Counterfactual Analysis Driven Unsupervised Image Editing in Stable Diffusion

Ziqian Liu Pietro Gori* Yunlong He*

*Supervisor: pietro.gori@telecom-paris.fr, yunlong.he@telecom-paris.fr

1. Introduction

Understanding the behavior of deep learning models has emerged as a central focus in artificial intelligence research. This is especially critical in image synthesis, where a growing body of work utilizes explanatory techniques to answer why a model generates specific images or features, and which factors might influence the model's outputs. Such insights can aid in enhancing model's controllability, exposing biases, and building more trustworthy generative systems.

One way to understand the behavior of a generative model is through **counterfactual analysis (CF)** [12], which focuses on discovering minimal changes to image features that result in different classifier outputs. Recently, diffusion models [6] have established themselves as the state-of-the-art (SOTA) for image synthesis, outperforming GANs [4] and VAEs [9] in both quality and stability. This has led to increasing efforts to use diffusion models for CF, as seen in several recent works exploring diffusion-based methods for counterfactual explanations [1, 8, 7, 11, 5]. Among them, an elegant strategy is to train a word embedding while keeping a pre-trained text-to-image model fixed—such as TIME [5]. At inference time, the image can be reconstructed by conditioning on the learned text embedding. This approach is computationally efficient, as it avoids retraining the full model and only requires optimization over the embedding space. However, using such framework for downstream model analysis still relies on a trained classifier. Besides, although counterfactual models like TIME can indicate which visual features are sufficient to alter a classifier's decision, they provide limited insight into the generative process itself. Specifically, they do not explain which components of the embedding space are responsible for generating features unrelated to classification (e.g., overall facial structure), and which ones encode class-specific attributes only (e.g., the presence of glasses). This lack of semantic meaning hinders interpretability in image generation.

Contrastive analysis (CA), on the other hand, offers an interesting alternative for analyzing a model's output by discovering the generative factors that are common across two datasets/domains, and distinguishing them from those that are unique to only one. Taking human face datasets as an example, one can split the data into images of faces with glasses and those without glasses. The objective of CA is to identify and separate the salient factors that are distinctive to a single domain (e.g., the presence of glasses) from the common factors shared across both domains (e.g., facial structure). Compared to CF, CA does not require the use of auxiliary classifiers during inference or a "target label" to specify the desired output domain. Instead, it controls the generation process by operating the common and salient factors—where common (e.g., facial structure or lighting) should appear in both datasets,

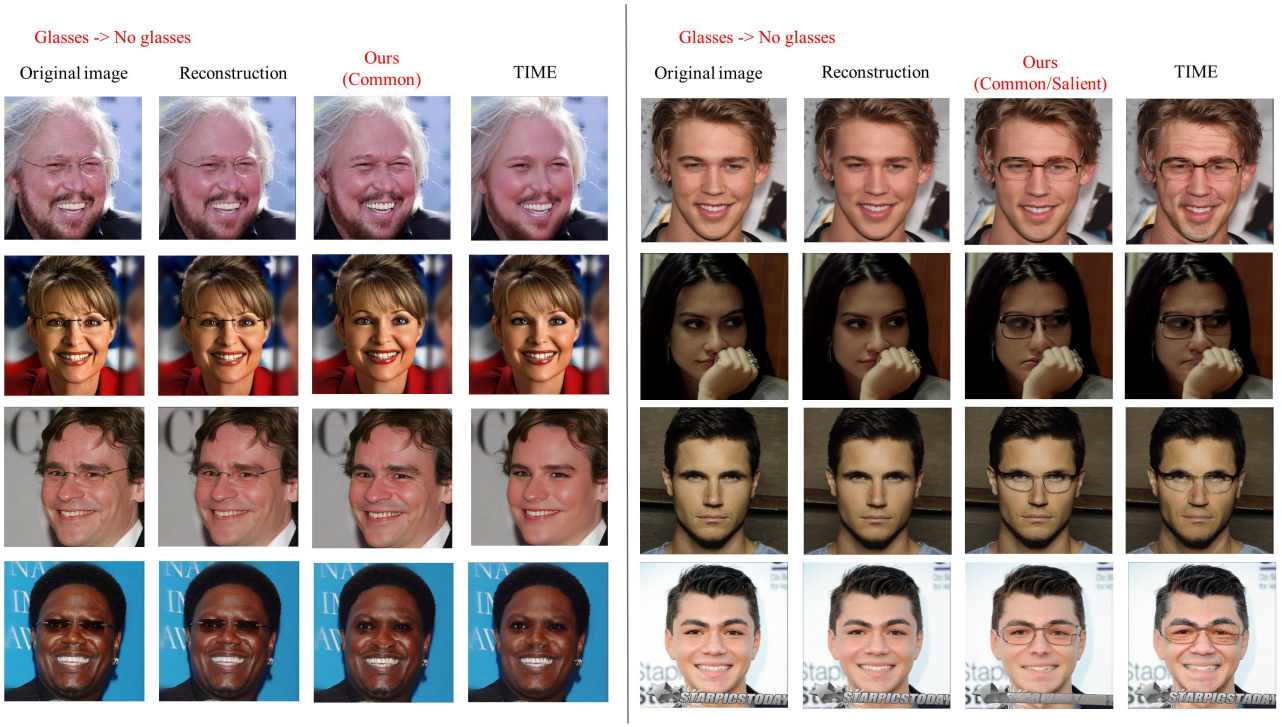

Figure 1: Visual comparison of glasses manipulation results using our method and TIME. Left side shows Glasses → No glasses transformations; right side shows No glasses → Glasses transformations. For each row, we show (from left to right): the original image, its reconstruction (TIME), the result of our method, and the result from TIME. Unlike TIME, which learns embeddings tied to two specific classes (with or without glasses), our method explicitly decomposes the embedding space into common and salient components.

and salient (e.g., the presence of glasses) is unique to one domain. While existing CA techniques [2, 13] remain based on VAE or GANs, with none employing diffusion-based approaches.

In this research, we aim to explore the embedding space (i.e., the conditioning input to diffusion models) through CA, to understand which components are common and which are salient in guiding image generation. Our experiments are based on TIME, a SOTA method that enables learning three new words in the textual embedding space of pre-trained text-to-image diffusion models as conditioning input: a context word, capturing dataset-specific structures (e.g., human face datasets), and two class words, associated with features of specific classes (e.g., faces with glasses vs. without glasses). We first reproduce image reconstruction and modification results using the context and class words learned by TIME across a range of datasets. Then, we modified the model of TIME to learn representations into a common and a salient embedding, enabling generation conditioned explicitly on shared or domain-specific factors.

