A. Validation for CLIP-guided Editing

Our methodology relies on CLIP-guided fine-grained image editing to provide adequate model diagnostics. It is critical to validate CLIP's capability of connecting language and visual representations. This section proposes two methods for CLIP's validation.

A.1. Visualization for edited images

In this section, we visualize the attribute edit method we employed from StyleCLIP [2].

Effect of λ. Fig. 7 shows the effect of λ in Equation 2 of the main text. Originally in StyleCLIP, this filter parameter (denoted as β in [2]) helps the style disentanglement for editing. Since we normalized the edit vectors which benefits the disentanglement in our framework, λ's effect on style disentanglement is lessened and it is mainly effective on intensity control and denoising.

Single-attribute editing. We show an extensive set of images of attribute editing based on the extracted global edit directions (mentioned in Section 3.2 of the main text) to demonstrate that the method is exactly editing the corresponding attribute. Fig. 9 and Fig. 10 visualize the results. We can observe that, based on the user's input attribute string, the edited image changes only in the attribute direction, while preserving the other attributes.

Multiple-attribute editing. To verify that our approach of editing multiple attributes by linear combination (Equation 3 of the main text) is valid, we show examples of combined edits in Fig 8.

A.2. User study for edited images

To validate that our counterfactual synthesis is effective and fine-grained, we conducted a user study validating two aspects: synthesis fidelity and attribute consistency.

User study for synthesis fidelity. We let the user classify which image is the counterfactual synthesis to verify that no unreal artifacts are introduced during the ZOOM process. Fig. 5a shows sample questions of this study. In theory, the worse case is that users can perfectly recognize the semantic modification and yield a 100% user recognition rate. Inversely, the best case should be that the users cannot recognize any counterfactual synthesis and do random guess to yield a 50% user recognition rate.

User study for attribute consistency. We ask users whether they agree that the counterfactual and original images are consistent on the ground truth w.r.t. the target classifier. For example, during the counterfactual synthesis for the cat/dog classifier, a counterfactual cat image should stay consistent as a cat. Fig. 5b shows another sample questions. The worst case is that the counterfactual changes the ground truth label to affect the target model, which makes the user agreement rate very low (even to zero).

The statistics of the user study are shown in Table 1, where we separate 34 volunteers (at least of undergraduate education level) by two collector links and receive responses from them. The group (i.e., the link clicked) is randomly chosen by the users themselves.

Name of Study	Domain	Group 1	Group 2
Synthesis Fidelity (	FFHQ	62.12	71.79
Recognition Rate ↓, %)	AFHQ	51.30	50.55
Attribute Consistency (	FFHQ	94.12	90.76
Agreement Rate ↑, %)	AFHQ	89.92	88.26

Table 1. User study results. We can see from the table that our counterfactual synthesis preserves the visual quality and maintains the ground truth labels from the user's perspective.

High quality counterfactual images were generated, as evidenced by the fact that users had trouble distinguishing between them. Most users also agree that the counterfactual images do not alter the ground truth with respect to the target classifier, proving that our methodology is producing meaningful counterfactuals. Please take into account that the nature of our recognition system makes human volunteers slightly more sensitive to human faces, therefore we see a little higher recognition rate in the human face (FFHQ) domain than in the animal face (AFHQ) domain.

A.3. Stability across CLIP phrasing/wording:

We notice that the resulting image depends on the prompt wording. In our framework, the neutral phrase (e.g. "a face") is subtracted after CLIP space encoding, to ensure the attribute edit direction is sufficiently unambiguous. Our tests revealed that, as long as the prompts' meanings are descriptive of the object, they will provide comparable outcomes. For example, on the perceived-age classifier, we have got similar sensitivity results on "a picture of a person with X", "a portrait of a person with X", or with other synonyms. Examples are shown in Fig. 2.

B. Additional Results of Model Diagnosis

B.1. Additional counterfactual images

Fig. 1 shows more examples of single-attribute counterfactual images on the Cat/Dog and Perceived Gender classifiers. The output prediction is shown in the top-right corner. It shows that the model prediction is flipped without changing the actual target attribute. In addition to binary classification and key-point detection in our manuscript, we further illustrate the extension of ZOOM counterfactuals on semantic segmentation, multi-class classification, and binary church classifier (BCC) in Fig. 3. Fig. 6 shows more examples of multiple-attribute counterfactual images.

