Language-Guided Trajectory Traversal in Disentangled Stable Diffusion Latent Space for Factorized Medical Image Generation

Zahra TehraniNasab[1,2*] Amar Kumar[1,2*] Tal Arbel[1,2] (*equal contribution)

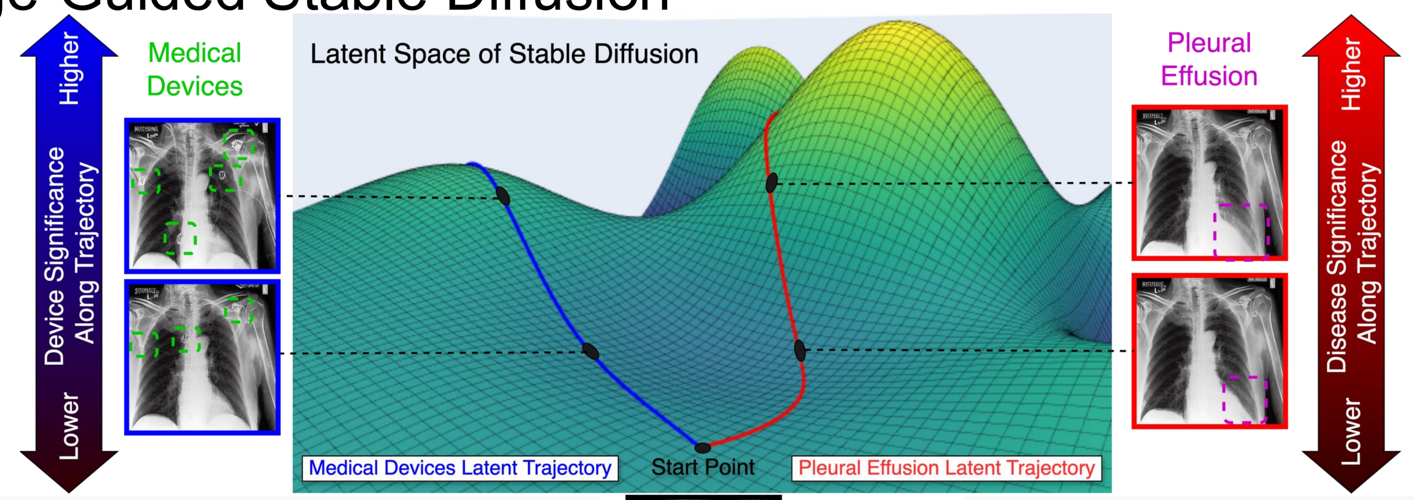
[1]McGill University [2]Mila-Quebec AI Institute

Website

CVPR Nashville JUNE 11-15, 2025

Introduction

Goal: Disentangling Latent Representations in Medical Images Using Language-Guided Stable Diffusion



Device Trajectory: `Chest x-ray showing support devices`

Disease Trajectory: `Chest x-ray showing Pleural Effusion`

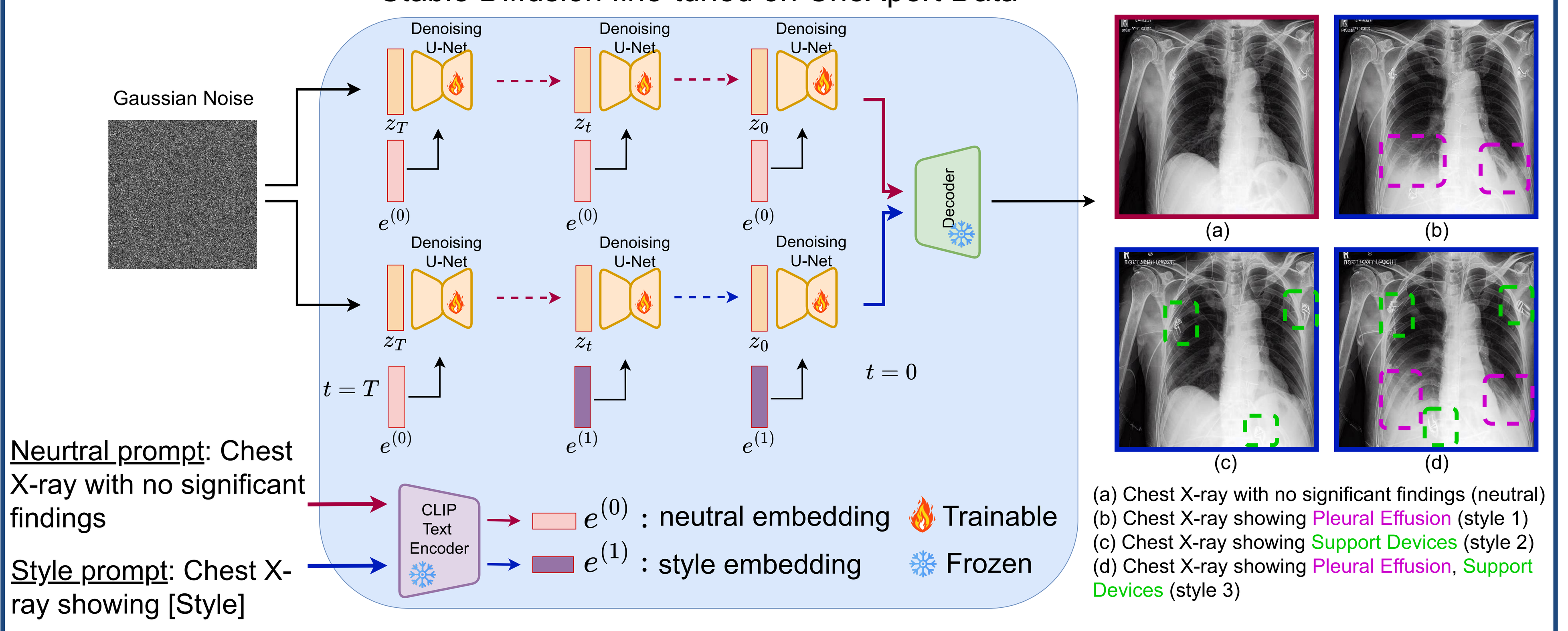
Text prompt for neutral image – `Normal chest x-ray with no significant finding`

