A FRAMEWORK FOR CONCEPT-BASED LOCAL UNIFIED EXPLANATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

There is an increasing need to integrate model-agnostic explanation techniques with concept-based approaches, as the former can explain models across different architectures while the latter makes explanations more faithful and understandable to end-users. However, existing concept-based model-agnostic explanation methods are limited in scope, mainly focusing on attribution-based explanations while neglecting diverse forms like sufficient conditions and counterfactuals, thus narrowing their utility. To bridge this gap, we propose a general framework *ConLUX* to elevate existing local model-agnostic techniques to provide concept-based explanations. Our key insight is that we can uniformly extend existing local model-agnostic methods to provide unified concept-based explanations with large pre-trained model perturbation. We have instantiated *ConLUX* to provide concept-based explanations in three forms: attributions, sufficient conditions, and counterfactuals, and applied it to popular text, image, and multimodal models. Our evaluation results demonstrate that *ConLUX* provides explanations more faithful than state-of-the-art concept-based explanation methods, and provides richer explanation forms that satisfy various user needs.

1 INTRODUCTION

The widespread application of machine learning models has created a strong demand for explanation methods. In particular, as the structure of machine learning models varies and evolves rapidly, and effective closed-source models (e.g., GPT-4 (Achiam et al., 2023) and Gemini (et al., 2024)) become more prevalent, model-agnostic explanation methods show their appeal (Wang, 2024). These methods treat target models as black boxes and thus do not require access to internal model information, and provide explanations in forms including attributions, sufficient conditions, and counterfactuals, satisfying various user needs (Wang, 2024). On the other hand, providing concept-based explanations has emerged as a promising direction (Poeta et al., 2023; Sun et al., 2023), which are well-known to be more faithful and understandable to end-users (Zhang et al., 2021a; Ghorbani et al., 2019b; Kim et al., 2018; Sun et al., 2023). As a result, there is a growing demand for providing concept-based model-agnostic explanations for machine learning models.

While providing concept-based model-agnostic explanations is promising, existing methods remain limited in scope. They are largely restricted to attribution forms (Poeta et al., 2023) and lack support for richer explanation types such as sufficient conditions and counterfactuals. On the other hand, feature-level model-agnostic methods support diverse forms other than attributions, but they lack the interpretability and fidelity brought by using high-level concepts (Zhang et al., 2021a; Ghorbani et al., 2019b; Kim et al., 2018; Sun et al., 2023).

To bridge this gap, we propose *ConLUX*, a general and lightweight framework to elevate existing feature-level model-agnostic methods from feature level to concept level, providing concept-based explanations in various forms beyond attributions, as well as concept-based attribution explanations of higher fidelity. Additionally, we focus on local explanations, as they are tractable for end-users when explaining complex models used in real-world applications.

We achieve this goal by making two key observations. First, elevating existing local model-agnostic methods to concept level does not require changing their core algorithms. Since these methods follow a common workflow (Liu & Zhang, 2025; Ribeiro et al., 2016; 2018; Guidotti et al., 2018;

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

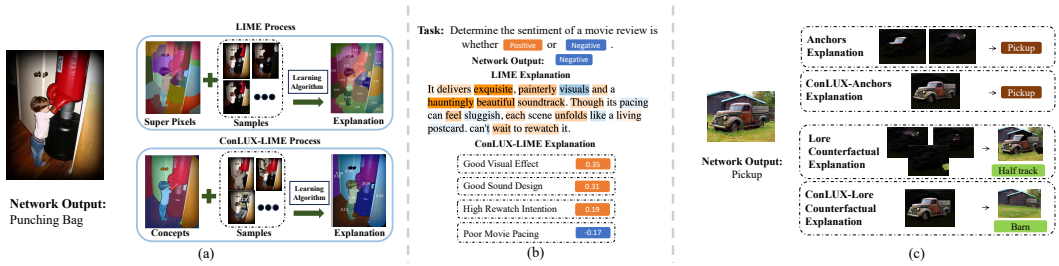

Figure 1: Examples of using *ConLUX* to lift feature-level local model-agnostic explanations to concept-level explanations. (a) For image classification, *ConLUX*-LIME replaces superpixel masking with concept-level perturbations and produces attributions over detected objects. (b) For text classification, *ConLUX*-LIME explains the prediction using high-level topics instead of individual words. (c) *ConLUX* also augments rule-based methods: *ConLUX*-Anchors identifies concept-level sufficient conditions, while *ConLUX*-LORE produces concept-level counterfactual explanations.

Lundberg & Lee, 2017), we only need to 1) extract high-level concepts from input data and 2) perform perturbations at concept level.

Second, given that several effective concept extraction methods already exist (Ghorbani et al., 2019a; El Shawi, 2024; Ludan et al., 2023), the only challenge is to perform perturbations at the concept level. To this end, we propose using large pre-trained models to handle concept-level perturbations.

Figure 1 shows examples of using *ConLUX* to provide concept-based explanations for image and text classification models. Let us delve into the details of Figure 1(a), where *ConLUX* elevates LIME (Ribeiro et al., 2016) to explain an image classification model. Unlike vanilla LIME using fragmented superpixels, *ConLUX*-augmented LIME generates explanations based on high-level concepts following Sun et al. (2023). In the perturbation phase, instead of masking fragmented superpixels, *ConLUX*-augmented LIME directly modifies the detected objects in the image. This augmentation make the resulting attributions grounded in semantically meaningful objects rather than low-level superpixels, making the explanations more interpretable for end-users.

Besides attribution-based methods, *ConLUX* can also extend local model-agnostic methods of various forms to provide concept-based explanations, satisfying diverse user needs and offering a more comprehensive understanding of target models. Figure 1(c) shows that *ConLUX* can extend Anchors (Ribeiro et al., 2018) and LORE (Guidotti et al., 2018) to provide concept-level rule-based sufficient conditions and counterfactual explanations, respectively. As *ConLUX* provides all these explanations in a unified manner, users can obtain their desired explanations by simply selecting the explanation type they need.

Moreover, although *ConLUX* is simple in design, it outperforms more elaborate, specially designed concept-based approaches, as our empirical evaluation results will later show.

In our evaluation, we first verified that large pre-trained models can faithfully perform concept-level perturbations. Then we evaluated *ConLUX* on explaining two text classification models, three image classification models, and one multimodal model. Our evaluation results show that *ConLUX* improves the fidelity of Anchors, LIME, LORE, and Kernel SHAP (Lundberg & Lee, 2017) explanations by 56.8%, on average, and *ConLUX*-augmented methods outperform state-of-the-art concept-based explanation methods specifically designed for text models (Ludan et al., 2023; Yu et al., 2024) and image models (Sun et al., 2023), respectively. We also ran a human evaluation, showing that *ConLUX* helps users better use the explanations in downstream tasks.

Contributions Our contributions are as follows:

- We introduce *ConLUX*, a general and lightweight framework that elevates existing local model-agnostic explanation methods to concept level with minimal user effort, providing concept-based explanations with state-of-the-art performance, and in various forms beyond attributions.
- We propose using large pre-trained models to perform concept-level perturbations, enabling model-agnostic methods to generate concept-based explanations, and empirically verify the fidelity of these perturbations.

- We instantiate *ConLUX* on four popular explanation methods: LIME, Kernel SHAP, Anchors, and LORE, and demonstrate its effectiveness through extensive experiments and human evaluations. *ConLUX* achieves state-of-the-art performance in generating concept-based explanations. Additionally, except the concepts we used in Section 1 and 4, *ConLUX* is flexible to work with various concepts, and we have shown examples in Appendix B.4.

2 Background and Related Work

Our work is related to model-agnostic explanation methods and concept-based explanation methods. We next introduce the background knowledge and related work.

2.1 Machine Learning Models

We consider a machine learning model as a function f that maps an input vector $\boldsymbol{x}$ to an output scalar $f(\boldsymbol{x})$. Formally, let $f : \mathbb{X} \to \mathbb{R}$, where $\mathbb{X} = \mathbb{R}^n$. Let $\boldsymbol{x}_i$ denote the i-th feature value of $\boldsymbol{x}$.

2.2 Local Model-Agnostic Explanation Methods

A local model-agnostic explanation method t takes a model f and an input $\boldsymbol{x}$, and generates a local explanation $g_{f,\boldsymbol{x}}$ to describe the behavior of f around $\boldsymbol{x}$, i.e., $g_{f,\boldsymbol{x}} := t(f, \boldsymbol{x})$. An explanation $g_{f,\boldsymbol{x}}$ (g for short) is an expression formed with predicates. Each predicate p maps an input $\boldsymbol{x}$ to a binary value, i.e., $p : \mathbb{X} \to \{0, 1\}$, indicating whether $\boldsymbol{x}$ satisfies a specific condition.

Existing local model-agnostic explanation methods follow a similar workflow:

1. **Producing Predicates**: These methods first generate a set of predicates $\mathbb{P}$ based on the input $\boldsymbol{x}$.
2. **Generating Samples**: The underlying perturbation model t_{per} generates a set of samples $\mathbb{X}_s$ and its corresponding predicate representation set $\mathbb{B}_s$.
3. **Learning Explanation**: The underlying learning algorithm generates the local explanation $g_{f,\boldsymbol{x}}$ consisting of predicates in $\mathbb{P}$ using $\mathbb{B}_s$ and $f(\mathbb{X}_s)$.

Mainstream local model-agnostic explanation methods like Anchors (Ribeiro et al., 2018), LIME (Ribeiro et al., 2016), LORE (Guidotti et al., 2018), and Kernel SHAP (Lundberg & Lee, 2017), all follow this workflow. In the following, we introduce the main components of these explanation methods.

