Unrevealing Hidden Relations Between Latent Space and Image Generations in Diffusion Models

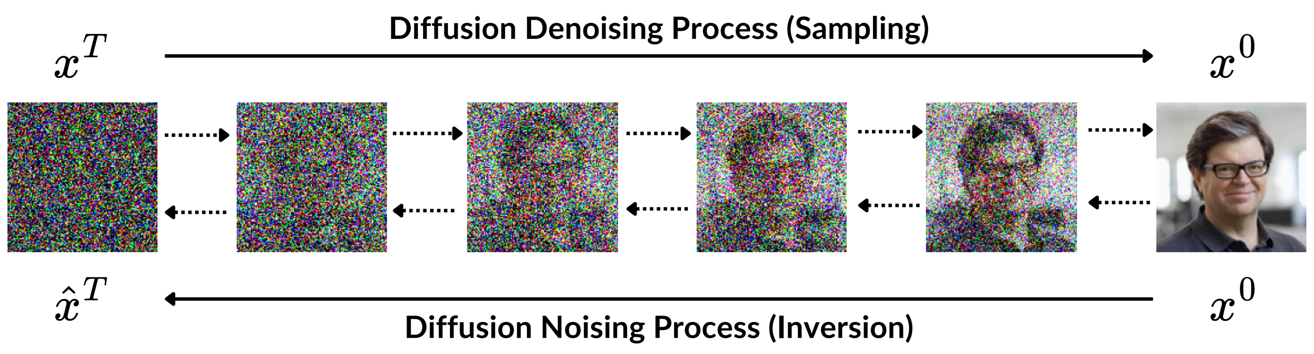
Łukasz Staniszewski[1] Łukasz Kuciński[2,3,4] Kamil Deja [1,2]

[1]Warsaw University of Technology [2]IDEAS NCBR [3]University of Warsaw [4]Polish Academy of Sciences

#TLDR

- We study relations between **Gaussian noises** $\mathbf{x}^\mathbf{T}$, **image samples** $\mathbf{x}^\mathbf{0}$ and their **latent encodings** $\hat{\mathbf{x}}^\mathbf{T}$ from the DDIM inversion procedure.
- We show that those encodings $\hat{\mathbf{x}}^\mathbf{T}$ manifold is between initial noise $\mathbf{x}^\mathbf{T}$ and image generations $\mathbf{x}^\mathbf{0}$.
- We show that noise $\mathbf{x}^\mathbf{T}$ to image $\mathbf{x}^\mathbf{0}$ mapping may be defined using the smallest $L2$ distance and that DMs learn important image features at the beginning of the fine-tuning.

Background



We can inverse the standard diffusion denoising procedure into the noising procedure:

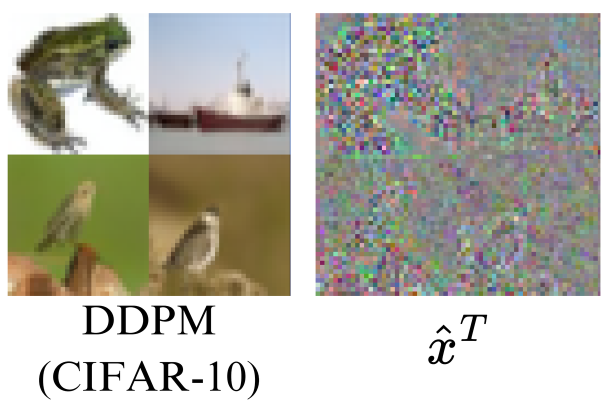
$$x_t = \gamma \cdot x_{t-1} + \eta \cdot \epsilon_\theta(x_t, t, c)$$

Due to circular dependency on $\epsilon_\theta(x_t, t, c)$, DDIM inversion approximates it:

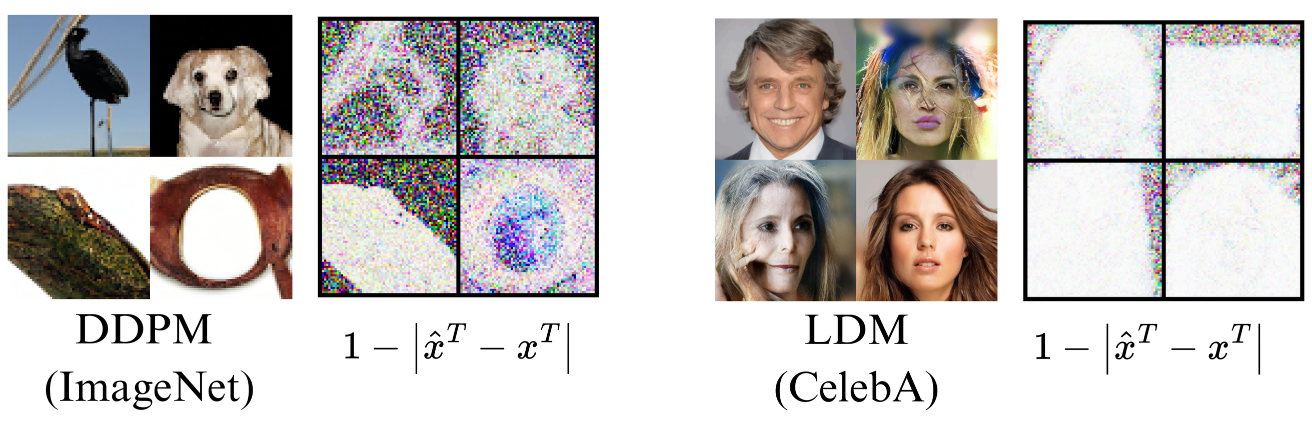
$$\epsilon_\theta(x_t, t, c) \approx \epsilon_\theta(\mathbf{x_{t-1}}, t, c).$$

Latent $\neq$ Noise

We can observe clear structures of original images $\mathbf{x}^\mathbf{0}$ in the inverted latents $\hat{\mathbf{x}}^\mathbf{T}$...



...or by showing the image difference between the latent $\hat{\mathbf{x}}^\mathbf{T}$ and the noise $\mathbf{x}^\mathbf{T}$.



Latent encodings $\hat{\mathbf{x}}^\mathbf{T}$ have correlated pixels.

	DDPM (CIFAR-10)	DDPM (ImageNet)	LDM (CelebA)
Noise ($\mathbf{x}^\mathbf{T}$)	0.159 ± 0.003	0.177 ± 0.007	
Latent ($\hat{\mathbf{x}}^\mathbf{T}$)	0.462 ± 0.009	0.219 ± 0.006	0.179 ± 0.008
Sample ($\mathbf{x}^\mathbf{0}$)	0.986 ± 0.001	0.966 ± 0.001	0.904 ± 0.005

Table 1. Top-10 correlation coefficients in random Gaussian noise vs. latent encoding.

If you enjoy this work...

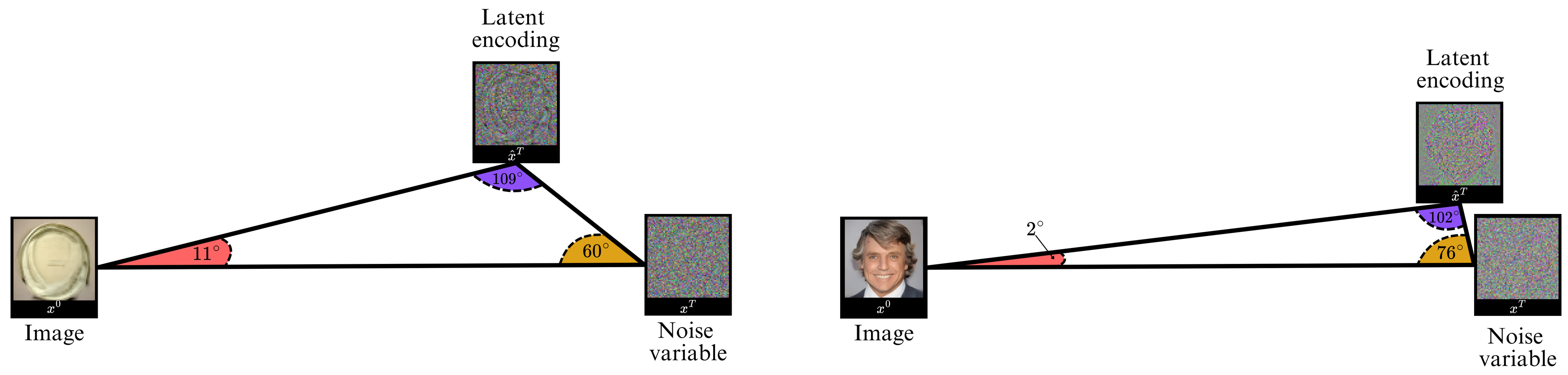
See the full paper for more details!

