ILVR: Conditioning Method for Denoising Diffusion Probabilistic Models

Jooyoung Choi[1], Sungwon Kim[1], Yonghyun Jeong[3], Yongjune Gwon[3], Sungroh Yoon[1,2,*]

[1] Data Science and AI Laboratory, Seoul National University, Korea
[2] ASRI, INMC, and Interdisciplinary Program in AI, Seoul National University, Korea
[3] AI Research Team, Samsung SDS

Accepted at ICCV

2023.05.04

Medical Imaging & Intelligent Reality Lab.
Convergence Medicine/Radiology,
University of Ulsan College of Medicine
Asan Medical Center

Jihoon Jung

Abstract

- DDPMs excel in unconditional image generation, but their inherent stochasticity makes it difficult to generate images **with specific desired semantics**.
- In this work, we propose Iterative Latent Variable Refinement (ILVR), a method to guide the generative process in DDPM to generate high-quality images **based on a given reference** image.
- The proposed ILVR method generates high-quality images **while controlling the generation**.
- The controllability of our method allows adaptation of a single DDPM **without any additional learning** in various image generation tasks.

Introduction

- There are mainly two approaches to control generative models to generate images as desired:
    1. One is by designing the conditional generative models for the desired purpose.
    2. The other is by **leveraging well-performed unconditional** generative models.
- The second approach involves using high-quality generative models(**StyleGAN** or **BigGAN**) to manipulate semantic attributes of images through latent space analysis or perform image editing by projecting images into the latent space.
- DDPM is an iterative generative model that performs **well in unconditional** image generation, but **controlling it** to generate images with desired semantics is **challenging** due to the **stochasticity** of transitions.
- The proposed **learning-free method**, iterative latent variable refinement (ILVR), utilizes a given reference image to refine each transition in sampling and ensure the given condition, resulting in **high-quality** images **sharing desired semantics**.

Introduction

- Our paper makes the following contributions:
  1. We propose ILVR, a method of refining each transition in the generative process(sampling) by **matching each latent variable with given reference image**.
  2. We investigate **several properties** that allows user controllability on semantic similarity to the reference.
  3. We demonstrate that our ILVR enables leveraging unconditional DDPM in various image generation tasks including **multi-domain image translation, paint-to-image, and editing with scribbles**.

Method

1. Iterative Latent Variable Refinement

   - In this section, ILVR is introduced as **a method for conditioning** the generative process in the unconditional DDPM model. This technique generates images **sharing high-level semantics from reference images** by sampling from the conditional distribution $p(x_0|c)$ with condition c.

$$p_\theta(x_0|c) = \int p_\theta(x_{0:T}|c)dx_{1:T},$$

$$p_\theta(x_{0:T}|c) = p(x_T)\prod_{t=1}^{T} p_\theta(x_{t-1}|x_t, c).$$

   - Our ILVR provides condition c to unconditional transition $p_\theta(x_{t-1}|x_t)$ without additional learning or models. Specifically, we refine each unconditional transition with a downsampled reference image.

Method

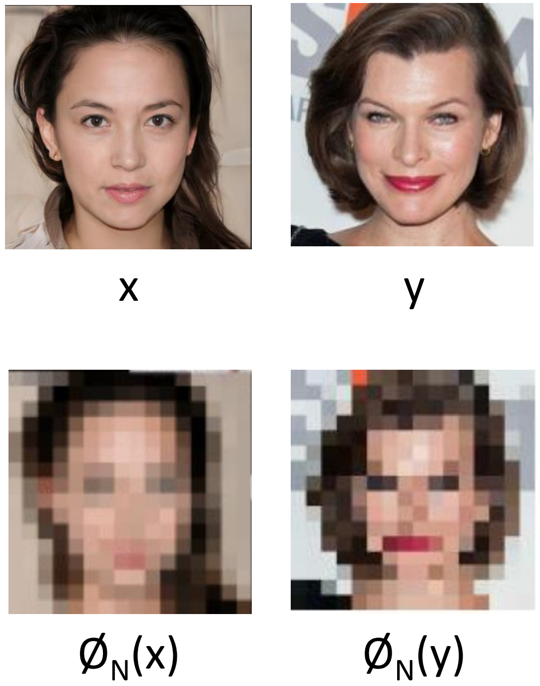
https://github.com/assafshocher/ResizeRight

1. Iterative Latent Variable Refinement

- Let $\phi_N(\cdot)$ denote a **linear low-pass filtering operation**, a sequence of downsampling and upsampling by a factor of N, therefore maintaining dimensionality of the image.