Contributions

- 1st demonstration of language-guided latent space traversal for medical images.
- Enable identification of attribute-specific trajectories in the latent space.
- Support continuous, interpolatable transitions between images while preserving semantic content.

Architecture: Latent Trajectory Traversal

Stable Diffusion fine-tuned on CheXpert Data



(a) Chest X-ray with no significant findings (neutral)
(b) Chest X-ray showing Pleural Effusion (style 1)
(c) Chest X-ray showing Support Devices (style 2)
(d) Chest X-ray showing Pleural Effusion, Support Devices (style 3)

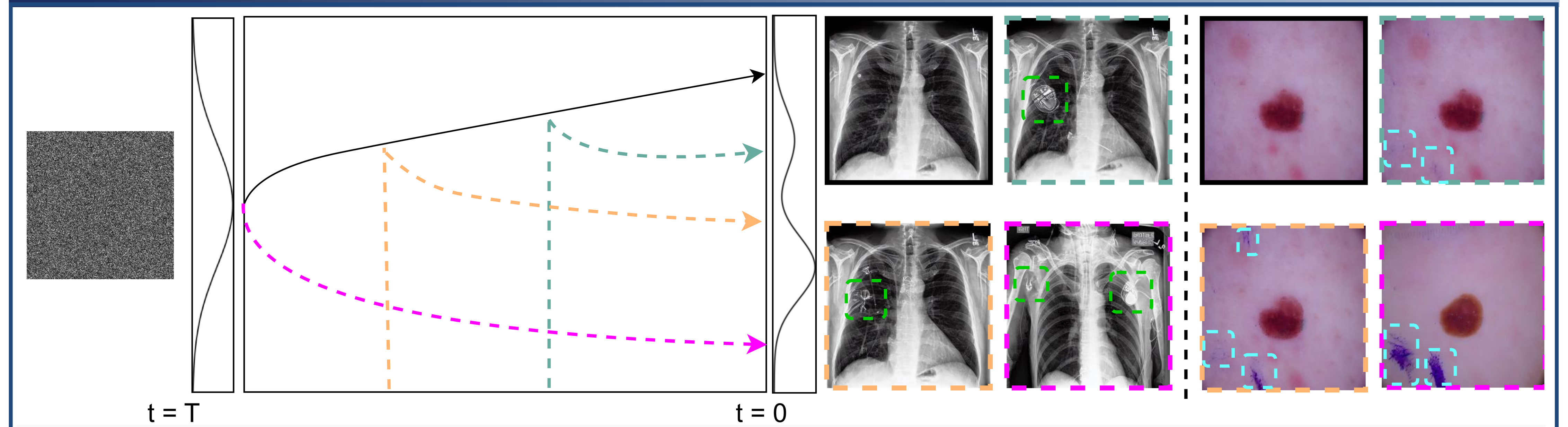
- **Finetuning:** Similar to PRISM[1], Stable Diffusion 1.5[2] is finetuned using public datasets - CheXpert and ISIC 2019.
  - Only the U-Net is trained while the VAE (encoder and decoder) remain frozen.
- **Inference:** A neutral image X_0 is generated from prompt embeddings e; attributes are added using modified embeddings e'.
  - During reverse diffusion, e' replaces the original text embeddings e at some timestep t.

