# Supplementary Material for From Visual Explanations to Counterfactual Explanations with Latent Diffusion

## A. Proof

In this section, we provide the proofs for the propositions. Our objective is to assess the degree of deviation in optimization when replacing the gradient $\boldsymbol{u}^*$, which is the fastest direction, with the pruned gradient $\boldsymbol{v}^*$. Additionally, we establish the relationship between convergence and the thresholds $\xi_d$. We restate the terminology previously mentioned in the main paper: Suppose $\xi^*$ is the optimal threshold to split foreground $\mathcal{F}^*$ and background $\mathcal{B}^*$. $\mathcal{A} = \mathcal{F} \cup \mathcal{B}$ represents the set of all pixel coordinates in the original image $\mathcal{I}^{\mathrm{F}}$. First, $c : \mathbb{R}^{H_2 \times W_2 \times 1} \to \mathbb{R}^{H_2 \times W_2 \times C}$ expands latent mask by concatenating this binary mask itself $C$ times along the channel dimension. Second, let $g : \mathbb{R}^{H_2 \times W_2 \times C} \to \mathbb{R}^{H_2 W_2 C}$ map the space $H_2 \times W_2 \times C$ to a vector of dimension $H_2 W_2 C$, where $H_2, W_2, C$ is the height, width, channel of latent space. $\mathcal{M}'$ represents the latent mask corresponding to the optimal threshold $\xi^*$.

**Proposition 1.** *Let $\boldsymbol{u}^* = g\left(\nabla_{\boldsymbol{z}^{init}} \mathcal{L}_{CE}\left(f_{cl}\left(\mathcal{I}^{(k)}\right), y^{CF}\right)\right)$ is the optimal vector, and $\boldsymbol{v}^* = \boldsymbol{m}' \odot \boldsymbol{u}^*$ represents the pruned adversarial gradient vector with $\boldsymbol{m}' = g(c(\mathcal{M}'))$. Then*

$$0^\circ \leq \angle(\boldsymbol{u}^*, \boldsymbol{v}^*) < 90^\circ, \tag{1}$$

*where $\boldsymbol{u}^* = (u_{1,1,1}, \ldots, u_{H_2,W_2,1}, \ldots, u_{H_2,W_2,4})$, $u_{i,j,s}$ represents the element corresponding to $\boldsymbol{\gamma}^{(k)}_{i,j,s}$.*

*Proof.*

$$\begin{aligned} \cos(\angle(\boldsymbol{u}^*, \boldsymbol{v}^*)) &= \frac{\boldsymbol{u}^* \cdot \boldsymbol{v}^*}{\|\boldsymbol{u}^*\| \|\boldsymbol{v}^*\|} \\ &= \frac{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{F}^*} u^2_{i,j,s}}{\underbrace{\sqrt{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{A}} u^2_{i,j,s}}}_{\text{constant}} \sqrt{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{F}^*} u^2_{i,j,s}}} \\ &= \gamma \sqrt{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{F}^*} u^2_{i,j,s}} > 0 \quad (\gamma > 0). \end{aligned}$$

Obviously:

$$0^\circ \leq \angle(\boldsymbol{u}^*, \boldsymbol{v}^*) < 90^\circ. \tag{2}$$

□

**Proposition 2.** *Suppose that $\{\xi_d\}_{d=1}^{\infty}$ is a sequence of thresholds satisfying:*

$$0 \leq \xi_d < \xi_{d+1} < 1. \tag{3}$$

*Then there exists a sequence of vectors $\left\{\boldsymbol{v}^{(d)}\right\}_{d=1}^{\infty}$ such that:*

$$0^\circ \leq \angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(1)}\right) \leq \cdots \leq \angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(\infty)}\right) < 90^\circ. \tag{4}$$