Predicate Sets Given an input $\boldsymbol{x}$, the corresponding predicate set $\mathbb{P}$ is defined as $\mathbb{P} = \{p_i | i \in [1, d]\}$, where d is the number of predicates in $\mathbb{P}$, a hyperparameter set by users or according to the input $\boldsymbol{x}$. Each p_i is a feature predicate that constrains the value of a set of feature values in $\boldsymbol{x}$, i.e. $p_i(\boldsymbol{z}) : \bigwedge_{j \in \mathbb{A}_i} 1_{\boldsymbol{x}_j = \boldsymbol{z}_j}$, where $\mathbb{A}_i$ is the set of indices of features that p_i constrains. Then, we define the predicate representation $\boldsymbol{b} \in \{0, 1\}^d$ of a sample input $\boldsymbol{z}$ as a binary vector where $\boldsymbol{b}_i = p_i(\boldsymbol{z})$.

Perturbation Models The perturbation model t_{per} first randomly selects $\mathbb{B}_s \subseteq \{0, 1\}^d$ as the predicate representations of the samples. Then, it transforms $\mathbb{B}_s$ back to the original input space to get $\mathbb{X}_s$. For each $\boldsymbol{b} \in \mathbb{B}_s$, if $\boldsymbol{b}_i = 1$, then for each $j \in \mathbb{A}_i$, it sets $\boldsymbol{z}_j = \boldsymbol{x}_j$; otherwise, it sets each $\boldsymbol{z}_j$ to a masked value, or a random value sampled from a user-defined distribution.

Learning Algorithms Existing local model-agnostic explanation methods use different learning algorithms. Anchors uses KL-LUCB (Kaufmann & Kalyanakrishnan, 2013) to learn sufficient conditions, LIME and Kernel SHAP use linear regression (McDonald, 2009) to learn attributions, and LORE uses decision trees (Ruggieri, 2004) to learn sufficient conditions and counterfactuals.

Explanation Forms Existing local model-agnostic explanation methods can provide various forms of explanations for different user needs:

- *Attributions*: An attribution explanation (Lundberg & Lee, 2017; Ribeiro et al., 2016; Tan et al., 2023; Shankaranarayana & Runje, 2019; Goldstein et al., 2015; Apley & Zhu, 2020) consists of a set of feature predicates p_i with importance weights w_i, i.e., $g = \{(p_i, w_i) \mid i \in [1, d]\}$. Attributions assist users in evaluating the trustworthiness of the model (Ribeiro et al., 2016).
- *Sufficient Conditions*: This form of explanation (Ribeiro et al., 2018) identifies a minimal set of conditions that are sufficient to produce the same output as the original input. Formally, $f(\boldsymbol{z}) = f(\boldsymbol{x})$

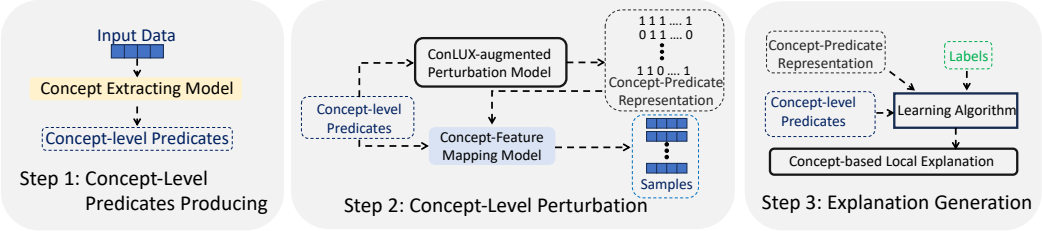

Figure 2: The workflow of *ConLUX*-augmented local model-agnostic explanation techniques.

if $g(\boldsymbol{z}) = 1$, where $g(\boldsymbol{z}) = \bigwedge_{p \in \mathbb{Q}} p(\boldsymbol{z})$ and $\mathbb{Q} \subseteq \mathbb{P}$. Sufficient conditions help users anticipate the model's behavior on unseen inputs (Ribeiro et al., 2018).

- *Counterfactuals*: A counterfactual explanation (Wachter et al., 2018; Guidotti et al., 2018) shows how a model's prediction changes under specific input modifications. It is defined as $f(\boldsymbol{z}) = y$ if $g(\boldsymbol{z}) = 1$, where $y \neq f(\boldsymbol{x})$ and $g(\boldsymbol{z}) = \bigwedge_{p \in \mathbb{Q}} p(\boldsymbol{z}) \wedge \bigwedge_{p \in \mathbb{C}} \neg p(\boldsymbol{z})$, with $\mathbb{Q}, \mathbb{C} \subseteq \mathbb{P}$. Counterfactuals provides actionable insights for decision-making and outcome reversal (Guidotti et al., 2018).

To our knowledge, existing model-agnostic explanation methods that provide explanations other than attributions mainly provide explanations at feature levels (Zhang et al., 2021c; Wang, 2024).

2.3 Concept-Based Explanations

Concept-based explanations primarily focus on attributing high-level concepts to model predictions (Poeta et al., 2023). While the definition of a concept may vary, Molnar (2020) broadly characterizes it as "an abstract idea, such as a color, an object, or even an idea," with the key requirement that concepts be meaningful to end-users (Ghorbani et al., 2019a), such as objects in images or topics in text.

Existing concept-based explanation methods have extensively explored the extraction of high-level concepts from input data, but the perturbation of high-level concepts remains largely unexplored, which is crucial for generating model-agnostic explanations. Specifically, these approaches, including various concept bottleneck models (Koh et al., 2020; Ludan et al., 2023; Oikarinen et al., 2023; Kim et al., 2023), leverage internal model representations (Zhang et al., 2021b; Yeh et al., 2019; Cunningham et al., 2023; Ghorbani et al., 2019c; Crabbé & van der Schaar, 2022; Fel et al., 2023a;b; Ghorbani et al., 2019a; Yeh et al., 2020; Varshney et al., 2025; Taparia et al.), external knowledge (El Shawi, 2024; Widmer et al., 2022; Ciravegna et al., 2023), or pre-trained models (Ludan et al., 2023; Sun et al., 2023) to extract concepts from input data. However, although Goyal et al. (2020) and Wu et al. (2022) try to map concept level changes back to feature level, they are designed for specific tasks, and need much effort to adapt to new datasets or models. Moreover, most concept-based model-agnostic methods focus primarily on attribution-based explanations, which limits their expressiveness and applicability. Although Ciravegna et al. (2023) provide rule-based sufficient conditions, they are limited to only provide global explanations, which is not tractable for complex models in practice.

3 The *ConLUX* Framework

In this section, we propose *ConLUX*, a general and lightweight framework to elevate existing local model-agnostic explanation methods to concept level without significantly changing their core components. As shown in Figure 2, *ConLUX*-augmented methods generate explanations in three steps: 1) concept-level predicate producing, 2) concept-level perturbation, and 3) explanation generation.

3.1 Concept-Level Predicate Producing

We use a **concept-extracting model** to extract high-level concepts from a given input $\boldsymbol{x}$ based on the target task and user needs. As existing work has proposed several effective methods for concept extraction (Ghorbani et al., 2019a; El Shawi, 2024; Ludan et al., 2023; Sun et al., 2023), users can select an appropriate method based on their specific needs.

ConLUX defines **concept predicates** (denoted as p^c) based on the extracted concepts. Each concept predicate p^c is a binary function that indicates whether the input $\boldsymbol{x}$ satisfies a specific concept. Subsequently, *ConLUX* replaces the vanilla predicates set $\mathbb{P}$ with the concept predicates set $\mathbb{P}^c$.

An Example For the text example in Figure 1(b), our approach follows Ludan et al. (2023) to extract concepts that correlate with sentiment from a professional movie reviewer's perspective, and the model generates the four concepts. For the i-th concept, *ConLUX* defines its corresponding predicate p_i^c as *"if the sentence has the i-th concept"*, and set the predicates set $\mathbb{P}^c = \{p_1^c, p_2^c, p_3^c, p_4^c\}$.

3.2 Concept-Level Perturbation

The ***ConLUX* augmented perturbation model** t_{per}^c changes high-level concepts directly to obtain samples $\mathbb{X}_s^c$ and their concept-predicate representations $\mathbb{B}_s^c$.

Specifically, t_{per}^c first generates samples in concept-predicate representations. Each **concept-predicate representation** $\boldsymbol{b}^c$ is a $|\mathbb{P}^c|$-dimension binary vector corresponding to a sample $\boldsymbol{z}$, indicating whether the sample satisfies the concept predicates in $\mathbb{P}^c$. Since concept predicates do not directly constrain the feature values, t_{per}^c requires a **concept-feature mapping model** $c: \{0,1\}^{|\mathbb{P}^c|} \to \mathbb{X}$ to transform the samples back to the feature level. We find that large pre-trained models are ideal to serve as the concept-feature mapping model in *ConLUX*, as they excel at generating content based on structured prompts. Besides, their generated samples are more meaningful compared to simply masking to changing feature values (Tan et al., 2023; Ribeiro et al., 2016; 2018; Liu & Zhang, 2025). Specifically, when mapping a sample's concept-predicate representation $\boldsymbol{b}^c$ back to the feature level, if $\boldsymbol{b}_i^c = 1$, we prompt the model to ensure that the generated sample satisfies the i-th concept; if $\boldsymbol{b}_i^c = 0$, the generated sample must not satisfy the i-th concept.

Examples Figure 1(a) shows some samples generated by the *ConLUX*-augmented perturbation model. For example, t_{per}^c generated the left-upper sample without a child in the image. Appendix A shows the prompts we used to generate the samples, which does not require much effort to design.

3.3 Explanation Generation

In this stage, *ConLUX* leverages the vanilla learning algorithms underlying existing methods to generate explanations. Using the concept-predicate representations $\mathbb{B}_s^c$ and their corresponding outputs $f(\mathbb{X}_s^c)$, the learning algorithm learns a concept-based explanation $g_{f,\boldsymbol{x}}$ composed of concept predicates in $\mathbb{P}^c$. Benefiting from *ConLUX*'s generality in augmenting various existing explanation methods, *ConLUX* inherits their ability to provide various forms of explanations, including attributions, sufficient conditions, and counterfactuals.

