Supplementary Materials
Latent Drifting in Diffusion Models for Counterfactual Medical Image Synthesis

Yousef Yeganeh[1,2*], Azade Farshad[1,2*†], Ioannis Charisiadis[1], Marta Hasny[1], Martin Hartenberger[1], Björn Ommer[2,3], Nassir Navab[1,2], Ehsan Adeli[4†]

[1]Technical University of Munich, Munich, Germany
[2]Munich Center for Machine Learning, Munich, Germany
[3]Ludwig Maximilian University of Munich, Munich, Germany
[4]Stanford University, Stanford, CA, USA

{y.yeganeh,azade.farshad,ge83cid,marta.hasny,martin.hartenberger,nassir.navab}@tum.de,
b.ommer@lmu.de, eadeli@stanford.edu

1. Algorithm

The reverse process of diffusion models generates new samples from pure Gaussian noise distribution $\epsilon \sim N(\mu, \sigma)$. This distribution has fixed mean μ and σ during training and inference time. Based on the Counterfactual assumption Eq. (5), if we aim to use a pre-trained model that has learned a particular distribution, we must find an x' that resembles x but can produce different labels or, here, different samples. In diffusion models, we show that ϵ can have a similar function, and by finding a new ϵ', we could generate samples with a domain drift compared to where we used ϵ. To find the appropriate ϵ', we used Monte Carlo sampling to match the mean μ_x and standard deviation σ_x of generated images from the diffusion model ($\mathcal{D}_\theta$) to the mean and standard deviation of the target dataset (($\mathcal{D}_{\mathcal{GT}}$)), and with that, we could find the best *LD* setting (δ) to ensure the generated samples belong to $\mathcal{D}_{\mathcal{GT}}$ as well. Fig. 10 demonstrates the comparison to pretraining and basic fine-tuning during the training and inference.

Algorithm 1: Optimizing δ for Diffusion Models

Input: Dataset $\mathcal{D}_{\mathcal{GT}}$, Pretrained Conditional Diffusion Model M, List of δ_i values
Output: Optimal δ_i
for δ_i *in List of* δ_i *values* **do**
  Fine-tune M;
  Generate images using $p_{\theta'}(c)$ conditioned on δ_i;
  Calculate mean μ_{gen} and standard deviation σ_{gen} of the generated images;
  Calculate mean μ_y and standard deviation σ_y of the training set;
  if μ_{gen} *and* σ_{gen} *are closer to* μ_y *and* σ_y **then**
    Move δ_i in the same direction;
  end
  else
    Move δ_i in the opposite direction;
  end
  Re-initialize M
end
Select the best-performing δ_i based on the convergence;

2. Theoretical Analysis

Interpolative Nature of Diffusion Models In diffusion models, the denoising process is represented as $x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \epsilon$ where x_t is the noisy image at timestep t, x_0 is the original image, and ϵ is Gaussian noise. By shifting ϵ's mean by δ, we bias the entire trajectory towards the target domain, increasing stability and control of the inference process. This shift effectively adapts the learned feature space of the pre-trained model to the new domain without extensive retraining.

Empirical Evidence In Fig. 2 in the main paper, we observe divergence in both pixel and latent spaces without LD, indicating that even small noise shifts cause significant disturbances in the output distribution. This is evidenced by the green and purple images at both ends of the generated spectrum. With LD, we achieve a more stable and controllable noise distribution that remains resilient to secondary changes. This stability is crucial for counterfac-

*Equal Contribution
†Project Lead

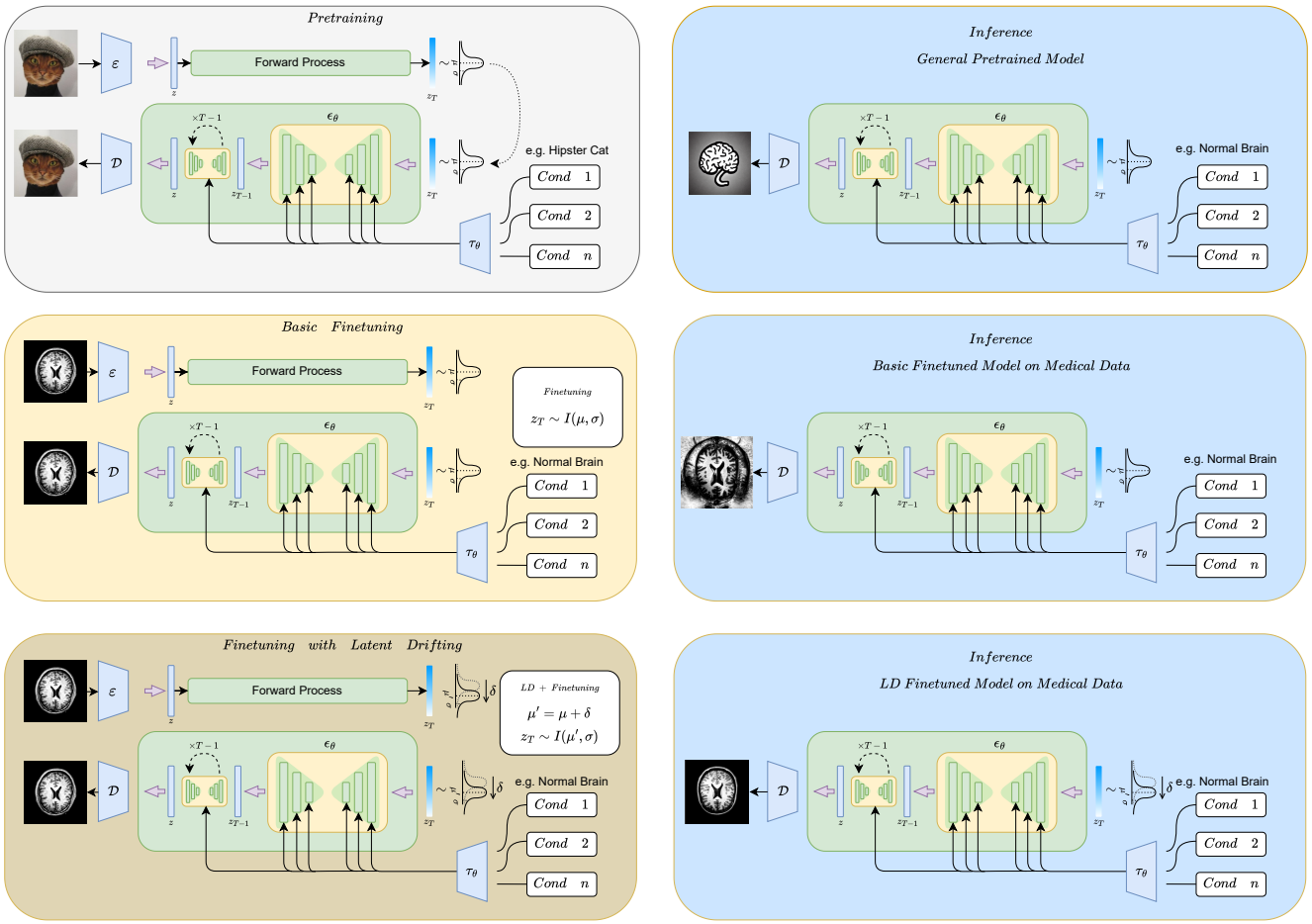


Figure 10. **An overview of Latent Drifting.** Training and Inference compared to Pretraining and Basic FT.