arXiv paper

Łukasz Staniszewski

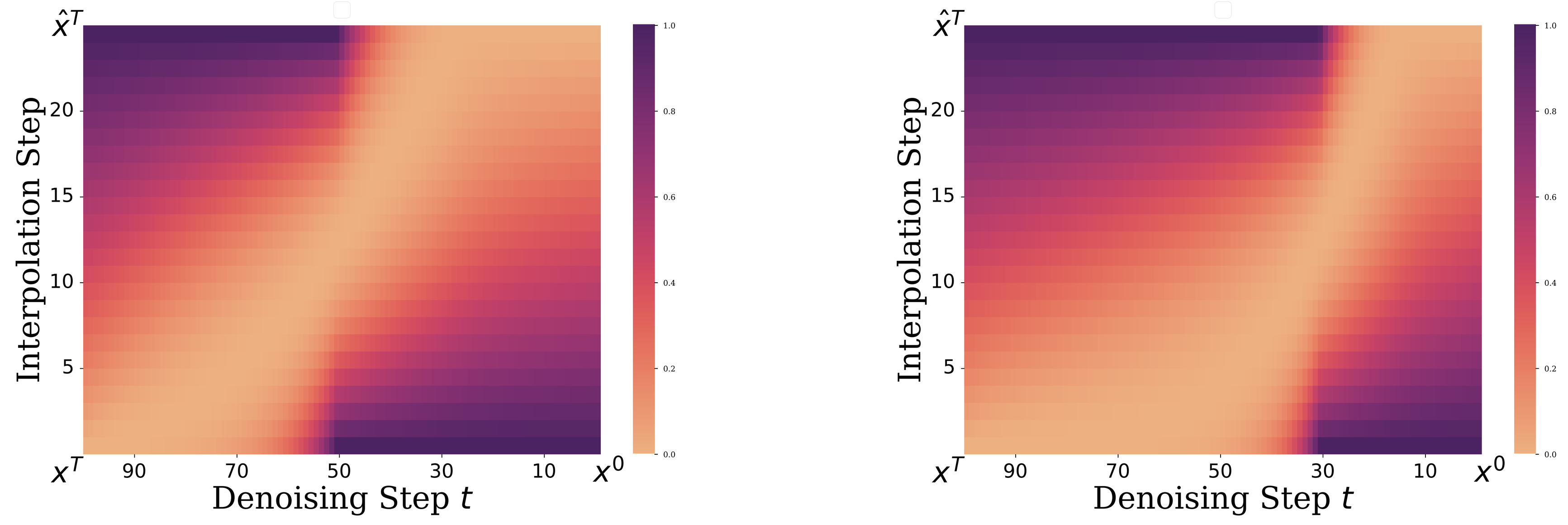
Where are the latents $\hat{\mathbf{x}}^\mathbf{T}$ located?

Latent encodings ($\hat{\mathbf{x}}^\mathbf{T}$) manifold is between random Gaussian noises ($\mathbf{x}^\mathbf{T}$) and their corresponding samples ($\mathbf{x}^\mathbf{0}$) manifolds.



(a) DDPM (ImageNet) (b) LDM (CelebA)

Diffusion model denoising trajectory is aligned with linear interpolation path between the Gaussian noise $\mathbf{x}^\mathbf{T}$ and latent encoding $\hat{\mathbf{x}}^\mathbf{T}$.