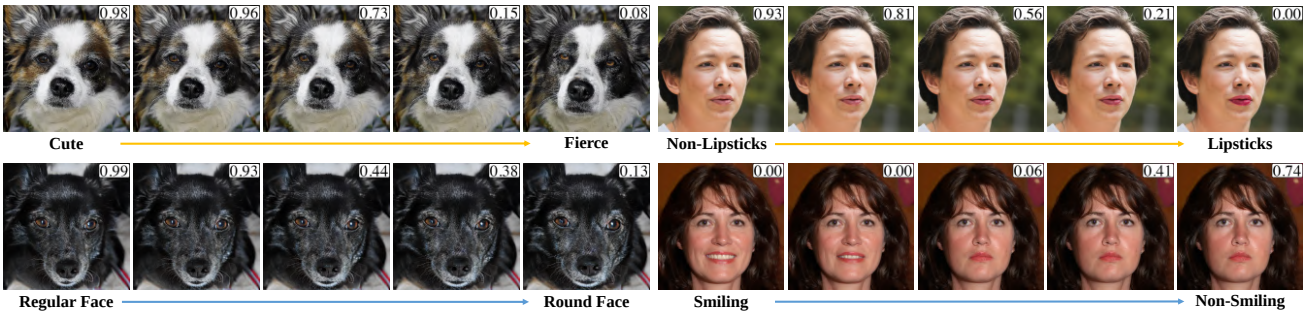


Figure 1. Effect of progressively generating counterfactual images on the Cat/Dog classifier (0-Cat / 1-Dog), and the Perceived Gender classifier (0-Female / 1-Male). Model probability prediction during the process is attached at the top right corner.

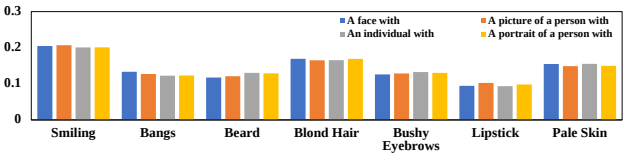


Figure 2. Sensitivity histograms when using four instances of phrases with a similar concept. Zoom in for better visibility.

Figure 3. ZOOM counterfactuals on more tasks (segmentation, multi-class classifier) and additional visual domains (cars, churches). Zoom in for better visibility.

B.2. Additional histograms

Fig. 4 shows more histograms on the classifiers trained on CelebA (top) and the classifiers that are intentionally biased (bottom). The models and datasets are created using the same method described in Section 4 of the main text.

C. Ablation of Optimization Method

When there are multiple attributes (i.e., $N > 1$) to optimize, linearizing the cost function as grid in high dimensional space will help to efficiently approximate convergence in limited epochs. Specifically, we have the option to adopt PGD [1] (i.e., update using $\eta \cdot \text{sign}(\nabla_{\mathbf{w}} \mathcal{L})$) for efficient optimization. We compared generating counterfactuals with and without projected gradients. Table 2 shows the visual quality and flip rate of the generated counterfactuals. We can observe that ZOOM-PGD image quality is finer under Structured Similarity Indexing Method (SSIM) [3], while ZOOM-SGD has a higher flip rate. The images from ZOOM-PGD is finer since the signed method stabilizes the optimization by eliminating problems of gradient vanishing and exploding.

Optimization	Classifier	SSIM (↑)	Flip Rate (%, ↑)
SGD	Perceived Age	0.5732	67.24
	Perceived Gender	0.5815	49.40
	Mustache	0.5971	36.33
PGD	Perceived Age	0.8065	50.19
	Perceived Gender	0.7035	42.84
	Mustache	0.7613	25.10

Table 2. The comparison of counterfactuals generated with stochastic gradient descent (SGD) and projected gradient descent (PGD) method. We can observe that ZOOM-PGD image quality is finer under SSIM (Structured Similarity Indexing Method) [3] metrics, while ZOOM-SGD has a higher flip rate.

