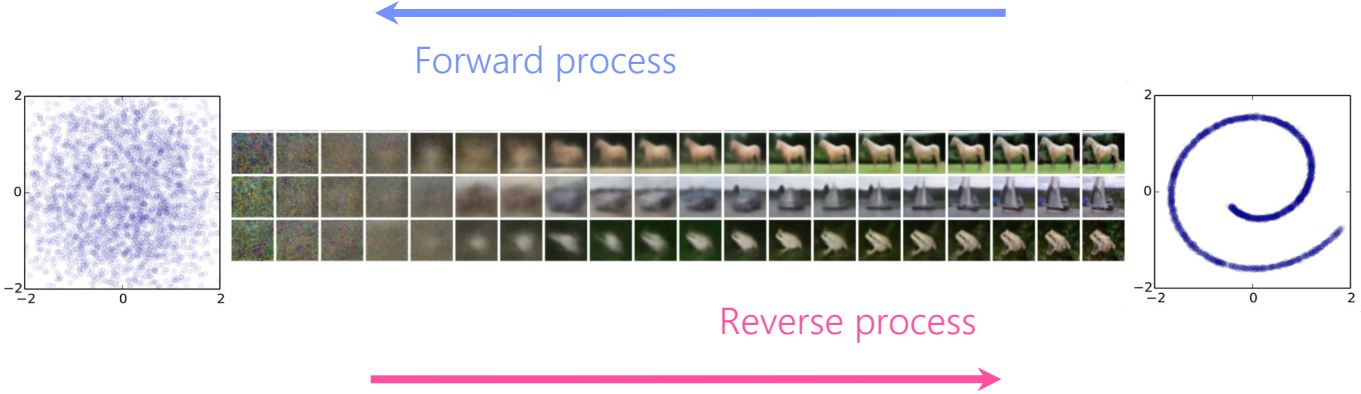
Semantic Editing in DM Latent

Lunit Research Seminar 23.06.09

Geonwoon Jang

Previously on Research Seminar... (Actually, 1 year ago)

22.03.11, (Fundamentals of) Diffusion Models [Link 1, Link 2]



Previously on Research Seminar...

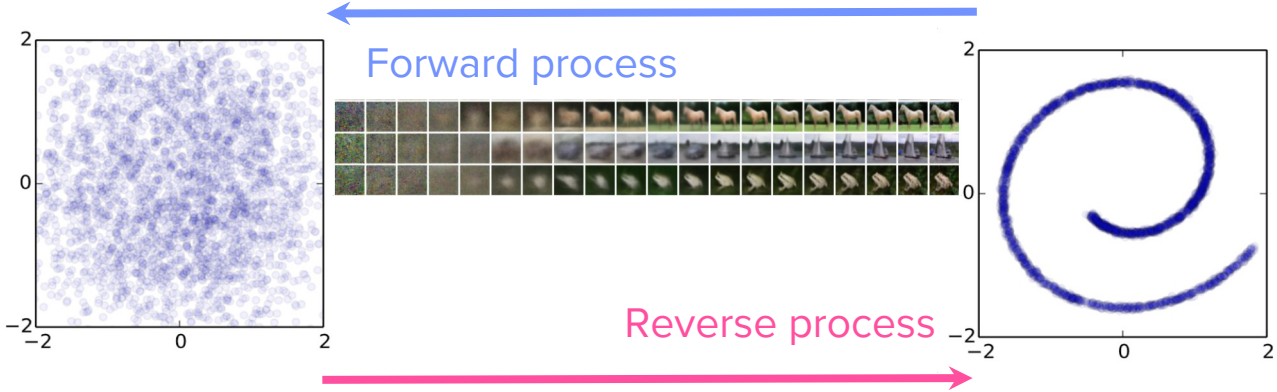
22.03.11, (Fundamentals of) Diffusion Models [Link 1, Link 2]

From the beginning,

- Deep Unsupervised Learning using Non-Equilibrium Thermodynamics (ICML '15)
- Denoising Diffusion Probabilistic Models (NeurIPS '20)
- Diffusion Models Beat GANs on Image Synthesis (NeurIPS '21)

Towards SoTA !

+) Quick Review of the Concept



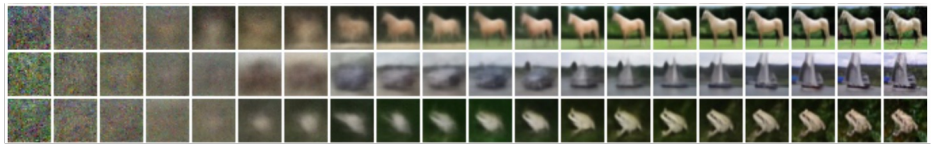
1. Set the forward process as **diffusion process**

   Adding noise gradually, micro-step repeatedly

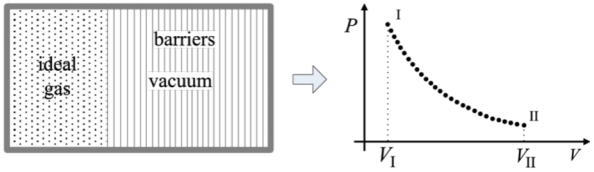
+) Quick Review of the Concept

But how to “reverse” it?

Forward process



Reverse process



1. Set the forward process as diffusion process with Gaussian
2. If the process is quasi-static, its reverse process can also be Gaussian
3. **Predict μ and Σ** for each steps in reverse process

Hype in Text-to-Image Generation

Midjourney (2020)

OpenAI
DALL·E 2

DALL-E 2 (2022)

Better multimodal feature guidance

Various methods for quality boost-up

Better LLM

...

stability.ai

Stable Diffusion (2022)

Google Research
Imagen

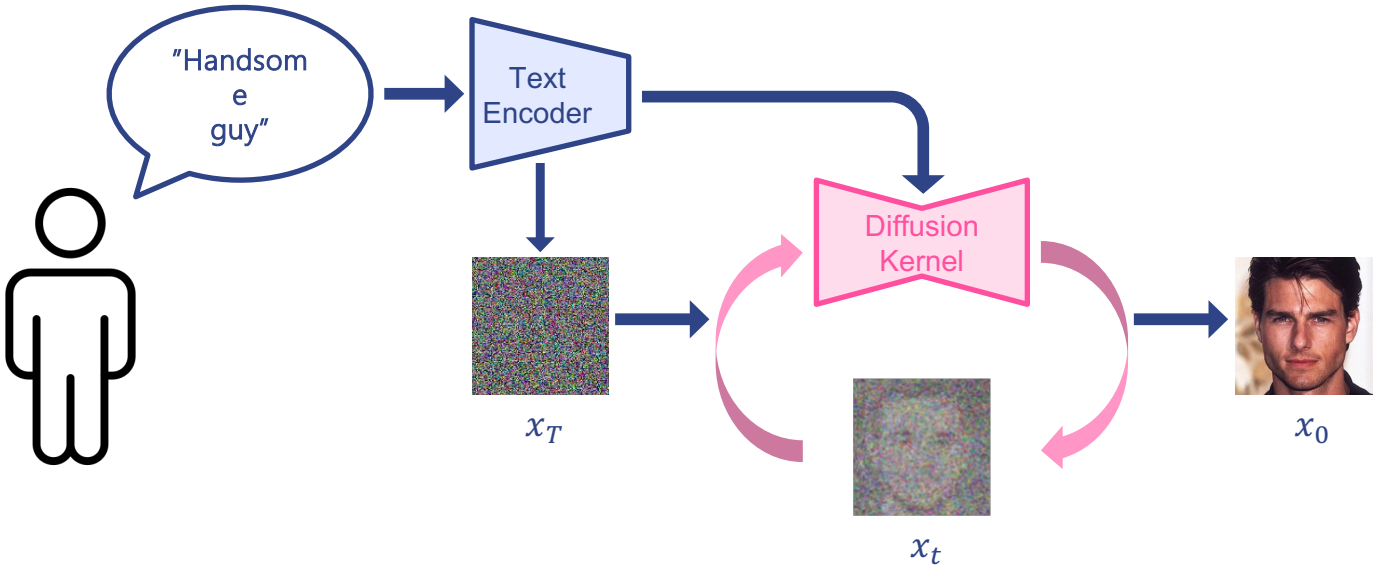
Imagen (2022)

On the other hand,

Think of this scenario:

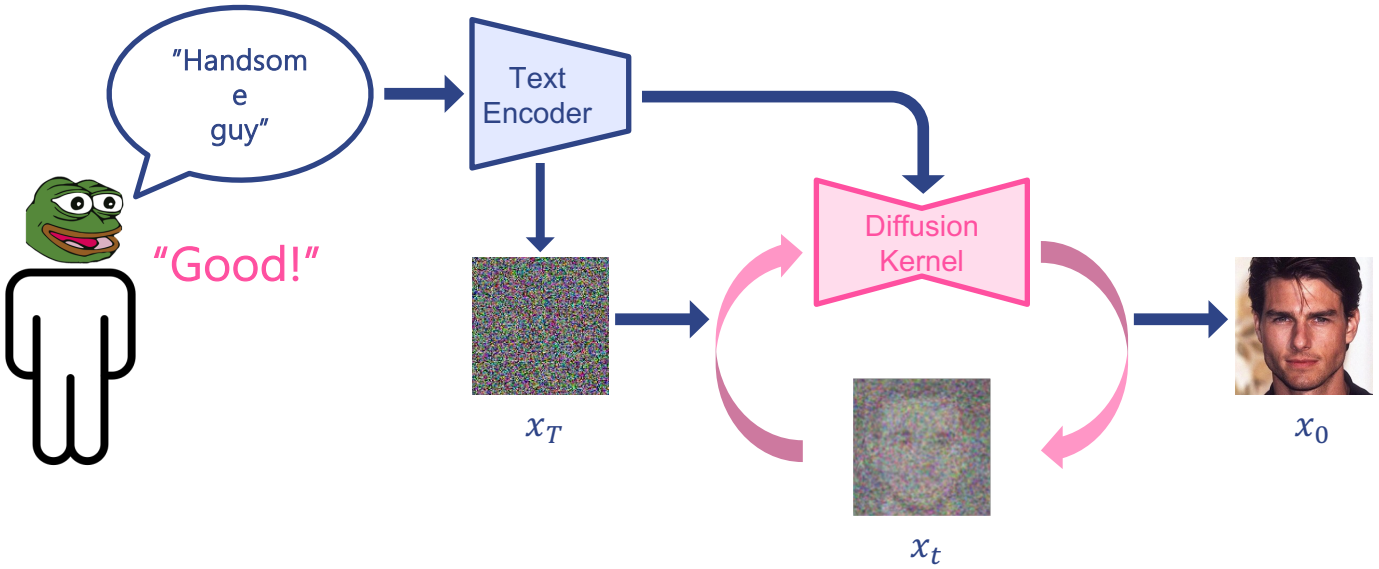
On the other hand,

Think of this scenario:



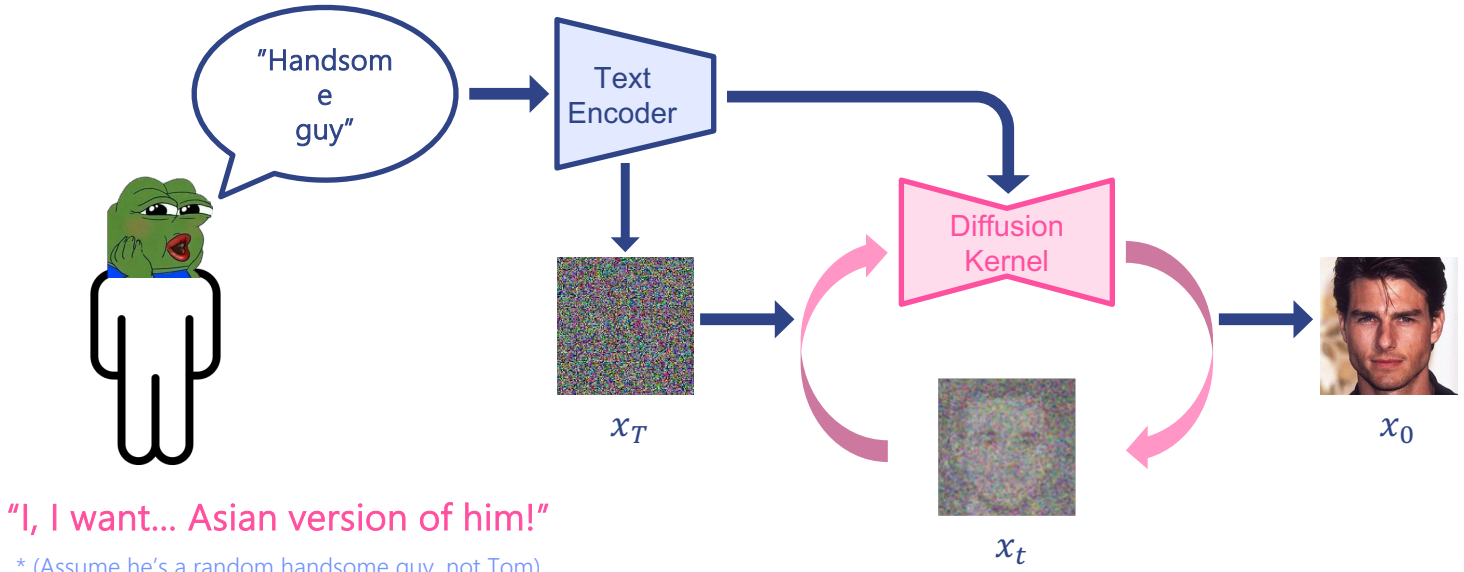
On the other hand,

Think of this scenario:



On the other hand,

Think of this scenario:

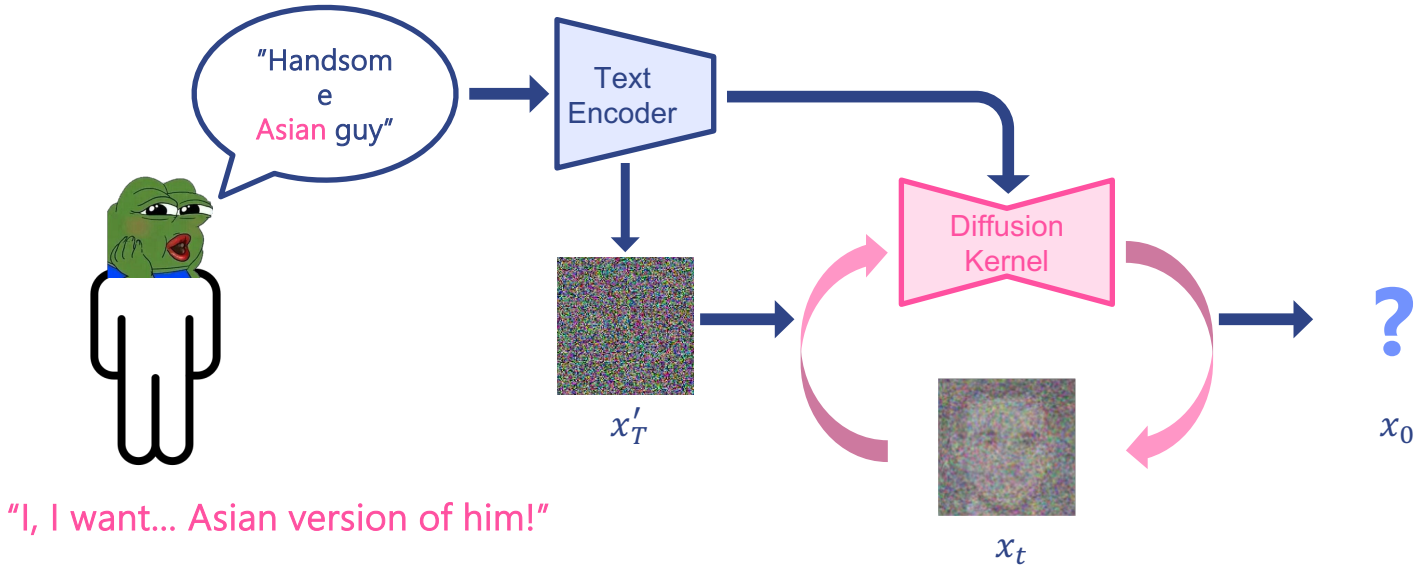


"I, I want... Asian version of him!"

* (Assume he's a random handsome guy, not Tom)

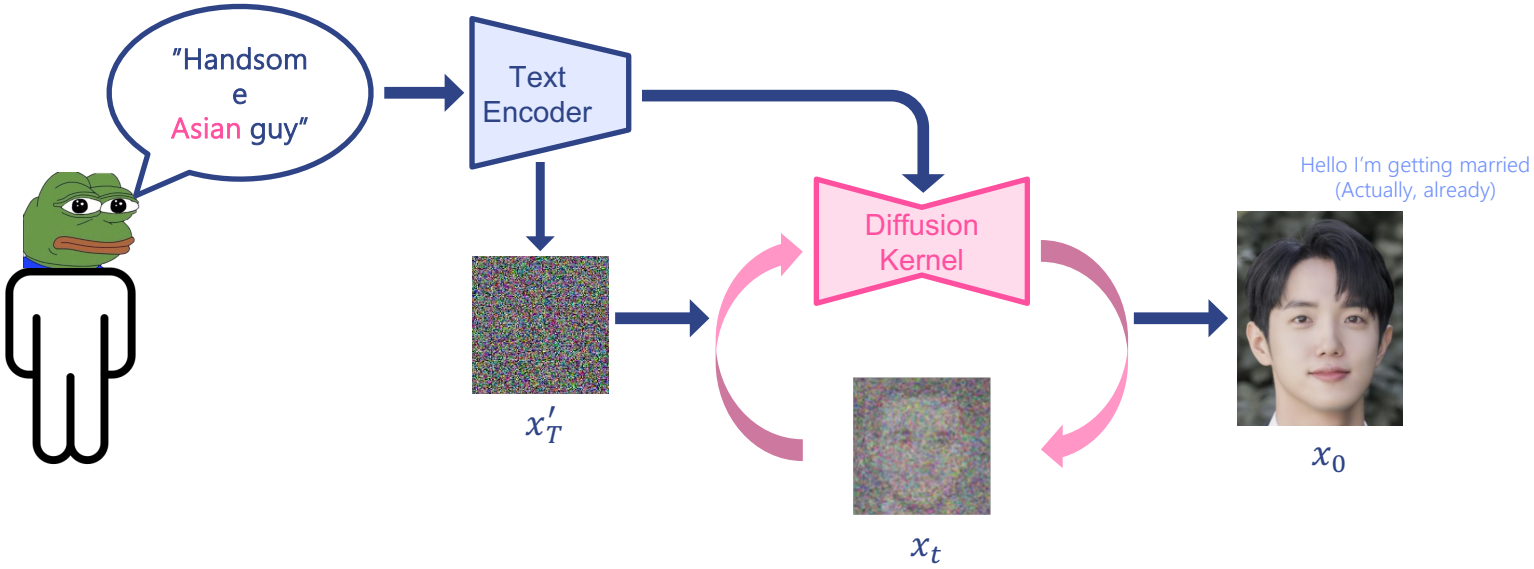
On the other hand,

Think of this scenario:



On the other hand,

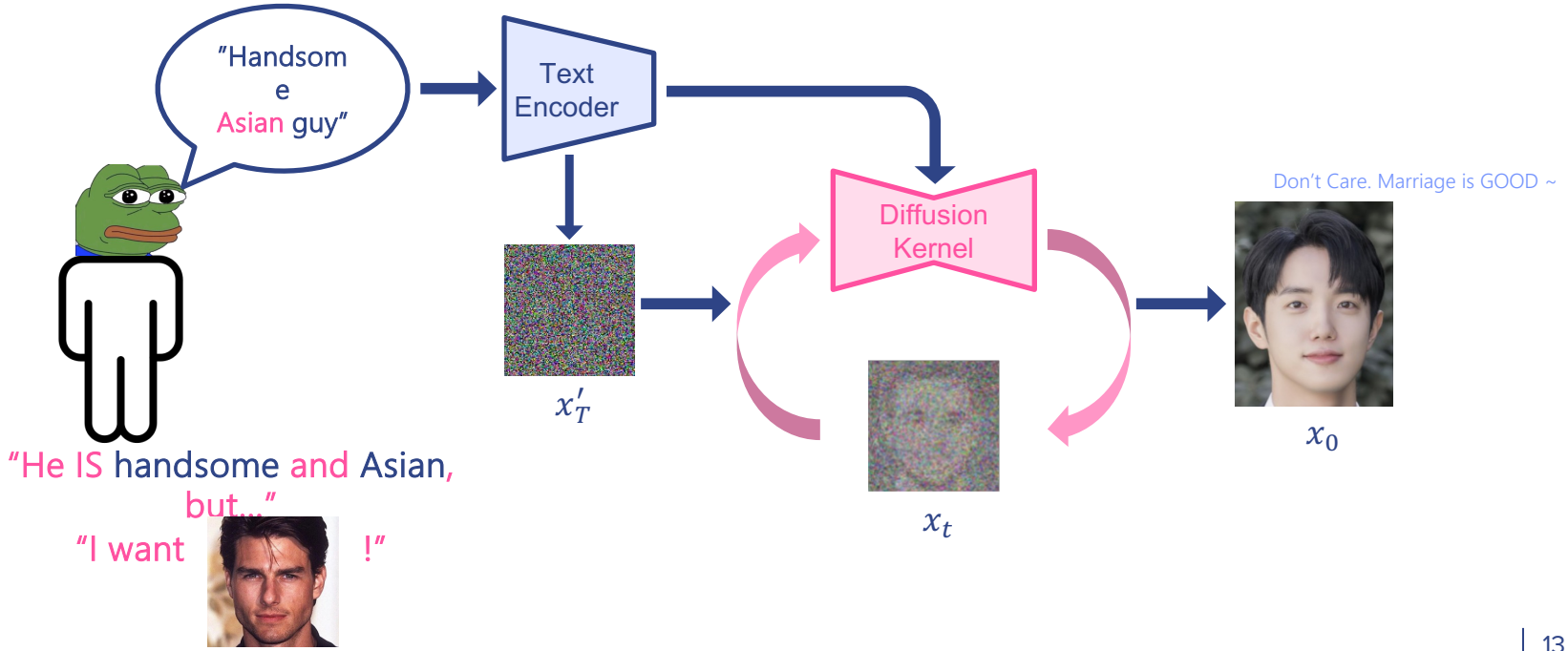
Think of this scenario:



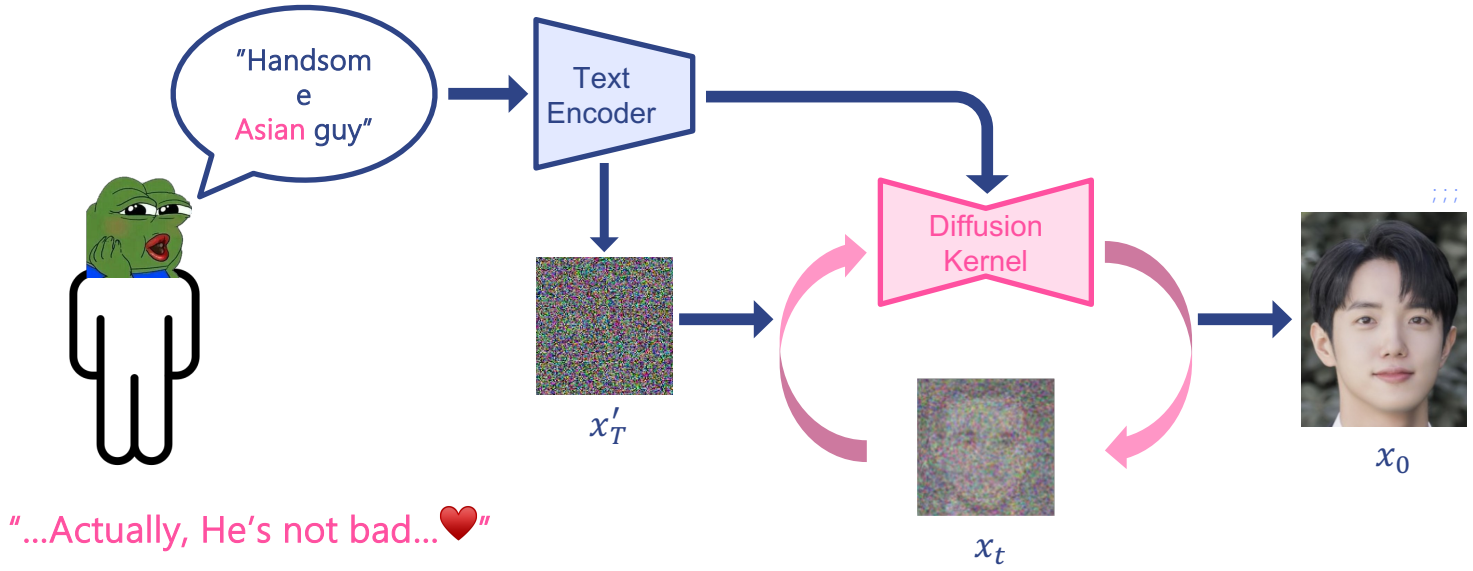
진짜허락받음

On the other hand,

Think of this scenario:

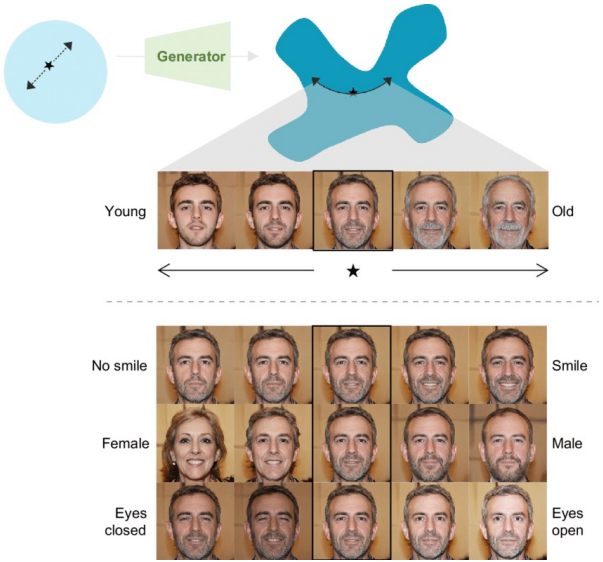


진짜허락받음


"Handsom
e
Asian guy"
Text
Encoder
Diffusion
Kernel
x'_T
x_t
x_0
"...Actually, He's not bad..."

진짜허락받음

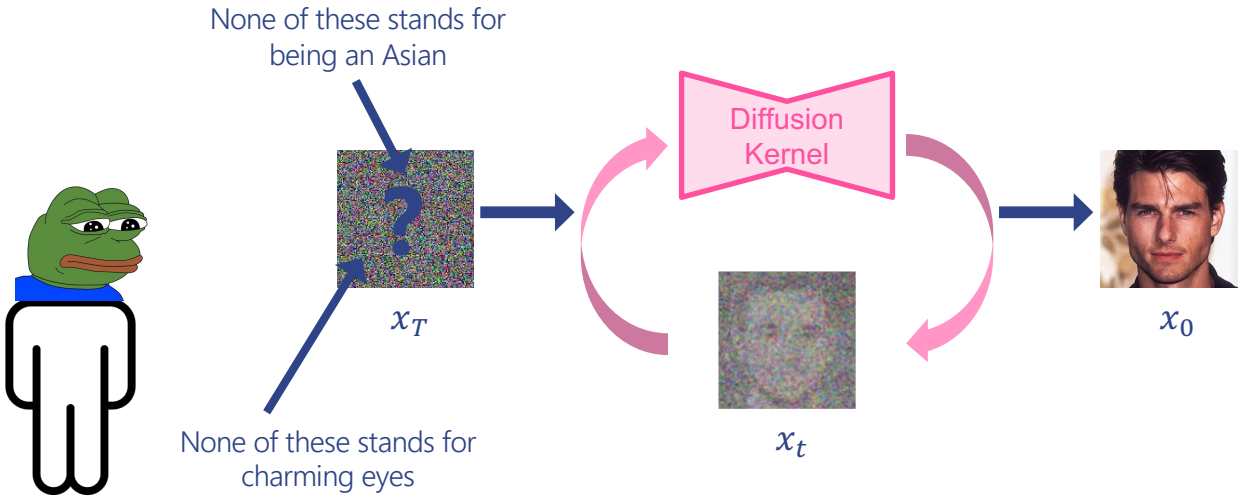
Why can’t simply control its latent?



If it was a GAN, we’d just control certain channel of input latent vector.

[image] 2021, Goetschalcks et al., Generative adversarial networks unlock new methods for cognitive science

Why can’t simply control its latent?



If it was a GAN, we’d just control certain channel of input latent vector.

But for DM, certain pixel of x_T (or x_t) doesn’t stand for any semantic.

Why can’t simply control its latent?

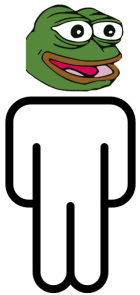
We lack in understanding of its latent manifold.

We cannot control its latent,

even though we can manipulate its intermediate state.

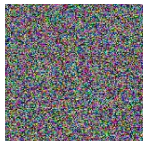
The Goal

Want to edit  ,



with the pretrained & frozen 

!!



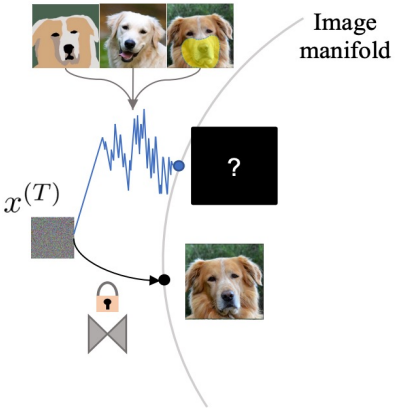
x_T

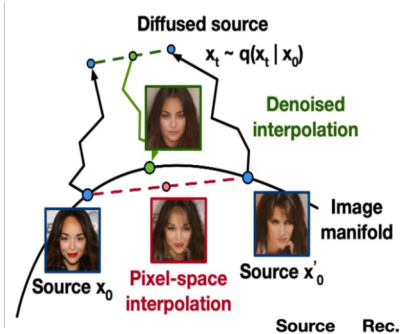
x_t

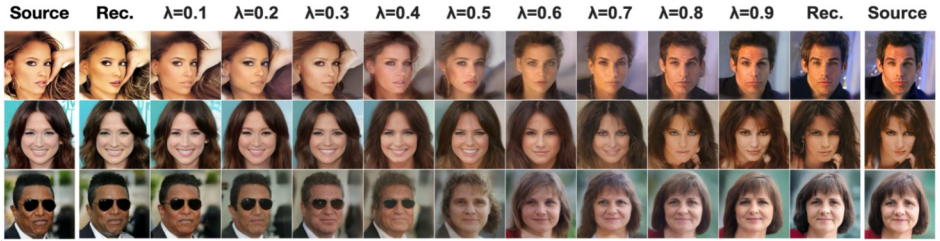
x_0

Other approaches

Interpolate?





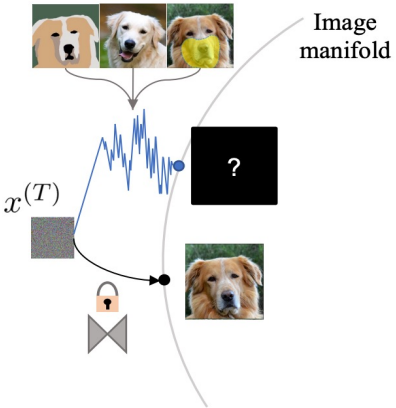


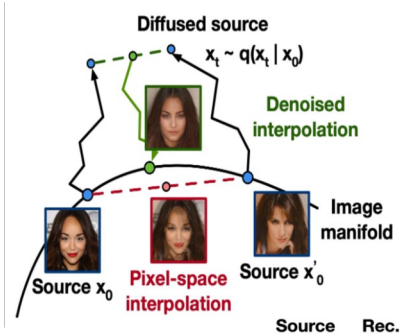
(Example of mixing them in T/2 step)

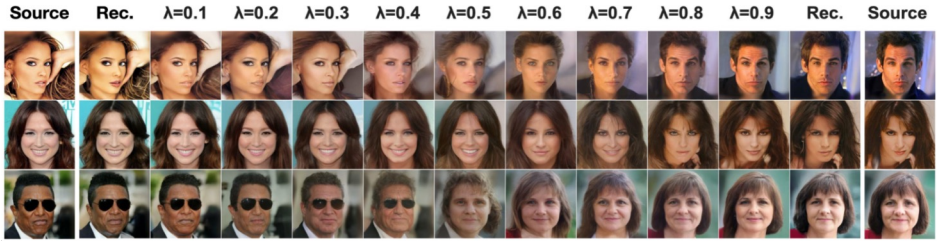
[left] 2023, Kwon et al., Asyrp (Diffusion Models Already Have a Semantic Latent Space)
[right] 2020, Ho et al., DDPM (Denoising Diffusion Probabilistic Models)

Other approaches

Interpolate?





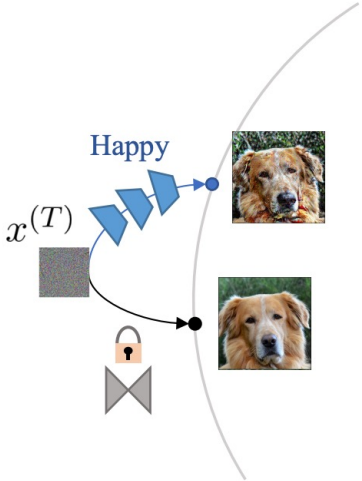


- Contents mixed up, unwanted semantics coming in

[left] 2023, Kwon et al., Asyrp (Diffusion Models Already Have a Semantic Latent Space)
[right] 2020, Ho et al., DDPM (Denoising Diffusion Probabilistic Models)

Other approaches

Classifier guidance?