ConLUX enables the generation of multiple explanation forms through a unified framework with a single click. As a result, users can flexibly choose the explanation type that best fits their needs. We refer to the resulting collection of explanations as a ***ConLUX* local unified explanation**.

4 Empirical Evaluation

We evaluate *ConLUX* from three aspects: (1) the fidelity of concept-level perturbation, (2) the fidelity of *ConLUX*-augmented explanations, and (3) how much *ConLUX* helps users use explanations in downstream tasks. We also discuss the time efficiency of *ConLUX* and the robustness of *ConLUX* to different generative model choices.

4.1 Perturbation Fidelity Evaluation

To ensure *ConLUX* can faithfully generate explanations, it is essential to ensure the underlying large models perform accurate concept-feature mapping.

Experimental Setup We conducted experiments on two types of input data: text and image.

For text data, we use DeepSeek-V3 (Liu et al., 2024) as the concept–feature mapping model. We verify its fidelity on perturbing on both human-annotated and model-extracted concepts. We use four datasets and randomly select 250 sentences from each of the datasets. For each sentence, we

Table 1: Accuracy of using different concept-level perturbation models to generate samples that satisfy certain concept predicates. We used DeepSeekV3 and Blended Latent Diffusion in our perturbation and explanation fidelity evaluation in Section 4.1 and 4.2.

Model	CoNLL	Onto.	Large.	Fake.	COCO
DeepSeekV3	97.5	98.1	97.7	94.7	–
Qwen2.5 72B	98.1	97.3	97.5	95.0	–
Qwen2.5 7B	92.4	91.9	92.3	91.3	–
Blended Latent Diffusion	–	–	–	–	98.1
Latent Consistency Model	–	–	–	–	97.5

Table 2: Average coverage and precision (higher is better) of Anchors, LORE, and their *ConLUX*-augmented versions.

Models	Coverage (%) ↑				Precision (%) ↑			
	Anchors	Anchors*	LORE	LORE*	Anchors	Anchors*	LORE	LORE*
DeepSeek-V3/Large.	4.7 ± 0.7	$\mathbf{23.5} \pm 2.0$	2.3 ± 0.3	$\mathbf{22.4} \pm 2.1$	79.2 ± 1.3	$\mathbf{96.2} \pm 0.1$	63.1 ± 1.7	$\mathbf{73.9} \pm 2.2$
DeepSeek-V3/Fake.	3.4 ± 0.7	$\mathbf{22.8} \pm 1.7$	2.4 ± 0.5	$\mathbf{18.5} \pm 2.0$	80.5 ± 0.3	$\mathbf{93.6} \pm 0.5$	60.4 ± 2.5	$\mathbf{73.1} \pm 2.4$
BERT/Large.	3.9 ± 0.4	$\mathbf{24.5} \pm 1.4$	3.6 ± 0.5	$\mathbf{20.3} \pm 1.8$	81.7 ± 0.2	$\mathbf{91.0} \pm 0.8$	64.2 ± 1.1	$\mathbf{79.4} \pm 1.1$
BERT/Fake.	4.1 ± 0.3	$\mathbf{24.2} \pm 1.7$	2.9 ± 0.1	$\mathbf{22.6} \pm 1.7$	78.2 ± 1.3	$\mathbf{89.8} \pm 0.8$	62.2 ± 2.5	$\mathbf{76.5} \pm 2.1$
YOLOv8/ImageNet	23.5 ± 1.5	$\mathbf{31.9} \pm 1.7$	19.4 ± 1.8	$\mathbf{26.1} \pm 2.1$	78.3 ± 3.2	$\mathbf{95.2} \pm 1.3$	68.3 ± 2.3	$\mathbf{81.8} \pm 2.5$
YOLOv8/Caltech101	16.2 ± 1.3	$\mathbf{23.6} \pm 1.6$	18.7 ± 1.3	$\mathbf{23.5} \pm 1.5$	84.3 ± 0.5	$\mathbf{96.7} \pm 0.8$	72.9 ± 1.3	$\mathbf{82.1} \pm 2.1$
YOLOv8/CUB	17.8 ± 1.5	$\mathbf{22.5} \pm 1.1$	20.1 ± 1.2	$\mathbf{24.1} \pm 1.6$	83.5 ± 0.1	$\mathbf{95.5} \pm 0.2$	68.8 ± 2.7	$\mathbf{75.8} \pm 1.7$
ViT/ImageNet	22.3 ± 1.2	$\mathbf{31.5} \pm 1.4$	20.7 ± 1.1	$\mathbf{23.9} \pm 1.5$	82.4 ± 0.5	$\mathbf{93.6} \pm 0.3$	75.2 ± 0.9	$\mathbf{84.3} \pm 1.2$
ViT/Caltech101	18.9 ± 1.8	$\mathbf{26.7} \pm 2.0$	19.4 ± 1.7	$\mathbf{24.1} \pm 1.4$	83.5 ± 0.8	$\mathbf{97.0} \pm 0.4$	71.2 ± 1.2	$\mathbf{80.9} \pm 0.7$
ViT/CUB	21.9 ± 1.0	$\mathbf{29.1} \pm 1.0$	26.4 ± 1.1	$\mathbf{30.7} \pm 1.0$	85.6 ± 0.3	$\mathbf{94.3} \pm 1.2$	74.5 ± 1.3	$\mathbf{82.0} \pm 1.0$
ResNet-50/ImageNet	21.5 ± 1.2	$\mathbf{29.9} \pm 1.6$	20.1 ± 1.4	$\mathbf{29.3} \pm 2.1$	81.8 ± 2.3	$\mathbf{97.9} \pm 0.5$	76.8 ± 2.1	$\mathbf{83.6} \pm 1.8$
ResNet-50/Caltech101	23.4 ± 1.2	$\mathbf{27.8} \pm 1.2$	19.7 ± 1.3	$\mathbf{28.4} \pm 1.2$	84.5 ± 4.4	$\mathbf{96.4} \pm 1.0$	77.3 ± 1.4	$\mathbf{85.1} \pm 1.0$
ResNet-50/CUB	17.0 ± 1.1	$\mathbf{26.1} \pm 1.2$	19.1 ± 1.4	$\mathbf{20.6} \pm 1.3$	78.3 ± 4.3	$\mathbf{89.5} \pm 0.2$	56.0 ± 2.0	$\mathbf{72.7} \pm 3.9$
Qwen2.5-VL	10.5 ± 0.9	$\mathbf{21.8} \pm 1.9$	8.5 ± 1.1	$\mathbf{21.8} \pm 1.9$	75.4 ± 2.3	$\mathbf{93.1} \pm 0.5$	64.7 ± 1.2	$\mathbf{73.2} \pm 1.4$

instruct DeepSeek-V3 to generate 10 perturbed samples by altering specific given concepts, and using a checker to verify whether the generated sentences meet the requirements.

Specifically, for human-annotated concepts, we use two named entity recognition datasets, CoNLL-2003 (Tjong Kim Sang & De Meulder, 2003) and OntoNotes 5.0 (Pradhan et al., 2013). The concepts include entity types such as *person*, *location*, and *organization*. We use a fine-tuned BERT model (Devlin et al., 2018) as the checker. For model-extracted concepts, as Ludan et al. (2023) have verified that GPT-based models can extract high-quality textual concepts and verify concept satisfaction, we use GPT-4o to extract concepts from sentences of the Large Movie Review (Maas et al., 2011) and Fake News (Pérez-Rosas et al., 2018) datasets, and also use GPT-4o as the checker.

For image data, we used Blended Latent Diffusion (Avrahami et al., 2023) as the concept-feature mapping model, and conducted our experiment on images from the validation set of COCO dataset (Lin et al., 2014), where objects in images are annotated by humans. For each image, we generated 10 samples by altering these concepts, and checked if the generated samples satisfy the requirements by an object-detection YOLO11 model (Jocher et al., 2023).

Metrics and Results We evaluated perturbation fidelity using accuracy, i.e. the proportion of generated samples that satisfy the given concept requirements. Table 1 shows that our perturbation models achieve an average accuracy of 96.8%, indicating that *ConLUX* can faithfully generate samples that meet the given concept requirements.

4.2 Explanation Fidelity Evaluation

Our evaluation consists of two parts: (1) assessing the fidelity improvement that *ConLUX* brings to existing local model-agnostic explanation methods (Anchors, LIME, LORE, and Kernel SHAP (KSHAP for short)), and (2) comparing the fidelity of *ConLUX* with state-of-the-art concept-based methods: TBM (Ludan et al., 2023) and LACOAT (Yu et al., 2024) for text models, and EAC (Sun et al., 2023) and ConceptLIME (Tan et al., 2024) for image models.

4.2.1 Experimental Setup

We conducted the experiments on two text classification tasks, three image classification tasks, and one multimodal task.

Table 3: Average AOPC (higher is better) and $\mathrm{accuracy_a}$ (lower is better) of LIME, KSHAP, and their *ConLUX*-augmented versions.