Overall, our work has the following contributions and innovations:

- We adapted the TIME model, originally developed for CF, as an interesting method for CA in diffusion-based generative modeling.
- We compare the results of CF and CA within the context of diffusion models, revealing their strengths and weaknesses in understanding the model's generation.
- We provide experimental results on different datasets using TIME, ranging from face images and medical imaging data. The results show that the adapted CA framework can produce interpretable outputs while eliminating the need for classifiers and text prompt in the whole process.

2. Background

This section introduces the basic concepts of detail involved in generating the counterfactual results in this paper, in the following order.

2.1 Training of Context, Common and Salient bias

In the original article, the authors point out that since models are subject to multiple biases during the learning process, there are two main types of bias that need to be trained.

• Context Bias: It mainly reflects the overall context structure of all images in the dataset, such as scene type and other information. We need to train it based on all the images in the dataset, so my understanding is that the overall general information of the dataset is all embedded into the weight matrix of CLIP's Embedding, so much so that the subsequent Class-0 and Class-1 embeddings can be more focused on learning the information unique to each.

• Class Bias: is primarily class-specific and relates to the fact that classifiers rely on semantic information when making decisions. For example, the classifier divides the dataset into Class-0 and Class-1, and then trains the respective CLIP weight matrices based on the above Context bias. Therefore, its goal is to learn specific semantic features related to the class predicted by the classifier for reflecting salient features in a specific class of images.

In the original article, the flow of training Context, class-0 and class-1 is shown in Figure 2. Where Class-0 and Class-1 are independent of each other, but both are trained based on Context(frozen)

To embed text into the CLIP visual space, the authors adopt a textual inversion approach to distill both the context bias and the target classifier's knowledge into the textual embedding space of Stable Diffusion. In a nutshell, textual inversion [3] links a new text-code c^* and an object (or style) such that when this new code is used, the generative model will generate this new concept. To achieve this, Gal et al. [3] proposed to instantiate a new text embedding e^*, associate it to the new text-code c^*, and then train e^* by minimizing the loss.

$$\mathbb{E}_{(x,C)\sim\mathcal{D},\, t\sim U[1,T],\, \epsilon\sim\mathcal{N}(0,I)}\left[L(x,t,\epsilon,C)\right] \tag{1}$$

Here, D is the set of images containing the concept to be learned, U is the uniform distribution of natural numbers between 1 and T, and C is a text prompt containing the new text code C^*.

To train pseudo-word embeddings for a given an image-text pair (x, C), each optimization step minimizes the reconstruction loss.The final trained pseudo-word will definitely capture the distribution of the data of the image.

$$L(x,\epsilon,t,C) = \|\epsilon - \epsilon_\theta(x_t(x,t,\epsilon),t,C)\|^2 \tag{2}$$

with

$$x_t(x,t,\epsilon) = \sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\epsilon. \tag{3}$$

In general, the training process used in the original TIME framework is illustrated in Fig. 2. However, this approach may not be suitable for CA, as it learns the Class-0 and Class-1 vectors independently without explicitly separating common from salient factors.

To address this limitation, we introduce three training strategies for learning common and salient components, as shown in Fig.3. For example, given a human face dataset with images with glasses (target) and without glasses (background), the strategies are as follows:

1. **Strategy 1 (Column 1 in Fig. 3):** Train the `"common"` vector using background (no-glasses) images. Then, freeze the `"common"` vector and continue training a `"salient"` vector to generate target (glasses) images.

2. **Strategy 2 (Column 2 in Fig. 3):** Similar

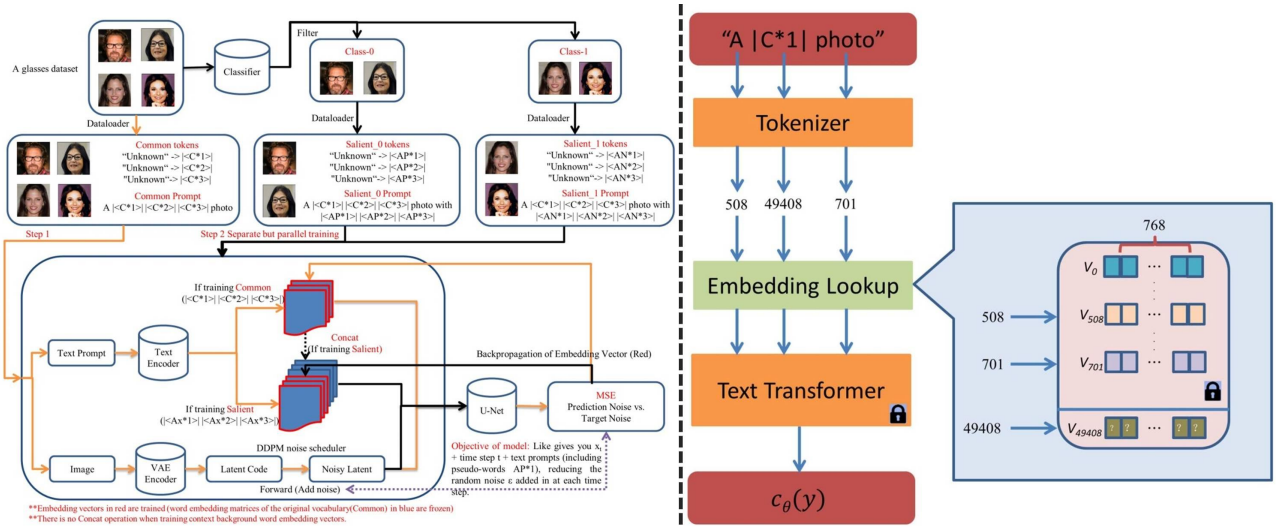

Figure 2: The training process of Context and Class bias. For right graph, which is the o utline of the text-embedding and inversion process. A string containing our placeholder word is first converted into tokens (*i.e.*, word or sub-word indices in a dictionary). These tokens are converted to continuous vector representations (the "embeddings", v). Finally, the embedding vectors are transformed into a single conditioning code $c_\theta(y)$ which guides the generative model. We optimize the embedding vector v_* associated with our pseudo-word S_*, using a reconstruction objective.

to Strategy 1, but during the second phase, the `"common"` vector is **not frozen**, allowing it to be updated while training the `"salient"` vector.

3. **Strategy 3 (Column 3 in Fig. 3):** Train both `"common"` and `"salient"` vectors jointly to generate both target and background images. The `"salient"` vector is explicitly *forced to approximate zero* (by minimizing L2 loss)when generating background images, ensuring it captures only class-specific information.

