

Explaining in Diffusion: Explaining a Classifier Through Hierarchical Semantics with Text-to-Image Diffusion Models

Tahira Kazimi[†] Ritika Allada[†] Pinar Yanardag
Virginia Tech

{tahirakazimi, ritika88, pinary}@vt.edu
explain-in-diffusion.github.io

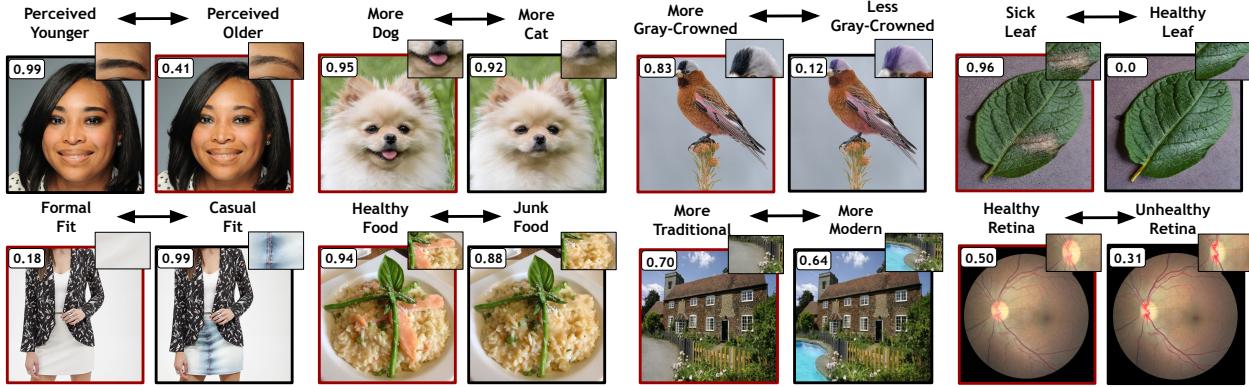


Figure 1. *DiffEx* explains the decisions of domain-specific classifiers by identifying the most influential semantics affecting their predictions. Classifier scores for each example are displayed in the top-left corner, demonstrating how classifier predictions change in response to the manipulation of different semantics (original images are shown with red borders). Our approach is capable of explaining classifiers that concentrate on individual concepts such as faces or animals (top row) as well as those that manage complex scenes involving multiple objects, such as a formal/casual fit in a fashion context (bottom row).

Abstract

Classifiers are important components in many computer vision tasks, serving as the foundational backbone of a wide variety of models employed across diverse applications. However, understanding the decision-making process of classifiers remains a significant challenge. We propose *DiffEx*, a novel method that leverages the capabilities of text-to-image diffusion models to explain classifier decisions. Unlike traditional GAN-based explainability models, which are limited to simple, single-concept analyses and typically require training a new model for each classifier, our approach can explain classifiers that focus on single concepts (such as faces or animals) as well as those that handle complex scenes involving multiple concepts. *DiffEx* employs vision-language models to create a hierarchical list of semantics, allowing users to identify not only the overarching semantic influences on classifiers (e.g., the ‘beard’ semantic in a facial classifier) but also their sub-types, such as ‘goatee’ or ‘Balbo’ beard. Our experiments demonstrate that *DiffEx* is able to cover a significantly

broader spectrum of semantics compared to its GAN counterparts, providing a hierarchical tool that delivers a more detailed and fine-grained understanding of classifier decisions.

1. Introduction

Classifiers are fundamental to computer vision tasks, forming the backbone of many models used in a broad spectrum of applications [13, 18, 25, 27, 46]. Their ability to generalize across tasks and adapt to new domains with minimal retraining makes them highly transferable, and thus they are employed extensively in fields such as healthcare [32, 34, 37, 60], finance [33, 54, 56], security [1, 24, 28], and autonomous systems [4, 53, 58]. Despite their versatility and widespread utility, understanding the decision-making process of classifiers remains a significant challenge [2, 39, 44, 65, 68]. This challenge stems largely from their “black box” nature. As images traverse through the deep, interconnected layers of the network, the features used by the classifier to make a decision become increasingly abstract and challenging to interpret. The lack of interpretable features in such models raises critical concerns, particularly in high-stakes

[†]Joint first authors.

environments such as medical diagnosis, where understanding the reasoning behind a model’s prediction is crucial for ensuring trust, accountability, and informed decision-making [9, 29, 41, 51, 69]. Explaining classifier decisions is crucial for enhancing the transparency and reliability of these models. Prior research [26] has used generative adversarial networks (GANs) [16] to interpret classifier decisions by generating counterfactual examples that manipulate GAN latent semantics. These manipulations illustrate how changes in specific attributes, like the addition of the *eyeglasses* semantic, impact classifier outputs. However, GANs are often limited to single domains, such as facial images, and typically require training a new model per classifier, which is resource and time-consuming. Moreover, in GAN-based methods, understanding which latent semantics affect classifier decisions often requires manual intervention to identify and interpret relevant features, such as recognizing that a discovered semantic controls the *eyeglasses* attribute. This manual process is not only time-consuming but also less feasible in specialized fields like medicine, where identifying intricate attributes requires substantial domain expertise, making the approach impractical in critical scenarios.

This limitation highlights the need for more automated and versatile approaches to interpret classifier decisions. Text-to-Image (T2I) diffusion models [40] emerge as a compelling alternative, widely recognized for their ability to generate high-quality images across various domains, which makes them a promising tool for explaining classifier decisions. These models offer the potential for a richer and more diverse set of semantic features compared to traditional methods. However, their ability to interpret and utilize latent space semantics remains limited in the context of diffusion models. Existing techniques for identifying meaningful semantics rely largely on supervised approaches [6, 7], which require users to craft detailed text prompts to specify particular features for editing, such as *mustache*. This process is labor-intensive and requires significant domain expertise, as users must carefully define prompts to edit the desired attributes. To make diffusion models effective for explaining classifier decisions, it is crucial to construct a comprehensive semantic corpus that covers large-scale semantics across various domains, thus minimizing reliance on manual input. In this paper, we first employ Vision-Language Models (VLMs) [31] to extract a large-scale corpus covering domain-specific hierarchical semantics (see Fig. 2). Then, we introduce a training-free method, DiffEx, which leverages this hierarchical corpus and text-to-image diffusion models to explain the decision-making process of classifiers by identifying the most influential semantics. Our method provides explanations for both coarse and fine-grained

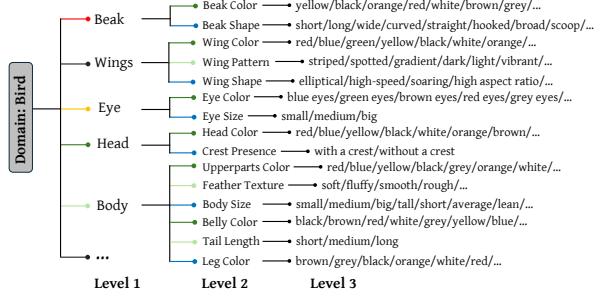


Figure 2. Hierarchical List of Attributes for the Bird Domain. We use VLMs to extract a hierarchical corpus of semantics within a given domain. This structured representation helps to illustrate how different attributes are grouped and their relationships within the broader domain, facilitating a better understanding of how each semantic contributes to the overall decision-making process of a classifier.

semantics. For example, it can recognize ‘beard’ as a coarse semantic influencing age classification scores and also demonstrate how specific beard types (such as ‘Balbo’ or ‘Anchor’ beards) impact the classifier’s scores. This hierarchical strategy provides users with an overview of the most significant semantics for a classifier, allowing them to dig deeper into particular fine-grained semantics that are essential for understanding classifier behavior. Our qualitative and quantitative experiments reveal that DiffEx offers considerably richer and more comprehensive explanations for binary and multi-class classifiers across various domains, including facial features, retinal health, and plant pathology. Our contributions are as follows:

- We propose DiffEx, a training-free approach using VLMs and T2I diffusion models to explain classifier decisions. To the best of our knowledge, this is the first hierarchical approach that explains classifier decisions.
- Our method employs a VLM to develop a comprehensive semantic corpus that spans multiple domains in a hierarchical form. We make this corpus publicly available to support future research.
- Unlike GAN-based methods, our approach can address classifiers that focus on single concepts (such as an ‘age’ classifier analyzing a headshot of a person) and also extend to classifiers that assess complex scenes (such as a ‘Modern/Traditional Architecture’ classifier evaluating entire scenes).
- We demonstrate that DiffEx offers more comprehensible and detailed explanations for classifiers in applications ranging from facial recognition to retinal health compared to prior approaches. Moreover, our method is adaptable and efficiently works with both binary and multi-classifiers.

2. Related Work

Traditional research focuses mainly on heat maps and patch-based extractions to explain classifier decisions. Specifically, class activation and saliency maps attempt to emulate human visual strategies by focusing on the image regions most relevant to a specific class [38, 43, 47, 50, 63, 66, 70]. While these maps highlight the object or part of an image that most influences the classifier’s decision, they fail to reveal which specific, fine-grained object attributes (e.g., color, pattern, or texture) impact the classification. Other approaches try to explain classifier outputs by analyzing extracted image patches [14, 64]; however, these methods can only reveal spatially localized attributes.

Previous studies have leveraged generative models such as variational autoencoders (VAEs) [17] or GANs [15, 42, 48, 49]. The most similar research to ours, StylEx [26], introduces a GAN-based approach to identify various attributes that influence a classifier’s decisions. However, this method has several limitations. Firstly, it requires training a new GAN for each classifier, which can be resource-intensive and time-consuming. Secondly, each identified attribute requires manual labeling, which can require domain expertise. Lastly, StylEx only uncovers a limited range of semantic attributes, potentially overlooking others that might significantly affect a classifier’s scores. Additionally, a notable limitation of the StylEx method and similar GAN models is their focus on single concepts, such as *human* or *animal* faces, or individual objects like *leaves*, rather than entire scenes. In contrast, diffusion models, known for their robust capability to generate complex and detailed entire scenes, offer a significant advantage. Our approach leverages the power of diffusion models to cover a wider spectrum of visual contexts, and enhances the versatility and applicability of our method across various classifiers that assess not only individual elements but also the interaction and composition of entire scenes.

Recent approaches have begun using diffusion models to generate counterfactual examples. One method utilizes shortcut learning to generate counterfactual images but fails to make semantically meaningful edits for certain attributes [59]. Another study explores modifying the diffusion process via adaptive parametrization and cone regularization to produce realistic counterfactual images; however, this approach depends on a robust model, which can be difficult to train [3]. [22] explored counterfactual image generation, however their approach is computationally demanding and uses DDPM [19] models trained on single domains. As a result, it does not take advantage of large-scale latent diffusion models like Stable Diffusion, which can han-

dle more complex scenes. Furthermore, even though some studies have leveraged diffusion models to generate realistic counterfactuals in high-stakes domains [21], such as the medical field, they use a less efficient base model for the image generation process and focus predominantly on single-attribute modifications. Furthermore, to the best of our knowledge, there is no existing research that explores a hierarchical explanation of classifiers. Such an approach would systematically unpack the layers of influence that different semantic levels have on a classifier’s scoring mechanism. This gap highlights a significant opportunity to enhance understanding by detailing how various semantic categories and their subtypes contribute to the decisions made by classifiers.