Our empirical observation during the experiment is that ZOOM-PGD frequently oscillates around a local minima of edit weights and fails to reach an optimal counterfactual. We hypothesize that the reason of lower flip rates from the signed method is that the edit weight search is constrained on nodes of a grid space (the grid unit length is step-size η), which loses precision and underperforms during counterfactual search.

References

[1] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*, 2018.

[2] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *ICCV*, 2021.

[3] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image Quality Assessment: from Error Visibility to Structural Similarity. In *IEEE TIP*, 2004.

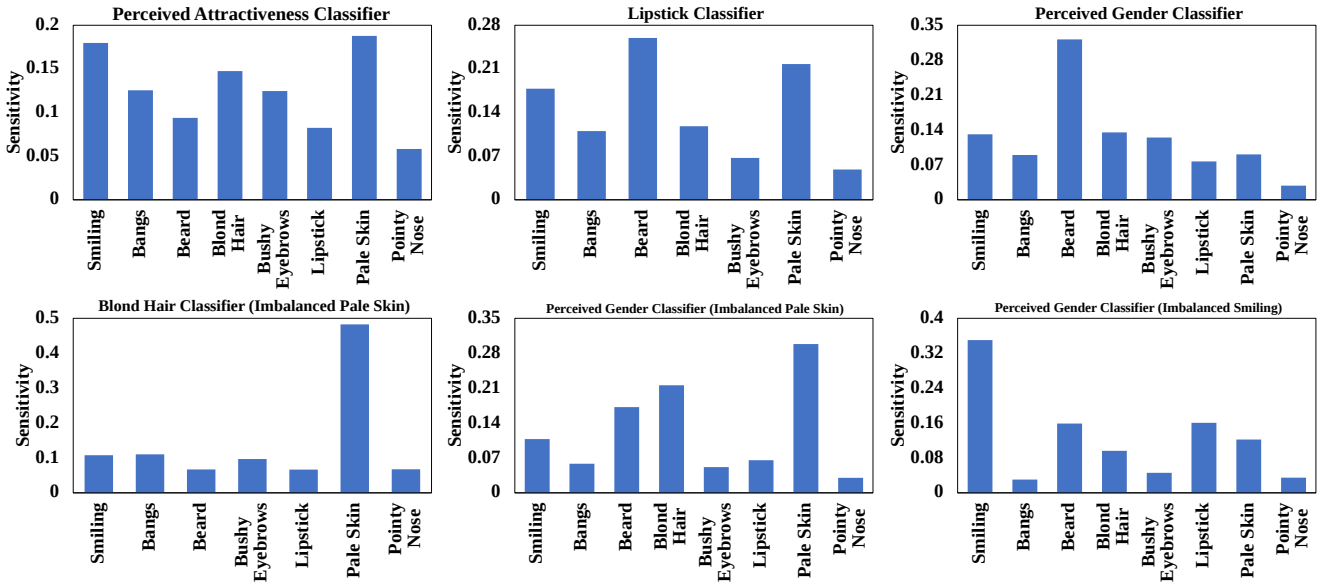
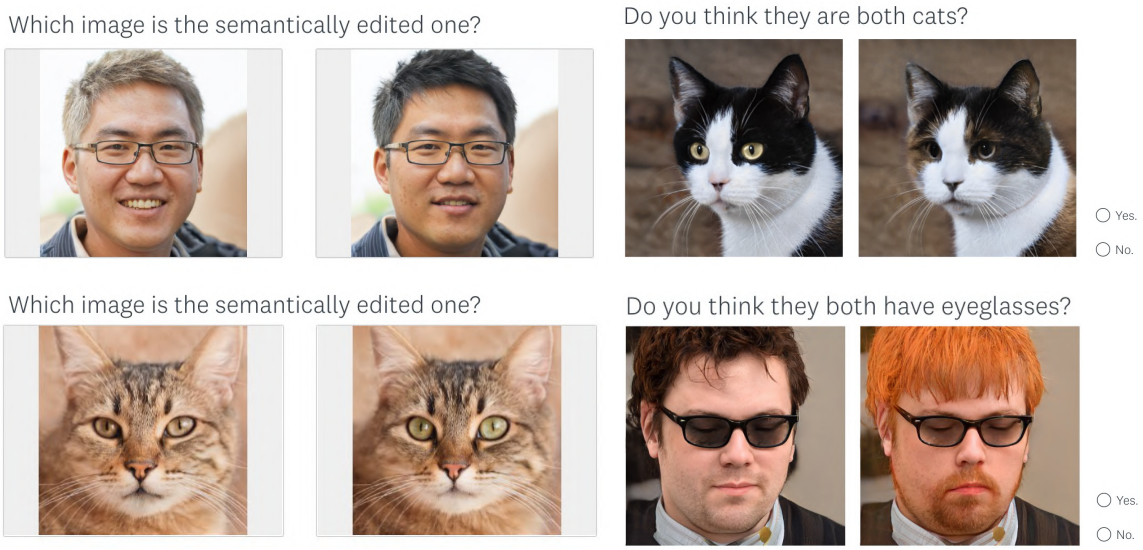