*Proof.* For Equation 3, we have the foreground and background sets corresponding to each threshold $\xi$ that satisfy the following conditions: $\mathcal{F}_d \supseteq \mathcal{F}_{d+1}$, $\mathcal{B}_d \subseteq \mathcal{B}_{d+1}$, $\mathcal{A} = \mathcal{F}_d \cup \mathcal{B}_d$, where $d \in \mathbb{N}^*$. Thanks to Proposition 1, we continue to analyze:

$$\cos\left(\angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(d)}\right)\right) = \gamma \sqrt{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{F}_d} u^2_{i,j,s}} \tag{5}$$

$$\geq \gamma \sqrt{\sum_{s=1}^{4} \sum_{(i,j)\in\mathcal{F}_{d+1}} u^2_{i,j,s}} \tag{6}$$

$$= \cos\left(\angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(d+1)}\right)\right) \tag{7}$$

Therefore:

$$0^\circ \leq \angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(d)}\right) \leq \angle\left(\boldsymbol{u}^*, \boldsymbol{v}^{(d+1)}\right) \leq 90^\circ, \forall d \in \mathbb{N}^*. \tag{8}$$

This proof is complete. □

## B. Overview of ECED

In this section, we provide the following information and material.

- Algorithms of our ECED.
- The implementation details and hyperparameter configurations of Latent Diffusion.
- Fine-tuning strategy for Stable Diffusion on the CelebA-HQ dataset.

- The ability to leverage the context of Latent Diffusion.
- The efficiency of blending latents.
- The preservation strategy of our method.
- More qualitative results.

### B.1. Algorithms

**Algorithm 1** Identifying key region

```
Require: Initial image $\mathcal{I}^{\text{F}}$, original label $y^{\text{F}}$, pretrained classifier $f_{cl}$, classifier's specified layer $l$, threshold $\xi$ spliting two parts of the image, visual explanation algorithm $ScoreCAM$
function IDENTIFY-FG($\mathcal{I}^{\text{F}}$, $y^{\text{F}}$, $l$, $\xi$)
    $\mathbf{U} = ScoreCAM(\mathcal{I}^{\text{F}}, y^{\text{F}}, f_{cl}, l)$    ▷ Extract the attention map
    $\mathcal{M} = \mathbf{U}[u_{i,j} > \xi]$    ▷ Get the binary mask
    $\mathcal{M}' = downsample(\mathcal{M})$    ▷ Get the latent mask
    return $\mathcal{M}'$
end function
```

**Algorithm 2** Preserving background during image synthesis

```
Require: Initial image $\mathcal{I}^{\text{F}}$, latent mask $\mathcal{M}'$, $VAE = (\mathcal{E}(\cdot), \mathcal{D}(\cdot))$, noise coefficient $\alpha_0$, distance loss $\mathcal{L}_{dis}$, number of update iterations $N$, optimization algorithm $Adam$
function PRESERVE-BG($\mathcal{I}^{\text{F}}$, $\mathcal{M}'$)
    $\boldsymbol{z}^{\text{F}} = \boldsymbol{z}^{\text{bg}} = \mathcal{E}(\mathcal{I}^{\text{F}})$    ▷ Init latents
    for $i = 1, \ldots, N$ do
        $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
        $\boldsymbol{z}^{\text{init}} \leftarrow \boldsymbol{z}^{\text{F}} \odot \mathcal{M}' + \boldsymbol{z}^{\text{bg}} \odot (\mathbf{1} - \mathcal{M}')$    ▷ Blend latents
        $\boldsymbol{z}_0 \leftarrow \sqrt{\alpha_0}\boldsymbol{z}^{\text{init}} + \sqrt{1-\alpha_0}\boldsymbol{\epsilon}$    ▷ Add noise
        $\mathcal{I}' \leftarrow \mathcal{D}(\boldsymbol{z}_0)$
        $grad_1 \leftarrow \nabla_{\boldsymbol{z}^{\text{bg}}}\mathcal{L}_{dis}(\mathcal{I}', \mathcal{I}^{\text{F}}, \mathcal{M}')$
        $grad_2 \leftarrow \nabla_{\theta_{\mathcal{D}}}\mathcal{L}_{dis}(\mathcal{I}', \mathcal{I}^{\text{F}}, \mathcal{M}')$
        $\boldsymbol{z}^{\text{bg}} \leftarrow Adam(\boldsymbol{z}^{\text{bg}}, grad_1)$    ▷ Update
        $\theta_{\mathcal{D}} \leftarrow Adam(\theta_{\mathcal{D}}, grad_2)$    ▷ Update
    end for
    return $\boldsymbol{z}^{\text{bg}*}, \theta^*_{\mathcal{D}}$
end function
```