3. Methodology

We first discuss the background on identifying attributes that influence classifier decisions through changes in logits. Following this, we introduce our method, which includes curating hierarchical attributes using vision-language models and a novel algorithm inspired by beam search to pinpoint attributes that affect classifier scores. Our pipeline is detailed in Fig. 3.

3.1. Background

StylEx [26] identifies semantics that meaningfully influence classifier decisions by ranking each attribute based on its impact on the classifier’s logit outputs. This ranking process aims to identify and select attributes that best explain the classifier’s behavior in a given context, such as understanding the factors influencing *age* classification.

Given a semantic corpus \mathcal{S} and a set of N images to analyze the influence of various semantic attributes on classifier decisions, counterfactual images are first generated. For each original image x_i , an edited version $x'_i = g(x_i, s)$ is generated by applying a semantic attribute $s \in \mathcal{S}$ through a transformation function g . This transformation highlights the effect of each semantic attribute on a classifier’s output. Specifically, the logit difference for each attribute is computed to measure how the classifier’s score changes due to the presence of s .

The influence of each attribute s is quantified as the average logit difference between the original and edited images across a set of sample images, defined as:

$$I(s) = \frac{1}{N} \sum_{i=1}^N |f(x'_i, y) - f(x_i, y)| \quad (1)$$

where $f(x, y)$ denotes the classifier’s logit score for target class y on image x . Here, y represents the specific target class aimed to be explained, such as “age” or “gender.” This influence score $I(s)$ captures the aver-

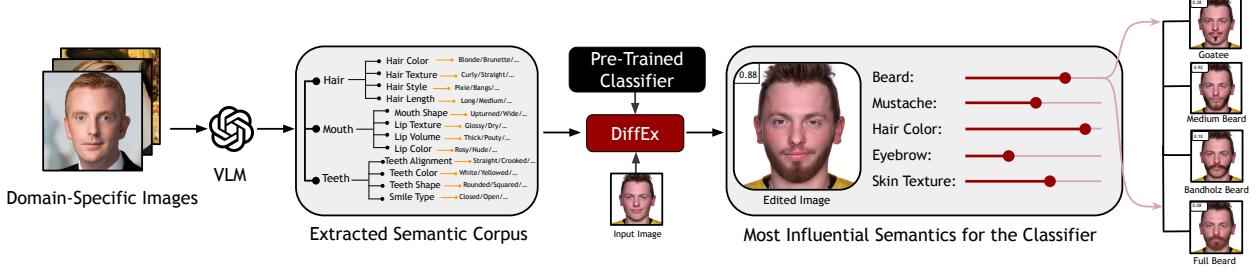


Figure 3. **An Overview of DiffEx.** Our pipeline processes a set of sample domain-specific images and a text prompt using a VLM to generate a hierarchical semantic corpus of attributes relevant to a specific domain. Based on this corpus, DiffEx identifies and ranks the most influential features affecting the classifier’s decisions, sorting them from most to least impactful (rightmost image). The hierarchical explanation of semantics (such as beard and its subtypes) provides a fine-grained understanding of which features drive classifier outputs.

age impact of each semantic attribute on classifier decisions by reflecting the logit score change due to attribute manipulation.

3.2. Our Method

We first employ Vision-Language Models (VLMs) [31] to extract a large-scale corpus covering domain-specific hierarchical semantics. Then, we introduce a training-free method, DiffEx, which leverages this hierarchical corpus and text-to-image diffusion models to explain the decision-making process of classifiers by identifying the most influential semantics.

3.2.1 VLM-Based Semantic Space

While the StyLex method outlined in Section 3.1 employs a logit-based approach to identify semantics, it requires manual labeling of each attribute extracted from the trained GAN model. Moreover, this method does not support the explanation of hierarchical attributes. Therefore, we first compile a large-scale set of semantics using VLMs. Given a domain d , such as the ‘facial domain,’ our objective is to identify a comprehensive set of domain-specific attributes from a collection of domain-specific images, denoted as N_d . To accomplish this, we utilize a Vision-Language Model (VLM) [31] to extract a range of relevant features, represented by \mathcal{H} . Employing in-context learning [8, 52], we prompt the VLM with a small set of images N_d , along with a detailed task description and examples of desired outputs. This process allows us to generate a substantial semantic corpus of keywords, \mathcal{H} , that captures the fine-grained attributes relevant to the domain. The resulting corpus, \mathcal{H} , comprises a comprehensive collection of keywords and phrases that covers the full spectrum of the domain’s attributes. We leverage this rich dataset to provide a hierarchical explanation of classifier decisions, systematically explaining how different attributes influence out-

Algorithm 1 DiffEx

Require: Hierarchical structure \mathcal{H} with semantic groups and features, beam width B , classifier or scoring function f , scoring threshold δ

Ensure: Optimal semantics maximizing f

- 1: Initialize $S \leftarrow$ root-level groups in \mathcal{H} {Initial candidate set at top-level groups}
- 2: Initialize beam $\mathcal{B} \leftarrow \emptyset$
- 3: Score each candidate $s \in S$ using the scoring function $f(s)$
- 4: Select top B candidates with $f(b) \geq \delta$ and store in beam \mathcal{B} {Apply thresholding to filter relevant candidates}
- 5: **while** $S \neq \emptyset$ **do**
- 6: Initialize $S_{\text{next}} \leftarrow \emptyset$
- 7: **for** each candidate $b \in \mathcal{B}$ **do**
- 8: Expand b by adding sub-features from its next level in \mathcal{H} to form new candidates
- 9: **for** each new combination b' generated from b **do**
- 10: **if** $f(b') > f(b)$ **then**
- 11: Add b' to S_{next}
- 12: **end if**
- 13: **end for**
- 14: **end for**
- 15: Set $S \leftarrow S_{\text{next}}$ {Move to next level in hierarchy}
- 16: **end while**
- 17: Return highest-scoring combination from final \mathcal{B} as the optimal joint semantic combination

comes.

3.2.2 DiffEx

Considering the extensive number of semantics identified using a VLM, it is computationally expensive to evaluate every possible semantic or combination of at-

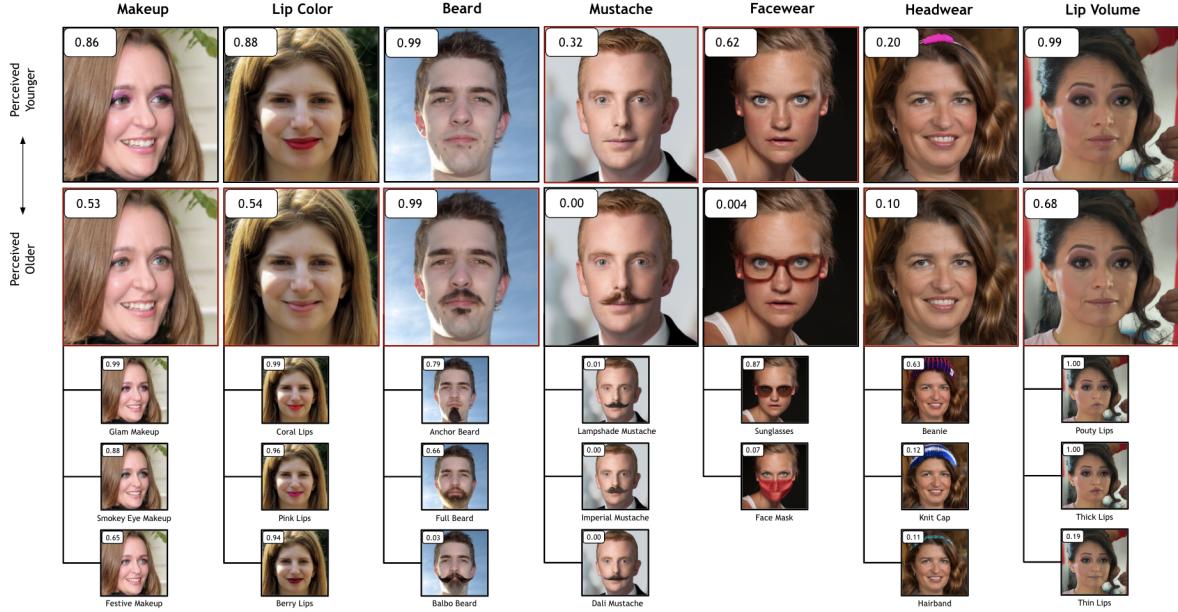


Figure 4. Top-7 Discovered Facial Attributes for the Age Classifier. DiffEx identifies key attributes and their top hierarchical subtypes for a perceived age classifier in the facial domain. For each attribute, the edited images and their respective subtypes are displayed in a hierarchical structure, outlined with a black border. The score for the “young” label is shown in the top-left corner of each image.

tributes. We introduce DiffEx, an efficient approach inspired by beam search to explain classifier decisions (see Algorithm 1). This method leverages our hierarchical semantic corpus to streamline the process. Our approach iteratively refines candidate attributes by expanding only the most impactful semantic paths at each hierarchical level, with each path’s relevance guided by a scoring function that assesses the classifier’s response to each semantic feature. Formally, this scoring function calculates the average classifier C score across N sample images generated using each semantic attribute s , as defined in Eq. (2), where g represents the generative diffusion model applied to introduce the attribute in each sample image:

$$f(s) = \frac{1}{N} \sum_{i=1}^N C(g(x_i, s)) \quad (2)$$

We begin with an initial candidate set S , which includes high-level groups from the semantic hierarchy (e.g., broader categories like ‘mouth features’ or ‘eye features’). Each candidate in S is evaluated by the scoring function $f(s)$, which quantifies the influence of the candidate on the classifier’s output. Only the top B candidates that surpass a predefined score threshold δ are retained, setting a beam width that limits the search to the most impactful candidates. For each candidate in the beam, the algorithm proceeds by expanding to the

next hierarchical level, incorporating more specific sub-features (e.g., for “mouth features,” sub-features such as “beard” and “mustache” are included). For each expanded candidate b' , the scoring function in Eq. (2) is re-evaluated. Only candidates with a score exceeding that of their parent $f(b') > f(b)$ are retained in the next candidate set S_{next} , ensuring that only those additions that yield significant incremental impact are added. This process of expansion, scoring, and filtering continues iteratively, moving from general to more specific semantic attributes at each hierarchical level. By dynamically adjusting the candidate set based on the beam width B and incremental scoring threshold δ , our method remains computationally efficient, focusing on high-impact combinations rather than exhaustively evaluating all possible attribute pairings. For generating counterfactual images, we utilize an off-the-shelf editing tool for diffusion models, Ledit++ [7]. However, our approach is versatile enough to accommodate any editing method, allowing for flexibility in application and integration with different tools.