Figure 4. The above histograms show ZOOM on three regularly trained classifiers on CelebA, and the bottom histograms show ZOOM successfully detects the bias in the manually-crafted imbalanced classifiers.



(a) Evaluating visual fidelity. We show two images and let users choose the one that they think is edited. The counterfactuals are generated on Eyeglasses classifier and Cat/Dog classifier.

(b) Evaluating attribute consistency. The user classifies whether the ground truth is flipped. Example of counterfactual images on Cat/Dog classifier and Eyeglasses classifier is shown above.

Figure 5. Sample questions in the user study. Each user answers 10 questions for each of the two user studies.

(a) Multiple-attribute counterfactual for cat/dog classifier.

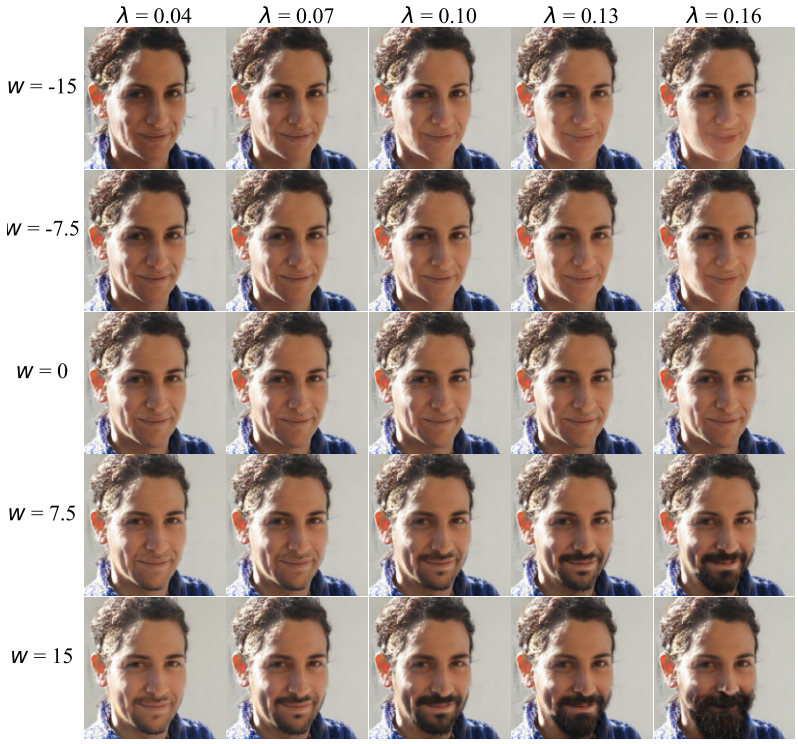
(b) Multiple-attribute counterfactual for eyeglasses classifier.

(c) Multiple-attribute counterfactual for perceived gender classifier.