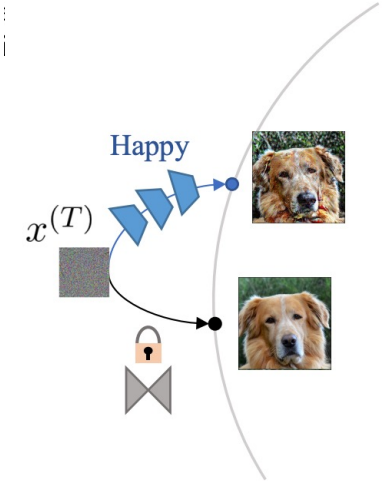
Guiding *x_t* of each timestamp **class-conditioned**:

$$\mu_\theta(x_t) \rightarrow \mu_\theta(x_t) + \Sigma_\theta(x_t) \cdot s \nabla_{x_t} \log \underbrace{f(x_t)}_{\text{Trained classifier}}$$

[left] 2023, Kwon et al., Asyrp (Diffusion Models Already Have a Semantic Latent Space)
[right] 2021, Dhariwal et al., ADM (Diffusion Models Beat GANs on Image Synthesis)

Other approaches

Classifier guidance?



Guiding *x_t* of each timestamp **class-conditioned**:

$$\mu_\theta(x_t) \rightarrow \mu_\theta(x_t) + \Sigma_\theta(x_t) \cdot s \nabla_{x_t} \log \underbrace{f(x_t)}_{\text{Trained classifier}}$$

- Needs extra classifier trained
- Only attributes(classes) of the training set are available

[left] 2023, Kwon et al., Asyrp (Diffusion Models Already Have a Semantic Latent Space)
[right] 2021, Dhariwal et al., ADM (Diffusion Models Beat GANs on Image Synthesis)

+) “Wait, heard of DM becoming deterministic”

“If so, we can keep the contents!”

Yes, diffusion model can be deterministic.

(Actually, may not be a “true” diffusion, strictly- theoretically- speaking?)

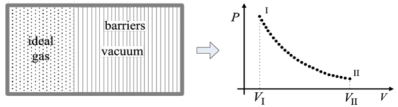
DENOISING DIFFUSION IMPLICIT MODELS

Jiaming Song, Chenlin Meng & Stefano Ermon
Stanford University
{tsong,chenlin,ermon}@cs.stanford.edu

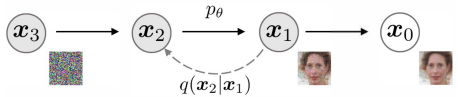
DDIM (ICLR ‘21)

+) Prerequisite: DDIM

Classic DM



Quasi-static process
: Slooooooooow



Markov process

T ~= 1000

!!

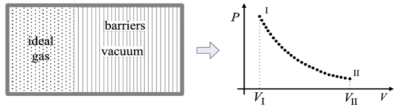
$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right)$$

$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)\right) + \sigma_t \boldsymbol{z}_t$$

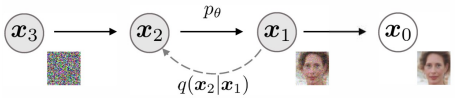
Yep, ignore these

+) Prerequisite: DDIM

Classic DM



Quasi-static process
: Slooooooooow



Markov process

T ~= 1000

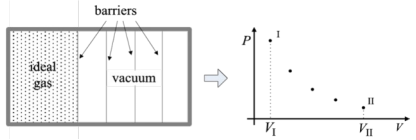
!!

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}\left(\boldsymbol{x}_t; \sqrt{1-\beta_t}\boldsymbol{x}_{t-1}, \beta_t \mathbf{I}\right)$$

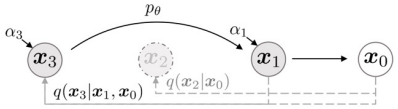
$$\boldsymbol{x}_{t-1} = \frac{1}{\sqrt{1-\beta_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t)\right) + \sigma_t \boldsymbol{z}_t$$

Yep, ignore these

DDIM



Generalized Langevin dynamics ~~blah blah~~
: **Fast!**



Non-Markovian

S	10	20	50	100	200	500	1000
Error	0.014	0.0065	0.0023	0.0009	0.0004	0.0001	0.0001

Much shorter schedule

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right)}_{\text{“predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$

+) Prerequisite: DDIM

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$

Adjustable “stochasticity”

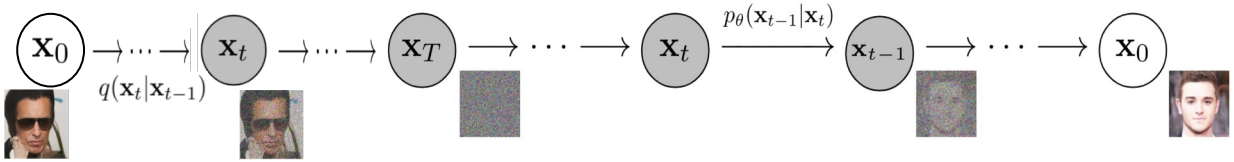
+) Prerequisite: DDIM

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}$$

where $\sigma_t = \eta\sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$

Adjustable “stochasticity”

With η = 1, it becomes a classic DM



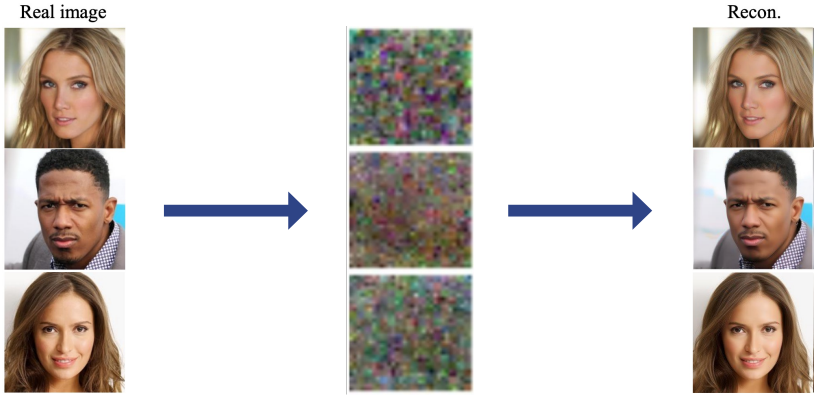
+) Prerequisite: DDIM

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}} \right)}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(\boldsymbol{x}_t)}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}$$

where $\sigma_t = \eta \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)} \sqrt{1-\alpha_t/\alpha_{t-1}}$

Adjustable “stochasticity”

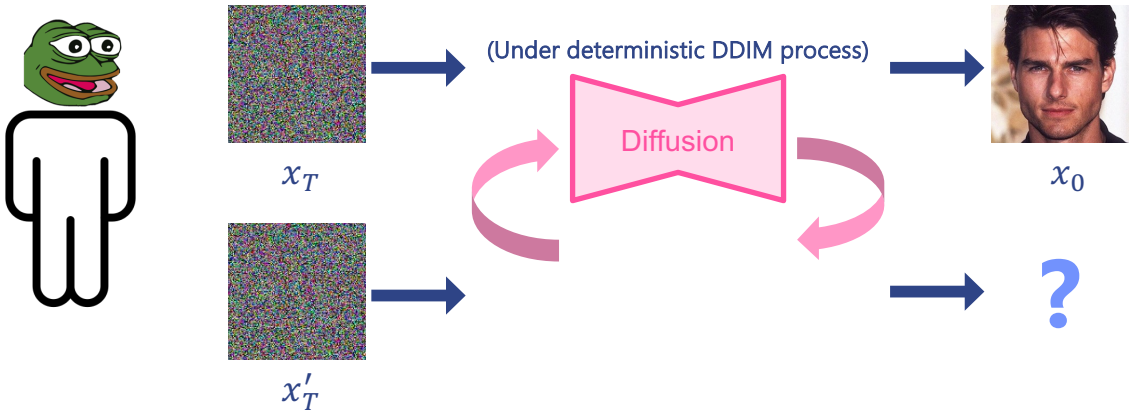
With η = 1, it becomes a classic DM
With η = 0, it becomes “perfect-inversion”, which is Deterministic.



Still no control in semantic

"So, we can keep the contents!"

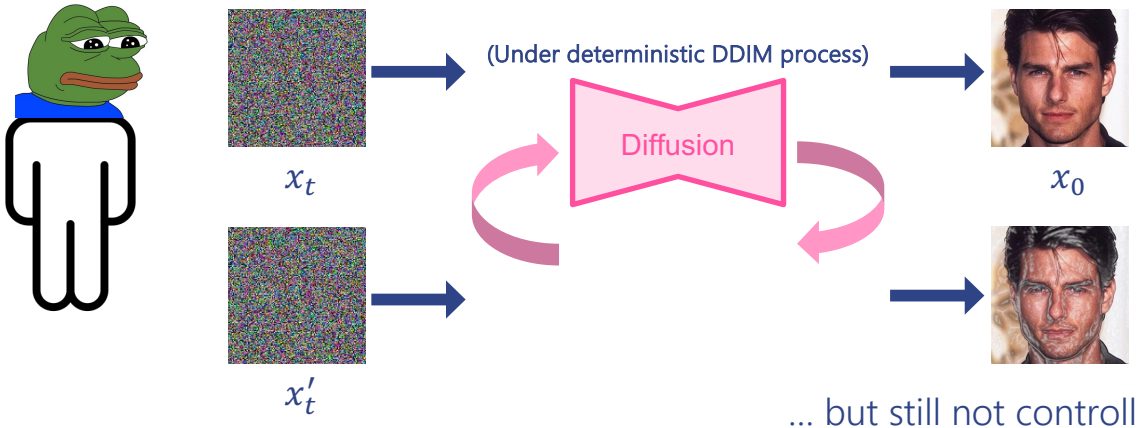
Yes we can perturbate x_t and get some meaningful output…



Still no control in semantic

"So, we can keep the contents!"

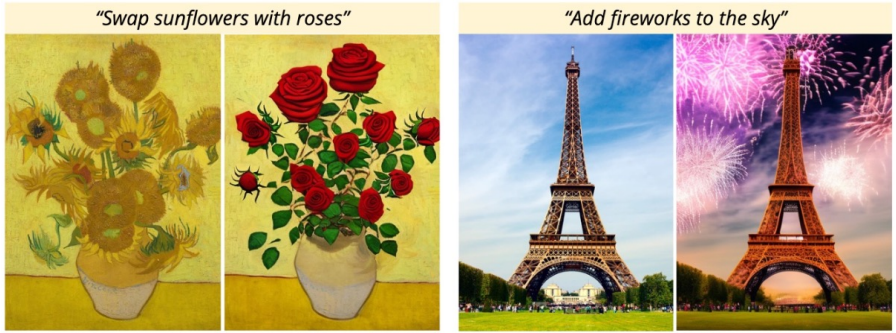
Yes we can perturbate x_t and get some meaningful output…



… but still not controlling specific attribute,
And might get a degraded result *.

* Since we don't know the space of x_t other than x_T.

