Multi-Rationale Explainable Object Recognition via Contrastive Conditional Inference (CCI)

Ali Rasekh[1], Sepehr Kazemi Ranjbar[1], Simon Gottschalk

MOTIVATION & PROBLEM

Goal: Achieve explainable object recognition using multiple rationales (e.g., “A cat because it has large eyes, a furry body, and triangle face”) rather than a single vague explanation.

Limitations of prior work:
- CLIP’s text encoder struggles with prompt-based conditioning (“because”).
- Existing benchmarks (DROR, ECOR) provide only one rationale per image.
- Weak multi-rationale reasoning and poor interpretability.

We propose: A Contrastive Conditional Inference (CCI) framework — a training-free, task-agnostic method that models the probabilistic relationship between image, rationales, and category embeddings.

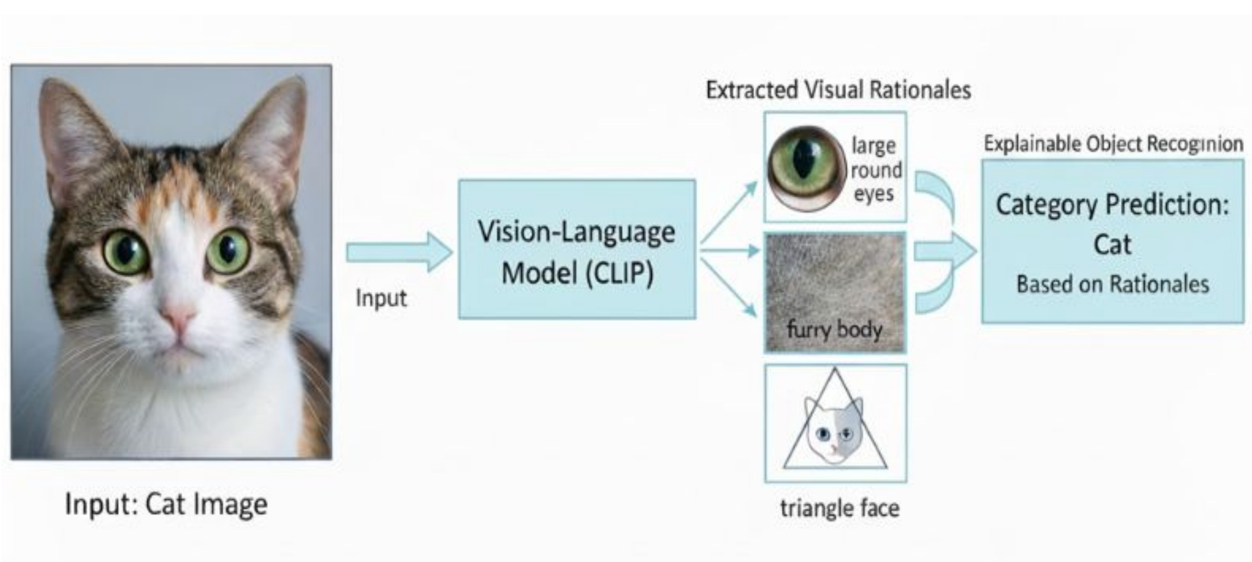

Figure 1: Problem Illustration

We extend the Doubly Right Object Recognition (DROR) dataset into a multi-rationale benchmark with ~3 rationales per image.
- **Source Datasets:** CIFAR-10, CIFAR-100, Food-101, Caltech-101, SUN, ImageNet.
- Metrics per image:
  - **RR**: **R**ight category + **R**ight rationales (maximize)
  - **RW, WR, WW:** various error types (minimize)

KEY IDEA

Contrastive Conditional Inference (CCI): Traditional CLIP models compute *P(c|x)* by similarity between image and text embeddings. CCI generalizes this by theoretical guarantees to $P(c|x, r_1, \ldots, r_M)$ by constructing a hyperplane combining image and rationale embeddings. CCI can generalize any number of conditions with any modality (text or vision).

[1] Equal contribution

METHOD OVERVIEW

Multi-Rationale Object Recognition Task with CCI.
Step 1. Construct the conditioning hyperplane: $S(x, r_1, \ldots, r_M)$
Step 2. Define the desirable direction: $d = \frac{(x + r_1 + \ldots + r_M)}{|x + r_1 + \ldots + r_M|}$
Step 3. Project each category embedding c onto the hyperplane and compute similarity with desirable direction.

Features: Training-free, Generalizes CLIP inference, Improves rationale-based conditioning.

