Explaining in Diffusion: Explaining a Classifier with Diffusion Semantics

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 F to GPT-4. A more detailed example of the text prompt for the face domain can be found in Table G.

S3. Hierarchy of Attributes

Figure S2 depict the hierarchical structures of various attributes for the bird and retinal disease domains respectively. Tables C, D and E 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. Failure Cases

Figure S1 shows some instances where the image editing model used (Ledits++) cannot edit the fine-grained attribute in the image.

S5. Algorithmic Comparison

We compared the runtime for DiffEx against StylEx algorithm. Our ranking method takes only 0.6 min and editing takes 16.1 min for 200 semantics, totaling 16.7 min, while StylEx takes 30.3 min. Faster editing methods can be integrated for further speed-ups.

Figure S1. **Failure Cases.** The image editing model we used has limitations in terms of capturing fine-grained semantics.

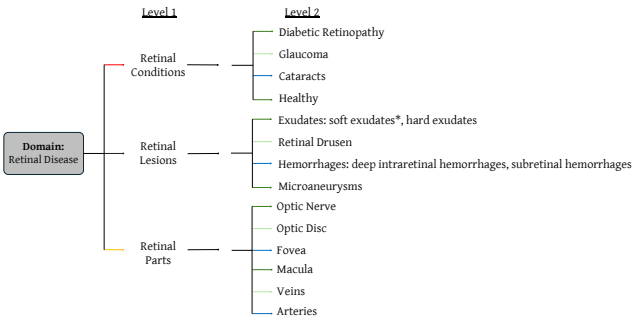

Figure S2. **Hierarchical List of Attributes for the Retinal Disease Domain.** The diagram above illustrates the hierarchical organization of various attributes within the retina scan domain, showing the levels to which they belong. Note: The asterisk next to "soft exudates" denotes that they are also referred to as "cotton wool spots," and the sub-categories under "exudates" are part of the level 3 hierarchy.

S6. 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

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.

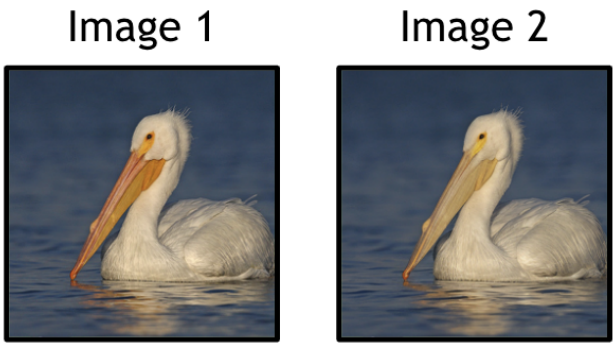

Figure S3. **Example of Original and Edited Image Comparison from User Study 1.** The image pair above serves as an example from the quantitative study on edit quality and disentanglement, specifically for the "beak color" attribute.

ensure that our evaluation captures real-world factors across diverse domains. Subsections S6.1 and S6.2 include some example questions from these user studies.

S6.1. 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 S3) 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.

Figure S4. **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.

S6.2. User Study 2: Comparisons with Grad-CAM and StylEx

To quantitatively evaluate our method, DiffEx, against other explainability metrics, we conducted another user 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 S4 were:

a.) Feather Texture
b.) Beak Shape
c.) Beak Color
d.) Eye Color

S6.3. User Study 3: Disentanglement in Images with Multiple Attributes

Figures S5 and S6 show examples of questions that were asked as part of user study 3 (as described in the main paper). Eight questions were asked in total (four regarding the disentanglement of edits with single attributes, and four regarding the disentanglement of edits with multiple attributes).

S7. 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.

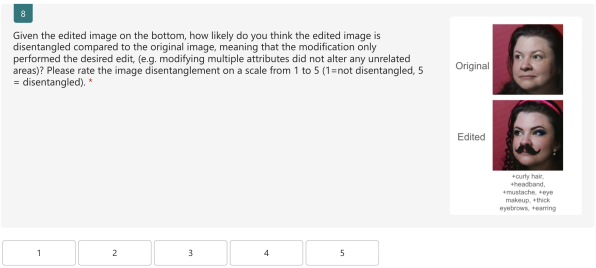

Figure S5. **User Study 3 Multiple Attributes Question**. The image above provides an example of a question from our third user study, where participants were asked to evaluate the quality of edits that added multiple attributes to an existing image.