From this, actually there are existing approaches (and good)



InstructPix2Pix
Learning to Follow Image Editing Instructions

Tim Brooks*, Aleksander Holynski*, Alexei A. Efros
University of California, Berkeley
*Denotes equal contribution

CVPR 2023 (Highlight)

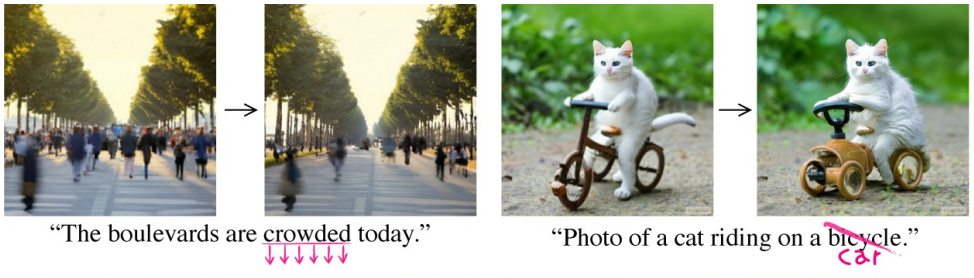
Prompt-to-Prompt Image Editing with Cross Attention Control

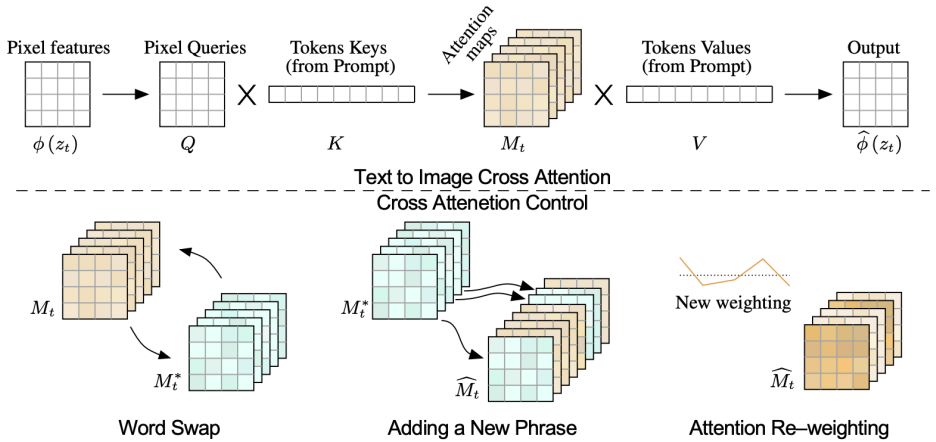
Amir Hertz*[1,2], Ron Mokady*[1,2], Jay Tenenbaum[1], Kfir Aberman[1], Yael Pritch[1], and Daniel Cohen-Or*[1,2]

[1] *Google Research*
[2] *The Blavatnik School of Computer Science, Tel Aviv University*

[top] 2023, InstructPix2Pix: Learning to Follow Image Editing Instructions
[bottom] 2022, Prompt-to-Prompt Image Editing with Cross Attention Control

From this, actually there are existing approaches (and good)





Constrains the new CA map in the kernel
To keep spatial similarity with the original

[image] 2022, Prompt-to-Prompt Image Editing with Cross Attention Control

Actually, the following works are about...

We lack in understanding of its latent manifold.

We want to directly touch the latent, just as we normally do.

Want to know its characteristic. Want to access its space.

Practically, urm..... Well, couldn't it be adopted in future works?

Asyrp (Asymmetric Reverse Process)

Diffusion Models Already Have A Semantic Latent Space

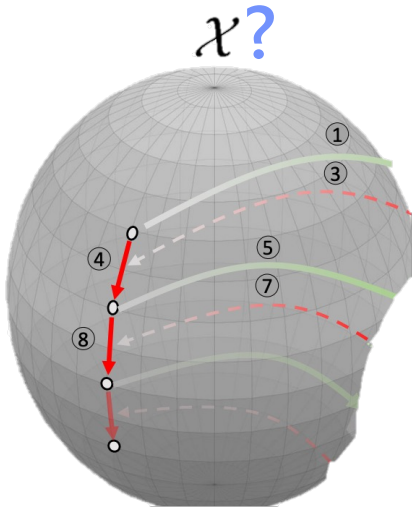
ICLR 2023 (notable-top-25%)

Mingi Kwon Jaeseok Jeong Youngjung Uh

Yonsei University, Korea

Key Idea & Notation

The space of x_t, X is intractable.
-> We cannot achieve editing by manipulating certain x_t of single timestep.



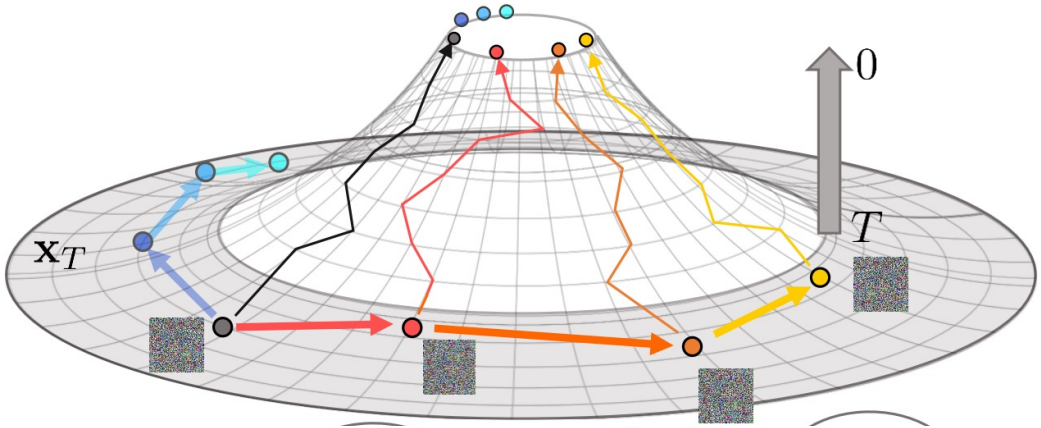
Can’t directly do this.

[figure] 2023, Park et al., Unsupervised Discovery of Semantic Latent Directions in Diffusion Models

Key Idea & Notation

However, it might be possible to guide the inversion trajectory to head into the resulting x_0 with the desired attribute.

-> We might displace updates(=noise ϵ_t) of each timesteps,



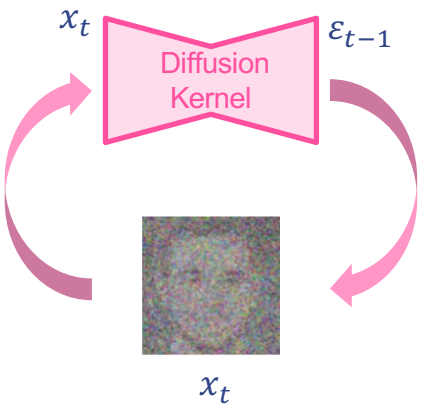
[figure] 2023, Park et al., Unsupervised Discovery of Semantic Latent Directions in Diffusion Models

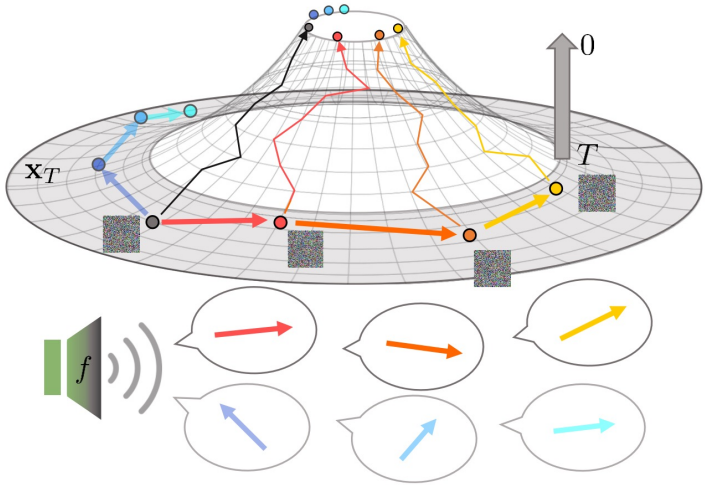
Key Idea & Notation

However, it might be possible to guide the inversion trajectory
to head into the resulting x_0 with the desired attribute.

-> We might displace updates(=noise ϵ_t) of each timesteps,

since we can access some high-level feature in diffusion kernel to manipulate the updates.





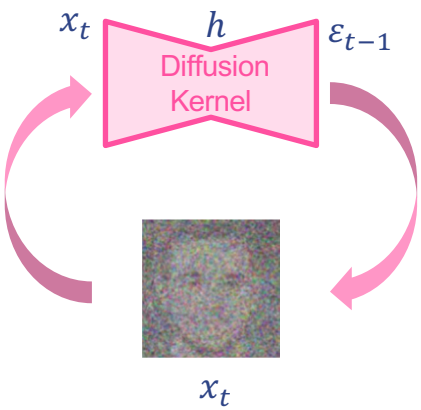
[right] 2023, Park et al., Unsupervised Discovery of Semantic Latent Directions in Diffusion Models

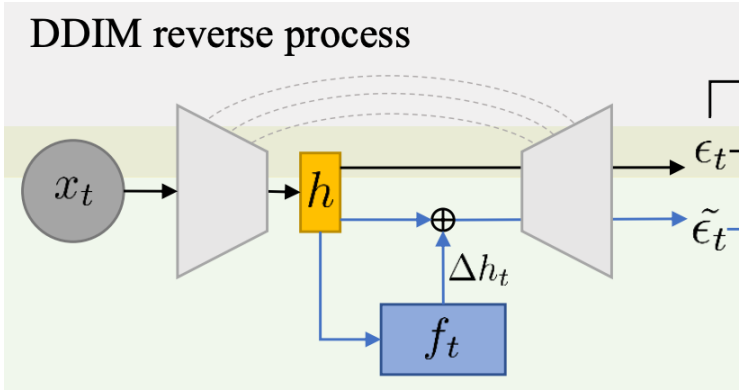
Key Idea & Notation

We access high-level feature in the kernel to manipulate the updates.

Most of diffusion-based methods use basic U-Net for the kernel.

We pick the bottleneck feature h, and somehow train a editor function f_t
which makes Δh_t over h.





[right] 2023, Kwon et al., Asyrp (Diffusion Models Already Have a Semantic Latent Space)

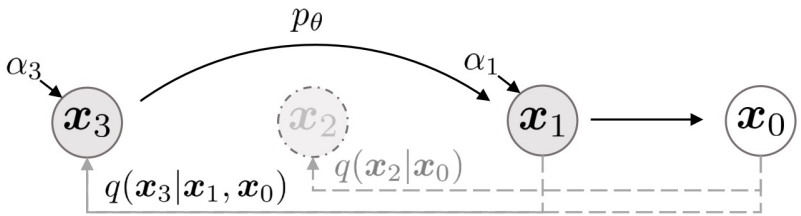
“Asymmetric” Reverse Process

But directly making perturbation in ϵ_t would result in severe distortion.
-> We need to keep a valid diffusion trajectory, yet displacing into the desired semantic direction.

“Asymmetric” Reverse Process

But directly making perturbation in ϵ_t would result in severe distortion.
-> We need to keep a valid diffusion trajectory, yet displacing into the desired semantic direction.

Back to the DDIM process,

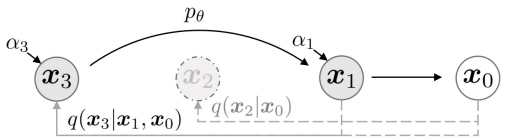


Being an non-Markovian, each update at t ($p_\theta(x_{t-1}|x_t)$)
is dependent to x_t and x_0.