RESULTS

Dataset	Method	RR ↑ ZS	RR ↑ FT	RW ↓ ZS	RW ↓ FT	WR ↓ ZS	WR ↓ FT	WW ↓ ZS	WW ↓ FT
CIFAR-10†	DROR	36.94	20.61	60.62	77.82	0.67	0.09	1.78	1.48
	ECOR	34.02	20.00	64.14	77.29	55.04	0.44	1.28	2.27
	Ours	**41.18**	**58.92**	57.29	39.55	0.00	0.06	1.53	1.47
CIFAR-100†	DROR	17.99	9.45	70.02	78.66	1.50	0.96	10.49	10.93
	ECOR	21.41	11.93	67.56	74.59	1.40	1.53	9.63	11.96
	Ours	**22.56**	**28.48**	46.12	50.54	3.18	2.22	28.14	18.77
Food-101†	DROR	15.21	8.10	70.90	76.91	1.63	0.69	12.26	14.31
	ECOR	21.88	11.72	65.41	73.05	1.80	1.25	10.92	13.80
	Ours	**22.15**	**26.26**	39.91	44.37	3.23	3.86	34.70	25.52
Caltech-101†	DROR	19.81	11.94	68.81	75.56	1.90	1.06	9.48	11.44
	ECOR	**24.85**	14.35	64.38	70.27	1.91	2.19	8.86	13.20
	Ours	23.00	**28.80**	36.40	41.15	4.46	2.85	36.15	27.19
SUN†	DROR	11.96	5.82	46.14	44.39	5.18	2.53	38.73	42.26
	ECOR	12.78	6.59	45.00	45.06	5.42	3.67	36.80	44.68
	Ours	**12.84**	**15.73**	33.00	34.56	4.84	4.35	49.32	45.36
ImageNet†	DROR	2.92	4.75	56.95	55.58	1.88	2.22	38.26	37.45
	ECOR	4.83	6.99	51.39	49.43	3.29	3.59	40.48	39.99
	Ours	**8.35**	**12.61**	20.82	28.86	3.92	3.58	66.91	54.96

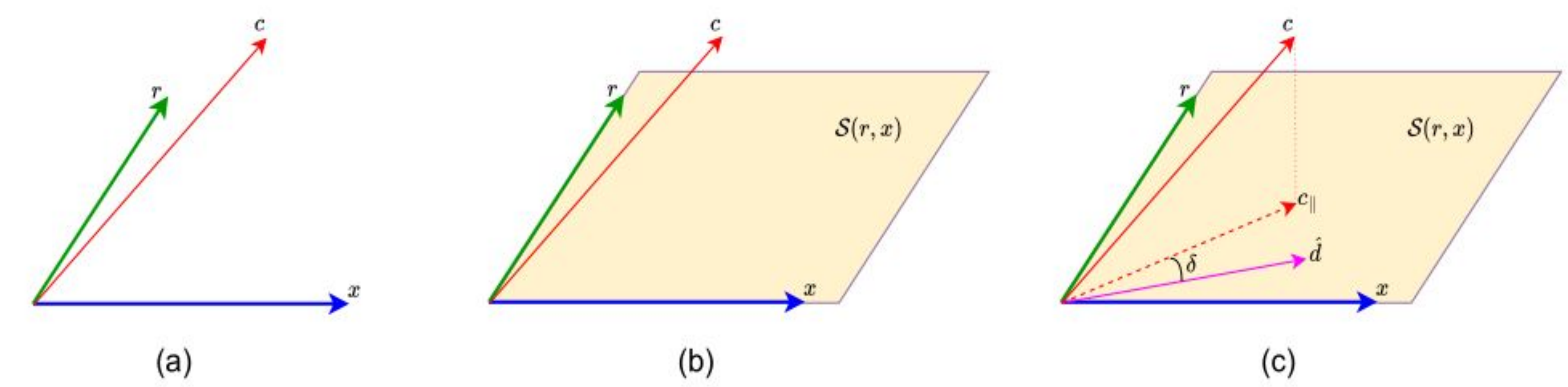

Figure 2: Illustration of Contrastive Conditional Inference (CCI). (a) We consider a triplet *(x, c, r)* and aim to estimate *P(c|x, r)*. (b) Constructing the hyperplane *S(x, r)*. (c) Projecting c onto the hyperplane, denoted c||, and computing its alignment with a desirable direction d with an arrow over it using $|c_{||}| \cos \delta$. This is followed by a softmax step to compute *P(c|x, r)*.