tual image generation, allowing more precise control over prompt or image conditioning. It also limits the inherent diversity in diffusion models, which is favorable for counterfactual image generation. We have demonstrated the effectiveness of this method across various datasets, including Brain MR, Chest X-ray (Main Paper), Faces, Retinal images, and Histopathology datasets.

Optimization Approach We formulate δ selection as minimizing a function $f(\delta) = E[d(G(z+\delta), X_{GT})]$, where G is the generator, z is the latent variable, and X_{GT} represents target domain samples.To avoid complex optimization, we use a surrogate function $s(\delta) = ||\mu_G(\delta) - \mu_T||_1 + ||\sigma^2_G(\delta) - \sigma^2_T||_1$ based on the L1 norm on pixels' values. This surrogate function provides a computationally efficient way to estimate domain adaptation quality, focusing on the critical color distribution shift between source and target domains. The results on the Face, Retinal, and Histopathology datasets show that this method works in datasets with similar color distribution to the source dataset.

Theoretical Guarantee Assuming probabilistic Lipschitz continuity of both f and s, we can provide an approximation guarantee: $|f(\delta^*_s) - f(\delta^*)| \leq C \times (L_f + L_s) \times \epsilon$ where δ^* and δ^*_s are global minimizers of f and s respectively, L_f and L_s are Lipschitz constants, and ϵ is the maximum discrepancy between f and s. This guarantee ensures that our simplified optimization approach yields near-optimal results for the true objective function.

3. Distribution Analysis

In Figs. 11 and 12, we show how *LD* can effectively shift the distribution of generated images in the spatial domain. This is particularly suitable in medical imaging, where the images often follow certain patterns and textures, such as bony structures or soft tissues. In Fig. 11, the mean μ and standard deviation σ of data are represented by circles, where the center is μ and the radius is σ. It is shown that the fine-tuning techniques improved the performance compared to the Basic fine-tuning of Stable Diffusion, but combined with *LD*, they can represent the distribution of the target dataset much more efficiently. In Fig. 12, we present the

impact of different LD values (δ) on the fine-tuning to show how they can be optimized to generate samples more similar to the target distribution.

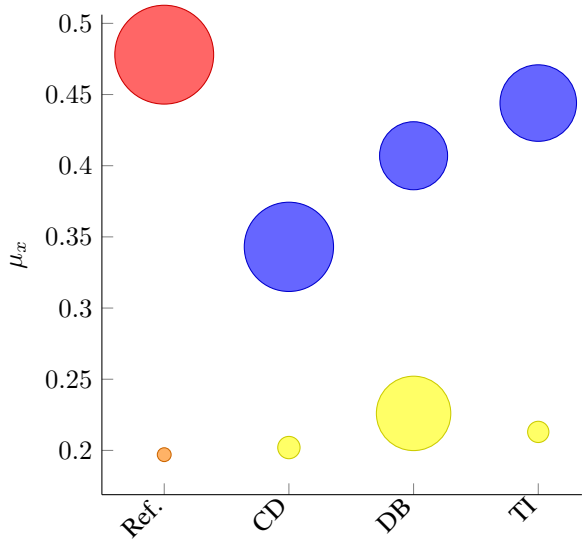


Figure 11. Distribution of generated image pixels using different methods with *LD* and without *LD*. The x-axis shows the mean of the pixels, and the bubble size shows the standard deviation of the pixels. : Real Data ($\mu = 0.198, \sigma = 0.013$), : Stable Diffusion + Basic FT w/o LD ($\mu = 0.478, \sigma = 0.093$), Ref.: Reference, CD: Custom Diffusion [10], DB: DreamBooth [16], TI: Textual Inversion [6].

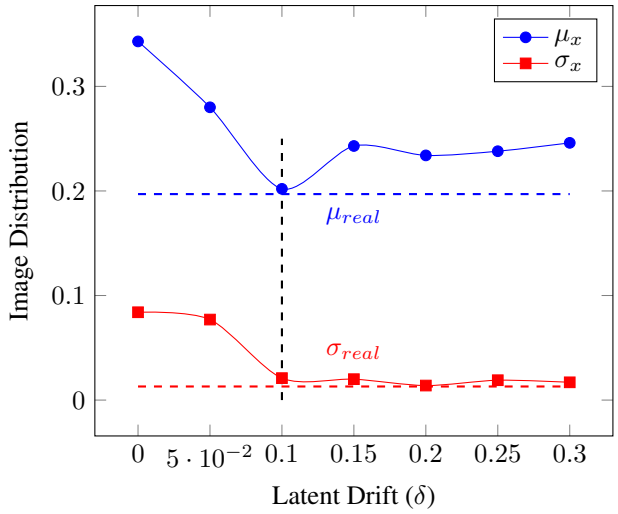


Figure 12. Image pixel distribution (mean and standard deviation of image pixels) with different *LD* values δ using SD + CD [10] + *LD* on Brain MRI generation.

4. Experimental Setup

4.1. Datasets

We utilize the Chexpert [7] dataset and two longitudinal brain MRI datasets, ADNI-1 [17], and OASIS-3 [11]. CheXpert [7] is a large dataset containing 224,316 chest radiographs of 65,240 patients with disease classification labels. ADNI-1 [17] is a longitudinal, multicenter dataset for the early detection and tracking of Alzheimer's disease (AD) with 400 subjects with early mild cognitive impairment (MCI), 200 subjects with early AD, and 200 normal control subjects. OASIS-3 [11] includes PET and MR clinical data for 1098 participants collected in 15 years from 605 cognitively normal adults and 493 individuals at various stages of cognitive decline ranging in age from 42 to 95 years. Over 2000 MR sessions with their metadata on age, sex, and diagnosis were extracted from both datasets. The included diagnoses are Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD). The ADNI and OASIS datasets have custom licenses under ADNI and OASIS websites, respectively. CheXpert is under CC BY 4.0.

