中国科学院大学
University of Chinese Academy of Sciences
中国科学院信息工程研究所
INSTITUTE OF INFORMATION ENGINEERING,CAS
中山大學
SUN YAT-SEN UNIVERSITY

Interpretation of the Foundation Model: Concepts, Challenges, and Applications

Ruoyu Chen
University of Chinese Academy of Sciences
https://ruoyuchen10.github.io/

2024.06.27

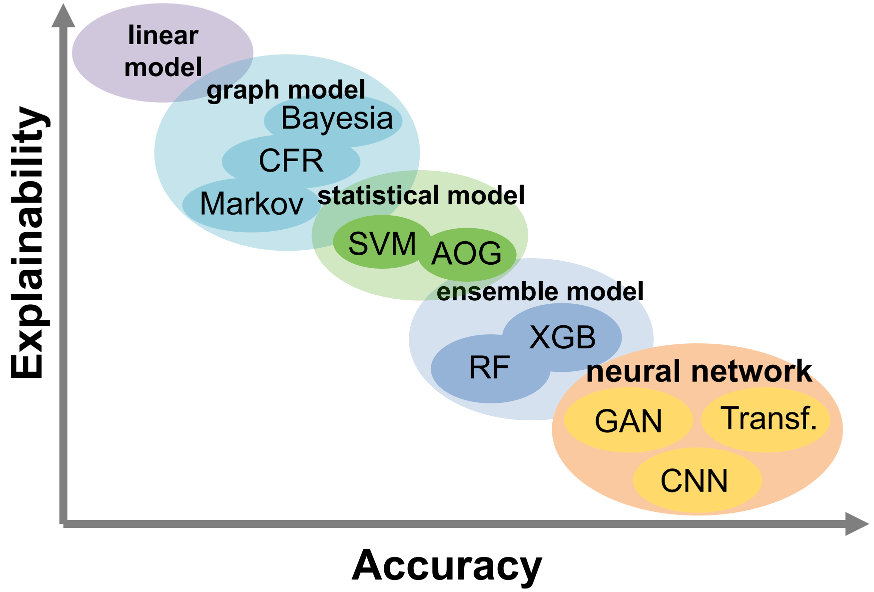
Outline

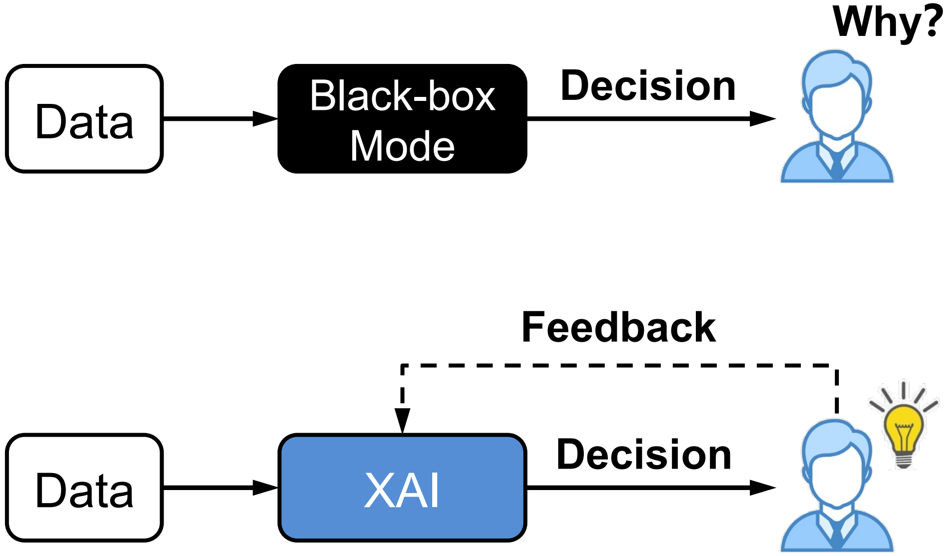
1. Why We Need Interpretable AI?
   - Introduction & Conception
   - How to Apply XAI - Research Routes
2. Interpretation for Large Model
   - Traditional Method
   - Category and Challenge
   - CLIP Interpretation
   - Explainable Generative AI
   - Interpret and Enhance Model Performance During Training
3. AI Agent and XAI
   - Related Work
   - What can we interpret
4. World Model and Challenges in XAI
   - Related Work
   - What can we interpret
5. Future Outlook

1. Why We Need Interpretable AI?

- ☐ Introduction & Conception
- ☐ How to Apply XAI - Research Routes

1.1 Introduction & Conception





The AI black box model has the risk of making decisions that are unreasonable, illegal, or without detailed explanations.

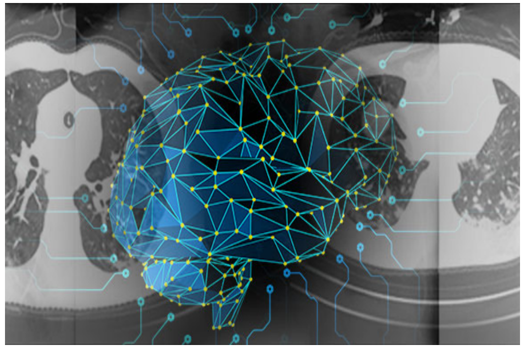
Explainable AI helps humans understand model decisions, trust the model more, and improve the AI model based on continuous feedback.

The huge success of ML has led to an explosion in the capabilities of AI, but its effectiveness will be limited by the machine's inability to explain its decisions and actions to human users. XAI is critical for users to understand, properly trust and effectively manage this new generation of artificial intelligence.

Autonomous driving

Education

Financial risk



Medical health

1.1 Introduction & Conception

Interpretation

- The actual operating mechanism behind the model;
- Accurately link model causes to effects;
- Determine what the model actually learned;
- Correct under certain conditions.

Explanation

- Represent the decision-making process or results in a human-understandable manner;
- Associating various feedback modalities and controlling the degree of semantic expression;
- Not necessarily correct.

Ante-hoc & Self-explainability

- Directly interpretable white-box models;
- Interpretability has been generated during the decision-making process of the model.

Post-hoc

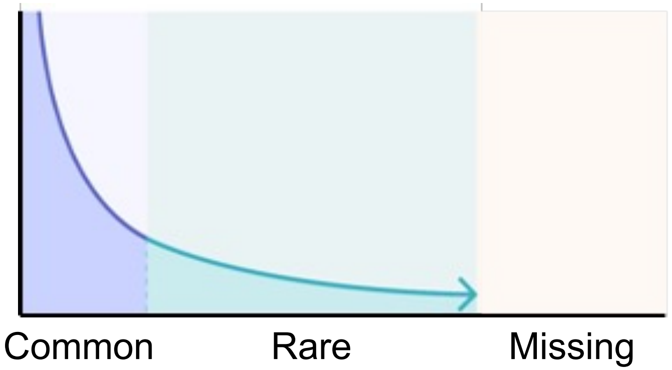
- Interpret the results of a pretrained model or its decisions;
- An explanation provided after the model has made one or several decisions.

1. Why We Need Interpretable AI?

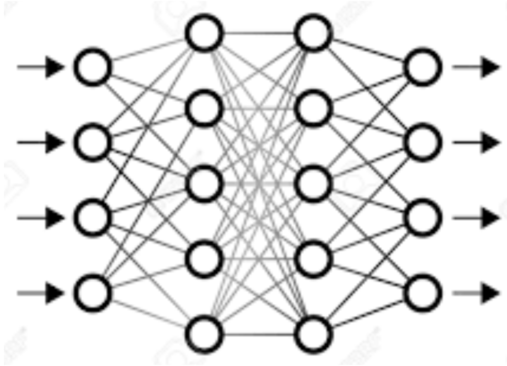
- ☐ Introduction & Conception
- ☐ How to Apply XAI - Research Routes

1.2 How to Apply XAI

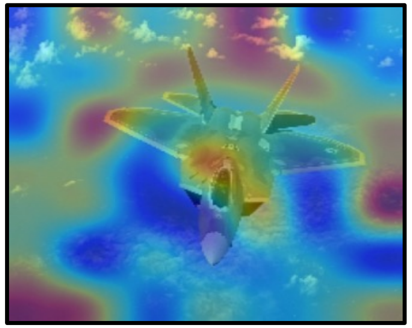
Why do AI models still have errors?



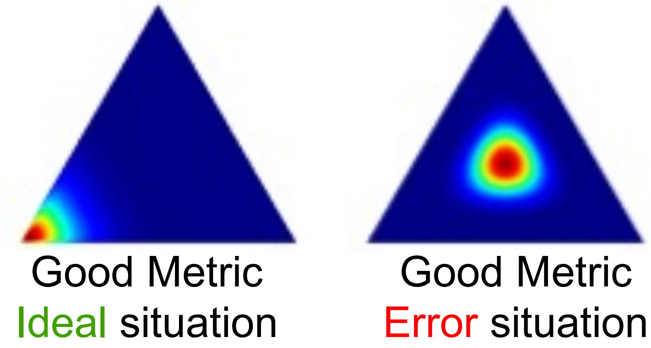
Data distribution is uneven



Defects in the model itself



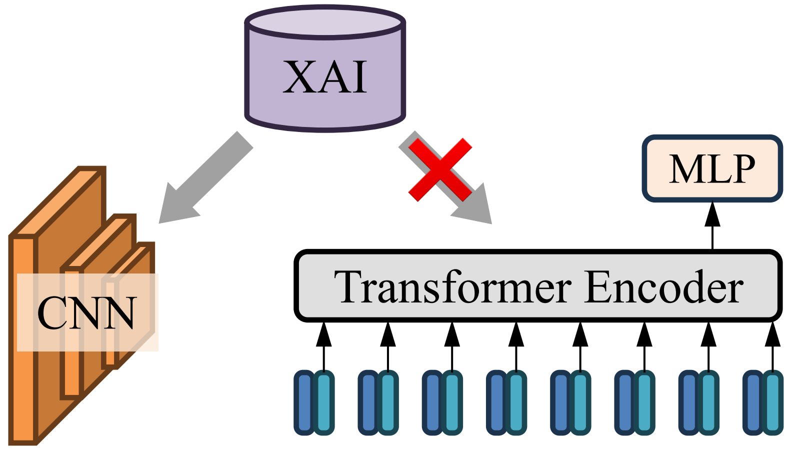
Less supervision information



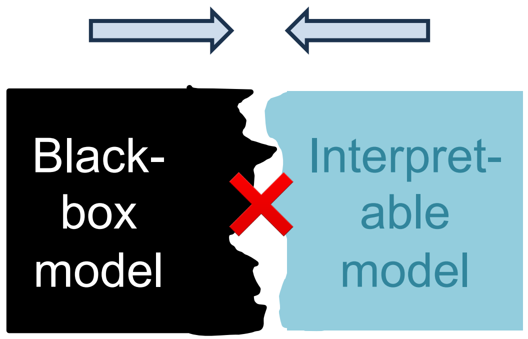
Evaluation metric defects

So we need interpretation!

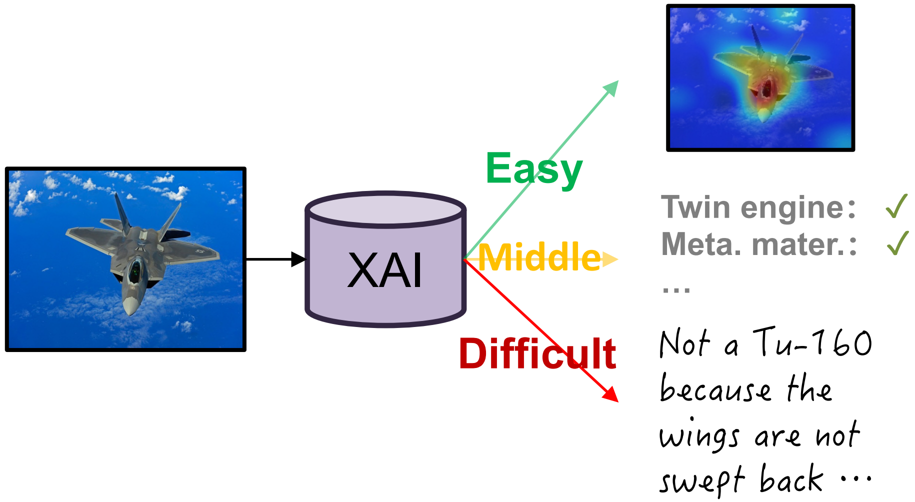
1.2 How to Apply XAI



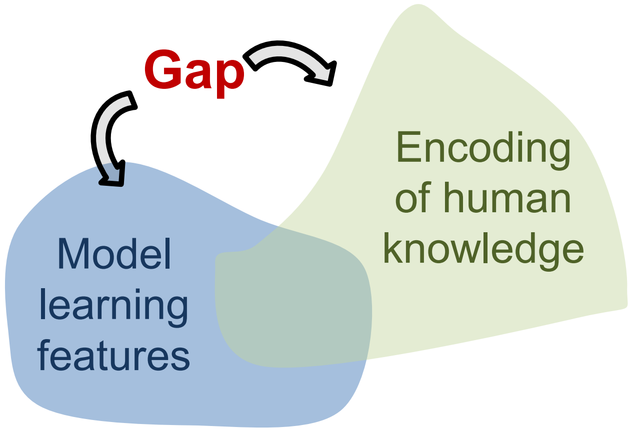
1. Interpretation paradigm is not universal



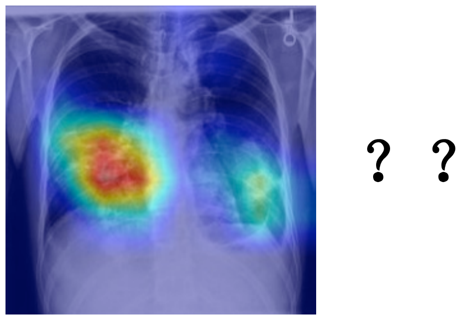
2. Interpretable models are difficult to design



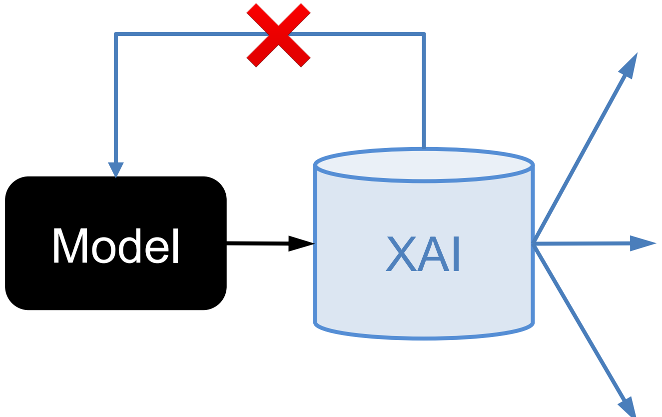
3. High degree of semantic feedback is difficult to interpret



4. Human knowledge is difficult to integrate

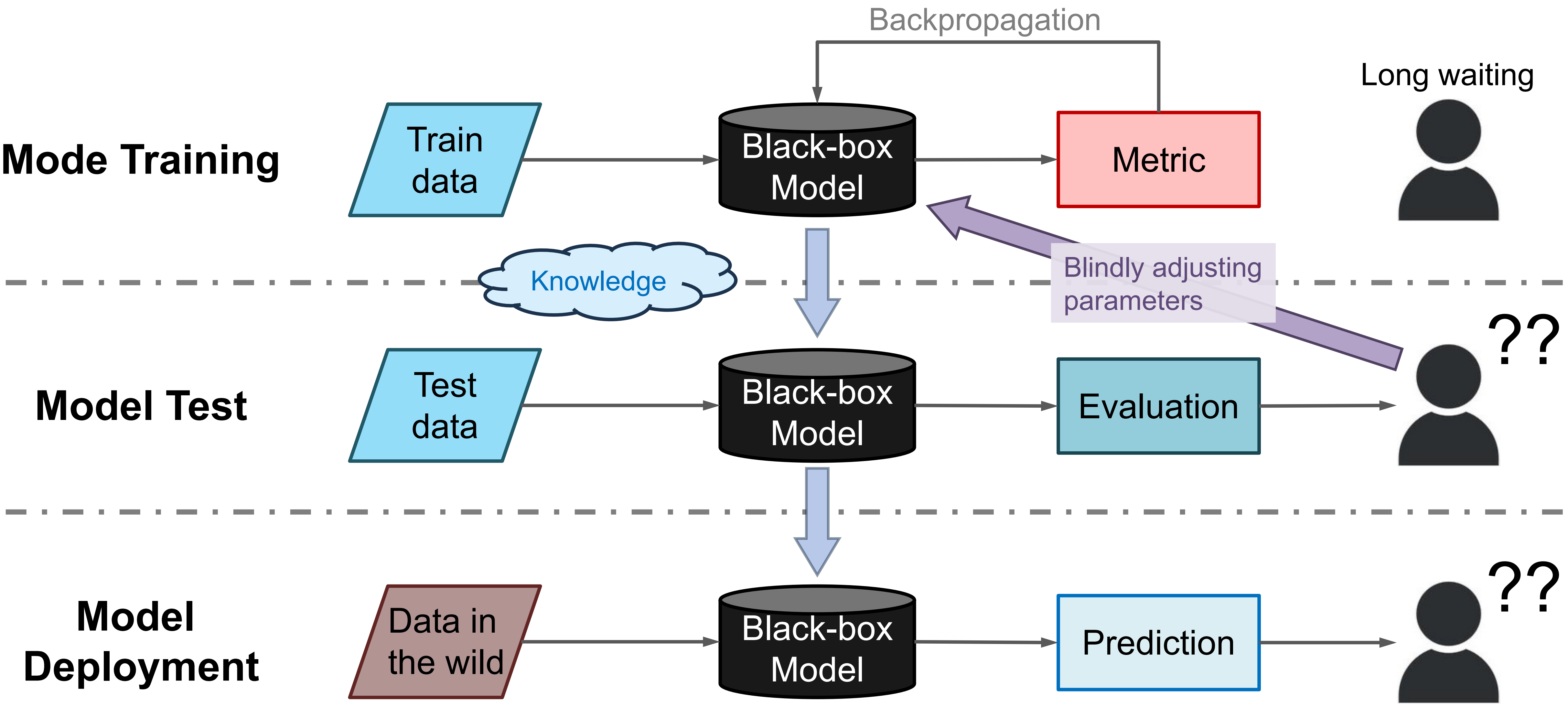


5. Interpretation results is difficult to evaluate

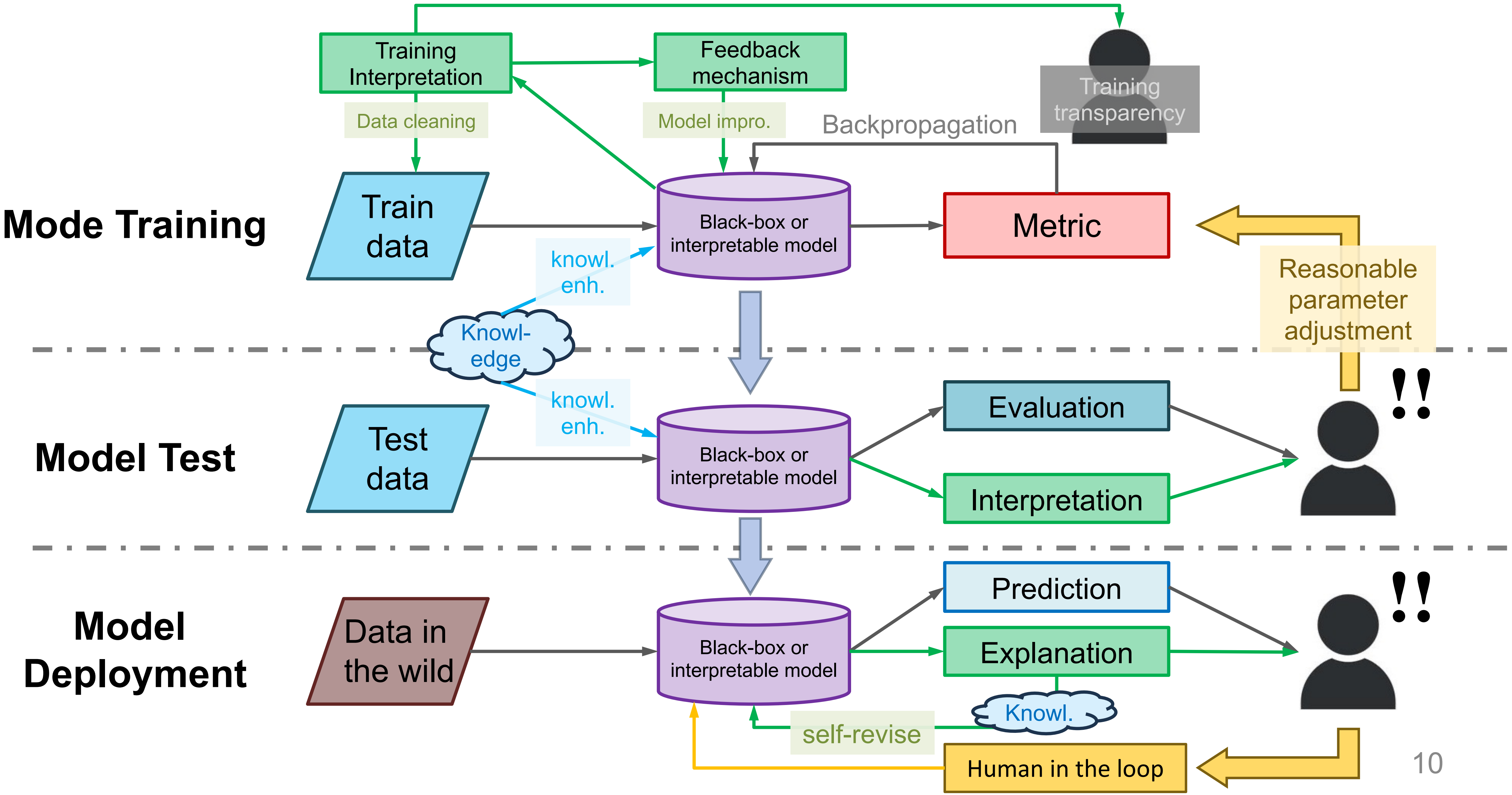


6. Model feedback is difficult to construct

1.2 How to Apply XAI

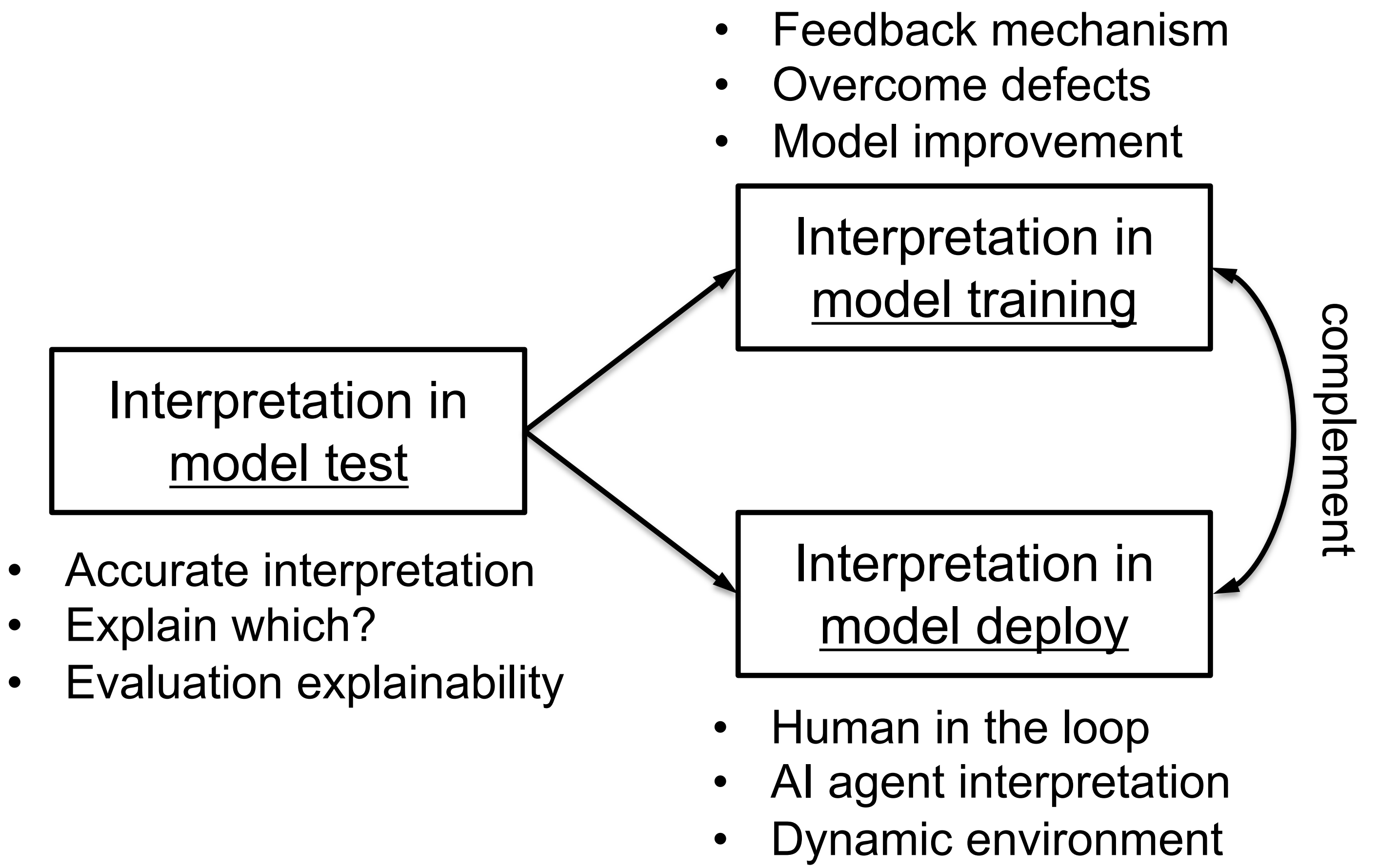


1.2 How to Apply XAI



1.2 How to Apply XAI

How to design?