“Asymmetric” Reverse Process

But directly making perturbation in ϵ_t would result in severe distortion.
-> We need to keep a valid diffusion trajectory, yet displacing into the desired semantic direction.

Back to the DDIM process, (Again, leave it at an intuitive level !)



$$x_{t-1} = \sqrt{\alpha_{t-1}} \underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}} \right)}_{\text{“predicted } x_0\text{”}} + \underbrace{\sqrt{1-\alpha_{t-1}-\sigma_t^2} \cdot \epsilon_\theta^{(t)}(x_t)}_{\text{“direction pointing to } x_t\text{”}} + \underbrace{\sigma_t \epsilon_t}_{\text{random noise}}$$

Being an non-Markovian, each update at t ($p_\theta(x_{t-1}|x_t)$)
is dependent to x_t and ~~x_0~~.

predicted x_0

(Think it as a plain denoising result of x_t)

“Asymmetric” Reverse Process

But directly making perturbation in ϵ_t would result in severe distortion.
-> We need to keep a valid diffusion trajectory, yet displacing into the desired semantic direction.

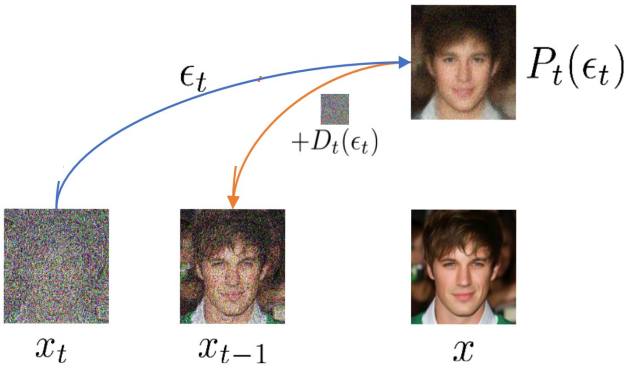
Back to the DDIM process,(Again, leave it at an intuitive level !)

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\boxed{\mathbf{P}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t))} + \boxed{\mathbf{D}_t(\boldsymbol{\epsilon}_t^\theta(\boldsymbol{x}_t))} + \sigma_t \boldsymbol{z}_t$$

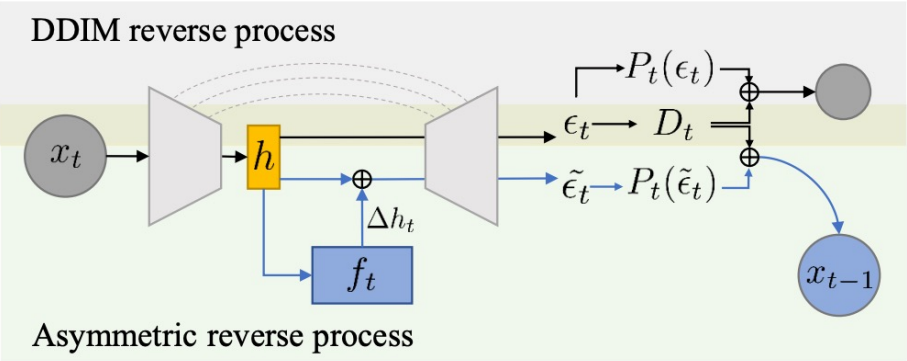
$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\boxed{\left(\frac{\boldsymbol{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}{\sqrt{\alpha_t}}\right)}}_{\text{“ predicted } \boldsymbol{x}_0\text{”}} + \underbrace{\boxed{\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^{(t)}(\boldsymbol{x}_t)}}_{\text{“direction pointing to } \boldsymbol{x}_t\text{”}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}$$

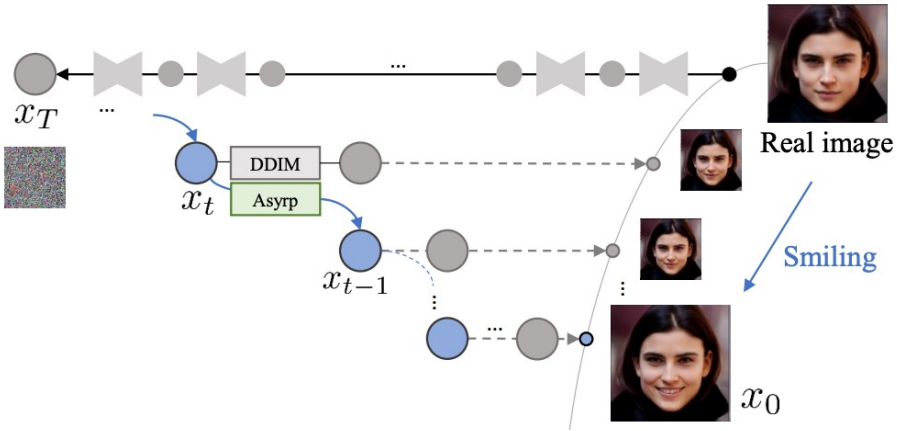
1) “From the very contents of this x_t, plain denoising result would look like this!”

2) “Grounded on that, update(add up) detailed contents towards x_{t-1}.”

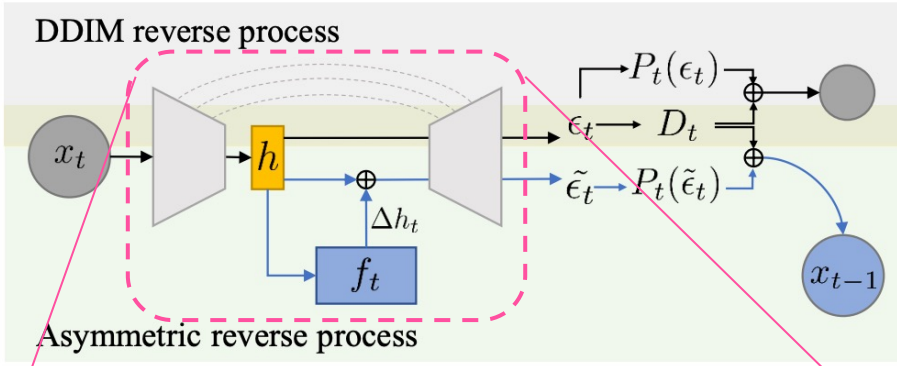


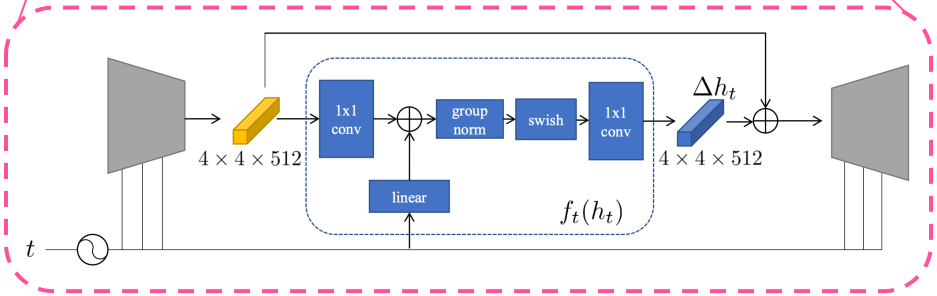
Finally, the process of Asyrp



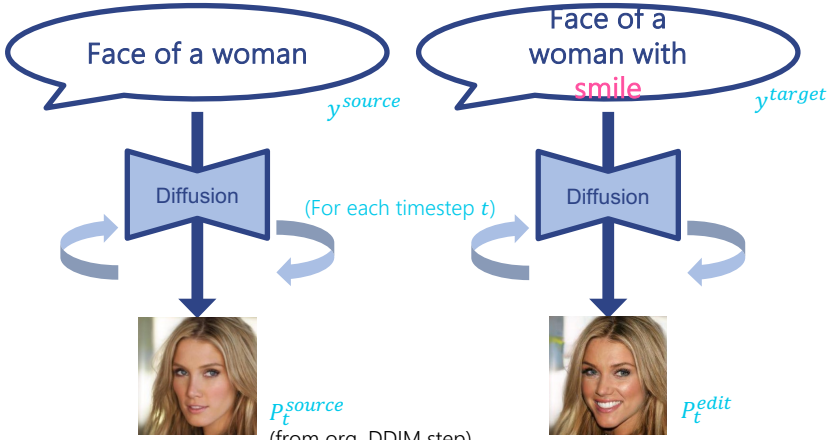


Training f_t





Using CLIP encoder,



$$\mathcal{L}_{\text{direction}}\left(\boldsymbol{x}^{\text{edit}}, y^{\text{target}}; \boldsymbol{x}^{\text{source}}, y^{\text{source}}\right) := 1 - \frac{\Delta I \cdot \Delta T}{\|\Delta I\| \|\Delta T\|}$$

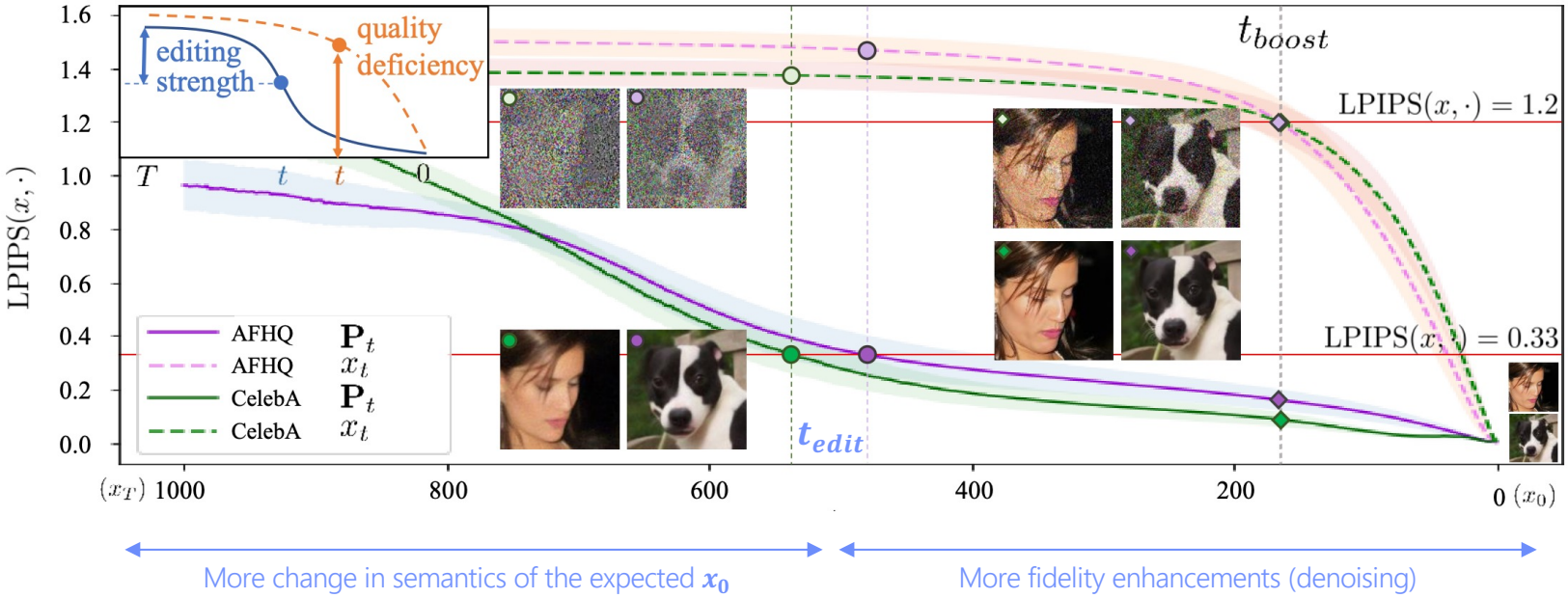
where $\Delta T = E_T\left(y^{\text{target}}\right) - E_T\left(y^{\text{source}}\right)$ and $\Delta I = E_I\left(\boldsymbol{x}^{\text{edit}}\right) - E_I\left(\boldsymbol{x}^{\text{source}}\right)$

$$\mathcal{L}^{(t)} = \lambda_{\text{CLIP}} \mathcal{L}_{\text{direction}}\left(\mathbf{P}_t^{\text{edit}}, y^{\text{ref}}; \mathbf{P}_t^{\text{source}}, y^{\text{source}}\right) + \lambda_{\text{recon}} \left|\mathbf{P}_t^{\text{edit}} - \mathbf{P}_t^{\text{source}}\right|$$

