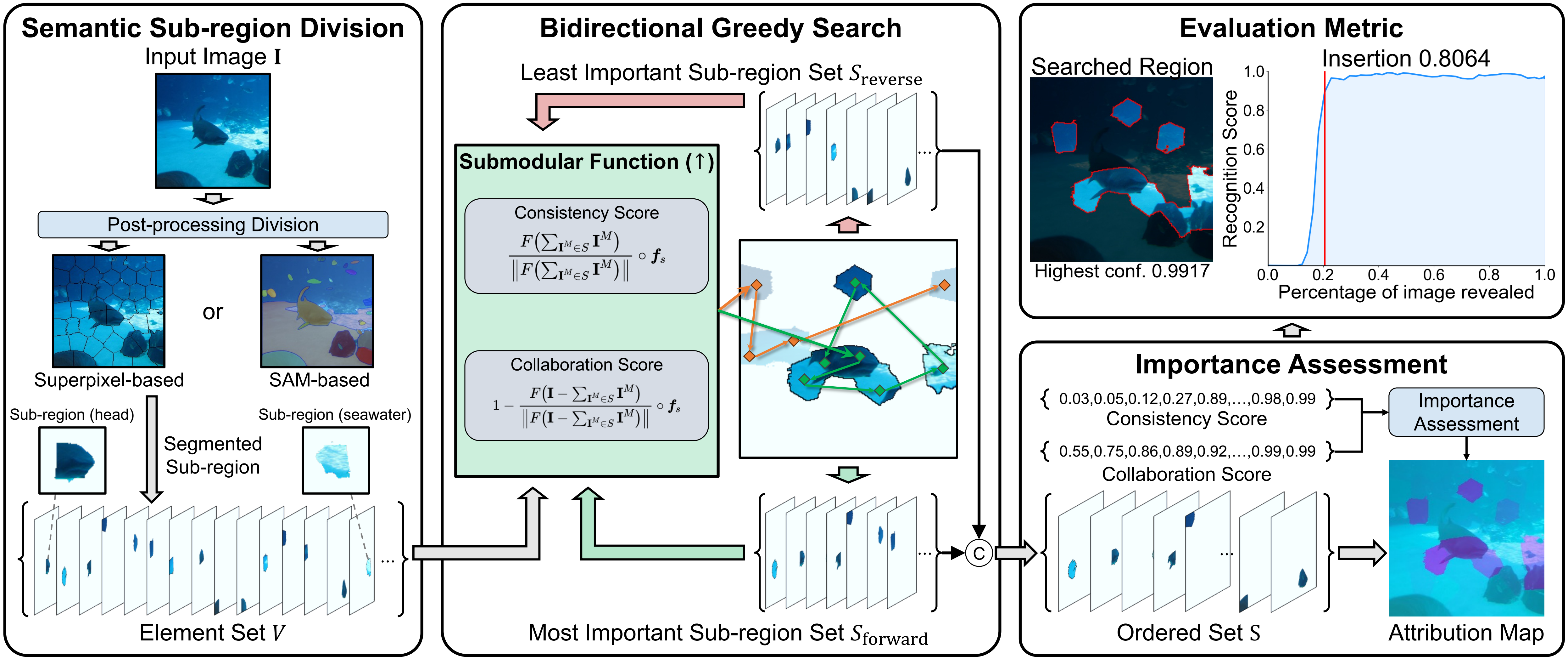
2 Subset Ranking-based Attribution — Method



Semantic Sub-region Division
Input Image $\mathbf{I}$
Post-processing Division
or
Superpixel-based
SAM-based
Sub-region (head)
Sub-region (seawater)
Segmented Sub-region
...
Element Set V
Bidirectional Greedy Search
Least Important Sub-region Set S_{reverse}
Submodular Function (↑)
Consistency Score
$\frac{F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$
Collaboration Score
$1-\frac{F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$
...
C
Most Important Sub-region Set S_{forward}
Strong combinatorial effects
Limited marginal impact
Submodular Value
Rapid score changes
Search Steps
Importance Assessment
{ 0.03,0.05,0.12,0.27,0.89,...,0.98,0.99 }
Consistency Score
{ 0.55,0.75,0.86,0.89,0.92,...,0.99,0.99 }
Collaboration Score
Importance Assessment
...
Ordered Set S
Attribution Map

- **Ruoyu Chen**, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection." ICLR 2024. (Oral Presentation, 1.16%)
- **Ruoyu Chen**, et al. "Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection." Preprint 2025.

2 Subset Ranking-based Attribution — Method



Semantic Sub-region Division

Input Image $\mathbf{I}$

Post-processing Division

Superpixel-based

or

SAM-based

Sub-region (head)

Sub-region (seawater)

Segmented Sub-region

Element Set V

Bidirectional Greedy Search

Least Important Sub-region Set S_{reverse}

Submodular Function (↑)

Consistency Score

$$\frac{F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$$

Collaboration Score

$$1-\frac{F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)}{\left\|F\left(\mathbf{I}-\sum_{\mathbf{I}^M\in S}\mathbf{I}^M\right)\right\|}\circ \boldsymbol{f}_s$$

Most Important Sub-region Set S_{forward}

Evaluation Metric

Searched Region

Highest conf. 0.9917

Insertion 0.8064

Importance Assessment

{ 0.03,0.05,0.12,0.27,0.89,…,0.98,0.99 }
Consistency Score

{ 0.55,0.75,0.86,0.89,0.92,…,0.99,0.99 }
Collaboration Score

Importance Assessment

Ordered Set S

Attribution Map

- **Ruoyu Chen**, et al. “Less is More: Fewer Interpretable Region via Submodular Subset Selection.” **ICLR 2024**. (**Oral Presentation**, **1.16%**)
- **Ruoyu Chen**, et al. “Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection.” Preprint 2025.

2 Subset Ranking-based Attribution — Evaluation Metrics

Faithfulness Evaluation Metrics

Deletion AUC score

Deletion AUC measures the decrease in the model score when important variables are set to a baseline state. Intuitively, a sharp drop indicates that the explanation method has effectively identified the variables that are critical to the model's decision.

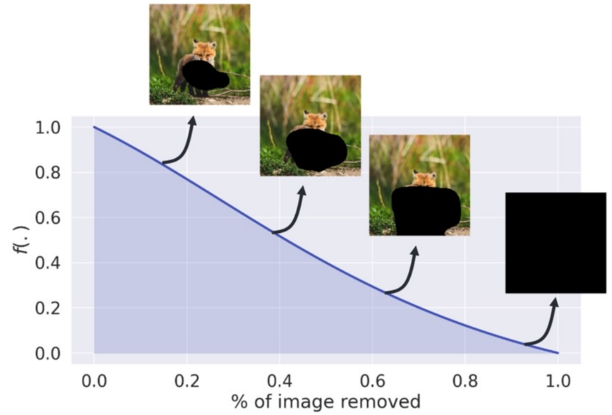
$$\text{Deletion}^{(k)} = f\left(x_{[x_{\boldsymbol{u}}=x_0]}\right)$$

Insertion AUC score

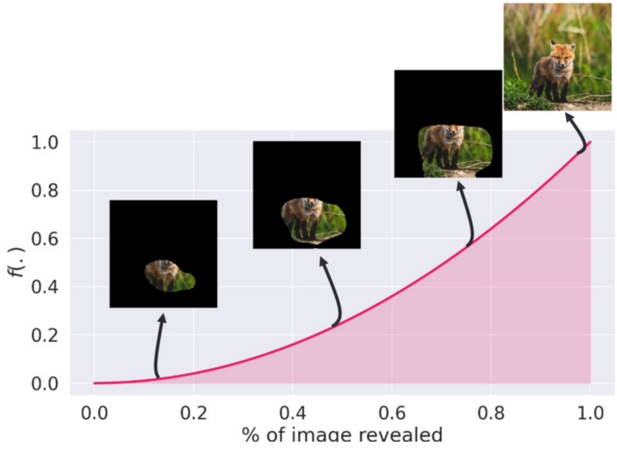
Insertion AUC follows the reverse procedure of Deletion, beginning with a baseline image and gradually inserting the most important variables. A faster increase in the model score indicates that the explanation method more accurately identifies decision-relevant evidence.

$$\text{Insertion}^{(k)} = f\left(x_{[x_{\overline{\boldsymbol{u}}}=x_0]}\right)$$

Deletion (low AUC = better faithfulness)



Insertion* (high AUC = better faithfulness)



Petsiuk, Vitali, Abir Das, and Kate Saenko. "Rise: Randomized input sampling for explanation of black-box models." *BMVC*. 2018.

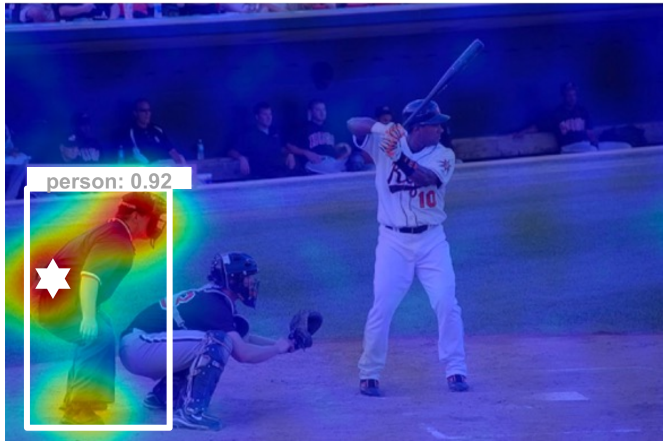
2 Subset Ranking-based Attribution — Evaluation Metrics

Location Metrics

Point Game

PG accuracy is computed by locating the **most salient coordinate** in the attribution map and recording a hit if it falls within the ground-truth object region (either a bounding box or an instance mask), after which the final score is obtained by averaging the hit indicator over all test objects.

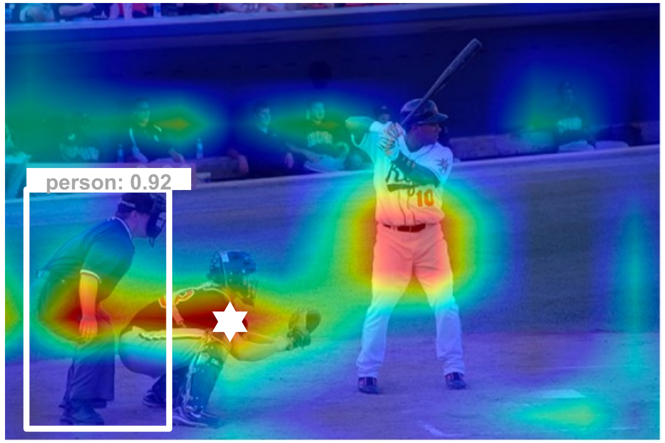
Explanation A



The most salient point is located inside the explained target object.

Better explanation under PG metric

Explanation B



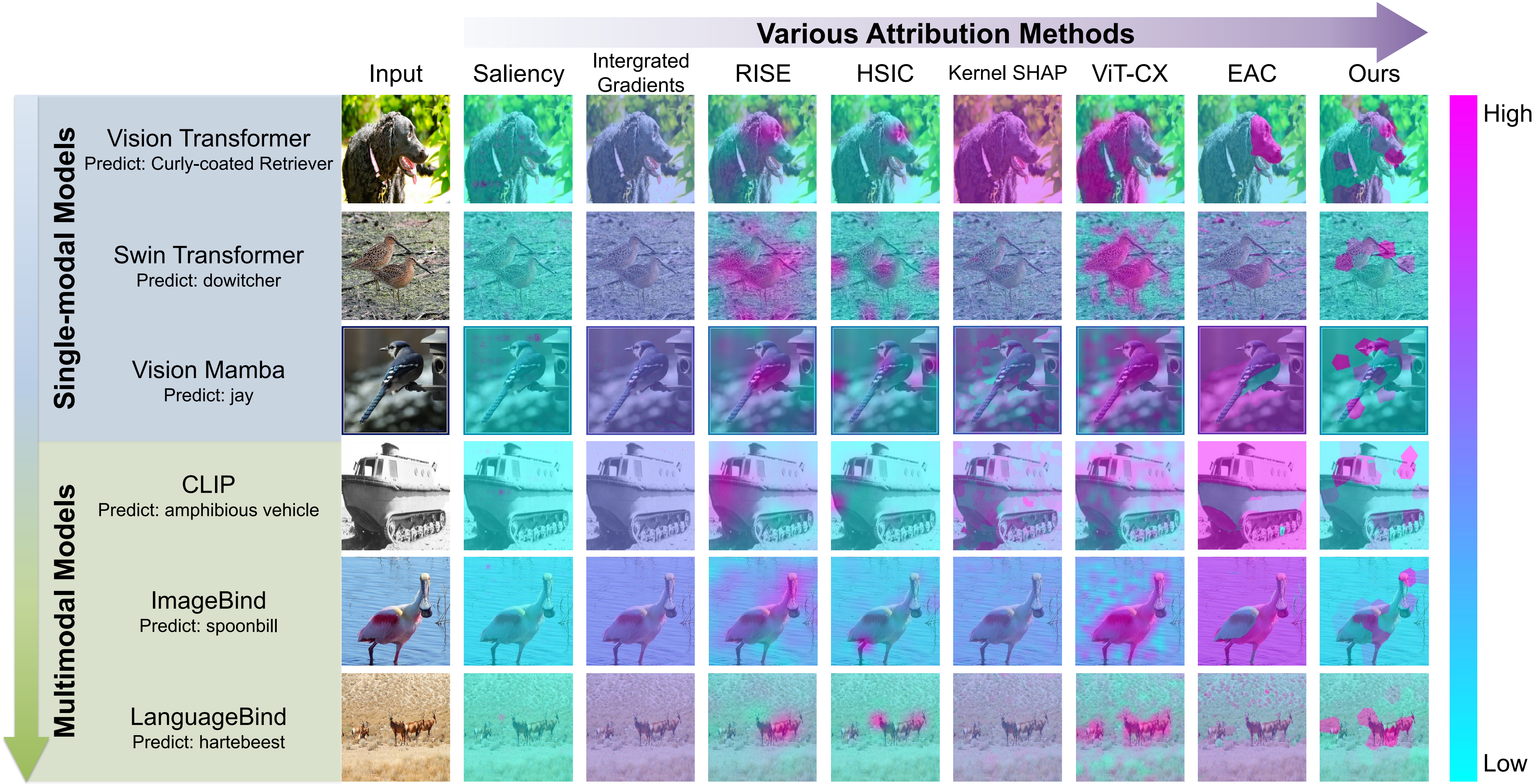
The most salient point is located outside the explained target object.

Poor explanation under PG metric

Note that this metric is only meaningful when the model achieves sufficiently strong performance and remains free of bias.

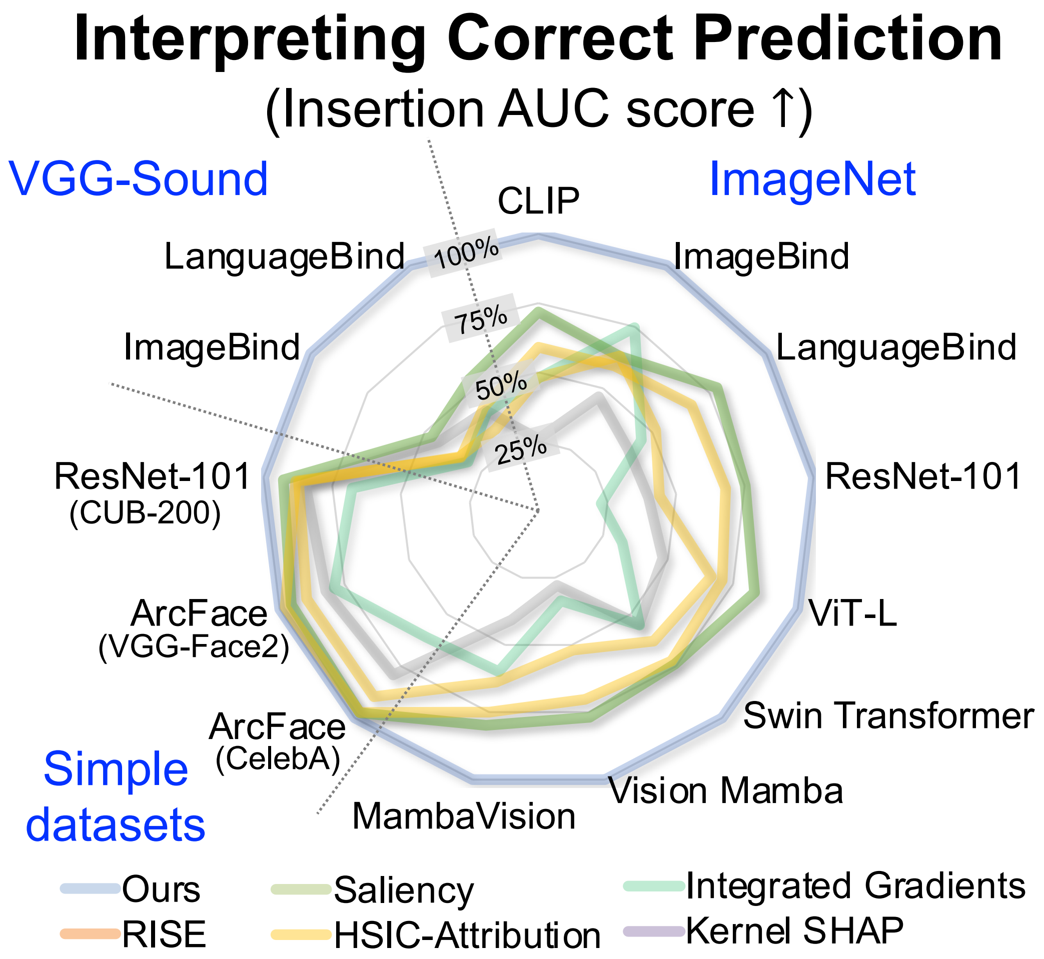
Zhang, Jianming, et al. "Top-down neural attention by excitation backprop." *International Journal of Computer Vision* 126.10 (2018): 1084-1102.

