V-CECE: Visual Counterfactual Explanations via Conceptual Edits

Introduction

- Counterfactual image explanations often flip labels with **uninterpretable** or **dispersed** pixel edits, good for metrics, bad for humans.

- Many methods ignore semantics or assume models reason with human concepts, creating misleading explanations.

- White box methodologies also induce unintended bias in the generative process, due to exposing the underlying learned distribution to the dataset

Motivation

- We ask two questions:

1. Do models reason at a **human-semantic** level? For example, do neural networks actively find semantically relevant information during the training

2. What is the **active minimum conceptual edit set** that flips the label?

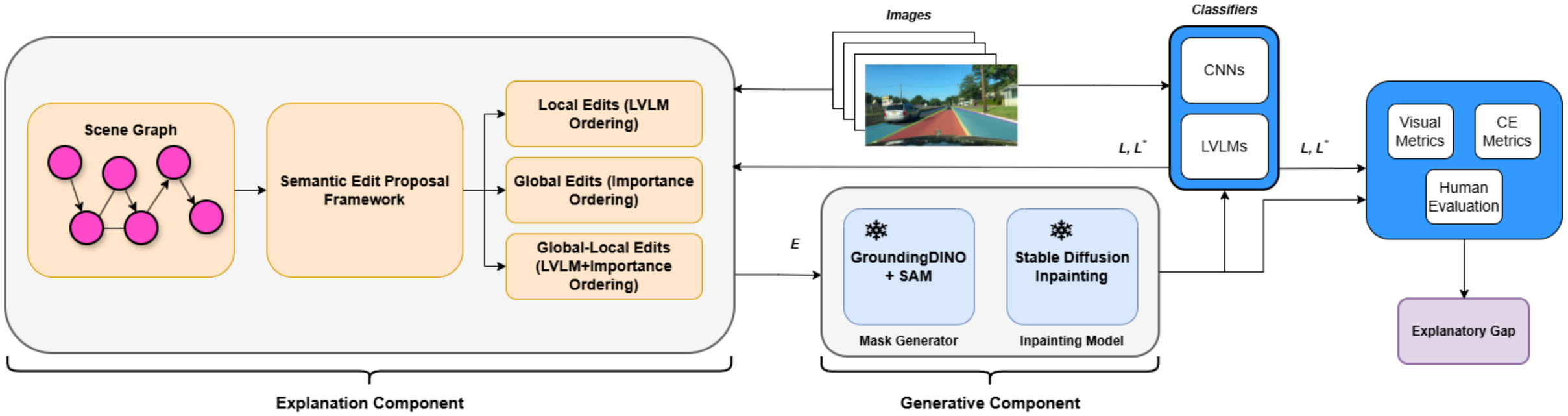
- In order to answer these questions, we propose **V-CECE**, a black-box counterfactual generation tool for evaluating the explainability of artificial intelligence models

Framework

- Our framework consists of two parts:

1. An explanation component, which proposes the edits based on different factors

2. A generative component, which implements the edits in the respective image



Explanation Component

- We compare two scene graphs from the respective classes. Through implementing Hungarian Matching, we find the provably minimum set edits to turn a graph from L to L*

- After acquiring the edits, we propose three different ways of ordering and evaluate each separately

1. **Local Ordering**: Order is proposed through an LVLM

2. **Global Ordering**: Order is proposed through an **importance metric**, which is based on how often an object flips an image to the other class

3. **Local-Global Ordering**: Local edits are applied through global importance ordering

Generative Component

- We utilize Stable Diffusion v1.5 Inpainting to exact the edits in each image. We use 40 steps for generation and employ mask expansion to avoid mis-generation of objects, by incorporating background information

- The mask of the objects are prompted through a pipeline of GroundingDINO for localization and Segment Anything Model (SAM) for mask extraction

- After the edits, we evaluate each image on whether or not it induced a label flip. If it has not, we repeat the process for the next edit in the proposed order

Results

- We evaluate using common image fidelity metrics, such as FID and CMMD, as well counterfactual distribution metrics, such as S3 and flip rate. We compare with prior works on the number of edits required for flipping, and evaluate using a human survey to recognize the alignment between models and humans

Method	FID (↓)	CMMD (↓)	S3 (↑)	SR (↑)	Avg. $\lvert E\rvert$ (↓)	Access	Training	Optimiz.
			Prior Work					
STEEX	58.8	—	—	99.5	—	white-box	days	✓
DiME	7.94	—	0.9463	90.5	—	white-box	days	✓
ACE ℓ_1	1.02	—	0.9970	99.9	—	white-box	days	✓
ACE ℓ_2	1.56	—	0.9946	99.9	—	white-box	days	✓
TIME	51.5	—	0.7651	81.8	—	**black-box**	hours	✗
			V-CECE					
V-CECE — Best CNN	67.27	0.767	0.6950	98.07	3.76	**black-box**	**N/A**	✗
V-CECE — Best ViT	82.93	1.027	0.6160	94.06	4.71	**black-box**	**N/A**	✗
V-CECE — Best LVLM	42.76	0.364	0.7970	98.10	2.44	**black-box**	**N/A**	✗

Table 1: BDD100K vision metrics. Previous methods and the best CNN/ViT/LVLM (by lowest FID).

Results

	Avg. $\lvert E\rvert$ Model (↓)	Avg. $\lvert E\rvert$ Human (↓)	Visually correct images (%)
DenseNet	5.22	2.21	59.71
ConvNext	7.35	2.27	34.24
EfficientNet	5.96	2.66	30.17
Swin	6.31	2.25	56.66
Claude-3-Haiku	2.91	1.88	69.58
Claude-3.5-Sonnet	**2.19**	**1.33**	**81.20**
Claude-3.7-Sonnet	2.50	1.37	79.98
Claude-3.7-Sonnet Thinking	4.33	2.69	70.01

Framework Example

Conclusions

- V-CECE closes the human–model gap by proposing minimal, semantics-aligned edits in a black-box setup and verifying them end-to-end.
- Neural networks (CNNs and ViTs) appear to make decisions based on statistical dependencies, not accounting form semantic information
- LVLMs appear more aligned to humans, recognizing semantic edits much sooner than neural networks.
- Compared with prior counterfactual methods, we flip labels with far fewer edits and no training or white-box access, producing cleaner, more interpretable explanations on shared benchmarks.

Thank you for your attention!

Poster session: Wed 3 Dec 7:30 p.m. EST — 10:30 p.m. EST