Note that the phases for training context are all 'A |C*1| Picture'. In 1) and 3), the training phases for Common are 'A |C*1| Picture showing |AN*1|' and the training phases for Salient are 'A |C*1| Picture showing |AN*1| and |AP*1|'. In addition in 2), the phase for training Common and Salient is both 'A |C*1| Picture showing |AN*1| and |AP*1|'. The above pseudo-word embeddings, such as AN*1 and AP*1, are based on Equation 3 and are learnt during image reconstruction. The specific learning process is shown in right graph of Fig.2.

2.2 Inversion Generation of DDPM

Formally, in formula (4), given a diffusion model ϵ_θ and a fixed set of steps T, ϵ_θ takes as input a noisy image x_t, the current step t to compute a residual shift, and a textual conditioning C in our case. For the generation, ϵ_θ updates x_t following, and the diagram is also shown below.

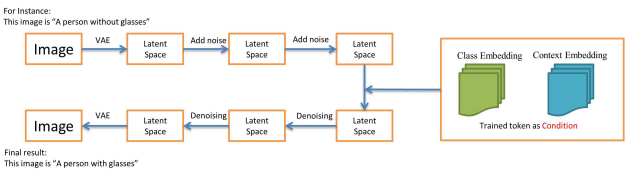

Figure 4: Schematic diagram of addition and removal of noise in DDPM

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t, C)\right) + \sigma_t\epsilon, \quad (4)$$

where α_t, α_t and $\bar{\alpha}_t$ are some predefined constants, and ϵ and x_T are extracted from a Gaussian distribution. This process is repeated until $t = 0$. To train a DDPM, for a given image-text pair (x, C), each optimization step minimizes the loss as follows.

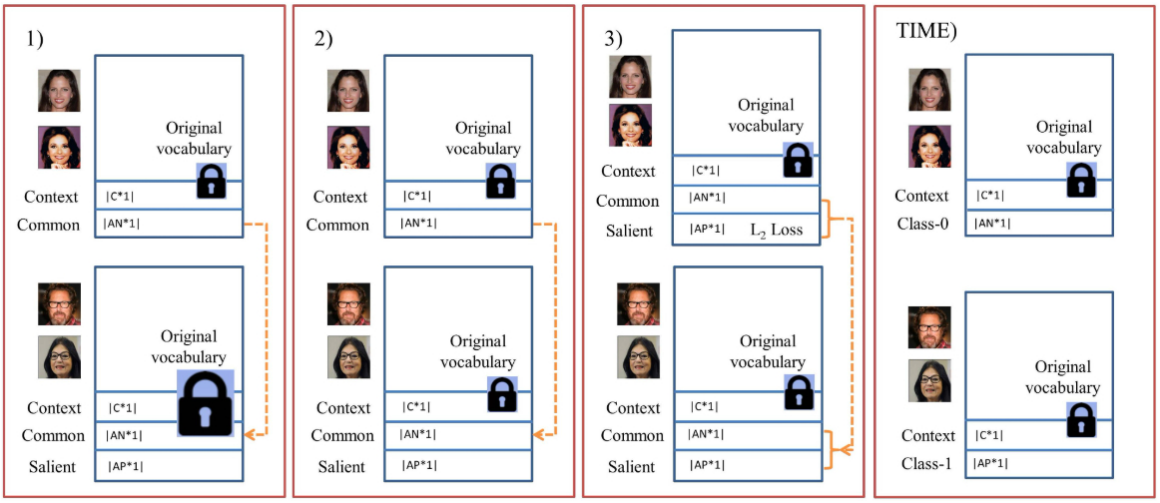

Figure 3: Comparison of three adapted encoders for CA (columns 1–3), alongside the original text encoder from TIME (column 4). The embedding vectors "C" (context), "AN" (common), and "AP" (salient) are optimized to guide the generative process during training. The original vocabulary embeddings are kept fixed (frozen) throughout the entire training process. Note that TIME's process can also be considered as a form of CA, where the "class-1" vector implicitly combines both common and salient factors, rather than separating them explicitly.

Most importantly, a latent diffusion model (coding the image into the latent space followed by noise addition and denoising) is used in the TIME article to reduce the computational cost by performing the diffusion process in the latent space of the quantised self-encoder. The above trained pseudo-word text (|C*1|, |AP*1|, |AN*1|) is then used to guide the editorial diffusion process. The specific process is shown in the following figure.

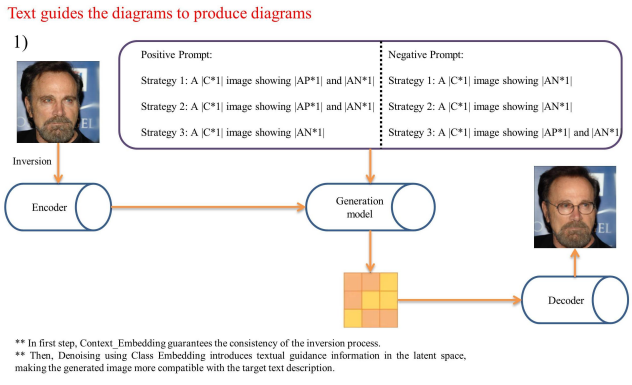

Figure 5: The generation process of Stable Diffusion. The serial numbers correspond to the positive and negative cue words entered for the different training modalities

The Classifier-Free Guidance [10] is then also invoked here to enhance the generation quality by linearly combining the predictions with and without the text condition to strengthen the impact of prompt. The specific principle is shown in Equation 5.

The CFG [10]'s core modifies the sampling strategy in Eq.2 by replacing ϵ_θ with ϵ_θ^f, a shifted version defined as follows:

$$\epsilon_\theta^f(x_t, t, C) := (1 + w)\,\epsilon_\theta(x_t, t, C) - w\,\epsilon_\theta(x_t, t, \varnothing) \quad (5)$$

where $\varnothing$ is the empty conditioning and w is a weighting constant.

As can be seen from Eq. 5 enhances conditional generation by introducing weights between conditional and unconditional generation. When w is higher, the model relies more on the conditional generation part; on the contrary when w is lower, the model relies more on the prediction when there is no prompts, and the generated image is more likely to be based on the prior distribution of the training data, thus generating features with higher frequency in the training data. So the constraint mechanism of CFG ensures that the model generates high quality images that meet the objective

even when the prompts are ambiguous or invalid.

3. Experiment

To evaluate the effectiveness of the Common and Salient embeddings trained under different strategies in the contrastive analysis above, we conducted experiments based on the "glasses" attribute in the CelebA-HQ dataset.