Evaluating Conditionally Generated Images

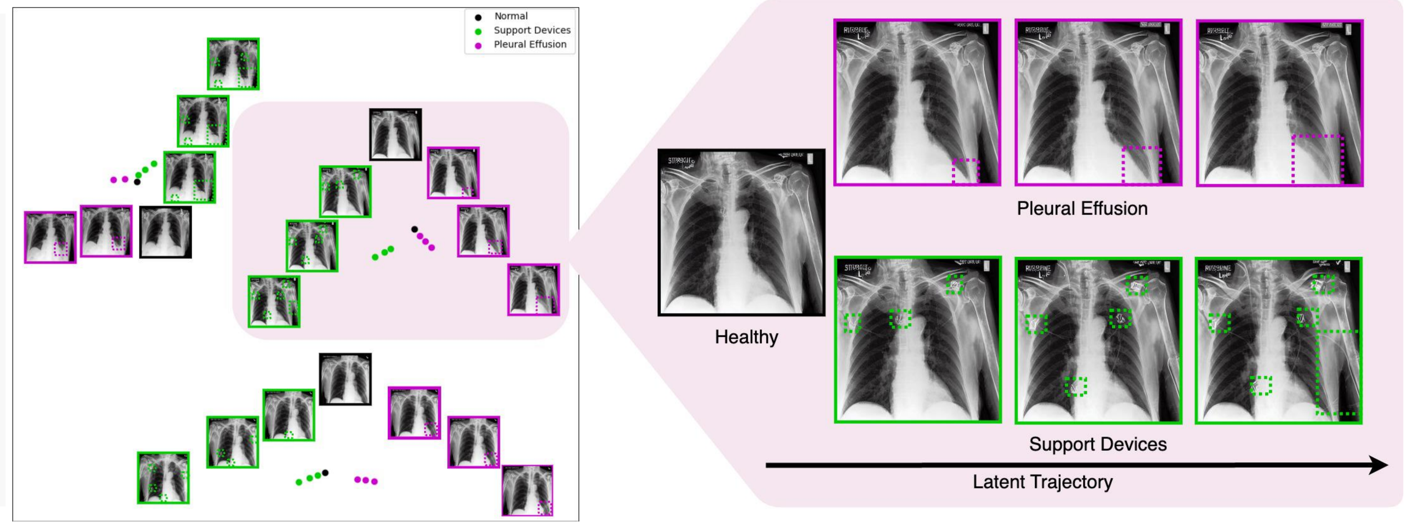
$$\mathrm{CFRT}_{\mathcal{A}} = \frac{1}{|X|}\sum_{x\in X} \mathbb{1}\left[|f(x) - f(x'_{\mathcal{A}})| > \max_{j\neq\mathcal{A}} |f(x) - f(x'_j)| \;\; \wedge y(x) = y(x'_{\mathcal{A}}) \wedge \; \forall k \neq \mathcal{A}, x_k = x'_{\mathcal{A}}(k) \right]$$

- We propose a new metric, Classifier Flip Rate along a Trajectory (CFRT), to validate disentanglement along the specified (style) trajectory.
- X is set of all the samples *x*, x_A' is the conditionally synthesized images where attribute *A* is flipped.
- f is the classifier and $\mathbb{1}[.]$ is the indicator function with value 1 if the condition is true and 0 otherwise.