CheXpert Preprocessing For CheXpert [7], we sample 100 AP-view radiographs for the following four categories: No Finding (Healthy), Cardiomegaly, Pleural Effusion, and Pneumonia as in [5]. By following the same settings as in [5], each radiograph is cropped to non-zero borders, the longest edge is resized to 512, and the aspect ratio is kept fixed. Finally, the image is zero-padded to a resolution of 512×512 px.

Brain MR Preprocessing To extract similar 2D slices from all brain MRIs, we rigidly register the MRIs to a 1mm isotropic MNI template and crop them to a common size of 256x256x192 using UniRes [2]. For our experiments, we used the 97th axial slice of the processed MRIs. Background noise was removed from the slices using a threshold, and the image intensities were scaled to 0 and 1. Scans with poor contrast or incorrect affine matrix were excluded from the experiments. The final dataset consisted of 3269 scans (414 AD, 634 MCI, 2214 CN) from 1461 patients. From these scans, 200 samples (100 AD, 100 CN) were separated as a test set. Additionally, 2658 image pairs were generated. Each pair includes a younger image as the source and an aged image as the target. 2045 pairs are used for training and 613 for testing. The pairs are generated together with editing prompts, which include the patient's information on sex, age, and diagnosis.

Implementation Details We tuned the learning rate and the number of steps required to fine-tune each method. DreamBooth [16] is fine-tuned for $1.5K$ steps for each concept ($3K$ in total), using a batch size of 1 and a learning rate of $2e{-}6$. Custom diffusion [10] is trained for $1.5K$ steps, with

a learning rate of $1e-5$, and a batch size of 1. Textual inversion [6] is trained for $1K$ steps per concept, with a learning rate of $5e-4$, a batch size of 1, and a vector embedding representation of 64. The Stable Diffusion basic fine-tuning was done with a learning rate of $1e-5$ and a batch size of 16. The best overall performance for fine-tuning Stable Diffusion was achieved after $10K$ epochs. The results with different values of epochs are presented in Fig. 13. For all of these models, we have used a single 24GB NVIDIA RTX-4090 GPU. The training phase for each finetuning model took roughly 2 hours for a single *LD* parameter and inference roughly half and hour for a batch of 100 images.

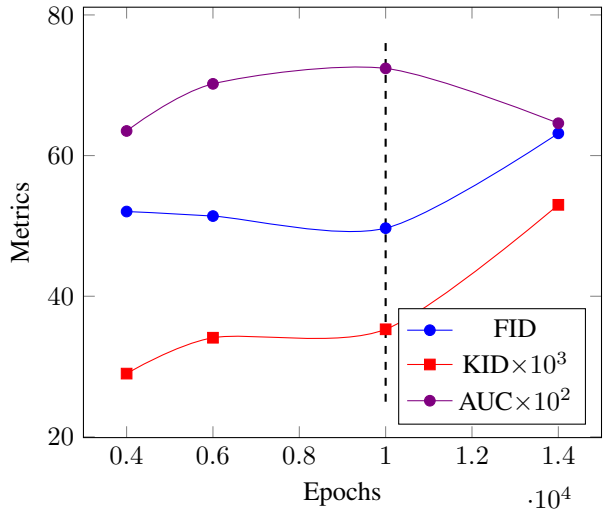


Figure 13. Counterfactual Brain MR generation performance using SD + Basic FT + *LD* at different epochs.

5. Additional Results

5.1. Domain Expert Study

We have conducted a study involving 3 medical and neuroimaging experts, each evaluating 400 images (200 synthetic and 200 real). The results are summarized in Tab. 7, where sensitivity indicates the detection rate for synthetic images, and specificity represents the accuracy in identifying real images. Interestingly, while participants excelled at recognizing real images, they found it challenging to identify synthetic ones. Participants reported that some synthetic images were detectable due to their 'overly perfect' appearance and symmetry, whereas real images were often distinguishable by their inherent artifacts and imperfections. This feedback suggests that our synthetic images closely mimic real data, with occasional instances of idealized features that may hint at their artificial nature.

Table 7. Domain Expert Study Results

Sensitivity	Specificity	Accuracy
20.53%	86.00%	54.19%

5.2. Other domains

We extend our experiments to adapt a model trained on the LAION-5B dataset to one general domain dataset on faces and two non-radiology medical datasets. We show that different data distributions can have different optimal values of δ. For example, $\delta = -0.05$ produces the best FID and KID to adapt the pretrained model to the CelebHQ dataset. At the same time, $\delta = 0$ produces cartoonish images, and with $\delta = 0.1$, despite producing realistic looking faces, the images contain some artifacts. In the Retinal Fundus images, it is evident that FID favors $\delta = 0.1$, and KID prefers the images produced by $\delta = -0.05$. In terms of histopathology images, we see a consistent improvement in $\delta = 0.05$. Referring to Fig. 14, we see the changes, specifically in color values.

Table 8. Textual Inversion with and without LD on various datasets.

Dataset	δ	FID ↓	KID ↓
CelebAHQ [9]	-0.05	**100.69**	**0.02**$_{\pm 0.01}$
	0	118.75	0.03$_{\pm 0.02}$
	0.1	107.64	0.03$_{\pm 0.01}$
Fundus [3]	-0.05	119.48	**0.10**$_{\pm 0.02}$
	0	189.52	0.18$_{\pm 0.03}$
	0.1	**117.81**	0.12$_{\pm 0.02}$
Histopathology [8]	-0.05	**113.09**	**0.12**$_{\pm 0.01}$
	0	169.83	0.16$_{\pm 0.02}$
	0.1	128.51	0.13$_{\pm 0.02}$

5.3. Ablation Study

We evaluate the image generation performance of the SD + Basic FT + *LD* at different steps. The best overall performance was achieved after $10,000$ epochs of fine-tuning as shown in Fig. 13.