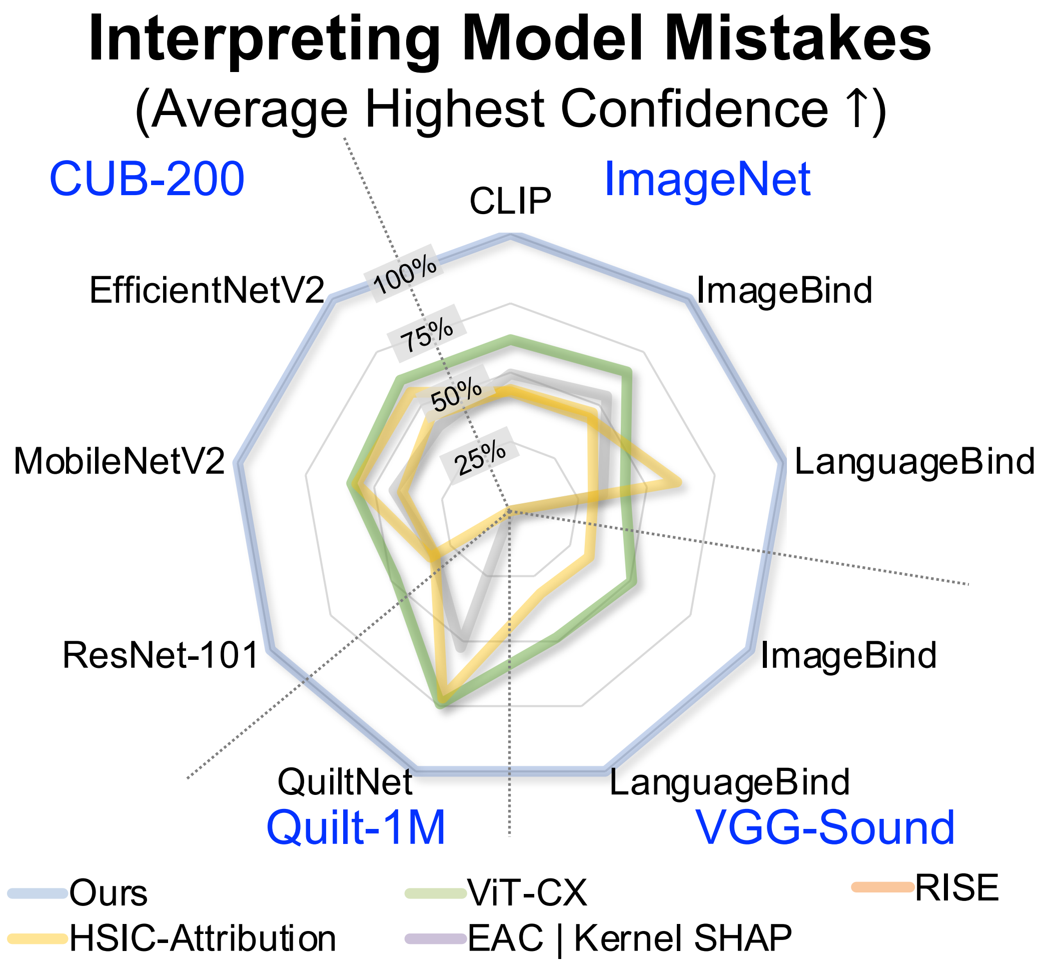
2 Subset Ranking-based Attribution — Experiments



- **Ruoyu Chen**, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection." ICLR 2024. (**Oral Presentation**, **1.16%**)
- **Ruoyu Chen**, et al. "Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection." Preprint 2025.

2 Subset Ranking-based Attribution — Experiments



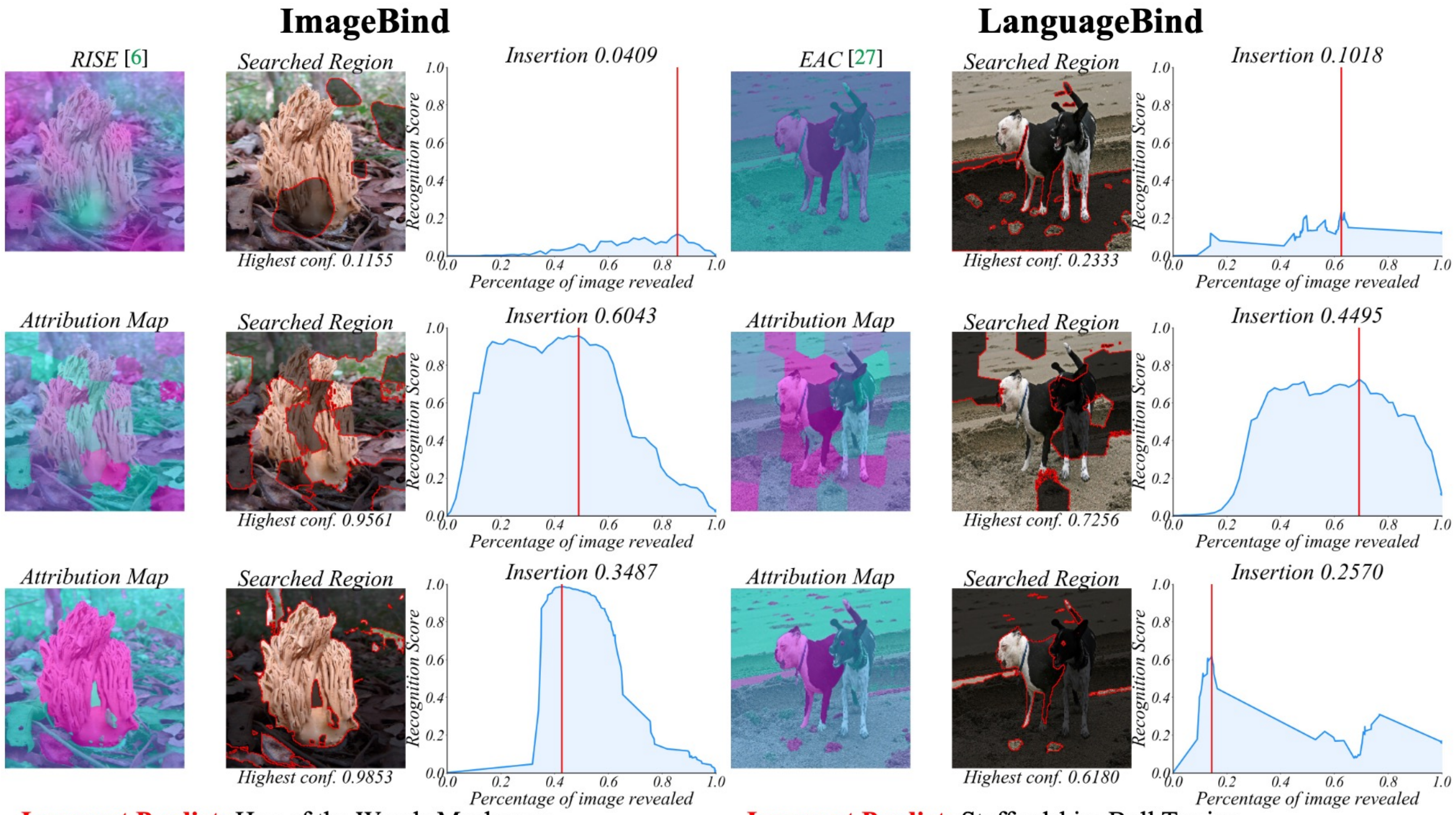


Phenomenon: The larger the model and pre-training scale, the more wrong the prediction results are, the more complex the internal interactions are, and the more difficult the attribution is.

- **Ruoyu Chen**, et al. “Less is More: Fewer Interpretable Region via Submodular Subset Selection.” ICLR 2024. (**Oral Presentation**, **1.16%**)
- **Ruoyu Chen**, et al. “Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection.” Preprint 2025.

2 Subset Ranking-based Attribution — Experiments

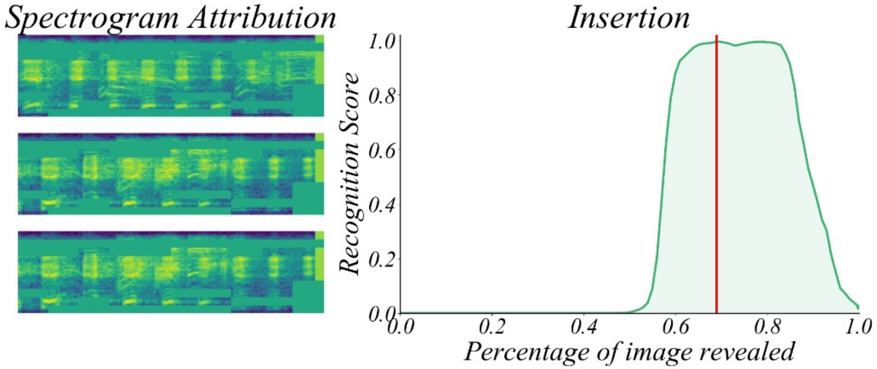
Alleviate the problem of insufficient granularity of attribution regions, thereby improving the fidelity of existing attribution algorithms (deletion\insertion) by **30.9%** and **41.7%**; discover the cause of model misprediction, and improve attribution performance (highest confidence\insertion) by **63.8%** and **127.2%**



Incorrect Predict: Hen of the Woods Mushroom
Ground Truth: Coral Fungus

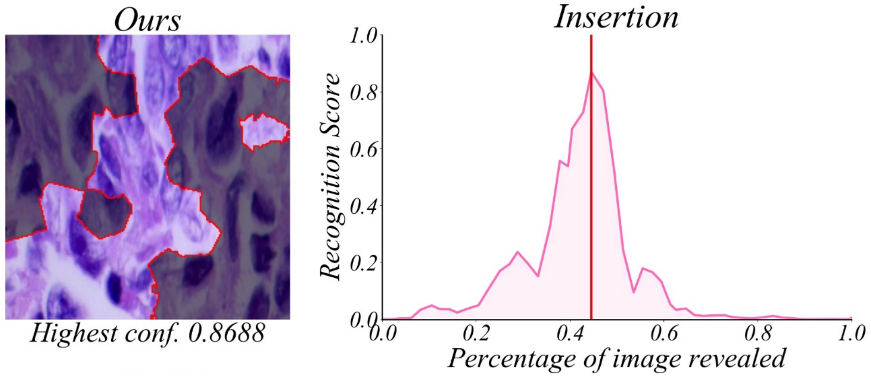
Incorrect Predict: Staffordshire Bull Terrier
Ground Truth: Boston Terrier

Natural Image Modality



Incorrect Predict: *People Belly Laughing* **Ground Truth:** *Baby Laughter*

Audio Modality



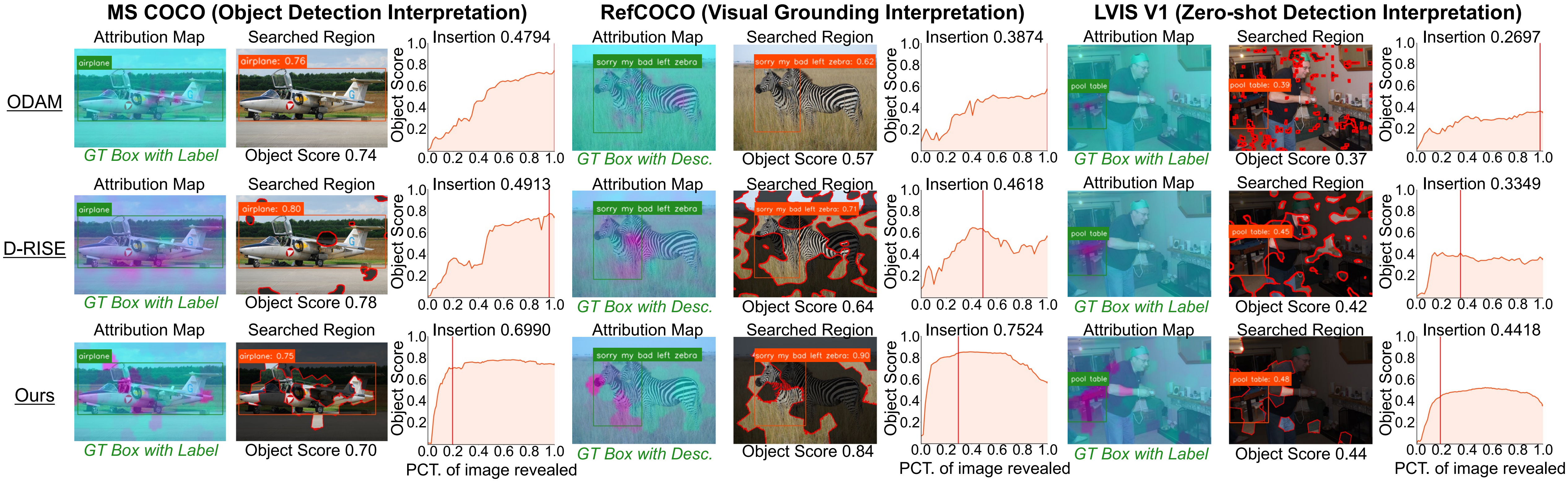
Incorrect Predict: *Benign Lung*
Ground Truth: *Lung Adenocarcinoma*

Medical Image Modality

- **Ruoyu Chen**, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection." **ICLR 2024**. (**Oral Presentation**, **1.16%**)
- **Ruoyu Chen**, et al. "Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection." Preprint 2025.

2 Subset Ranking-based Attribution — Experiments

- □ On Grounding DINO, the faithfulness of MS COCO, LVIS, and RefCOCO is improved by 23.7%, 31.6%, and 20.1%, respectively.



Ruoyu Chen, et al. "Interpreting Object-level Foundation Models via Visual Precision Search." **CVPR 2025**. (**Highlight**, **2.98%**)

2 Subset Ranking-based Attribution — Experiments

- **Explaining Failures (Hallucinations):** Explaining failure examples in visual localization and object detection tasks, outperforming existing methods on multiple evaluation metrics.

Visual Grounding Task:

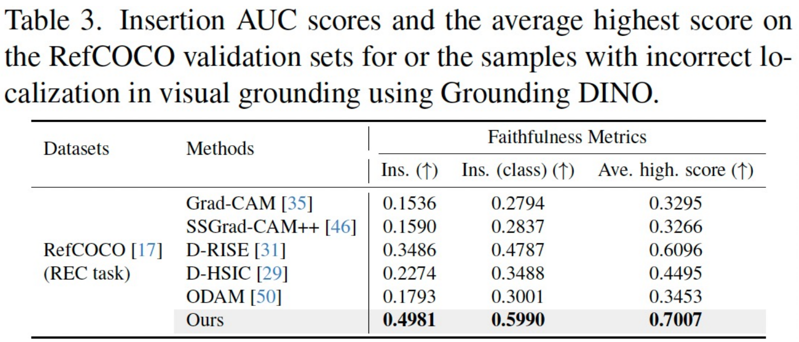
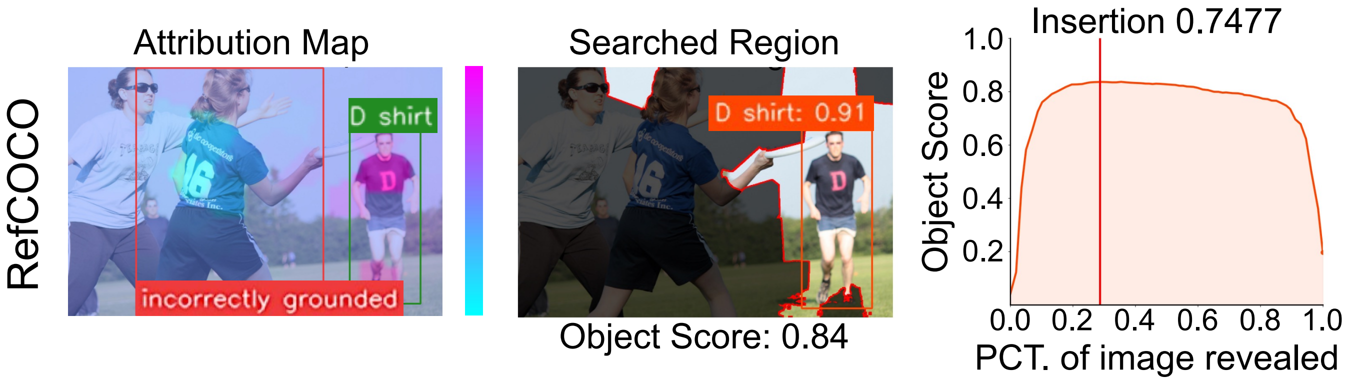


Table 3. Insertion AUC scores and the average highest score on the RefCOCO validation sets for or the samples with incorrect localization in visual grounding using Grounding DINO.