Algorithm 1 Iterative Latent Variable Refinement

Input: Reference image y
Output: Generated image x
$\phi_N(\cdot)$: low-pass filter with scale N
Sample $x_T \sim N(\mathbf{0}, \mathbf{I})$
for $t = T, ..., 1$ **do**
    $\mathbf{z} \sim N(\mathbf{0}, \mathbf{I})$
    $x'_{t-1} \sim p_\theta(x'_{t-1}|x_t)$ ▷ unconditional proposal
    $y_{t-1} \sim q(y_{t-1}|y)$ ▷ condition encoding
    $\boxed{x_{t-1} \leftarrow \phi_N(y_{t-1}) + x'_{t-1} - \phi_N(x'_{t-1})}$
end for
return x_0



Method

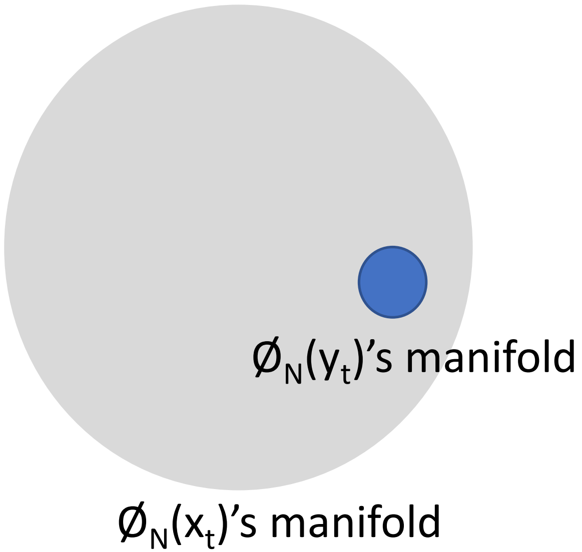
1. Iterative Latent Variable Refinement

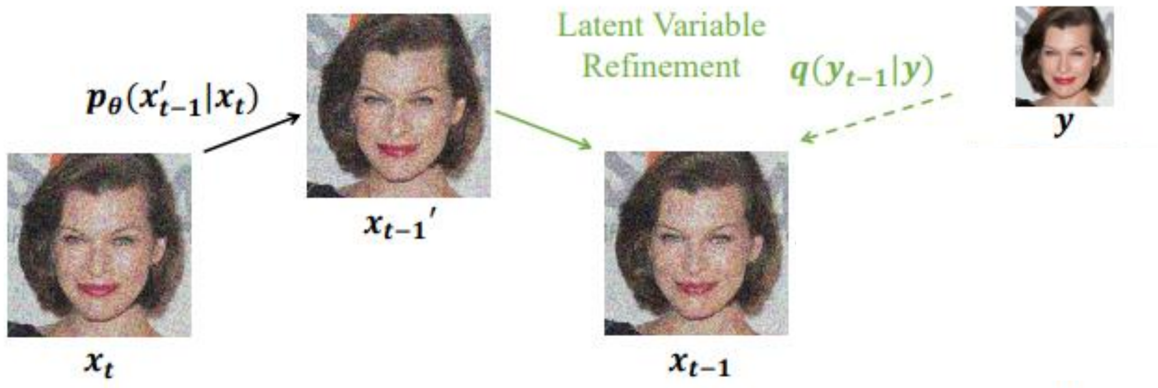
- Utilizing the forward process $q(x_t|x_0)$ and the linear property of ϕ_N , each Markov transition under the condition c is approximated as follows:

  $$p_\theta(x_{t-1}|x_t,\ c) \approx p_\theta(x_{t-1}|x_t, \phi_N(x_{t-1}) = \phi_N(y_{t-1}))$$

- If we perform sampling within the manifold of $\phi_N(y)$,

  then “ $p_\theta(x_{t-1}|x_t,\ c)$ ” approximates “ $p_\theta(x_{t-1}|x_t, \phi_N(x_{t-1}) = \phi_N(y_{t-1}))$ ”.