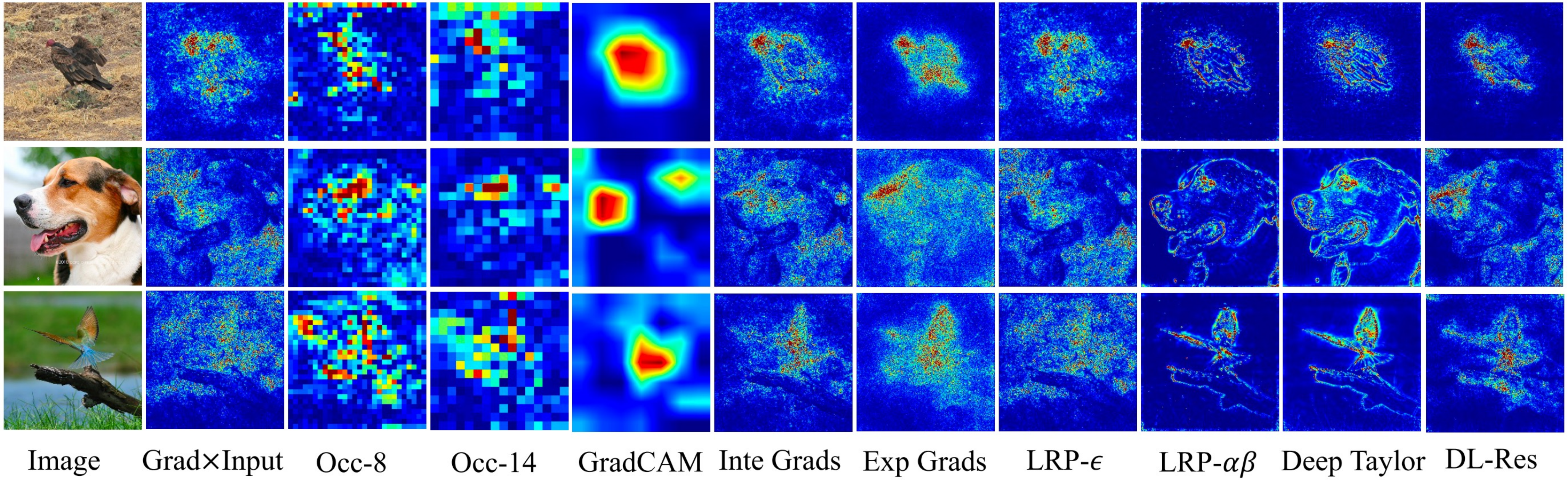


2. Interpretation for Large Model

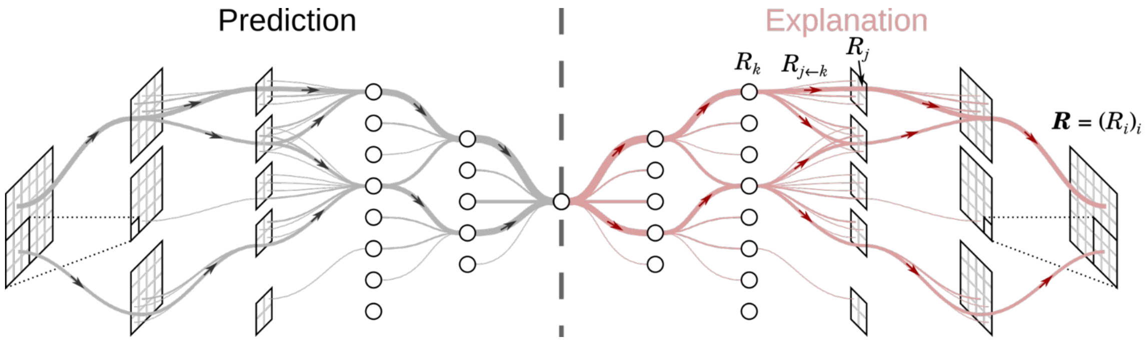
- Tradition Method
- Category and Challenge
- CLIP Interpretation
- Explainable Generative AI
- Interpret and Enhance Model Performance During Training

2.1 Traditional Method

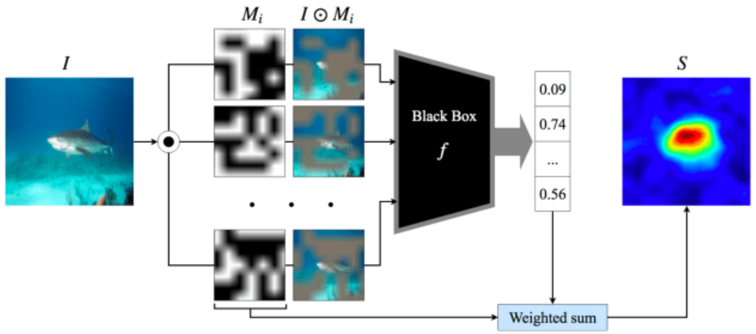
Attribution-based Methods



Based on the internal mechanism of the model (white box)

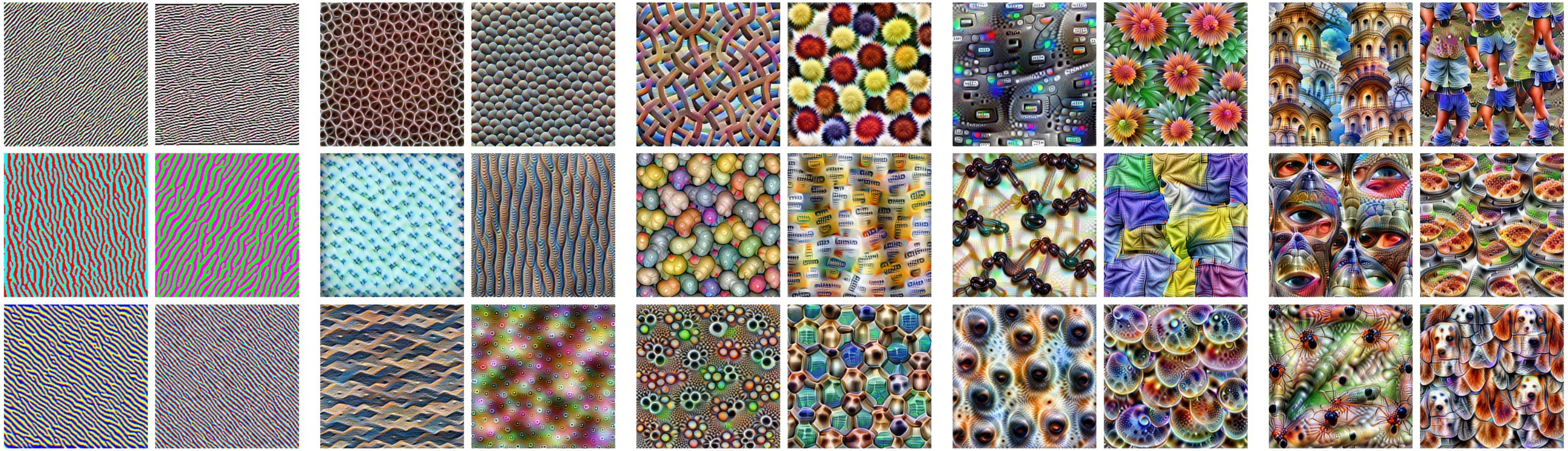


Based on perturbation (black box)

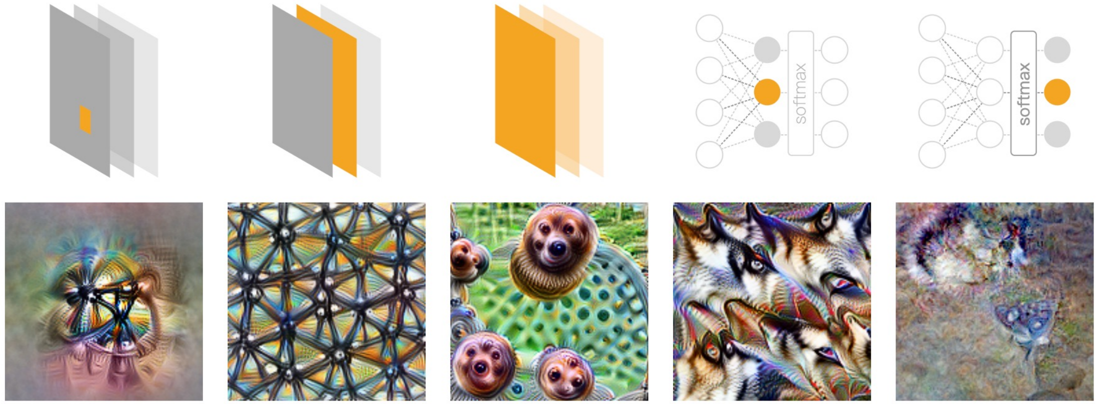


2.1 Traditional Method

Feature visualization-based Methods



Edges (layer conv2d0) **Textures** (layer mixed3a) **Patterns** (layer mixed4a) **Parts** (layers mixed4b & mixed4c) **Objects** (layers mixed4d & mixed4e)



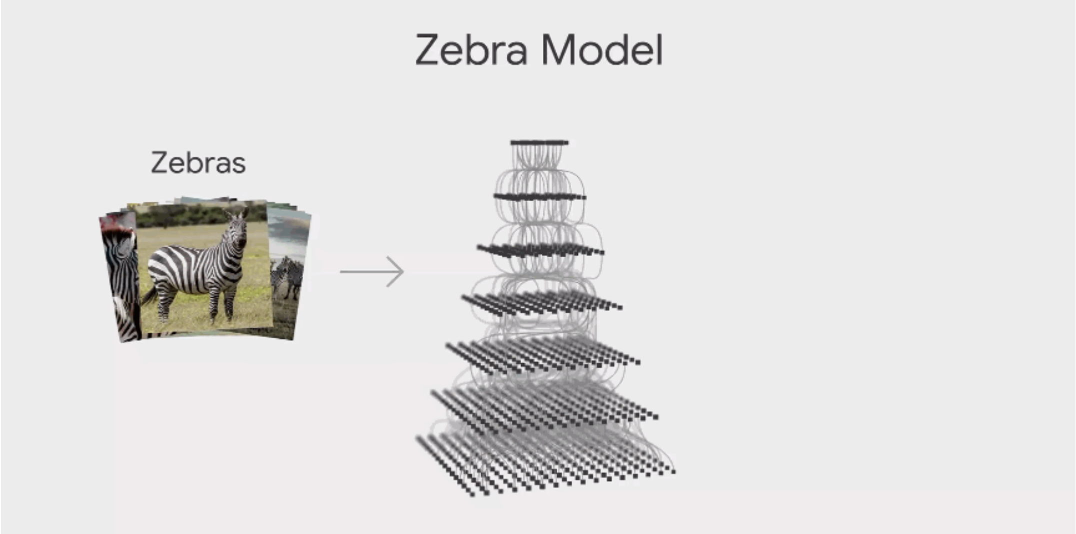
Neuron $\texttt{layer}_n\texttt{[x,y,z]}$ **Channel** $\texttt{layer}_n\texttt{[:,:,z]}$ **Layer**/DeepDream $\texttt{layer}_n\texttt{[:,:,:]}^2$ **Class Logits** `pre_softmax[k]` **Class Probability** `softmax[k]`

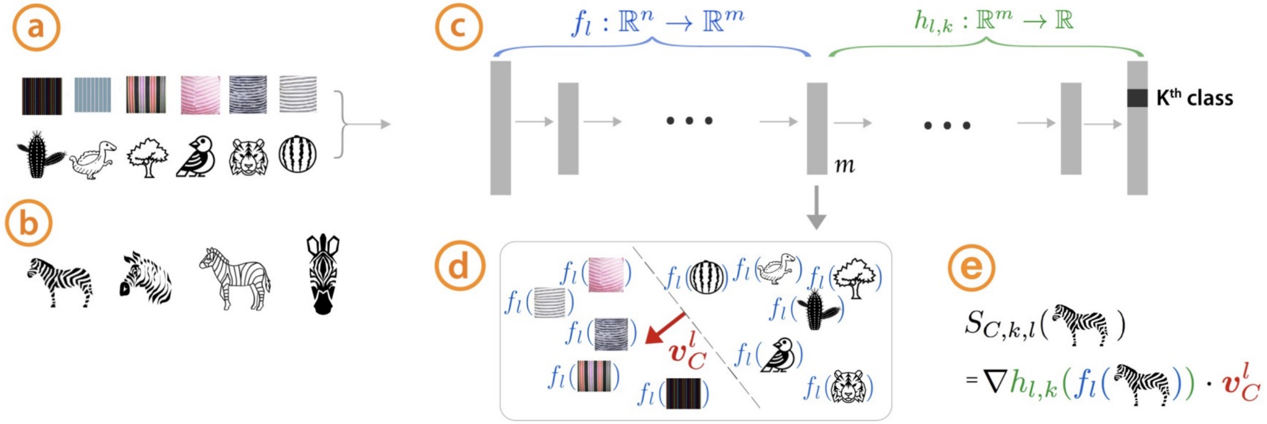
Feature Visualization:
Specify the intermediate unit, optimize the input so that the target unit has the maximum activation response, and observe the optimized input image.

Feature Visualization, https://distill.pub/2017/feature-visualization/

2.1 Traditional Method

Concept-based Methods





TCAV: For a concept activation vector v_l in the f_l layer of the model, the categories are c, the predicted score is f_c. Thus:

$$S_c(x) = v_l \cdot \frac{\partial f_c(x)}{\partial f_l(x)},$$

The TCAV score is the percentage of elements in category c that have a positive score S_c:

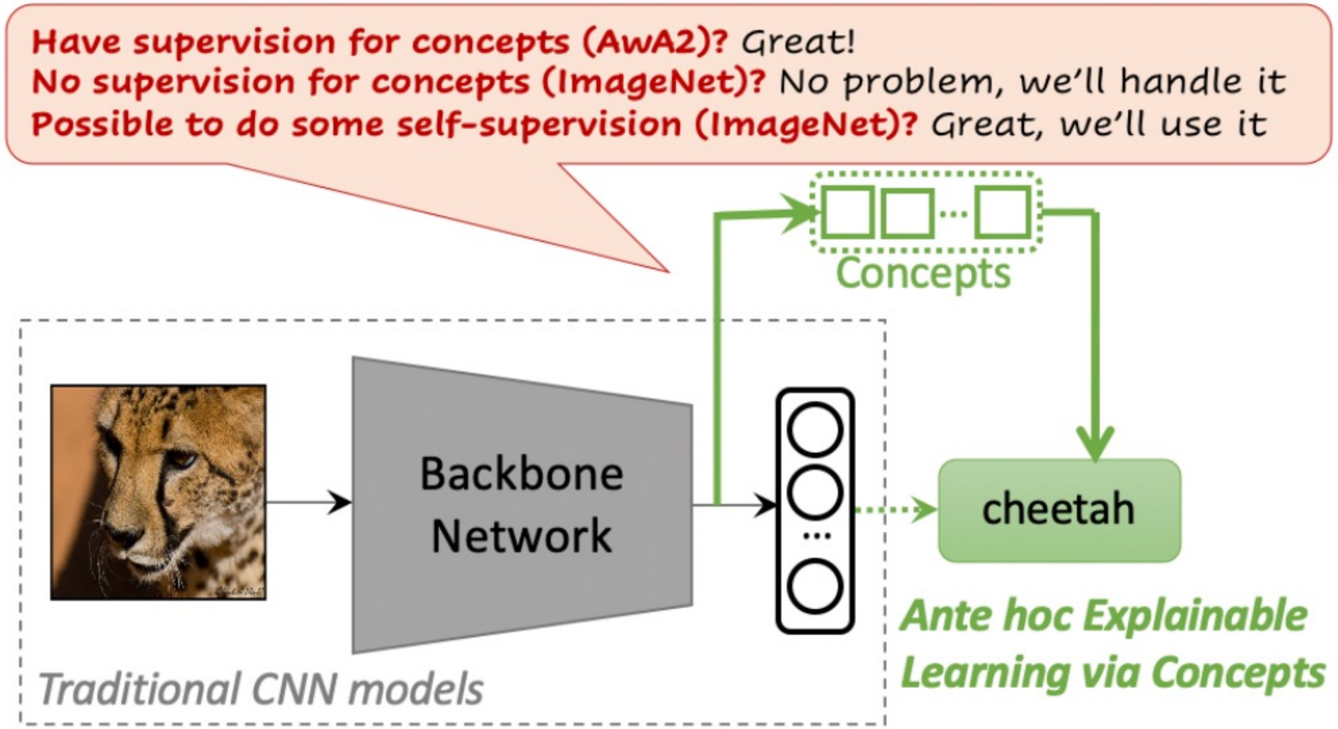
$$TCAV_c = \frac{|x \in \chi^c : S_c(x) > 0|}{|\chi^c|}.$$

Ramaswamy *et al.*: Conceptual information in data sets is often less salient and more difficult to learn than the class of information they purport to explain.

Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV)." *ICML*, 2018.
Ramaswamy, Vikram V., et al. "Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability." *CVPR*. 2023.

2.1 Traditional Method

Concept-based Methods

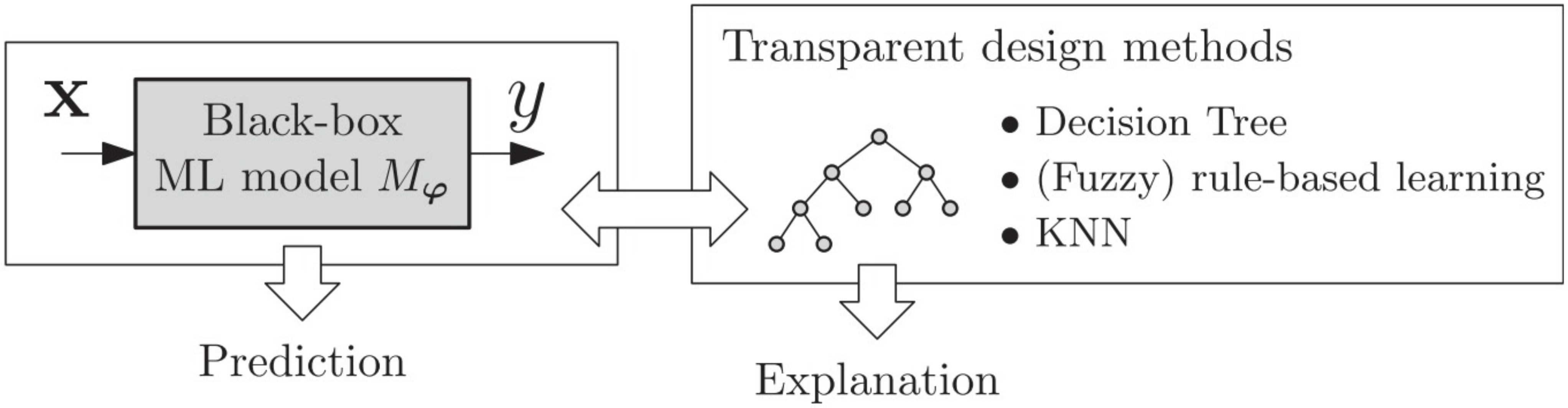


Self-Explaining Neural Networks
Annotated semantic concepts are explicitly learned during the model learning process, and category features and concept information are combined when inferring categories. Its interpretability lies in the semantic concepts generated when the model makes decisions.

Sarkar, Anirban, et al. "A framework for learning ante-hoc explainable models via concepts." *CVPR*. 2022.

2.1 Traditional Method

Agent model-based Methods

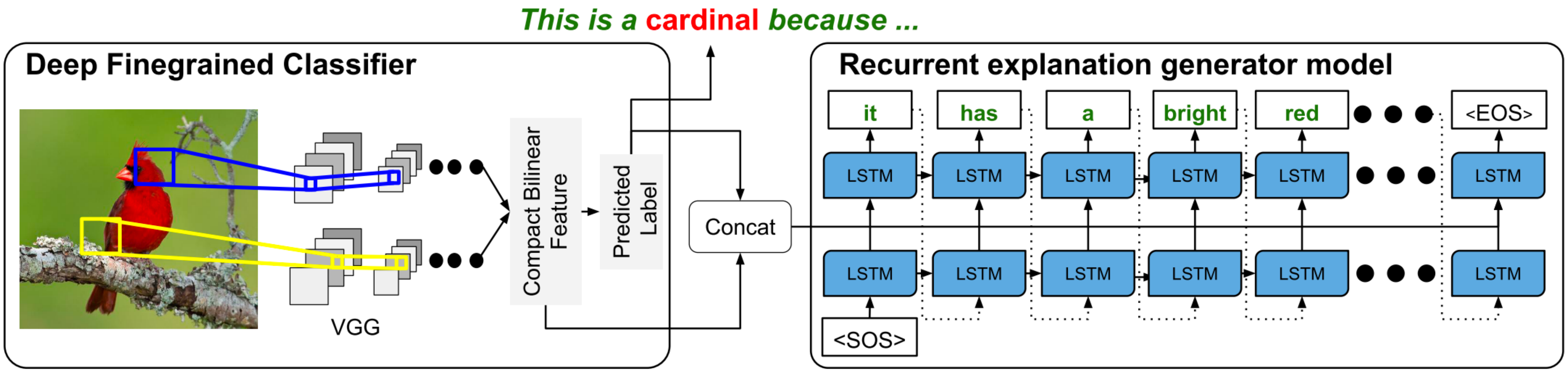


Mapping an uninterpretable black-box system into a white-box twin that is easier to explain. But it usually affects the performance of the final model.

Arrieta, Alejandro Barredo, et al. "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI." *Information Fusion* 58 (2020): 82-115.