Cross Attention Guidance Scale We evaluate the method using the aforementioned prompt style by varying the cross attention guidance parameter τ from 0.1 to 0.3 in Tab. 9 using the Pix2Pix Zero model for disease-conditioned image editing. The cross-attention guidance parameter is used as a learning rate in the optimizer of latent space parameters of the model where the loss is the Mean Square Error Loss between the cross-attention maps of encoded target prompt with intermediate features and the cross-attention maps of the encoded generated caption by BLIP [12] with intermediate features. For all cross-attention guidance values, the

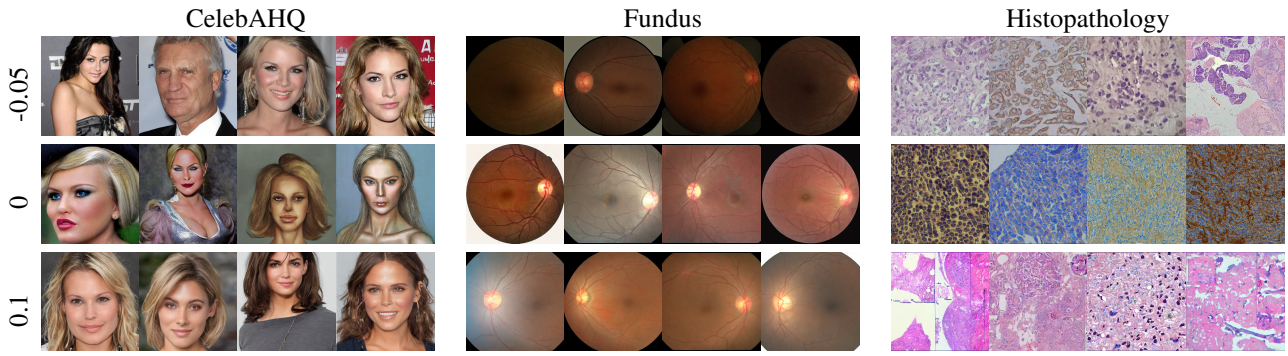


Figure 14. Results with different δ values on different domains: face, fundus, histopathology ($\delta = 0$ denotes fine-tuning without LD).

Table 9. Quantitative evaluation of counterfactual image generation using Pix2Pix Zero + *LD* with different guidance scales (τ).

	Counterfactuals			
τ	FID ↓	KID ↓	AUC ↑	SSIM ↑
0.1	36.76	0.0163	**0.77**	0.86
0.2	**36.12**	**0.0151**	0.56	0.895
0.3	37.6	0.0170	0.43	**0.9**
	Real Images			
Source class	38.78	0.0192	0.13	1
Target class	23.11	0.0002	0.87	1

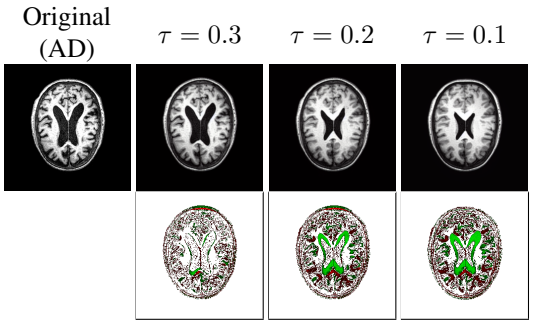


Figure 15. Effect of guidance scale on counterfactual brain MRI generation with Pix2Pix Zero + *LD*. Source Class: Alzheimer's Disease, Target Class: Cognitively Normal. ■: Removal, ■: Addition.

quantitative metrics of the counterfactuals improve compared to the source images. The generated counterfactuals successfully altered the disease diagnosis towards the target class. Using $\tau = 0.1$, the AUC score of the classifier evaluated on the counterfactuals gets closer to the score from real data of the target class. However, this comes at the trade-off of losing parts of the subject's identity. As expected, a low guidance value results in stronger image manipulation. This can be visually seen in Fig. 15, and quantitatively with the decrease in SSIM in Tab. 9.

Effect of Different δ Values The value of δ is optimized through a simple 1-D grid search in the range of $\delta \in [-0.1, 0.3]$, which remains constant during the fine-tuning and inference process. We present the additional ablation results in Tab. 10 and analyze the generated data distribution given different δ values in Fig. 12.

5.4. Additional Comparison

In Tab. 12, we compare Latent Drifting to Causal-Gen [14], which is the SOTA in counterfactual medical image generation on CheXpert. As it can be seen, Latent Drifting outperforms Causal-Gen by a large margin in both image quality metrics and the AUC. Furthermore, we fine-tune and evaluate ControlNet [18] on the Brain MR data with different LD values and present the FID values in Tab. 11. The results show that LD drastically improves image generation performance.

5.5. Results on Chest X-ray Image Generation

In order to evaluate our method on other organs and modalities, we utilize the CheXpert [7] dataset. We finetune the Stable Diffusion [15] using Textual Inversion [6] with and without *LD*. For these experiments, we set the guidance scale value to 5 and the number of sampling steps to 100. We empirically found that the best image generation performance with different prompt settings is achieved using the initialization token "radiation" and with placeholder tokens, each corresponding to the respective observation ("<healthy-chest-radiation>", "<cardiomegaly-chest-radiation>", "<fusion-chest-radiation>", "<pneumonia-chest-radiation>"). The quantitative results are reported in Tab. 2 of the main paper. Qualitative results are provided in Figs. 16 and 17. The results in Fig. 16 illustrate that Latent Drifting enhances the realism of the generated images. It is worth noting that the pre-trained classifier evaluates if the images correspond to their conditions.

Table 10. Ablation study on different δ values on Brain MR

δ		-0.1	0	0.05	0.1	0.15	0.2	0.25	0.3
Healthy	FID ↓	114.54	137.39	82.1	68.18	65.77	**56.89**	73.58	83.45
	KID ↓	0.12	0.15	0.0804	0.0727	0.0785	**0.0538**	0.0815	0.0978
AD	FID ↓	126.93	121.03	98.11	**58.97**	83.29	93.77	116.25	120.17
	KID ↓	0.118	0.110	0.0812	0.057	**0.0538**	0.0949	0.1177	0.1249
Mean	FID ↓	120.73	129.21	90.11	**63.58**	74.53	75.33	94.91	101.81
	KID ↓	0.119	0.13	0.0808	**0.0648**	0.066	0.074	0.099	0.111

Table 11. Results on ControlNet [18] with and without LD for Brain MR Synthesis. $\delta = 0$ denotes fine-tuning without LD.

δ	-0.7	-0.6	-0.5	-0.4	-0.2	-0.1	0	0.1	0.2	0.4	0.5	0.6
FID	162.37	154.95	**141.17**	159.15	170.80	206.14	274.31	217.83	180.37	162.72	151.11	164.1

Table 12. Quantitative evaluation of counterfactual image generation using Pix2Pix Zero with *LD* compared to Causal-Gen [14] on CheXpert.