Method

1. Iterative Latent Variable Refinement

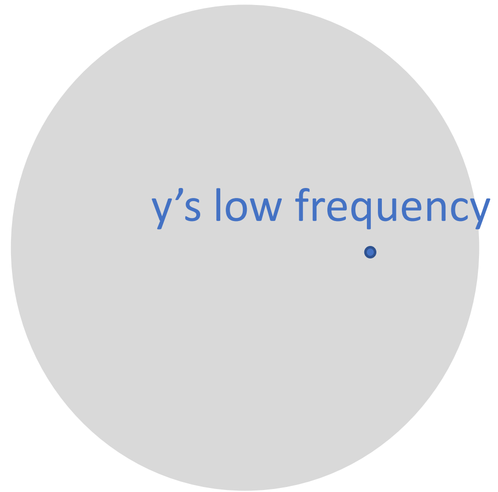
- The condition c in each transition from x_t to x_{t-1} **can be replaced with a local condition**, wherein latent variable x_{t-1} and corrupted reference $\varnothing_N(y_{t-1})$ share low-frequency contents.
- Then, since operation $\varnothing$ maintains dimensionality,

  we refine $p_\theta(x'_{t-1} \mid x_t)$ by matching $\varnothing(x'_{t-1})$ of the x'_{t-1} with $\varnothing(y_{t-1})$ of y_{t-1} as follows :

$$\boxed{\begin{aligned} x'_{t-1} &\sim p_\theta(x'_{t-1}|x_t), \\ x_{t-1} &= \phi(y_{t-1}) + (I - \phi)(x'_{t-1}). \end{aligned}}$$

- ILVR ensures local conditions by matching latent variables,

  enabling conditional generation with unconditional DDPM.



X's low frequency manifold

Method

2. Reference selection and user controllability

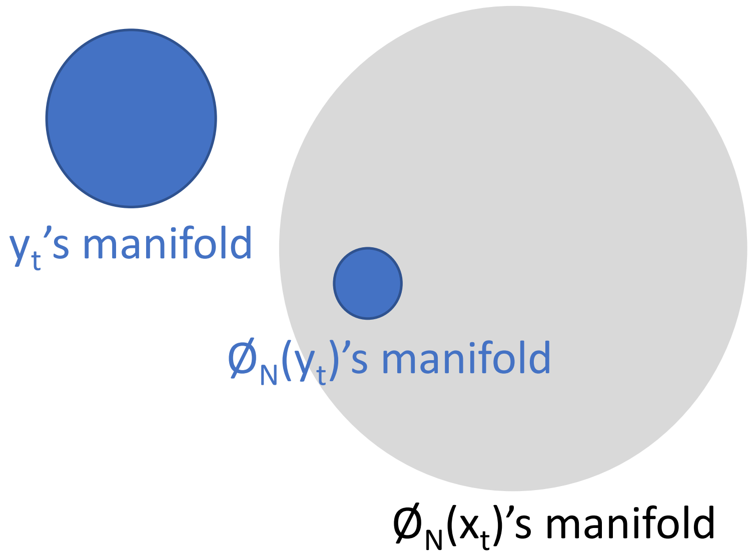
- Our method allows us to sample images from **the subset of images dictated by the reference** image.
- We present some properties to control the reference image.

Property 1.

$$Y = \{y : \phi_N(y) = \phi_N(x), x \in \mu\}$$

The reference image only needs to match the low-resolution space of learned data distribution.