**Algorithm 3** Generating counterfactual explanation

```
Require: Initial image $\mathcal{I}^{\text{F}}$, target label $y^{\text{CF}}$, latent mask $\mathcal{M}'$, original latent $\boldsymbol{z}^{\text{F}}$, optimal background latent $\boldsymbol{z}^{\text{bg}*}$, Latent Diffusion model $SD = \{(\mathcal{E}(\cdot), \mathcal{D}^*(\cdot)), DiffusionModel = (noise(\boldsymbol{z}, t), denoise(\boldsymbol{z}, \mathbf{C}, t))\}$, text encoder CLIP, sequence of noise coefficients $\{\alpha_t\}_{t=0}^{T_1}$, diffusion steps $\tau$, classifier $f_{cl}$, Cross-Entropy loss $\mathcal{L}_{CE}$, number of update iterations $T_2$, optimization algorithm $Adam$
function GENERATE-CE($\mathcal{I}^{\text{F}}$, $y^{\text{CF}}$, $\mathcal{M}'$, $\boldsymbol{z}^{\text{bg}*}$)
    $\mathbf{C}^{\text{CF}} = CLIP(y^{\text{CF}})$
    $\boldsymbol{z}^{\text{init}} \leftarrow \boldsymbol{z}^{\text{F}} \odot \mathcal{M}' + \boldsymbol{z}^{\text{bg}*} \odot (\mathbf{1} - \mathcal{M}')$
    ▷ Attack iteration steps
    for $t_2 = 0, \ldots, T_2$ do
        ▷ Blended latent diffusion algorithm
        $\boldsymbol{z}_\tau \sim noise(\boldsymbol{z}^{\text{init}}, \tau)$
        for $t = \tau, \ldots, 0$ do
            $\boldsymbol{z}^{\text{fg}} \sim denoise(\boldsymbol{z}_t, \mathbf{C}^{\text{CF}}, t)$
            $\boldsymbol{z}_t^{\text{bg}*} \sim noise(\boldsymbol{z}^{\text{bg}*}, t)$
            $\boldsymbol{z}_t \leftarrow \boldsymbol{z}^{\text{fg}} \odot \mathcal{M}' + \boldsymbol{z}_t^{\text{bg}*} \odot (\mathbf{1} - \mathcal{M}')$
        end for
        $\mathcal{I}^{(t_2)} \leftarrow \mathcal{D}^*(\boldsymbol{z}_0)$
        $grad \leftarrow \nabla_{\boldsymbol{z}^{\text{init}}}\mathcal{L}_{CE}\left(f_{cl}\left(\mathcal{I}^{(t_2)}\right), y^{\text{CF}}\right)$
        $\boldsymbol{z}^{\text{init}} \leftarrow Adam(\boldsymbol{z}^{\text{init}}, grad \odot \mathcal{M}')$    ▷ Update with pruning-based attack
    end for
    $\mathcal{I}^{\text{CF}} = \mathcal{I}^{(T_2)}$    ▷ Counterfactual explanation
    return $\mathcal{I}^{\text{CF}}$
end function
```

### B.2. Implementation Details

In this work, we implement the blended latent diffusion algorithm proposed in [1]. To reiterate, this algorithm blends latent representations at each timestep $t$, defined as follows:

$$\boldsymbol{z}_t^{\text{bg}} = \sqrt{\bar{\alpha}_t}\boldsymbol{z}^{\text{F}} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t, \tag{9}$$

$$\boldsymbol{z}_t^{\text{fg}} \approx \sqrt{\alpha_{t-1}}\hat{\boldsymbol{z}}_0 + \beta_{t-1}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, \mathbf{C}^{\text{CF}}) + \sigma_t\boldsymbol{\epsilon}_t. \tag{10}$$