2.1 Traditional Method

Multi-modal-based Methods

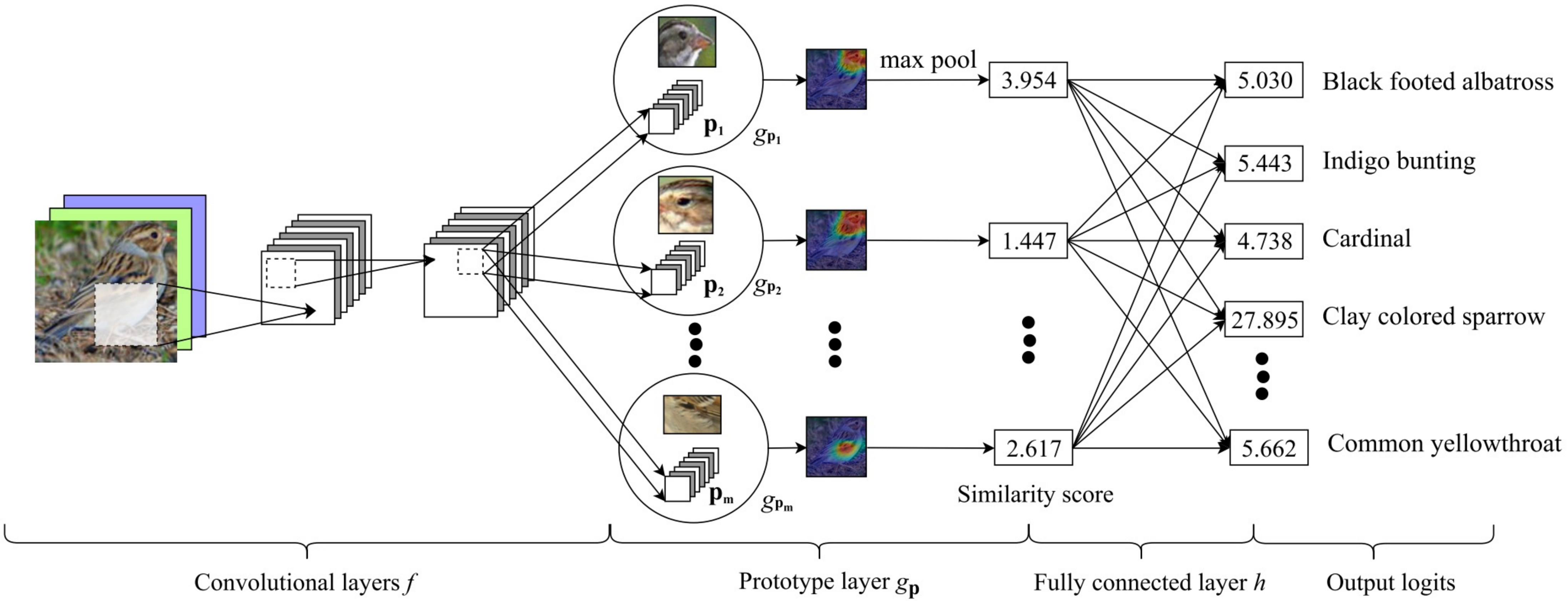


Interpreting a black box model with an uninterpretable model is worrisome.

Hendricks, Lisa Anne, et al. "Generating visual explanations." *ECCV*, 2016.

2.1 Traditional Method

Prototype-based Methods

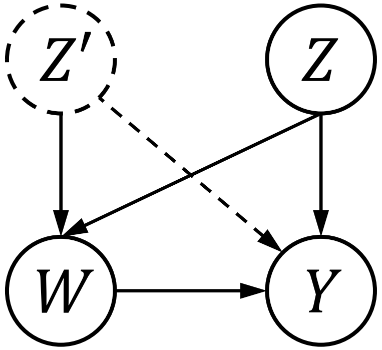


It needs to specify the characteristics of the concept prototype, and has poor versatility and scalability.

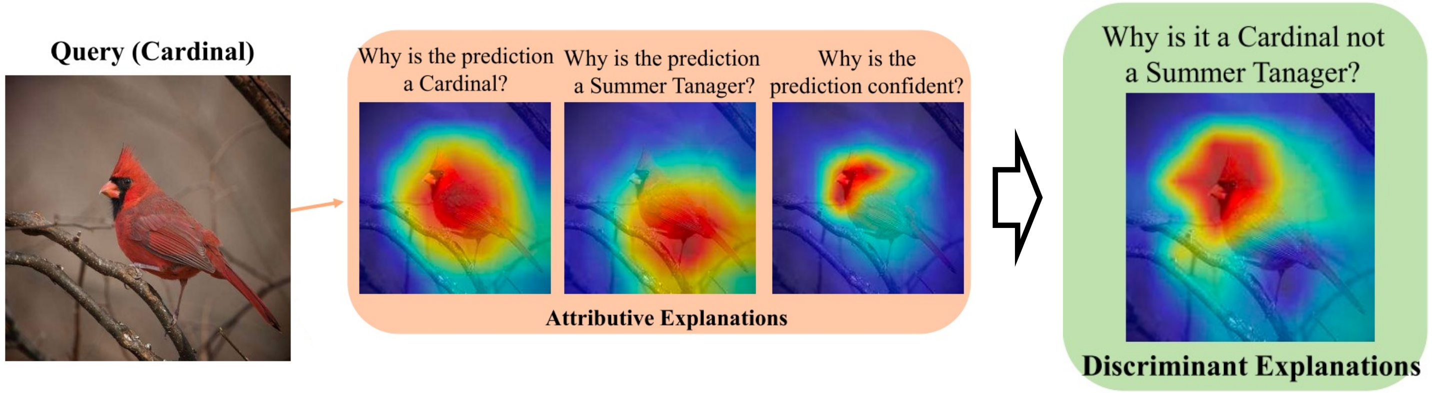
Chen, Chaofan, et al. "This looks like that: deep learning for interpretable image recognition." *NeurIPS* 32 (2019).

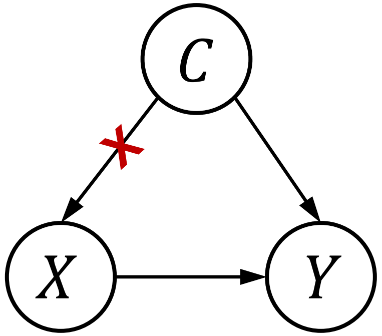
2.1 Traditional Method

Causal-based Methods



Counterfactual Inference





Causal Intervention

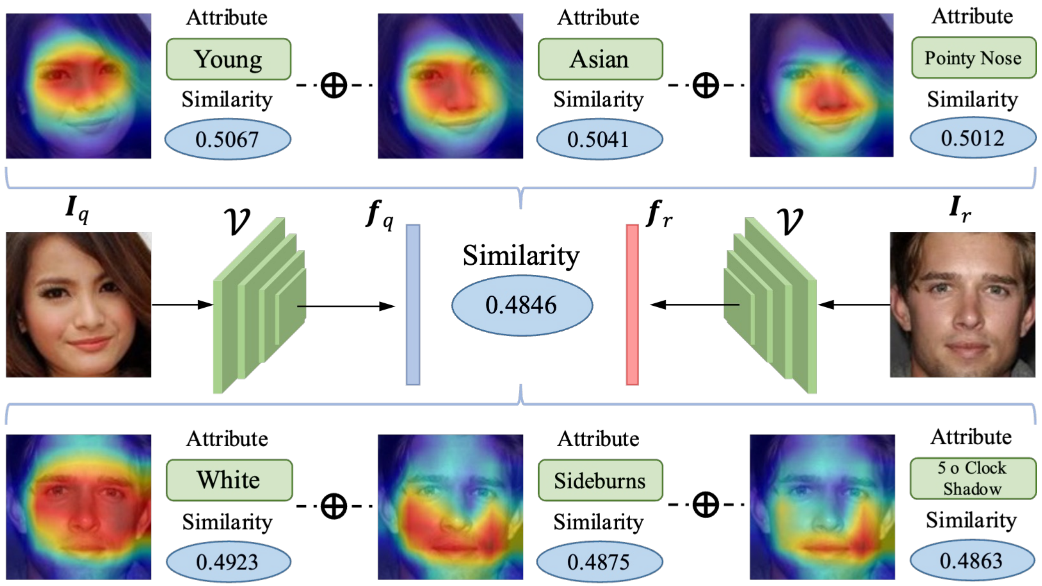
$$P(Y|do(X)) = \sum_{t \in \mathcal{T}} P(Y|X,t)P(t)$$

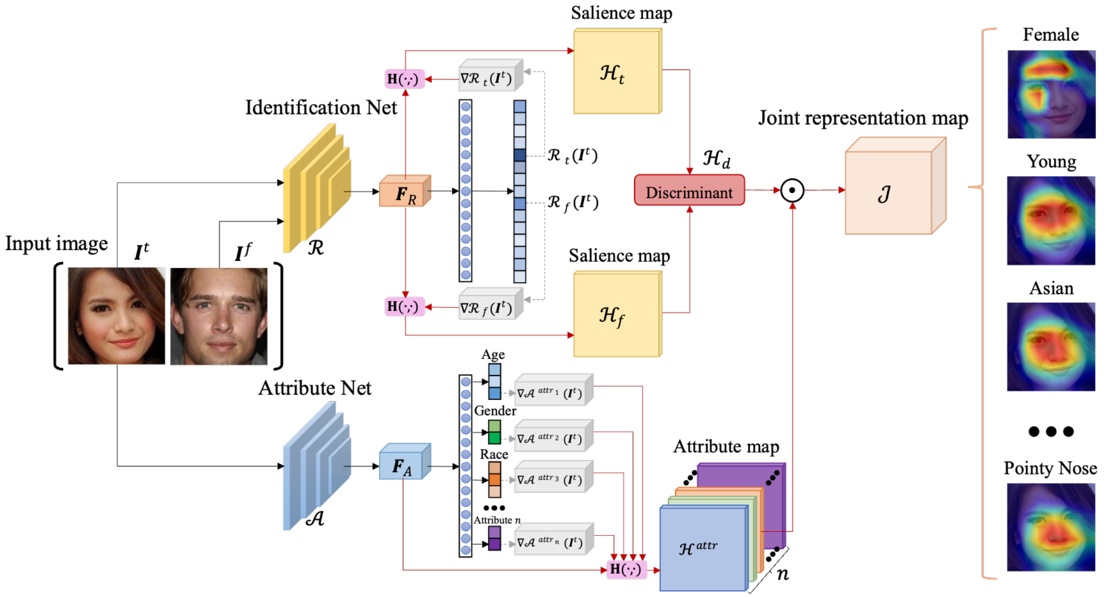
A priori factors that could bias the model

Wang, Pei, and Nuno Vasconcelos. "Scout: Self-aware discriminant counterfactual explanations." *CVPR*. 2020.
Wang, Tan, et al. "Causal attention for unbiased visual recognition." *ICCV*. 2021.

2.1 Traditional Method

Multi explanations output





Sim2Word interprets model via

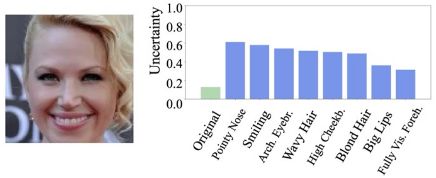
- Salience Maps

What regions does the model focus on?

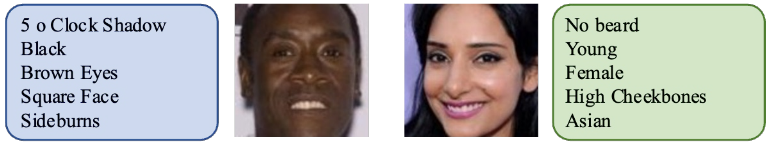
- Textual Description

The most characteristic attribute is the **pointy**

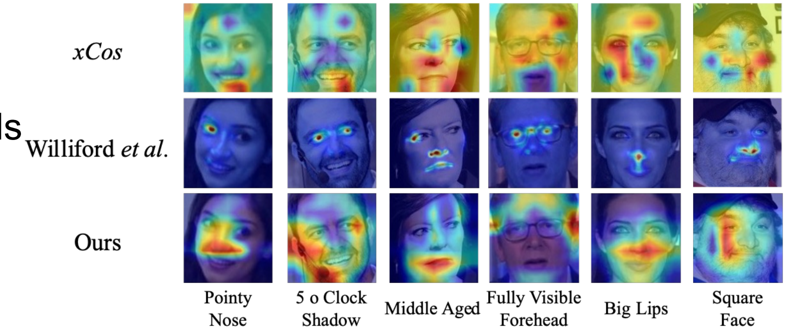
- Numerical Score



Top-5 most characteristic attribute



SOTA methods comparison

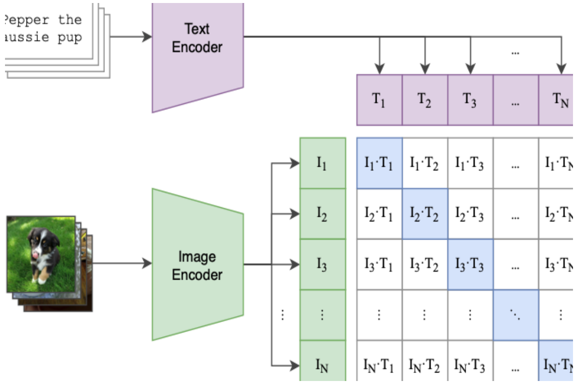


Ruoyu Chen *et al.* "Sim2Word: Explaining Similarity with Representative Attribute Words via Counterfactual Explanations." *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* (2022).

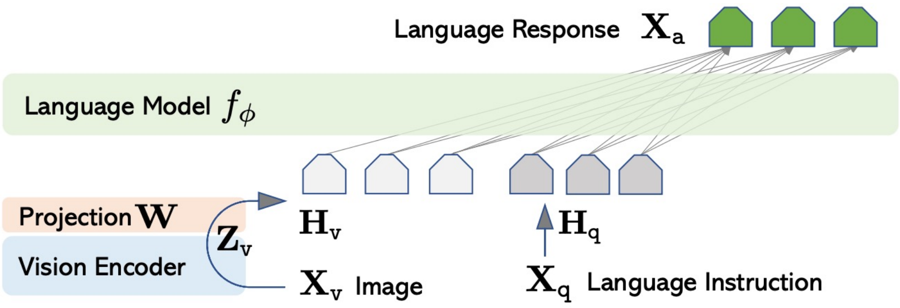
2. Interpretation for Large Model

- Tradition Method
- Category and Challenge
- CLIP Interpretation
- Explainable Generative AI
- Interpret and Enhance Model Performance During Training

2.2 Category and Challenge



Multimodal Embedding Representation Foundation Model



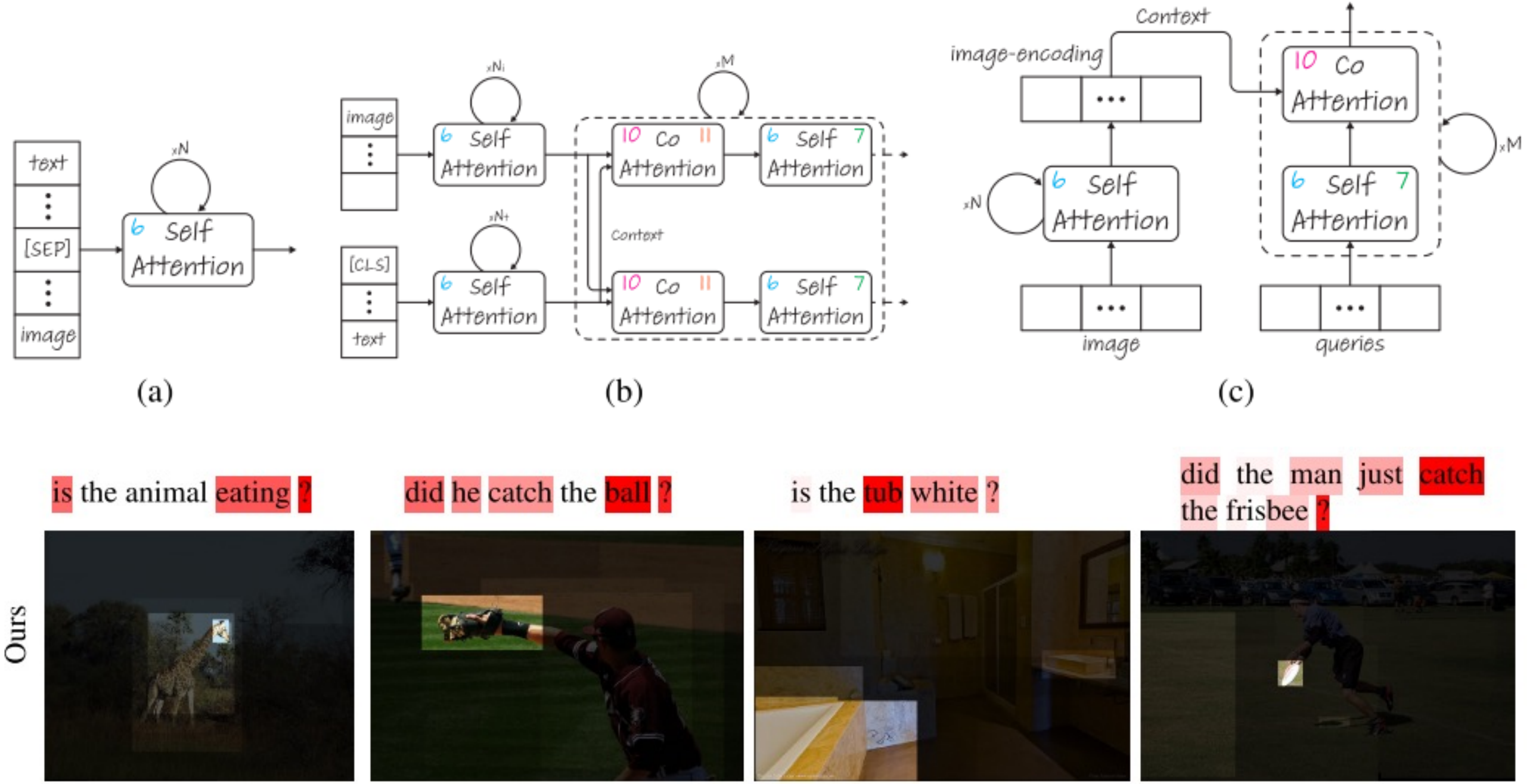
Generative Foundation Model

Characteristic	✓ Multi-stream architecture ✓ Transformer architecture ✓ Encoder model ✓ Zero-shot ability	✓ Large parameter amount ✓ Generative model ✓ Multimodal input ✓ Prompt learning ability
Shortcomings of traditional methods	☐ May not be suitable for explaining models that handle multi-modal inputs. ☐ Fail to consider the unique properties of multimodal models ☐ Methods of ViT and CNN are not universal!	☐ The parameter amount is very large ☐ There is a relative lack of interpretation research ☐ The internal structure is very complicated ☐ Unable to quantitatively metric the generated results
Advantages brought by new models	☐ Higher level semantic understanding capabilities ☐ Can explain any concept to enhance understandability	☐ Rich dialogue content to assist explanations ☐ More convenient human-computer interaction ☐ The generated outputs are more diverse and semantic

2. Interpretation for Large Model

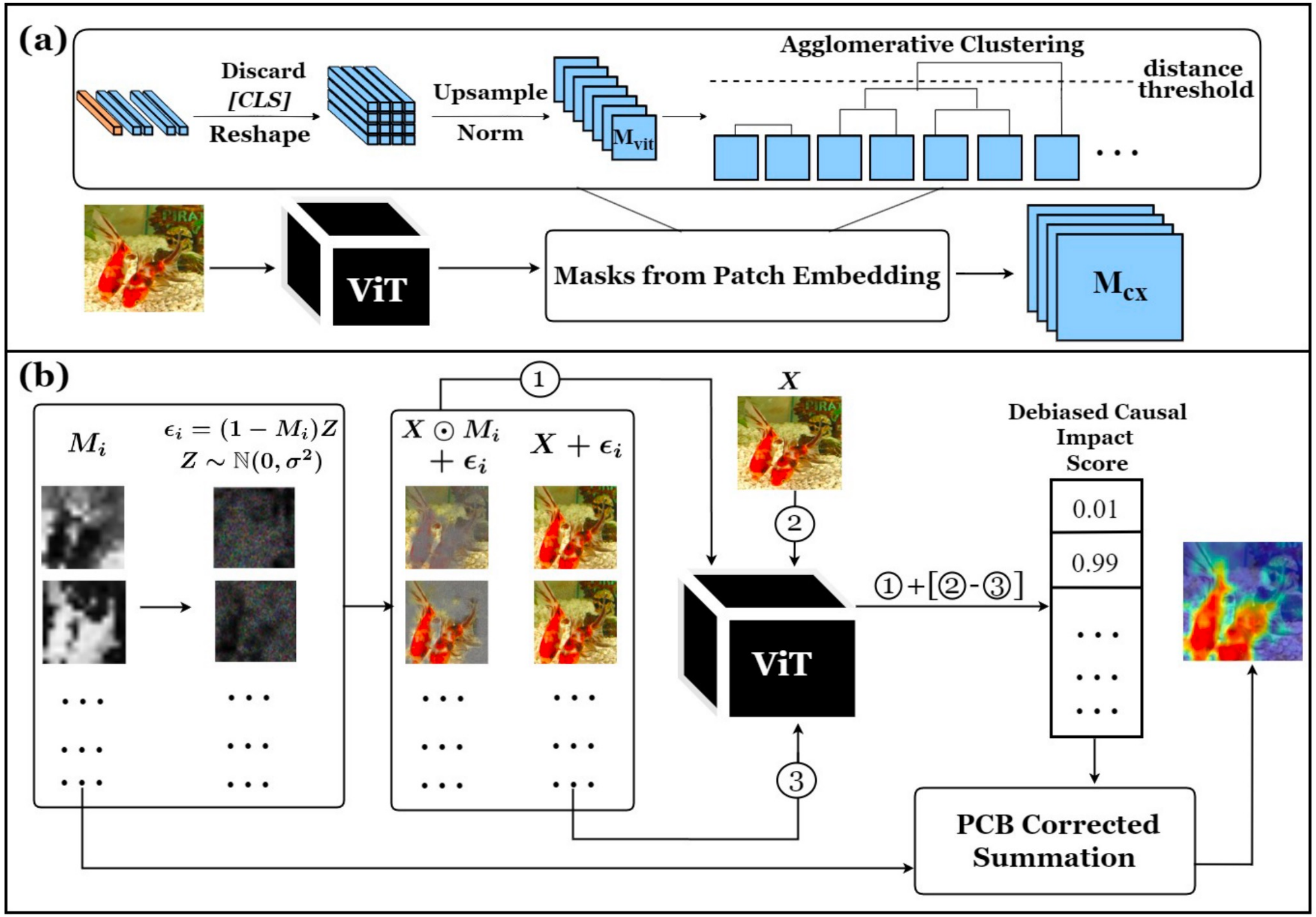
- Tradition Method
- Category and Challenge
- CLIP Interpretation
- Explainable Generative AI
- Interpret and Enhance Model Performance During Training

2.3 CLIP Interpretation



Chefer, Hila, Shir Gur, and Lior Wolf. "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers." *ICCV*. 2021.

2.3 CLIP Interpretation



Xie, Weiyan, et al. "ViT-CX: causal explanation of vision transformers." *IJCAI*. 2023.