4. Experiments

4.1. Experimental Setup

Our experiments utilize Ledit++ [7] and Stable Diffusion XL (SDXL) [35] for generating counterfactual images. Twenty-five time steps are omitted to boost com-

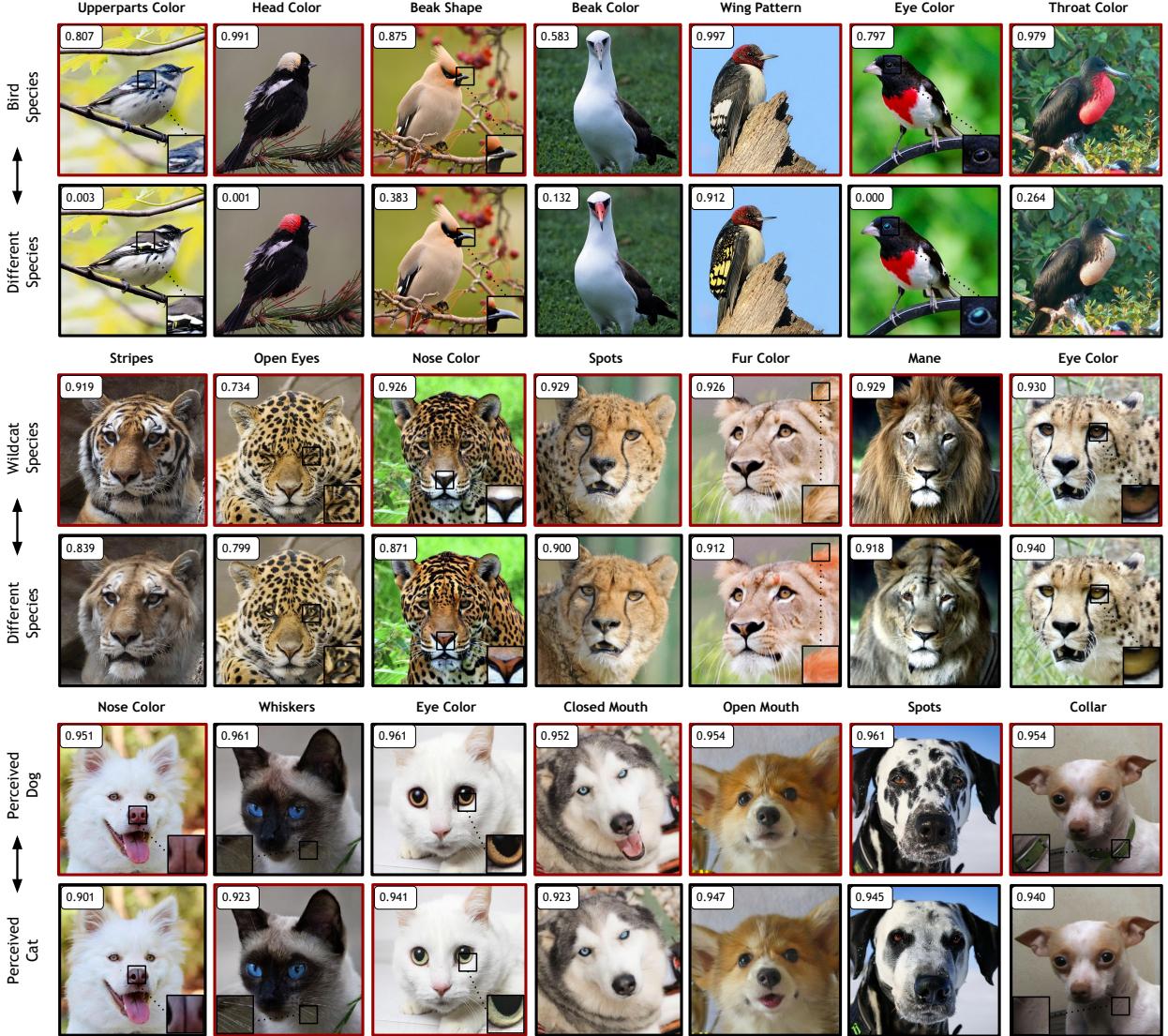


Figure 5. Top-7 Discovered Attributes Across Different Animal Domains. Our method successfully identifies key attributes for multiple domains, such as bird, wildcat, and pet species. The original images are depicted with red borders while the edited images are depicted with black borders. For the pet species domain, we used a binary classifier and for the bird and wildcat species domains, we used a multi-classifier. The impact of each attribute on the classifier’s score is shown in the top-left corner of each image. For attributes that caused subtle changes, we provided a zoomed-in view of the edit, displayed in the bottom left or right corner of the image.

putational efficiency while preserving the quality of edits. The edit threshold, a hyperparameter that dictates the global application scope of edits, is adjusted based on each domain. We test our method across diverse classifiers to evaluate its effectiveness in explaining model behavior through semantic influence. Specifically, we utilize classifiers trained on facial attributes data (i.e. age and gender classifiers) [23] as well plant health [20], retinal disease [11], bird species [57], wildcat/pet species [10], fashion [30], places [71], and food data [5]. The experiments for the face and plant health domains

utilize a CLIP-based classifier [36] to evaluate and interpret edits, while experiments within the bird, wildcat, pet, food, fashion, and places domains use CNN-based classifiers. These CNN classifiers were built on the EfficientNet [55] architecture and achieved an accuracy of over 95 percent on their test sets. For the retinal disease domain, we utilized the FLAIR model [45], which is based on a pre-trained vision-language model, to classify various retina scans.

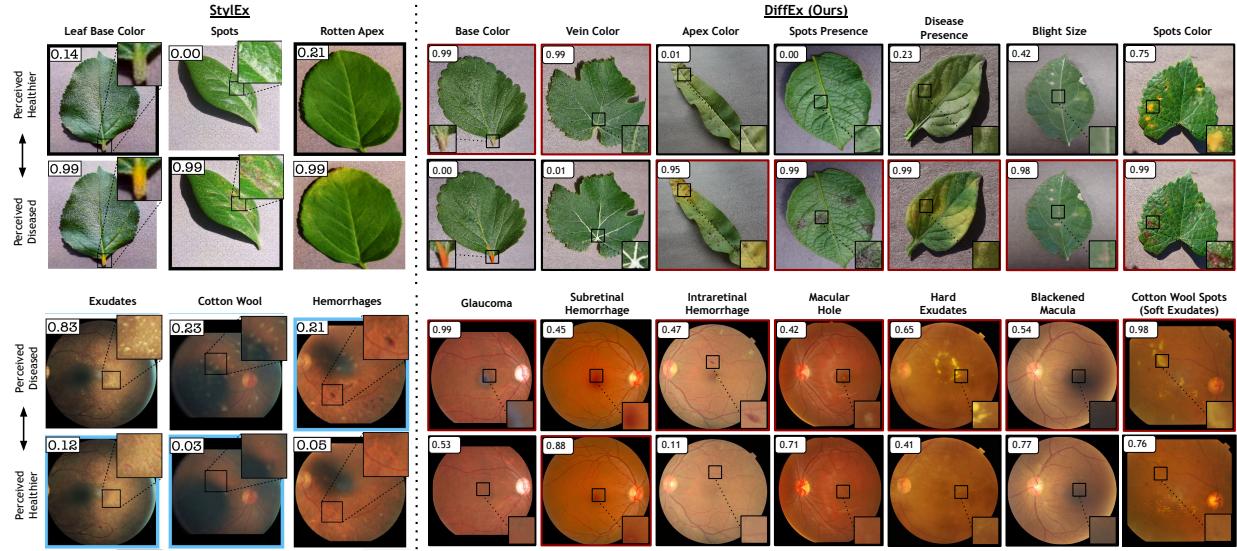


Figure 6. Visual Comparison of Key Attributes Identified by StylEx and DiffEx in the Plant Health and Retinal Disease Domains. This figure illustrates the enhanced capability of our method (DiffEx) in identifying a broader set of significant attributes compared to StylEx within the plant health and retinal disease domains. DiffEx successfully uncovers more detailed and diagnostically relevant features, such as “leaf vein color” and “macular hole,” which provide deeper insights into leaf and retina health. In contrast, StylEx primarily identifies general attributes like “leaf base color” and “exudates.” For DiffEx, images with black borders represent the counterfactual images, while those with red borders represent the original images. For StylEx, images with blue or black borders are counterfactuals. For a comprehensive comparison of the top attributes discovered by StylEx and DiffEx across various domains, please refer to Table 1.

4.2. Qualitative Experiments

4.2.1 Explaining Classifiers with Single Attributes

In the facial domain, we used a classifier to analyze fine-grained semantic features influencing the person’s perceived age, while in other domains, we examined classifiers for species identification, leaf health assessment, retinal disease detection, food categorization, place perception, and clothing type.

Face Domain: In the facial domain, as illustrated in Fig. 4, DiffEx identifies and ranks key attributes impacting age classification. For example, features such as “makeup styles,” “lip volume,” and “accessories” (ex.“hairbands”) are associated with perceived youthfulness, whereas attributes like “facial hair” and “eyeglasses” are linked with perceived older age. Notably, the eyeglasses attribute consistently reduces the classifier’s score for the “young” label, reflecting its association with older demographics. Additionally, DiffEx uncovers hierarchical attribute structures, demonstrating how subcategories within a feature can have varying effects on classifier outcomes. For instance, as shown in Fig. 4, different beard styles impact the perceived age differently (e.g. a “Balbo beard” significantly increases the age classification score more than a “full beard”). Additional examples of hierarchical explanations of age classifiers are detailed in the appendix (S7).

Animal Domain: Furthermore, DiffEx explains classifiers in a variety of animal types. Fig. 5 highlights the top-7 most influential attributes in the bird, wildcat, and pet domains where our method was able to identify fine-grained semantics to explain classifier behavior. For example, in the Wildcat classifier, attributes like *stripes* and *manes* are critical in distinguishing wildcats from other species. In contrast, in the Cat/Dog classifier, features such as *whiskers*, *mouth position (open or closed)*, and *spots* are among the key differentiators. Explaining the workings of these classifiers also reveals potential biases. For instance, the presence of a *collar* significantly increases the likelihood of an animal being classified as a dog. This bias may stem from the training dataset where images of dogs more frequently featured collars compared to those of cats.

Medical and Plant Health Domains: Fig. 6 demonstrates the results for retinal disease and plant health domains. For the retinal disease domain, we use the FLAIR model to classify images of diseased and healthy retinal fundus scans. This model utilizes detailed domain expert knowledge descriptions, such as “no hemorrhages, microaneurysms, or exudates” compared to general descriptions like “no diabetic retinopathy” to aid in its classification. For the plant health domain, the CLIP classifier we use looks at features such as the presence of spots, fungus, or discoloration to make its decision.