Datasets	Methods	Faithfulness Metrics		
		Ins. (↑)	Ins. (class) (↑)	Ave. high. score (↑)
RefCOCO [17] (REC task)	Grad-CAM [35]	0.1536	0.2794	0.3295
	SSGrad-CAM++ [46]	0.1590	0.2837	0.3266
	D-RISE [31]	0.3486	0.4787	0.6096
	D-HSIC [29]	0.2274	0.3488	0.4495
	ODAM [50]	0.1793	0.3001	0.3453
	Ours	**0.4981**	**0.5990**	**0.7007**



Object Detection Task:

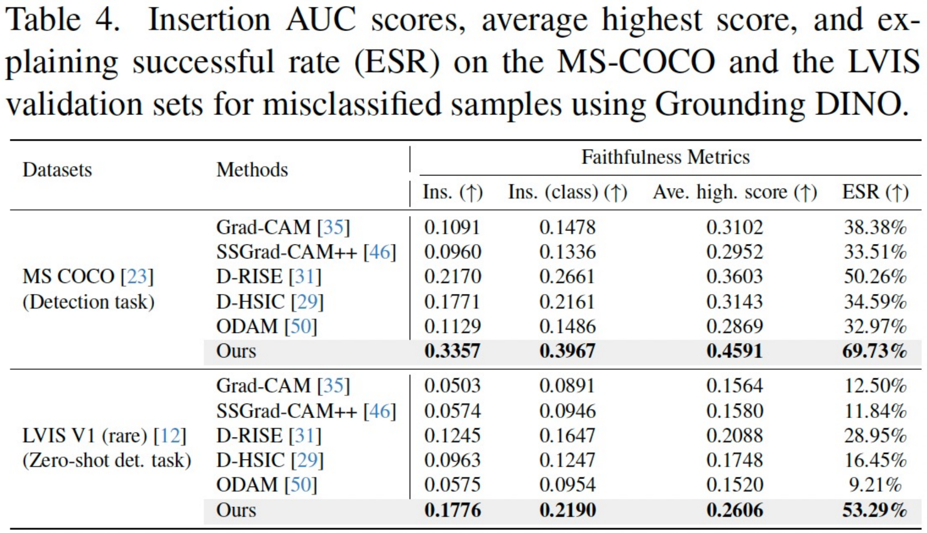
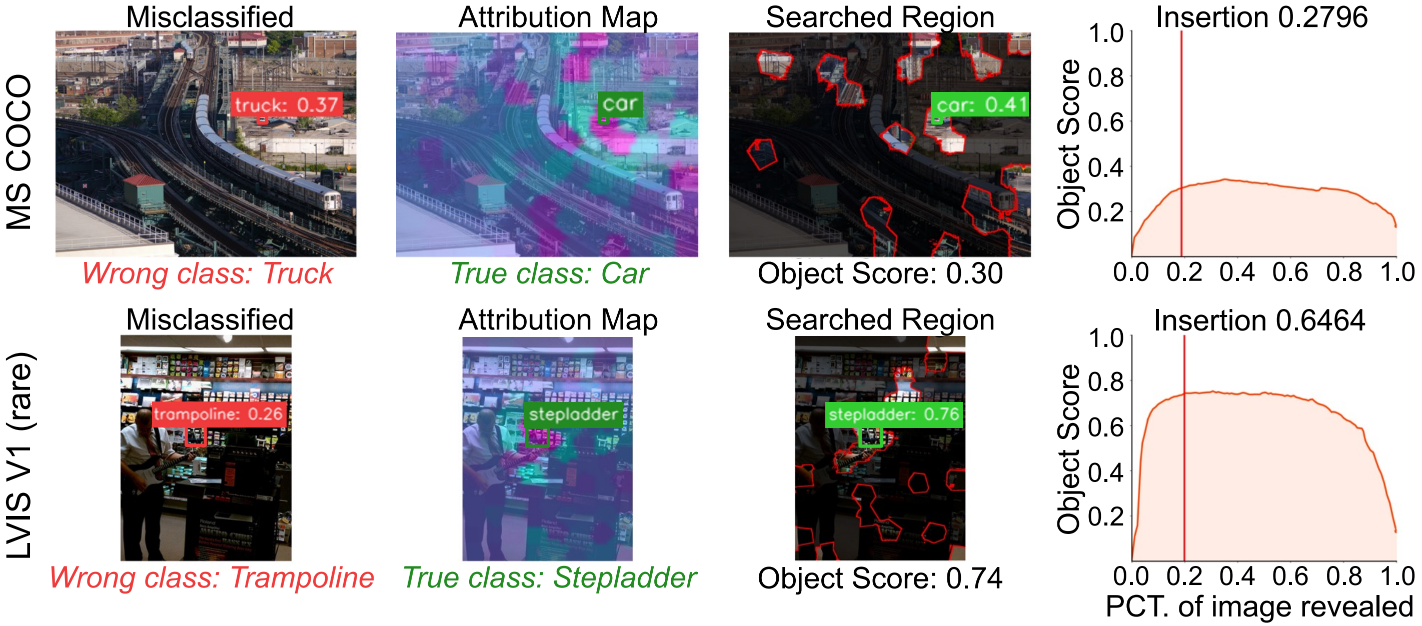


Table 4. Insertion AUC scores, average highest score, and explaining successful rate (ESR) on the MS-COCO and the LVIS validation sets for misclassified samples using Grounding DINO.

Datasets	Methods	Faithfulness Metrics			
		Ins. (↑)	Ins. (class) (↑)	Ave. high. score (↑)	ESR (↑)
MS COCO [23] (Detection task)	Grad-CAM [35]	0.1091	0.1478	0.3102	38.38%
	SSGrad-CAM++ [46]	0.0960	0.1336	0.2952	33.51%
	D-RISE [31]	0.2170	0.2661	0.3603	50.26%
	D-HSIC [29]	0.1771	0.2161	0.3143	34.59%
	ODAM [50]	0.1129	0.1486	0.2869	32.97%
	Ours	**0.3357**	**0.3967**	**0.4591**	**69.73%**
LVIS V1 (rare) [12] (Zero-shot det. task)	Grad-CAM [35]	0.0503	0.0891	0.1564	12.50%
	SSGrad-CAM++ [46]	0.0574	0.0946	0.1580	11.84%
	D-RISE [31]	0.1245	0.1647	0.2088	28.95%
	D-HSIC [29]	0.0963	0.1247	0.1748	16.45%
	ODAM [50]	0.0575	0.0954	0.1520	9.21%
	Ours	**0.1776**	**0.2190**	**0.2606**	**53.29%**

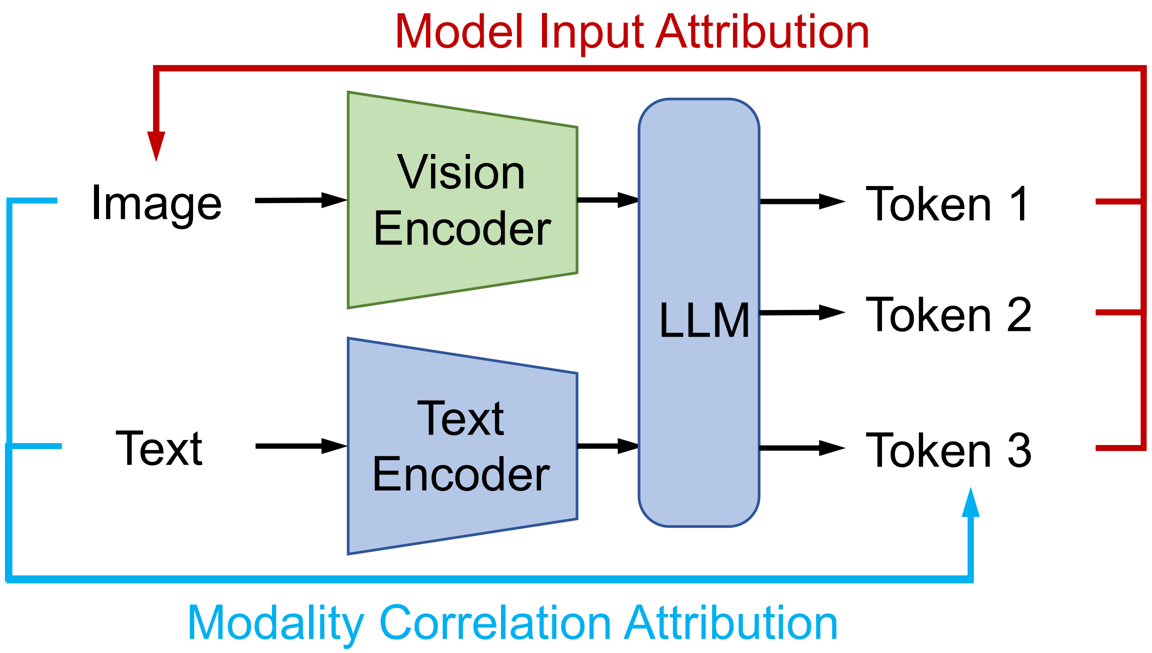


Ruoyu Chen, et al. "Interpreting Object-level Foundation Models via Visual Precision Search." **CVPR 2025**. (**Highlight**, **2.98%**)

2 Explaining Autoregressive MLLM

Question 1: Where MLLMs Attend?

Question 2: What They Rely On

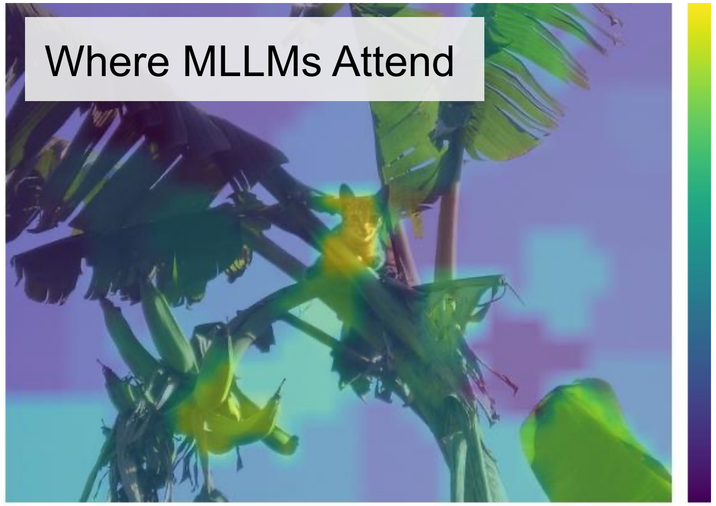


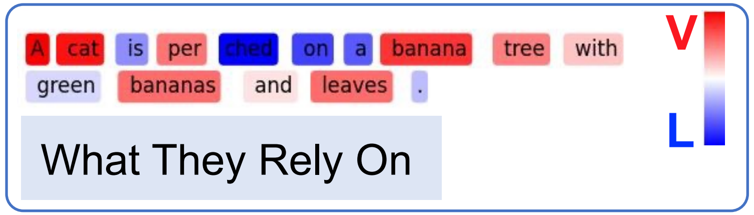
Answer: A cat is perched on a banana tree with green bananas and leaves.

Multimodal Large Language Models

Input Image

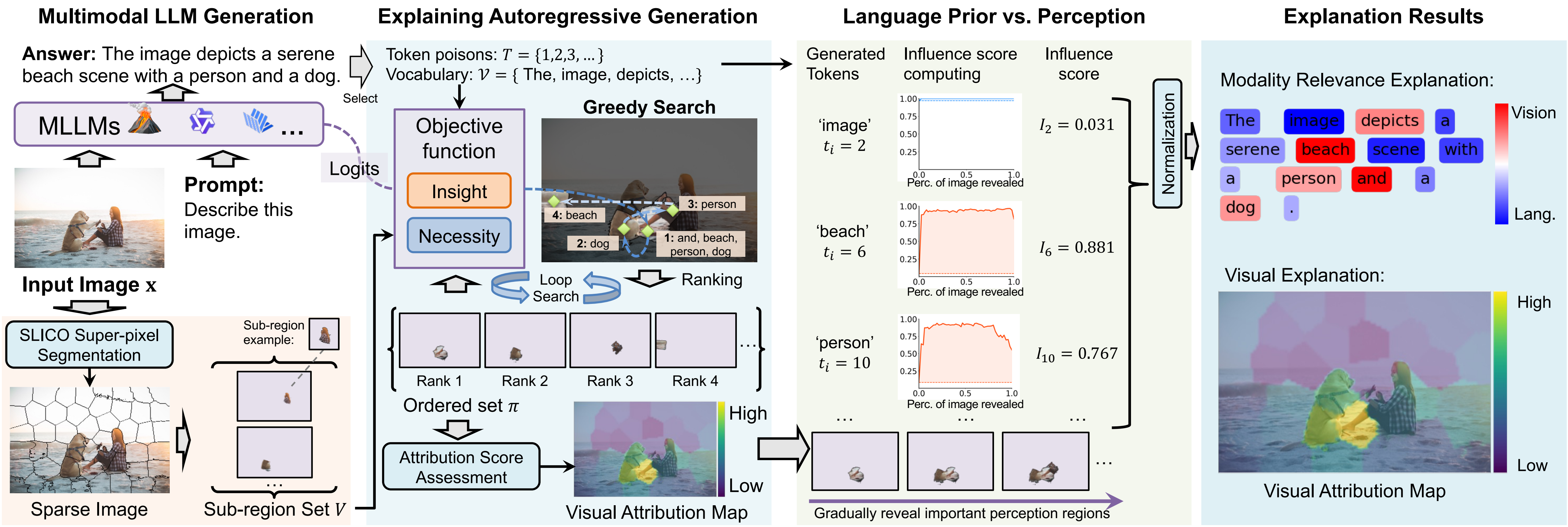
Prompt: Describe this image.





Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

2 Explaining Autoregressive MLLM — Method



Multimodal LLM Generation

Answer: The image depicts a serene beach scene with a person and a dog.

MLLMs ...

Prompt: Describe this image.

Input Image $\mathbf{x}$

SLICO Super-pixel Segmentation

Sub-region example:

Sparse Image

Sub-region Set V

Explaining Autoregressive Generation

Token poisons: $T = \{1,2,3,\ldots\}$

Vocabulary: $\mathcal{V} = \{$ The, image, depicts, …$\}$

Objective function

Insight

Necessity

Greedy Search

Loop Search

Ranking

Ordered set π

Attribution Score Assessment

Visual Attribution Map

Language Prior vs. Perception

Generated Tokens

Influence score computing

Influence score

‘image’ $t_i = 2$ — $I_2 = 0.031$

‘beach’ $t_i = 6$ — $I_6 = 0.881$

‘person’ $t_i = 10$ — $I_{10} = 0.767$

… … …

Normalization

Gradually reveal important perception regions

Explanation Results

Modality Relevance Explanation:

The image depicts a serene beach scene with a person and a dog .

Visual Explanation:

Visual Attribution Map

Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” Preprint 2025.

2 Explaining Autoregressive MLLM — Sentence-level Explanation

Image Caption Interpretation

LLaVA-1.5

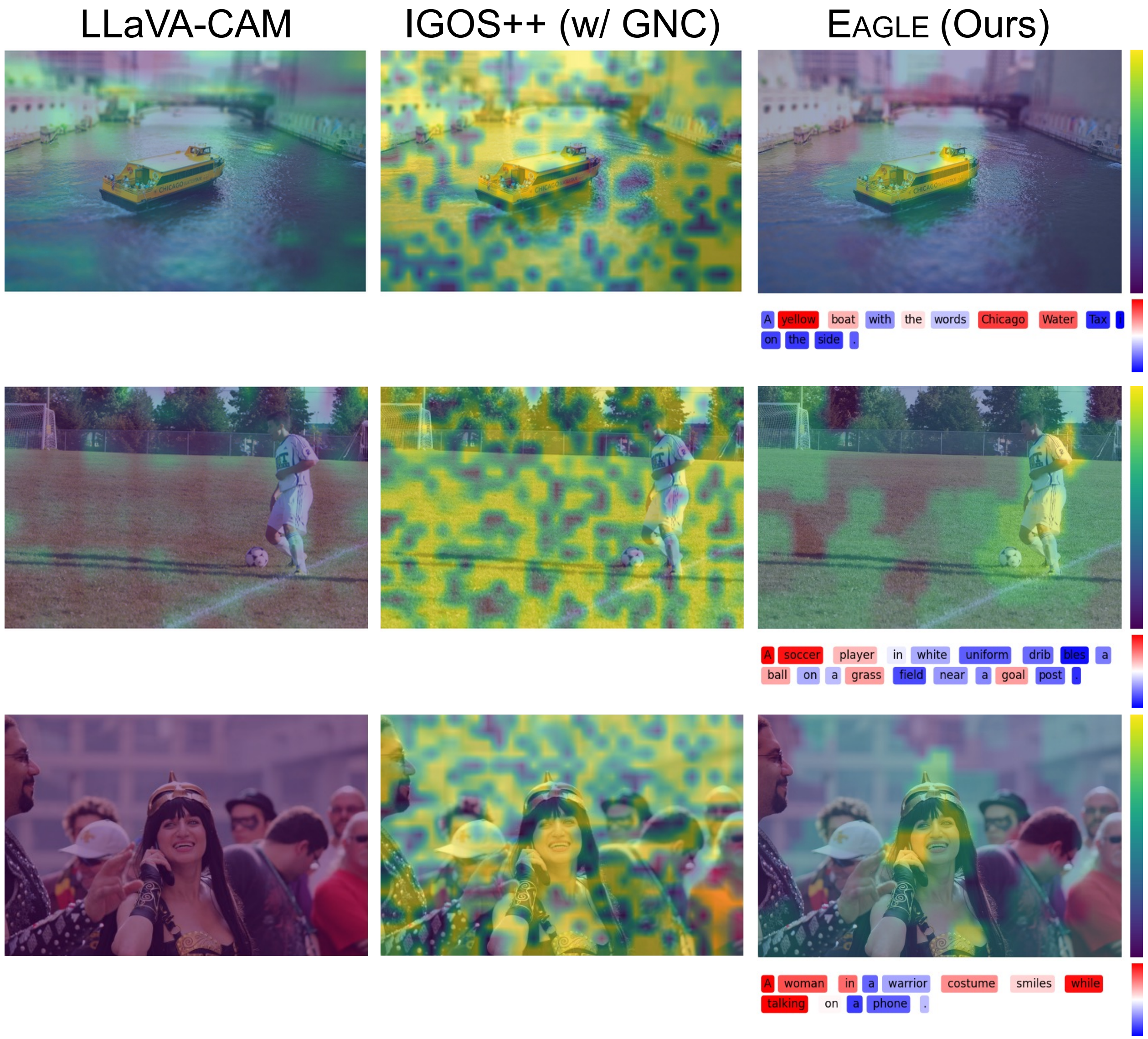
Captioning: A yellow boat with the words Chicago Water Taxi on the side.

Qwen2.5-VL

Captioning: A soccer player in white uniform dribbles a ball on a grass field near a goalpost.

InternVL3.5

Captioning: A woman in a warrior costume smiles while talking on a phone.


LLaVA-CAM
IGOS++ (w/ GNC)
EAGLE (Ours)
A yellow boat with the words Chicago Water Tax i on the side .
A soccer player in white uniform drib bles a ball on a grass field near a goal post .
A woman in a warrior costume smiles while talking on a phone .

Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

2 Explaining Autoregressive MLLM — Sentence-level Explanation

Visual Question Answering Interpretation

LLaVA-1.5 7B

Question: Is the door of the truck cab open?

Answering: Yes, the door of the truck cab is open.

Qwen2.5-VL 7B

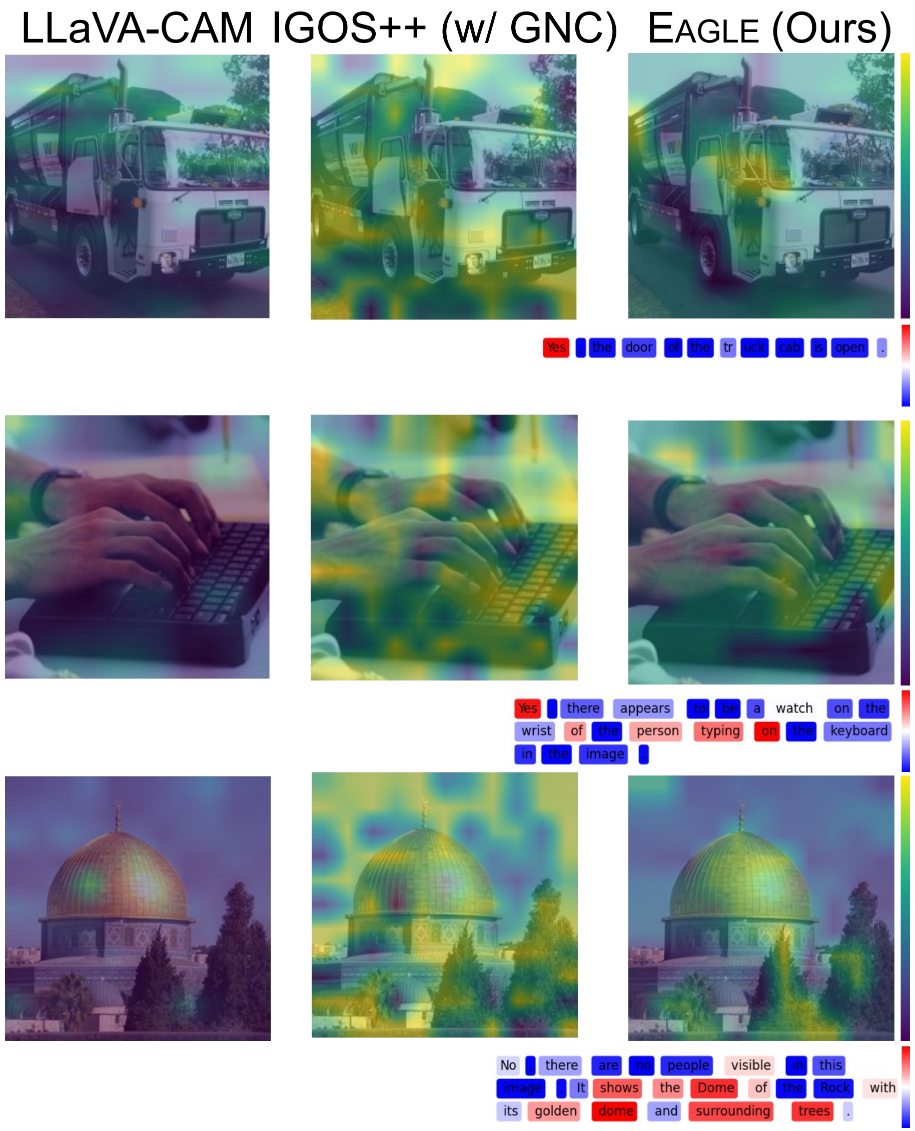
Question: Is there any accessory on the wrists?