Even reference images from unseen data domains are possible.



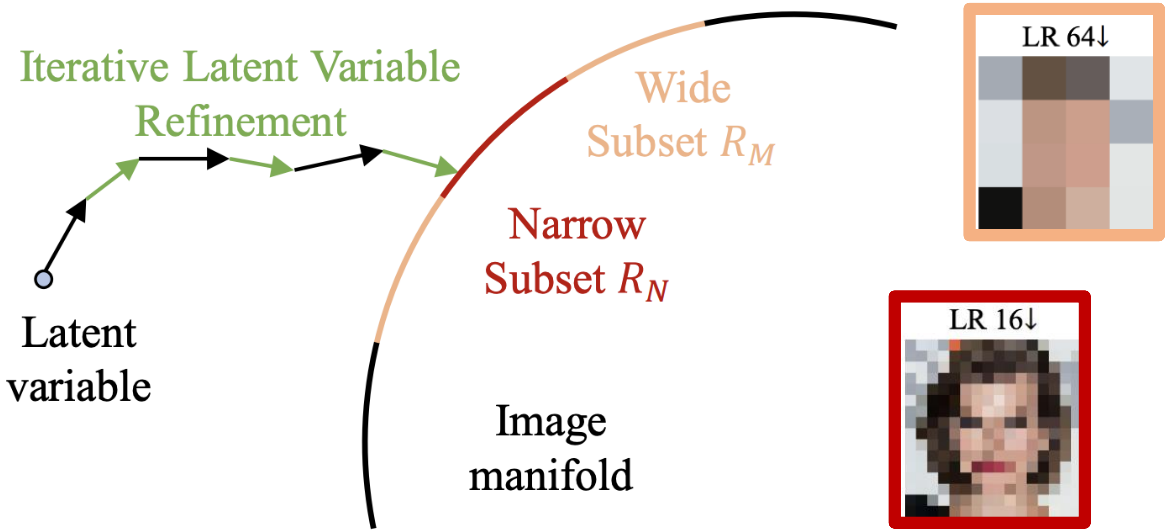
Method

2. Reference selection and user controllability

   Property 2. : Considering downsampling factors N and M where N ≤ M,

   $$R_N \subset R_M \subset \mu,$$

   The larger the downsampling fator, the smaller the effect of conditioning on the image.