According to the hyperparameter settings and configuration of Stable Diffusion, $\{\beta_d\}_{t=0}^{T_1}$ defines a linear noise scheduling with $\beta_0 = 0.00085$ and $\beta_{T_1} = 0.012$ ($\alpha_t = 1$ if $t < 0$), and $\bar{\alpha}_t = \prod_{s=0}^{t}\alpha_s$. The representations and coefficients in Equation 9 are: $\boldsymbol{\epsilon}_t \sim \mathcal{N}(0, \mathbf{I})$, $\hat{\boldsymbol{z}_0} = \frac{\boldsymbol{z}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t, \mathbf{C}^{\text{CF}})}{\sqrt{\bar{\alpha}_t}}$, $\beta_t = \sqrt{1-\alpha_{t-1}-\sigma_t^2}$, $\sigma_t = \sqrt{\frac{1-\alpha_{t-1}}{1-\alpha_t}}\sqrt{1-\frac{\alpha_t}{\alpha_{t-1}}}$.

To represent the target classes, these conditions are mapped to CLIP-style text prompts [8], as follows:

- ImageNet: `A photo of a/an {category}.` (`attribute` ∈ {cougar, cheetah, sorrel, zebra, Persian cat, Egyptian cat}).
- CelebA-HQ: `A photo of a/an {attribute} face.` (`attribute` ∈ {smiling, non smiling, young, old}).

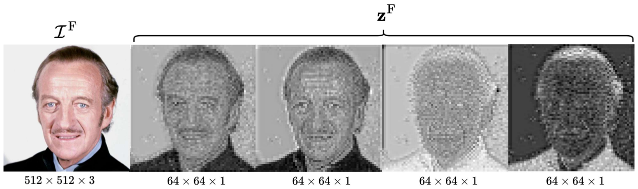

Figure 1. The visualization of the latent space.

For CelebA-HQ dataset, we fine-tuned the Stable Diffusion model to align the generated images with the desired conditions. Additionally, the purpose of this optimization is to generate a set of images that closely resemble the data distribution, thereby improving the FID score. Specifically, we optimized the weights of the UNet [9] to minimize the following loss:

$$\mathcal{L}_{LDM} = \mathbb{E}_{\mathcal{E}(\mathbf{x}),\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\boldsymbol{z}_t,\mathbf{C},t)\|_2^2\right]. \quad (11)$$

We utilized the AdamW optimizer [7] with $betas = (0.9, 0.999)$ and a learning rate of $10^{-4}$.

### B.3. Blending Approach in Latent Diffusion

We provide examples in Figure 1. The channels of the latent variable capture high-level features of the original image. Therefore, Avrahami *et al.*'s approach [1] effectively separated the latent subspaces of background and foreground using a latent mask rescaled from the original binary mask. Combined with the optimal latent $\boldsymbol{z}^{\text{bg}*}$, the details in the counterfactual image are likely to tightly align the context. This is primarily because Stable Diffusion integrated two attention mechanisms into the UNet model during the denoising process. Specifically, self-attention highlights the importance of latent features relative to the query feature, while cross-attention indicates the significance of positions on the generated image with respect to the content of the text prompt. These issues have been discussed in related works [3, 6].

### B.4. Preservation Strategy

Based on the blending strategy mentioned in Algorithm 3, we observed that the latent variable $\boldsymbol{z}_t$ retains a small amount of noise in the subspace related to the background at timestep $t = 0$, corresponding to the noise coefficient $\beta_0 = 0.00085$. This is why we simulate the noise addition process before decoding the latent $\boldsymbol{z}_0$ back to the pixel space in the second phase. In the experimental section, we verified the effectiveness of this approach by calculating the pixel-wise difference in the background region between the reconstructed image and the original image. Formally, we calculate the loss as follows:

$$\mathcal{L}_{bg} = \frac{1}{|\mathcal{B}|}\sum_{(i,j)\in\mathcal{B}} MSE(\mathcal{D}(\boldsymbol{z}_0)_{i,j}, \mathcal{I}^{\text{F}}_{i,j}), \quad (12)$$

where MSE denotes the Mean Squared Error, and $\mathcal{B}$ represents the locations of the pixels in the background. We conducted experiments and computed the average difference over 100 random samples, and then obtained the confidence interval based on the normal distribution.