Answering: Yes, there appears to be a watch on the wrist of the person typing on the keyboard in the image.

InternVL3.5 4B

Question: Can you see people in this image?

Answering: No, there are no people visible in this image. It shows the Dome of the Rock with its golden dome and surrounding trees.



Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

2 Explaining Autoregressive MLLM — Word-level Explanation

LLaVA-1.5 7B

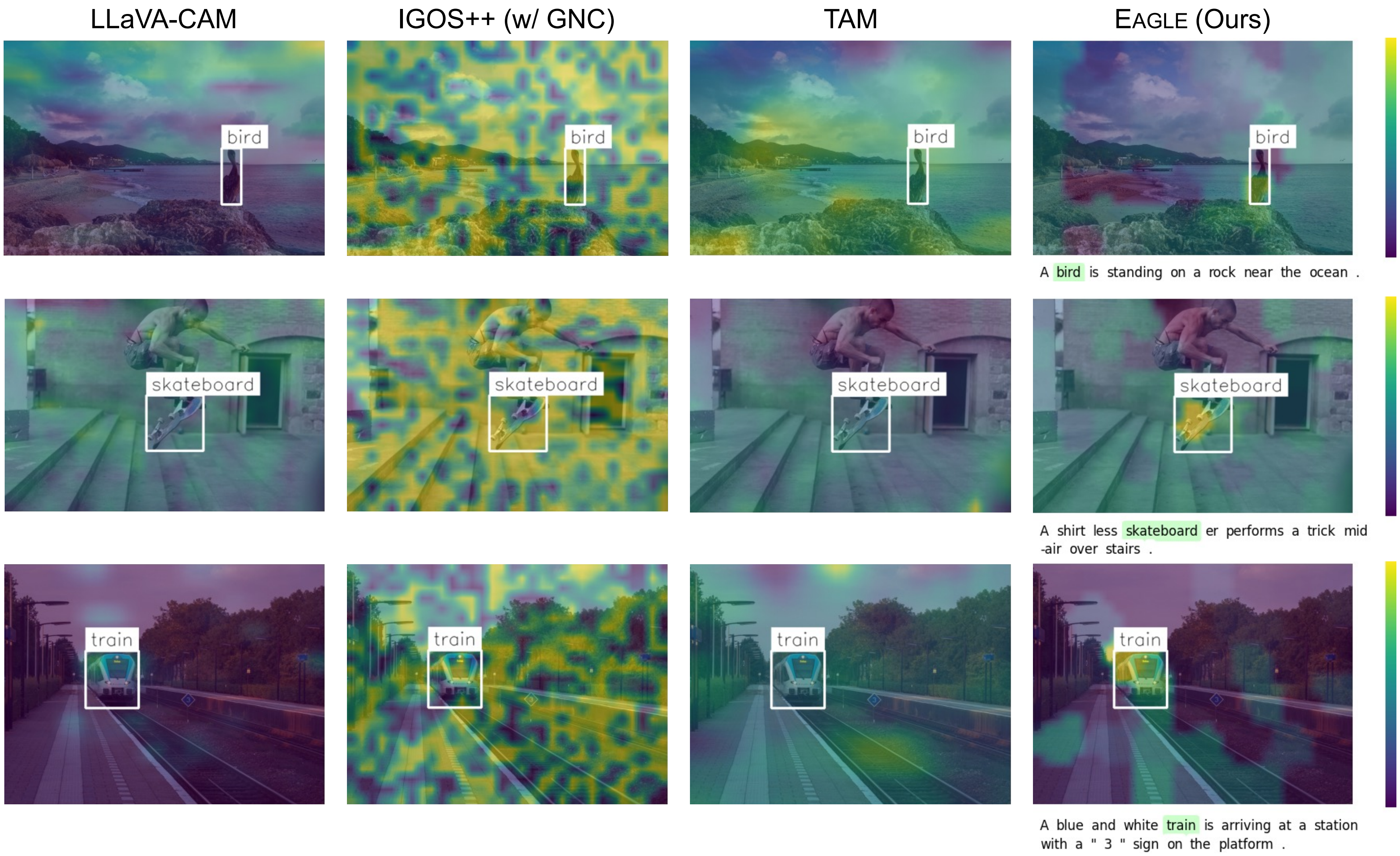
Captioning: A bird is standing on a rock near the ocean.

Qwen2.5-VL 7B

Captioning: A shirtless skateboarder performs a trick mid-air over stairs.

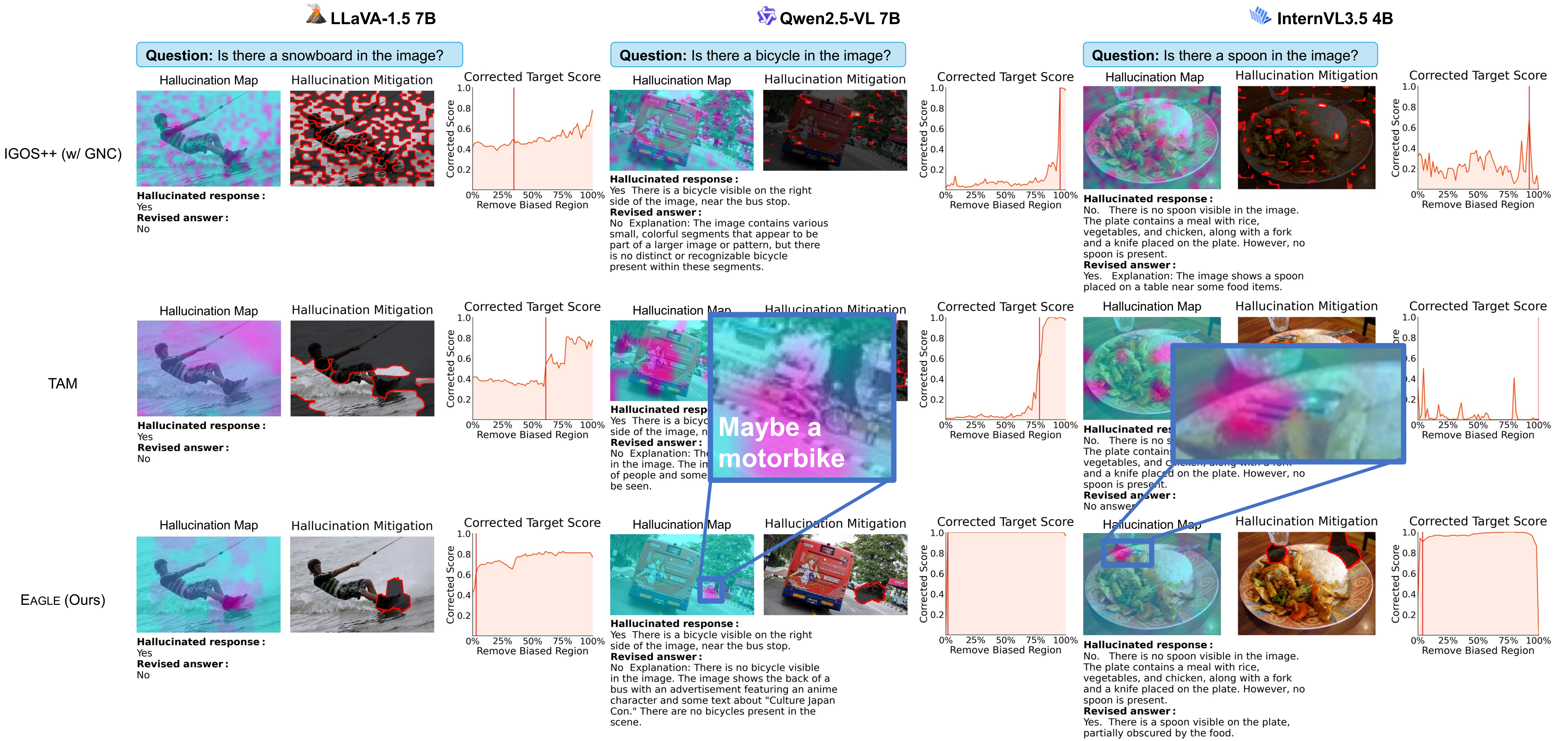
InternVL3.5 4B

Captioning: A blue and white train is arriving at a station with a \"3\" sign on the platform.



Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

2 Explaining Autoregressive MLLM — Explaining Hallucination



LLaVA-1.5 7B

Question: Is there a snowboard in the image?

IGOS++ (w/ GNC)

Hallucinated response :
Yes
Revised answer :
No

TAM

Hallucinated response :
Yes
Revised answer :
No

EAGLE (Ours)

Hallucinated response :
Yes
Revised answer :
No

Qwen2.5-VL 7B

Question: Is there a bicycle in the image?

IGOS++ (w/ GNC)

Hallucinated response :
Yes There is a bicycle visible on the right side of the image, near the bus stop.
Revised answer :
No Explanation: The image contains various small, colorful segments that appear to be part of a larger image or pattern, but there is no distinct or recognizable bicycle present within these segments.

TAM

Hallucinated resp
Yes There is a bicyc
side of the image, n
Revised answer :
No Explanation: The
in the image. The in
of people and some
be seen.

EAGLE (Ours)

Hallucinated response :
Yes There is a bicycle visible on the right side of the image, near the bus stop.
Revised answer :
No Explanation: There is no bicycle visible in the image. The image shows the back of a bus with an advertisement featuring an anime character and some text about "Culture Japan Con." There are no bicycles present in the scene.

InternVL3.5 4B

Question: Is there a spoon in the image?

IGOS++ (w/ GNC)

Hallucinated response :
No. There is no spoon visible in the image. The plate contains a meal with rice, vegetables, and chicken, along with a fork and a knife placed on the plate. However, no spoon is present.
Revised answer :
Yes. Explanation: The image shows a spoon placed on a table near some food items.

TAM

Hallucinated res
No. There is no s
The plate contain
vegetables, and c
and a knife placed on the plate. However, no spoon is present.
Revised answer :
No answer

EAGLE (Ours)

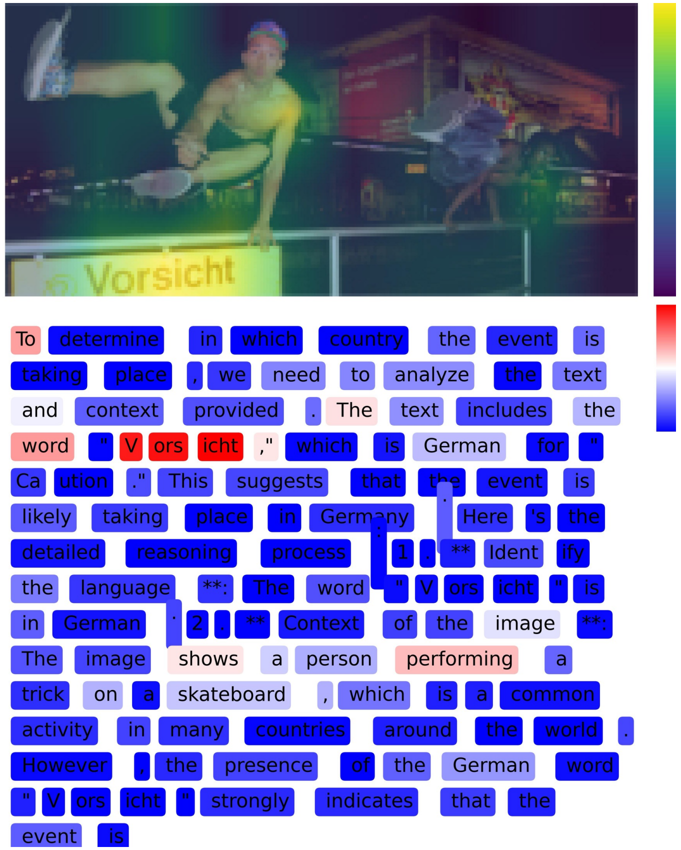
Hallucinated response :
No. There is no spoon visible in the image. The plate contains a meal with rice, vegetables, and chicken, along with a fork and a knife placed on the plate. However, no spoon is present.
Revised answer :
Yes. There is a spoon visible on the plate, partially obscured by the food.

Ruoyu Chen, et al. "Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation." **Preprint 2025**.

2 Explaining Autoregressive MLLM — More Tasks

Long-horizon reasoning explanation

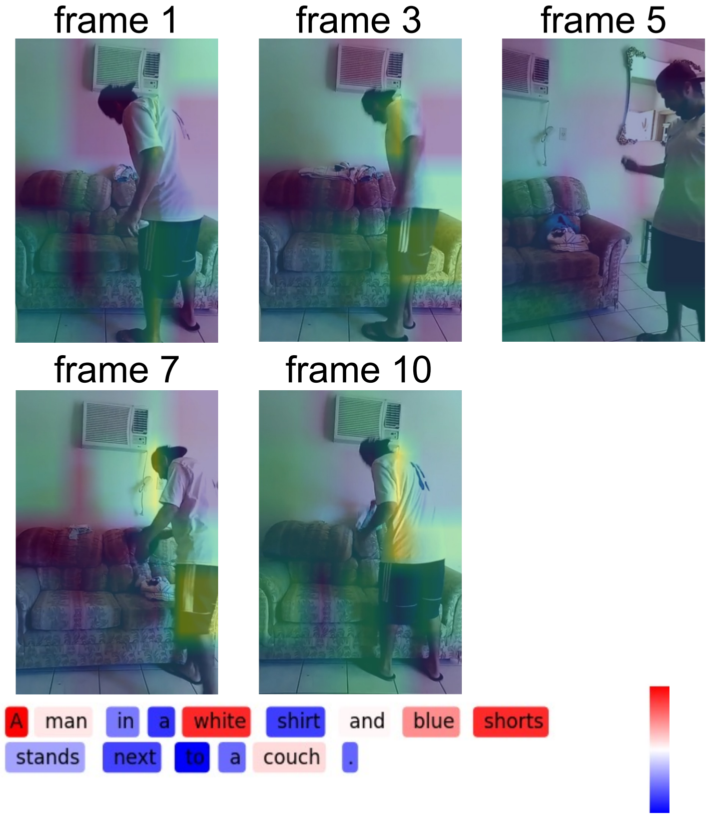
Q: In which country is this event taking place? Give me the detailed reasoning process.



To determine in which country the event is taking place, we need to analyze the text and context provided. The text includes the word "Vorsicht," which is German for "Caution." This suggests that the event is likely taking place in Germany. Here's the detailed reasoning process: 1. **Identify the language**: The word "Vorsicht" is in German. 2. **Context of the image**: The image shows a person performing a trick on a skateboard, which is a common activity in many countries around the world. However, the presence of the German word "Vorsicht" strongly indicates that the event is

Video-based explanation

Q: Describe the image in one factual English sentence of no more than 20 words. Do not include information that is not clearly visible.



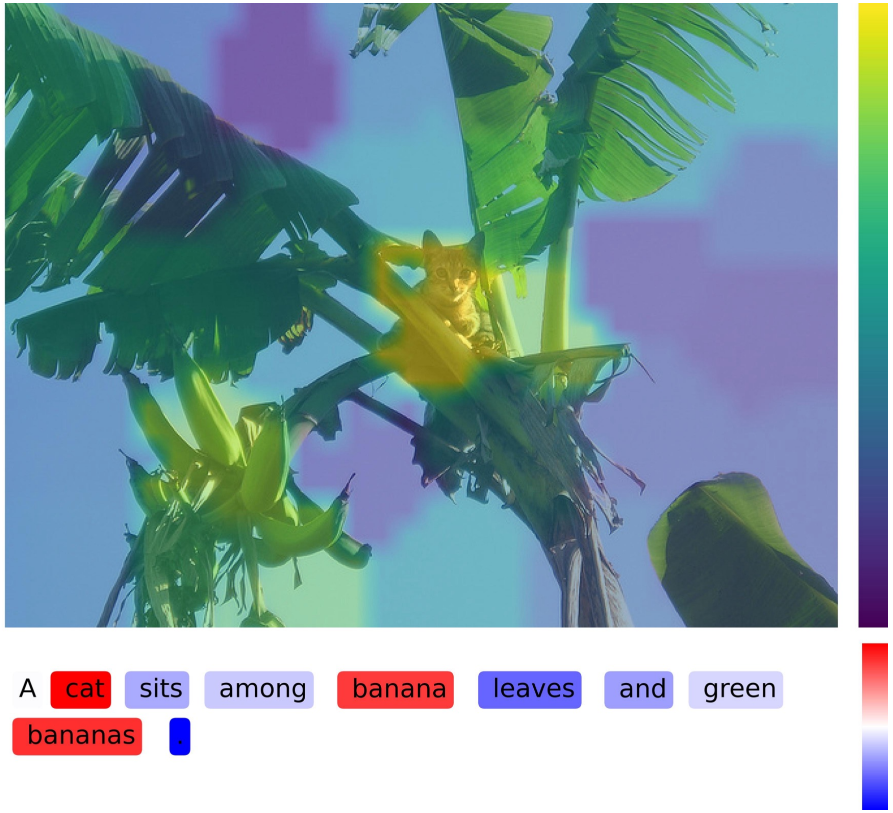
A man in a white shirt and blue shorts stands next to a couch .