Models	AOPC (%)↑				$\mathrm{Accuracy_a}$ (%) ↓			
	LIME	LIME*	KSHAP	KSHAP*	LIME	LIME*	KSHAP	KSHAP*
DeepSeek-V3/Large.	22.1 ± 4.1	$\mathbf{46.4} \pm 3.3$	32.7 ± 3.3	$\mathbf{52.5} \pm 3.3$	77.3 ± 1.9	$\mathbf{49.3} \pm 0.8$	73.1 ± 1.7	$\mathbf{46.3} \pm 1.2$
DeepSeek-V3/Fake.	21.7 ± 1.7	$\mathbf{48.3} \pm 2.1$	30.4 ± 1.4	$\mathbf{49.6} \pm 1.9$	76.4 ± 1.3	$\mathbf{51.1} \pm 1.1$	71.6 ± 1.4	$\mathbf{49.6} \pm 1.1$
BERT/Large.	24.3 ± 3.2	$\mathbf{45.6} \pm 3.6$	37.9 ± 4.1	$\mathbf{55.3} \pm 2.3$	75.7 ± 1.2	$\mathbf{47.7} \pm 1.9$	60.3 ± 1.7	$\mathbf{40.2} \pm 2.1$
BERT/Fake.	25.7 ± 2.7	$\mathbf{47.2} \pm 3.1$	33.8 ± 3.0	$\mathbf{51.9} \pm 3.4$	76.2 ± 1.3	$\mathbf{53.2} \pm 2.0$	67.9 ± 1.2	$\mathbf{45.1} \pm 1.5$
YOLOv8/ImageNet	38.9 ± 2.0	$\mathbf{51.4} \pm 1.8$	41.7 ± 1.5	$\mathbf{58.7} \pm 1.7$	14.9 ± 1.1	$\mathbf{4.7} \pm 1.3$	23.9 ± 1.3	$\mathbf{6.5} \pm 0.9$
YOLOv8/Caltech101	44.0 ± 4.0	$\mathbf{62.2} \pm 2.8$	43.1 ± 3.3	$\mathbf{59.6} \pm 3.6$	23.8 ± 1.5	$\mathbf{13.7} \pm 1.2$	21.1 ± 1.6	$\mathbf{11.3} \pm 1.6$
YOLOv8/CUB	51.8 ± 3.2	$\mathbf{63.7} \pm 3.0$	52.9 ± 3.1	$\mathbf{62.4} \pm 3.2$	7.0 ± 1.2	$\mathbf{1.0} \pm 0.5$	8.1 ± 1.4	$\mathbf{3.2} \pm 1.2$
ViT/ImageNet	45.8 ± 2.8	$\mathbf{56.3} \pm 1.7$	46.8 ± 2.1	$\mathbf{62.0} \pm 3.3$	22.4 ± 1.2	$\mathbf{5.4} \pm 1.0$	18.6 ± 1.4	$\mathbf{5.3} \pm 1.4$
ViT/Caltech101	47.3 ± 2.8	$\mathbf{64.5} \pm 3.5$	48.1 ± 3.0	$\mathbf{67.3} \pm 3.4$	22.3 ± 1.3	$\mathbf{13.4} \pm 1.2$	21.3 ± 1.9	$\mathbf{11.8} \pm 1.5$
ViT/CUB	59.2 ± 2.1	$\mathbf{70.8} \pm 2.1$	62.9 ± 2.2	$\mathbf{69.7} \pm 2.2$	6.5 ± 0.8	$\mathbf{1.6} \pm 0.3$	8.2 ± 1.0	$\mathbf{3.2} \pm 0.7$
ResNet-50/ImageNet	22.3 ± 3.0	$\mathbf{32.3} \pm 2.9$	24.1 ± 2.8	$\mathbf{33.9} \pm 2.4$	21.1 ± 1.4	$\mathbf{3.6} \pm 1.0$	19.4 ± 0.8	$\mathbf{3.9} \pm 0.9$
ResNet-50/Caltech101	29.7 ± 3.7	$\mathbf{48.1} \pm 3.2$	27.5 ± 3.4	$\mathbf{50.1} \pm 2.8$	17.7 ± 1.3	$\mathbf{7.9} \pm 1.2$	15.8 ± 1.3	$\mathbf{6.3} \pm 1.0$
ResNet-50/CUB	46.5 ± 3.0	$\mathbf{56.3} \pm 3.5$	45.2 ± 3.4	$\mathbf{56.7} \pm 3.0$	4.9 ± 1.1	$\mathbf{2.1} \pm 0.9$	7.4 ± 1.6	$\mathbf{4.5} \pm 1.0$
Qwen2.5-VL	24.5 ± 2.1	$\mathbf{40.7} \pm 2.4$	31.6 ± 2.8	$\mathbf{39.8} \pm 2.7$	21.2 ± 1.2	$\mathbf{5.1} \pm 0.9$	16.9 ± 1.4	$\mathbf{3.0} \pm 0.6$

Text tasks Text classification models take a text sequence as input and predict its category. We used the Large Movie Review (Maas et al., 2011) and Fake News (Pérez-Rosas et al., 2018) datasets, explaining validation-set predictions from a fine-tuned BERT (Devlin et al., 2018) and DeepSeek-V3 with in-context examples. For the baseline explanation methods, we kept the default settings. For *ConLUX*, we followed the settings in the perturbation fidelity evaluation.

Image tasks Image classification models take an image as input and predict its category. We conducted experiments on ImageNet (Deng et al., 2009), Caltech-101 (Fei-Fei et al., 2007), and CUB-200 (Welinder et al., 2010). For each dataset, we used a fine-tuned YOLOv8, Vision Transformer (ViT) (Oquab et al., 2023; Darcet et al., 2023), and ResNet-50 (He et al., 2016) to classify images from the validation set and explain the models' local behaviors. For the vanilla methods, we followed LIME to use superpixels obtained by Quickshift (Jiang et al., 2018) as predicates. For *ConLUX*, we used SAM (Kirillov et al., 2023) as the concept-extracting model, and Blended Latent Diffusion as the concept-feature mapping model. For EAC and ConceptLIME, we kept their default settings.

Multimodal tasks We follow the settings in text and image tasks, explain multimodal Qwen2.5-VL on 250 randomly selected Yes/No Question of the VQAv2 (Goyal et al., 2017) dataset, and compare the fidelity of *ConLUX*-augmented explanations with the vanilla ones.

As TBM, LACOAT, EAC, and ConceptLIME are not applicable to multimodal tasks, we compare the fidelity of these methods and *ConLUX*-augmented explanations only on text and image tasks.

4.2.2 Evaluation Metrics

Fidelity reflects how faithfully an explanation describes a target model. As the explanation form of these methods varies, we used different metrics to evaluate their fidelity.

For Anchors and LORE, following the settings of their original papers, we used **coverage** and **precision** as the fidelity metrics (which are named differently in the LORE paper). Given a target model f, an input $\boldsymbol{x}$, and a distribution $D_{\boldsymbol{x}}$ derived from the perturbation model, and the corresponding explanation g, we defined the coverage as $\mathrm{cov}(\boldsymbol{x}; f, g) = \mathbb{E}_{\boldsymbol{z} \sim D_{\boldsymbol{x}}}[g(\boldsymbol{z})]$, which indicates the proportion of inputs in the distribution that match the rule. We defined precision as $\mathrm{prec}(\boldsymbol{x}; f, g) = \mathbb{E}_{\boldsymbol{z} \sim D_{\boldsymbol{x}}}[\mathbf{1}_{f(\boldsymbol{z})=y} | g(\boldsymbol{z})]$, where y is the consequence of the rules in g with $y = f(\boldsymbol{x})$ for factual rules and $y \neq f(\boldsymbol{x})$ for counterfactual rules. Precision indicates the proportion of covered inputs for which g correctly predicts the model outputs.

As LIME and KSHAP are attribution-based local surrogates, we measured their explanation fidelity by **deletion experiments**. We used *Area Over most relevant first perturbation curve* (AOPC), and $\mathrm{accuracy_a}$ as the metrics (Samek et al., 2016; Modarressi et al., 2023). Given a target model f, an input $\boldsymbol{x}$, the model output $y = f(\boldsymbol{x})$, their corresponding explanation g, and $\boldsymbol{x}^{(k)}$ that is generated by masking the $k\%$ most important predicates in $\boldsymbol{x}$, AOPC and $\mathrm{accuracy_a}$ are defined as follows:

- **AOPC:** Let $\mathrm{AOPC}_k = \frac{1}{|\mathbb{T}|} \sum_{\boldsymbol{x} \in \mathbb{T}} (p_f(y|\boldsymbol{x}) - p_f(y|\boldsymbol{x}^{(k)}))$, where $p_f(y|\boldsymbol{x})$ is the probability of f to output y given the input $\boldsymbol{x}$, and $\mathbb{T}$ is the set of all test inputs. AOPC_k indicates the average change

Table 4: Average accuracy (%) (higher accuracy is better) of TBM, EAC, *ConLUX*-augmented KSHAP (denoted as KSHAP*), and *ConLUX* unified explanations.

Models	TBM	LACOAT	EAC	ConceptLIME	KSHAP*	*ConLUX*
DeepSeek-V3/Large.	78.3	65.8	–	–	87.1	**90.7**
DeepSeek-V3/Fake.	75.4	66.7	–	–	82.3	**87.3**
BERT/Large.	82.4	73.4	–	–	86.3	**92.6**
BERT/Fake.	81.4	75.2	–	–	85.6	**91.8**
YOLOv8/ImageNet	–	–	82.7	81.3	88.1	**92.4**
YOLOv8/Caltech101	–	–	84.6	83.9	88.7	**92.1**
YOLOv8/CUB	–	–	76.8	72.6	86.7	**89.3**
ViT/ImageNet	–	–	83.4	79.3	88.6	**91.6**
ViT/Caltech101	–	–	82.7	84.1	89.4	**92.6**
ViT/CUB	–	–	74.3	72.1	86.4	**90.3**
ResNet-50/ImageNet	–	–	78.1	77.9	85.0	**90.2**
ResNet-50/Caltech101	–	–	79.7	76.8	85.4	**91.5**
ResNet-50/CUB	–	–	77.4	77.1	83.1	**89.1**

of the model output when masking the $k\%$ most important predicates. A higher AOPC_k indicates a better explanation. AOPC is defined as $\mathrm{AOPC} = \sum_{k=1}^{100} \mathrm{AOPC}_k/100$.

- **Accuracy$_\mathbf{a}$**: $\mathrm{Accuracy_a}$ indicates the proportion of inputs among all $\boldsymbol{x}^{(k)}$ that the target model gives the same output as the original input $\boldsymbol{x}$, i.e. $\mathbb{E}[f(\boldsymbol{x}^{(k)}) = f(\boldsymbol{x})]$. Specifically, $\mathrm{accuracy_a}$ is different from the standard accuracy, and a lower $\mathrm{accuracy_a}$ indicates a better explanation.

Specifically, we only considered the predicates that positively contribute to $f(\boldsymbol{x})$.