2.3 CLIP Interpretation

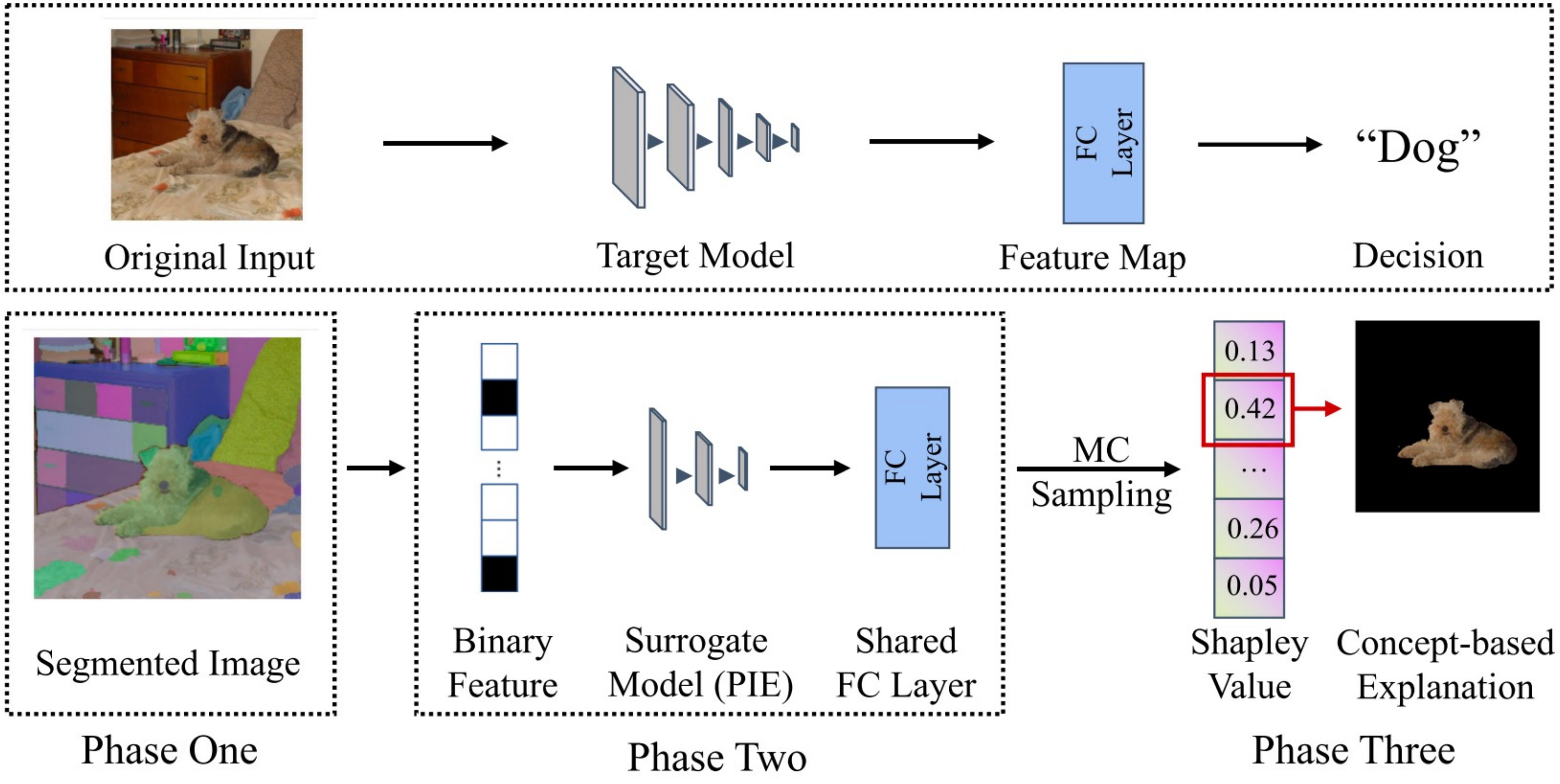
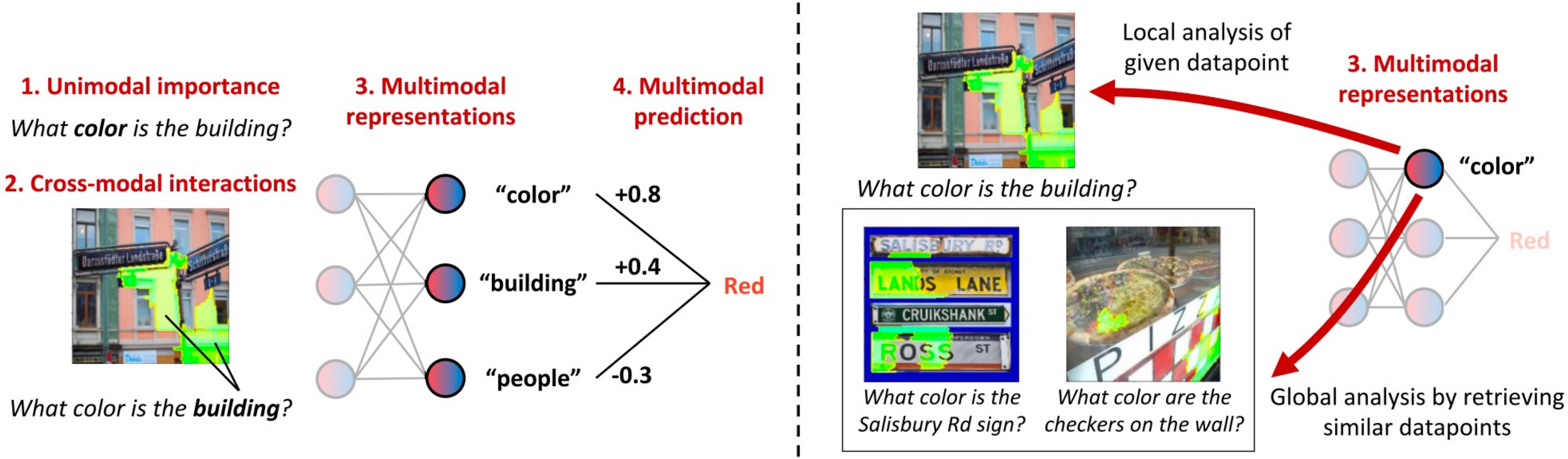


Figure 1: The technical pipeline of EAC in a three-phase form.

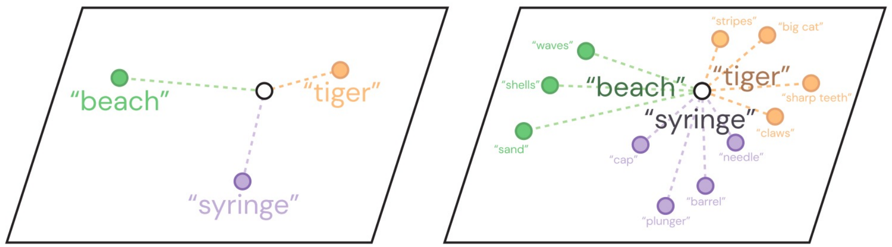
Sun, Ao, et al. " Explain Any Concept: Segment Anything Meets Concept-Based Explanation." *NeurIPS*. 2023.

2.3 CLIP Interpretation



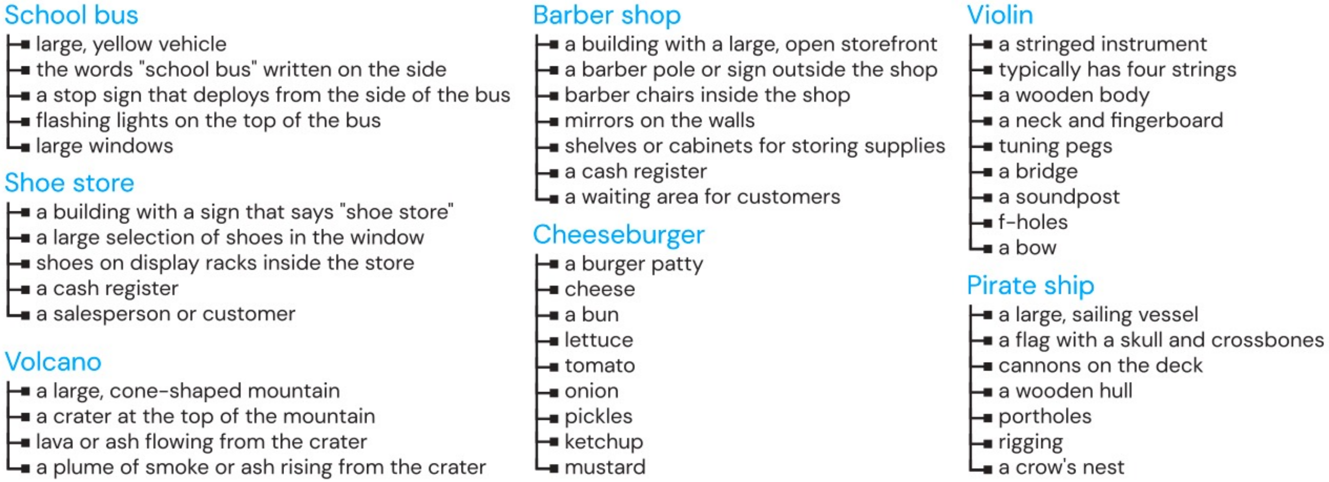
Paul Pu, et al. "MultiViz: Towards Visualizing and Understanding Multimodal Models." *ICLR*. 2023.

2.3 CLIP Interpretation



(a) (b)

Mining large language models to automatically build descriptors



Example of a descriptor pattern generated by GPT-3.

$$s(c, x) = \frac{1}{|D(c)|} \sum_{d \in D(c)} \phi(d, x)$$

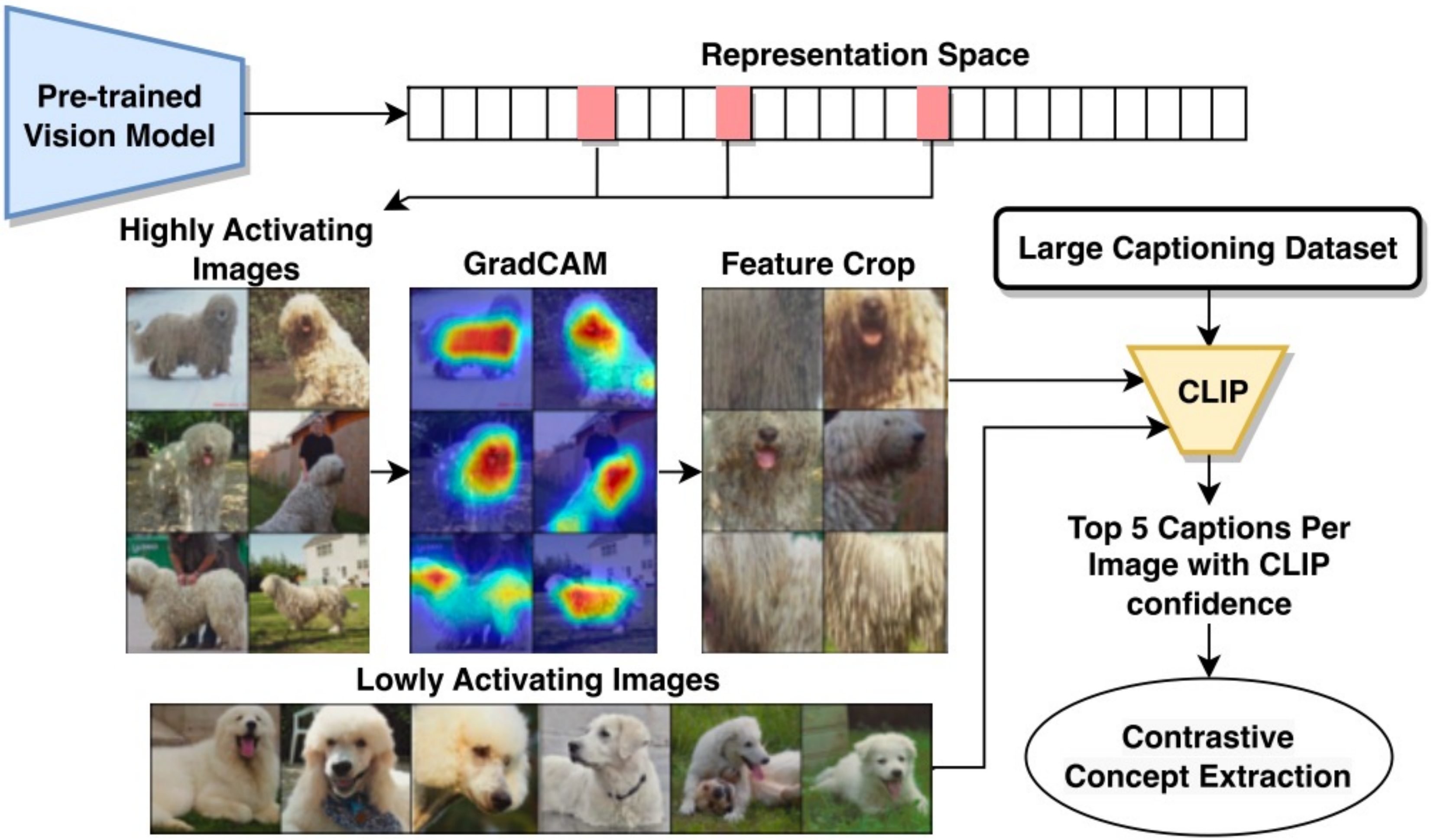
CLIP makes decisions through descriptors.

Architecture for ϕ		ImageNet Ours	ImageNet CLIP	ImageNet Δ	ImageNetV2 Ours	ImageNetV2 CLIP	ImageNetV2 Δ	CUB Ours	CUB CLIP	CUB Δ
Vision Transformers	ViT-B/32	**62.97**	58.46	4.51	**55.52**	51.90	3.62	**52.57**	51.95	0.62
	ViT-B/16	**68.03**	64.05	3.98	**61.54**	57.88	3.66	**57.75**	56.35	1.40
	ViT-L/14	**75.00**	71.58	3.42	**69.3**	65.33	3.97	**63.46**	63.08	0.38
	ViT-L/14@336px	**76.16**	72.97	3.19	**70.32**	66.58	3.74	**65.257**	63.41	1.847
ResNets	RN50	**59.44**	54.81	4.63	**52.98**	49.43	3.55	**48.91**	47.79	1.12
	RN101	**61.88**	57.65	4.23	**55.43**	51.13	4.30	**51.59**	49.46	2.13
	RN50x4	**66.05**	61.48	4.27	**59.23**	54.85	4.38	**55.97**	54.99	0.98
	RN50x16	**69.45**	66.28	3.17	**62.68**	58.8	3.88	**59.03**	57.59	1.44
	RN50x64	**73.19**	69.63	3.56	**66.82**	63.02	3.80	**64.62**	64.24	0.38

The ImageNet and ImageNetV2 models have consistent ~3-5% improvements, and CUB has ~1% improvements.

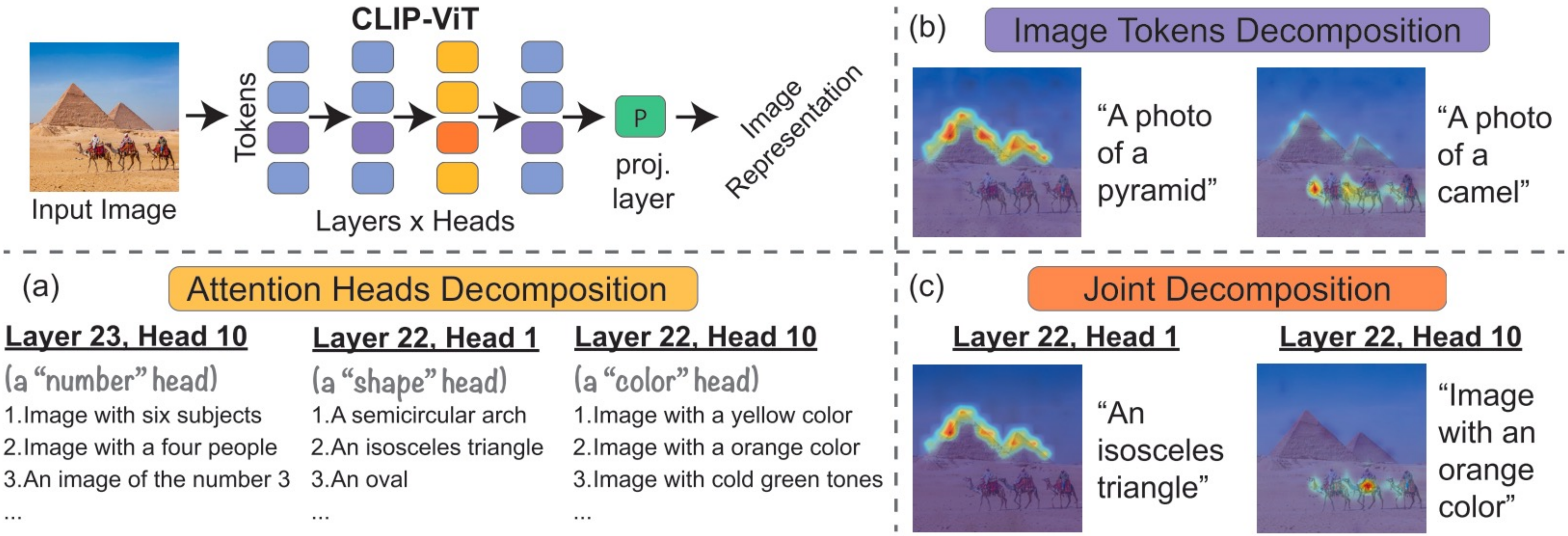
Menon, Sachit, and Carl Vondrick. "Visual Classification via Description from Large Language Models." *ICLR*. 2023.

2.3 CLIP Interpretation



Kalibhat, Neha, et al. "Identifying Interpretable Subspaces in Image Representations." (2023).

2.3 CLIP Interpretation



Gandelsman, Yossi, Alexei A. Efros, and Jacob Steinhardt. "Interpreting CLIP's Image Representation via Text-Based Decomposition." *ICLR*. 2024.

Summary

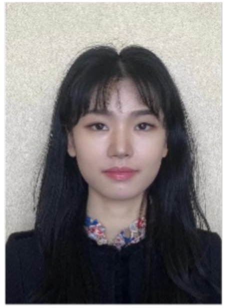
- How to take advantage of the characteristics of the multi-modal encoder foundation model and use text descriptions that are easy for humans to understand to assist interpretation?
- How to understand the internal operating mechanism of the multimodal basic model? Are some of the assumptions correct? Or should it be understood this way?
- How to build a unified causal graph model to cope with the challenges of huge parameter quantities and consumed parameter reasoning in large models?
- How to disentangle features to aid human understanding?
- How to design a more convenient and interpretable model result while adapting to the huge amount of training data in large models.

Less is More: Fewer Interpretable Region via Submodular Subset Selection

Ruoyu Chen

Hua Zhang

Siyuan Liang



Jingzhi Li

Xiaochun Cao

Paper

ICLR 2024

Selected as Oral Presentation (1.16%)

Code

Image Attribution

The main objective in attribution techniques is to highlight the discriminating variables for decision-making.

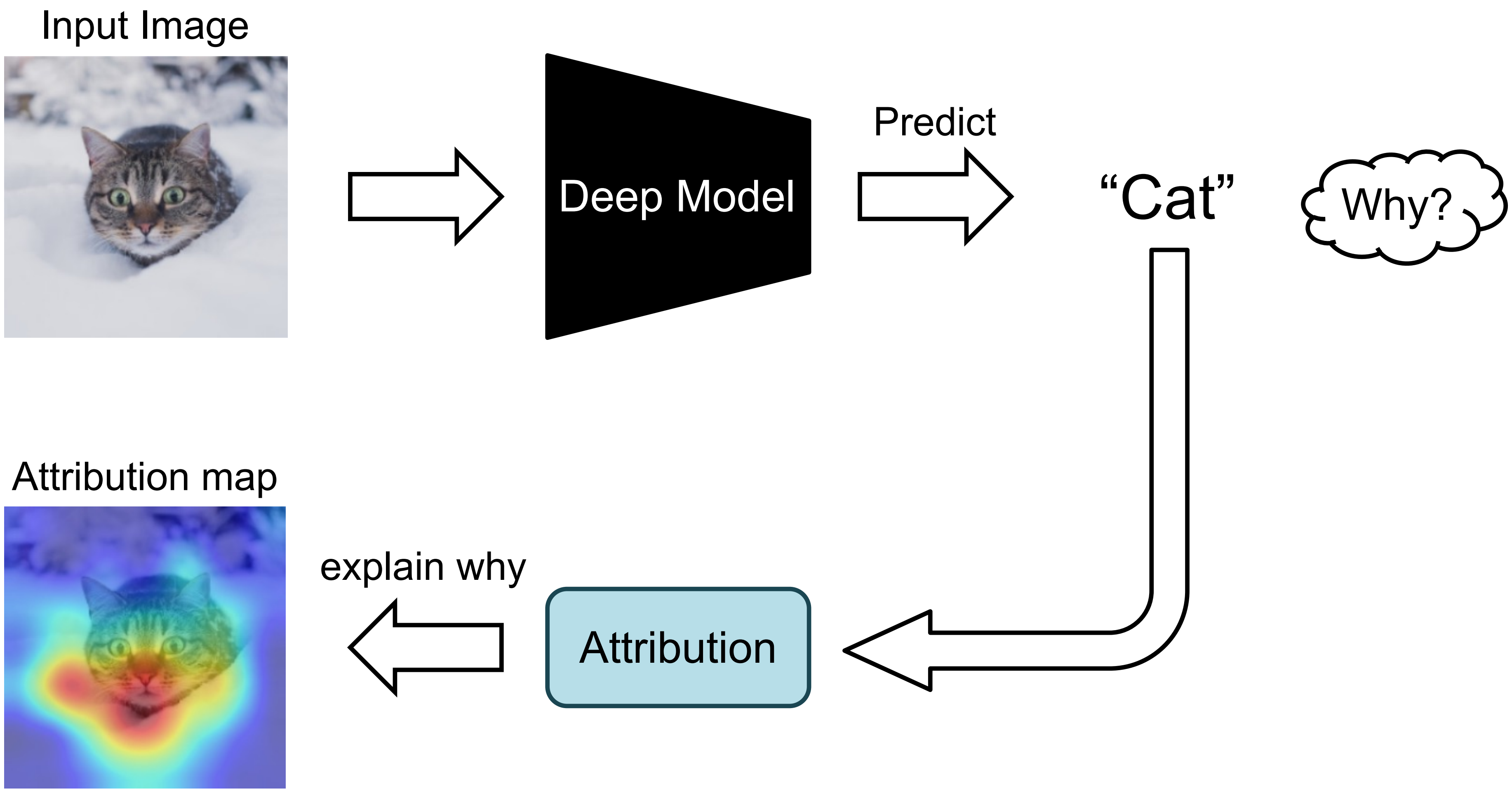
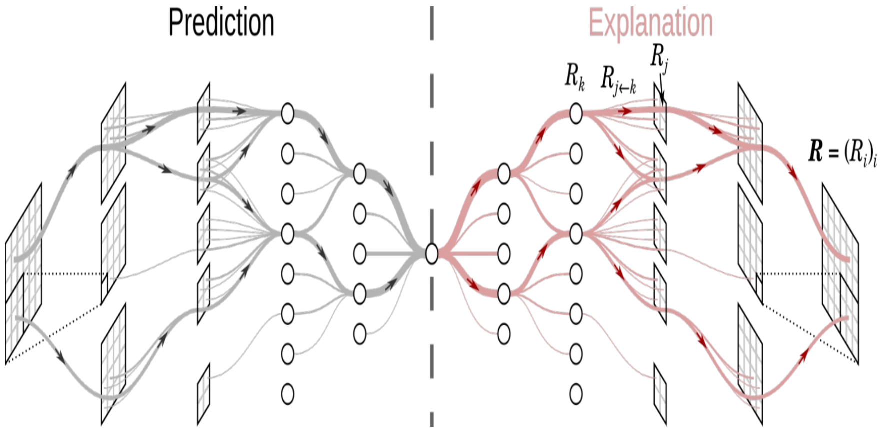


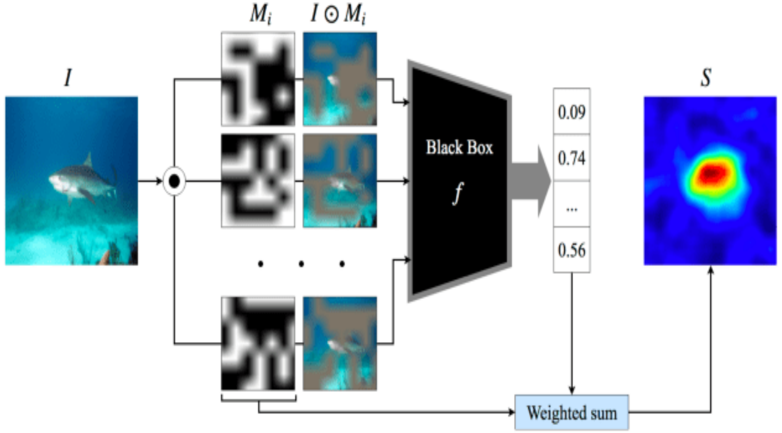
Image Attribution



Based on inner propagation, activation, or gradient

SHAP

Based on sharpley value estimation



Based on perturbation

Challenge in Attribution

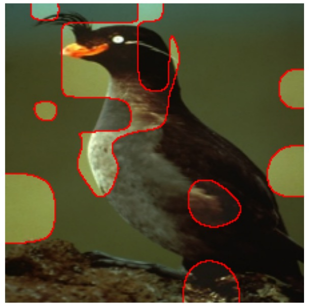
- Existing attribution methods generate <u>*inaccurate small regions*</u> thus misleading the direction of correct attribution.

Input Image $\mathbf{I}$

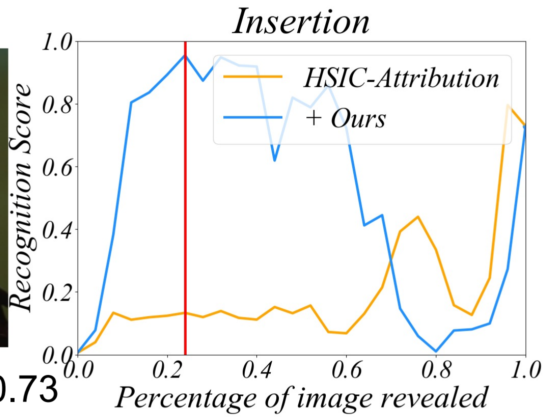
Attribution Map $\mathcal{A}$

Sample correctly predicted by the model

HSIC-Attribution



+ Ours

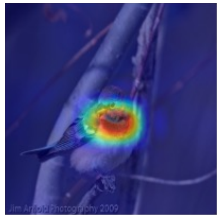


Original prediction confidence: 0.73
Confidence in the region we discovered: **0.95**

- They also can't produce good attribution results for samples with <u>*wrong predictions*</u>.