Ruoyu Chen, et al. "Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation." **Preprint 2025**.

2 Explaining Autoregressive MLLM — API Explanation

GPT-5.2

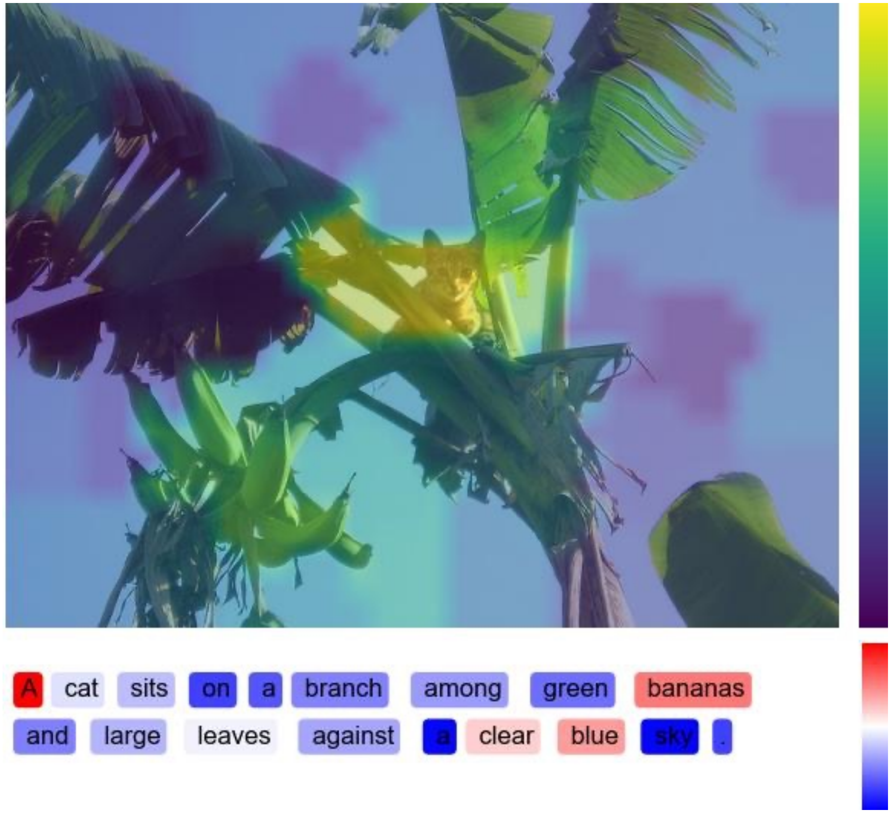
Q: Describe the image in one factual English sentence of no more than 10 words. Do not include information that is not clearly visible.



A cat sits among banana leaves and green bananas .

Qwen 2.5 VL

Q: Describe the image in one factual English sentence of no more than 20 words. Do not include information that is not clearly visible.

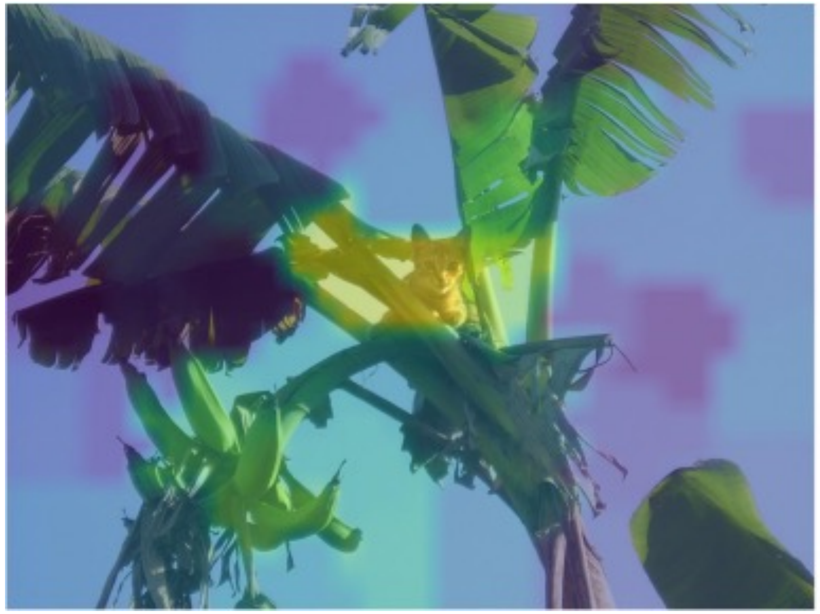


A cat sits on a branch among green bananas and large leaves against a clear blue sky .

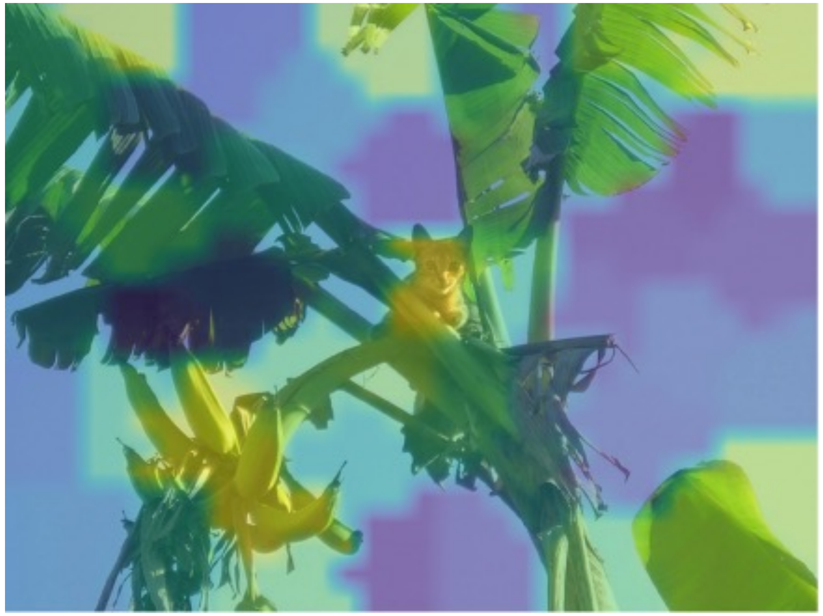
Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

2 Explaining Autoregressive MLLM — Acceleration

Q: Describe the image in one factual English sentence of no more than 20 words. Do not include information that is not clearly visible
A: A cat sits on a branch among green bananas and large leaves against a clear blue sky.



Greedy Search: 514.5 s

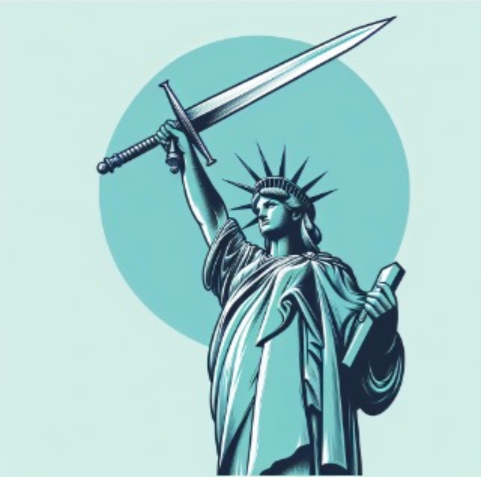


Early Stop (w/ Sliding window): 25.6 s

Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

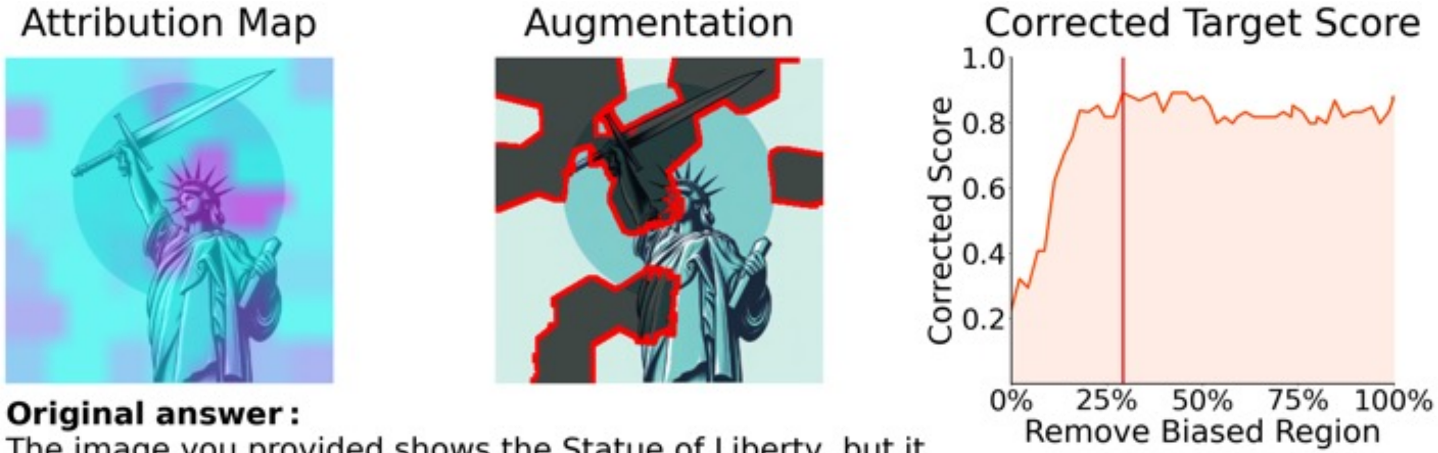
2 Explaining Autoregressive MLLM — Shortcut Discovery

Q: The Statue of Liberty is holding a torch?



Original Image

Counterfactual Explaining



Original answer :

✓ The image you provided shows the Statue of Liberty, but it does not depict her holding a torch. Instead, she is holding a sword in one hand and a tablet in the other. The statue is a symbol of freedom and democracy, representing the ideals of the United States.

Augmented answer :

✗ The image you provided shows the Statue of Liberty holding a torch, which is indeed one of her most iconic symbols. The torch represents enlightenment and freedom, symbolizing the ideals of liberty and democracy that the statue stands for.

Ruoyu Chen, et al. “Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation.” **Preprint 2025**.

3 Attribution-guided Model Training Enhancement

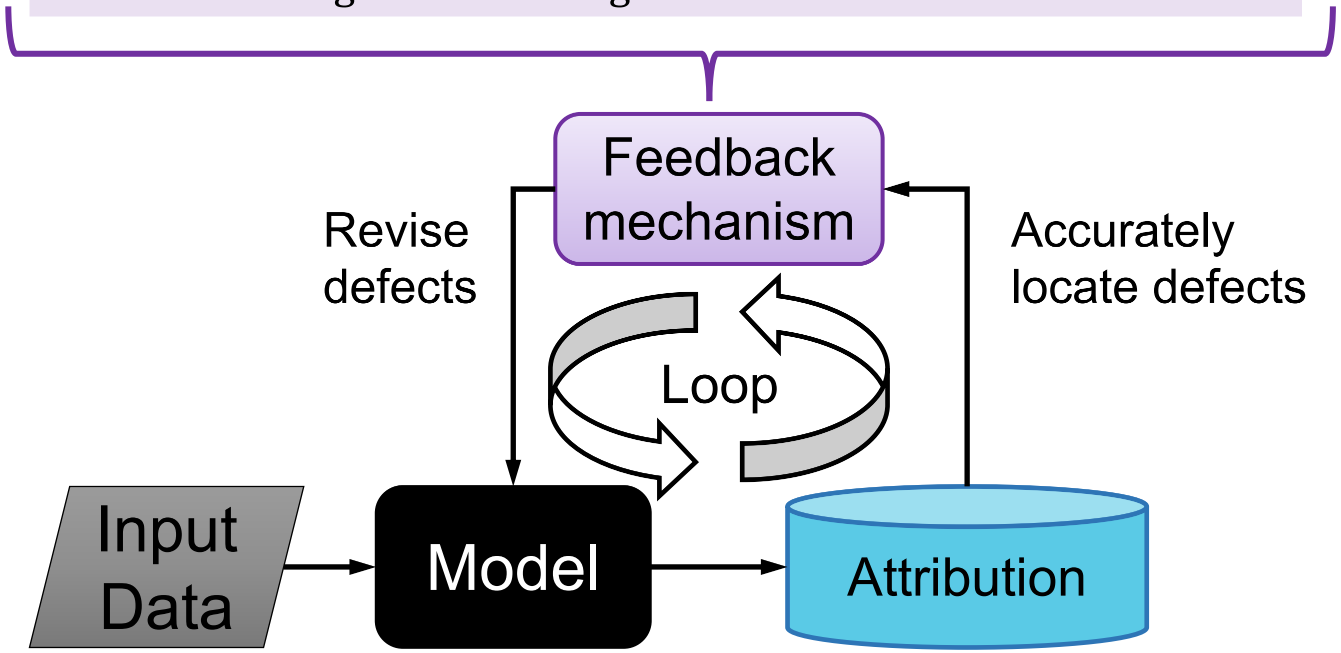
Conceptual:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\text{task}}(f_\theta(X), Y)}_{\text{task supervision}} + \underbrace{\lambda_1 \mathcal{L}_{\text{human}}(\mathcal{A}(f_\theta(X_i), Y), H_i)}_{\text{human prior alignment}} + \underbrace{\lambda_2 \mathcal{L}_{\text{re}}\big(\mathcal{A}(f_\theta(X_i), Y)\big)}_{\text{attribution regularization}}$$

$$+ \underbrace{\lambda_3 \varepsilon \mathcal{L}_{\text{task}}\big(f_\theta\big(T_{\text{Au}}(\mathcal{A}, X_i, Y\,)\big), Y\big)}_{\text{attribution−guided data augmentation}}$$

Attribution rationality enhancement

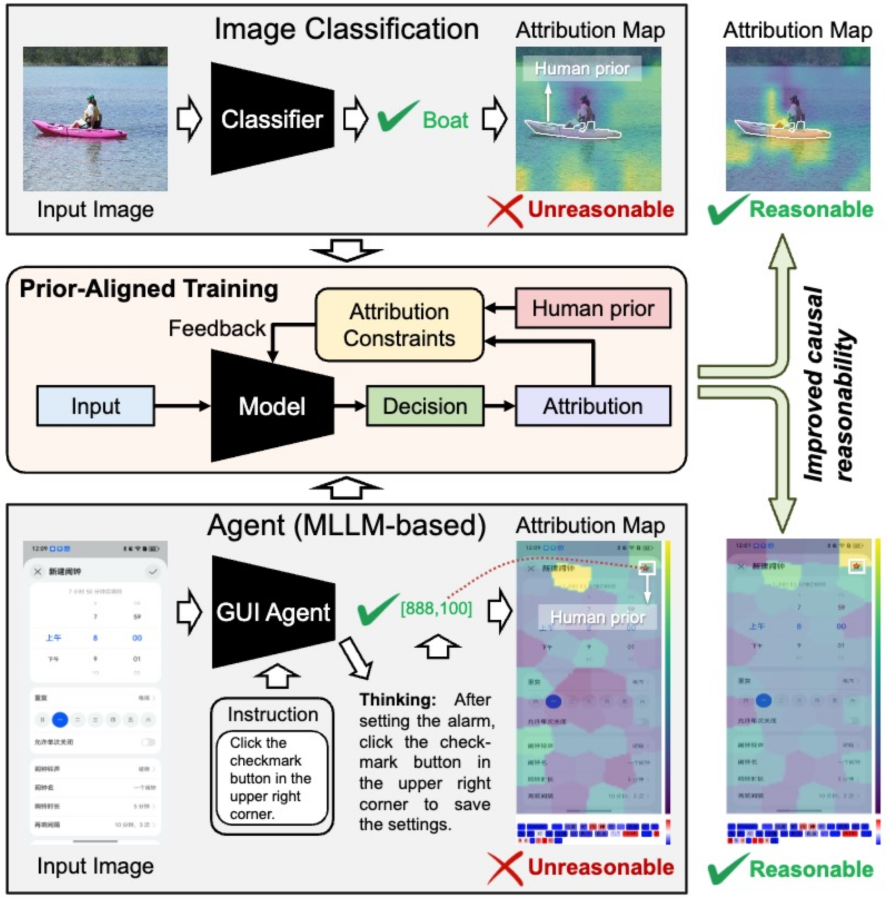
Model performance enhancement



Attribution-guided model training enhancement, using attribution methods to guide model training, so as to improve the rationality of model attribution or improve model performance.

3 Prior-Aligned Training with Attribution Constraints

- Reliable models should not only predict correctly, but also justify decisions with acceptable evidence.
- The causal reasonableness of model behavior can be regulated through constraints induced by human prior knowledge.



Click the checkmark button in the upper right corner.

Thinking: After setting the alarm, click the check-mark button in the upper right corner to save the settings.

Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

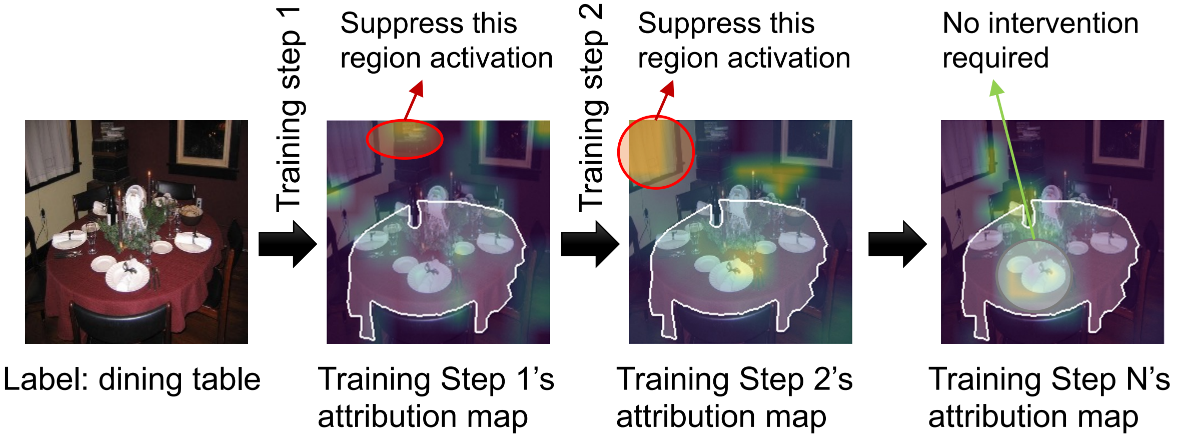
3 Prior-Aligned Training with Attribution Constraints

- When attribution methods are sufficiently faithful, they can be used to assess whether model decisions align with human cognition.
- When model decisions conflict with human common sense or perception, attribution helps identify and suppress untrustworthy predictions.