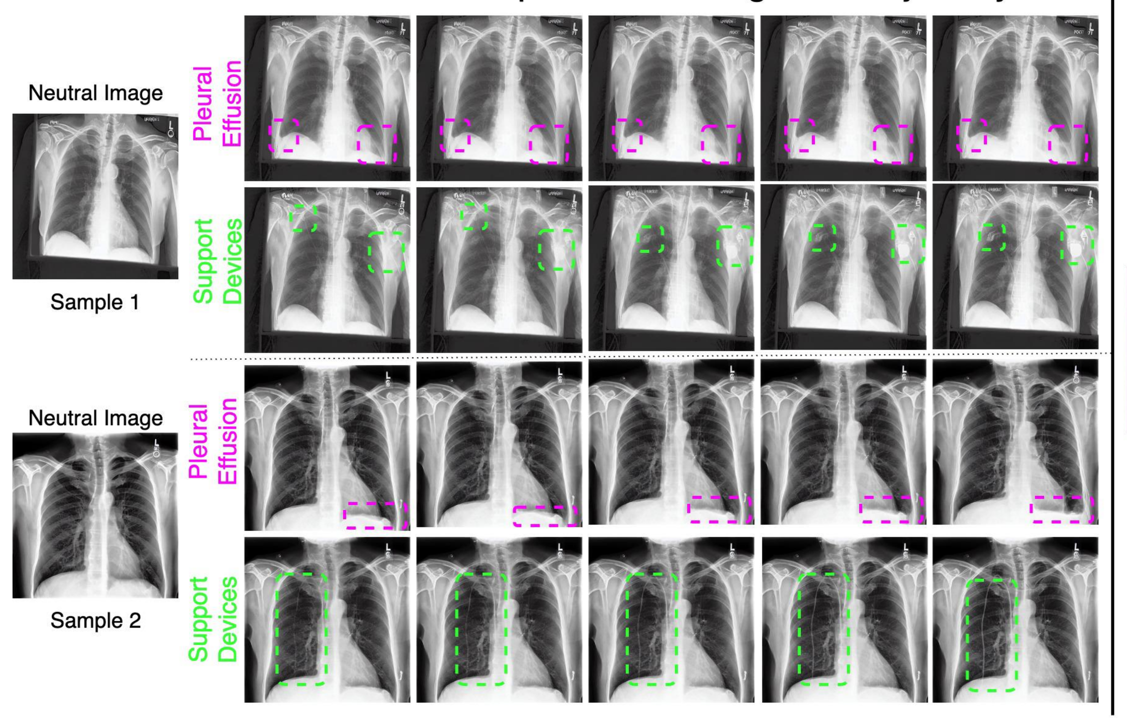
Qualitative Results



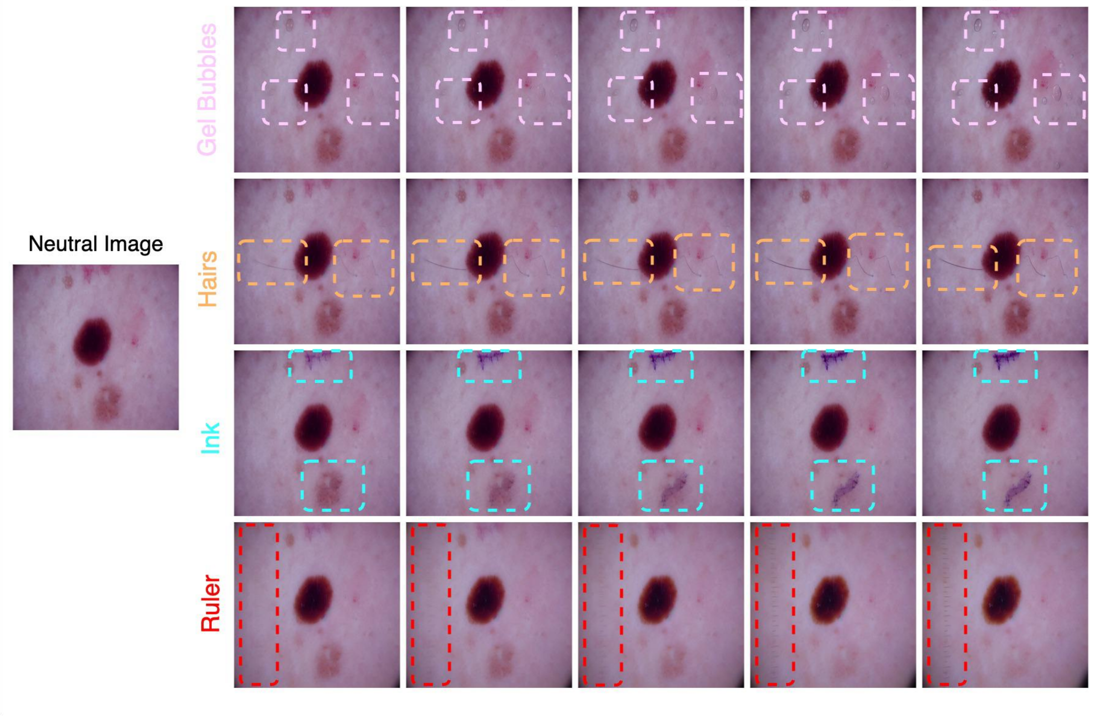
- Sampling closer to the timepoint t=0 results in a synthesized image similar to the original image.
- t-SNE plot of generated latent vectors of Stable Diffusion sampled from noise shows disentanglement.
- Traversal along the trajectory amplifies the desired attribute without altering confounding factors, indicating disentanglement in the Stable Diffusion latent space.



