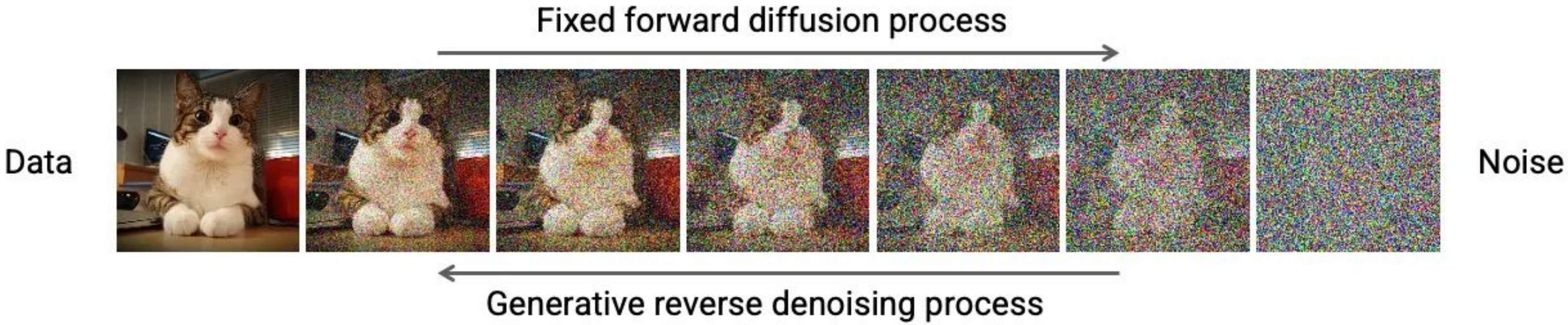
Inversion for Generative Image Editing

Alexey Arzhantsev
Konstantin Amelichev

Diffusion models

Let’s add noise to images (forward process)



and train a model to reconstruct images from pure noise (reverse process)

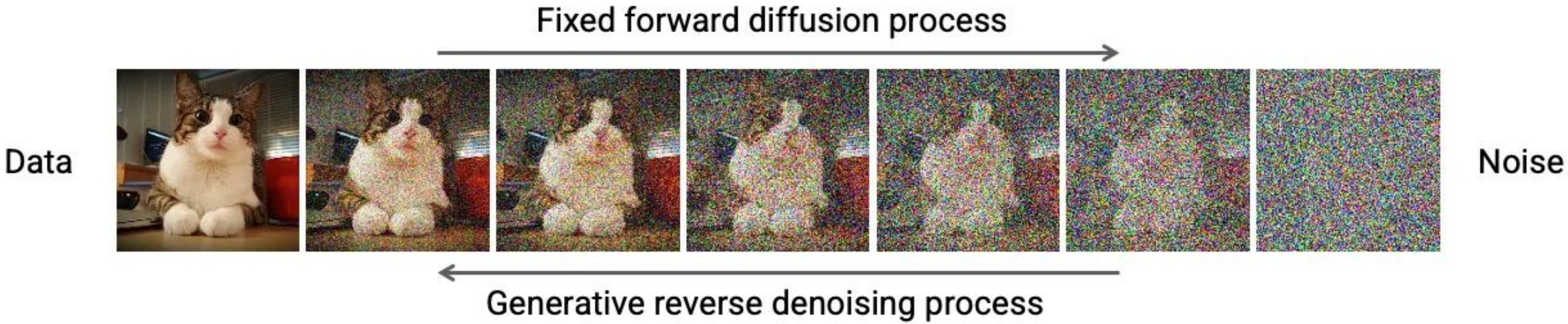
Classifier-free guidance

- We can use guidance to generate images: $\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_\lambda, \mathbf{c}) = (1+w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) - w\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda)$
- Not applicable for editing

+ “...with a hat…” =

DDIM inversion

Let’s try to reconstruct the noise that will generate the image we want



Here is the formula

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1} \right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C})$$

DDIM inversion

Let’s see how it works in practice

DDIM inversion

Solution – use guidance_scale=1 during inversion

DDIM-based editing

Perform guided generation from the inverted latents

“...zebra…”

“...elk…”

“...chicken…”

DDIM inversion

Let’s play and add more guidance during generation

“...zebra…”

“...elk…”