A training batch contains b samples.

$$\mathcal{L}_{\text{human}} = \sum_{i=1}^{b} \underbrace{\mathcal{F}(s_i)}_{\text{attribution value of the most important subregion}} \cdot \underbrace{I(s_i \in H_i)}_{\text{a binary indicator denoting whether } s_i \text{ is contained in the human-prior region}}$$

Physical interpretation: For each training sample i, no penalty is applied when the most important attribution region $\mathrm{s_i}$ lies within the human-prior region. Otherwise, its explanatory contribution is constrained by suppressing the corresponding submodular value $\mathcal{F}(\mathrm{s_i})$.


Training step 1
Suppress this region activation
Training step 2
Suppress this region activation
No intervention required
Label: dining table
Training Step 1's attribution map
Training Step 2's attribution map
Training Step N's attribution map

If the most important attribution region is outside the human prior, its activation is iteratively suppressed while other regions remain unchanged; once it falls within the prior region, no further constraint is applied.

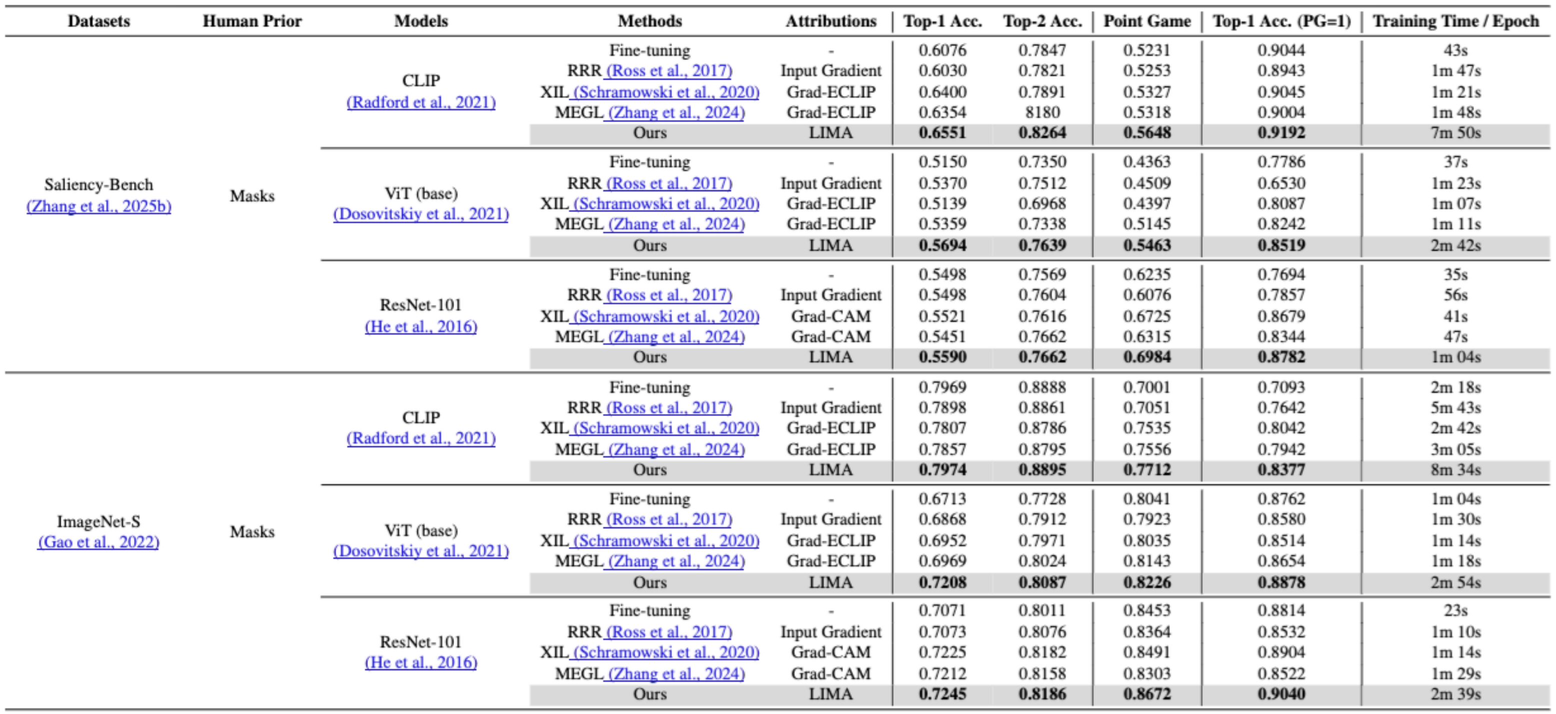
Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

3 Prior-Aligned Training with Attribution Constraints

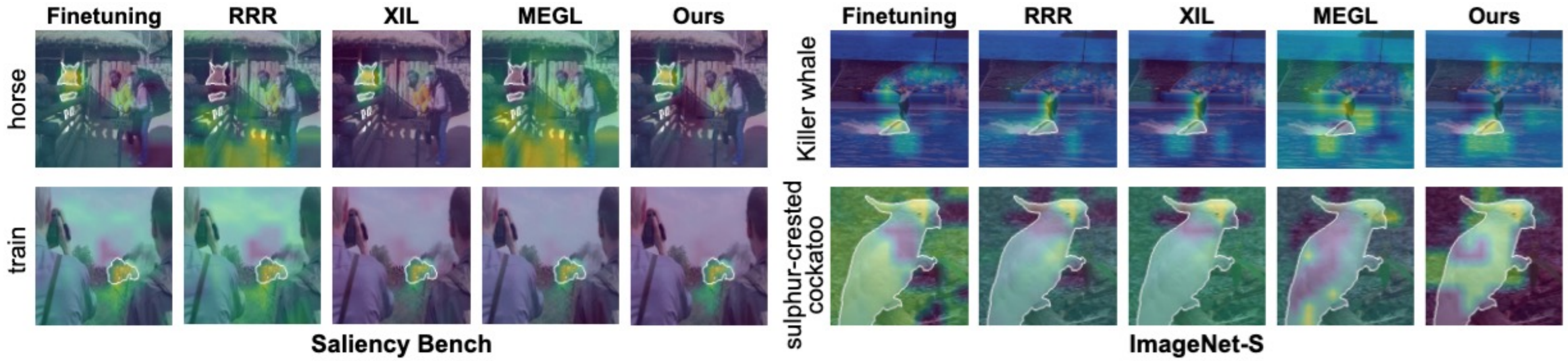
Evaluation on Image Classification

- ➢ Our method improves model performance.
- ➢ It also enhances the causal reasonableness of model decisions.

Table 1. Evaluation of attribution-based prior alignment methods for image classification models on the Saliency-Bench and ImageNet-S datasets. Both model performance (accuracy) and decision rationality are reported, with rationality measured by the Point Game and accuracy conditioned on successful Point Game outcomes.



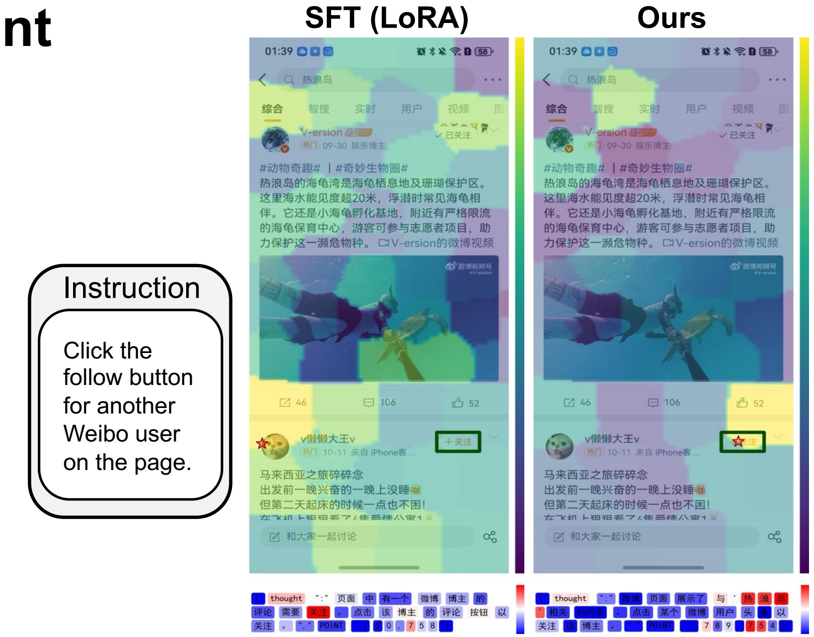
Datasets	Human Prior	Models	Methods	Attributions	Top-1 Acc.	Top-2 Acc.	Point Game	Top-1 Acc. (PG=1)	Training Time / Epoch
Saliency-Bench (Zhang et al., 2025b)	Masks	CLIP (Radford et al., 2021)	Fine-tuning	-	0.6076	0.7847	0.5231	0.9044	43s
			RRR (Ross et al., 2017)	Input Gradient	0.6030	0.7821	0.5253	0.8943	1m 47s
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.6400	0.7891	0.5327	0.9045	1m 21s
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.6354	8180	0.5318	0.9004	1m 48s
			Ours	LIMA	**0.6551**	**0.8264**	**0.5648**	**0.9192**	7m 50s
		ViT (base) (Dosovitskiy et al., 2021)	Fine-tuning	-	0.5150	0.7350	0.4363	0.7786	37s
			RRR (Ross et al., 2017)	Input Gradient	0.5370	0.7512	0.4509	0.6530	1m 23s
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.5139	0.6968	0.4397	0.8087	1m 07s
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.5359	0.7338	0.5145	0.8242	1m 11s
			Ours	LIMA	**0.5694**	**0.7639**	**0.5463**	**0.8519**	2m 42s
		ResNet-101 (He et al., 2016)	Fine-tuning	-	0.5498	0.7569	0.6235	0.7694	35s
			RRR (Ross et al., 2017)	Input Gradient	0.5498	0.7604	0.6076	0.7857	56s
			XIL (Schramowski et al., 2020)	Grad-CAM	0.5521	0.7616	0.6725	0.8679	41s
			MEGL (Zhang et al., 2024)	Grad-CAM	0.5451	0.7662	0.6315	0.8344	47s
			Ours	LIMA	**0.5590**	**0.7662**	**0.6984**	**0.8782**	1m 04s
ImageNet-S (Gao et al., 2022)	Masks	CLIP (Radford et al., 2021)	Fine-tuning	-	0.7969	0.8888	0.7001	0.7093	2m 18s
			RRR (Ross et al., 2017)	Input Gradient	0.7898	0.8861	0.7051	0.7642	5m 43s
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.7807	0.8786	0.7535	0.8042	2m 42s
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.7857	0.8795	0.7556	0.7942	3m 05s
			Ours	LIMA	**0.7974**	**0.8895**	**0.7712**	**0.8377**	8m 34s
		ViT (base) (Dosovitskiy et al., 2021)	Fine-tuning	-	0.6713	0.7728	0.8041	0.8762	1m 04s
			RRR (Ross et al., 2017)	Input Gradient	0.6868	0.7912	0.7923	0.8580	1m 30s
			XIL (Schramowski et al., 2020)	Grad-ECLIP	0.6952	0.7971	0.8035	0.8514	1m 14s
			MEGL (Zhang et al., 2024)	Grad-ECLIP	0.6969	0.8024	0.8143	0.8654	1m 18s
			Ours	LIMA	**0.7208**	**0.8087**	**0.8226**	**0.8878**	2m 54s
		ResNet-101 (He et al., 2016)	Fine-tuning	-	0.7071	0.8011	0.8453	0.8814	23s
			RRR (Ross et al., 2017)	Input Gradient	0.7073	0.8076	0.8364	0.8532	1m 10s
			XIL (Schramowski et al., 2020)	Grad-CAM	0.7225	0.8182	0.8491	0.8904	1m 14s
			MEGL (Zhang et al., 2024)	Grad-CAM	0.7212	0.8158	0.8303	0.8522	1m 29s
			Ours	LIMA	**0.7245**	**0.8186**	**0.8672**	**0.9040**	2m 39s



Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

3 Prior-Aligned Training with Attribution Constraints

Evaluation on GUI Agent

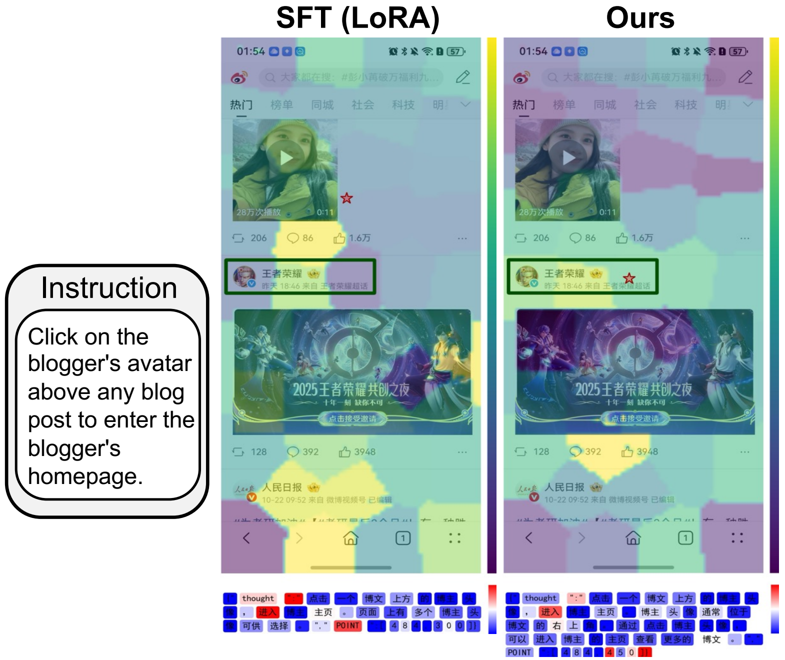


Instruction

Click the follow button for another Weibo user on the page.

{**"thought"**:"There is a Weibo post on this page that requires following. Tap the blogger's Comment button to follow.",**"POINT"**:[50,758]}

{**"thought"**:"The Weibo page shows content related to 'Hot Waves Island'. Tap a Weibo user's avatar to follow the uploader.",**"POINT"**:[789,754]}



Instruction

Click on the blogger's avatar above any blog post to enter the blogger's homepage.

{**"thought"**:"Tap the blogger's avatar above a post to enter the blogger's homepage. There are multiple blogger avatars on the page to choose from.",**"POINT"**:[484,300]}

{**"thought"**:"Tap the blogger's avatar above a post to enter the blogger's homepage. The avatar is usually at the top-right of the post. By tapping it, the blogger's profile opens to view more posts.",**"POINT"**:[484,450]}

Table 4. Evaluation on the GUI agent clicking task with AgentCPM-GUI. Standard SFT (LoRA) is compared with attribution-based alignment (LoRA). Task performance is reported by click success rate and distance error, and reliability is measured by Point Game and metrics conditioned on successful Point Game outcomes (click success rate and distance error when PG=1).

Methods	Task Performance		Point Game (↑)	Reliability Metrics	
	Click success rate (↑)	Distance error (↓)		Click success rate (PG=1) (↑)	Distance error (PG=1)
SFT (LoRA)	84.61%	94.71	0.8153	96.22%	7.11
Ours (LoRA)	**89.23%**	**78.64**	**0.8615**	**100%**	**0.0**

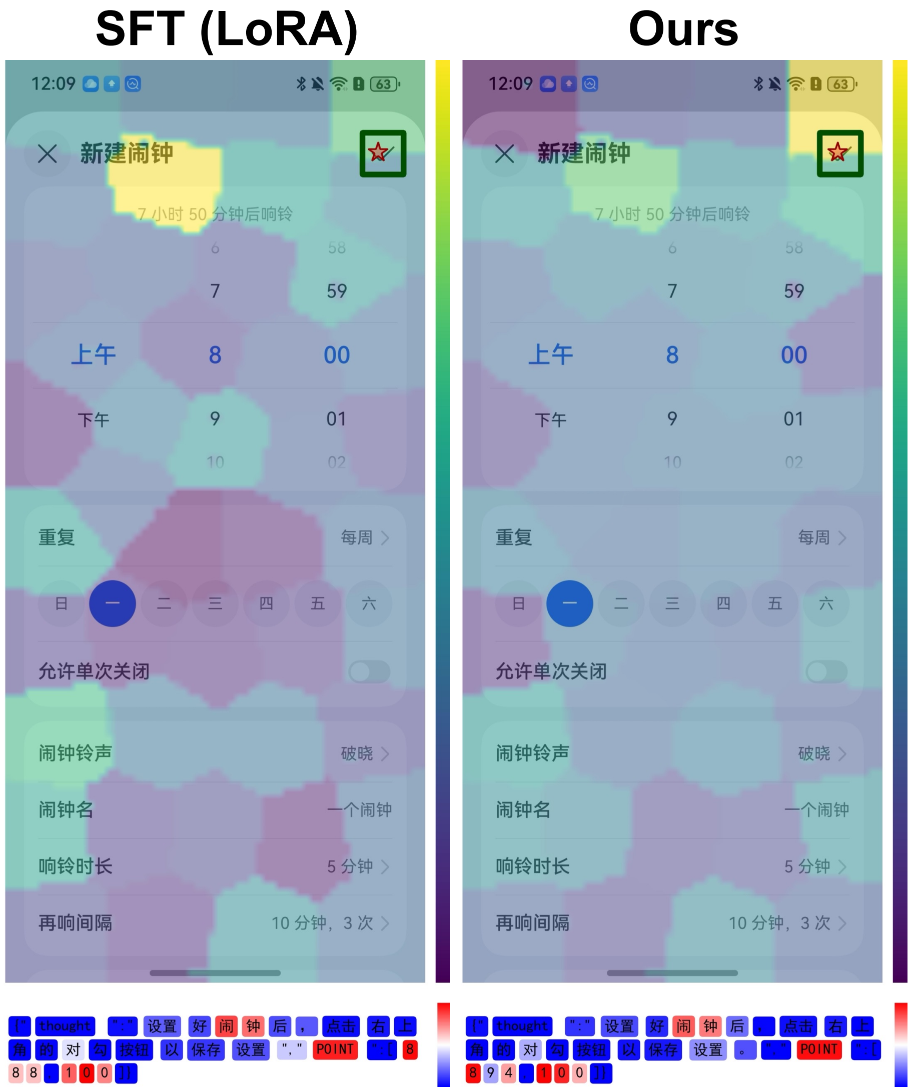
Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

3 Consistency between Decision and Attribution

Evaluation on GUI Agent

Instruction

Click the checkmark button in the upper right corner.