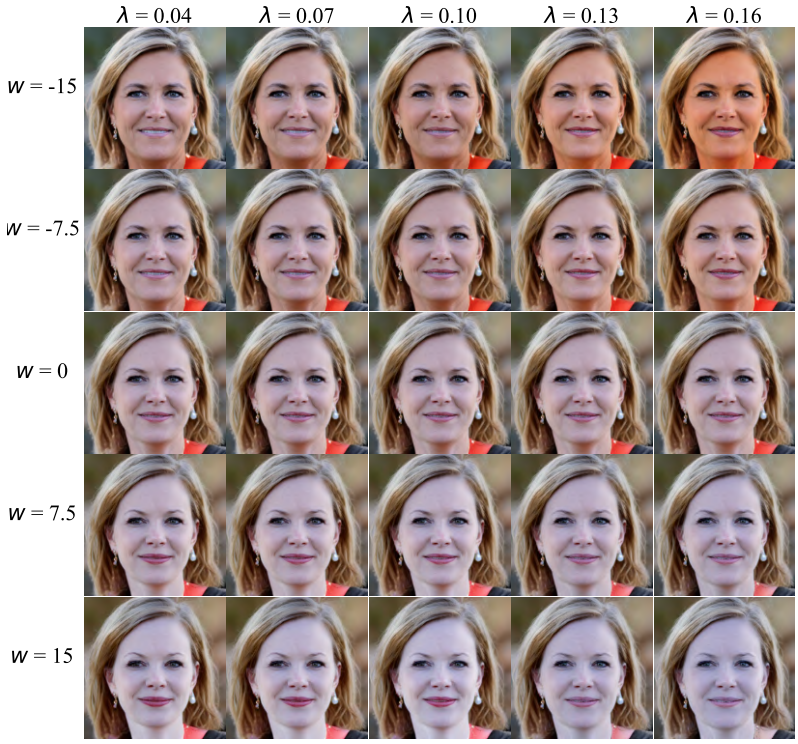
(d) Multiple-attribute counterfactual for mustache classifier.

(e) Multiple-attribute counterfactual for perceived age classifier.

Figure 6. Multi-attribute counterfactual in the human face and animal face domain. The right-up corner of each image records the model output prediction.

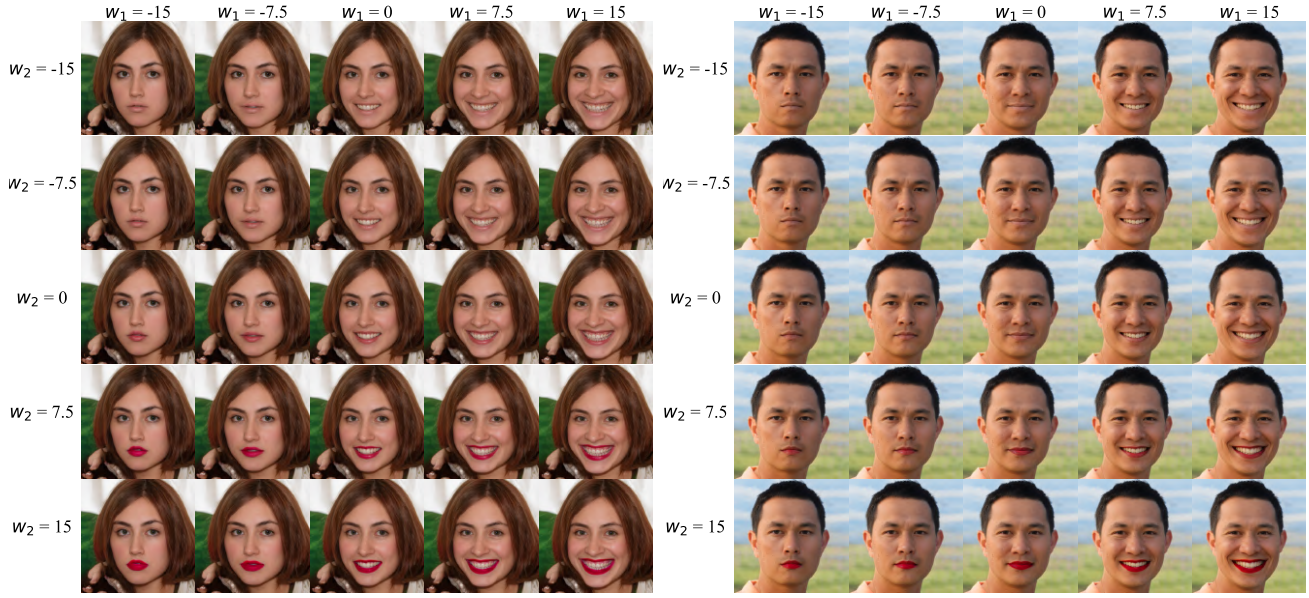


(a) Effect of λ values for editing beard.