When comparing to state-of-the-art concept-based explanation methods, we considered *ConLUX* unified explanations as one of the main advantages of *ConLUX* over existing concept-based approaches is offering rich forms of explanations. Since there are multiple forms of explanations, considering that they can all serve as local surrogate models, we defined the metrics as follows (Balagopalan et al., 2022; Ismail et al., 2021): given a target model f, an input $\boldsymbol{x}$, a perturbation distribution $\mathbb{D}_{\boldsymbol{x}}$, their corresponding explanation g, and a performance metric L (e.g. accuracy, F1 score, etc.), we defined the (in-)fidelity as $E_{\boldsymbol{z}\sim\mathbb{D}_{\boldsymbol{x}}}[L(\mathbf{1}_{\{\boldsymbol{z}|f(\boldsymbol{z})=f(\boldsymbol{x})\}}(\boldsymbol{z}), g(\boldsymbol{z}))]$, which indicates the performance of using g to approximate the local behavior of f, i.e., predict if f($\boldsymbol{z}$) is the same as $f(\boldsymbol{x})$. Here, we used accuracy as the performance metric.

For a fair comparison, we calculate fidelity in different neighborhoods for each setup: 1) When comparing *ConLUX*-augmented explanations with their feature-level versions, we use feature-level neighborhood. 2) When comparing *ConLUX* unified explanations with other concept-based explanations, we use concept-level neighborhood.

4.2.3 Results

Table 2 and 3 show the fidelity of Anchors, LORE, LIME, KSHAP, and their *ConLUX*-augmented versions. For Anchors and LORE, *ConLUX* improves the average coverage by 11.2% and 9.5%, and the average precision by 13.0% and 10.6%, respectively. For LIME and KSHAP, *ConLUX* improves the average AOPC by 0.164 and 0.151, and decreases the average $\mathrm{accuracy_a}$ by 14.8% and 13.8%, respectively. We did paired t-tests for setups that only differ in whether to apply *ConLUX*, which indicates with over 99% confidence the improvement is significant.

Table 4 shows the fidelity of *ConLUX*-augmented explanations and state-of-the-art concept-based explanation methods: TBM and LACOAT for text tasks, and EAC and ConceptLIME for image tasks, which shows that *ConLUX* helps KSHAP to achieve higher fidelity than these methods, and *ConLUX* local unified explanations further achieve 4.52% more fidelity than *ConLUX*-augmented KSHAP.

4.3 Human Evaluation

ConLUX provides explanations in various forms beyond attributions, including sufficient conditions and counterfactuals, which can more effectively support users in downstream decision-making tasks.

Specifically, sufficient conditions help users anticipate the model's behavior on unseen inputs, while counterfactuals help understand how the model's prediction changes under specific input modifications. To evaluate the usefulness of introducing these explanation forms, we conducted a

Table 5: Human Evaluation Results on EAC, *ConLUX*-augmented Anchors (denoted as Anchors*), and *ConLUX*-augmented LORE (denoted as LORE*).

Methods	Sufficient Conditions		Counterfactuals	
	EAC	Anchors*	EAC	LORE*
coverage_u (%)	57.0	60.0	29.3	36.1
Precision_u (%)	63.9	72.0	82.8	97.0

user study following the setup in Ribeiro et al. (2018); Warren et al. (2022); Hase & Bansal (2020), comparing our concept-based sufficient conditions and counterfactuals with concept-based attribution explanations provided by EAC (Sun et al., 2023) across two tasks.

We recruited 18 subjects, all graduate students with machine learning experience but no explainable AI expertise. Each subject completed a questionnaire with 10 tests per task.

For the sufficient condition task, in each test, participants were first shown an image randomly sampled from the ImageNet dataset, along with YOLOv8's prediction and an explanation randomly selected from EAC or *ConLUX*-augmented Anchors explanations. They were then shown 5 new images generated by concept-level perturbing the original image, and asked to predict whether the model would output the same label for each new image based on the explanation. They could answer "yes" or "no." We evaluate responses using two metrics: coverage_u, the proportion of "yes" responses, and precision_u, the proportion of correct predictions among the "yes" responses.

The counterfactual task follows a similar setup, except that participants were shown *ConLUX*-augmented LORE explanations instead of Anchors and were asked to predict whether the model would output a *different* label for each new image.

Table 5 shows the average coverage_u and precision_u of the two tasks. The results show that *ConLUX*-augmented explanations outperform EAC by 3.0% in coverage_u and 8.1% in precision_u for the sufficient condition task, and by 6.8% and 14.2% respectively for the counterfactual task. These findings indicate that *ConLUX*-augmented explanations are more effective in helping users apply explanations to reason about model behavior.

5 Limitations and Discussions

One practical consideration of *ConLUX* is its reliance on generative models for concept–feature mapping. Therefore, we examine whether its perturbation fidelity is robust to different choices of generative models. We replace the generative model used in *ConLUX* with alternative models and measure the resulting perturbation fidelity. Specifically, we test two additional Qwen2.5 models for text perturbations and Latent Consistency Model (Luo et al., 2023) for image perturbations. Table 1 summarizes the results. Across all tested models, *ConLUX* maintains high perturbation fidelity. These results suggest that *ConLUX* is flexible in the choice of generative models for concept–feature mapping. This flexibility brings two benefits: (1) users can select models based on their own preferences or resource constraints, and (2) users can employ multiple generative models to increase the diversity of generated samples and reduce potential biases from relying on a single model. Further results and discussion are provided in Appendix B.

6 Conclusion

We have proposed *ConLUX*, a general framework that elevates local model-agnostic explanation methods to the concept level. *ConLUX* offers unified explanations combining attributions, sufficient conditions, and counterfactuals. This satisfies diverse user needs and fills the current gap of concept-based explanations lacking forms beyond attributions. *ConLUX* achieves this by utilizing large pre-trained models to extract high-level concepts, and extending perturbation models to sample in the concept space. We have instantiated *ConLUX* on Anchors, LIME, LORE, and Kernel SHAP, and demonstrated the state-of-the-art performance of *ConLUX* by empirical evaluations. *ConLUX* shows that it is unnecessary to design concept-based explanation methods from scratch, as existing local model-agnostic methods can be easily elevated to concept level in a lightweight manner.

AI USE STATEMENT

For manuscript preparation, we use LLMs to refine and polish human-written text and use DeepResearch to identify related literature. We do not use LLMs to generate the main manuscript text or research ideas.

ETHICS STATEMENT

ConLUX aims to make black-box model behavior easier to understand by presenting local explanations with human-interpretable concepts. Such explanations can help practitioners and end-users diagnose model errors, compare different models, and decide when a model's prediction should not be trusted, especially in settings where low-level feature explanations are difficult to interpret.

At the same time, explanation methods are inherently dual-use. By revealing which concepts support or oppose a prediction, *ConLUX* may help malicious users infer local decision rules and craft inputs that evade a model or target a specific class. This risk is not unique to our framework, but it suggests that deployments in security-sensitive domains should control access to explanations, monitor suspicious querying behavior, and combine interpretability tools with robustness and abuse-mitigation measures.

Our human evaluation involved 18 graduate students with machine learning experience but no explainable AI expertise. The user study received ethics approval. Participation was voluntary, and participants were compensated at twice the local minimum hourly wage applicable at our institution's location. The participant screening procedure and example questionnaires are described in Appendix B.7.

REPRODUCIBILITY STATEMENT

We describe the framework in Section 3 and provide additional details on concept extraction and concept–feature mapping in Appendix A. Section 4 describes the experimental datasets, models, evaluation metrics, and comparisons. Appendix B provides further experimental details, including hardware, hyperparameters, prompts, robustness analyses, and runtime measurements. Participant screening and example questionnaires for the human evaluation are provided in Appendix B.7.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Daniel W Apley and Jingyu Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(4): 1059–1086, 2020.

Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Trans. Graph.*, 42(4), July 2023. ISSN 0730-0301. doi: 10.1145/3592450. URL https://doi.org/10.1145/3592450.

Aparna Balagopalan, Haoran Zhang, Kimia Hamidieh, Thomas Hartvigsen, Frank Rudzicz, and Marzyeh Ghassemi. The road to explainability is paved with bias: Measuring the fairness of explanations. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1194–1206, 2022.

Maria A Bravo, Sudhanshu Mittal, Simon Ging, and Thomas Brox. Open-vocabulary attribute detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7041–7050, 2023.

Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, and Stefano Melacci. Logic explained networks. *Artificial Intelligence*, 314:103822, January 2023. ISSN 0004-3702. doi: 10.1016/j.artint.2022.103822. arXiv:2108.05149 [cs].

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. (arXiv:2209.11222), September 2022. doi: 10.48550/arXiv.2209.11222. URL http://arxiv.org/abs/2209.11222. arXiv:2209.11222 [cs].

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. (arXiv:2309.08600), October 2023. doi: 10.48550/arXiv.2309.08600. URL http://arxiv.org/abs/2309.08600. arXiv:2309.08600 [cs].

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. (Apache 2.0).

Radwa El Shawi. Conceptglassbox: Guided concept-based explanation for deep neural networks. *Cognitive Computation*, 16(5):2660–2673, September 2024. ISSN 1866-9964. doi: 10.1007/s12559-024-10262-8.

Gemini Team et al. Gemini: A family of highly capable multimodal models, 2024. URL https://arxiv.org/abs/2312.11805.

Li Fei-Fei, Rob Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer vision and Image understanding*, 106(1):59–70, 2007.

Thomas Fel, Victor Boutin, Louis Béthune, Remi Cadene, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. *Advances in Neural Information Processing Systems*, 36: 54805–54818, December 2023a.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. (arXiv:2211.10154), March 2023b. doi: 10.48550/arXiv.2211.10154. URL http://arxiv.org/abs/2211.10154. arXiv:2211.10154.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. (arXiv:1902.03129), October 2019a. doi: 10.48550/arXiv.1902.03129. URL http://arxiv.org/abs/1902.03129. arXiv:1902.03129 [cs, stat].