SFT (LoRA)

Ours

12:09
新建闹钟
7 小时 50 分钟后响铃
6 58
7 59
上午 8 00
下午 9 01
10 02
重复 每周
日 一 二 三 四 五 六
允许单次关闭
闹钟铃声 破晓
闹钟名 一个闹钟
响铃时长 5 分钟
再响间隔 10 分钟，3 次
{" thought ":" 设置 好 闹 钟 后 ， 点击 右 上 角 的 对 勾 按钮 以 保存 设置 "," POINT ":[8 8 8 , 1 0 0]}
{" thought ":" 设置 好 闹 钟 后 ， 点击 右 上 角 的 对 勾 按钮 以 保存 设置 。 "," POINT ":[8 9 4 , 1 0 0]}

{**“thought”**:“After setting the alarm, tap the checkmark button in the top-right corner to save the settings.”, **"POINT"**:[888,100]}

{**"thought"**:"After setting the alarm, tap the checkmark button in the top-right corner to save the settings.",**"POINT"**:[894,100]}

Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

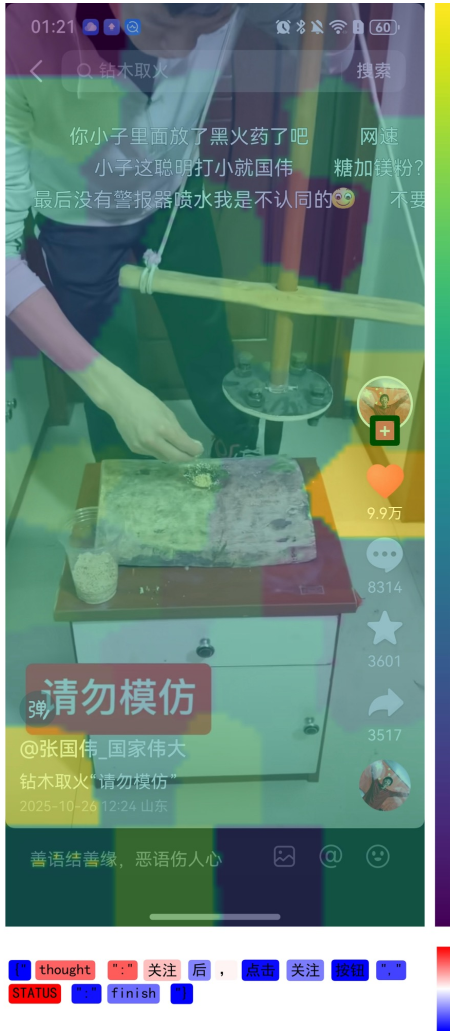
3 Failures

Evaluation on GUI Agent

Instruction

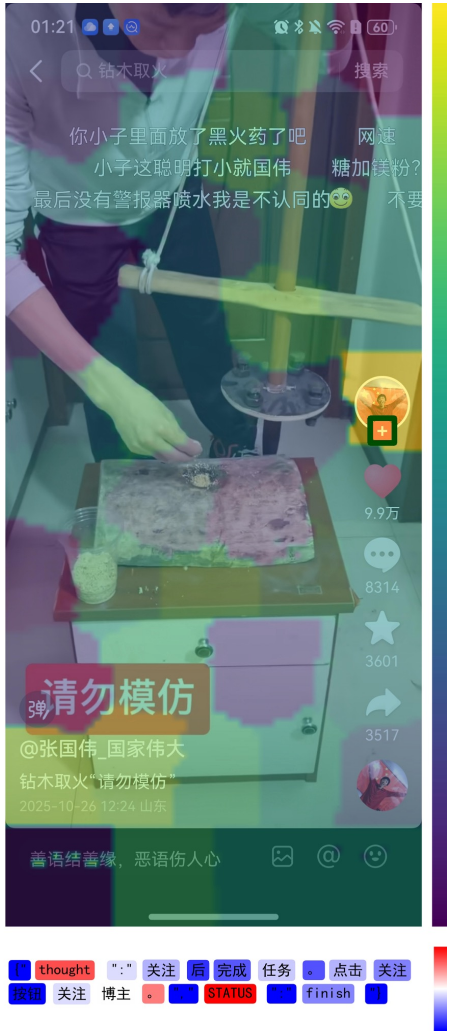
Follow the blogger by clicking the plus sign below their profile picture on the right side of the video.

SFT (LoRA)



{**“thought”**:“After following, tap the Follow button.”, **"STATUS"**:"finish"}

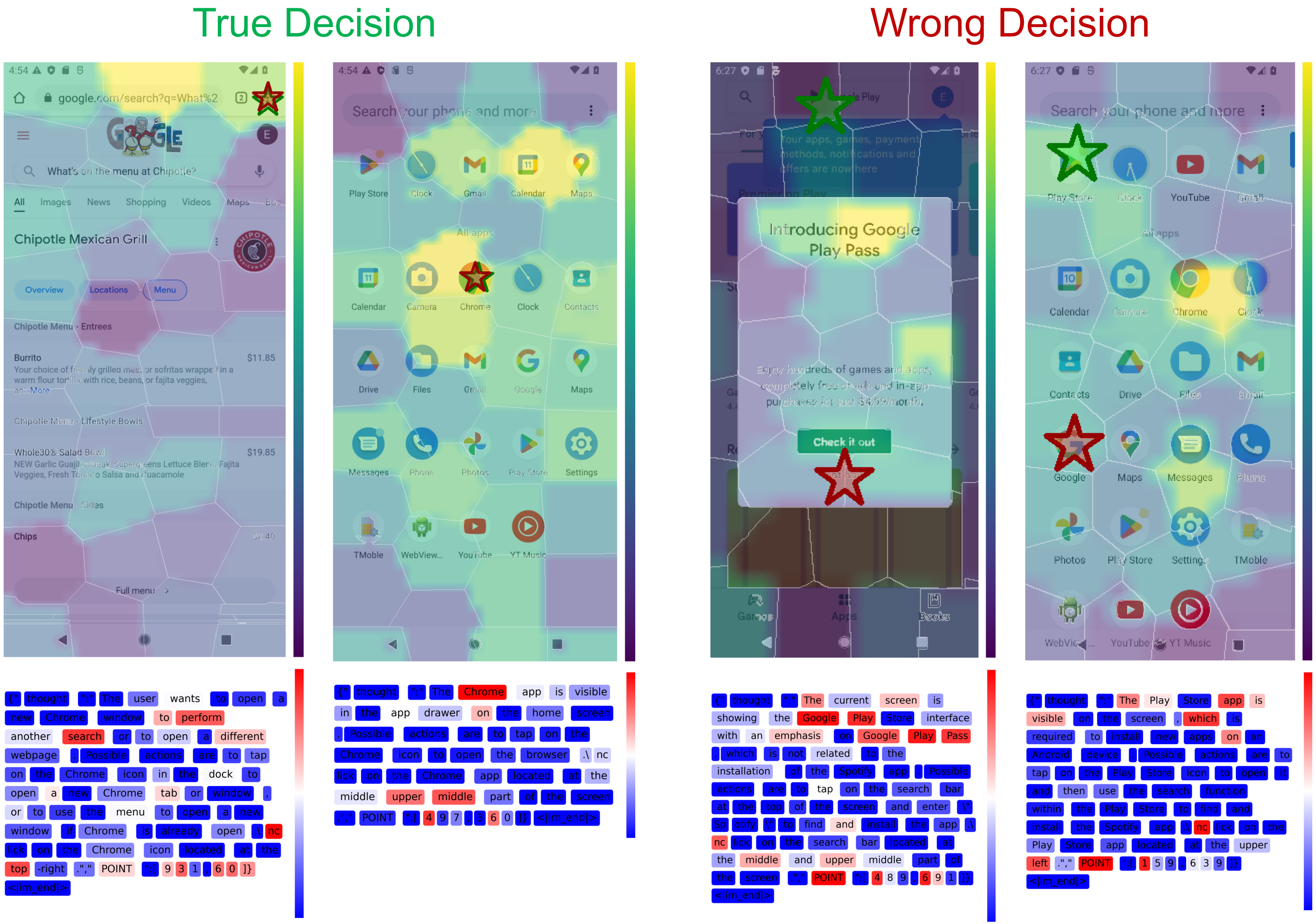
Ours



{**"thought"**:"After following, complete the task by tapping the Follow button to follow the uploader.",**"STATUS"**:"finish"}

3 Consistency Between Attribution and Decision-making

- □ We analyze the model’s reasoning process through attribution and compare the resulting attribution map with the model’s decision outcome.
- □ Low consistency between them implies a high probability of erroneous prediction, indicating potential use for hallucination detection.



```
{" thought ":" The user wants to open a
new Chrome window to perform
another search or to open a different
webpage . Possible actions are to tap
on the Chrome icon in the dock to
open a new Chrome tab or window ,
or to use the menu to open a new
window if Chrome is already open .\ nc
lick on the Chrome icon located at the
top -right ."," POINT ":[ 9 3 1 , 6 0 ]}
<|im_end|>
```

```
{" thought ":" The Chrome app is visible
in the app drawer on the home screen
. Possible actions are to tap on the
Chrome icon to open the browser .\ nc
lick on the Chrome app located at the
middle upper middle part of the screen
."," POINT ":[ 4 9 7 , 3 6 0 ]} <|im_end|>
```

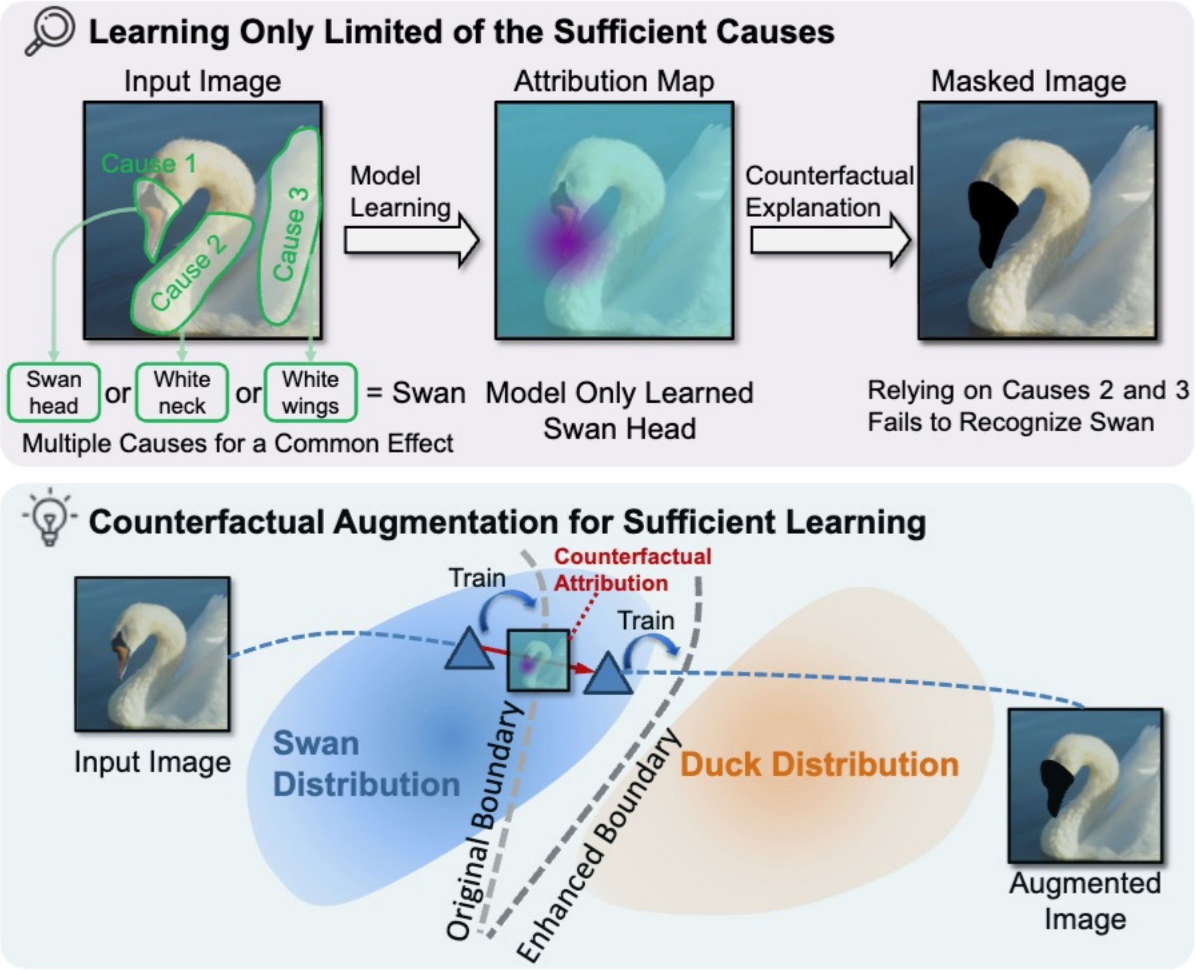
```
{" thought ":" The current screen is
showing the Google Play Store interface
with an emphasis on Google Play Pass
, which is not related to the
installation of the Spotify app . Possible
actions are to tap on the search bar
at the top of the screen and enter \"
Sp otify \" to find and install the app .\
nc lick on the search bar located at
the middle and upper middle part of
the screen ."," POINT ":[ 4 8 9 , 6 9 1 ]}
<|im_end|>
```

```
{" thought ":" The Play Store app is
visible on the screen , which is
required to install new apps on an
Android device . Possible actions are to
tap on the Play Store icon to open it
and then use the search function
within the Play Store to find and
install the Spotify app .\ nc lick on the
Play Store app located at the upper
left ."," POINT ":[ 1 5 9 , 6 3 9 ]}
<|im_end|>
```

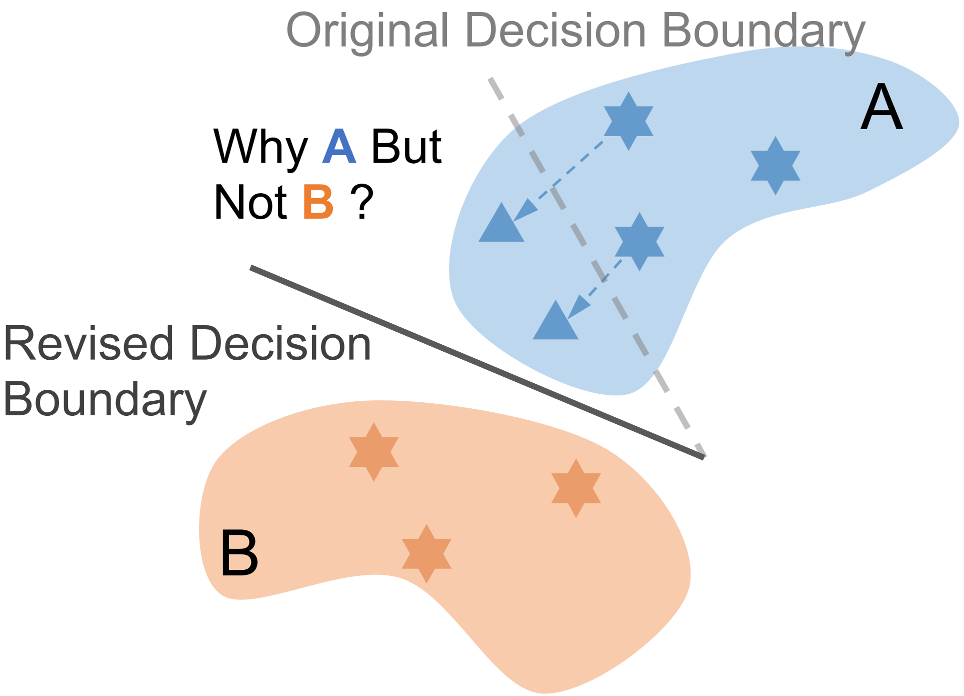
Ruoyu Chen, et al. Where Not to Learn: Prior-Aligned Training with Subset-based Attribution Constraints for Reliable Decision-Making. **Preprint 2026**.

3 Counterfactual Data Augmentation

During data-driven training, the model may rely on a subset of **underlying causes** rather than comprehensively capturing the full causal structure, which can result in biased representations and decisions.


Learning Only Limited of the Sufficient Causes
Input Image
Cause 1
Cause 2
Cause 3
Model Learning
Attribution Map
Counterfactual Explanation
Masked Image
Swan head
or
White neck
or
White wings
= Swan
Multiple Causes for a Common Effect
Model Only Learned Swan Head
Relying on Causes 2 and 3 Fails to Recognize Swan
Counterfactual Augmentation for Sufficient Learning
Train
Counterfactual Attribution
Train
Input Image
Swan Distribution
Original Boundary
Enhanced Boundary
Duck Distribution
Augmented Image

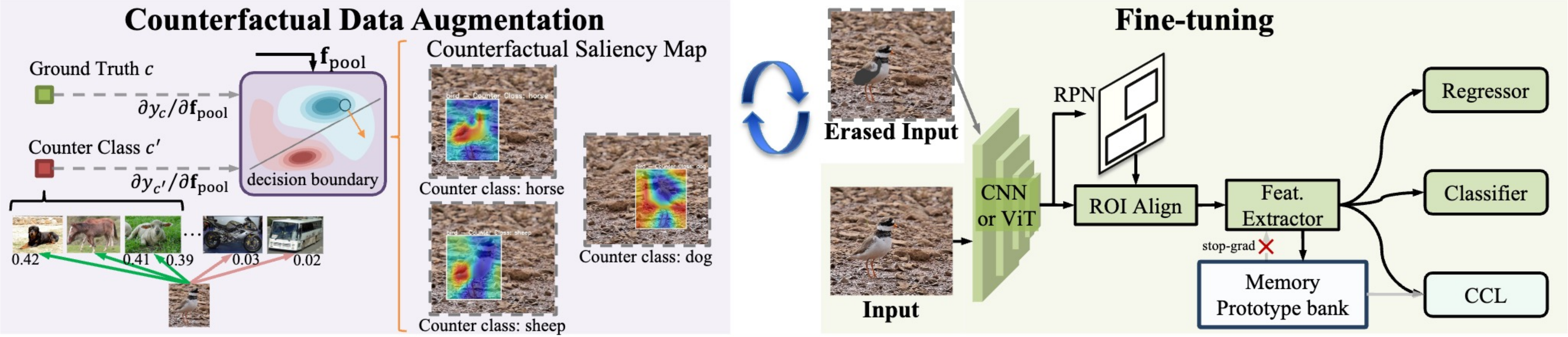
Few-shot Sceneries



Ruoyu Chen, et al. Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection. **IEEE Trans. Pattern Anal. Mach. Intell.** (2025).

3 Counterfactual Data Augmentation

Solution: We propose an interpretable feedback loop to make model training transparent, using explainable methods to locate and correct potential model flaws. A counterfactual explanation approach is designed to reveal bias information and refine the feature space through counterfactual augmentation. Theoretically, the empirical risk is proven to decrease relative to the baseline: $\left(1-\frac{\lambda}{\sqrt{1+N_a/N_r}}\right)\sqrt{\frac{\ln(4/\delta)}{2N_r}}$.