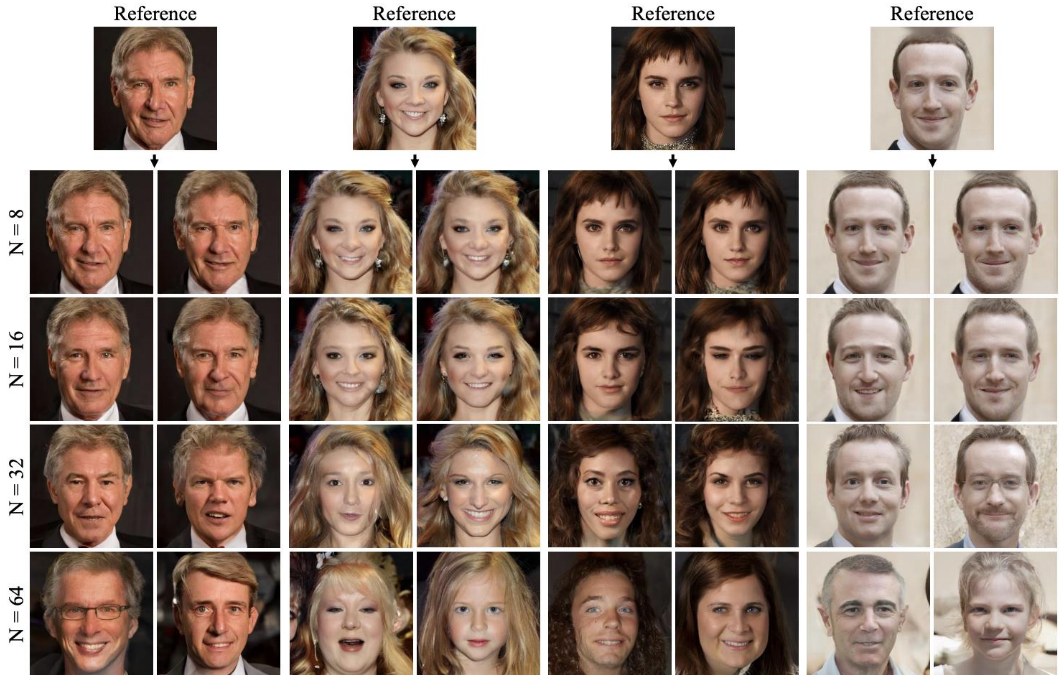
Method

2. Reference selection and user controllability

Property 2. : Considering downsampling factors N and M where N ≤ M,

$$R_N \subset R_M \subset \mu,$$

The larger the downsampling fator, the smaller the effect of conditioning on the image.

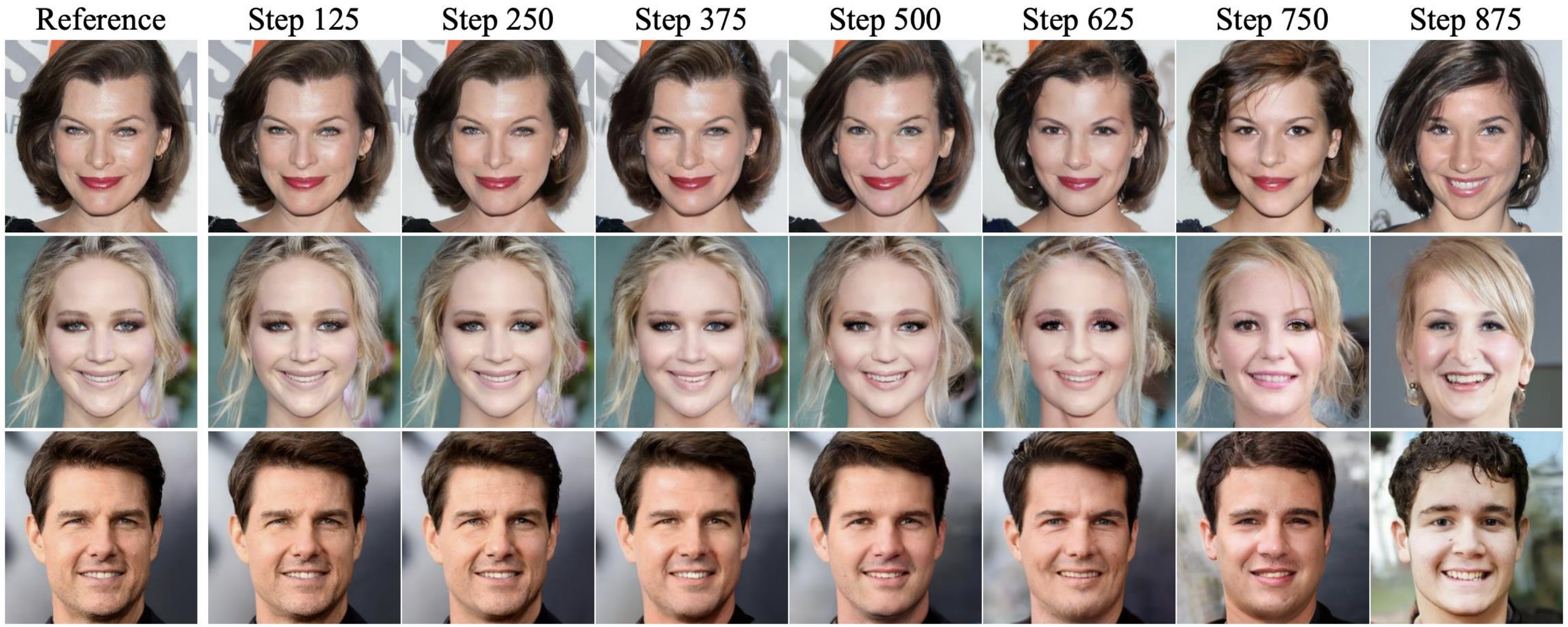


Method

2. Reference selection and user controllability

   Property 3. : Limiting the range of conditioning steps enables sampling from a broader subset, while sampling from learned image distribution is still guaranteed.