(a) DDPM (ImageNet) (b) LDM (CelebA)

Figure 5. Distances between next denoising steps and the $\mathbf{x}^\mathbf{T} \to \hat{\mathbf{x}}^\mathbf{T}$ interpolation points. Intermediate generations along the sampling trajectory initially get closer to the latent variable, and after approximately 50-70% of the path, they pass the latent.

Noise-to-Sample mapping

The mapping between initial Gaussian noise $\mathbf{x}^\mathbf{T}$ and its corresponding generation $\mathbf{x}^\mathbf{0}$ is secretly a $L2$-based nearest neighbor mapping.

T	ImageNet (DDPM) $x^0 \to x^T$	ImageNet (DDPM) $x^T \to x^0$	CelebA (LDM) $x^0 \to x^T$	CelebA (LDM) $x^T \to x^0$
10	$99.4_{\pm 0.0}$	$100_{\pm 0.0}$	$100_{\pm 0.0}$	$100_{\pm 0.0}$
100	$100_{\pm 0.0}$	$59.0_{\pm 7.1}$	$100_{\pm 0.0}$	$100_{\pm 0.0}$
1000	$99.8_{\pm 0.2}$	$44.6_{\pm 6.3}$	$100_{\pm 0.0}$	$100_{\pm 0.0}$
4000	$99.5_{\pm 0.3}$	$43.3_{\pm 6.7}$	-	-