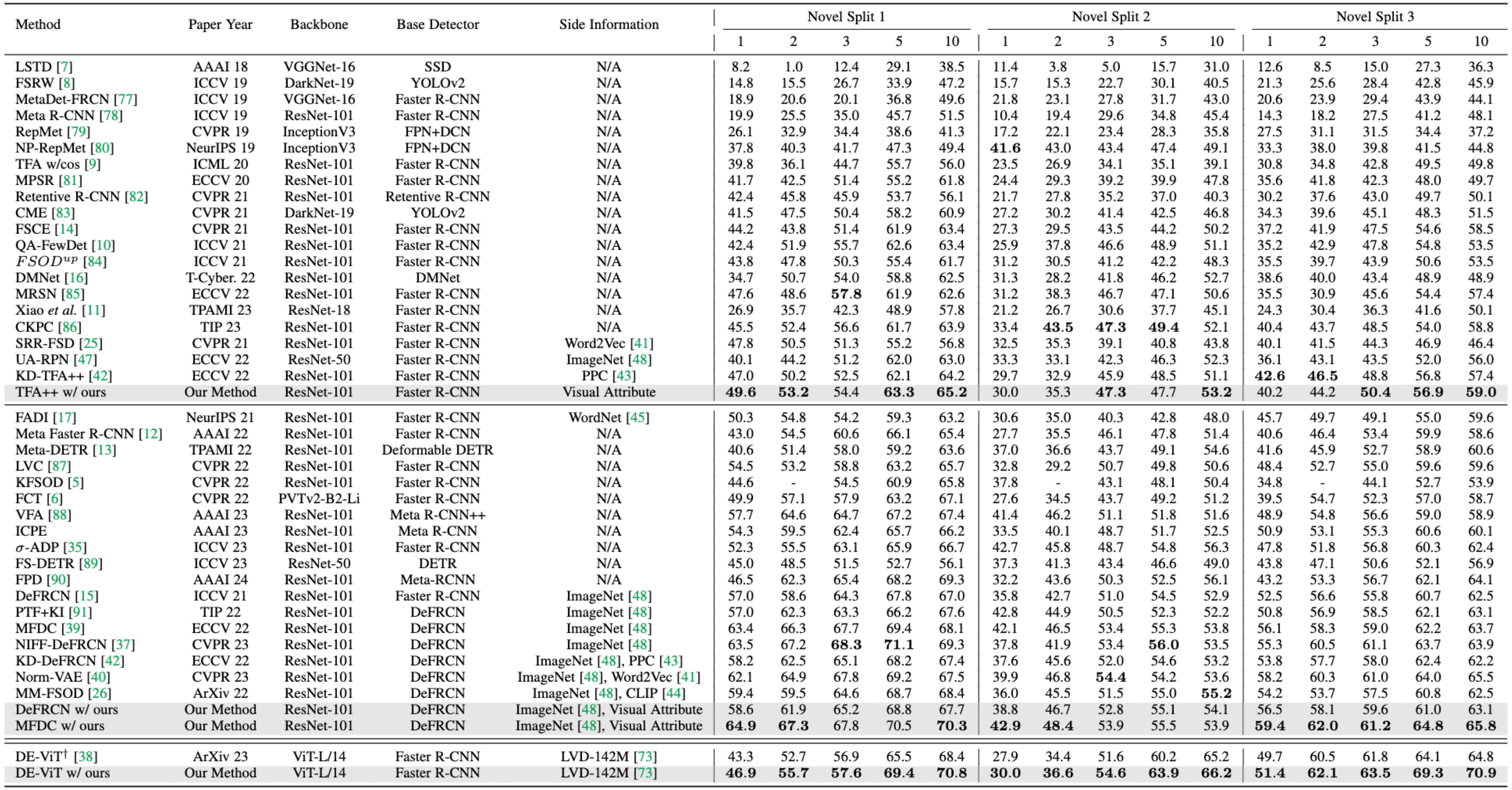
Ruoyu Chen, et al. Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection. IEEE Trans. Pattern Anal. Mach. Intell. (2025).

3 Counterfactual Data Augmentation

Experimental results show state-of-the-art performance on few-shot detection benchmarks, compatibility with multiple baselines and backbones (including CNNs and ViTs), and theoretical guarantees on the generalization error bound.

TABLE 6

Few-shot object detection evaluation results on PASCAL VOC [27]. The evaluation metric adopts the mean average precision (mAP@0.5). † denotes further fine-tuned on the novel categories.

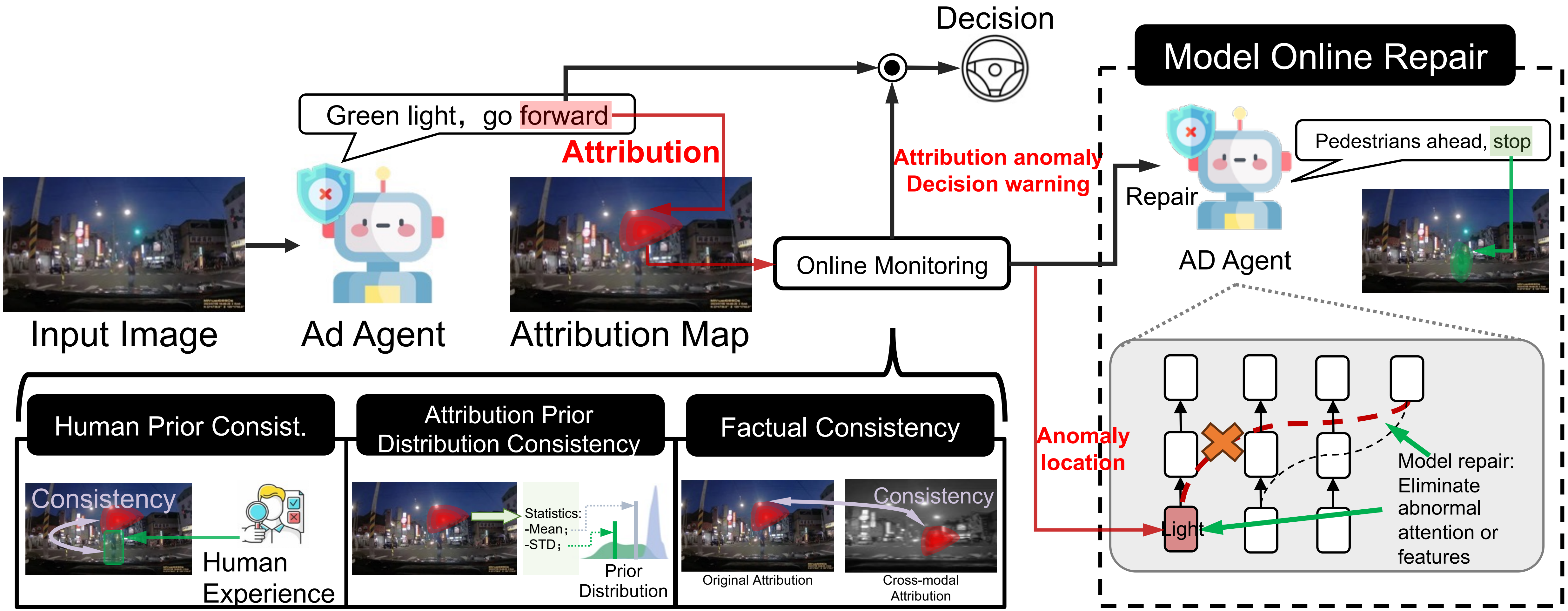


Method	Paper Year	Backbone	Base Detector	Side Information	Novel Split 1					Novel Split 2					Novel Split 3				
					1	2	3	5	10	1	2	3	5	10	1	2	3	5	10
LSTD [7]	AAAI 18	VGGNet-16	SSD	N/A	8.2	1.0	12.4	29.1	38.5	11.4	3.8	5.0	15.7	31.0	12.6	8.5	15.0	27.3	36.3
FSRW [8]	ICCV 19	DarkNet-19	YOLOv2	N/A	14.8	15.5	26.7	33.9	47.2	15.7	15.3	22.7	30.1	40.5	21.3	25.6	28.4	42.8	45.9
MetaDet-FRCN [77]	ICCV 19	VGGNet-16	Faster R-CNN	N/A	18.9	20.6	20.1	36.8	49.6	21.8	23.1	27.8	31.7	43.0	20.6	23.9	29.4	43.9	44.1
Meta R-CNN [78]	ICCV 19	ResNet-101	Faster R-CNN	N/A	19.9	25.5	35.0	45.7	51.5	10.4	19.4	29.6	34.8	45.4	14.3	18.2	27.5	41.2	48.1
RepMet [79]	CVPR 19	InceptionV3	FPN+DCN	N/A	26.1	32.9	34.4	38.6	41.3	17.2	22.1	23.4	28.3	35.8	27.5	31.1	31.5	34.4	37.2
NP-RepMet [80]	NeurIPS 19	InceptionV3	FPN+DCN	N/A	37.8	40.3	41.7	47.3	49.4	**41.6**	43.0	43.4	47.4	49.1	33.3	38.0	39.8	41.5	44.8
TFA w/cos [9]	ICML 20	ResNet-101	Faster R-CNN	N/A	39.8	36.1	44.7	55.7	56.0	23.5	26.9	34.1	35.1	39.1	30.8	34.8	42.8	49.5	49.8
MPSR [81]	ECCV 20	ResNet-101	Faster R-CNN	N/A	41.7	42.5	51.4	55.2	61.8	24.4	29.3	39.2	39.9	47.8	35.6	41.8	42.3	48.0	49.7
Retentive R-CNN [82]	CVPR 21	ResNet-101	Retentive R-CNN	N/A	42.4	45.8	45.9	53.7	56.1	21.7	27.8	35.2	37.0	40.3	30.2	37.6	43.0	49.7	50.1
CME [83]	CVPR 21	DarkNet-19	YOLOv2	N/A	41.5	47.5	50.4	58.2	60.9	27.2	30.2	41.4	42.5	46.8	34.3	39.6	45.1	48.3	51.5
FSCE [14]	CVPR 21	ResNet-101	Faster R-CNN	N/A	44.2	43.8	51.4	61.9	63.4	27.3	29.5	43.5	44.2	50.2	37.2	41.9	47.5	54.6	58.5
QA-FewDet [10]	ICCV 21	ResNet-101	Faster R-CNN	N/A	42.4	51.9	55.7	62.6	63.4	25.9	37.8	46.6	48.9	51.1	35.2	42.9	47.8	54.8	53.5
$FSOD^{up}$ [84]	ICCV 21	ResNet-101	Faster R-CNN	N/A	43.8	47.8	50.3	55.4	61.7	31.2	30.5	41.2	42.2	48.3	35.5	39.7	43.9	50.6	53.5
DMNet [16]	T-Cyber. 22	ResNet-101	DMNet	N/A	34.7	50.7	54.0	58.8	62.5	31.3	28.2	41.8	46.2	52.7	38.6	40.0	43.4	48.9	48.9
MRSN [85]	ECCV 22	ResNet-101	Faster R-CNN	N/A	47.6	48.6	**57.8**	61.9	62.6	31.2	38.3	46.7	47.1	50.6	35.5	30.9	45.6	54.4	57.4
Xiao *et al.* [11]	TPAMI 23	ResNet-18	Faster R-CNN	N/A	26.9	35.7	42.3	48.9	57.8	21.2	26.7	30.6	37.7	45.1	24.3	30.4	36.3	41.6	50.1
CKPC [86]	TIP 23	ResNet-101	Faster R-CNN	N/A	45.5	52.4	56.6	61.7	63.9	33.4	**43.5**	**47.3**	**49.4**	52.1	40.4	43.7	48.5	54.0	58.8
SRR-FSD [25]	CVPR 21	ResNet-101	Faster R-CNN	Word2Vec [41]	47.8	50.5	51.3	55.2	56.8	32.5	35.3	39.1	40.8	43.8	40.1	41.5	44.3	46.9	46.4
UA-RPN [47]	ECCV 22	ResNet-50	Faster R-CNN	ImageNet [48]	40.1	44.2	51.2	62.0	63.0	33.3	33.1	42.3	46.3	52.3	36.1	43.1	43.5	52.0	56.0
KD-TFA++ [42]	ECCV 22	ResNet-101	Faster R-CNN	PPC [43]	47.0	50.2	52.5	62.1	64.2	29.7	32.9	45.9	48.5	51.1	**42.6**	**46.5**	48.8	56.8	57.4
TFA++ w/ ours	Our Method	ResNet-101	Faster R-CNN	Visual Attribute	**49.6**	**53.2**	54.4	**63.3**	**65.2**	30.0	35.3	**47.3**	47.7	**53.2**	40.2	44.2	**50.4**	**56.9**	**59.0**
FADI [17]	NeurIPS 21	ResNet-101	Faster R-CNN	WordNet [45]	50.3	54.8	54.2	59.3	63.2	30.6	35.0	40.3	42.8	48.0	45.7	49.7	49.1	55.0	59.6
Meta Faster R-CNN [12]	AAAI 22	ResNet-101	Faster R-CNN	N/A	43.0	54.5	60.6	66.1	65.4	27.7	35.5	46.1	47.8	51.4	40.6	46.4	53.4	59.9	58.6
Meta-DETR [13]	TPAMI 22	ResNet-101	Deformable DETR	N/A	40.6	51.4	58.0	59.2	63.6	37.0	36.6	43.7	49.1	54.6	41.6	45.9	52.7	58.9	60.6
LVC [87]	CVPR 22	ResNet-101	Faster R-CNN	N/A	54.5	53.2	58.8	63.2	65.7	32.8	29.2	50.7	49.8	50.6	48.4	52.7	55.0	59.6	59.6
KFSOD [5]	CVPR 22	ResNet-101	Faster R-CNN	N/A	44.6	-	54.5	60.9	65.8	37.8	-	43.1	48.1	50.4	34.8	-	44.1	52.7	53.9
FCT [6]	CVPR 22	PVTv2-B2-Li	Faster R-CNN	N/A	49.9	57.1	57.9	63.2	67.1	27.6	34.5	43.7	49.2	51.2	39.5	54.7	52.3	57.0	58.7
VFA [88]	AAAI 23	ResNet-101	Meta R-CNN++	N/A	57.7	64.6	64.7	67.2	67.4	41.4	46.2	51.1	51.8	51.6	48.9	54.8	56.6	59.0	58.9
ICPE	AAAI 23	ResNet-101	Meta R-CNN	N/A	54.3	59.5	62.4	65.7	66.2	33.5	40.1	48.7	51.7	52.5	50.9	53.1	55.3	60.6	60.1
σ-ADP [35]	ICCV 23	ResNet-101	Faster R-CNN	N/A	52.3	55.5	63.1	65.9	66.7	42.7	45.8	48.7	54.8	56.3	47.8	51.8	56.8	60.3	62.4
FS-DETR [89]	ICCV 23	ResNet-50	DETR	N/A	45.0	48.5	51.5	52.7	56.1	37.3	41.3	43.4	46.6	49.0	43.8	47.1	50.6	52.1	56.9
FPD [90]	AAAI 24	ResNet-101	Meta-RCNN	N/A	46.5	62.3	65.4	68.2	69.3	32.2	43.6	50.3	52.5	56.1	43.2	53.3	56.7	62.1	64.1
DeFRCN [15]	ICCV 21	ResNet-101	Faster R-CNN	ImageNet [48]	57.0	58.6	64.3	67.8	67.0	35.8	42.7	51.0	54.5	52.9	52.5	56.6	55.8	60.7	62.5
PTF+KI [91]	TIP 22	ResNet-101	DeFRCN	ImageNet [48]	57.0	62.3	63.3	66.2	67.6	42.8	44.9	50.5	52.3	52.2	50.8	56.9	58.5	62.1	63.1
MFDC [39]	ECCV 22	ResNet-101	DeFRCN	ImageNet [48]	63.4	66.3	67.7	69.4	68.1	42.1	46.5	53.4	55.3	53.8	56.1	58.3	59.0	62.2	63.7
NIFF-DeFRCN [37]	CVPR 23	ResNet-101	DeFRCN	ImageNet [48]	63.5	67.2	**68.3**	**71.1**	69.3	37.8	41.9	53.4	**56.0**	53.5	55.3	60.5	61.1	63.7	63.9
KD-DeFRCN [42]	ECCV 22	ResNet-101	DeFRCN	ImageNet [48], PPC [43]	58.2	62.5	65.1	68.2	67.4	37.6	45.6	52.0	54.6	53.2	53.8	57.7	58.0	62.4	62.2
Norm-VAE [40]	CVPR 23	ResNet-101	DeFRCN	ImageNet [48], Word2Vec [41]	62.1	64.9	67.8	69.2	67.5	39.9	46.8	**54.4**	54.2	53.6	58.2	60.3	61.0	64.0	65.5
MM-FSOD [26]	ArXiv 22	ResNet-101	DeFRCN	ImageNet [48], CLIP [44]	59.4	59.5	64.6	68.7	68.4	36.0	45.5	51.5	55.0	**55.2**	54.2	53.7	57.5	60.8	62.5
DeFRCN w/ ours	Our Method	ResNet-101	DeFRCN	ImageNet [48], Visual Attribute	58.6	61.9	65.2	68.8	67.7	38.8	46.7	52.8	55.1	54.1	56.5	58.1	59.6	61.0	63.1
MFDC w/ ours	Our Method	ResNet-101	DeFRCN	ImageNet [48], Visual Attribute	**64.9**	**67.3**	67.8	70.5	**70.3**	**42.9**	**48.4**	53.9	55.5	53.9	**59.4**	**62.0**	**61.2**	**64.8**	**65.8**
DE-ViT† [38]	ArXiv 23	ViT-L/14	Faster R-CNN	LVD-142M [73]	43.3	52.7	56.9	65.5	68.4	27.9	34.4	51.6	60.2	65.2	49.7	60.5	61.8	64.1	64.8
DE-ViT w/ ours	Our Method	ViT-L/14	Faster R-CNN	LVD-142M [73]	**46.9**	**55.7**	**57.6**	**69.4**	**70.8**	**30.0**	**36.6**	**54.6**	**63.9**	**66.2**	**51.4**	**62.1**	**63.5**	**69.3**	**70.9**

Ruoyu Chen, et al. Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection. **IEEE Trans. Pattern Anal. Mach. Intell.** (2025).

Future Outlook

Anomaly Monitoring: Evaluate the reliability of the current model decision by explaining whether the attribution is abnormal, and use online repair methods to dynamically repair model defects at low cost.



Thanks for listening!

Any questions?

Ruoyu Chen