Method	FID ↓	KID ↓	AUC ↑	SSIM ↑
Causal-Gen [14]	51.75	0.0370	0.59	0.85
Pix2Pix Zero + *LD* (Ours)				
$\tau = 0.1$	36.76	0.0163	**0.77**	0.86
$\tau = 0.2$	**36.12**	**0.0151**	0.56	0.89
$\tau = 0.3$	37.60	0.0170	0.43	**0.90**
Real Images				
Source class	38.78	0.0192	0.13	1
Target class	23.11	0.0002	0.87	1

5.6. Additional Qualitative Results on Brain MR Generation

Longitudinal data on Brain MR has been used for the evaluation of counterfactual image generation [14], and we used two datasets of Brain MR [11, 17] for the task of counterfactual image generation. In Fig. 19, it is observed that fine-tuning of Stable Diffusion [15] leads to images that contain diverse shapes, which evidently cannot be considered as brain MR images; on the other hand, in Fig. 20, by additional conditioning with *LD*, the same pre-trained model can be fine-tuned with similar steps of training to generate images that follow the general pattern of the brain. The fidelity of the images to the class they are conditioned on (Alzheimer vs. Healthy) is evaluated by a pre-trained classifier. However, even visually, the results have a drastic improvement in contrast to Fig. 19. In Fig. 21, we compare how *LD* can introduce Alzheimer's disease to a healthy brain. It can be seen that most of the brain remained intact, and only parts of the brain have deteriorated. In Fig. 22, we prompt the model to generate healthy brain images from developed cases of Alzheimer's disease, and with *LD*, the model is able to grow certain areas of the brain while preserving the general structure of the input brain images. In

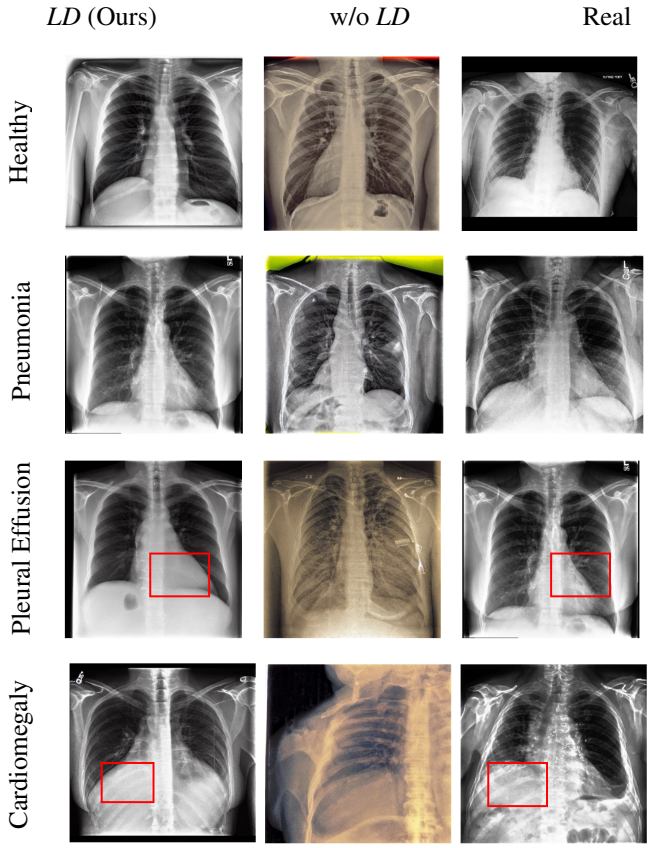
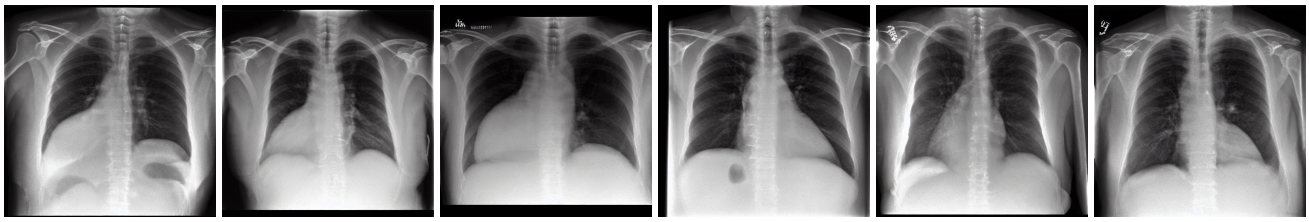


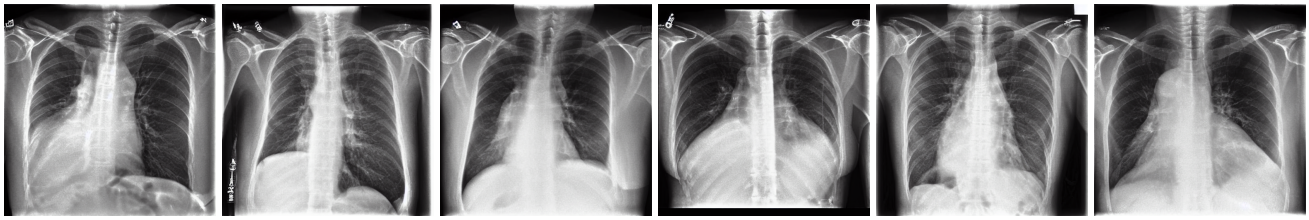
Figure 16. Comparison of generated images with and without *LD* during fine-tuning on CheXpert [7] dataset . Examples generated using Textual Inversion [6]. From left to right: With *LD*, Without *LD*, Real Chest X-ray.

Figs. 21 and 22, we also compare Latent Drifting to Causal-Gen [14] and StarGAN v2 [4]. As it can be seen in the difference, Latent Drifting preservers the bony structure while generating deteriorated ventricles in the brain, which cor-

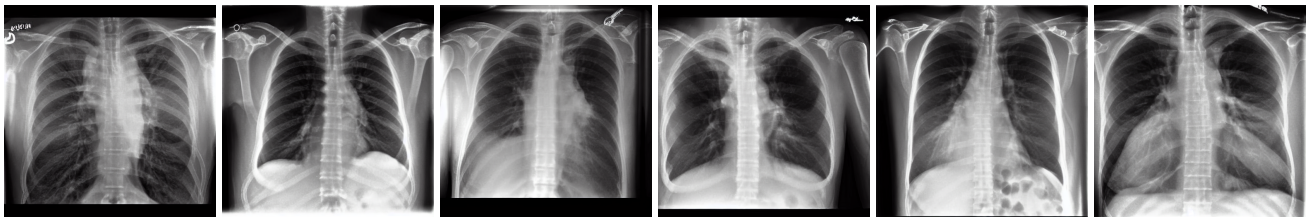
(a) **Cardiomegaly**: Enlargement of the heart