### B.5. More qualitative results

We provide counterfactual explanations, presented in Figure 2 and 3. Additionally, we examine the diversity in generating CEs by ECED by setting different thresholds $\xi$, as shown in Figure 4.

## C. Evaluation Protocols for Counterfactual Explanations

Visual counterfactual explanations are evaluated based on three key criteria: Closeness/Sparsity, Validity, and Realism.

### C.1. Sparsity

*Euclidean distance* matches the $p$-th order discrepancy of pixel values at each corresponding position between the original image $\mathcal{I}^{\text{F}}$ and the counterfactual image $\mathcal{I}^{\text{CF}}$:

$$\begin{aligned}\mathcal{L}_p &= \frac{1}{N}\sum_{i=1}^{N}\|d_i\|_p & (13)\\ &= \frac{1}{N}\sum_{i=1}^{N}(\sum_{c=1}^{C}\sum_{h=1}^{H_1}\sum_{w=1}^{W_1}|\mathcal{I}^{\text{F}}_{i,c,h,w}-\mathcal{I}^{\text{CF}}_{i,c,h,w}|^p)^{\frac{1}{p}}, & (14)\end{aligned}$$

where $N$ is the number of images, and $C, H_1, W_1$ are the number of channels, height, and width of the images, respectively, with $p > 0$.

*SimSiam Similarity* measures the cosine similarity between the counterfactual image $\mathcal{I}^{\text{CF}}$ and the corresponding original image $\mathcal{I}^{\text{F}}$ in the feature space extracted by the self-supervised $SimSiam$ model [2].

$$S^3(\mathcal{I}^{\text{CF}},\mathcal{I}^{\text{F}}) = \frac{\mathcal{S}(\mathcal{I}^{\text{CF}})\cdot\mathcal{S}(\mathcal{I}^{\text{F}})}{\|\mathcal{S}(\mathcal{I}^{\text{CF}})\|\|\mathcal{S}(\mathcal{I}^{\text{F}})\|}. \quad (15)$$

*Correlation Difference* (CD) and *Mean Number of Attribute Changes* (MNAC) measure the average number of attributes modified in the counterfactual explanation, where MNAC addresses the limitations of CD.

$$\text{MNAC} = \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}}\left[\mathbb{I}\left(\mathbb{I}\left(O_a(\mathcal{I}_i^{CF})>\beta\right)\neq\mathbb{I}\left(O_a(\mathcal{I}_i^{F})>\beta\right)\right)\right], \quad (16)$$