Input Image $\mathbf{I}$

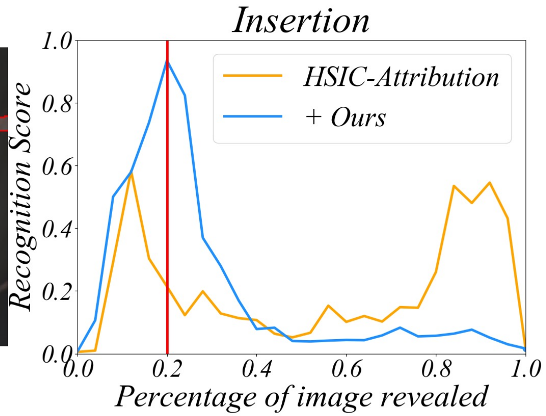
Attribution Map $\mathcal{A}$



Sample incorrectly predicted by the model

HSIC-Attribution

+ Ours



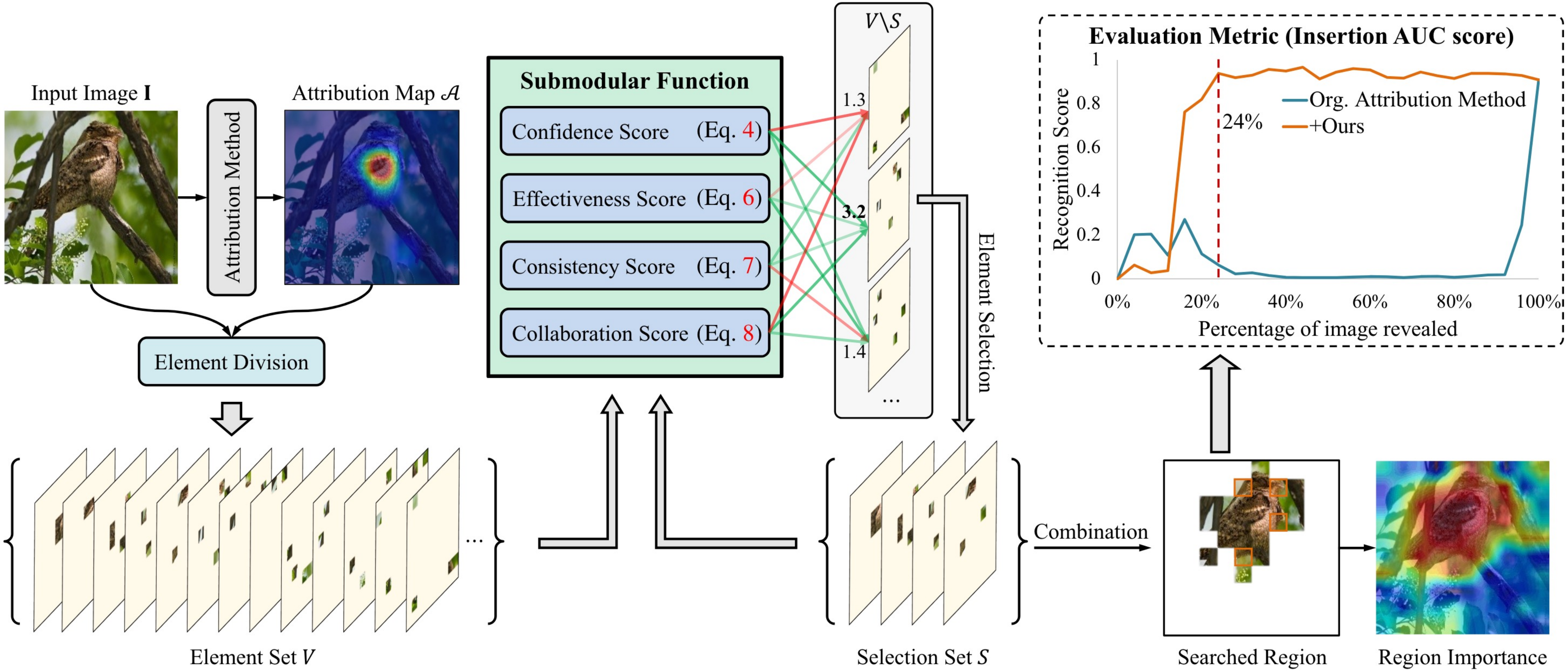
Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting

Our Solution

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

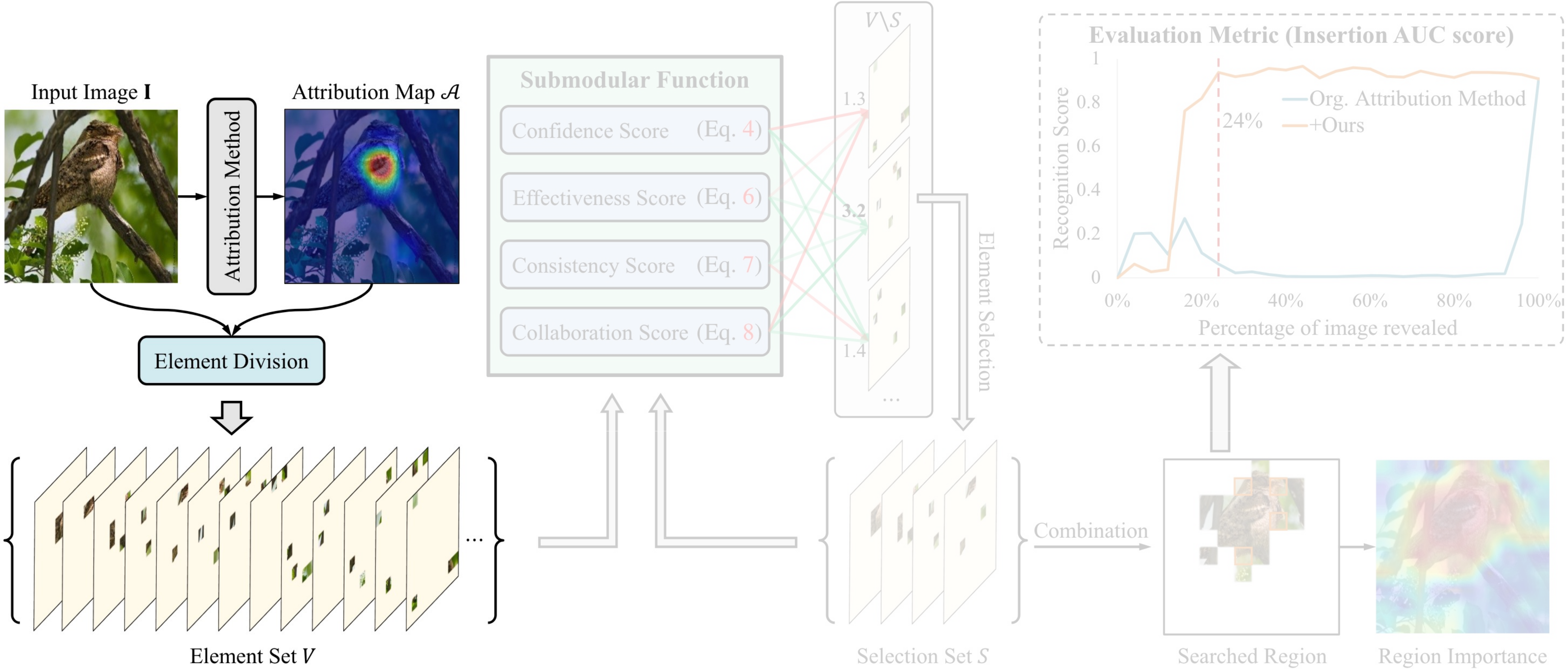
- Reformulate the attribution problem as a *submodular subset selection problem*;
- Employ regional *search* to expand the sub-region set to *alleviate the insufficient dense of the attribution region*;
- A novel *submodular mechanism* is constructed to *limit the search for regions with wrong class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

Method



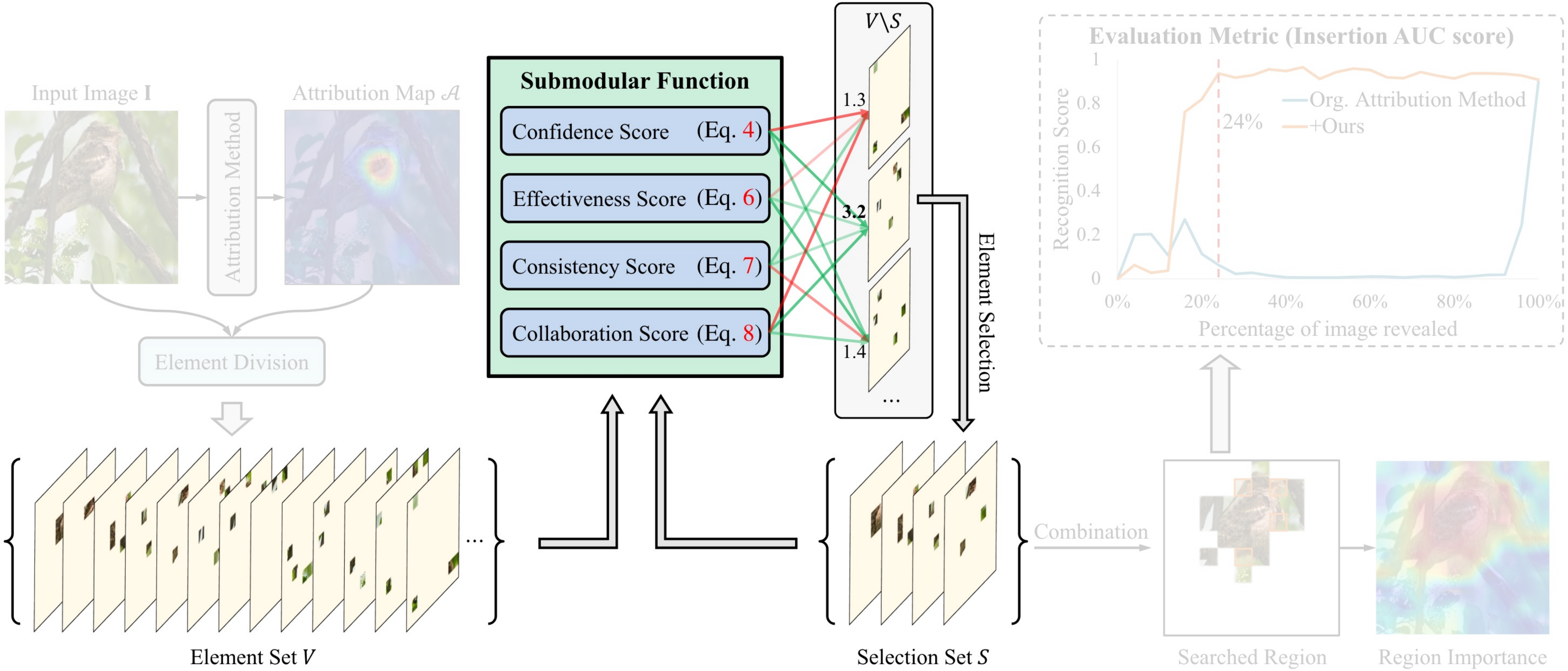
Input Image I
Attribution Method
Attribution Map $\mathcal{A}$
Element Division
Element Set V
Submodular Function
Confidence Score (Eq. 4)
Effectiveness Score (Eq. 6)
Consistency Score (Eq. 7)
Collaboration Score (Eq. 8)
$V \backslash S$
1.3
3.2
1.4
Element Selection
Selection Set S
Combination
Searched Region
Region Importance
Evaluation Metric (Insertion AUC score)
Recognition Score
Percentage of image revealed
Org. Attribution Method
+Ours
24%
0 0.2 0.4 0.6 0.8 1
0% 20% 40% 60% 80% 100%

Method



1. Sub-Region Division

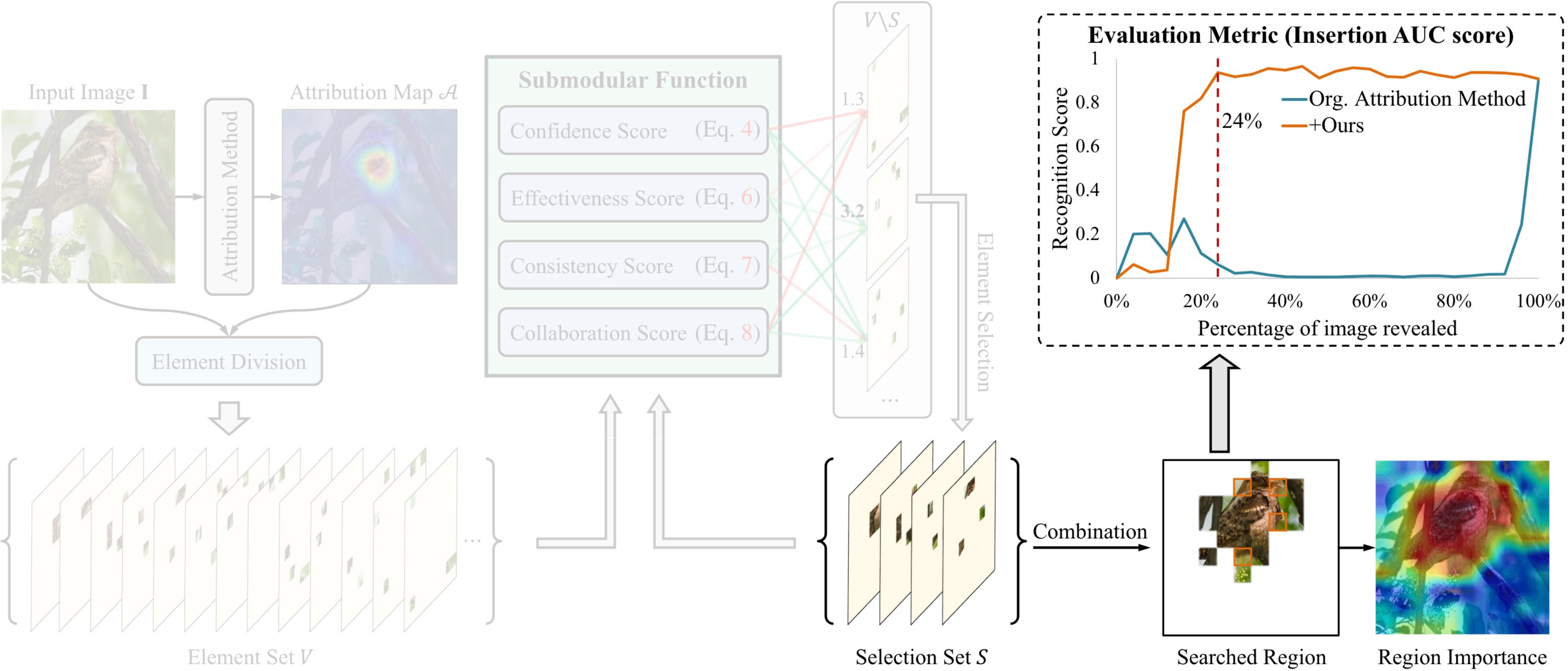
Method



1. Sub-Region Division

2. Sub-Region Selection

Method

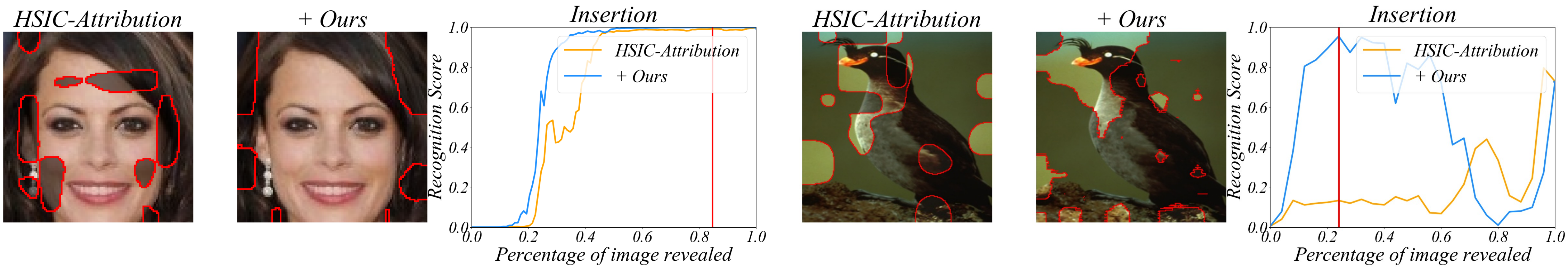


1. Sub-Region Division

2. Sub-Region Selection

3. Combination and Evaluation

Advanced Attribution Results



Use fewer image region but get higher prediction confidence.

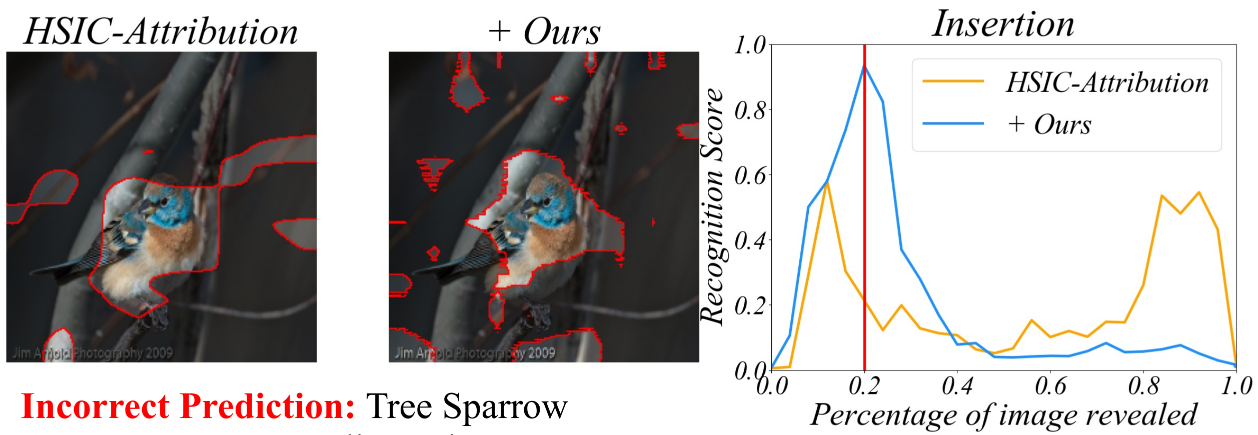
Table 1: Deletion and Insertion AUC scores on the Celeb-A, VGG-Face2, and CUB-200-2011 validation sets.

Method	Celeb-A		VGGFace2		CUB-200-2011	
	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)	Deletion (↓)	Insertion (↑)
Saliency (Simonyan et al., 2014)	0.1453	0.4632	0.1907	0.5612	0.0682	0.6585
Saliency (w/ ours)	**0.1254**	**0.5465**	**0.1589**	**0.6287**	**0.0675**	**0.6927**
Grad-CAM (Selvaraju et al., 2020)	0.2865	0.3721	0.3103	0.4733	0.0810	0.7224
Grad-CAM (w/ ours)	**0.1549**	**0.4927**	**0.1982**	**0.5867**	**0.0726**	**0.7231**
LIME (Ribeiro et al., 2016)	0.1484	0.5246	0.2034	0.6185	0.1070	0.6812
LIME (w/ ours)	**0.1366**	**0.5496**	**0.1653**	**0.6314**	**0.0941**	**0.6994**
Kernel Shap (Lundberg & Lee, 2017)	0.1409	0.5246	0.2119	0.6132	0.1016	0.6763
Kernel Shap (w/ ours)	**0.1352**	**0.5504**	**0.1669**	**0.6314**	**0.0951**	**0.6920**
RISE (Petsiuk et al., 2018)	0.1444	0.5703	0.1375	0.6530	0.0665	0.7193
RISE (w/ ours)	**0.1264**	**0.5719**	**0.1346**	**0.6548**	**0.0630**	**0.7245**
HSIC-Attribution (Novello et al., 2022)	0.1151	0.5692	0.1317	0.6694	0.0647	0.6843
HSIC-Attribution (w/ ours)	**0.1054**	**0.5752**	**0.1304**	**0.6705**	**0.0613**	**0.7262**

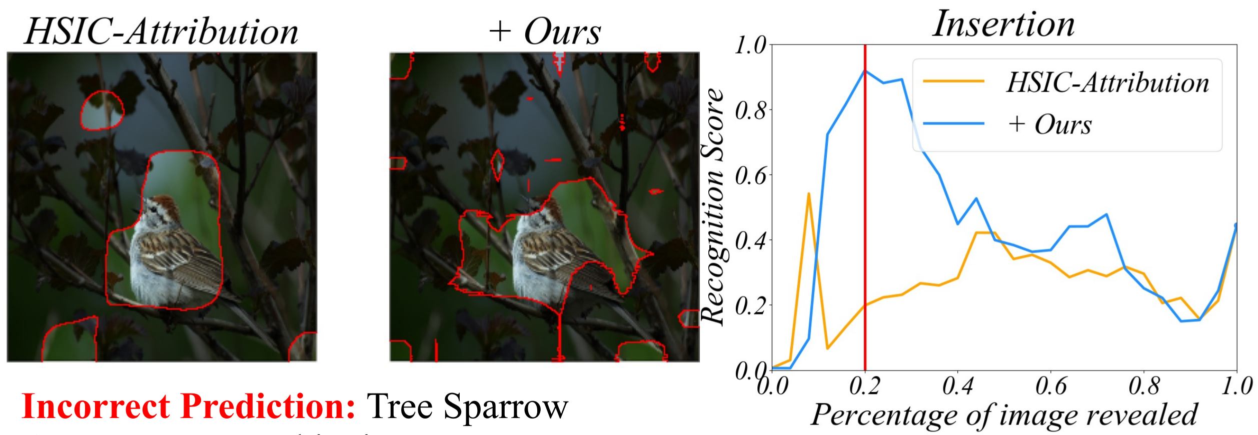
Deletion: *4.9%* improvement

Insertion: *2.5%* improvement

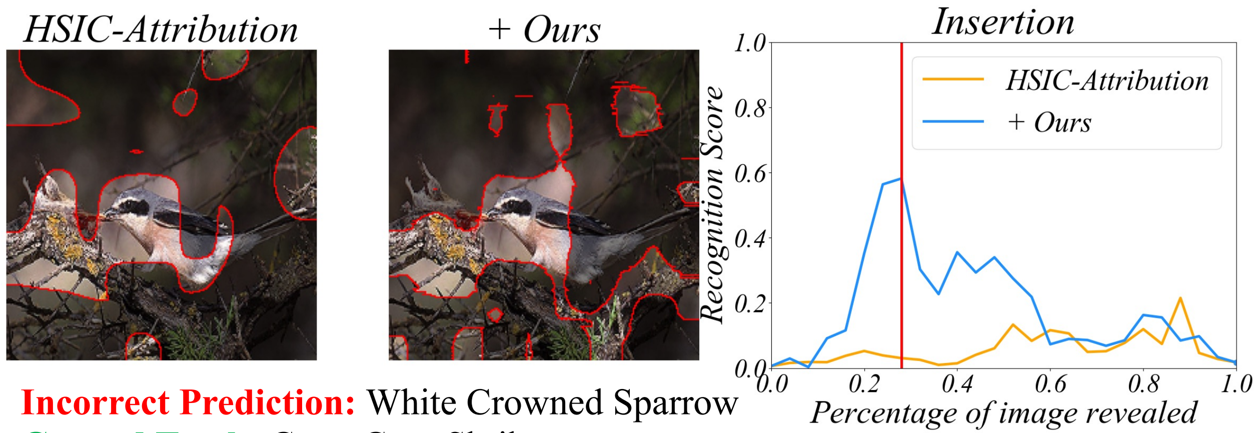
Debugging Model Prediction Errors



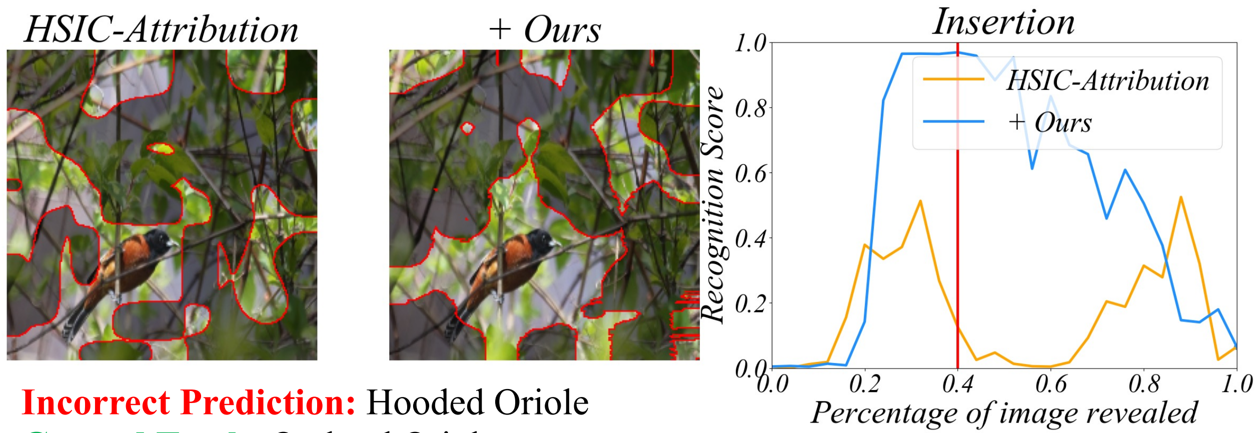
Incorrect Prediction: Tree Sparrow
Ground Truth: Lazuli Bunting