(b) **Pleural Effusion:** fluid buildup, blurring on the lower part



(c) **Pneumonia:** infection, appear as hazy area of white or gray



(d) **No Finding**

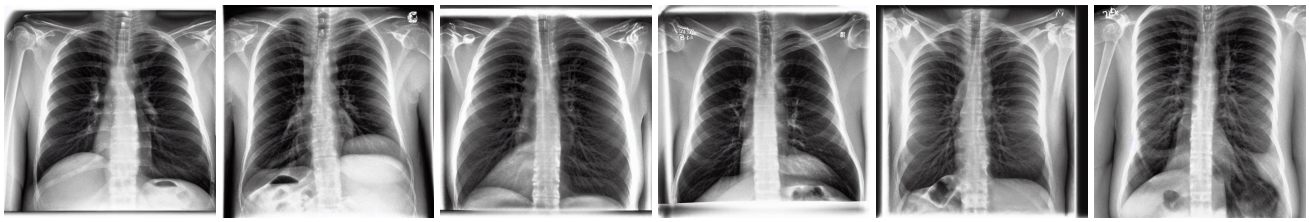


Figure 17. Chest X-ray generated images conditioned on the label using Textual Inversion [6] with *LD*.

respond to Alzheimer's disease, while on the other hand, Causal-Gen also affects the bony structure of the brain. Compared to the latter, StarGAN v2 generates the least amount of difference between the source and the counterfactual. Furthermore, we present the image synthesis results using LD for the MCI (Mild cognitive impairment) class in Fig. 18.

5.7. Additional Qualitative Results on Brain Aging

Fig. 23 shows additional results on brain aging using the InstructPix2Pix [1] model fine-tuned with *LD*. In this experiment, we condition the model on the source brain MR and the target age, as explained in the main paper. The generated image with *LD* should correspond to the target image, which is the correct counterfactual image as a result of brain aging. As can be seen in the results, the brain MR images generated by *LD* are visually similar to the target brain MR and show minor deterioration of the brain matter.

6. Discussions

Ethical Considerations Following a Human Subjects Research (HSR) Determination and utilizing publicly de-identified data, we sought approval from a reputable Ethics

Figure 18. Randomly sampled real and generated data for Brain MR, MCI class

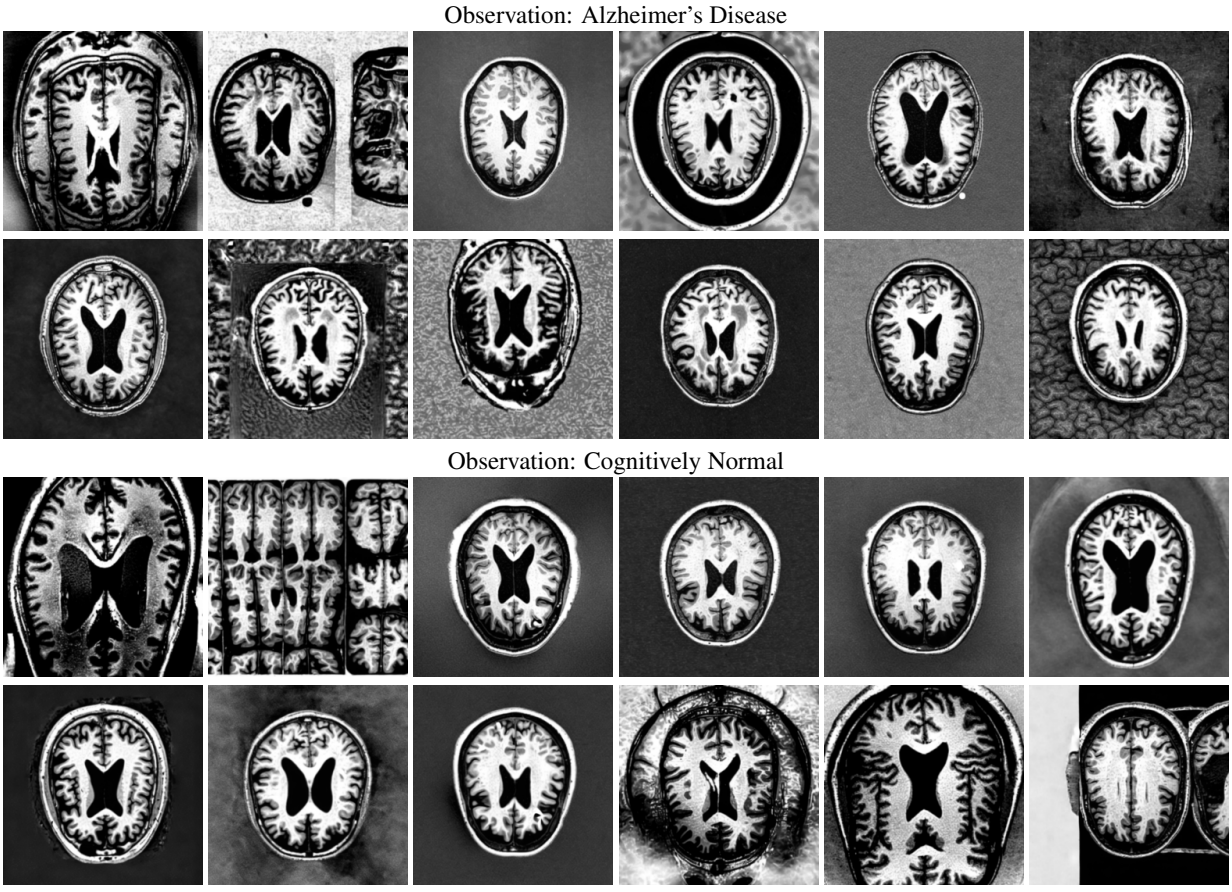


Figure 19. Brain MRI slice generation using SD + Basic FT without *LD*.

Review Board (ERB). Note that our synthetic counterfactual data were *exclusively* employed for training other models, with subsequent rigorous evaluation using *real* data, ensuring sufficient representation of under-represented conditions, particularly during testing. Despite this, we emphasize the need for transparency during model deployment by explicitly stating that the models were trained on synthetic data.