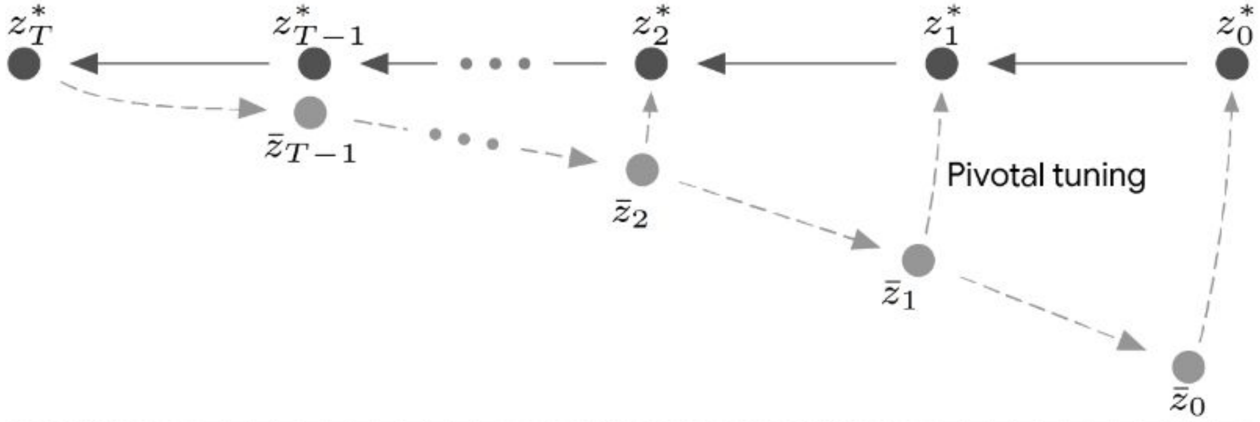
“...chicken…”

DDIM Inversion: Issues

- For complex scenes: can’t recover even the initial image, so result differs greatly from the source

Null-text inversion

We see that the generated image is not exactly the original.

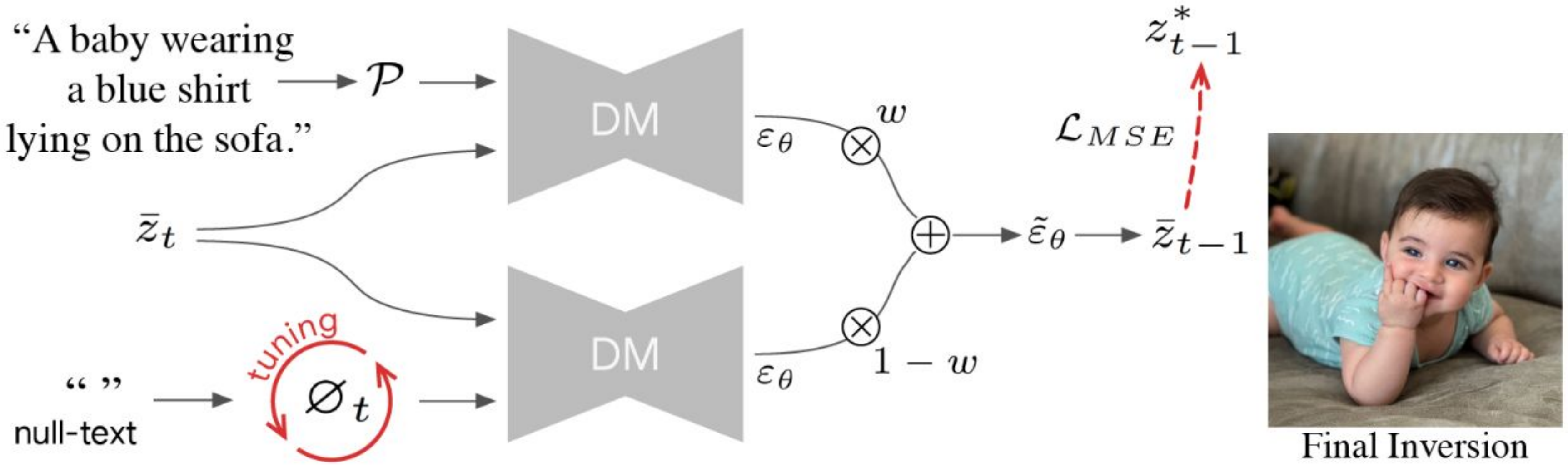


Input Image

Initial Inversion

Null-text inversion

Let's **train** unconditional embeddings to fix it.