And also to evaluate the validity of this counterfactual method, this report uses several datasets for evaluation ***(as shown in appendix)***, including CelebA-HQ (local conversion), Dogs and cats (category conversion), Street Cleanliness and littered(global conversion), and popular medical datasets BraTS and BraTS2023(Healthy and Tumor class). The workflow of this study's model is relatively straightforward, primarily leveraging pre-trained models (e.g., Stable Diffusion, CLIP, VAE, etc.).

Whatever the CA or CF, we still based our approach on Stable Diffusion V1.4. For all dataset, we trained three textual embeddings for the context and class biases for 3000 iterations with a learning rate of 0.01, a weight decay of $1e^{-4}$, and a batch size of 64. For the inference, we used the default EDICT's hyperparameter $p = 0.93$ and a total of 50 steps.

4. Main result

Here we experimented with the three Common/Salient training approaches we proposed above based on glasses attribute in CelebA-HQ. Experiments on all of the following datasets for the methods used in this paper consist of these three steps: (Step 1) We learn a context token for the whole dataset using textual inversion. (Step 2) Learning Common and Salient embeddings separately based on different classes of images. (Step 3) Finally, to generate the counterfactual interpretation, we edit the image to denoise it using the target embedding.

Here, we found that the best performance of Common and Salient was trained using the 3rd training way, i.e., training Common was based on Context(frozen), and then both Common (loaded with the last training result) and Salient were trained based on Context.The inverse result graph is shown below.

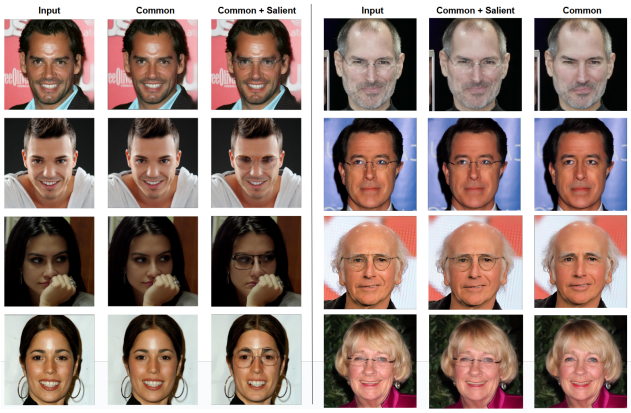

Figure 6: Qualitative results generated by the adapted CA framework. The model successfully identify "common" and "common + salient" factors, enabling controlled image generation by operating in the embedding space.

Of course we also used Common and Salient trained by the first and second strategy to generate the images, but the images either failed to be inverted or had too many artifacts, so I consider this a failure. The result of the failure is shown below.

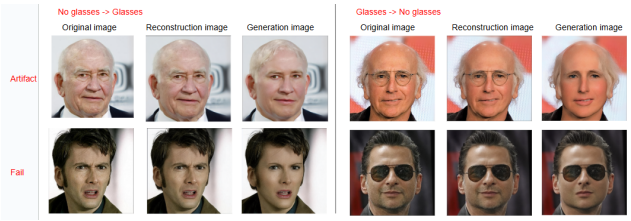

Figure 7: Using Common and Salient to interpret the Inversion result (For 1st training way, when need add glasses, positive prompt="Context+Common", negative prompt="Context+Common+Salient". When need remove glasses, positive prompt="Context+Common+Salient", negative prompt="Context+Common". For 2rd training way, When need add glasses, positive prompt="Context+Common+Salient", and negative prompt="Context+Common". When need remove glasses, positive prompt="Context+Common", and negative prompt="Context+Common+Salient"))

By looking at the generated results, most of the resultant images generated by the inversion of 1st and 2st strategy have the result of inversion failure and the rest have the problem of heavy artifacts. Then here we also used Common and Salient trained under different strategies to invert the edited image. The following figure shows the result of editing on the same image.

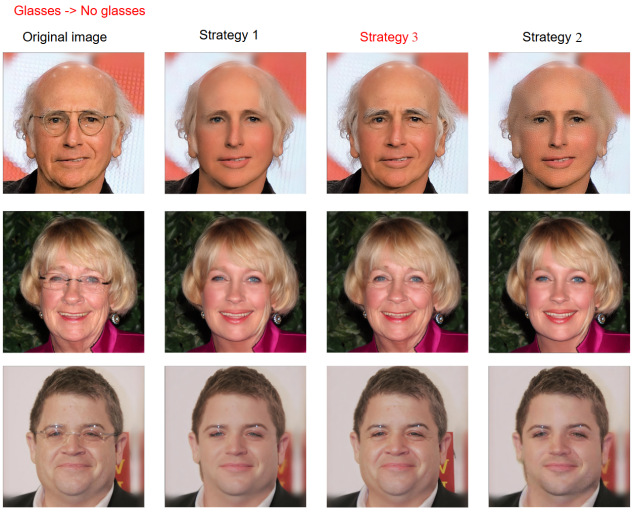

Figure 8: The results of image editing using Common and Salient representations obtained under different training strategies.**(None of strategy 1 and strategy 2 even succeeded in editing the sample results for No glasses -> glasses.)**

From the above figure, it is evident that the three training strategies differ significantly in their ability to remove the salient factor—glasses. While Strategy 1 and Strategy 2 succeed in removing glasses in some cases, they often suffer from blurred facial details, loss of identity, or noticeable artifacts, indicating insufficient disentanglement of representations. In contrast, Strategy 3 consistently removes glasses while preserving common factors such as facial structure, expression, and lighting. The generated images appear more natural and realistic, with higher identity consistency. Overall, Strategy 3 demonstrates superior controllability and disentanglement capability, validating its potential to more effectively model Common and Salient factors in contrastive analysis tasks.

5. Conclusion

In this work, we build upon the ideas proposed in the original TIME paper and integrate a contrastive analysis perspective to propose a diffusion-based CA framework. This framework extends the original textual inversion scheme to learn a triad of pseudo-word embeddings—Context, Common, and Salient—thus eliminating the need for class-specific black-box classifiers. Moreover, we follow the embedding learning procedure introduced in TIME to enable pseudo-word embedding training without requiring explicit prompt inputs.

Compared with prior approaches that either condition Stable Diffusion on manually crafted cues or rely on a pre-trained classifier to locate discriminative directions, our method achieves a fully self-contained paradigm: Context captures the shared visual bias of the entire dataset, Common models class-agnostic appearance factors, and Salient isolates the class-specific cues that drive a decision change. Among the three alternative training strategies we investigated, the sequential regimen that first freezes Context, then jointly refines Common and Salient (Strategy 3) consistently yielded the highest-fidelity inversions, while Strategies 1–2 suffered from mode collapse and severe artefacts.