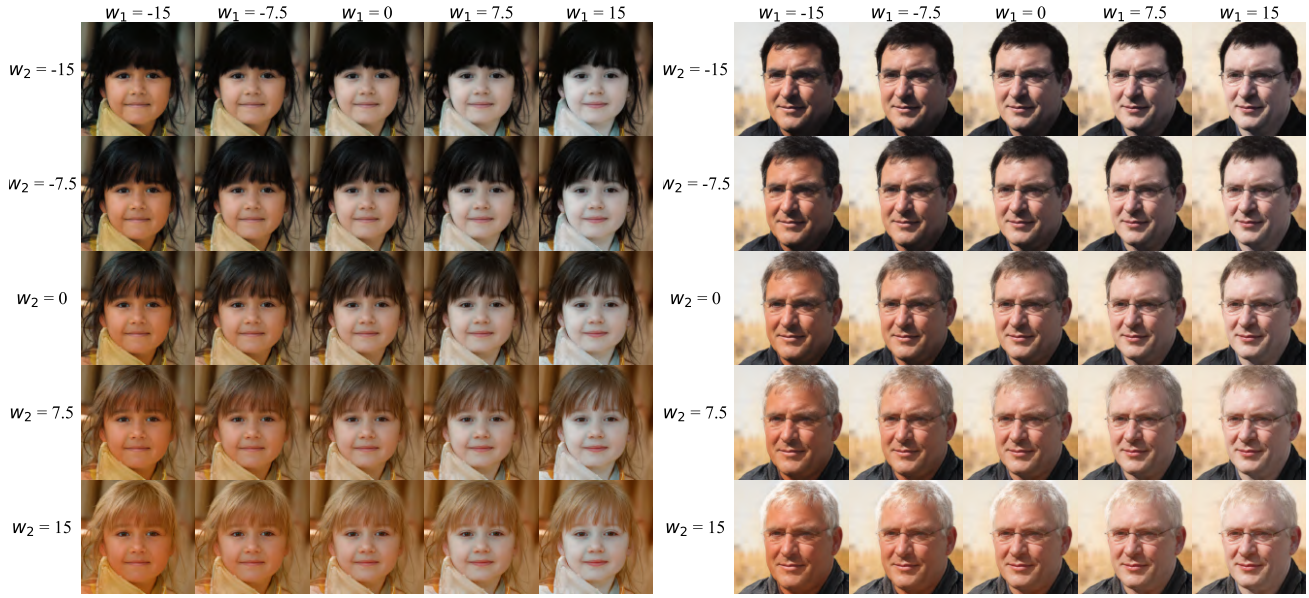


(b) Effect of λ values for editing pale skin.

Figure 7. Visualization of the effect of different λ values.



(a) Combination of smiling (w_1) and lipstick (w_2).



(b) Combination of pale skin (w_1) and blond hair (w_2).

Figure 8. Visualization of traversing on directional (attribute) style vectors to validate the effectiveness of multiple attribute editing.

(a) Attribute editing: a cat with green eyes.

(b) Attribute editing: a cute cat.

(c) Attribute editing: a dog with round face.

(d) Attribute editing: a cute dog.

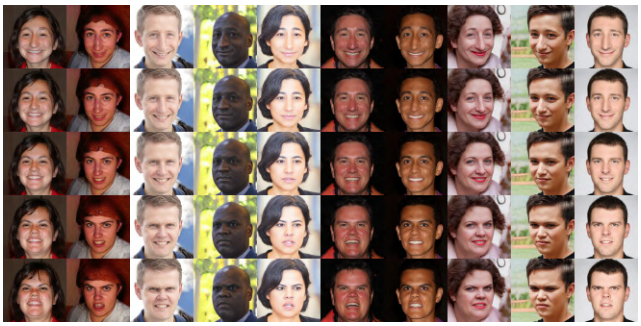
(e) Attribute editing: a cat with round face.