Incorrect Prediction: Tree Sparrow
Ground Truth: Chipping Sparrow



Incorrect Prediction: White Crowned Sparrow
Ground Truth: Great Grey Shrike

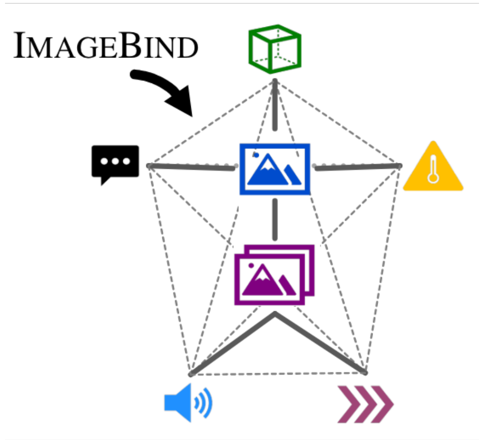


Incorrect Prediction: Hooded Oriole
Ground Truth: Orchard Oriole

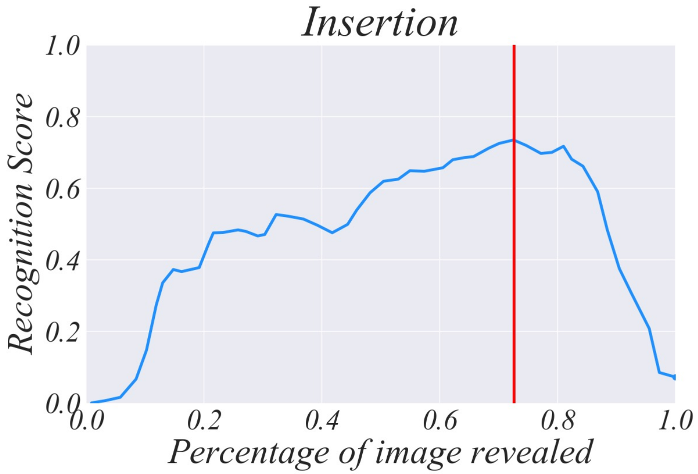
Dark regions are the cause of model prediction errors

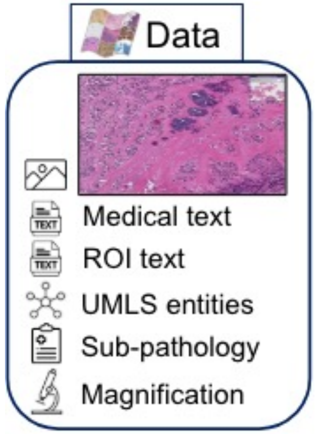
Scale to Large Model

Explaining multimodal foundation model

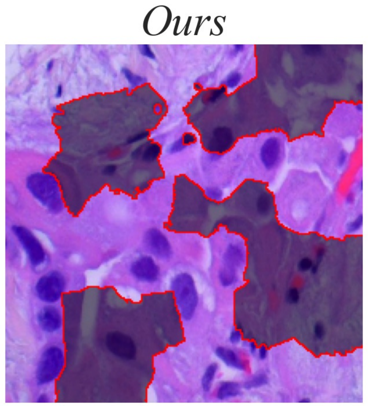


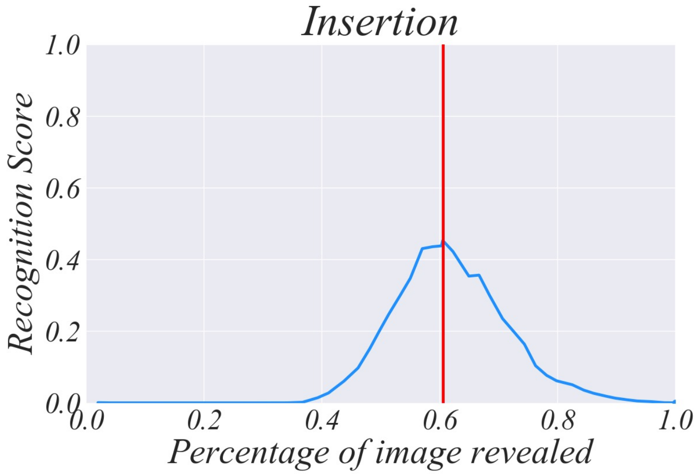
ImageBind is a Transformer-based multimodal model that can generate joint embeddings across seven modalities





Quilt-1M is a medical multimodal model, which outperforms state-of-the-art models on both zero-shot and linear probing tasks for classifying new histopathology images





Easy to scale to large model.

Chen, Ruoyu, et al. "Less is More: Fewer Interpretable Region via Submodular Subset Selection." *ICLR*. 2024. **(Oral, 1.16%)**

Summary

- A new perspective on image attribution: submodular subset selection
- A general attribution method for image classification problems that can be easily scaled to large models
- Can effectively discover potential regions that cause model’s wrong prediction

2. Interpretation for Large Model

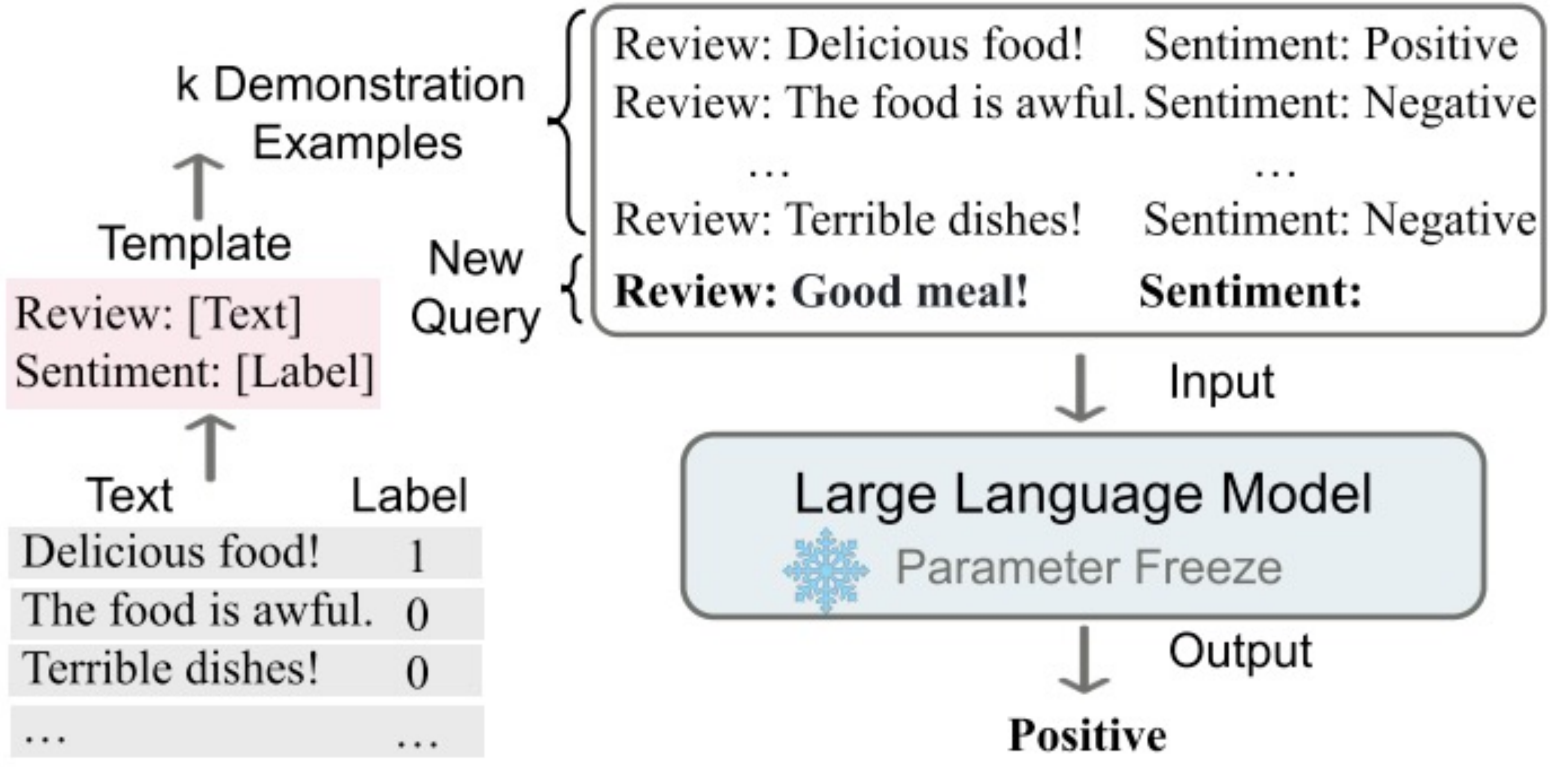
- Tradition Method
- Category and Challenge
- CLIP Interpretation
- Explainable Generative AI
- Interpret and Enhance Model Performance During Training

2.4 Explainable Generative AI

Examples of generative AI

Input/Output	Description	Example
Text to Text	Input: Raw text. Output: Processed or generated text.	ChatGPT-3.5
Text to Image/Video	Input: Descriptive text or prompt. Output: Generated image/video.	DALL-E  Sora 
Image/Video to Text	Input: Image/video and text. Output: Textual interpretation and answer.	GPT-4
Images, Actions to Actions	Input: Images depicting actions. Output: Generated action sequences.	Gato
Image to Image	Input: Image/noise. Output: Generated images.	Stable Diffusion 
Text to 3D	Input: Text describing object. Output: 3D representation of object.	Magic3d

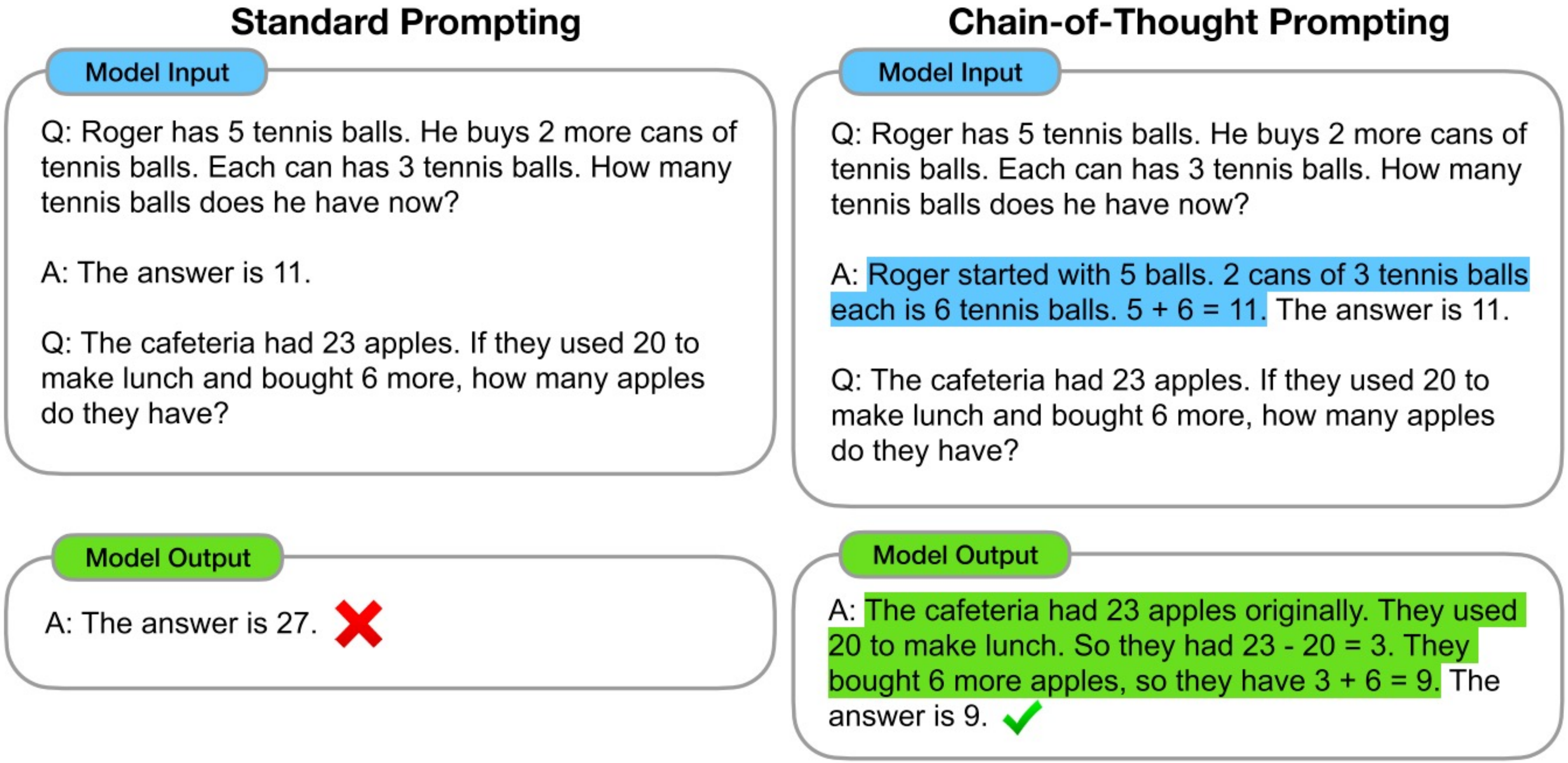
2.4 Explainable Generative AI



In-context learning

Dong, Qingxiu, et al. "A survey for in-context learning." *arXiv preprint arXiv:2301.00234* (2022).

2.4 Explainable Generative AI

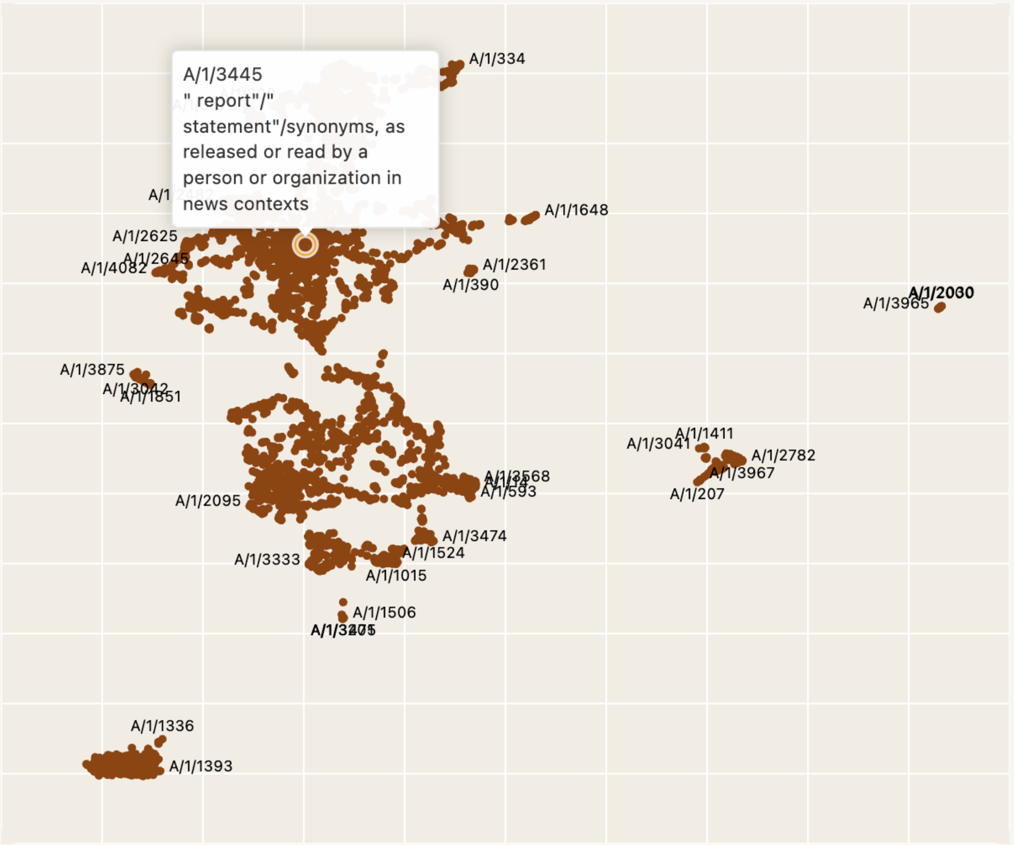


Wei, Jason, et al. "Chain-of-thought prompting elicits reasoning in large language models." *NeurIPS* 2022.

2.4 Explainable Generative AI

Anthropic, the company behind Claude, releases Poster, using sparse autoencoders, a large number of interpretable features are extracted from a single-layer Transformer.

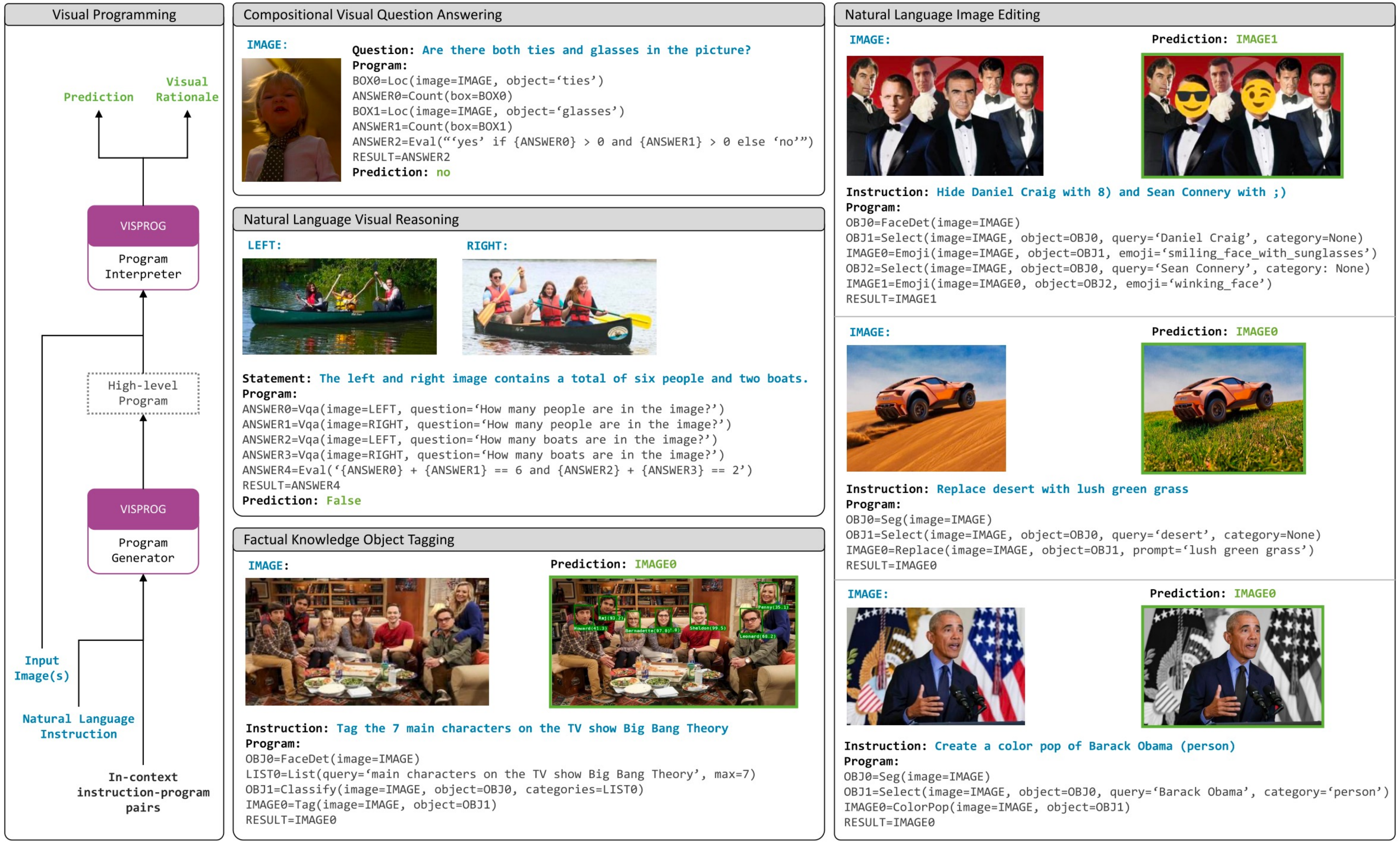
Cluster #49	A/0/307	This feature fires for references to citations in scientific pa...
	A/0/311	This feature fires for reference citations in academic paper...
	A/1/776	Years in some citation notation
	A/1/1538	Citations in a [@author] or [@authoryear] format
	A/1/1875	Markdown Citation (Predict year)
	A/1/2252	" [@"
	A/1/2237	[Ultralow density cluster]
Cluster #42	A/0/126	This feature seems to fire on section headings, specifically ...
	A/1/357	"ref" in [context]
	A/1/1469	"s"/"sec" after "{#", section reference in some markup
	A/1/3841	"Sec"
	A/1/3898	Section number in {#SecX}
	A/1/4083	" {#"
	A/1/2129	"." in [context]
	A/1/553	"](#" in [context]
Cluster #43	A/0/8	This feature attends to text formatting markups such as ref...
	A/0/398	This feature attends to references to figures and tables.
	A/0/454	This feature fires on reference/bibliographic citations in LaT...
	A/1/35	"){"
	A/1/366	"type"
	A/1/945	"ref" in [context]
	A/1/1895	"-" in [context]
	A/1/2176	"fig"



Trenton Bricken, *et al.*, "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning." https://transformer-circuits.pub/2023/monosemantic-features. 2023.

2.4 Explainable Generative AI

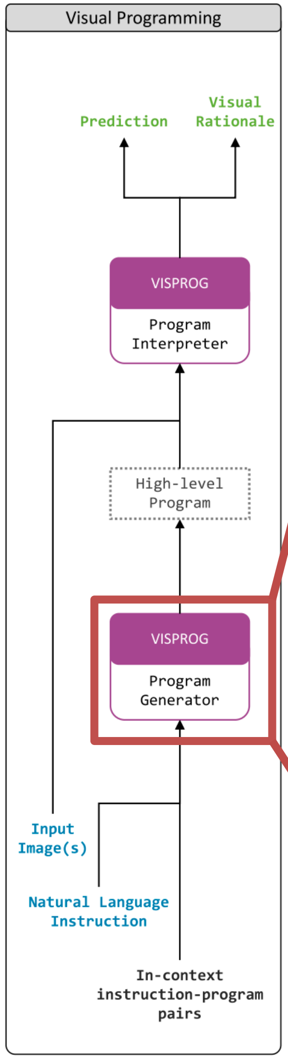
VisProg CVPR 2023 Best Paper



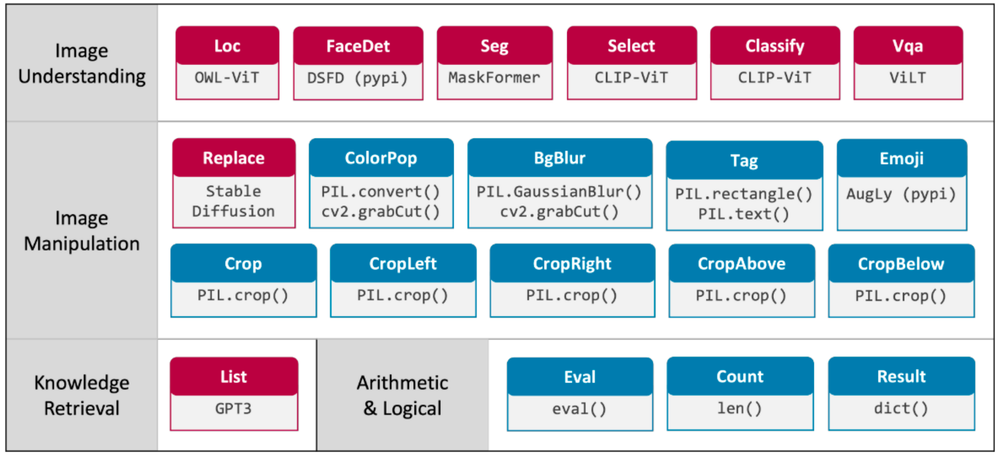
Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023.