Crucially, all qualitative samples in the Appendix were generated exclusively with TIME's original "Context + Class" setting and therefore serve as a baseline rather than evidence of our approach. A visual comparison between the main-text CelebA-HQ glasses results (ours) and the Appendix (TIME) highlights three advantages of the proposed embeddings: higher reconstruction fidelity, finer control over attribute transfer, and a marked reduction in over-smoothed textures and colour shifts. The ability to perform attribute manipulations even when the positive prompt is replaced by an "Unknown" token confirms that the learned latent codes alone encode sufficient semantic content to guide the diffusion trajectory.

Taken together, the reported results demonstrate that By explicitly disentangling domain-invariant and domain-specific factors in the embedding space, we remove the need for auxiliary classifiers and prompt engineering at train and inference time and obtain a self-contained, more interpretable generation pipeline, marking a step toward scalable and interpretable generative explainability. Future work could be deepened along the following path: introducing cascading or multi-scale comparison targets to capture finer-grained saliency factors.

A. Extra Qualitative Results

Here, we provide editorial results using the training and generation logic from the original TIME article to process diverse datasets. Then here I'll present some of my findings and insights in using TIME technology.

Firstly shown below is an ablation experiment. The second column is the result when only Context is used as Positive prompt in image inversion generation, the third column is the result when only Class is used as Positive prompt, and the last column is the result when the two embeddings(Context+Class) work together.

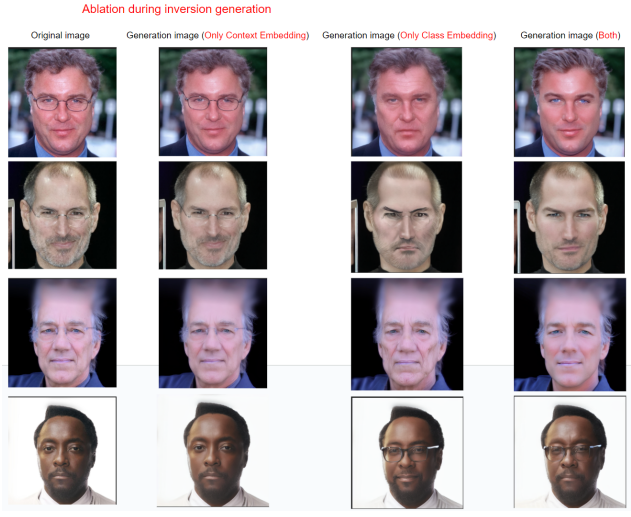

Figure 9: Abalation result (The explicit prompts for training Context bias is *"Centered, Realistic, Celebrity"*. The explicit prompts for training Class bias is *"Bare face, No glasses, Clear eyes"*)

From the figure above, we can observe that when only the Context bias is used as the positive prompt (column 1), the image does not successfully add or remove glasses. However, it still preserves most of the original visual features. This suggests that the Context bias primarily captures general information such as background context and lighting conditions.

When only the Class bias is used as the positive prompt (column 2), the image inversion does take place. However, as seen in the second and third rows, the generated results often suffer from artifacts and visual distortions. This indicates that using only the Class bias may lead to instability in the image inversion process.

In contrast, the third column, which uses both embeddings together, achieves the target inversion successfully without introducing unwanted changes elsewhere in the image.

Additionally, it is important to note that we tested the generation results resulting from inputs without explicit semantic prompts. In the third column of each graph is the result. They have the word 'Unknown' in both training process of Context and Class.

A.0.1 CelebAHQ dataset (local conversion)

In this dataset, we used two attributes (Glasses and Smile) separately for the experiment. When classifying the Glasses attribute, the DenseNet121 classification does not work well. However for this counterfactual interpretation task, the accuracy of the classifier is crucial. As shown in Fig. 1, since we need to filter the images when subsequently training Class bias embeddings, we need to train separate embeddings based on different classes so that the proprietary features of each class can be captured. So I replaced it with Yolov8 (which is a target detection model along the lines of the model already trained in Research Project 1. Here it is sufficient to select the candidate box category with the highest confidence as the final classification result) for

classification, and the metrics comparison results are shown below.

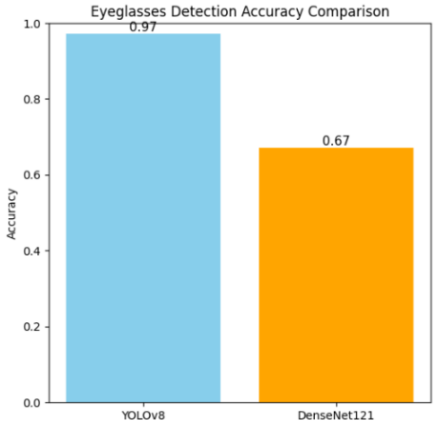

Figure 10: Classification accuracy comparison

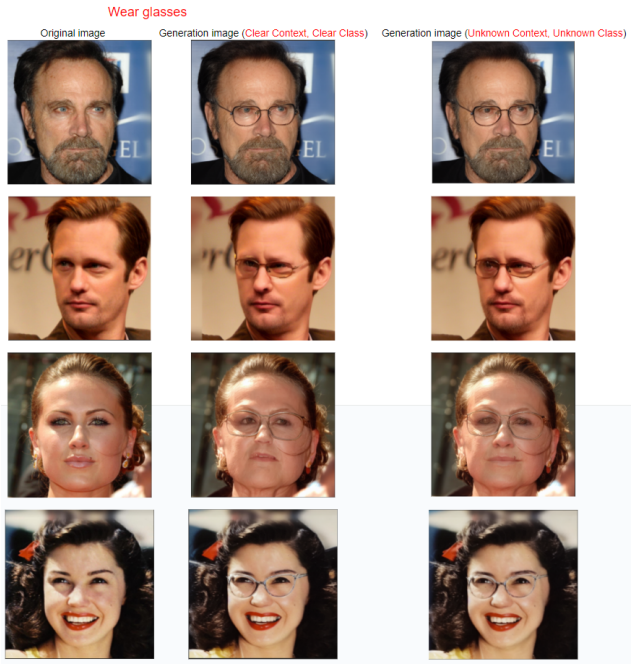

Figure 11: The result of the inversion of No Glasses -> Glasses (The explicit prompts for training Context bias is *"Centered, Realistic, Celebrity"*. The explicit prompts for training Class bias is *"Glasses, Spectacles, Eyewear"*)