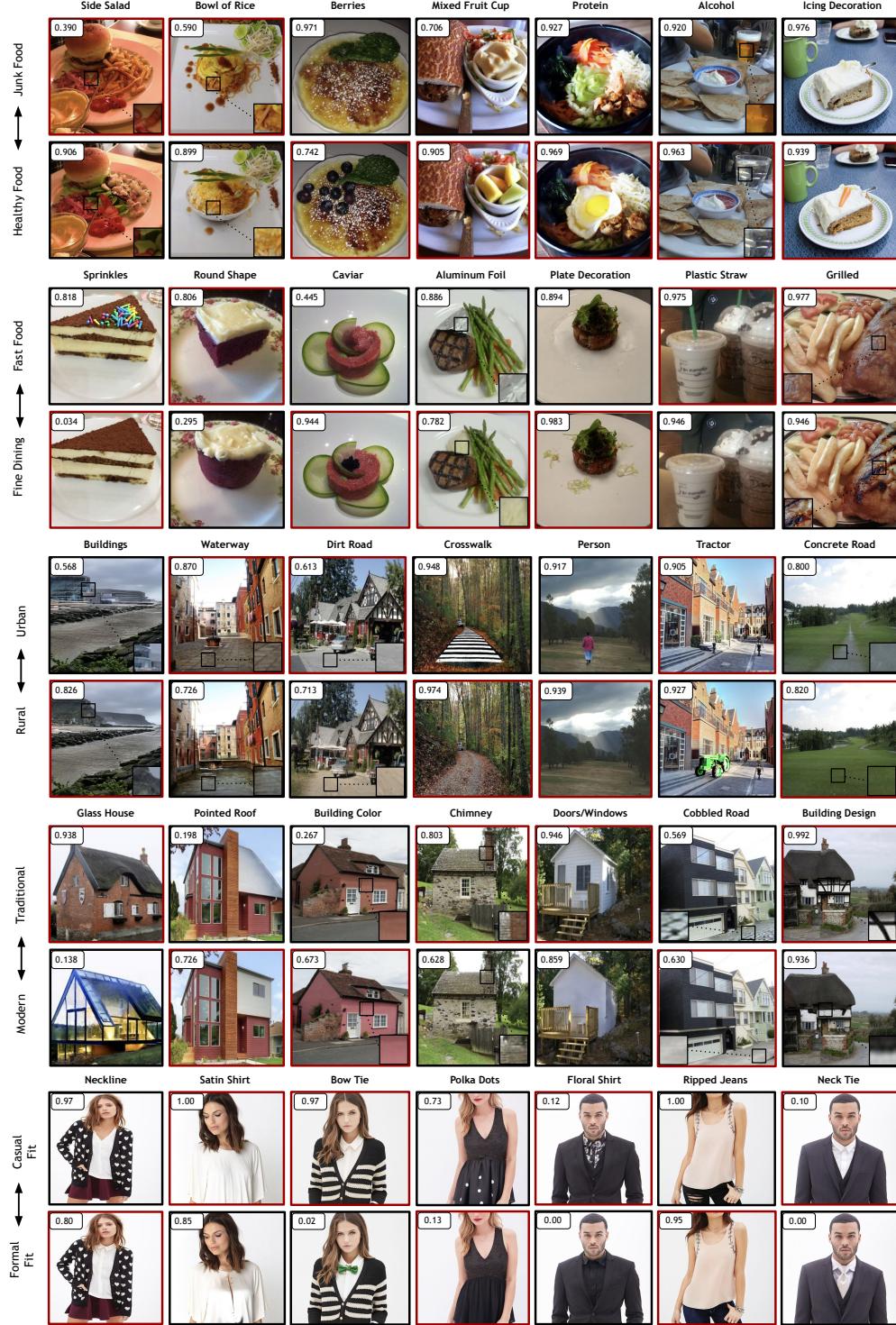


Figure 7. Top-7 Attributes Discovered by DiffEx for Multi-Object Domains involving Complex Scenes. Compared to existing methods, DiffEx was used to identify meaningful attributes for multi-object domains, such as “food,” “places,” and “fashion.” By identifying semantic features that make food appear “healthier” or part of “fine dining,” as well as attributes that give a place a more “rural” or “modern” feel, DiffEx can serve as a powerful tool for understanding and modifying perceptions through targeted edits. As with the other examples in this paper, images with black borders indicate the original, unedited versions, while those with red borders represent the edited versions. For attributes with subtle changes, a zoomed-in view of the modification is shown in the image’s bottom left or right corner.

Multi-Object Domains involving Complex Scenes:

Unlike traditional GAN-based methods that primarily focus on single-object scenarios like a cropped face, DiffEx extends its utility by providing a list of relevant features for domains encompassing multiple objects, such as places, food, and fashion. Classifiers within these domains often evaluate multiple elements simultaneously. For example, a fashion classifier might determine the ‘Formal or Casual Fit’ of an outfit by considering various components such as hairstyle, top, and bottom. Similarly, an architectural classifier assessing whether a building appears ‘Urban or Rural’ may base its evaluation on not just the structure itself but also its exteriors and surroundings. Explaining the decision-making process of such classifiers is crucial, as they analyze multiple objects within a single image, adding complexity to their interpretative frameworks. Fig. 7 illustrates the top-7 attributes identified by DiffEx for the food, place, and fashion domains. Specifically, for the places domain, DiffEx identifies the top-7 attributes for the “urban” vs. “rural” and “modern” vs. “traditional” classifiers, while for the food domain, it does so for the “healthy” vs. “junk food” and “fine dining” vs. “fast food” classifiers. Fig. 7 also demonstrates the top-7 attributes for “casual” vs. “formal” apparel classifier discovered by our method.

DiffEx is able to uncover interesting observations across various domains. For example, for the food domain, DiffEx was able to discover that removing “caviar” or a “plate decoration” from the image made the food appear to be more perceived as fast food compared to fine dining. For the places domain, DiffEx was able to find that adding a “tractor” to an image or a “dirt road” made the place seem more rural than urban. In the fashion domain, the style of a neckline significantly influences whether an outfit is classified as “formal” or “casual.” For instance, a V-neckline is often associated with a more casual look, whereas a classic boat neckline is generally perceived as more formal. Since DiffEx leverages text-to-image diffusion models, it can handle classifiers that interpret complex scenes involving multiple objects and provide detailed explanations for their decisions. This capability is essential for advancing understanding classifier decisions in diverse domains.

Explaining Classifiers with Joint Attributes Our method goes beyond analyzing individual attributes by identifying attribute combinations that collectively improve classifier interpretation. While single attributes may have a minimal impact on logits, their combined effect can substantially influence the classifier’s output, uncovering subtle interactions that shape decision-making. For example, as shown in Fig. 8, individual changes in “lip color” or “eye makeup” result in min-

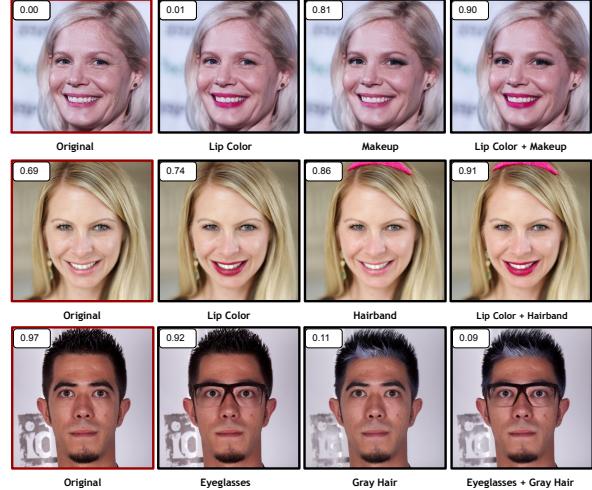


Figure 8. Impact of Joint Attributes on Classifier Decisions. DiffEx effectively explains classifier decisions by analyzing the combined influence of joint attributes. The score indicating the perceived age (“younger” label) is displayed in the top-left corner of each image, highlighting how specific attribute pairings can amplify or alter classifier outputs.

imal score changes, but when modified together, they make the subject appear significantly younger. Similarly, adding a “hairband” or changing the “lip color” alone has little impact, but their combination shifts the classifier score markedly toward a younger age. On the other hand, pairing “gray hair” with “eyeglasses” strongly increases the perceived age, more so than each feature individually. This analysis highlights how classifiers may respond more robustly to specific attribute combinations, providing deeper insights into feature interactions. Additional examples of joint attributes for the age, bird species, and plant health classifiers are provided in the appendix (S4, S5, S6).

Extending DiffEx for Multi-Classifier Analysis We adapt DiffEx for multi-classifier applications to uncover semantic attributes essential for tasks like retinal disease classification (Fig. 6) and bird and wildcat species identification (Fig. 5). This approach highlights key features such as “upperparts color” and “beak shape” for bird species, “stripes” and “nose color” for wildcat species, and types of hemorrhages and exudates for retinal conditions. The results in Fig. 6 for the retinal diseases domain and Fig. 5 for the bird/wildcat species domain demonstrate that modifying these attributes significantly alters classifier scores, impacting assessments for species identification and disease likelihood.

Face (Age)		Bird (Species)		Leaves (Health)		Retina Scans (Disease)		Wildcat (Species)		Pet (Cat/Dog)	
StylEx	Ours	StylEx	Ours	StylEx	Ours	StylEx	Ours	StylEx	Ours	StylEx	Ours
Skin Pigmentation		Eyebrow	Belly Color	Upperparts Color	Base Leaf Color	Base Color	Exudates	Glaucoma	Stripes	Open Mouth	Nose Color
Eyebrow Thickness		Makeup	Upperparts Color	Head Color	Apex Color	Vein Color	Cotton Wool Spots	Subretinal Hemorrhage	Closed Mouth	Whiskers	
Eyeglasses		Mustache Type	Wing Pattern	Beak Shape	Spots	Apex Color	Hemorrhages	Intraretinal Hemorrhage	Open Eyes	Eye Shape	Eye Color
Hair Color		Teeth	Beak Color	Beak Color	Blight	Spots Presence	Clustered Exudates	Macular Hole	Nose Color	Dropped Ears	Closed Mouth
Lip Thickness + Position		Lip Volume	Head Color	Wing Pattern	Halos	Disease Presence	x	Hard Exudates	Spots	Pointed Ears	Open Mouth
Bangs		Lip Color	Breast Color	Eye Color	x	Bright Size	x	Blackened Macula	Fur Color	Tongue	Spots
Eye Makeup		Eyelash	x	Throat Color	x	Leaf Texture	x	Soft Exudates	Mane	x	Collar
Facial Hair Color		Beard Type	x	Wing Color	x	Spots Color	x	Retinal Drusen	Eye Color	x	Pointed Ears
x		Facewear	x	Crest Presence	x	Discoloration	x	Optic Disc Hemorrhage	Tongue	x	Mouth Color
x		Headwear	x	Feather Texture	x	Leaf Orientation	x	Cataract	Pupil Size	x	Fur Pattern

Table 1. **Comparison of Top Attributes Across Different Domains and Classifiers.** The table above contains a list of the top attributes discovered by DiffEx (Ours) vs. StylEx. The **X** in the table indicates attributes that were not mentioned in StylEx. It is also important to note that “cotton wool spots” and “soft exudates” refer to the same condition within the retinal disease domain.

Method	Crest Presence	Beak Shape	Throat Color	Feather Texture	Eye Color	Beak Color	Head Color	Upperparts Color	Avg. Correct Response
Grad-CAM	36%	50%	56%	35%	47%	65%	59%	76%	53%
StylEx	68%	85%	79%	82%	74%	68%	91%	65%	76.5%
DiffEx (Ours)	88%	91%	88%	91%	82%	82%	97%	88%	88.4%

Table 2. **Comparison with Other Explainability Methods.** The table above displays the percentage of correct attribute selections for the bird class, as chosen by users when viewing outputs from different explainability methods. It also includes the average percentage of correct responses across all attributes for each method. As shown, for each attribute presented, the majority of users identified the correct attribute when viewing the output generated by DiffEx.

Rating	Bird Domain	Face Domain
Edit Quality	3.386 ± 0.223	3.659 ± 0.248
Disentanglement	3.163 ± 0.197	3.204 ± 0.213

Table 3. **Edit Quality and Disentanglement Ratings.** The table above provides the average edit quality and faithfulness ratings across different domains from User Study 1. The scoring is done on a scale from 1 to 5.