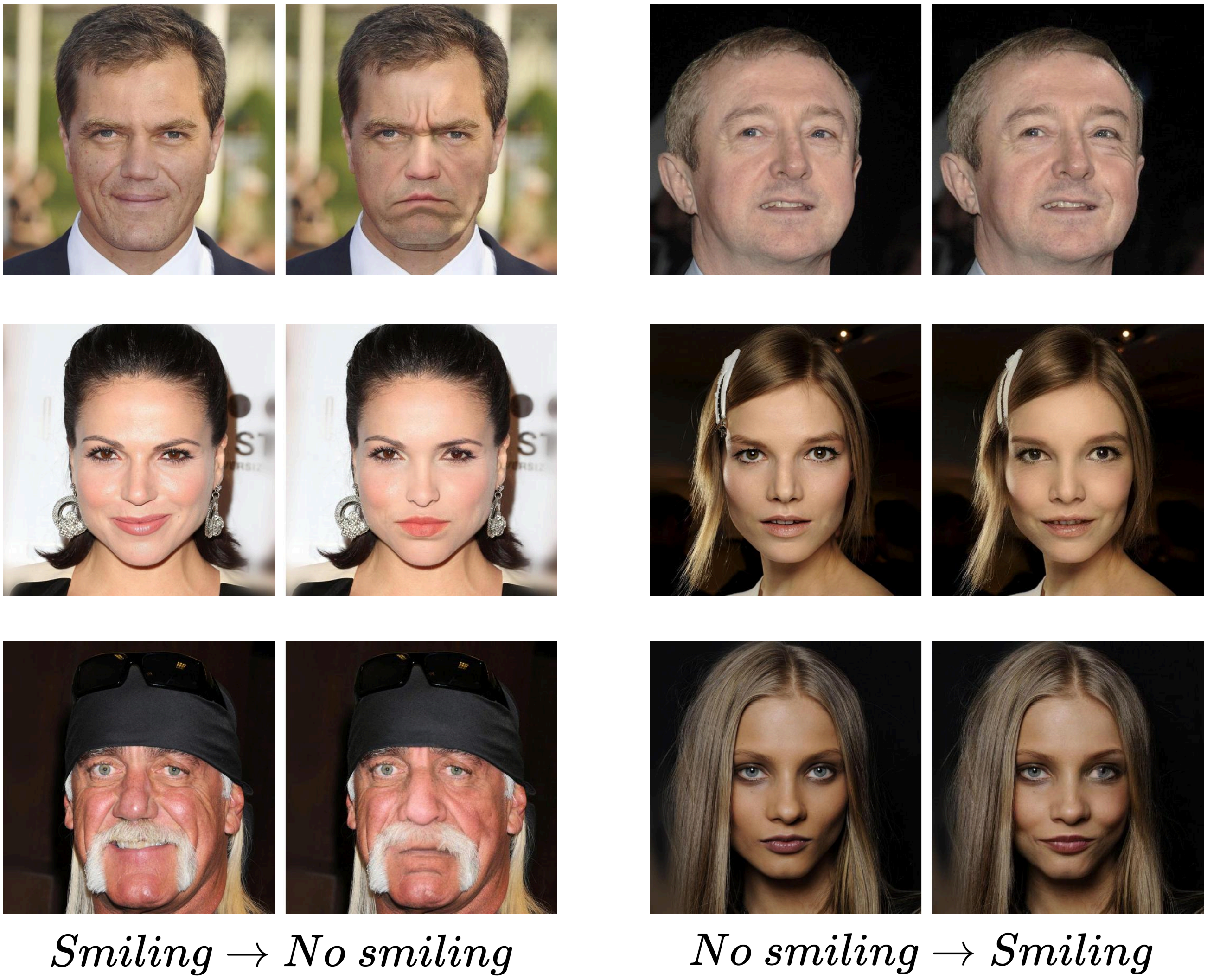

Figure 2. **Qualitative results for 'Smile' attribute with VGG-16.** Left to right: original image, counterfactual image generated by ECED.

$$\mathrm{CD}_q = \frac{1}{N}\sum_{i=1}^{N}\sum_{a\in\mathcal{A}} |c^{q,a}(\mathcal{I}_i^{CF}) - c^{q,a}(\mathcal{I}_i^{F})|. \tag{17}$$

*Counterfactual Transition* (COUT) [5] measures the sparsity of changes in counterfactual explanations. It quantifies the impact of perturbations applied to the factual image $\mathcal{I}^{\mathrm{F}}$ by using a normalized mask $m = \delta(||x^F - x^{CF}||1)$ that represents the relative change compared to the counterfactual image, where $\delta$ normalizes the absolute values to the range $[0, 1]$. The computation of COUT is performed incrementally by gradually inserting the highest-ranked pixel groups from $\mathcal{I}^{\mathrm{CF}}$ based on these sorted mask values.