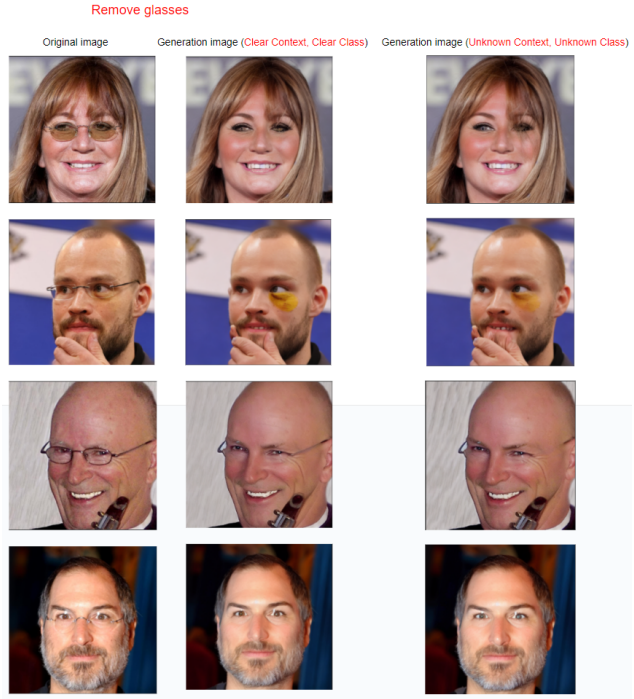

Figure 12: The result of the inversion of Glasses -> No Glasses (The explicit prompts for training Context bias is *"Centered, Realistic, Celebrity"*. The explicit prompts for training Class bias is *"Bare face, No glasses, Clear eyes"*)

By carefully observing the above figure, it can be found that Unknown and other unambiguous semantics as prompts can also generate good results. Secondly, there are some minor differences in the inverse results generated by the Diffusion denoising with and without explicit semantic prompts, for example, in the third line of the result graph for glasses, the pupil colour of the woman's eyes is different and the direction of gaze is different, etc. In the third line of the result graph for glasses removal, only the lenses are removed, whereas the result generated by the explicit prompts is only the lenses are removed. In the third line of the glasses removal graph, the explicit prompts removes only the lenses, whereas Unknown as a prompts removes not only the lenses but also part of the frames.

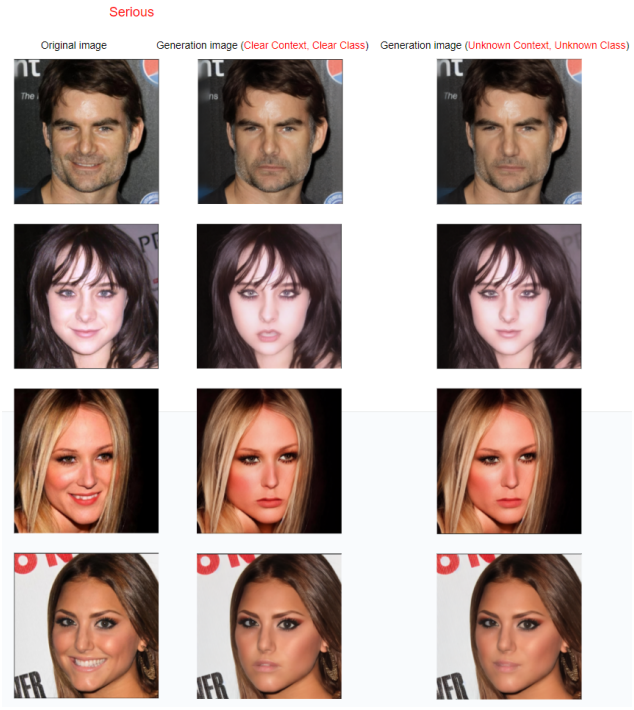

Figure 13: The result of the inversion of Smile -> Serious (the explicit prompts for training Context bias is *"Centered, Realistic, Celebrity"*. The explicit prompts for training the Class bias is *"Serious, Serious, Serious"*)

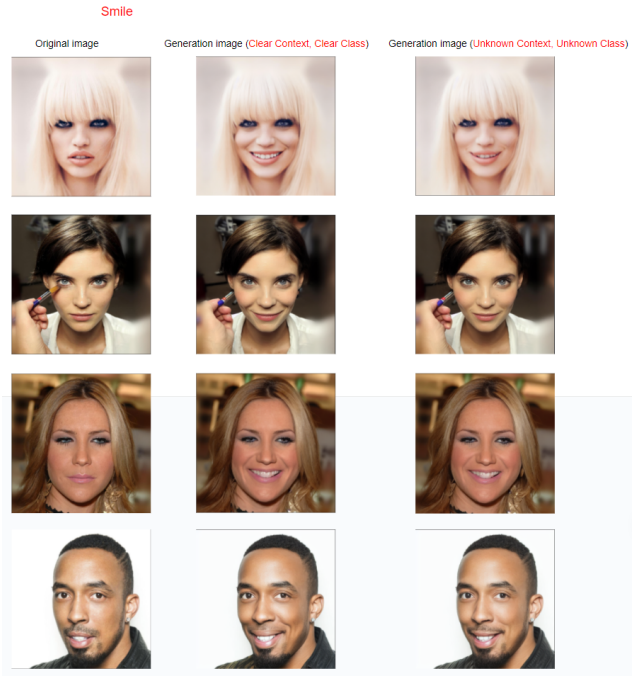

Figure 14: The result of the inversion of Serious -> Smile (the explicit prompts for training Context bias is *"Centered, Realistic, Celebrity"*. The explicit prompts for training the Class bias is *"Smile, Smile, Smile"*)

By looking at the above figure, the phenomenon that occurs is almost identical to the "Glasses" property. In order to evaluate the robustness and stability of the algorithmic framework, as well as to assess whether no semantic words such as "Unknown" can be used as prompts to successfully generate the target results, we continue to conduct experiments on other datasets using the same training procedure and hyper-parameter configurations.

A.0.2 Dog and Cat dataset (category conversion)

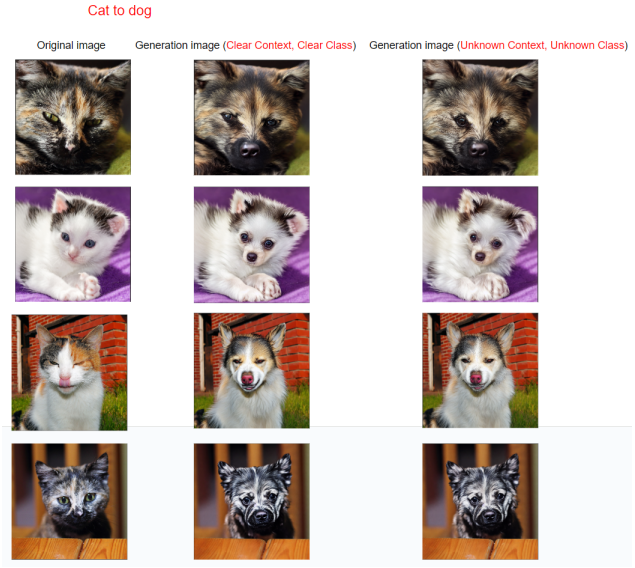

Figure 15: The result of the inversion of Cat -> Dog (the explicit prompts for training Context bias is *"Centered, Realistic, pet"*. The explicit prompts for training the Class bias is *"Floppy ears, Muscular, Canine"*)