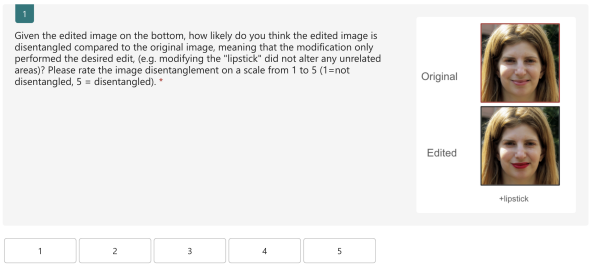

Figure S6. **User Study 3 Single Attributes Question**. The image above provides an example of a question from our third user study, where participants were asked to evaluate the quality of edits that added single attributes to an existing image.

S7.1. Experiments with Single Attributes

To effectively illustrate the hierarchical structure of the edited features, additional experimental results are presented in Figure S10, 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.

S7.2. 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

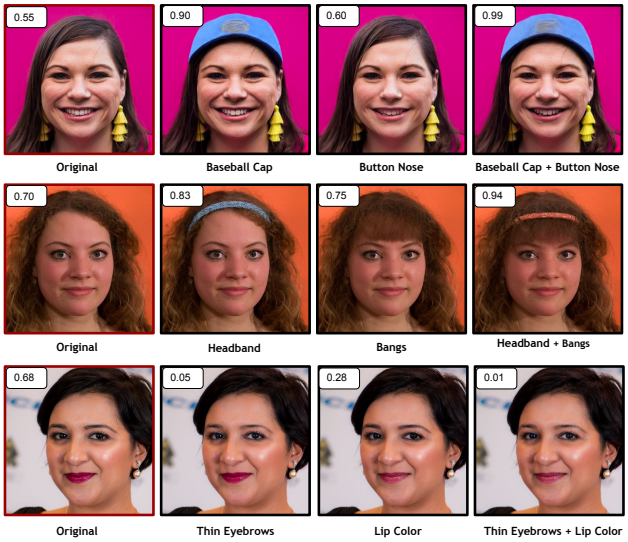

Figure S7. **Joint Attributes Experiments for the Facial Domain.** This figure showcases some edited facial attributes and their individual and collective effects on the age classifier's decision. The original images, marked with red frames, are compared to their edited counterparts, marked with black 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.

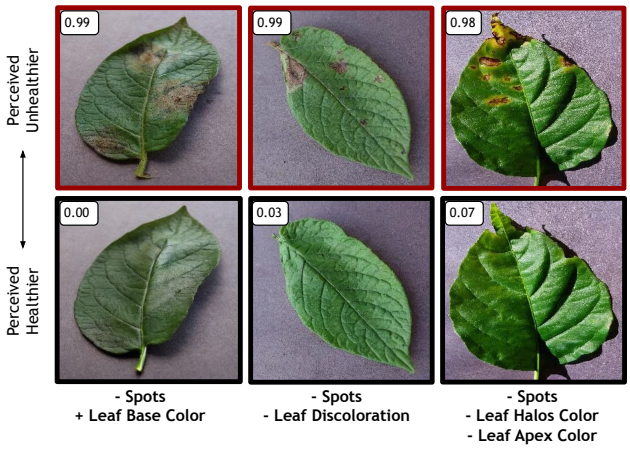

Figure S8. **Joint Attributes Experiments for the Plant Health Domain.** Here, we present three images edited using joint attributes in the plant health domain. A "+" sign indicates that an attribute was added to the image, while a "-" sign signifies that an attribute was removed. Images with red frames 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.

in classifier scores, are presented in Figures S7, S8, and S9.

S7.3. 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 B 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.

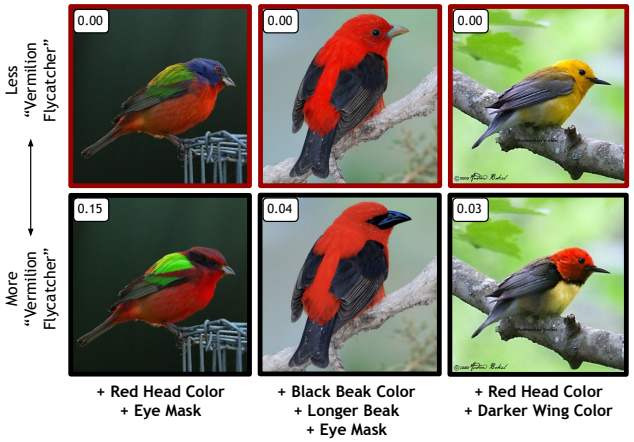

Figure S9. **Joint Attributes Experiments for the Bird Domain**. Here, we present three images from the bird domain that were edited to resemble the Vermilion Flycatcher. The focus was on adding attributes to make the birds appear more similar to the Vermilion Flycatcher. As seen in the joint attributes 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 B. **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, Sunglasses, 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, Eyeshadow Makeup, Eyeliner Makeup, Blush Makeup, Lipstick Makeup, Highlighter Makeup, Mascara Makeup, ...