Broad Impacts Our Latent Drift (*LD*) method can impact the field of medical imaging by generating high-quality synthetic images where real data is limited. It overcomes the need for extensive datasets, addresses fine-tuning challenges in pre-trained models, and supports creating realis-

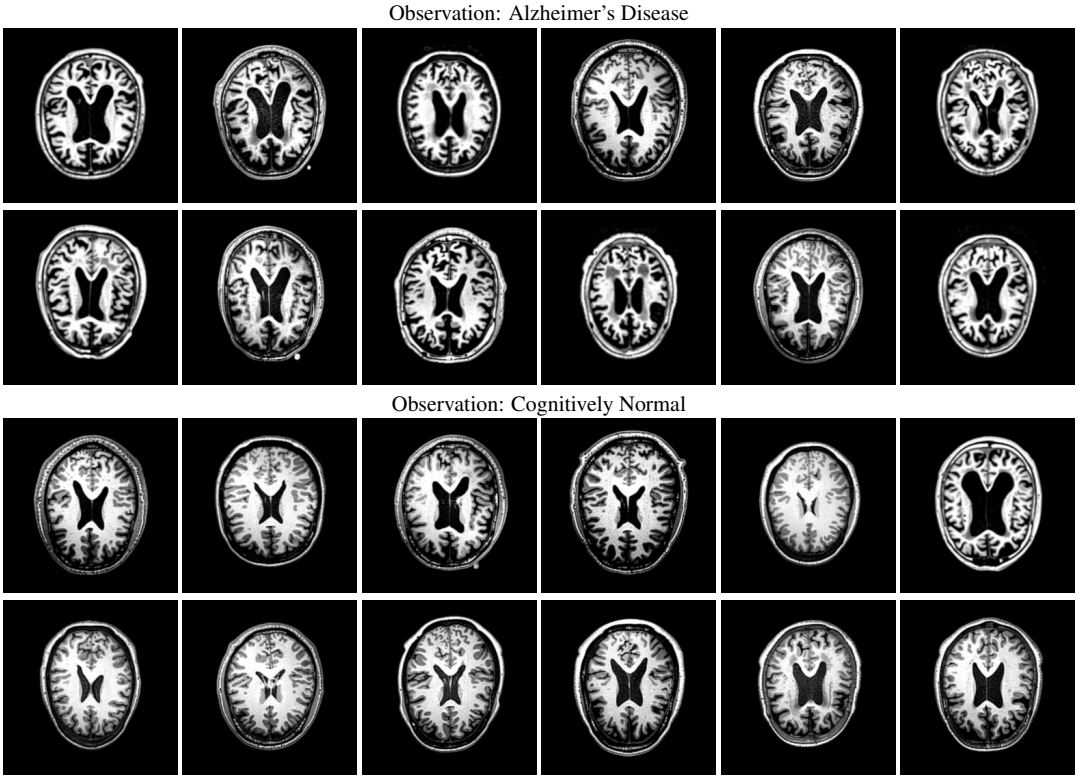


Figure 20. Brain MRI slice generation using SD + Basic FT with *LD*.

tic counterfactual images. The ability to produce realistic counterfactual images based on text and image conditions while maintaining fidelity is crucial for simulating patient-specific scenarios and contributes to the development of personalized treatment strategies, as evidenced by improved classifier performance on real datasets and evaluations on diverse medical benchmarks.

Limitations Evaluating the authenticity of synthetic medical images is complex and typically necessitates expert review, which is impractical for large datasets and large-scale training. Our study utilized a classifier to screen the extensive data, recognizing that traditional clinical validation methods are unfeasible. Although the AUC was reported for classifiers trained on synthetic images and tested on real data, these metrics may not be adequate for clinical utility. Therefore, there is a pressing need for innovative metrics capable of assessing the authenticity of counterfactual medical images without the need for heavy expert involvement to confirm their clinical value.

Figure 21. Comparison of counterfactual MR slice generation from healthy to Alzheimer's Disease using Pix2Pix Zero + *LD* to Causal-Gen [14] and StarGAN-v2 [4]. ■: Removal, ■: Addition. P2P-Z stands for Pix2Pix Zero.

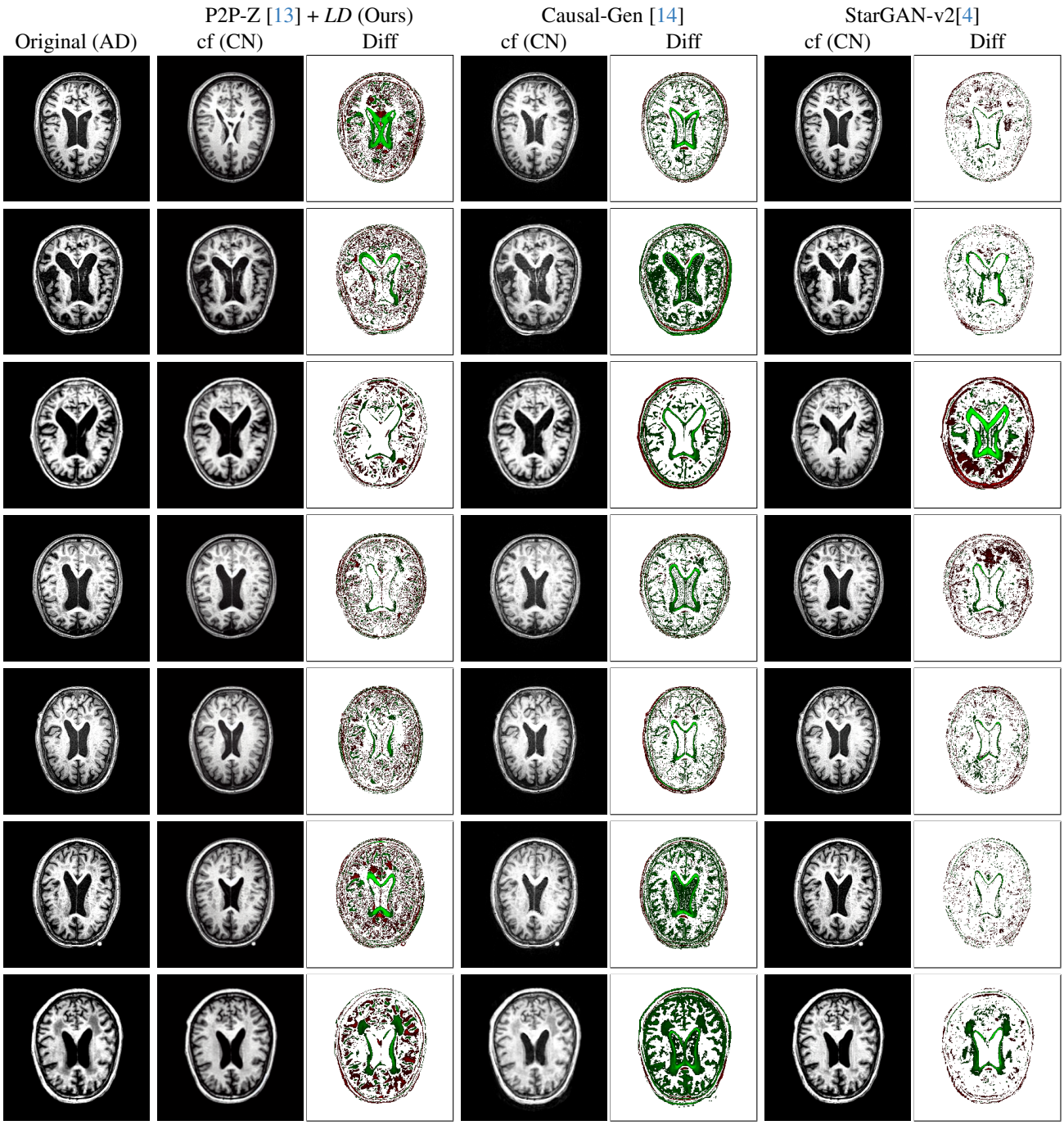


Figure 22. Comparison of counterfactual MR slice generation from Alzheimer's Disease to healthy using Pix2Pix Zero + *LD* to Causal-Gen [14] and StarGAN-v2 [4]. ■: Removal, ■: Addition. P2P-Z stands for Pix2Pix Zero.