Null-text inversion

Final pipeline is

1. Perform DDIM inversion to get latents (with *guidance_scale* = 1!)
2. Train unconditional embeddings
3. Perform denoising process with new unconditional embeddings and target prompt

Algorithm 1: Null-text inversion

Input: A source prompt embedding $\mathcal{C} = \psi(\mathcal{P})$ and input image $\mathcal{I}$.
Output: Noise vector z_T and optimized embeddings $\{\varnothing_t\}_{t=1}^{T}$.

Set guidance scale $w = 1$;
Compute the intermediate results $z_T^*, \ldots, z_0^*$ using DDIM inversion over $\mathcal{I}$;
Set guidance scale $w = 7.5$;
Initialize $\bar{z_T} \leftarrow z_T^*$, $\varnothing_T \leftarrow \psi(\text{""})$;
for $t = T, T-1, \ldots, 1$ **do**
    for $j = 0, \ldots, N-1$ **do**
        $\varnothing_t \leftarrow \varnothing_t - \eta \nabla_{\varnothing} \left\| z_{t-1}^* - z_{t-1}(\bar{z}_t, \varnothing_t, \mathcal{C}) \right\|_2^2$;
    end
    Set $\bar{z}_{t-1} \leftarrow z_{t-1}(\bar{z}_t, \varnothing_t, \mathcal{C})$, $\varnothing_{t-1} \leftarrow \varnothing_t$;
end
Return $\bar{z_T}$, $\{\varnothing_t\}_{t=1}^{T}$

Null-text inversion

First, let’s compare the reconstruction abilities. These examples use w = 7.5

original

DDIM inversion

Null-text inversion

Recovery of the initial image: sheep example

Null-text inversion

Now, let’s compare actual editing abilities.

“...zebra…”

“...elk…”

“...chicken…”

Null-text inversion

Now, let’s compare actual editing abilities.

“...zebra…”

“...horse…”

“...sheep…”

Null-text inversion

Let’s compare DDIM and null-text inversion.

original

DDIM inversion

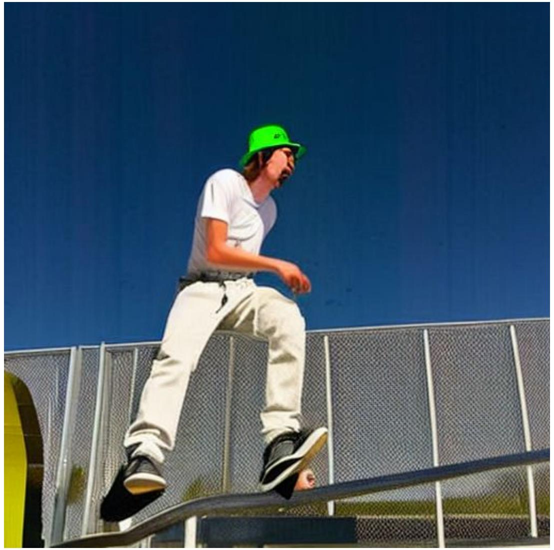
Null-text inversion

Null-text inversion

Let’s compare DDIM and null-text inversion.

original

DDIM inversion



Null-text inversion

Null-text inversion

Let’s compare DDIM and null-text inversion.

original

DDIM inversion

Null-text inversion

Results & Conclusion

Metric	Generated	DDIM Inversion	Null-text inversion
SSIM	0.2151±0.0825	0.4644±0.1318	**0.4894**±0.0971
CLIP Img2Img Target	0.7483±0.0997	**0.7930**±0.0960	0.7418±0.0835
CLIP Img2Img Source	0.6991±0.1061	**0.7509**±0.1055	0.7060±0.1018
CLIP Img2Text Target	**0.3212**±0.0288	0.3181±0.0318	0.3090±0.0218
CLIP Img2Text Source	**0.2911**±0.0384	0.2906±0.0394	0.2643±0.0374
Generation time	**30 seconds**	60 seconds	7 minutes

- Comparison of DDIM, Null-text Inversion
  - Some tradeoff in quality vs structural similarity
  - Much better than generating everything from sampled noise!
- Hyperparameters matter a lot

Q&A

References

- Mokady, R., Hertz, A., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6038-6047).
- Song, J., Meng, C., & Ermon, S. (2020). Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in neural information processing systems*, *33*, 6840-6851.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PmLR.