Table C. **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 D. **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 E. **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 F. **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",}
       "earrings", "necklace, glasses, sunglasses",
}]
```

Table G. **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.

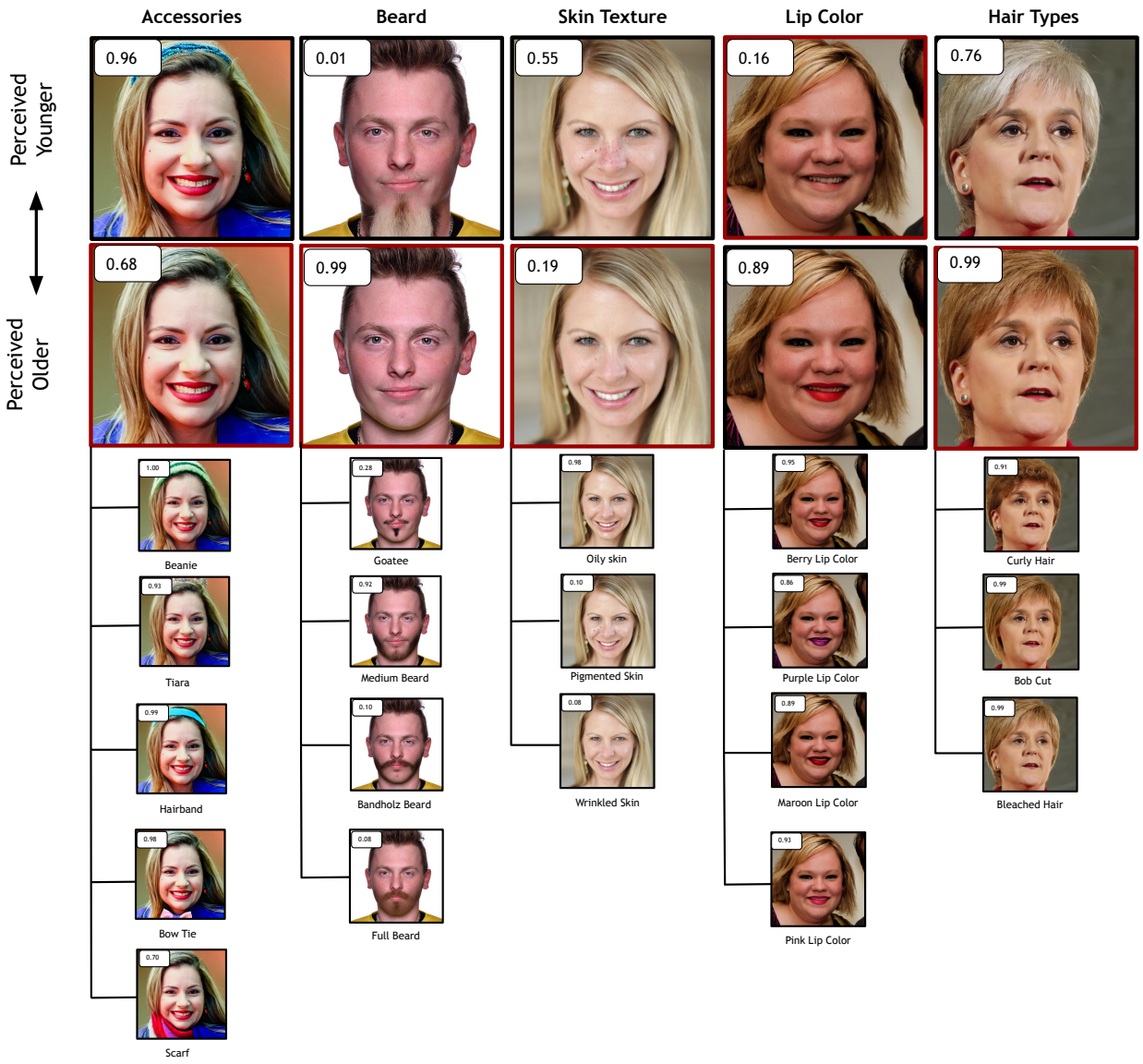

Figure S10. **Hierarchical Structure of the Top Facial Attributes and their Impact on Age Classifier Scores.** The figure demonstrates how DiffEx organizes fine-grained attribute categories and their influence on classifier decisions. Logit scores in the top-left corners represent the scores for the "young" label.