At each step of adding pixel groups $t \in \{0, \dots, T\}$, the measure calculates the probability $f_{cl}(\cdot)$ for the original label and the desired label, $y^{\mathrm{F}}$ and $y^{\mathrm{CF}}$, through the transition from $x_0 = \mathcal{I}^{\mathrm{F}}$ to $x_T = \mathcal{I}^{\mathrm{CF}}$. From this, the COUT score is defined as:

$$\mathrm{COUT} = \mathrm{AUPC}(y^{\mathrm{CF}}) - \mathrm{AUPC}(y^{\mathrm{F}}) \in [-1, 1]. \tag{18}$$

The perturbation area under the curve for each label $y \in yF, yCF$ is calculated as follows:

$$\mathrm{AUPC}(y) = \frac{1}{T}\sum_{t=0}^{T-1}\frac{1}{2}\left(f_{cl}\left(x_t, y\right) + f_{cl}\left(x_{t+1}, y\right)\right) \tag{19}$$

$$\mathrm{AUPC}(y) \in [0, 1]. \tag{20}$$

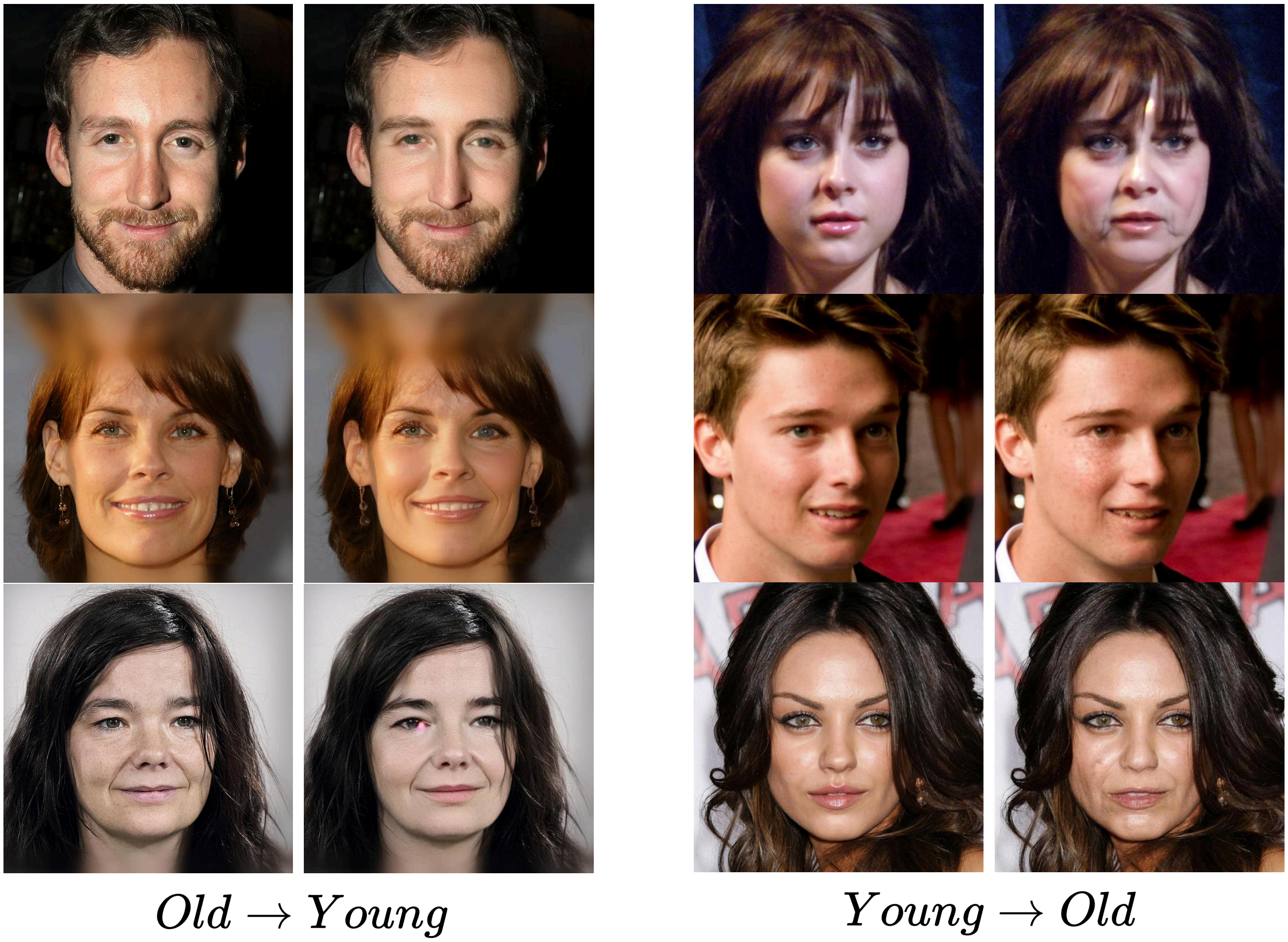

Figure 3. **Qualitative results for ‘Age’ attribute with VGG-16.** Left to right: original image, counterfactual image generated by ECED.

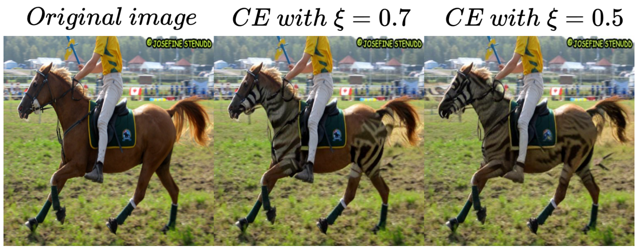

Figure 4. The example of diversity in generating counterfactual explanations by ECED.

## C.2. Validity

*Flip Ratio* (FR) measure is commonly used to assess the authenticity of counterfactual outcomes for the desired label. This criterion focuses on evaluating the validity of $N$ counterfactual explanations by measuring the extent to which the original label $y_i^{\mathrm{F}}$ of the $i$-th original image $\mathcal{I}^{\mathrm{F},i}$ shifts the classification model's prediction to the counterfactual target class $y_i^{\mathrm{CF}}$ for the counterfactual image $\mathcal{I}^{\mathrm{CF},i}$.

$$FR = \frac{\sum_{i=1}^{N} \mathbb{I}\left(f_{cl}(\mathcal{I}^{\mathrm{CF},i}) = y_i^{\mathrm{CF}}\right)}{N}, \tag{21}$$

where $\mathbb{I}$ is the indicator function.

## C.3. Realism

*Fréchet Inception Distance* (FID) assesses the realism of generated images by measuring the FID distance between the distributions of features (extracted via the InceptionV3 network [10]) in the original dataset and the counterfactual image set:

$$\mathrm{FID} = \|\mu_F - \mu_{CF}\|_2^2 + \mathrm{Tr}(\Sigma_F + \Sigma_{CF} - 2\sqrt{\Sigma_F \Sigma_{CF}}), \tag{22}$$

where $\mu_F, \mu_{CF}$ represent the mean vectors, and $\Sigma_F, \Sigma_{CF}$ represent the covariance matrices derived from the feature distributions of the InceptionV3 model for the real and counterfactual image sets, respectively. However, due to the nature of this type of explanation, which creates very subtle changes (differing only at a few pixels), this introduces significant bias. To address this issue, [4] proposed a solution by splitting the dataset to compute FID, termed sFID. The basic idea is to split both the real and counterfactual image sets into two subsets, then compute the FID value across the cross subsets (i.e., real image subsets not paired with their respective real counterparts). Finally, the average of the two FID values is taken to obtain the sFID value.

## References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42(4):1–11, 2023. 2, 3

[2] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021. 3

[3] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 3

[4] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16425–16435, 2023. 6

[5] Saeed Khorram and Li Fuxin. Cycle-consistent counterfactuals by latent transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10203–10212, 2022. 4

[6] Bingyan Liu, Chengyu Wang, Tingfeng Cao, Kui Jia, and Jun Huang. Towards understanding cross and self-attention in stable diffusion for text-guided image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7817–7826, 2024. 3

[7] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 3

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3

[10] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 5