Training f_t

Editing has been taken during $[T, t_{edit}]$

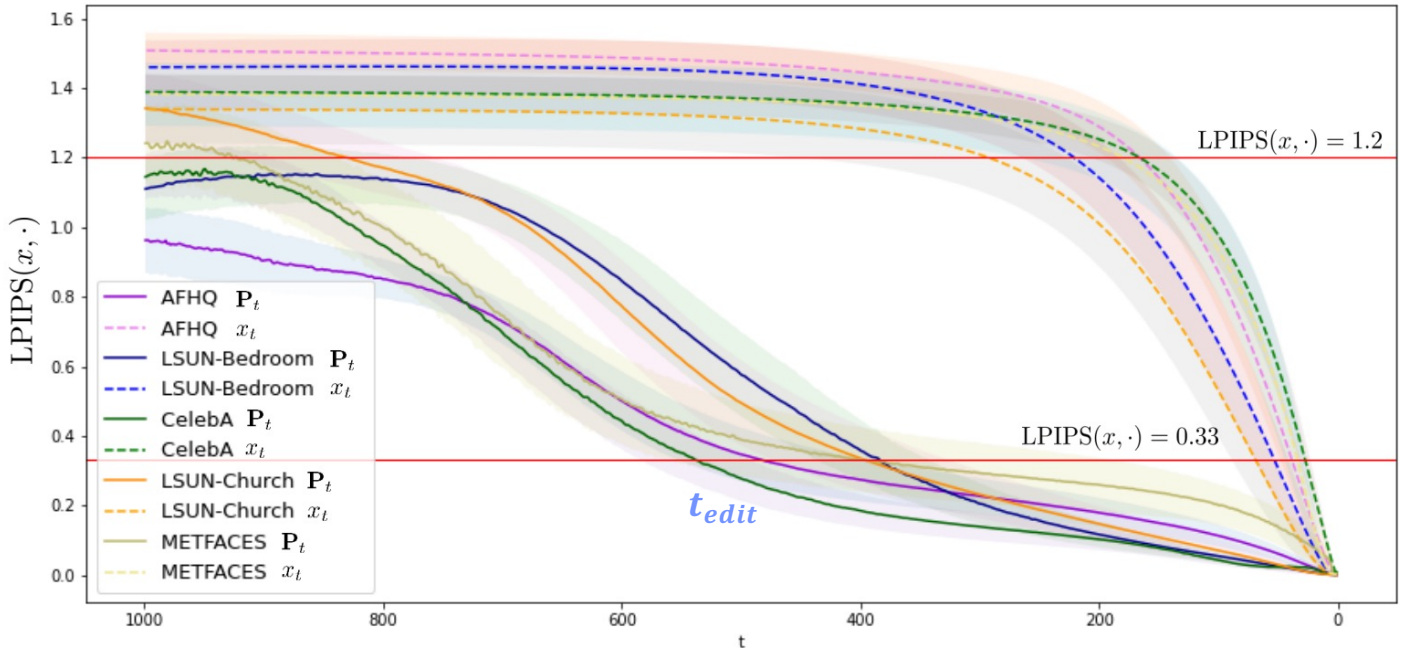
Empirically got t_{edit} with diff(Perceptual metric) becomes near max between x_t and P_t.



Training f_t

Editing has been taken during $[T, t_{edit}]$

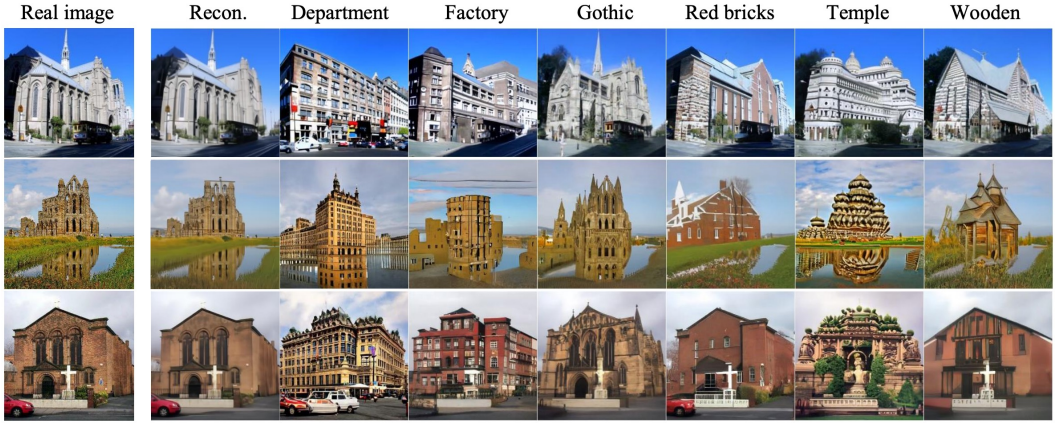
Empirically got t_{edit} with diff(Perceptual metric) becomes near max between x_t and P_t.

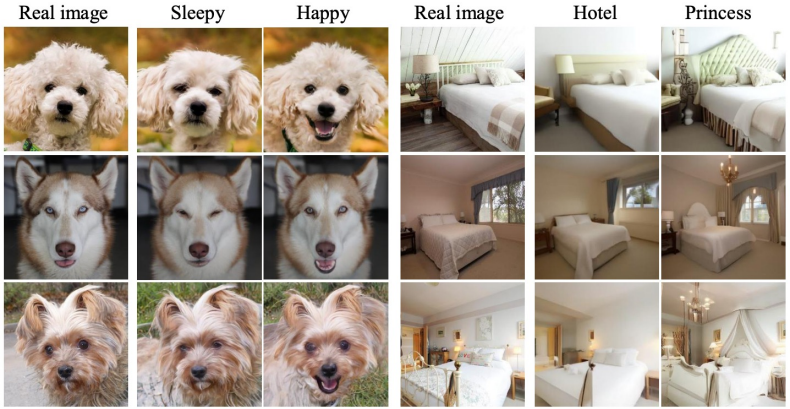


Results

Experimented on the pre-trained, frozen {DDPM++ (2020), iDDPM (2021), ADM (2021)} released

Results





Same Δh_t works for different samples. The "semantic direction" is global.

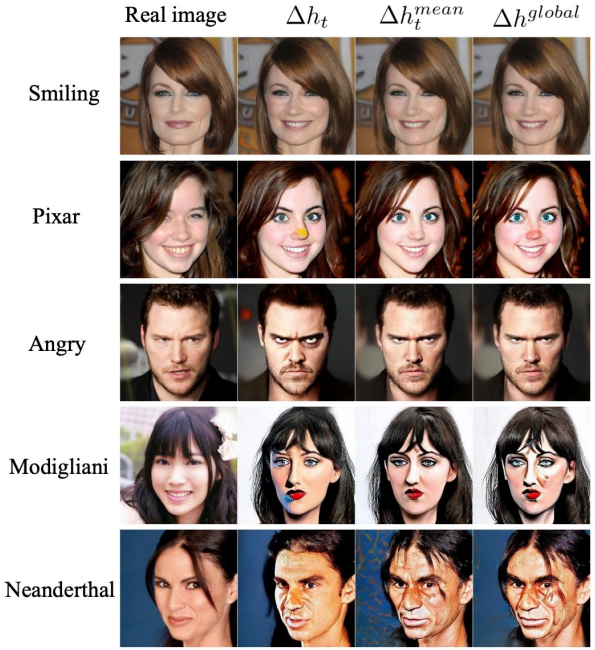
The attributes that were not/rarely available in the pre-training set

- Guiding kernel feature h with CLIP-encoded text
- **Wouldn't be available with classifier-guided method or interpolating in several x_t s.**

Results

Linearly scaling $\Delta \boldsymbol{h}_t$

Linear combination of $\Delta \boldsymbol{h}_t$



H space being:

1) Locally-linear

2) Global though samples

and might be? - 3) Not severely-changing through t

Taking avg. through

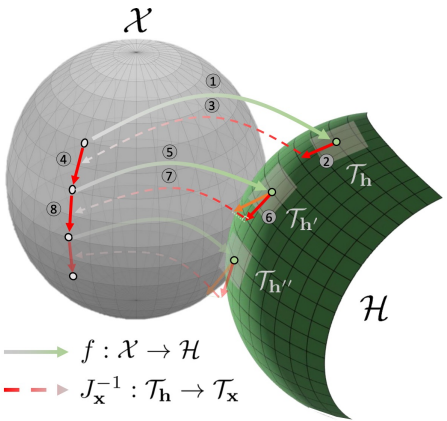
- Samples: Δh_t^{mean}
- Time: Δh^{global}

Unsupervised Discovery of Directions

Little more attempts by themselves, rather quickly just mentioning -

Unsupervised Discovery of Semantic Latent Directions in Diffusion Models

Yong-Hyun Park [*1] **Mingi Kwon** [*2] **Junghyo Jo** [1] **Youngjung Uh** [2]



Based on the previous finding that H space being:

1) Locally-linear

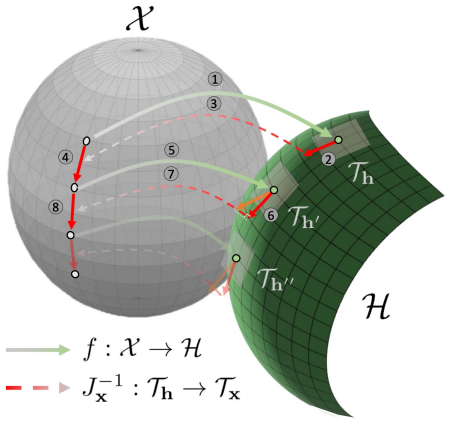
2) Global though samples

and might be? - 3) Not severely-changing through t

Trying to find the semantic direction without training any editor model

Unsupervised Discovery of Directions

Little more attempts by themselves, rather quickly just mentioning -



Trying to find the semantic direction without training any editor model

1. Given the diffusion update, get Jacobian of $\mathcal{H}$ -> $\mathcal{X}$ `torch.autograd.functional.jacobian()`, maybe?
2. Get max N eigenvectors, like taking SVD to the Jacobian…

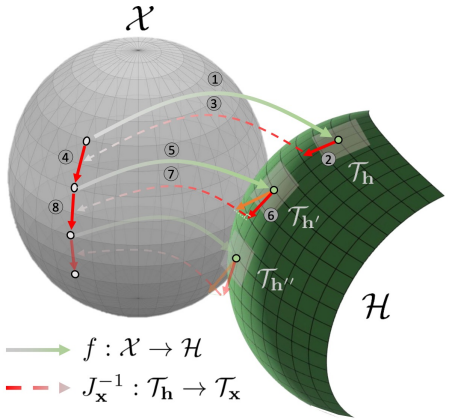
   Find v among tangent at x (T_x) that maximizes $\|v\|^2 \triangleq \langle u, u \rangle_h = v^T \mathcal{J}_x^T \mathcal{J}_x v$

   Find next v, keeping orthogonal to the found ones

   …
3. Moving through the found directions, apply some more cautious mathematical considerations…… curvature…… remapping to tangent space… blah blah……

Unsupervised Discovery of Directions

Little more attempts by themselves, rather quickly just mentioning -



Trying to find the semantic direction without training any editor model

...Okay! Let's just say,

It's possible to analyze directions in DM latent space,

even though it's intractable (unlike GANs)

leveraging that directions in $\mathcal{H}$ is global.