4.2.2 Qualitative Comparison

In Fig. 6, we present a visual comparison of the top attributes and classification scores identified by DiffEx against those identified by StylEx for the plant health and retinal disease domains. As illustrated, DiffEx successfully uncovers a broader set of semantically meaningful attributes compared to StylEx. For instance, DiffEx identifies detailed attributes such as “leaf vein color” and “spots color,” in addition to the more general attributes found by StylEx, like “leaf base color” and “apex color.” This expanded set of attributes enables a more detailed understanding of key diagnostic features, especially relevant in applications such as plant health assessment. Table 1 provides a comprehensive comparison of the top features identified by StylEx and our method, highlighting DiffEx’s superior ability to uncover semantics that are crucial for the classifier’s decision-making process. In particular, the table presents an extended list of features across domains such as faces, bird species, plant health, retina scans, wildcat species, and pet types, all of which influence a classifier’s score—further emphasizing DiffEx’s ability to uncover fine-grained and contextually rich features. This highlights DiffEx’s advantage in providing a more comprehensive understanding of classifier behavior by cap-

turing both general and specific feature variations within a given category.

In addition to identifying common features found by StylEx, such as “eyebrow thickness,” “makeup,” and “facial hair,” DiffEx demonstrates a broader and more generalized ability to detect relevant features in the facial domain. Unlike StylEx, which tends to focus on specific attribute variations, such as “facial hair color,” DiffEx captures a wider range of feature types and their detailed effects. For example, DiffEx is capable of distinguishing between different beard shapes and understanding how each individual style influences the classifier’s decision, as shown in Fig. 4. While StylEx may highlight “facial hair color” as a key feature, DiffEx’s approach covers the diversity within the category, such as differentiating between a “full beard,” “Balbo beard,” and “anchor beard,” and analyzing their respective impacts on perceived age.

4.3 Quantitative Experiments

Baselines To quantitatively evaluate the effectiveness of DiffEx compared to other explainability methods, such as StylEx [26] and Grad-CAM [43], we conducted a series of comprehensive user studies to assess how easily participants could identify the relevant features extracted by our method.

User Study 1: Visual Quality and Disentanglement

To evaluate the visual quality and disentanglement of the edited images for the face and bird domains, we conducted a user study with 50 participants on Prolific*. Specifically, for each domain, we showed pairs of unedited and edited images for ten attributes and asked the users to assess whether the edited image contained

*Prolific, <https://www.prolific.com>

the desired attribute and if the edit appeared to be disentangled. For each pair of images, we asked the users to rate the edit and disentanglement from one to five, with five representing the highest score. Our results indicate that our edits successfully reflected the intended attributes while minimizing any unrelated changes (see Table 3). Please refer to Appendix for more details about our user study.

User Study 2: Comparison with Grad-CAM and StylEx To evaluate how effectively our method, DiffEx, explains various semantics within a specific domain compared to other explainability methods (Grad-CAM and StylEx), we conducted a user study with 35 participants on Prolific. We focused on images from the bird domain, generating three sets of three images per attribute: one original image, one edited image, and one image illustrating the explainability method. For Grad-CAM, participants were shown a heatmap overlay on the edited image. For StylEx, we displayed the edited image alone, as this method requires users to manually label the edits. For DiffEx, participants viewed the edited image along with the attribute automatically assigned by the VLM. We then asked users to select the attribute (e.g., “beak color,” “crest presence”) that best explained the edited image from four answer choices. Results from this study indicate that DiffEx significantly outperforms Grad-CAM and StylEx in explaining edited attributes in images (see Table 2). Please refer to Appendix for more details about our user study.

5. Limitation

While our approach offers significant insights into classifier behavior through semantic edits, there are a few limitations to consider. First, since our method relies on semantics curated through VLMs, it is constrained by the quality and scope of the initial semantic corpus. This corpus may not fully capture all relevant or nuanced features, especially in specialized domains where unique or context-specific attributes are necessary for accurate interpretation. Another limitation is that since we are utilizing an off-the-shelf image editing model, such as Ledit+++, to target specific attributes, some edits may be entangled (a general issue in image editing algorithms [62]), and might introduce confounding factors that could affect classifier scores. This is particularly important in high-stakes domains, such as medical imaging, where even minor unintended changes may impact interpretability. Nevertheless, our framework is flexible and can be improved with domain-specific semantic adjustments such as task-specific RAGs [61] or other editing methods [6, 12].

6. Conclusion

In this work, we introduce DiffEx, a novel approach for explaining classifier decisions by utilizing semantic edits within diffusion models. By harnessing the power of vision language models, we curate a comprehensive, hierarchical semantic corpus across various domains and propose a novel algorithm inspired by beam search to filter and rank the most impactful features. DiffEx ranks these semantic features based on their influence on classifier logits, capturing both individual and joint attribute effects, which are crucial for understanding complex classifier behaviors. Through experiments conducted on a wide range of domains—including face, bird, and medical classifiers—we showcase the robustness and adaptability of our approach. This work provides a powerful tool for interpreting model decisions across diverse applications, promoting transparency, and building trust in AI-driven classification systems.

References

- [1] Mostofa Ahsan, Rahul Gomes, Md. Minhaz Chowdhury, and Kendall E. Nygard. Enhancing machine learning prediction in cybersecurity using dynamic feature selector. *Journal of Cybersecurity and Privacy*, 1(1):199–218, 2021. 1
- [2] Leila Amgoud. Explaining black-box classifiers: Properties and functions. *International Journal of Approximate Reasoning*, 155:40–65, 2023. 1
- [3] Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022. 3
- [4] Ouahiba Azouaoui and Amine Chohra. [no title found]. *Applied Intelligence*, 16(3):249–272, 2002. 1
- [5] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101-mining discriminative components with random forests. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6–12, 2014, proceedings, part VI 13*, pages 446–461. Springer, 2014. 6
- [6] Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 2, 11
- [7] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledit+++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024. 2, 5
- [8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Ad-*

- vances in neural information processing systems*, 33: 1877–1901, 2020. 4
- [9] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This looks like that: deep learning for interpretable image recognition. *Advances in neural information processing systems*, 32, 2019. 2
- [10] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. *CoRR*, abs/1912.01865, 2019. 6
- [11] Jorge Cuadros and George Bresnick. Eyepacs: An adaptable telemedicine system for diabetic retinopathy screening. *Journal of Diabetes Science and Technology*, 3(3): 509–516, 2009. 6
- [12] Yusuf Dalva and Pinar Yanardag. Noiseclr: A contrastive learning approach for unsupervised discovery of interpretable directions in diffusion models. *arXiv preprint arXiv:2312.05390*, 2023. 11
- [13] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 1
- [14] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in neural information processing systems*, 32, 2019. 3
- [15] Lore Goetschalckx, Alex Andonian, Aude Oliva, and Phillip Isola. Ganalyze: Toward visual definitions of cognitive image properties. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5744–5753, 2019. 3
- [16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2
- [17] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace). *arXiv preprint arXiv:1907.07165*, 2019. 3
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. (*arXiv:1512.03385*), 2015. *arXiv:1512.03385*. 1
- [19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3
- [20] David P. Hughes and Marcel Salathe. An open access repository of images on plant health to enable the development of mobile disease diagnostics. (*arXiv:1511.08060*), 2016. *arXiv:1511.08060* [cs]. 6
- [21] Indu Ilanchezian, Valentyn Boreiko, Laura Kühlewein, Ziwei Huang, Murat Seçkin Ayhan, Matthias Hein, Lisa Koch, and Philipp Berens. Generating realistic counterfactuals for retinal fundus and oct images using diffusion models. *arXiv preprint arXiv:2311.11629*, 2023. 3
- [22] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pages 858–876, 2022. 3
- [23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 4396–4405, Long Beach, CA, USA, 2019. IEEE. 6
- [24] Ziv Katzir and Yuval Elovici. Quantifying the resilience of machine learning classifiers used for cyber security. *Expert Systems with Applications*, 92:419–429, 2018. 1
- [25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1
- [26] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: Training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 2, 3, 10, 1
- [27] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 1
- [28] Joffrey L. Leevey, John Hancock, Richard Zuech, and Taghi M. Khoshgoftaar. Detecting cybersecurity attacks across different network features and learners. *Journal of Big Data*, 8(1):38, 2021. 1
- [29] Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond. *Knowledge and Information Systems*, 64(12):3197–3234, 2022. 2
- [30] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaozhou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6
- [31] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya. Gpt-4 technical report, 2024. 2, 4
- [32] Akin Ozciit and Arif Gulten. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer Methods and Programs in Biomedicine*, 104(3): 443–451, 2011. 1
- [33] Yi Peng, Guoxun Wang, Gang Kou, and Yong Shi. An empirical study of classification algorithm evaluation for financial risk prediction. *Applied Soft Computing*, 11(2): 2906–2915, 2011. 1
- [34] Emmanuel Pintelas, Meletis Liaskos, Ioannis E Livieris, Sotiris Kotsiantis, and Panagiotis Pintelas. Explainable machine learning framework for image classification problems: case study on glioma cancer prediction. *Journal of imaging*, 6(6):37, 2020. 1
- [35] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 5
- [36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolu-

- tional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 6
- [37] R. Joshua Samuel Raj, S. Jeya Shobana, Irina Valeryevna Pustokhina, Denis Alexandrovich Pustokhin, Deepak Gupta, and K. Shankar. Optimal feature selection-based medical image classification using deep learning model in internet of medical things. *IEEE Access*, 8:58006–58017, 2020. 1
- [38] Sylvestre-Alvise Rebuffi, Ruth Fong, Xu Ji, and Andrea Vedaldi. There and back again: Revisiting backpropagation saliency methods. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8839–8848, 2020. 3
- [39] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery. 1
- [40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2
- [41] Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022. 2
- [42] Axel Sauer and Andreas Geiger. Counterfactual generative networks. (arXiv:2101.06046), 2021. arXiv:2101.06046 [cs]. 3
- [43] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 3, 10
- [44] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, page 5103–5111. AAAI Press, 2018. 1
- [45] Julio Silva-Rodriguez, Hadi Chakor, Riadh Kobbi, Jose Dolz, and Ismail Ben Ayed. A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. *Medical Image Analysis*, 99: 103357, 2025. 6
- [46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. 2014. 1
- [47] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 3
- [48] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. Explanation by progressive exaggeration. (arXiv:1911.00483), 2020. arXiv:1911.00483 [cs]. 3
- [49] Sumedha Singla, Motahare Eslami, Brian Pollack, Stephen Wallace, and Kayhan Batmanghelich. Explaining the black-box smoothly- a counterfactual approach. (arXiv:2101.04230), 2022. arXiv:2101.04230 [cs, eess]. 3
- [50] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019. 3
- [51] Gregor Stiglic, Primož Kocbek, Nino Fijacko, Marinka Zitnik, Katrien Verbert, and Leona Cilar. Interpretability of machine learning-based prediction models in healthcare. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 10(5):e1379, 2020. 2
- [52] Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A Smith, et al. Selective annotation makes language models better few-shot learners. *arXiv preprint arXiv:2209.01975*, 2022. 4
- [53] Prabu Subramani, Khalid Sattar, Rocío De Prado, Balasubramanian Girirajan, and Marcin Wozniak. Multi-classifier feature fusion-based road detection for connected autonomous vehicles. *Applied Sciences*, 11(17): 7984, 2021. 1
- [54] Jie Sun and Hui Li. Financial distress prediction based on serial combination of multiple classifiers. *Expert Systems with Applications*, 36(4):8659–8666, 2009. 1
- [55] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks, 2020. 6
- [56] Manoj Thakur and Deepak Kumar. A hybrid financial trading support system using multi-category classifiers and random forest. *Applied Soft Computing*, 67:337–349, 2018. 1
- [57] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge J. Belongie. The caltech-ucsd birds-200-2011 dataset. 2011. 6
- [58] Wahyono, Laksono Kurnianggoro, Joko Hariyono, and Kang-Hyun Jo. Traffic sign recognition system for autonomous vehicle using cascade svm classifier. In *IECON 2014 - 40th Annual Conference of the IEEE Industrial Electronics Society*, pages 4081–4086, 2014. 1
- [59] Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pages 338–357. Springer, 2025. 3
- [60] M. Wiggins, A. Saad, B. Litt, and G. Vachtsevanos. Evolving a bayesian classifier for ecg-based age classification in medical applications. *Applied Soft Computing*, 8(1):599–608, 2008. 1
- [61] Junde Wu, Jiayuan Zhu, and Yunli Qi. Medical graph rag: Towards safe medical large language model via graph retrieval-augmented generation. *arXiv preprint arXiv:2408.04187*, 2024. 11
- [62] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1900–1910, 2023. 11