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards automatic concept-based explanations. (arXiv:1902.03129), October 2019b. doi: 10.48550/arXiv.1902.03129. URL http://arxiv.org/abs/1902.03129. arXiv:1902.03129 [cs, stat].

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, pp. 9273–9282, 2019c.

Alex Goldstein, Adam Kapelner, Justin Bleich, and Emil Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015. doi: 10.1080/10618600.2014.907095.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). (arXiv:1907.07165), February 2020. doi: 10.48550/arXiv.1907.07165. URL http://arxiv.org/abs/1907.07165. arXiv:1907.07165 [cs].

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. Local rule-based explanations of black box decision systems. *CoRR*, abs/1805.10820, 2018. URL http://arxiv.org/abs/1805.10820.

Peter Hase and Mohit Bansal. Evaluating explainable AI: Which algorithmic explanations help users predict model behavior? In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5540–5552, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.491. URL https://aclanthology.org/2020.acl-main.491/.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Aya Abdelsalam Ismail, Hector Corrada Bravo, and Soheil Feizi. Improving deep learning interpretability by saliency guided training. *Advances in Neural Information Processing Systems*, 34: 26726–26739, 2021.

Heinrich Jiang, Jennifer Jang, and Samory Kpotufe. Quickshift++: Provably good initializations for sample-based mean shift, 2018. URL https://arxiv.org/abs/1805.07909.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Conference on Learning Theory*, pp. 228–251. PMLR, 2013.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, and Fernanda Viegas. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pp. 2668–2677. PMLR, 2018. URL http://proceedings.mlr.press/v80/kim18d.html.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. (arXiv:2306.01574), June 2023. doi: 10.48550/arXiv.2306.01574. URL http://arxiv.org/abs/2306.01574. arXiv:2306.01574.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348. PMLR, November 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014. URL http://arxiv.org/abs/1405.0312.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

Junhao Liu and Xin Zhang. Rex: A framework for incorporating temporal information in model-agnostic local explanation techniques. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 18888–18896, 2025.

Josh Magnus Ludan, Qing Lyu, Yue Yang, Liam Dugan, Mark Yatskar, and Chris Callison-Burch. Interpretable-by-design text classification with iteratively generated concept bottleneck. (arXiv:2310.19660), October 2023. doi: 10.48550/arXiv.2310.19660. URL http://arxiv.org/abs/2310.19660. arXiv:2310.19660 [cs].

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 4765–4774, 2017.

Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module, 2023. URL https://arxiv.org/abs/2311.05556.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P11-1015.

Gary C McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1 (1):93–100, 2009.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2649–2664, Toronto, Canada, July 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.acl-long.149.

Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. (arXiv:2304.06129), June 2023. doi: 10.48550/arXiv.2304.06129. URL http://arxiv.org/abs/2304.06129. arXiv:2304.06129 [cs].

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 3391–3401, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1287/.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. (arXiv:2312.12936), December 2023. URL http://arxiv.org/abs/2312.12936. arXiv:2312.12936 [cs].

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In Julia Hockenmaier and Sebastian Riedel (eds.), *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pp. 143–152, Sofia, Bulgaria, August 2013. Association for Computational Linguistics. URL https://aclanthology.org/W13-3516/.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. In Balaji Krishnapuram, Mohak Shah, Alexander J. Smola, Charu C. Aggarwal, Dou Shen, and Rajeev Rastogi (eds.), *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144. ACM, 2016.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In Sheila A. McIlraith and Kilian Q. Weinberger (eds.), *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of*

Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018, pp. 1527–1535. AAAI Press, 2018.

Salvatore Ruggieri. Yadt: Yet another decision tree builder. In *16th IEEE International Conference on Tools with Artificial Intelligence*, pp. 260–265. IEEE, 2004. URL https://ieeexplore.ieee.org/abstract/document/1374196/.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016.

Sharath M. Shankaranarayana and Davor Runje. Alime: Autoencoder based approach for local interpretability. (arXiv:1909.02437), September 2019. doi: 10.48550/arXiv.1909.02437. URL http://arxiv.org/abs/1909.02437. arXiv:1909.02437 [cs, stat].

Ao Sun, Pingchuan Ma, Yuanyuan Yuan, and Shuai Wang. Explain any concept: Segment anything meets concept-based explanation. (arXiv:2305.10289), May 2023. doi: 10.48550/arXiv.2305.10289. URL http://arxiv.org/abs/2305.10289. arXiv:2305.10289 [cs].

Lidan Tan, Changwu Huang, and Xin Yao. A concept-based local interpretable model-agnostic explanation approach for deep neural networks in image classification. In *International Conference on Intelligent Information Processing*, pp. 119–133. Springer, 2024.

Zeren Tan, Yang Tian, and Jian Li. Glime: General, stable and local lime explanation, 2023.

Aditya Taparia, Som Sagar, and Ransalu Senanayake. Explainable concept generation through vision-language preference learning for understanding neural networks' internal representations. In *Forty-second International Conference on Machine Learning*.

Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pp. 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

Payal Varshney, Adriano Lucieri, Christoph Balada, Andreas Dengel, and Sheraz Ahmed. Generating counterfactual trajectories with latent diffusion models for concept discovery, 2025. URL https://arxiv.org/abs/2404.10356.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr, 2018. URL https://arxiv.org/abs/1711.00399.

Xudong Wang, Shufan Li, Konstantinos Kallidromitis, Yusuke Kato, Kazuki Kozuka, and Trevor Darrell. Hierarchical open-vocabulary universal image segmentation, 2023. URL https://arxiv.org/abs/2307.00764.

Yifei Wang. A comparative analysis of model agnostic techniques for explainable artificial intelligence. *Research Reports on Computer Science*, pp. 25–33, August 2024. ISSN 2811-0013. doi: 10.37256/rrcs.3220244750.

Greta Warren, Mark T Keane, and Ruth M J Byrne. Features of explainability: How users understand counterfactual and causal explanations for categorical and continuous features in xai, 2022. URL https://arxiv.org/abs/2204.10152.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. 2010.

Cara Widmer, Md Kamruzzaman Sarker, Srikanth Nadella, Joshua Fiechter, Ion Juvina, Brandon Minnery, Pascal Hitzler, Joshua Schwartz, and Michael Raymer. Towards human-compatible xai: Explaining data differentials with concept induction over background knowledge, 2022. URL https://arxiv.org/abs/2209.13710.

Zhengxuan Wu, Karel D'Oosterlinck, Atticus Geiger, Amir Zur, and Christopher Potts. Causal proxy models for concept-based model explanations. (arXiv:2209.14279), 2022. doi: 10.48550/arXiv.2209.14279. URL http://arxiv.org/abs/2209.14279. arXiv:2209.14279 [cs].

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Pradeep Ravikumar, and Tomas Pfister. On concept-based explanations in deep neural networks. September 2019. URL https://openreview.net/forum?id=BylWYC4KwH.

Chih-Kuan Yeh, Been Kim, Sercan Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *Advances in neural information processing systems*, 33:20554–20565, 2020.

Xuemin Yu, Fahim Dalvi, Nadir Durrani, Marzia Nouri, and Hassan Sajjad. Latent concept-based explanation of nlp models. *arXiv preprint arXiv:2404.12545*, 2024.

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors. (arXiv:2006.15417), June 2021a. doi: 10.48550/arXiv.2006.15417. URL http://arxiv.org/abs/2006.15417. arXiv:2006.15417 [cs].

Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A. Ehinger, and Benjamin I. P. Rubinstein. Invertible concept-based explanations for cnn models with non-negative concept activation vectors, 2021b. URL https://arxiv.org/abs/2006.15417.

Yu Zhang, Peter Tiňo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, October 2021c. ISSN 2471-285X. doi: 10.1109/TETCI.2021.3100641. arXiv:2012.14261 [cs].

A THE *ConLUX* FRAMEWORK (CONTINUED)

In this section, we introduce the details of incorporating *ConLUX* into existing local explanation methods.

We first follow Section 3 to introduce how we extend each part for text models in detail.

A.1 PRODUCING CONCEPT

ConLUX provides predicates that describe high-level concepts **following existing concept-based methods**.

Specifically, we follow Ludan et al. (2023) for text data and Sun et al. (2023) for image data, and keep all theirs settings.

A.2 CONCEPT-FEATURE MAPPING

Text Data. We the the following prompt for predicate-feature mapping:

> Generate a sentence similar to a given sentence from the domain of {} dataset. The dataset's description is that {}.
>
> The generated sentence satisfies given concepts. Before generating the sentence, carefully read the description of each concept to understand the properties that the generated sentence must satisfy, think about how the sentence satisfies the concepts first, and then create the sentence that satisfies the concepts.
>
> Format your response as a JSON with string keys and string values. Below is an example of a valid JSON response. The JSON contains keys `thoughts`, and `answer`. End your response with \###
>
> ---
>
> Concepts:
> 1. Concept 1
> 2. Concept 2
> ...
>
> Response JSON:
> {
> "thoughts": "In this section, you explain which snippets in your text support the concepts. Be as objective as possible and ignore irrelevant information. Focus only on the snippets and avoid making guesses.",
> "answer": "A sentence that satisfies the concepts."
> }
> ###
>
> Two examples of this task being performed can be seen below. Note that the answer should be in 5 to 20 words and should be a single sentence.
>
> Example 1:
>
> Concepts:
> 1. The plot of the text is exciting, captivating, or engrossing. It may have unexpected twists, compelling conflicts, or keep the reader eagerly turning pages.
> 2. The characters in the movie are portrayed in a realistic and convincing manner. Their actions, dialogue, emotions, motivations, and development feel authentic and relatable, making them believable to the audience.
> 3. The narrative structure of the text is confusing or unclear, making it difficult to

Table 6: Explanation fidelity (accuracy) of using *ConLUX* with different concept-feature mapping models.

Model	Accuracy
DeepSeekV3	92.6
Qwen2.5 72B	91.8
Qwen2.5 7B	90.3
Blended Latent Diffusion	91.6
Latent Consistency Model	91.4

follow or comprehend the events happening within the story.
4. The text introduces some original elements or takes minor risks in the plot development, but overall, it lacks a truly unique or innovative narrative.