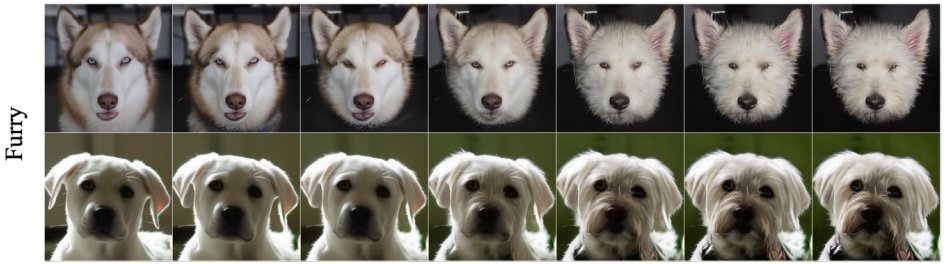
It's about giving more evidence for the properties of DM latent, rather than for practical use.

Unsupervised Discovery of Directions

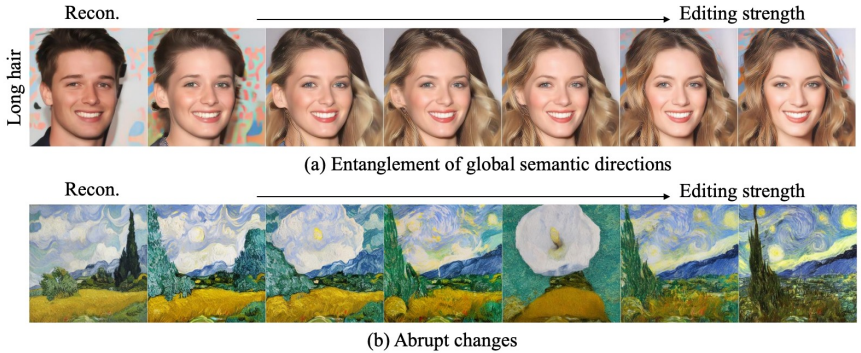
- With “manually found” directions, they could catch semantics.
- Was able to globally, time-consistently edit the generation process.

(Again, note that this is not a special skill for usual AI.
It’s a proxy to enable the things only DMs could not.)

Unsupervised Discovery of Directions



Show success on some,



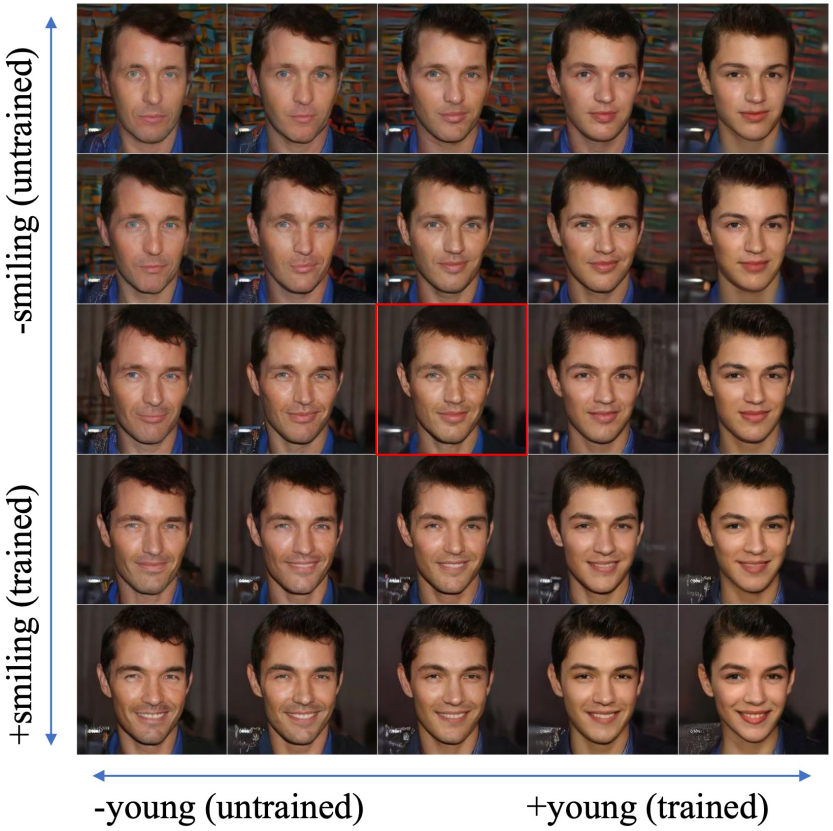
(a) Entanglement of attributes

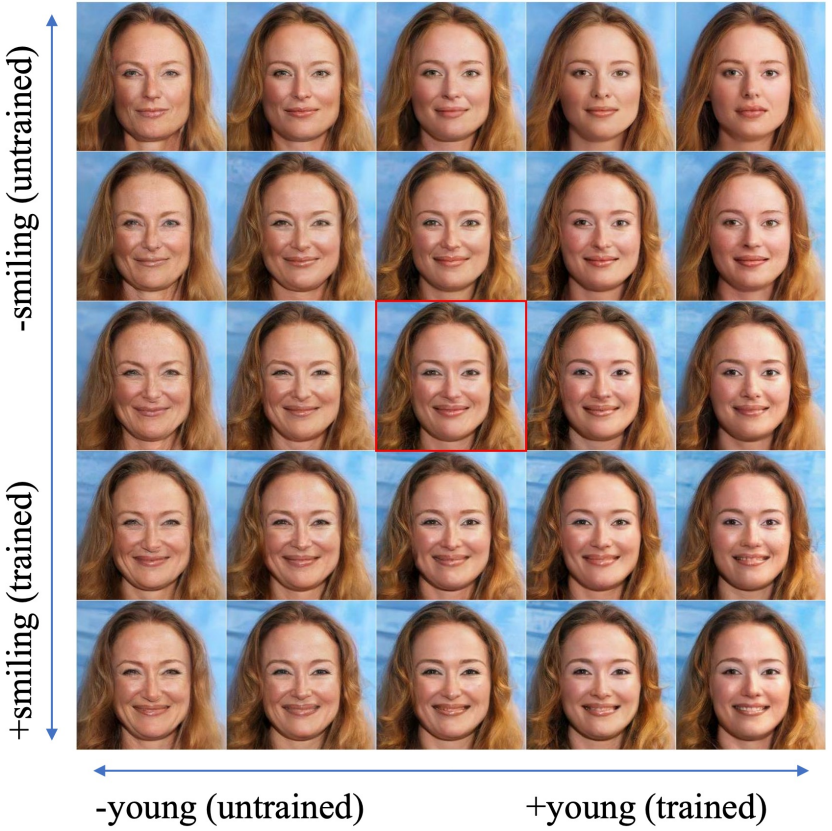
(b) Sudden changes in models like Stable Diffusion

- maybe since it sends the latent into a more complex space

But limitations also

Back to… Just more results form Asyrp





Back to… Just more results form Asyrp

Back to... Just more results form Asyrp

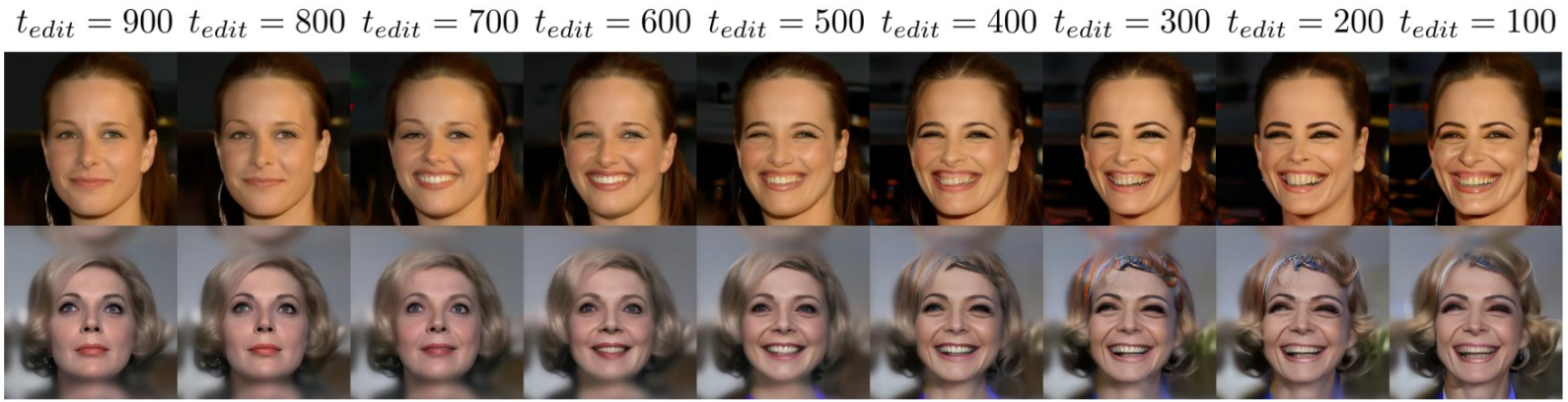


Figure 18: **Importance of choosing proper t_{edit}.** We explore various t_{edit} with `smiling`. Too short editing interval struggles to manipulate attributes. Excessive editing strength results in degraded images.

Again,

We lack in understanding of its latent manifold.

We want to directly touch the latent, just as we normally do.

Want to know its characteristic. Want to access its space.

Practically, urm..... Well, couldn't it be adopted in future works?

- The h editor could be also trained for general attributes (getting E_text)
  Still, it needs another training step

- Lacks in experiments for larger sets / varying domain with captions given

- The unsupervised way cannot target certain attribute a user want
  Hence cannot argue about practical use

Still,

We lack in understanding of its latent manifold.

We want to directly touch the latent, just as we normally do.

- The same $\Delta \boldsymbol{h}$ leads to the same effect on different samples.
- Linearly scaling $\Delta \boldsymbol{h}$ controls the magnitude of attribute change, even with negative scales.
- Adding multiple $\Delta \boldsymbol{h}$ manipulates the corresponding multiple attributes simultaneously.
- $\Delta \boldsymbol{h}$ preserves the quality of the resulting images without degradation.
- $\Delta \boldsymbol{h}_t$ is roughly consistent across different timesteps t.

More effective future method?
Enable sub-tasks (say, transfer learning of DM kernel?)
Enable a way of control regardless of LLM?

Thanks!

Thanks!

Thanks!

Thanks!

Thanks!

Thanks!

Thanks!

Thanks!