- [63] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in scale and space. (arXiv:2004.03383), 2020. arXiv:2004.03383 [cs]. 3
- [64] 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. 3
- [65] Muhammad Bilal Zafar, Michele Donini, Dylan Slack, Cédric Archambeau, Sanjiv Das, and Krishnaram Kenthapadi. On the lack of robust interpretability of neural text classifiers. *arXiv preprint arXiv:2106.04631*, 2021. 1
- [66] MD Zeiler. Visualizing and understanding convolutional networks. In *European conference on computer vision/arXiv*, 2014. 3
- [67] Aston Zhang, Zachary C Lipton, Mu Li, and Alexander J Smola. Dive into deep learning. *arXiv preprint arXiv:2106.11342*, 2021. 1
- [68] Quanshi Zhang, Ying Nian Wu, and Song-Chun Zhu. Interpretable convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1
- [69] Yu Zhang, Peter Tišo, Aleš Leonardis, and Ke Tang. A survey on neural network interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, 2021. 2
- [70] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 3
- [71] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 6

Explaining in Diffusion: Explaining a Classifier Through Hierarchical Semantics with Text-to-Image Diffusion Models

Supplementary Material

S1. Overview

In this appendix, we present further details on our methodology, user studies, and experiments. Specifically, we include the prompt template used with our chosen vision-language model, GPT-4, as well as a more detailed hierarchy of attributes across various domains. We also share further insights from our user studies, including the questions asked and examples of images presented during the evaluations. Finally, we include additional examples of single- and joint-attribute editing and demonstrate how integrating the counterfactual images we generated into the training data of a classifier can improve its accuracy and robustness.

S2. GPT-4 Prompt Template

To extract a list of potential attributes for each domain, we provided the prompt template in Table G to GPT-4. A more detailed example of the text prompt for the face domain can be found in Table H.

S3. Hierarchy of Attributes

Figures 2 and S1 depict the hierarchical structures of various attributes for the bird and retinal disease domains respectively. Tables D, E and F show an extensive list of potential attributes for the face, plant health, and bird domains respectively. Level 1 attributes refer to “broader” categories while Level 2 and Level 3 attributes refer to “finer-grained” categories. It is important to note that the attributes listed in these figures represent only a subset of all the attributes provided by GPT-4. Additionally, Table A features the top-10 attributes for the face, bird, plant health, and retinal disease domains, and their corresponding ranking scores.

S4. Algorithmic Comparison

The big-O algorithmic comparisons between StyleEx [26] and DiffEx are approximated in Table B. Both algorithms perform at $O(n^3)$ in the worst case, where n denotes the number of sample images and s denotes the number of style vectors for the GAN-based approach and the set of semantics in the diffusion-based approach. StyleEx uses a greedy search approach to find the most relevant attribute. However, the average complexity of this algorithm is heavily dependent on the number of sample images and the length of the style vectors. If the number of images n is much smaller than the length of

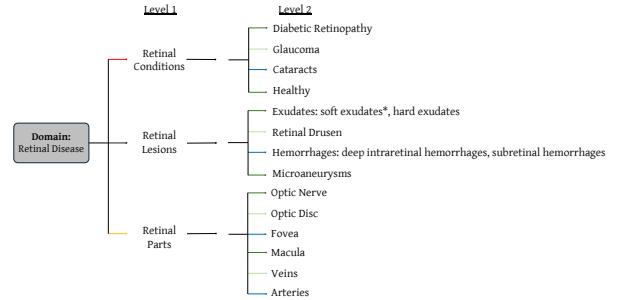


Figure S1. **Hierarchical List of Attributes for the Retinal Disease Domain.** The diagram above illustrates the hierarchical organization of various attributes within the retina scan domain, showing the levels to which they belong. Note: The asterisk next to “soft exudates” denotes that they are also referred to as “cotton wool spots,” and the sub-categories under “exudates” are part of the level 3 hierarchy.

style vector s , ($n \ll s$), then the complexity becomes $O(s^2)$. However, if $n \gg s$, then the time complexity becomes $O(n^2)$. On the other hand, DiffEx is linear with respect to the number of semantics s , the number of sample images n , and the beam width b . For $n \gg s$, DiffEx performs at $O(b \cdot n)$ whereas for $n \ll s$, the complexity becomes $O(b \cdot s)$. Furthermore, the beam search approach strikes a compromise between the efficiency of greedy search and optimality of exhaustive search [67]. Therefore, beam search can produce more optimal outcomes with the help of its beam width logic compared to the greedy method.

S5. Quantitative Evaluation

For our primary quantitative evaluation, we conducted two user studies to assess different aspects of our approach. **User Study 1** focused on evaluating edit quality and disentanglement, while **User Study 2** compared our method, DiffEx, against two explainability techniques: Grad-CAM and StylEx. We chose user studies as the main quantitative assessment because they directly evaluate the human-centric goals of our explainability method. Explainability is ultimately about making AI systems more interpretable and useful for humans, which user studies are well-suited to measure. There is also no universally accepted benchmark for explainability, and by focusing on user studies, we ensure that our evaluation captures real-world factors across diverse domains. Subsections S5.1 and S5.2 include some

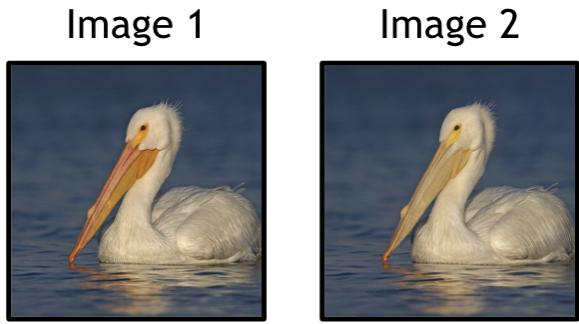


Figure S2. Example of Original and Edited Image Comparison from User Study 1. The image pair above serves as an example from the quantitative study on edit quality and disentanglement, specifically for the “beak color” attribute.

example questions from these user studies.

S5.1. User Study 1: Evaluating Edit Quality and Disentanglement

In this study, participants were shown eight pairs of images per domain (specifically the “face” and “bird” domains). Each pair consisted of an edited image and its corresponding unedited version to highlight the change in a specific attribute. Participants were then asked to evaluate the edits by answering the following questions, which quantitatively assessed the quality and disentanglement of the modifications. For example, for the *beak color* attribute in the bird domain, participants were shown a sample pair of images (as seen in Figure S2) and asked the following questions:

- **Edit Quality:** *Given the original image (image 1) and edited image (image 2), how likely do you think the modified image reflects the intended change (e.g., beak color)?*

Scale: 1 = Not Likely, 5 = Very Likely

- **Disentanglement:** *Given the original image (image 1) and edited image (image 2), how likely do you think the edited image is disentangled compared to the original image? Disentanglement means the modification performed only the desired edit (e.g., modifying the “beak color” without altering unrelated areas).*

Scale: 1 = Not Disentangled, 5 = Fully Disentangled
These questions enabled us to systematically evaluate the effectiveness of the edits in achieving the desired modifications while maintaining disentanglement. The results of the study are included in Table 3 in the Quantitative Experiments section of our paper.

S5.2. User Study 2: Comparisons with Grad-CAM and StylEx

To quantitatively evaluate our method, DiffEx, against other explainability metrics, we conducted another user



Figure S3. Comparison of Original, Edited, and Grad-CAM Images from User Study 2. Image 1 depicts the original image of a bird, Image 2 shows the edited version of the bird, and Image 3 illustrates the Grad-CAM explainability metric, highlighting the most important attribute(s) in the edited image. In this example, “beak shape” was the attribute that was edited (as seen in image 2); however, Grad-CAM highlights both the bird’s wing and beak, making it unclear which attribute is the primary focus of the image.

study. Participants were presented with 3 sets of 3 images per attribute: the original image, an edited (counterfactual) image, and a third image explained using a comparable metric, such as Grad-CAM. For each set, participants were asked the following question, with modified answer choices and corresponding images: “Given three images (image 1, image 2, and image 3), select the attribute that best describes the feature highlighted in image 3.” A sample set of answer choices provided for the question accompanying Figure S3 were:

- Feather Texture
- Beak Shape
- Beak Color
- Eye Color

S6. Additional Experiments

In this section, we present some further experiments that demonstrate the impact of single and joint attributes on the classifier’s output. We also explore how training the classifier on counterfactual examples can enhance its robustness.

S6.1. Experiments with Single Attributes

To effectively illustrate the hierarchical structure of the edited features, additional experimental results are presented in Figure S7, focusing on the facial domain. These results provide a clearer understanding of how specific modifications within different feature categories influence the classifier’s output. For instance, as demonstrated in the illustrations, distinct subtypes within a single category, such as various beard styles (e.g., “stubble,” “goatee,” or “full beard”), exhibit varied impacts on the classifier’s score. This highlights the subtle relationship between fine-grained feature variations and their respective contributions to the classification process, showcasing the importance of understanding these

hierarchical relationships for improving model interpretability and performance.

S6.2. Experiments with Joint Attributes

To examine the impact of combining multiple attributes on classifier scores, we conducted a series of experiments. Specifically, we generated images featuring joint attributes for the face, bird, and plant health classes. The attributes used in these experiments, along with the resulting changes in classifier scores, are presented in Figures S4, S5, and S6.

S6.3. Improving Classifier Accuracy with Counterfactual Images

After generating counterfactual images for the face domain, we integrated them into the training dataset of image classifiers designed to predict one’s gender and age, with the goal of improving their accuracy. The experiments in Table C demonstrate how the classifier’s performance changes when 100 counterfactuals containing the “bangs” and “makeup” attributes are added to the training data. The original classifiers were convolutional neural networks based on EfficientNet and trained with 1000 images from the FFHQ dataset. Both classifiers achieved an overall accuracy of 95 percent on their test sets. Compared to the other domains, we decided to retrain a classifier with counterfactual images of edited human faces because these images maintain contextually relevant attributes that align with the real-world variations that a classifier will encounter. On the other hand, counterfactuals of edited birds do not reflect realistic bird species (although they can help identify which features of a bird are significant for its overall classification). Thus, these types of edited images introduce features and contexts that are far removed from the target domain, making them unsuitable for training.