(f) Attribute editing: a cat with pointed ears.

(g) Attribute editing: a dog with open mouth.

(h) Attribute editing: a black dog.

Figure 9. Visualization of global edit directions by utilizing the StyleCLIP channel relevance matrix. Images are sampled from the AFHQ domain using StyleGAN2-ADA. Every column demonstrates an edited image from edit weight $w = -30$ to $w = 30$. Weights of five images are linearly interpolated as $\{-30, -15, 0, 15, 30\}$. We can see that global edit directions are generalizable on multiple images.

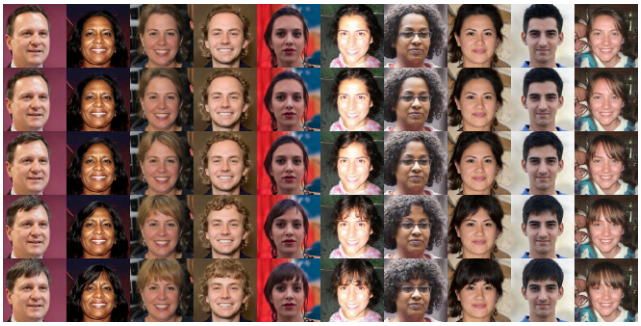


(a) Attribute editing: an angry face.

(b) Attribute editing: a face with eyeglasses.

(c) Attribute editing: a cute face.

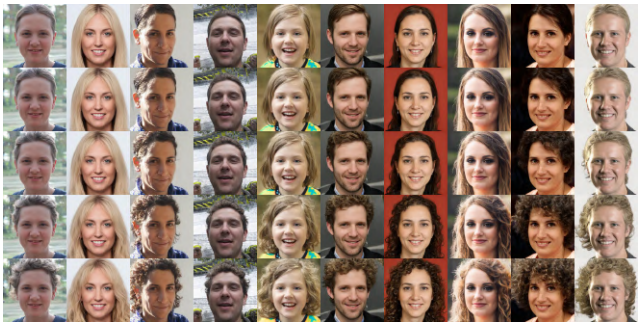
(d) Attribute editing: a face with blond hair.



(e) Attribute editing: a face with bangs.

(f) Attribute editing: a smiling face.

(g) Attribute editing: a happy face.

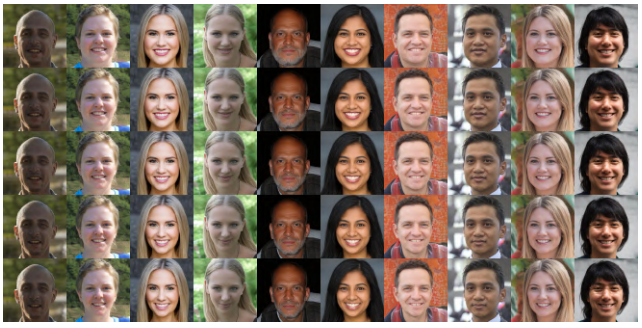


(h) Attribute editing: a face with curly hair.

(i) Attribute editing: a face with beard.

(j) Attribute editing: a face with lipstick.

(k) Attribute editing: a tired face.



(l) Attribute editing: a skinny face.

(m) Attribute editing: a male face.

(n) Attribute editing: a surprised face.

(o) Attribute editing: a face with long hair.

(p) Attribute editing: a face with pale skin.

Figure 10. Visualization of global edit directions by utilizing the StyleCLIP channel relevance matrix. Images are sampled from the FFHQ domain using StyleGAN2-ADA. Every column demonstrates an edited image from edit weight $w = -30$ to $w = 30$. Weights of five images are linearly interpolated as $\{-30, -15, 0, 15, 30\}$. We can see that global edit directions are generalizable on multiple images.