Bezier Interpolations Along The Trajectory



Bezier Interpolations Along The Trajectory



Quantitative Results

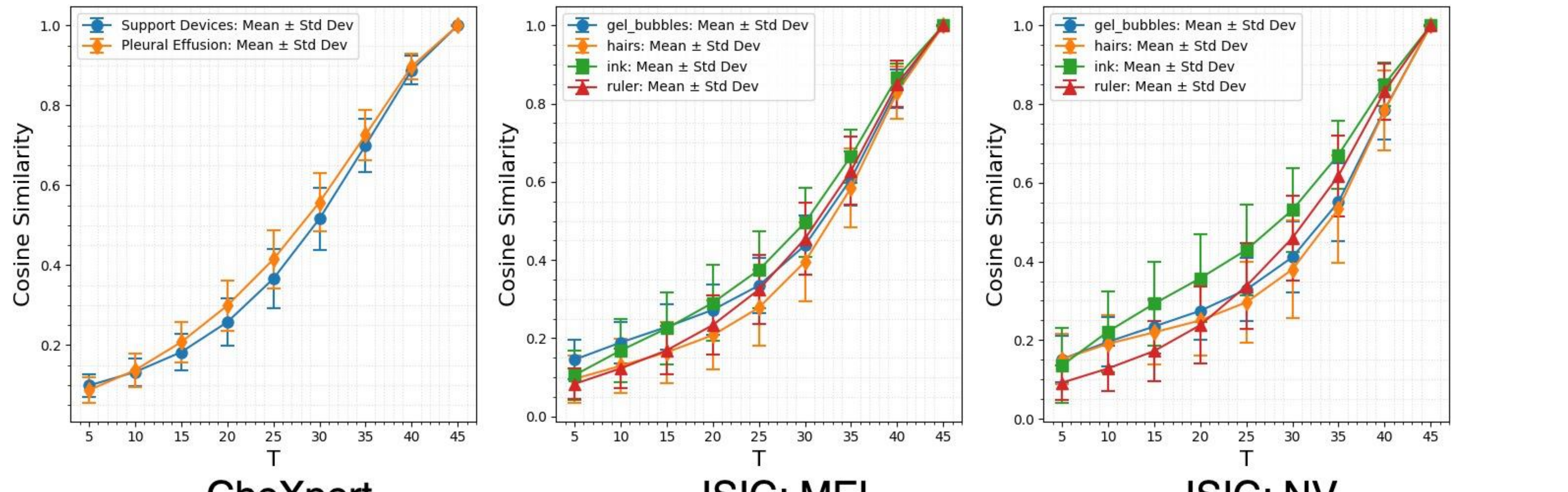
- Efficient-Net[3] is trained on real data for disease or artifact classification.

	CheXpert		ISIC				
	Support Devices	Pleural Effusion	MEL / NV	Hair	Gel Bubbles	Ink	Ruler
Accuracy	0.86	0.80	0.91	0.93	0.94	0.96	0.97
F1-score	0.88	0.79	0.88	0.91	0.78	0.89	0.88

- Evaluating synthesized images, 2500 samples per sub-class.
- Learned Perceptual Image Patch Similarity (LPIPS) shows visual quality of these images.

	CheXpert		ISIC				
Style→	Pleural Effusion	Support Devices		Hair	Gel Bubbles	Ink	Ruler
CFRT↑	0.78	0.89	MEL	0.91	0.99	0.59	0.74
			NV	0.86	0.97	0.71	0.95
LPIPS↓	0.24	0.05	MEL	0.08	0.09	0.12	0.11
			NV	0.05	0.09	0.06	0.10
Interpolations							
CFRT↑	0.73	0.86	MEL	0.88	0.99	0.62	0.79
			NV	0.93	0.99	0.72	0.97
LPIPS↓	0.22	0.04	MEL	0.05	0.08	0.09	0.08
			NV	0.04	0.07	0.04	0.07

- Cosine similarity between the direction of the latent representation of the conditionally generated image at a timestep relative to the latent of the original ("neutral") image.
- The cosine similarities indicate the non-linearity of trajectories for different attributes.



CheXpert ISIC: MEL ISIC: NV

Acknowledgments

The authors are grateful for funding provided by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, Mila - Quebec AI Institute, Google Research, Calcul Québec, and the Digital Research Alliance of Canada.

References

1. Kumar et. al, PRISM: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion, MIDL 2025.
2. Rombach et. al, High-resolution image synthesis with latent diffusion models, CVPR 2022.
3. Tan et. al, EfficientNet: Rethinking model scaling for convolutional neural network, ICML 2019.