Response JSON:
{
"thoughts": "The snippet 'the silly and crude storyline' mentions a storyline that is described as silly and crude, indicating a lack of creativity and reliance on clichéd plot devices. The phrase 'real issues tucked between the silly and crude storyline' suggests a potentially confusing structure. It also implies real conflicts, satisfying the exciting plot concept. The term 'real issues' also supports the realistic character portrayal.",
"answer": "it's about issues most adults have to face in marriage and i think that's what i liked about it – the real issues tucked between the silly and crude storyline."
}
###

Example 2:
...
###

Perform the task below, keeping in mind to limit the response to 5 to 20 words and a single sentence. Return a valid JSON response ending with ###

Concepts:
{}

Response JSON:

Image Data. We use the following prompt for predicate-feature mapping:

Integrate this area into the background of the image.

B Empirical Evaluation Details and Discussions

B.1 Robustness to Concept-Feature Mapping Models (Continued)

Besides the results in Section 5, we additionally conducted experiments to show the explanation fidelity of using different generative models for concept-feature mapping. We explain a BERT model on the Large Movie Review dataset, and a ViT modle on ImageNet using *ConLUX* local unified explanation. Table 6 shows the results, which futher support that *ConLUX* is robust to different generative model choices.

Table 7: Accuracy of using different prompts to generate perturbations for text data.

Prompt	CoNLL	Onto.	Large.
Original	97.5	98.1	97.7
Paraphrased	97.8	98.0	97.9
Shortened	96.2	96.8	96.5

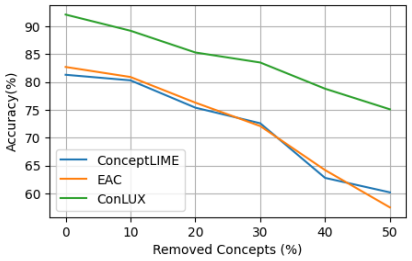

Figure 3: Effect of concept absence on explanation fidelity.

B.2 Robustness to Prompts

We also conducted an experiment using different prompts to generate perturbations. We applied two other settings, one to ask GPT-4o to make our prompt template to at only 70% of the original perturbation, and the other to ask GPT-4o to generate a prompt that is similar to the original prompt, but paraphrased. The results are shown in Table 7. We can see that the perturbation model is robust to different prompts.

B.3 Robustness to Concepts

To show the robustness of *ConLUX* to when concepts are not ideal, we conducted two ablation studies to investigate the effects of concept absence and concept correlation on explanation fidelity using ImageNet.

B.3.1 Concept absence

When explaining YOLOv8 on ImageNet, we randomly removed a certain percentage of concepts from the concept set used for generating explanations. Figure 3 shows how explanation fidelity changes as we vary the percentage of removed concepts. The results show that *ConLUX* exhibits greater robustness to concept removal compared to baseline concept-based explanation methods.

B.3.2 Concept Correlation

We conducted experiments on text data, specifically explaining BERT on the Large Movie Review dataset. In these experiments, we added a correlated concept to the concept set and measured the explanation fidelity both before and after adding the correlated concept. The results show that the accuracy slightly decreases from 92.6% to 91.9%, indicating that *ConLUX* is robust to concept correlation.

B.4 Flexibility of Supporting Various Concepts

ConLUX is a flexible framework to generate explanations with various concepts. Besides those we have shown in Section 4, we additionally implmented *ConLUX* for two more types of concepts to demonstrate its flexibility.

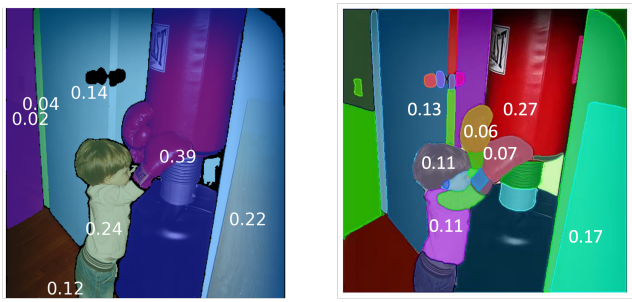

Figure 4: *ConLUX*-LIME explanation with hierarchical concepts.



Figure 5: *ConLUX*-LIME perturbation samples with attributes as concepts and the corresponding explanation.

B.4.1 Hierarchical concepts

HIPIE Wang et al. (2023) can extract hierarchical concepts, which can then be directly used in *ConLUX* to produce explanations. We have implemented this, and Figure 4 shows an example of hierarchical *ConLUX*-LIME explanations.

B.4.2 Attributes as Concepts

The Open-Vocabulary Attribute Detection dataset Bravo et al. (2023) provides attribute-level annotations (e.g., color, texture) for objects in COCO dataset. This allows *ConLUX* to generate explanations based on these object attributes instead of replacing objects entirely. Figure 5 shows some perturbation samples generated by *ConLUX*-LIME using attributes as concepts and the corresponding explanation.

B.5 Runtime Overhead

As *ConLUX* relies on large pre-trained models to perform concept–feature mapping, it inevitably introduces additional computational overhead compared to vanilla explanation methods. We show the execution time of the explanation fidelity experiments on image data in Figure 6. Here, we use Latent Consistency Model (Luo et al., 2023) for concept–feature mapping, and run the experiments on a machine with four RTX 4090 GPUs. As shown in Figure 6(a), *ConLUX* does require additional computation time due to generative model calls. However, the running time of *ConLUX*-augmented methods is practically acceptable.

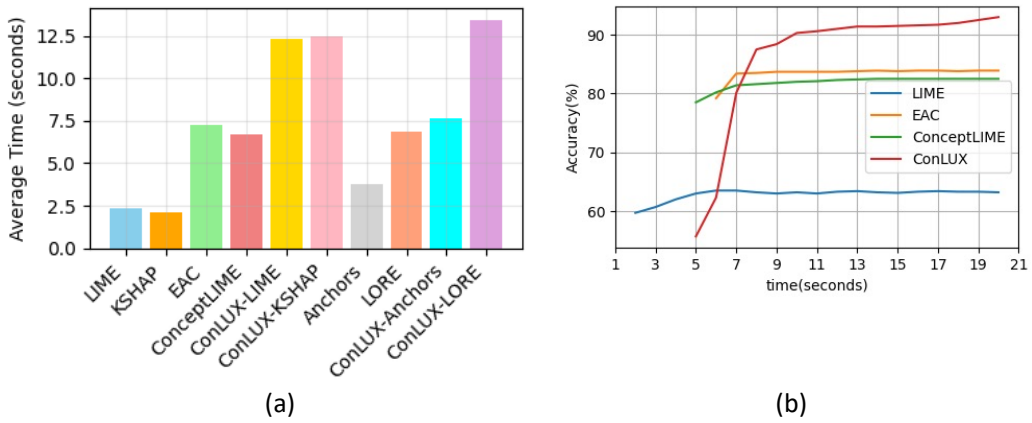

Figure 6: (a) Average running time of explanation methods we used on image data. (b) Matched-budget comparisons between LIME, EAC, ConceptLIME and *ConLUX* local unified explanations.

Table 8: Average runtime overhead and monetary cost of different explanation methods on image data.

Method	Avg Perturbations	Avg LM Calls	Avg Time (s)	Avg Cost ($)
LIME	1000	0	2.35	0
ConLUX-LIME	1000	1000	12.31	0
EAC	1000	0	7.25	0
ConceptLIME	1000	1000	6.67	0
KSHAP	1000	1000	2.11	0
ConLUX-KSHAP	1000	1000	12.45	0
Anchors	193	0	3.75	0
ConLUX-Anchors	71	71	7.67	0
LORE	851	0	6.87	0
ConLUX-LORE	873	873	13.45	0

We further conducted matched-budget comparisons (measured in running time) to evaluate the utility of each method under the same computational constraints. We explain ViT on the ImageNet dataset with LIME, EAC, and ConceptLIME, and *ConLUX* local unified explanations. Figure 6(b) shows the results. While *ConLUX* may start with lower fidelity at very low budgets due to the overhead of generative model calls, it surpasses baseline methods as the budget increases. This indicates that when users have sufficient computational time, *ConLUX* can effectively utilize it to provide higher-fidelity explanations. Especially when *ConLUX* surpasses baseline methods, the absolute time cost is still not high.

we also provide the moneytary cost and counts the calls of the generative models in Table 8. As shown, *ConLUX* does require additional computation time due to generative model calls. However, we note:

- Without compromising performance, all our experiments (Section 4 and Table 6) can use LLMs that can run on a single GPU (e.g. Qwen2.5-7B, Latent Consistency Model) without relying on commercial APIs, resulting in negligible monetary cost.
- The running time of *ConLUX*-augmented methods is practically acceptable.

B.6 Explanation Fidelity Evaluation Details

B.6.1 Setup Details

We experimented on two machines, one with an Intel i9-13900K CPU, 128 GiB RAM, and RTX 4090 GPU, and another with Intel(R) Xeon(R) Silver 4314 CPU, 256GiB RAM, and 4 RTX 4090 GPUs.

To measure the fidelity improvement brought by *ConLUX*, we keep all hyperparameters the same for both vanilla and augmented methods.

For LIME and KSHAP, we set the number of sampled inputs to 1000 except for explaining DeepSeekV3.

For Anchors, we follow the default settings.

For LORE, we set $ngen = 5$.