2.4 Explainable Generative AI

VisProg CVPR 2023 Best Paper



VisProg's program generation process.

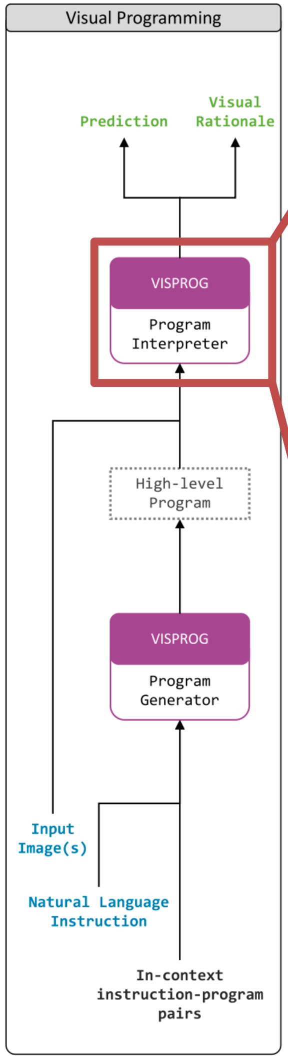


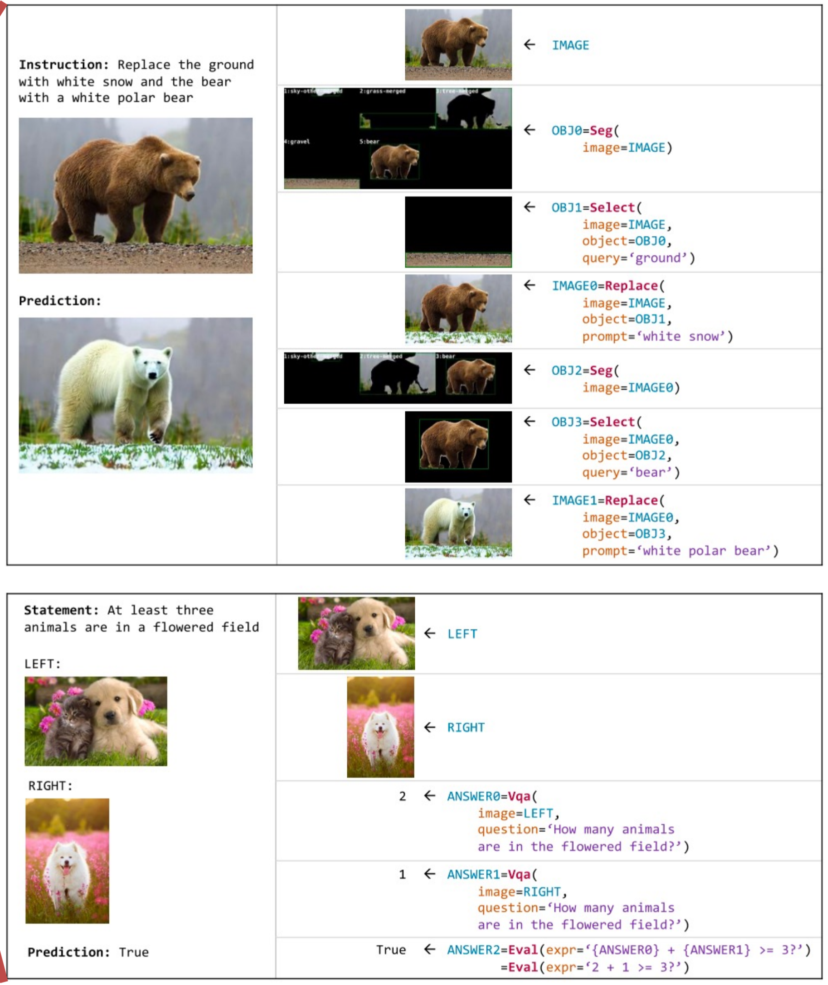
Function modules already supported by VisProg.

Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023.

2.4 Explainable Generative AI

VisProg CVPR 2023 Best Paper





Visual principles generated by VisProg.

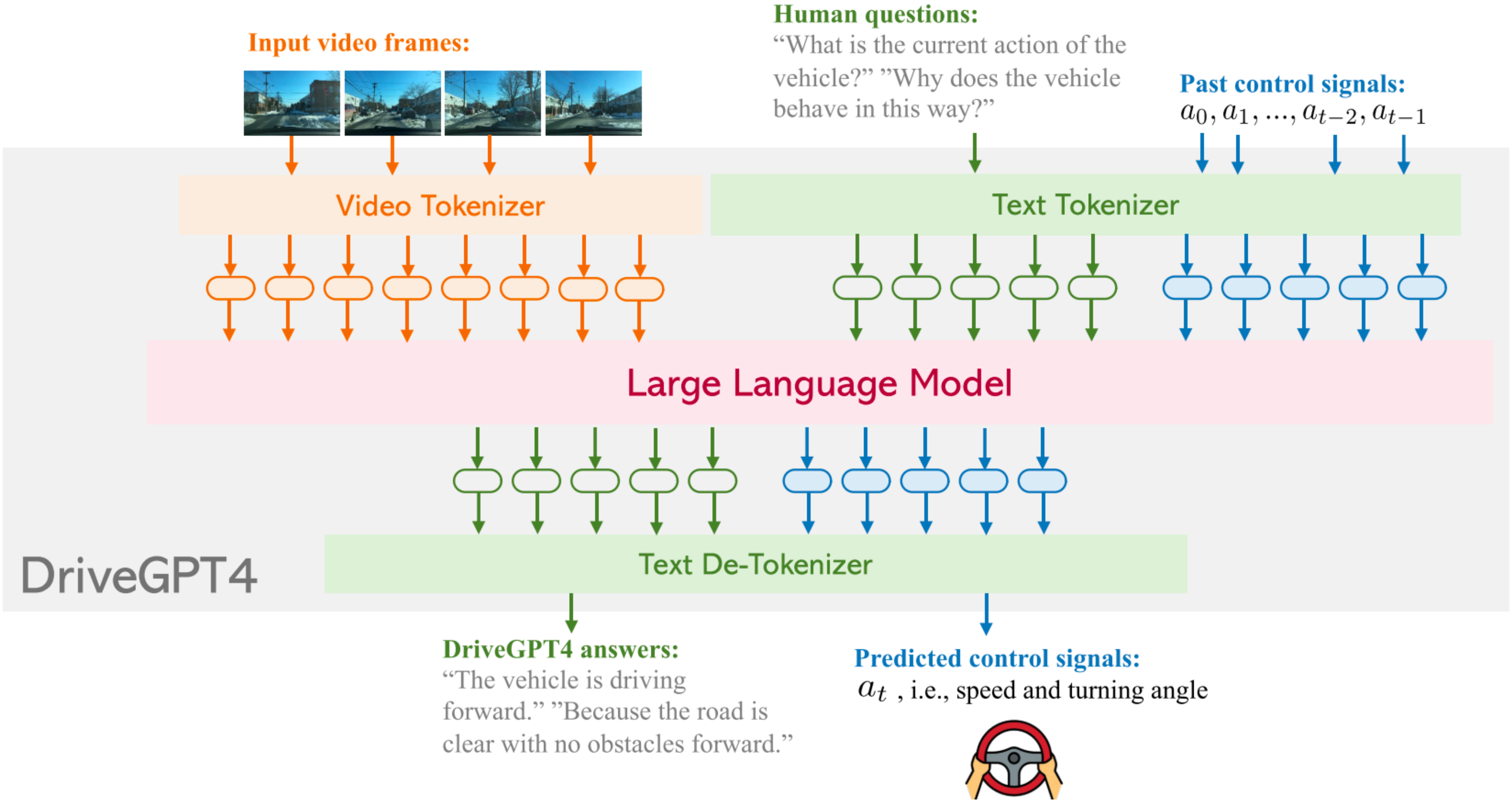
Task	Input	Output	Modules
Compositional Visual QA (GQA)	Image + Question	Text	Loc, Vqa, Eval, Count, Crop, CropLeft, CropRight, CropAbove, CropBelow
Reasoning on Image Pairs (NLVR)	Image Pair + Statement	True/False	Vqa, Eval
Factual Knowledge Object Tagging	Image + Instruction	Image	FaceDet, List, Classify, Loc, Tag
Image Editing with Natural Language	Image + Instruction	Image	FaceDet, Seg, Select, Replace, ColorPop, BgBlur, Emoji

Evaluate VisProg on a range of different tasks.

Gupta, Tanmay, and Aniruddha Kembhavi. "Visual programming: Compositional visual reasoning without training." *CVPR*. 2023.

2.4 Explainable Generative AI

DriveGPT4



Xu, Zhenhua, et al. "DriveGPT4: Interpretable End-to-end Autonomous Driving via Large Language Model." *arXiv preprint arXiv:2310.01412* (2023).

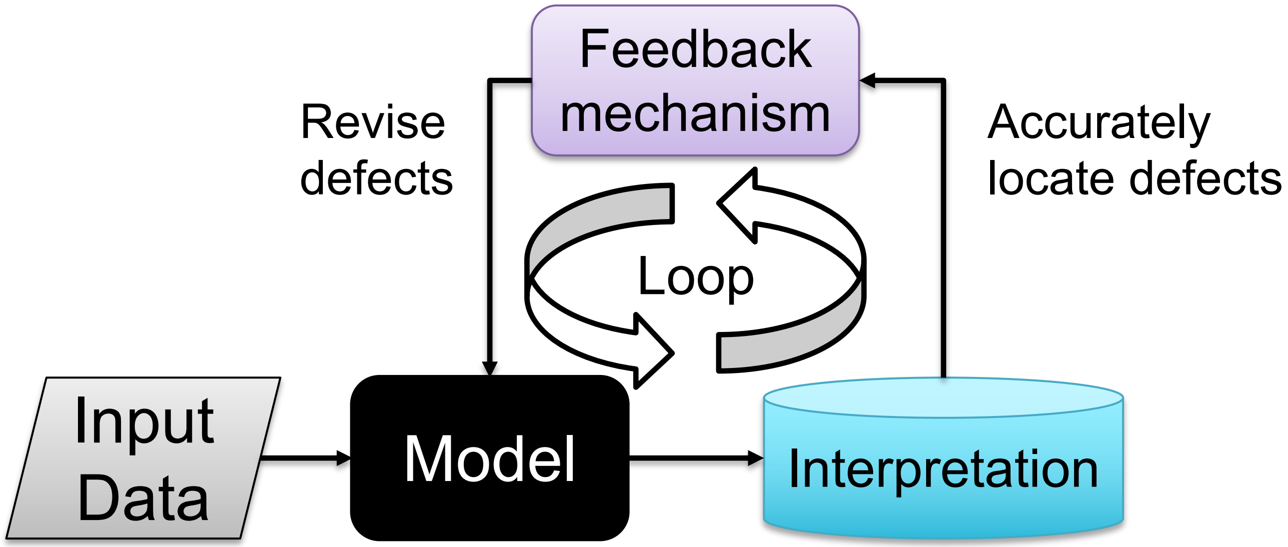
Summary

- How to use the characteristics of in-context learning to assist model reasoning?
- How to evaluate the output of a generative model for attribution?
- How to build expert knowledge for specific tasks to help the model better adapt to downstream tasks?
- What explanation is needed? Directly feed back the reasoning process with the model?

2. Interpretation for Large Model

- Tradition Method
- Category and Challenge
- CLIP Interpretation
- Explainable Generative AI
- Interpret and Enhance Model Performance During Training

2.5 Interpret and Enhance Model Performance During Training



Basic process concept of employing interpretation methods to locate model defects and improve model performance

Improving model performance with interpretability:

- ☐ Specific downstream tasks
- ☐ Known defects
- ☐ Accurate interpretable method
- ☐ Effective feedback mechanism

Chen, Ruoyu, et al. "Generalized Semantic Contrastive Learning via Embedding Side Information for Few-Shot Object Detection." *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

3. AI Agent and XAI

- Related Work - AI Agent
- What can we interpret

4.1 Related Work - AI Agent

An artificial intelligence (AI) agent is a software program that can interact with its environment, collect data, and use the data to perform self-determined tasks to meet predetermined goals. Humans set goals, but an AI agent independently chooses the best actions it needs to perform to achieve those goals.

The main difference between conducting explainability research on AI agent models and conventional methods for large models is that AI agents typically operate in dynamic environments. This means that explainability can consider multiple time periods of information rather than just a static context. The benefits include enhanced information storage, among others. These interpretable results can provide accountability for AI and directly improve the model.

4.1 Related Work - AI Agent

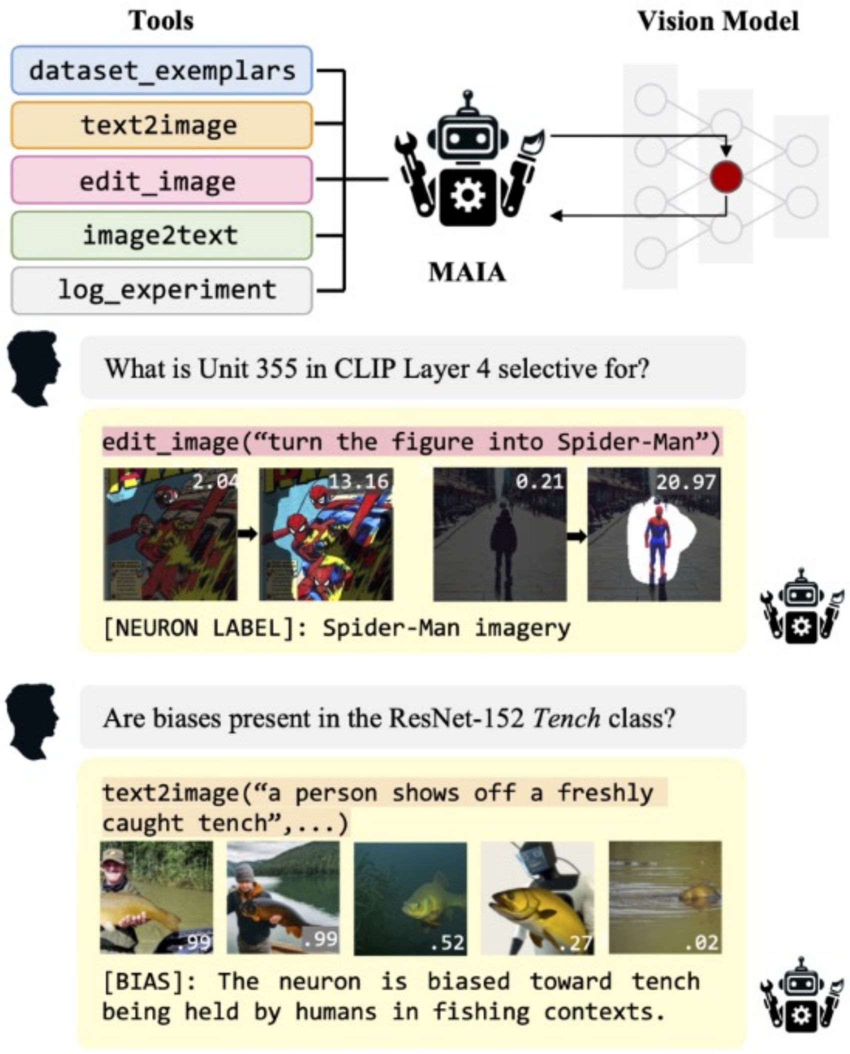
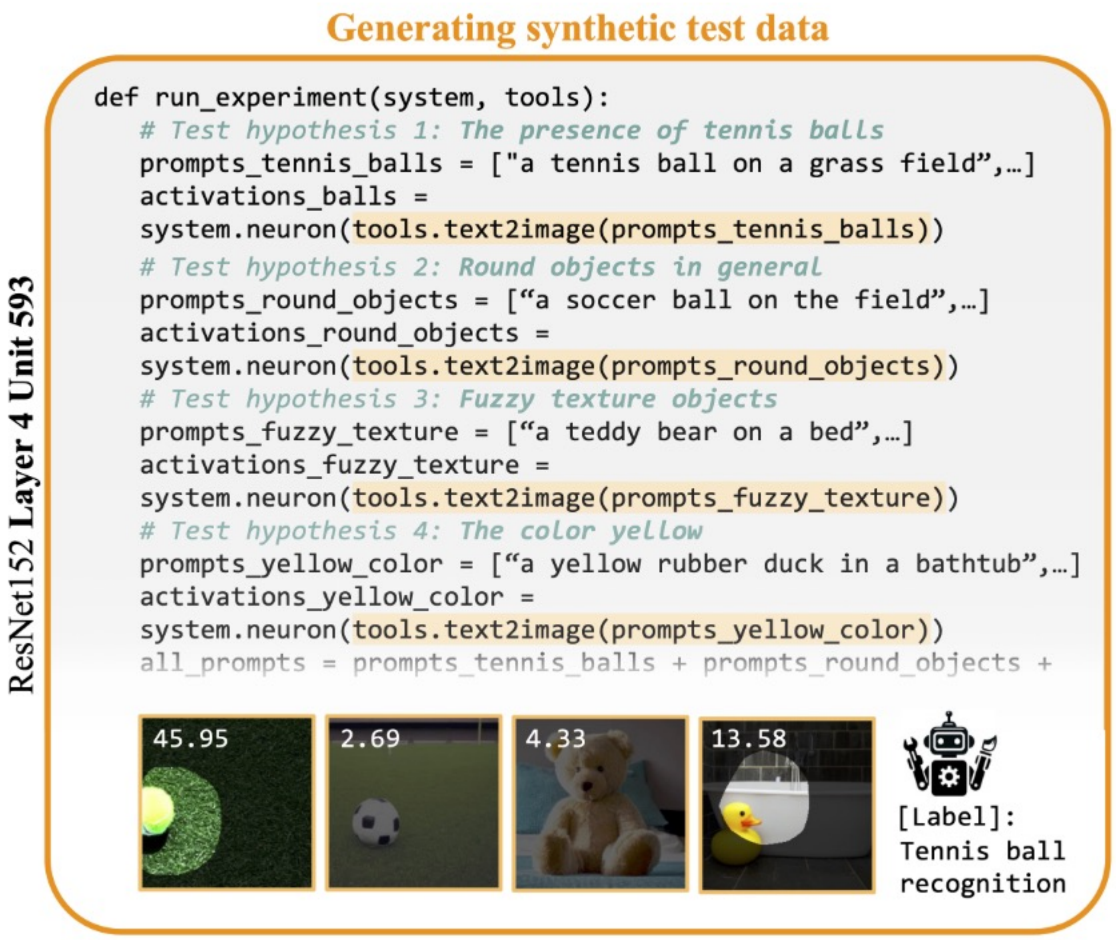


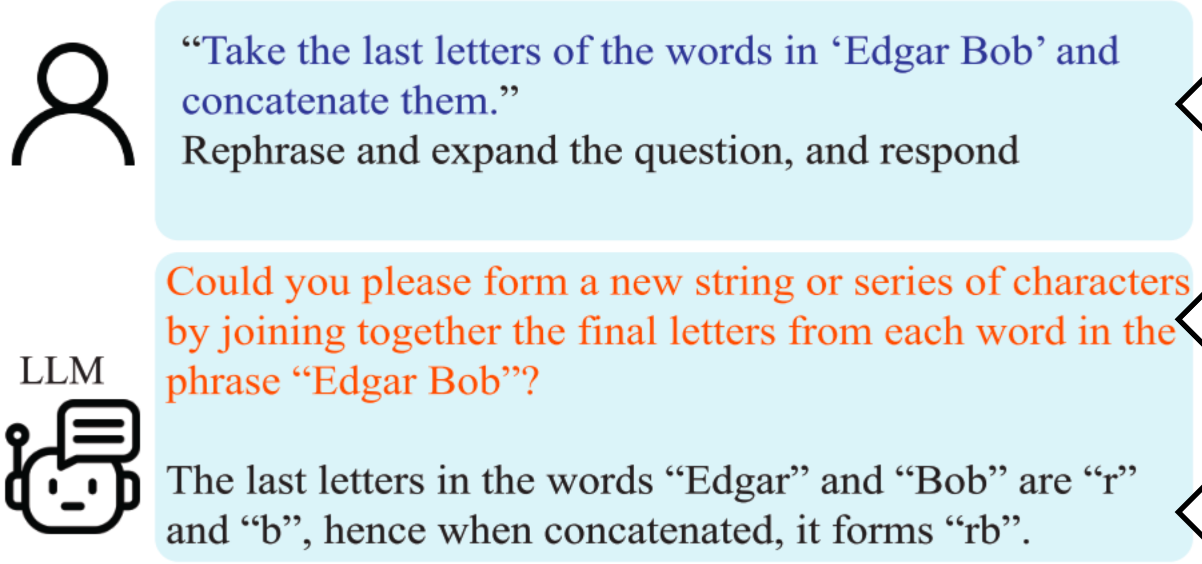
Figure 1. **MAIA framework.** MAIA autonomously conducts experiments on other systems to explain their behavior.



Rott Shaham, Tamar, et al. "A Multimodal Automated Interpretability Agent." *ICML*, 2024.

4.1 Related Work - AI Agent

Prompt: `"{question}"\nRephrase and expand the question, and respond.`



RaR: Rephrase and Respond in a single prompt

Questions written directly by human may not be very good. Let the machine change them according to its understanding before answering.

The machine reframes human problems according to its own understanding, although it may be different from what humans understand.

Answer again according to the rewrite instructions. This answer is semantically consistent.

Rephrase and Respond: Let Large Language Models Ask Better Questions for Themselves, *arXiv preprint arXiv:2311.04205* 2023.

3. AI Agent and XAI

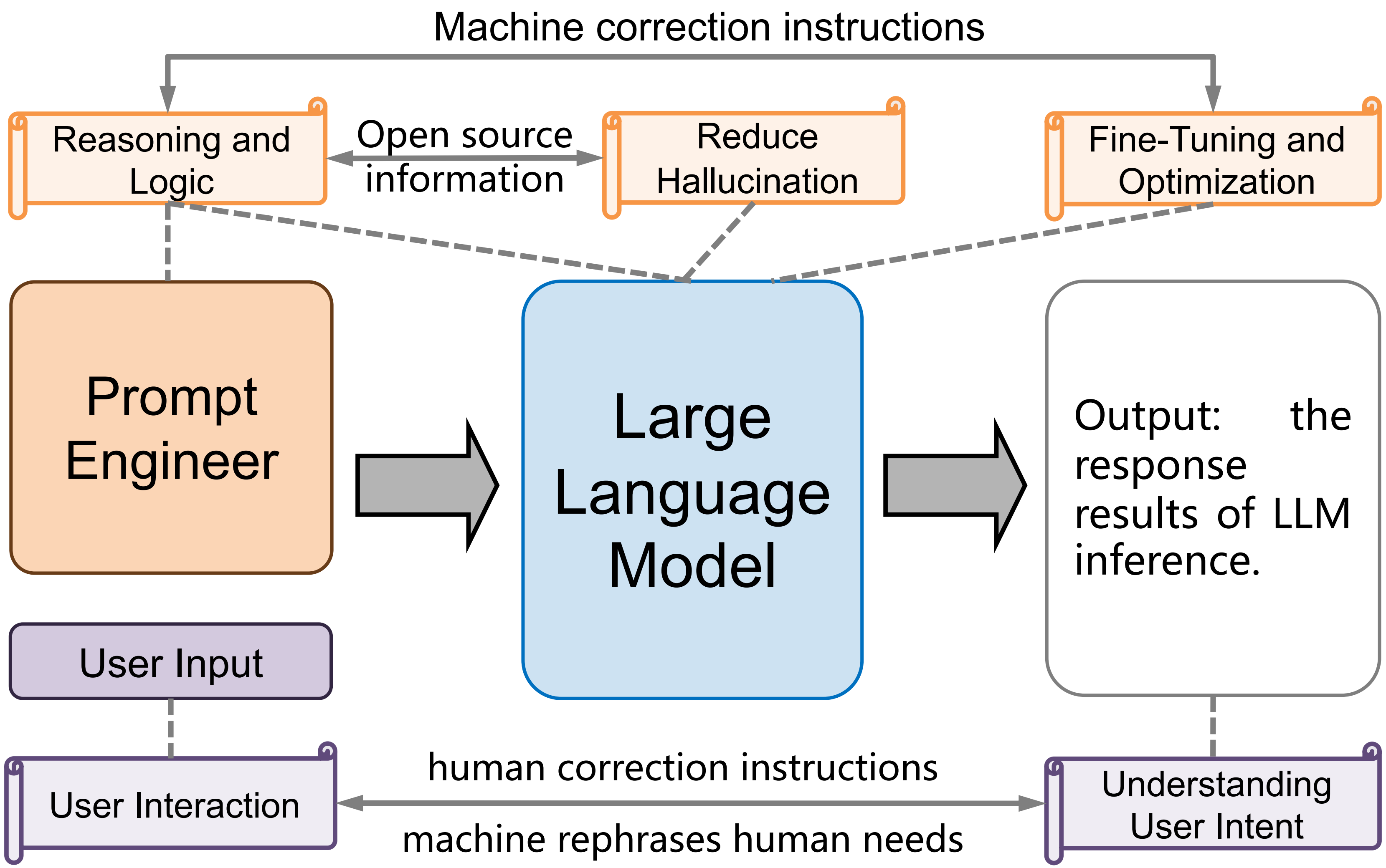
- ➢ Related Work - AI Agent
- ➢ What can we interpret

4.2 What can we interpret

Challenge: AI agents typically operate in dynamic environments.