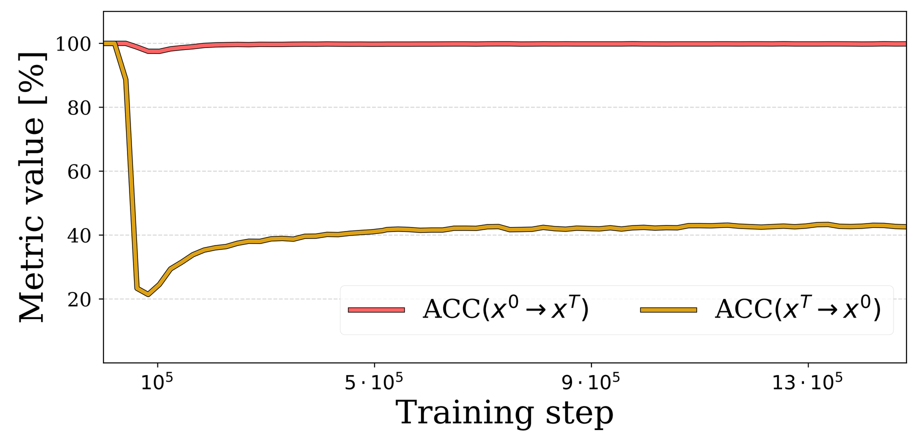
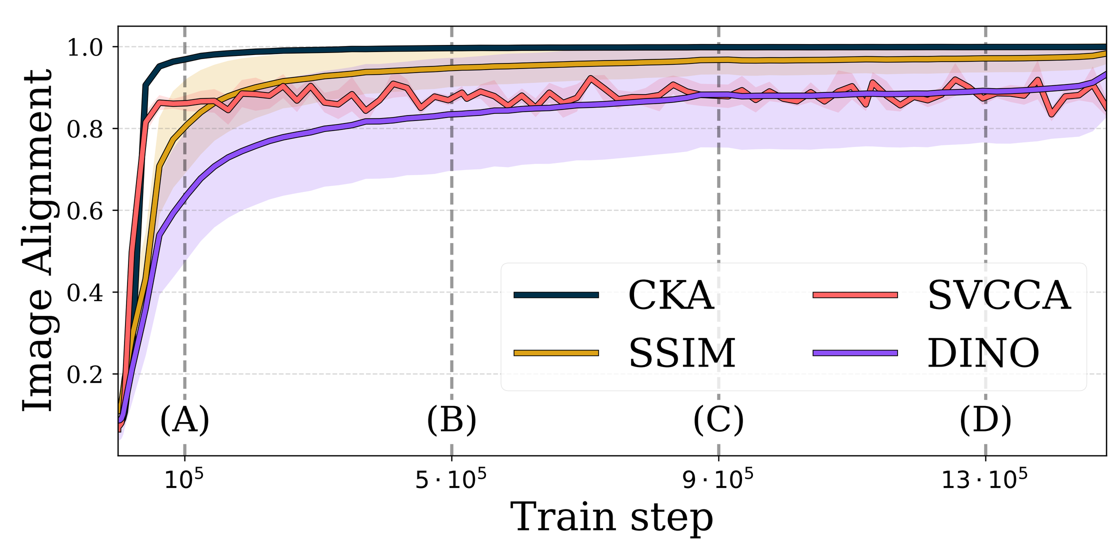
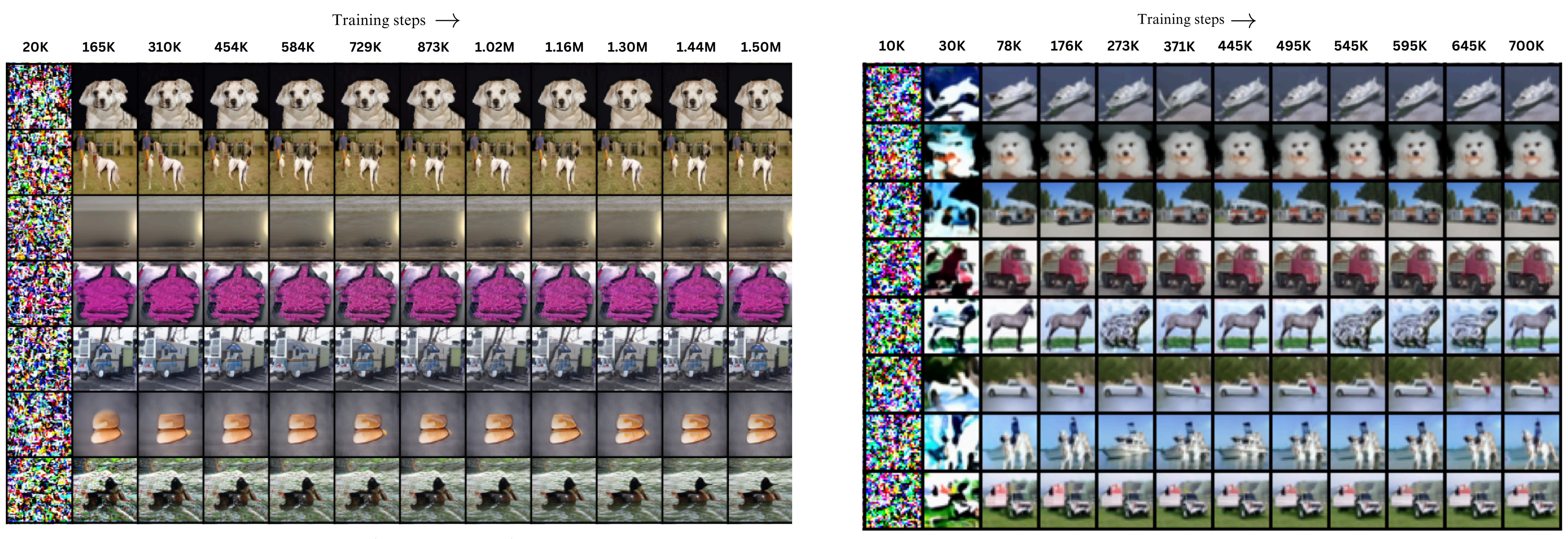


Figure 6. We are able to correctly select the original noise ($\mathbf{x}^\mathbf{T}$) for a given sample ($\mathbf{x}^\mathbf{0}$) by indicating the one with the closest $L2$ distance (left). Moreover, we show (right) that this mapping is established at the beginning of fine-tuning.

DMs generate the most important image features right at the beginning of fine-tuning, with only small details added further.





DDPM (ImageNet) DDPM (CIFAR-10)

Supported by PL-Grid Infrastructure grant no. PLG/2024/017266.