For the DeepSeek-V3 model, when applying it to the movie review sentiment analysis task, we simply use the following prompt:

> From now on, you should act as a sentiment analysis neural network. You should classify the sentiment of a movie review as positive or negative. If the sentence is positive, you should reply 1. Otherwise, if it's negative, you should reply 0. There may be some words that are masked in the sentence, which are represented by <UNK>. The input sentence may be empty, which is represented by <EMPTY>. You will be given the sentences to be classified, and you should reply with the sentiment of the sentence by 1 or 0.
> There are two examples:
> Sentence:
> {sentence 1 from training set}
> Sentiment:
> {label 1}
> Sentence:
> {sentence 2 from training set}
> Sentiment:
> {label 2}
> You must follow this format. Then I'll give you the sentence. Remember Your reply should be only 1 or 0. Do not contain any other content in your response. The input sentence may be empty.
> Sentence:
> {The given sentence}
> Sentiment:

When using Deepseek-V3 for fake news detection, we apply similar prompts to those used in other tasks. We obtain the classification probabilities by applying a softmax function to the raw probabilities of returning the two tokens: 0 and 1.

For Qwen2.5-VL, we directly prompt it to answer questions in VQAv2, and obtain the classification probabilities by applying a softmax function to the raw probabilities in Yes/No questions.

B.6.2 Results (continued)

In this section, we present the fidelity results that are not present in the main paper limited by space.

As Table 4 shows, either *ConLUX*-augmented KSHAP or *ConLUX* unified explanation can outperform TBM and EAC, two state-of-the-art concept-based methods, in terms of fidelity. We additionally provide the results of deletion experiments on TBM, EAC, and *ConLUX*-augmented KSHAP in Table 9.

B.7 Human Evaluation

We introduce more detail about human evaluation in this section.

Ethics and Compensation. The user study received ethics approval. Participation was voluntary, and participants were compensated at twice the local minimum hourly wage applicable at our institution's location.

User Screening. We recruited the users from students of Computer Science school in our university, and screened them before filling out the questionnaires. Therefore, the set is relatively small. To make sure our users understood the meaning of the explanation and would predict the model output

Table 9: Average accuracy (%) (higher accuracy is better) of TBM, EAC, *ConLUX*-augmented KSHAP (denoted as KSHAP*), and *ConLUX* unified explanations.

Models	AOPC↑					Accuracy$_a$(↓)				
	TBM	LACOAT	EAC	ConceptLIME	KSHAP*	TBM	LACOAT	EAC	ConceptLIME	KSHAP*
DeepSeek-V3/Large.	47.8	46.3	–	–	**54.3**	49.6	49.8	–	–	**44.1**
DeepSeek-V3/Fake.	44.2	45.2	–	–	**51.4**	47.4	48.6	–	–	**45.7**
BERT/Large.	51.0	49.6	–	–	**57.7**	46.5	43.1	–	–	**38.2**
BERT/Fake.	47.4	47.8	–	–	**53.1**	49.1	45.9	–	–	**39.3**
YOLOv8/ImageNet	–	–	52.9	51.1	**62.1**	–	–	6.2	5.9	**5.3**
YOLOv8/Caltech101	–	–	54.9	51.6	**62.0**	–	–	8.3	8.4	**7.2**
YOLOv8/CUB	–	–	52.4	51.5	**60.5**	–	–	8.9	9.1	**7.8**
ViT/ImageNet	–	–	57.1	54.8	**65.4**	–	–	7.1	6.4	**4.5**
ViT/Caltech101	–	–	62.9	59.4	**69.3**	–	–	11.5	11.9	**10.3**
ViT/CUB	–	–	54.1	56.8	**64.7**	–	–	6.3	6.1	**5.1**
ResNet-50/ImageNet	–	–	22.1	25.3	**34.9**	–	–	4.6	4.7	**2.9**
ResNet-50/Caltech101	–	–	41.4	43.6	**53.4**	–	–	7.4	7.8	**5.2**
ResNet-50/CUB	–	–	32.6	30.7	**39.1**	–	–	5.2	6.0	**4.1**



Figure 7: An question in the task of comparing EAC and *ConLUX*-augmented Anchors. This question provides user an *ConLUX*-augmented Anchors explanation.

based on the explanation rather than their own feelings, we made up a test with a counter-intuitive explanation. Then we let our users predict the model outputs of ten perturbed images. We remain the users that follow the counter-intuitive explanation.

Questionnaires details. The questionnaires are similar for all users with minor variations in the order of presentation. We presented the ten questions, Q1, Q2,. . . , and Q10 in a random order. For each question, we presented one of the explanations generated by EAC and *ConLUX*. For example, for the first question, user A received an explanation generated by EAC, while user B may received an explanation generated by *ConLUX*. The users were asked to predict the model output based on the explanation. Also, we presented the five images to be predicted in a random order. Figure 7, 8, 9, and 10 show example test questions in the questionnaire.

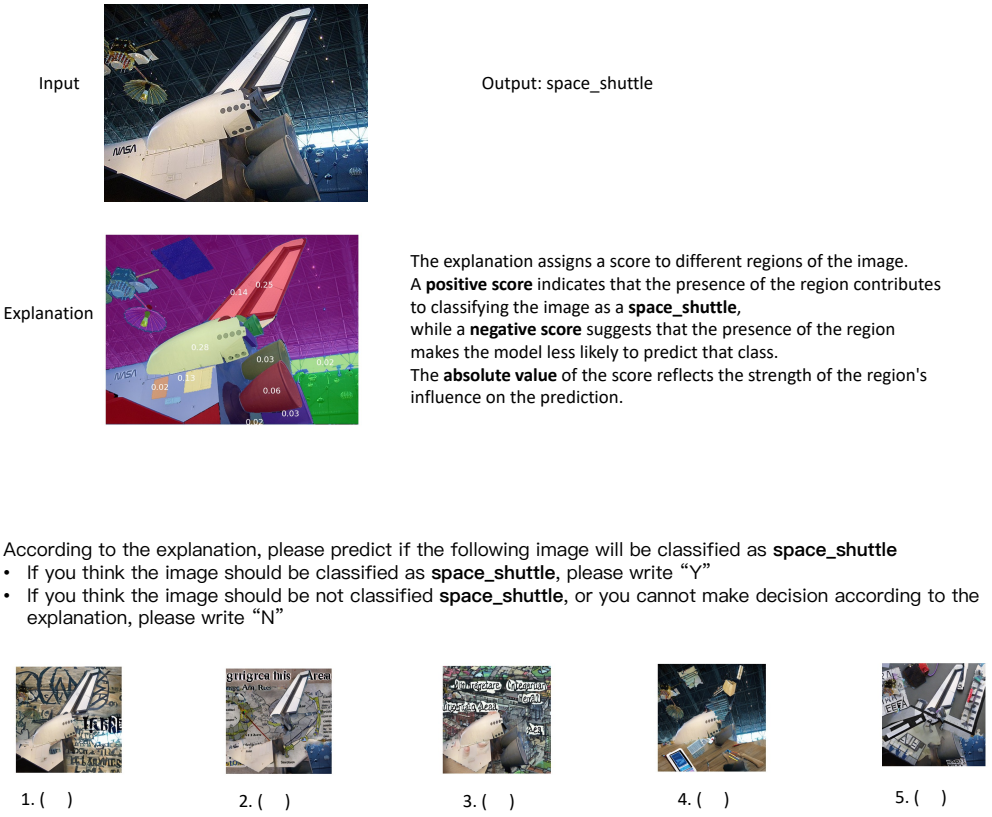

Figure 8: A question in the task of comparing EAC and *ConLUX*-augmented Anchors. This question provides the user with an EAC explanation.

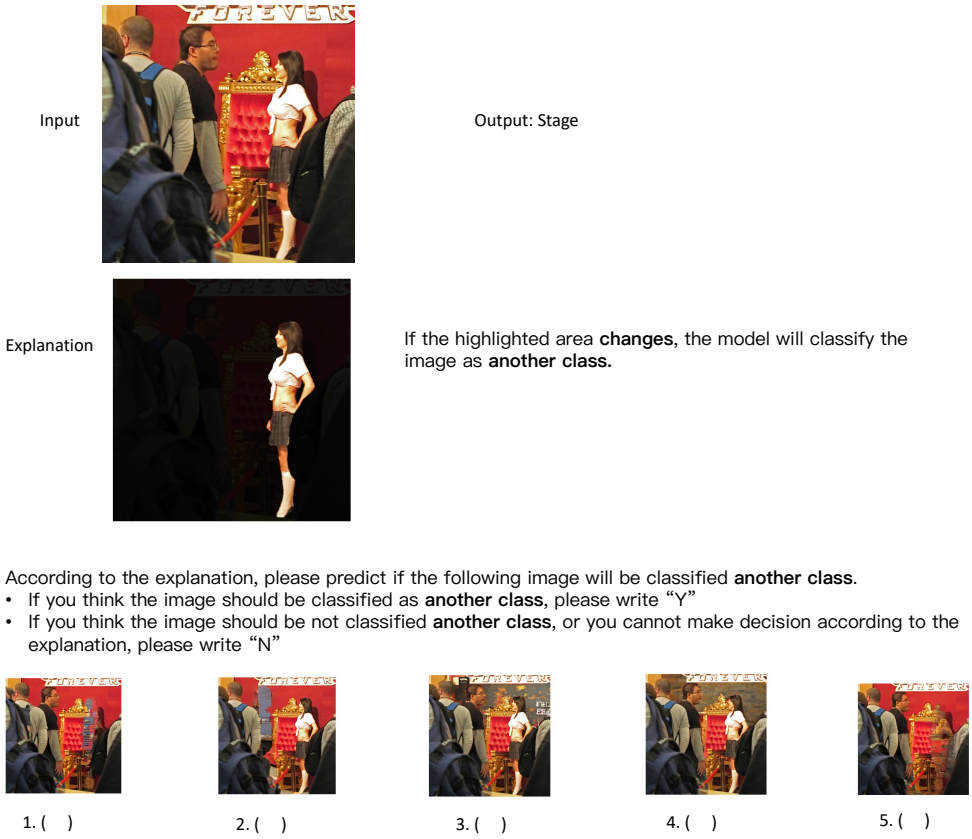

Figure 9: A question in the task of comparing EAC and *ConLUX*-augmented LORE. This question provides the user a *ConLUX*-augmented LORE explanation.



Figure 10: An question in the task of comparing EAC and *ConLUX*-augmented LORE. This question provides user an EAC explanation.