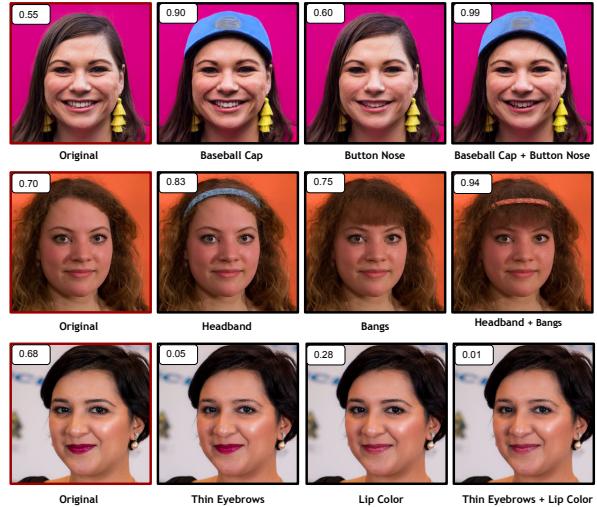


Figure S4. Joint Attributes Experiments for the Facial Domain. This figure showcases some edited facial attributes and their individual and collective effects on the age classifier’s decision. The original images, marked with red frames, are compared to their edited counterparts, marked with black frames. The classifier scores displayed in the top-left corner of each image represent how strongly the edited attributes influence the classifier’s output. Higher scores indicate a stronger impact of an attribute on a specific domain.

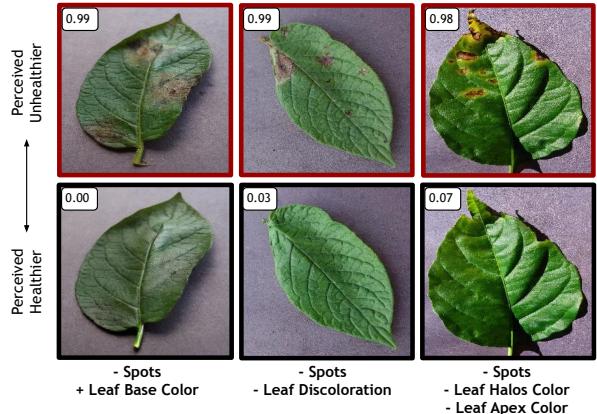


Figure S5. Joint Attributes Experiments for the Plant Health Domain. Here, we present three images edited using joint attributes in the plant health domain. A “+” sign indicates that an attribute was added to the image, while a “-” sign signifies that an attribute was removed. Images with red frames represent the original, unedited versions, while those with black frames are the edited versions. The numbers in the corners reflect the classifier’s score, indicating the perceived level of the leaf’s unhealthiness.

Domain	Top-10 Attributes	Score
Face	Eyebrow	0.74
	Makeup	0.50
	Mustache	0.47
	Teeth (Smile)	0.44
	Lip Volume	0.37
	Headwear	0.33
	Lip Color	0.25
	Beard	0.15
	Facewear	0.11
	Hair	0.10
Bird	Upperparts Color	0.55
	Head Color	0.55
	Beak Shape	0.38
	Beak Color	0.37
	Wing Pattern	0.29
	Eye Color	0.28
	Throat Color	0.27
	Wing Color	0.26
	Crest Presence	0.13
	Feather Texture	0.04
Plant Health	Leaf Base Color	0.97
	Leaf Vein Color	0.91
	Leaf Apex Color	0.89
	Leaf Spots	0.84
	Leaf Disease	0.77
	Leaf Blight Size	0.16
	Leaf Spots Color	0.10
	Leaf Texture	0.07
	Leaf Discoloration	0.04
	Leaf Orientation	0.03
Retinal Disease	Glaucoma	0.43
	Subretinal Hemorrhage	0.42
	Intraretinal Hemorrhage	0.35
	Macular Hole	0.33
	Hard Exudates	0.33
	Blackened Macula	0.23
	Soft Exudates	0.21
	Retinal Drusen	0.13
	Optic Disc Hemorrhage	0.05
	Cataract	0.04

Table A. Top-10 Attributes and their Respective Scores Across Various Domains. The table above displays the top 10 attributes for the face, bird, plant health, and retinal disease domains, ranked from highest to lowest based on their scores. These scores were derived by calculating the average difference between the classification scores of the edited and unedited images.

Algorithm	Worst Case	Average Case
StylEx	$O(n^3)$	$O(n^2)$
DiffEx (Ours)	$O(s^2 \cdot n)$	$O(b \cdot n \cdot s)$

Table B. Big-O Comparisons Between Greedy-Based StylEx and Beam-Based DiffEx. The table above provides estimations of the algorithmic efficiency of both methods. In StyleEx, n represents the number of sample images and s is the length of style coordinates. On the other hand, s in DiffEx represents the hierarchical structure of keywords extracted from the VLM.

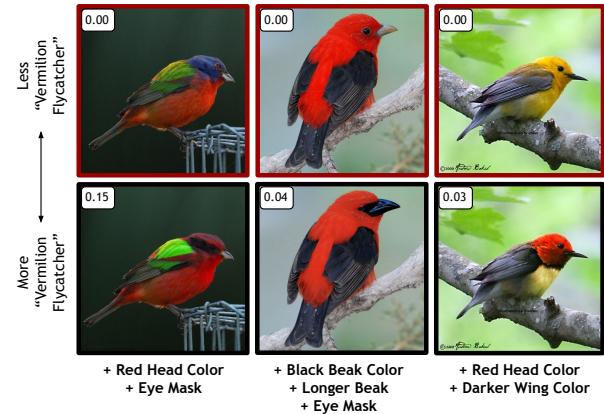


Figure S6. Joint Attributes Experiments for the Bird Domain. Here, we present three images from the bird domain that were edited to resemble the Vermilion Flycatcher. The focus was on adding attributes to make the birds appear more similar to the Vermilion Flycatcher. As seen in the joint attributes figure for the plant health domain, images with red frames represent the original versions, while those with black frames are the edited versions. A higher classifier score indicates a greater resemblance to the Vermilion Flycatcher.

Attribute	Classifier Type	Original	Updated
Makeup	Gender Classifier	91%	95%
Bangs	Age Classifier	68%	96%

Table C. Improvement in Classifier Accuracy. The table illustrates the improvement in the average accuracy of two classifiers in predicting the ages and genders of individuals with makeup and bangs, following the inclusion of counterfactual examples in the face dataset. The “Original” column presents the average classification scores for individuals with makeup and bangs before the incorporation of counterfactual examples, while the “Updated” column shows the improved average accuracy scores after adding counterfactual examples to the training data.

Level 1 Attributes	Level 2 Attributes	Level 3 Attributes
Face Features	Face Shape, Beard, Mustache	Oval Face, Round Face, Square Face, Heart-Shaped Face, Rectangular Face, Diamond-Shaped Face, Oblong Face, Triangular Face, Long Face, Narrow Face, Wide Face, Broad Face, Full Face, Chunky Face, Wide-Set Face, Expansive Face, Larger Face, Flatter Face, Goatee Beard, Full Beard, Short Beard, Long Beard, Classic Mustache, Handlebar Mustache, Horseshoe Mustache, Pencil Mustache, ...
Hair Features	Hair Color, Hair Texture, Hair Length, Hair Style	Black Hair, Brown Hair, Blonde Hair, Red Hair, Gray Hair, White Hair, Auburn Hair, Straight Hair, Wavy Hair, Curly Hair, Pixie Cut Hair, Bob Cut Hair, Bangs, Permed Hair, Bleached Hair, ...
Eyebrow Features	Eyebrow Shape, Eyebrow Density, Eyebrow Style	Arched Eyebrows, Straight Eyebrows, Thick Eyebrows, Thin Eyebrows, Curved Eyebrows, Flat Eyebrows, Angled Eyebrows, Sparse Eyebrows, Dense Eyebrows, Brushed-Up Eyebrows, Plucked Eyebrows, Threaded Eyebrows, ...
Mouth Features	Mouth Shape, Lip Volume, Lip Color, Smile Type	Full Lips, Thin Lips, Thick Lips, Wide Mouth, Narrow Mouth, Pouty Lips, Red Lip, Pink Lip, Nude Lip, Coral Lip Color, Berry Lip Color, Brown Lip Color, Purple Lip, Orange Lip, Maroon Lips, ...
Eyelash Features	Eyelash Length, Eyelash Volume, Eyelash Curl	Short Eyelashes, Medium Eyelashes, Long Eyelashes, Sparse Eyelashes, Dense Eyelashes, Straight Eyelashes, Curled Eyelashes, ...
Nose Features	Nose Shape, Nose Tip, Nostril Shape	Straight Nose, Curved Nose, Button Nose, Hooked Nose, Flat Nose, Wide Nose, Narrow Nose, Upturned Nose, Long Nose, Broad Nose, Pointed Nose, Roman Nose, Snub Nose, Aquiline Nose, Crooked Nose, Rounded Tip Nose, Pointed Tip Nose, Wide Nostril, Narrow Nostril, Flared Nostril, ...
Skin Features	Skin Texture, Skin Color	Smooth Skin, Rough Skin, Oily Skin, Dry Skin, Combination Skin, Sensitive Skin, Acne-Prone Skin, Wrinkled Skin, Freckled Skin, Blemished Skin, Porous Skin, Flaky Skin, Fair Skin, Light Skin, Medium Skin, Dark Skin, Olive Skin, Tan Skin, ...
Accessories	Jewelry, Facewear, Headwear	Earrings, Necklace, Bracelet, Ring, Glasses, Sun-glasses, Face Mask, Hat, Scarf, Headband, Bow Tie, Hairband, Beanie, Beaded Headband, Tiara, ...
Makeup	Makeup Style, Makeup Type	Natural Makeup, Glam Makeup, Smoky Eye Makeup, Dewy Makeup, Matte Makeup, Bold Lip Makeup, Bridal Makeup, Festive Makeup, Eye-shadow Makeup, Eyeliner Makeup, Blush Makeup, Lipstick Makeup, Highlighter Makeup, Mascara Makeup, ...

Table D. **Examples of Attribute Candidates Proposed for the Face Domain.** The table above shows potential level 1, level 2, and level 3 attributes for the face domain. Due to limited space, we include a sample list of level 3 attributes for the first level 2 attribute listed in each row.