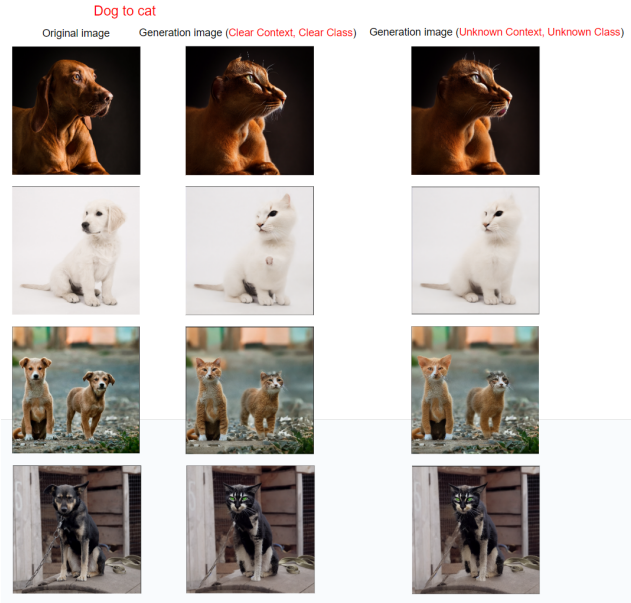

Figure 16: The result of the inversion of Dog -> Cat (the explicit prompts for training Context bias is *"Centered, Realistic, pet"*. The prompts for Class are also *"Pointed ears, Slender, Cat"*)

A.0.3 Clean and littered dataset (global conversion)

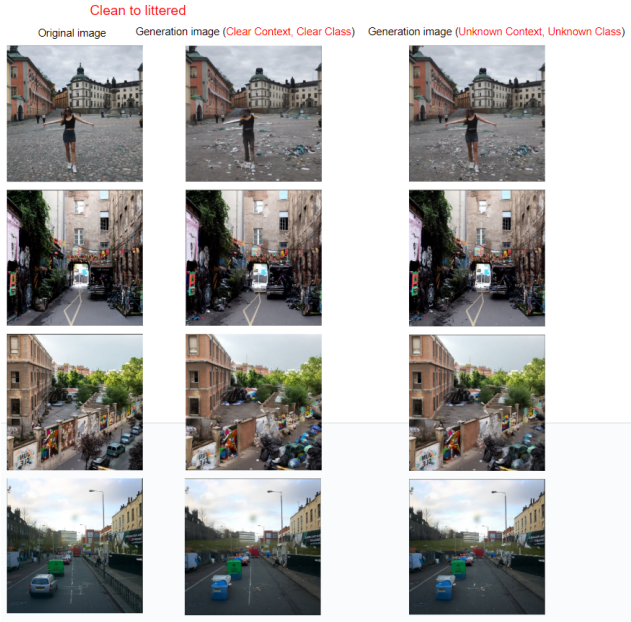

Figure 17: The result of the inversion of Clean -> Littered (the explicit prompts for training Context bias is *"Scenery, Realistic, vivid"*. The prompts for Class is *"Trash, Littered, Dirty"*)

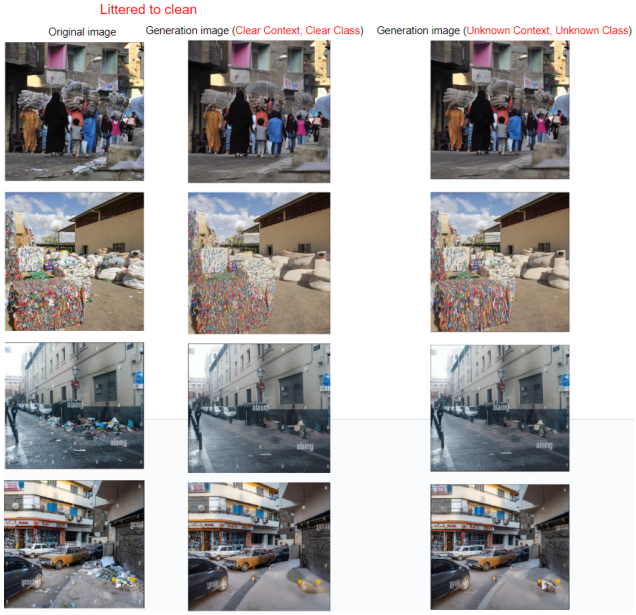

Figure 18: The result of the inversion of Littered -> Clean (the explicit prompts for training Context bias is *"Scenery, Realistic, vivid"*. The explicit prompts for training the Class bias is *"Tidy, Fresh, Orderly"*)

A.0.4 BraTS dataset

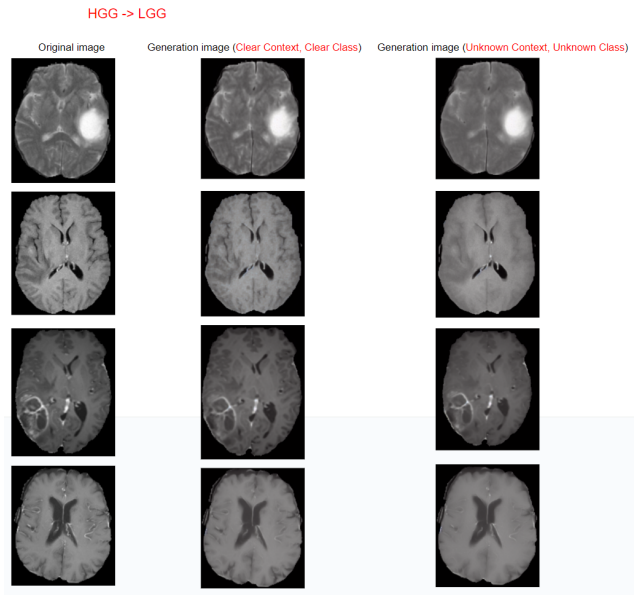

Figure 19: The result of the inversion of HGG -> LGG (the explicit prompts for training Context bias is *"Centered, Realistic, Anatomical brain scan"*. The prompts for Class = *"LGG" are "Well-defined, Homogeneous, Minimal Edema"*)