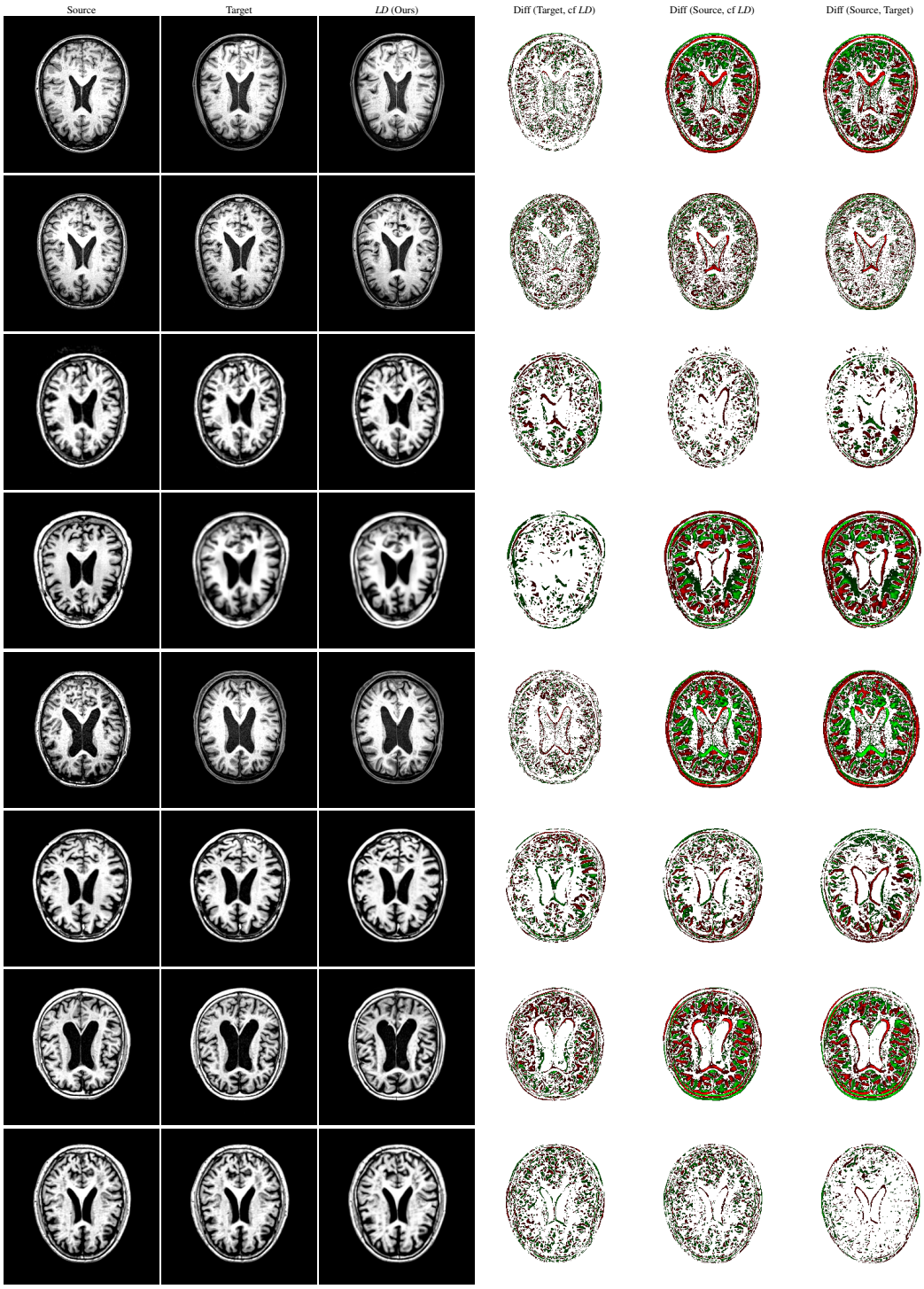


Figure 23. Counterfactual Image Generation for Brain Aging using InstructPix2Pix [1] + *LD*. Diff$(a, b) = b - a$. ■: Removal, ■: Addition.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 7, 12

[2] Mikael Brudfors, Yael Balbastre, Parashkev Nachev, and John Ashburner. A tool for super-resolving multimodal clinical mri. *arXiv preprint arXiv:1909.01140*, 2019. 3

[3] Attila Budai, Rüdiger Bock, Andreas Maier, Joachim Hornegger, and Georg Michelson. Robust vessel segmentation in fundus images. *International journal of biomedical imaging*, 2013(1):154860, 2013. 4

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8188–8197, 2020. 6, 10, 11

[5] Bram de Wilde, Anindo Saha, Richard PG ten Broek, and Henkjan Huisman. Medical diffusion on a budget: textual inversion for medical image generation. *arXiv preprint arXiv:2303.13430*, 2023. 3

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations*, 2022. 3, 4, 5, 6, 7

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 3, 5, 6

[8] Andrew Janowczyk and Anant Madabhushi. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *Journal of pathology informatics*, 7(1):29, 2016. 4

[9] Tero Karras. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 4

[10] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023. 3

[11] Pamela J LaMontagne, Tammie LS Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G Vlassenko, et al. Oasis-3: longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *MedRxiv*, pages 2019–12, 2019. 3, 6

[12] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 4

[13] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023. 10, 11

[14] Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. *arXiv preprint arXiv:2306.15764*, 2023. 5, 6, 10, 11

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 5, 6

[16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023. 3

[17] Michael W. Weiner, Dallas P. Veitch, Paul S. Aisen, Laurel A. Beckett, Nigel J. Cairns, Robert C. Green, Danielle Harvey, Clifford R. Jack, William Jagust, Enchi Liu, John C. Morris, Ronald C. Petersen, Andrew J. Saykin, Mark E. Schmidt, Leslie Shaw, Judith A. Siuciak, Holly Soares, Arthur W. Toga, and John Q. Trojanowski. The alzheimer's disease neuroimaging initiative: a review of papers published since its inception. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 8(1 Suppl):S1–68, 2012. 3, 6

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5, 6