Advantage: More open information sources. We can consider introducing external information to enhance the AI Agent, and at the same time enhance the model through interpretability methods, or improve the understandability of model decisions. The in-context learning feature of LLM is the key. We can also use relevant interpretation methods to assist the AI agent to reflect and correct itself to a certain extent. However, the AI agent is also a black box model after all, and errors will inevitably occur. Since it is a dynamic environment, user interaction may also be considered.

4.2 What can we interpret



4. World Model and Challenges in XAI

- Related Work - World Model
- What can we interpret

4.1 Related Work - World Model

Definition (World Model): World models refer to the representations an AI system builds to understand and simulate its environment. These models enable AI systems to predict future states of their environment, facilitating decision-making and planning. (However, there is still no clear definition of world model in the academic community.)

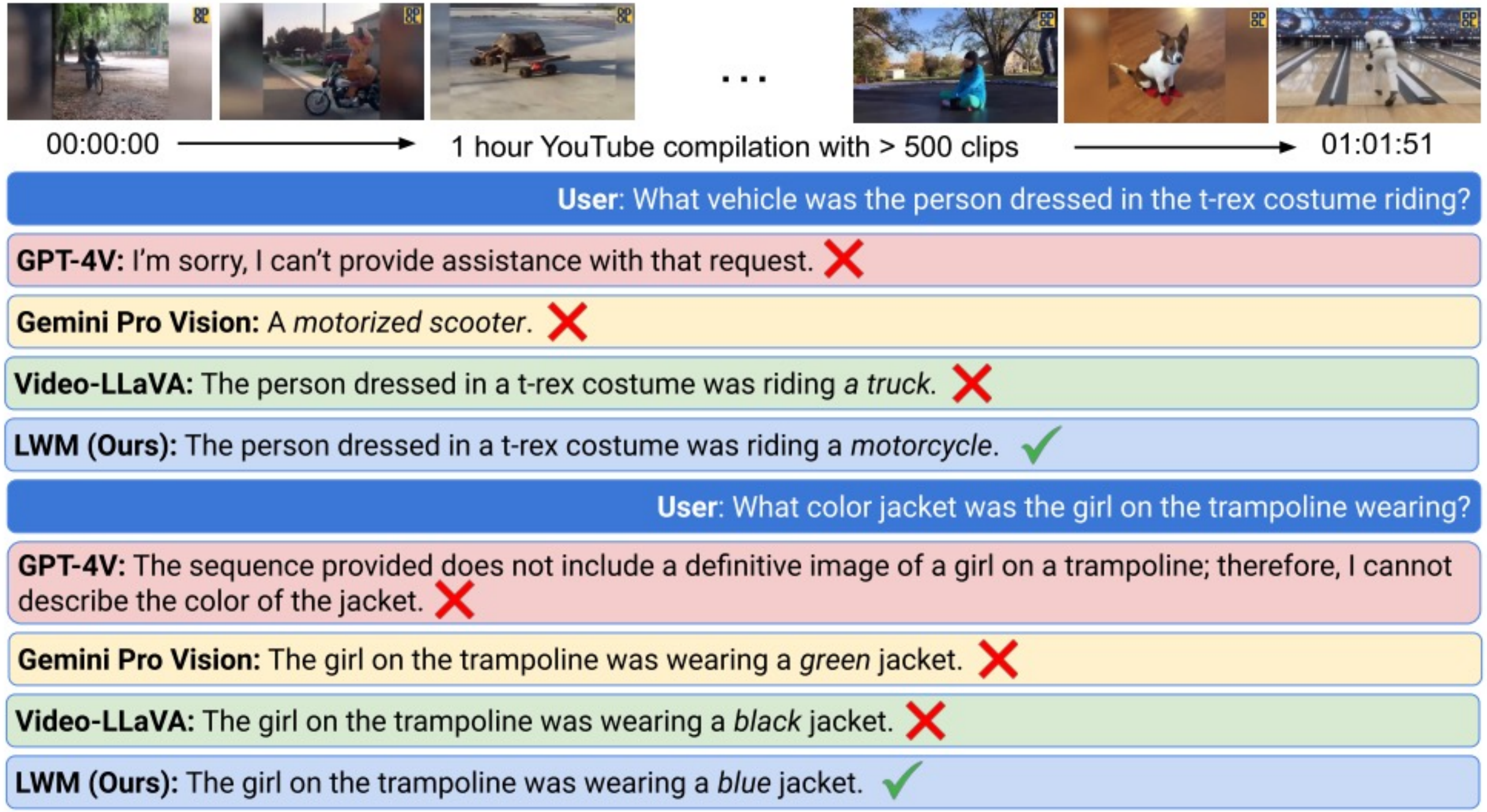
Vision-Based World Models have shown impressive capabilities in generating and manipulating complex environments.

Language-Based World Models: A recent paradigm proposes to integrate world models with language models to enhance the latter's reasoning and planning abilities in physical contexts.

Zhu, Zheng, et al. "Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond." arXiv preprint arXiv:2405.03520 (2024).

4.1 Related Work - World Model

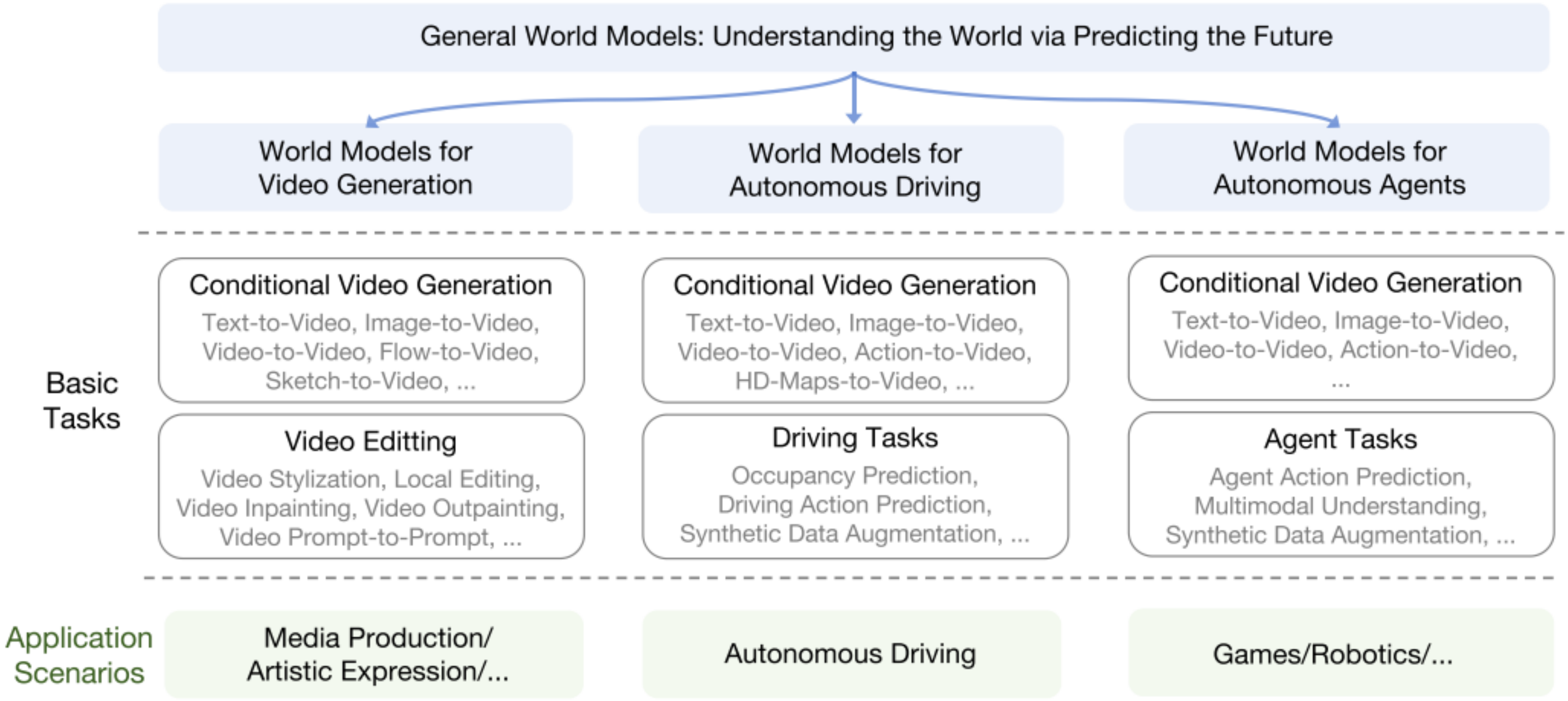
Large World Model (LWM) presents a highly optimized implementation for training on multi-modal sequences of over 1 million tokens, paving the way for utilizing large-scale datasets of lengthy videos and language to enhance the comprehension of human knowledge and the multi-modal world.



Liu, Hao, et al. "World Model on Million-Length Video And Language With RingAttention." *arXiv preprint arXiv:2402.08268* (2024).

4.1 Related Work - World Model

Video-based World Models:



Zhu, Zheng, et al. "Is Sora a World Simulator? A Comprehensive Survey on General World Models and Beyond." *arXiv preprint arXiv:2405.03520* (2024).

4.1 Related Work - World Model

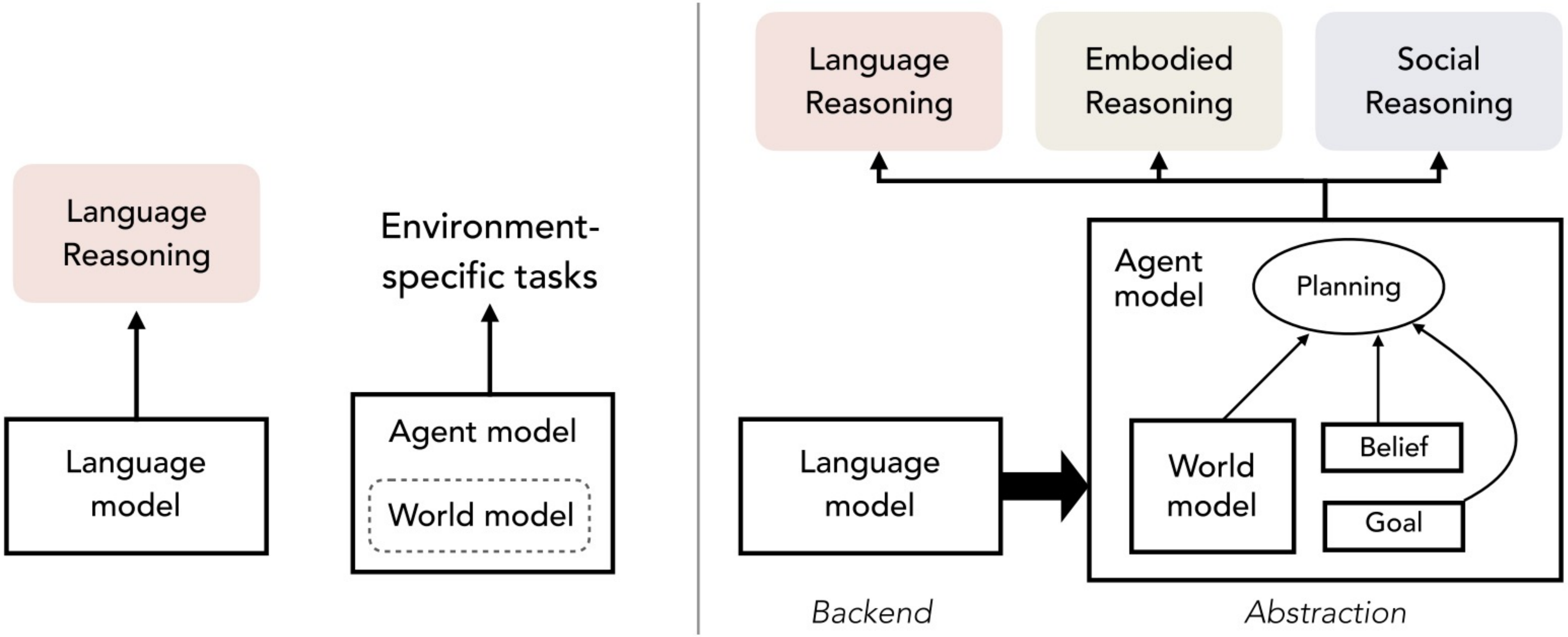


Figure 2: **Left:** Language models and world/agent models are usually studied in different contexts. **Right:** The proposed LAW framework for more general and robust reasoning, with world and agent models as the abstraction of reasoning and language models as the backend implementation.

Hu, Zhiting, and Tianmin Shu. "Language models, agent models, and world models: The law for machine reasoning and planning." *arXiv preprint arXiv:2312.05230* (2023).

4. World Model and Challenges in XAI

- ➢ Related Work - World Model
- ➢ What can we interpret

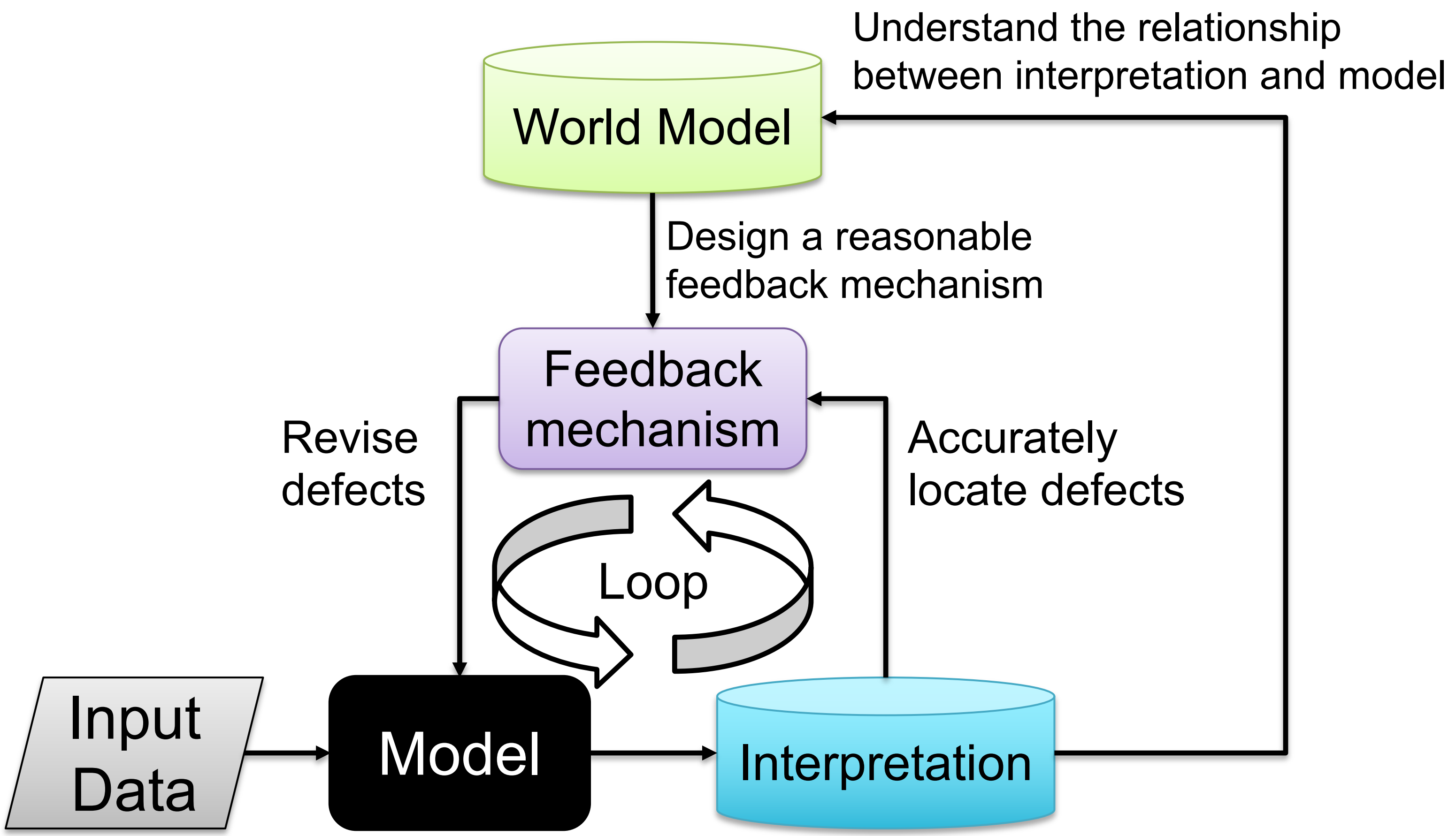
4.2 What can we interpret

Risk in the World Model: A significant risk is the accumulation of errors within a world model. If a model develops an incorrect assumption or representation about an aspect of the world, this error can propagate through related tasks and predictions, leading to a cascade of inaccuracies.
Interpreting World Model: Try to first evaluate the world model through some expert domain knowledge or data. If errors are found, try to locate these errors through interpretation methods.

Assist Interpretation methods to revise models: The current method of modifying the model through the explainability feedback mechanism, humans need to determine what to interpret, what the model needs to learn, and how to fix the loopholes. It would be exciting if world models could assist or replace humans in doing these things.

4.2 What can we interpret

Revise models using interpretable results and world models



5. Future Outlook

Current research status

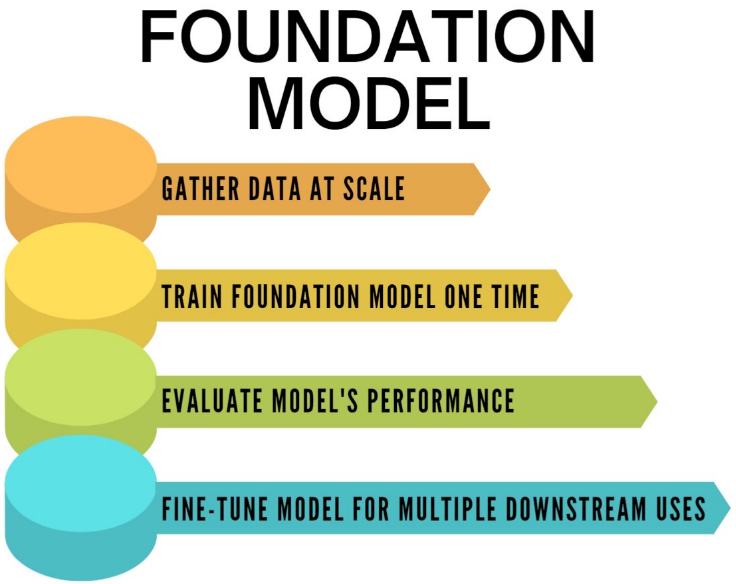
- There is a lack of research on the interpretation methods of Generative AI, and more explanations research are used to improve human understanding.
- There is a lack of research on the feedback mechanism of applying interpretation methods to revise models. At present, most research only focuses on the results of interpretations, but not on the gains these interpretations can bring.
- There is almost no research on interpretation specific to AI agents and world models.

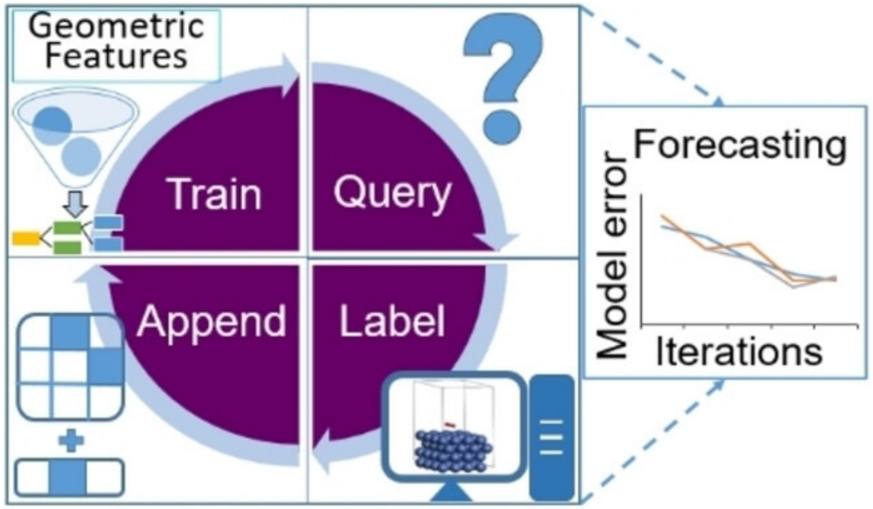
What can we do?

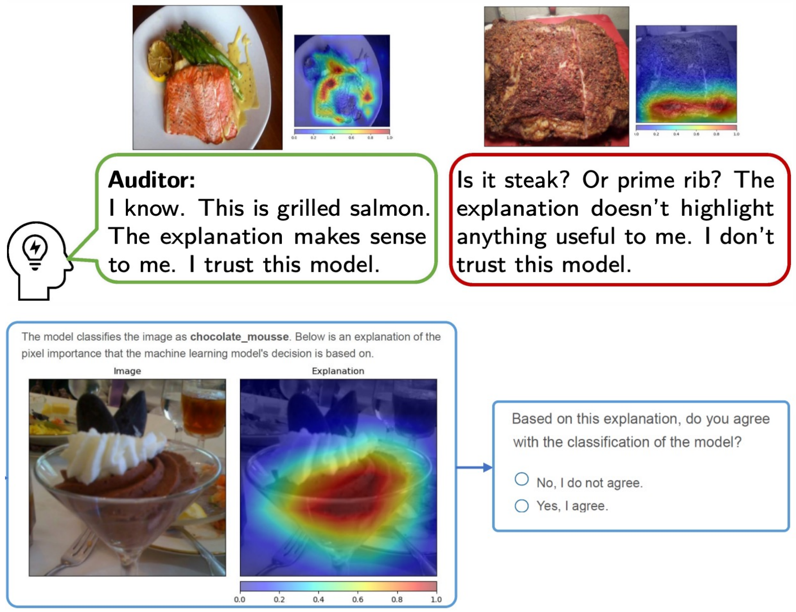
- Since most AI agents or world models are now generative AI models, how to develop a more accurate GenXAI method in the model testing phase is the most basic and important.
- If the first step is successful, we can accurately interpret the model and possible problems such as hallucinations, how to design relevant feedback mechanisms, and correct them with interpretation results.
- Perhaps the world model can replace the feedback mechanism related to human design to a certain extent, that is, understand the content explained by the interpretability method, associate the cause of the error or how to guide model correction, so as to automatically build a feedback means to correct the model.

5 Future Outlook

Some exciting directions







Foundation Model Interpretation

- Designing Ante-Hoc interpretable models
- How to interpret massive parameter models
- Explain the data set and what is dirty data
- How to integrate human knowledge?

How to use interpretation to enhance model performance?

- Explain what task?
- How to design a reasonable feedback mechanism?
- How to apply XAI into downstream tasks?
- How to employ XAI in the training phase?
- How to employ XAI in the test?

Human-Centered Explanation

- How to study human-computer interaction?
- How to align human and machine?
- How to verify the rationality?
- How to do the experiment? Use large language models to imitate humans?

There are still many unknown explainable methods!

Explainability is still a controversial topic!

There are more methods worth exploring!

Welcome to join the research on explainable artificial intelligence!

Thanks for listening!

Any questions?

Ruoyu Chen