Level 1 Attributes	Level 2 Attributes
Leaf Base Color	Green, Yellow, Light Green, Dark Green, Orange, Red, Brown, Purple, Pink, White, Light Yellow, Dark Red, Burgundy, Copper, Chartreuse, Ivory, Olive, Black, Tan, ...
Leaf Apex Color	Green, Yellow, Red, Purple, Brown, Orange, Pink, White, Light Green, Dark Green, Light Yellow, Dark Red, Rust, Burgundy, Violet, Lime Green, Chartreuse, Copper, Amber, Ivory, ...
Leaf Spots	With Spots, Without Spots
Leaf Disease	Spots, Lesions, Discoloration, Necrosis, Blight, Mold, Mildew, Rust, Canker, Wilting, Decay, Yellowing, Browning, Pustules, Fungal Infection, Bacterial Infection, Viral Infection, Chlorosis, Fungal Growth, Powdery Mildew, Downy Mildew, ...
Leaf Blight Size	Small Blight, Medium Blight, Large Blight, Tiny Blight, Extensive Blight, Minor Blight, Moderate Blight, Severe Blight, Pinpoint Blight, Patchy Blight
Leaf Spots Color	Brown Spots, Yellow Spots, Black Spots, Red Spots, Orange Spots, Green Spots, White Spots, Purple Spots, Light Green Spots, Dark Brown Spots, ...
Leaf Shape	Oblong Shape, Ovate Shape, Lanceolate Shape, Cordate Shape, Elliptical Shape, Linear Shape, Palmate Shape, Pinnate Shape, Lobed Shape, Tamarisk Shape, Sagittate Shape, Triangular Shape, Denticulate Shape, Wedge Shape, Reniform Shape, Setaceous Shape, Circinate Shape, Falcate Shape, Acicular Shape, Subulate Shape, ...
Leaf Symmetry	Bilateral Symmetry, Radial Symmetry, Asymmetrical, Mirror Symmetry, Transverse Symmetry, Rotational Symmetry, ...

Table E. **Examples of Attribute Candidates Proposed for the Plant Health Domain.** The table above lists potential attributes for the plant health domain. However, not all of these attributes are relevant for describing a leaf or would result in effective edits. Therefore, DiffEx filters this list, selecting only the most meaningful and applicable attributes.

Level 1 Attributes	Level 2 Attributes	Level 3 Attributes
Beak	Beak Color, Beak Shape, Beak Size	Yellow, Orange, Black, Red, Brown, Pink, White, Blue, Green, Grey, Ivory, Cream, Purple, Beige, Tan, Light Pink, Dark Brown, Light Yellow, Dark Green, ...
Wings	Wing Shape, Wing Color, Wing Pattern	Pointed, Rounded, Elongated, Broad, Narrow, Oval, Triangular, Crescent, Oval-Shaped, Square, Short, Long, Fan-Shaped, Forked, Tapered, Slender, Angular, Spade-Shaped, Elliptical, High-Speed, Soaring, High-Aspect Ratio, Cambered, Alula, Swept-Back, V-Shaped, Bent, ...
Eye	Eye Shape, Eye Size, Eye Color	Round, Oval, Almond, Circular, Slit, Horizontal, Vertical, Hooded, Wide, Narrow, Protruding, Sunken, Large, Small, Bulging, Beady, Piercing, Squinted, Deep-Set, Prominent, ...
Head	Head Color, Crest Presence	Black, White, Yellow, Red, Blue, Brown, Green, Grey, Orange, Pink, Purple, Cream, Beige, Tan, Violet, Charcoal, Silver, Rust, Burgundy, Golden, Copper, ...
Body	Feather Texture, Upperparts Color, Body Size, Belly Color, Tail Length, Leg Color	Soft, Coarse, Smooth, Rough, Fluffy, Silky, Woolly, Feathery, Stiff, Shiny, Matted, Glossy, Velvet, Harsh, Prickly, Fuzzy, Curled, Frizzy, Downy, Crisp, ...

Table F. **Examples of Attribute Candidates Proposed for the Bird Domain.** The table above shows potential level 1, level 2, and level 3 attributes for the bird domain. Due to limited space, we include a sample list of level 3 attributes for the first level 2 attribute listed in each row.

```
[{"role": "system",
"content": "You are an expert at finding features important for text-based image editing using diffusion models, given a set of images. Upon receiving a set of images, analyze the given inputs and extract important features and keywords that can be used for text-based image editing using diffusion models. Analyze the set of images and identify key features that define or are significant within the specified domain. These features are encoded to guide generative diffusion model for fine-grained image editing of subjects.

List all different categories related to that specific feature. For example, for DOMAIN_NAME features, it ranges from ATTRIBUTE_1 to ATTRIBUTE_2, ATTRIBUTE_3, ATTRIBUTE_4, etc. Output must be in the format given, a sample output is given below, give the output only without any other descriptive text. Do not restrict your answers to the given sample, come up with all features. I want detailed fine-grained features.

[ {
    "ATTRIBUTE_1": {"sub_attribute_1_1" , "sub_attribute_1_2", "sub_attribute_1_3", },
    "ATTRIBUTE_2": {"sub_attribute_2_1", "sub_attribute_2_2", "sub_attribute_2_3" },
    "ATTRIBUTE_3": {"sub_attribute_3_1", "sub_attribute_3_2" },
    "ATTRIBUTE_4": {"sub_attribute_4_1", "sub_attribute_4_2" },
    "ATTRIBUTE_5": {"sub_attribute_5_1", "sub_attribute_5_2", "sub_attribute_5_3" },
} ]
```

Table G. Prompt Template for Keyword-Extraction. The text above illustrates the standard format used to input text prompts into GPT-4 for extracting potential attributes across different domains. “DOMAIN_NAME” refers to a specific domain, such as facial features, bird species, etc. “ATTRIBUTE_1, ATTRIBUTE_2, etc.” refer to the Level 1 (broad) categories, while “sub_attribute_1_1, sub_attribute_1_2, etc.” refer to Level 2 (finer-grained) categories.

```
[{"role": "system",
"content": "You are an expert at finding features important for text-based image editing using diffusion models, given a set of images. Upon receiving a set of images, analyze the given inputs and extract important features and keywords that can be used for text-based image editing using diffusion models. Analyze the set of images and identify key features that define or are significant within the specified domain. These features are encoded to guide generative diffusion model for fine-grained image editing of subjects.

List all different categories related to that specific feature. For example, for human features, it ranges from skin texture to expression, accessories, eyebrow shape, etc. Output must be in the format given, a sample output is given below, give the output only without any other descriptive text. Do not restrict your answers to the given sample, come up with all features. I want detailed fine-grained features.

[{"Face": {"oval face", "rectangular face", "round face"}, "Skin Texture": {"smooth skin", "freckled skin", "blemish skin", "scar skin"}, "Skin Color": {"light colored skin", "dark colored skin"}, "Eyes Shape": {"round eyes", "almond eyes"}, "Eyes Color": {"blue colored eyes", "green colored eyes", "hazel colored eyes"}, "Eyebrows": {"thin eyebrows", "bushy eyebrows"}, "Hair Color": {"dark colored hair", "light colored hair", "blonde hair", "brunette hair"}, "Hair Texture": {"straight hair", "curly hair", "wavy hair"}, "Hair Length": {"short hair", "long hair", "medium hair"}, "Nose Shape": {"button nose", "straight nose", "prominent nose"}, "Mouth Shape": {"full lip", "thin lip"}, "Lip Color": {"matte lip", "glossy lip"}, "accessories": {"earrings", "necklace", "glasses", "sunglasses"}]}]
```

Table H. Face Domain Keyword-Extraction Prompt Used in GPT-4. The text above shows the prompt we fed into the VLM in order to find potential attributes in the face domain.

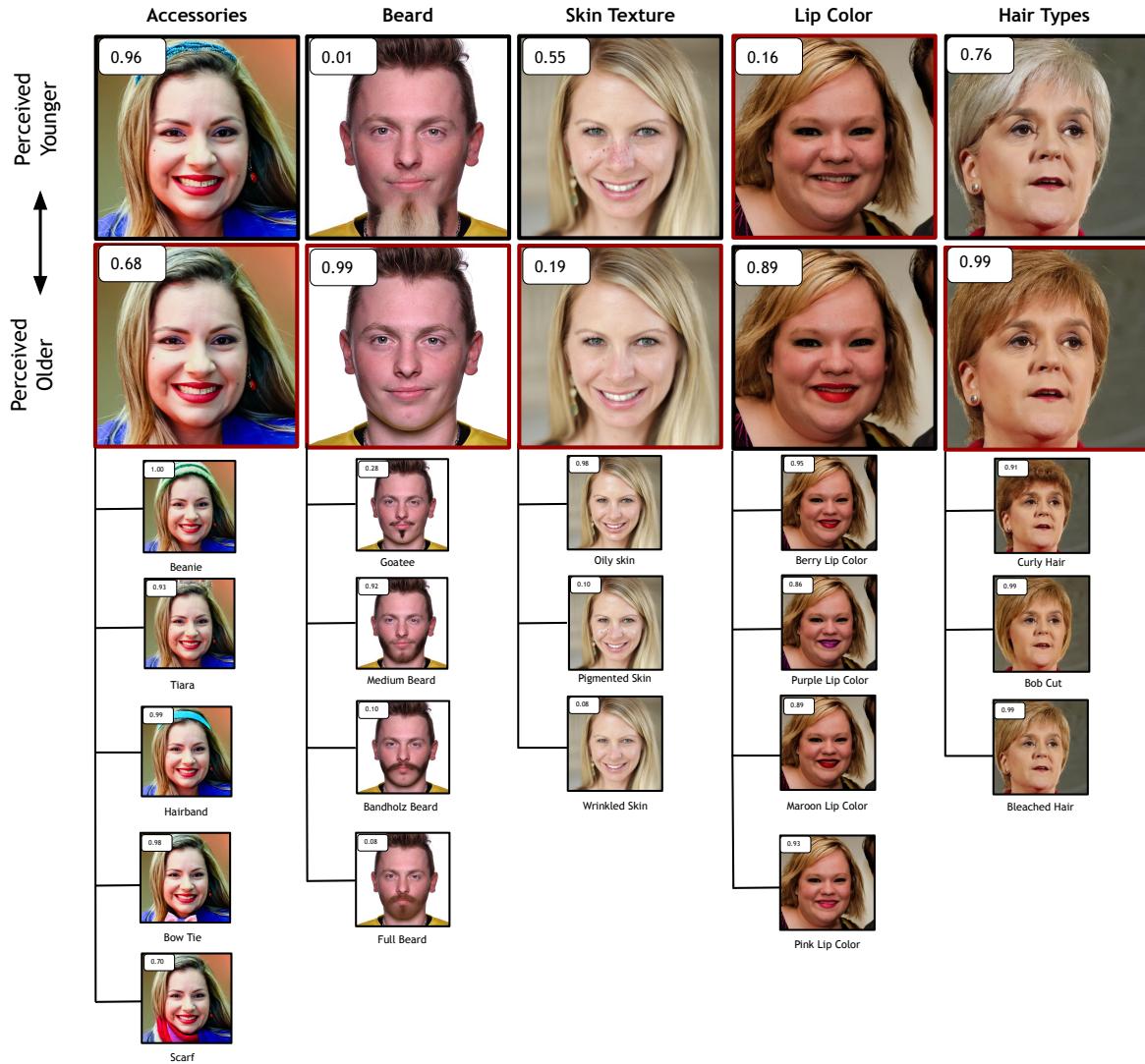


Figure S7. Hierarchical Structure of the Top Facial Attributes and their Impact on Age Classifier Scores. The figure demonstrates how DiffEx organizes fine-grained attribute categories and their influence on classifier decisions. Logit scores in the top-left corners represent the scores for the “young” label.