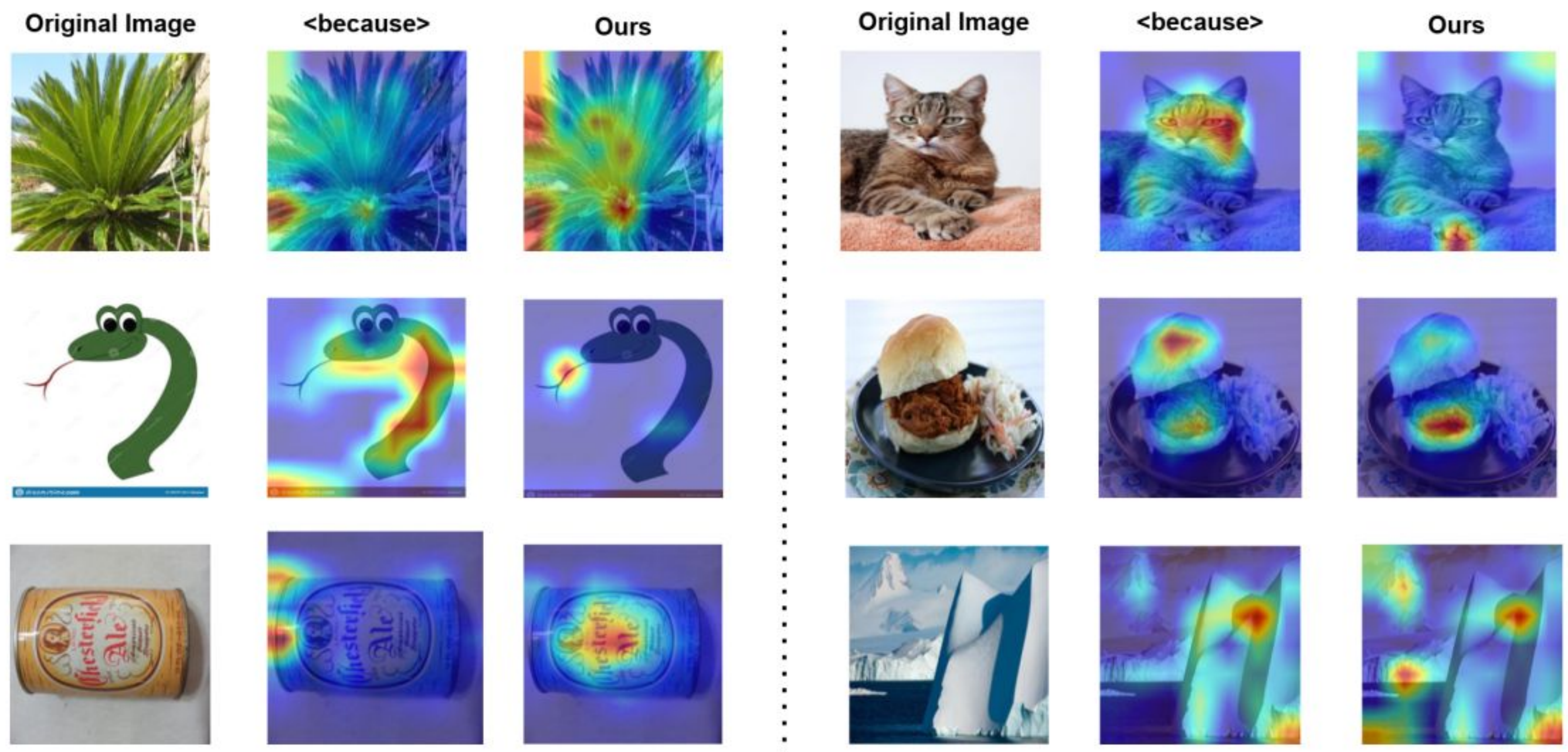

Figure 3: Saliency Maps. The left column of each part shows the original image, the middle columns show the saliency map with the prompt "A photo of a **<category>** because there is **<rationale>**". The right column shows the saliency of the **ground-truth category and rationale** via our method. **Ground-truth categories and rationales** for the left part: **<palm tree, long leaves>** , and **<snake, long tongue>**, **<can, information label>**. Ground-truth categories and rationales for the right part: **<cat, furry body>** , **<pulled pork sandwich, pulled pork>** , and **<iceberg, jagged and irregular shapes of ice>.**

L3S AI Research

Kontakt: Ali Rasekh
Forschungszentrum L3S
Appelstraße 9a
30167 Hannover
Mail: ali.rasekh@l3s.de

Projektpartner: MOTORES AI

www.yourdomain.de