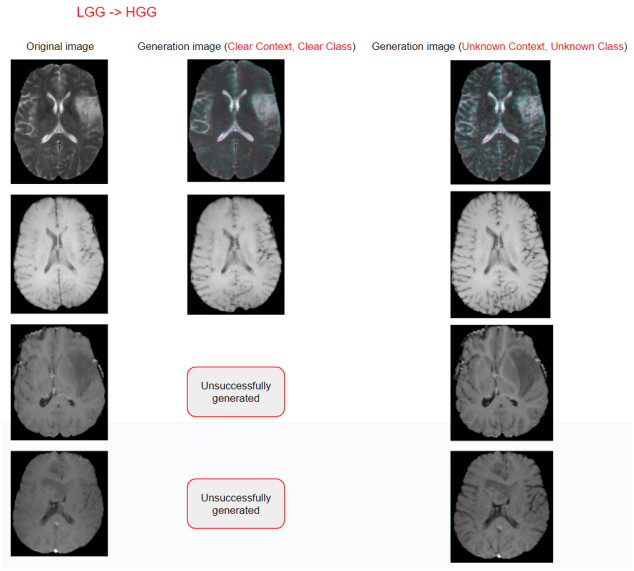

Figure 20: The result of the inversion of LGG -> HGG (the explicit prompts for training Context bias is *"Centered, Realistic, Anatomical brain scan"*. The explicit prompts for training the Class bias is *"Necrosis, Edema, Heterogeneity"*)

It can be observed from the above image that the generated LGG fake image is less and lighter in texture compared to the original HGG image. However, for LGG->HGG, the result is not satisfactory, that is, it cannot even generate the correct image under clear prompts. And for the fake HGG image generated compared to the original LGG image, its texture wrinkles become deeper and more obvious. So eventually we will be able to tell which features the classifier is relying on to make decisions.

A.0.5 BraTS2023 dataset

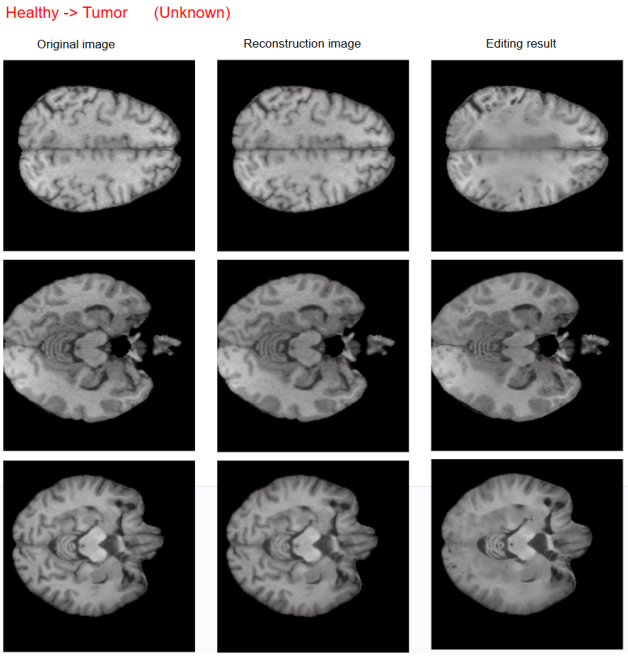

Figure 21: The result of the inversion of Healthy -> Tumor (Pick randomly)

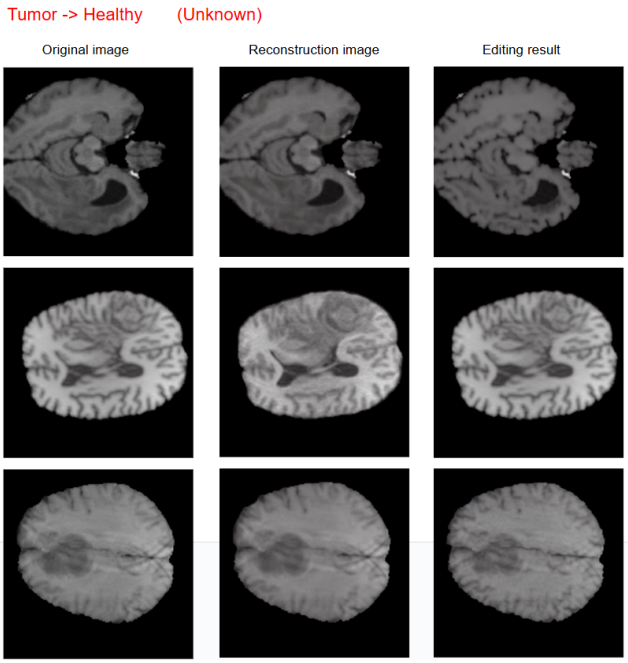

Figure 22: The result of the inversion of Tumor -> Healthy(Pick randomly)

References

[1] Maximilian Augustin et al. “Diffusion visual counterfactual explanations”. In: *Advances*

in Neural Information Processing Systems 35 (2022), pp. 364–377.

[2] Florence Carton, Robin Louiset, and Pietro Gori. "Double infogan for contrastive analysis". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 172–180.

[3] Rinon Gal et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion". In: 2022. URL: https://arxiv.org/abs/2208.01618.

[4] Ian J Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[5] Frédéric Jurie Guillaume Jeanneret Loïc Simon. "Text-to-Image Models for Counterfactual Explanations: a Black-Box Approach". In: *arXiv preprint arXiv:2306.15415* (2023). URL: https://arxiv.org/abs/2309.07944.

[6] Jonathan Ho, Ajay Jain, and Pieter Abbeel. "Denoising diffusion probabilistic models". In: *Advances in neural information processing systems* 33 (2020), pp. 6840–6851.

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

[8] Guillaume Jeanneret, Loïc Simon, and Frédéric Jurie. "Diffusion models for counterfactual explanations". In: *Proceedings of the Asian conference on computer vision*. 2022, pp. 858–876.

[9] Diederik P Kingma, Max Welling, et al. *Auto-encoding variational bayes*. 2013.

[10] William Peebles and Aleksander Holynski. "Classifier-Free Diffusion Guidance". In: *arXiv preprint arXiv:2207.12598* (2022). URL: https://arxiv.org/abs/2207.12598.

[11] Pedro Sanchez and Sotirios A Tsaftaris. "Diffusion causal models for counterfactual estimation". In: *arXiv preprint arXiv:2202.10166* (2022).

[12] Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Counterfactual explanations without opening the black box: Automated decisions and the GDPR". In: *Harv. JL & Tech.* 31 (2017), p. 841.

[13] Ethan Weinberger, Nicasia Beebe-Wang, and Su-In Lee. "Moment matching deep contrastive latent variable models". In: *arXiv preprint arXiv:2202.10560* (2022).