$$R_N \subset R_{N,(T,k)} \subset \mu.$$



Experiments

- **Dataset and training :**

  Here we describe datasets and training details. For all datasets, we trained at 2562 resolution with a batch size 8.

  a) **FFHQ** consists of 70,000 high-resolution **face images**. We trained a model for 1.2M steps.

     METFACES consists of 1,000 high-resolution **portrait images**. To avoid overfitting, we fine-tuned a model pre-trained on FFHQ, for 20k steps.

  b) **AFHQ** consists of 15,000 high-resolution animal face images, which are equally split into three categories: dog, cat, and wild. We trained on the **train set of dog** category, then used **test sets of three categories** as reference images to demonstrate multi-domain image translation.

  c) **Places365** consists of 10M images of over 400 scene categories. We **trained a model on a waterfall category**, which consists of 5,000 images. We used this model to **paint-to-image task**. Paintings used for paint-to-image task are collected from the web.

  d) **LSUN Church** consists of 126,227 images of churches. We trained a model for 1M steps.

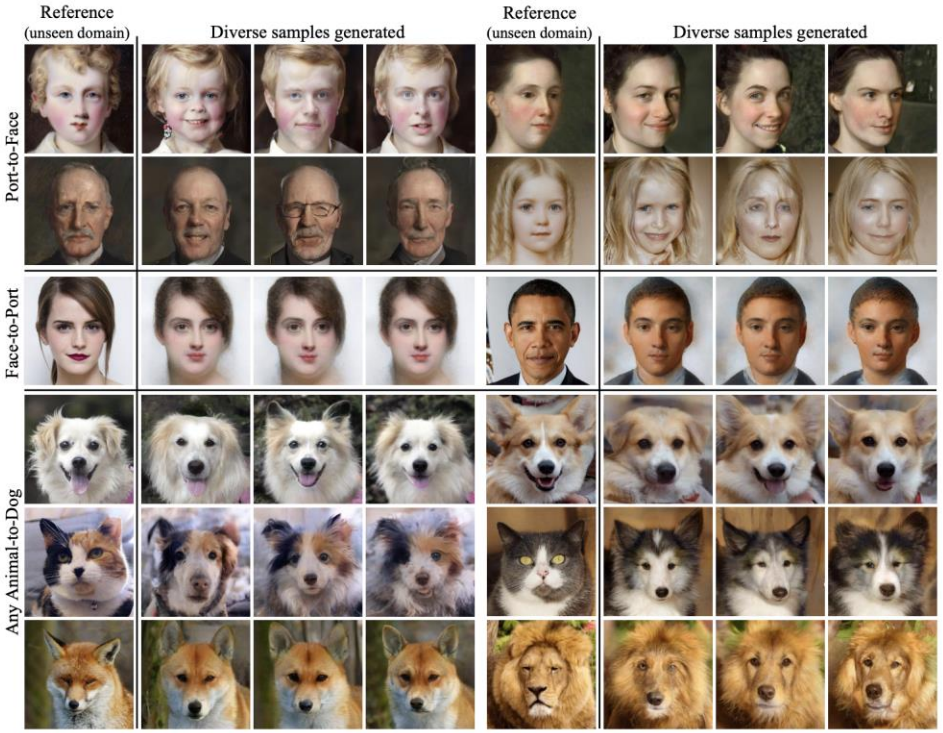
Experiments

- **Image translation from various source domains.**

  Row1-2: portrait to face. Row3: face to portrait. Row4-6: any animal (dog, cat, wildlife) to dog.

  Our method enables translation from any source domains, unseen in the training phase.

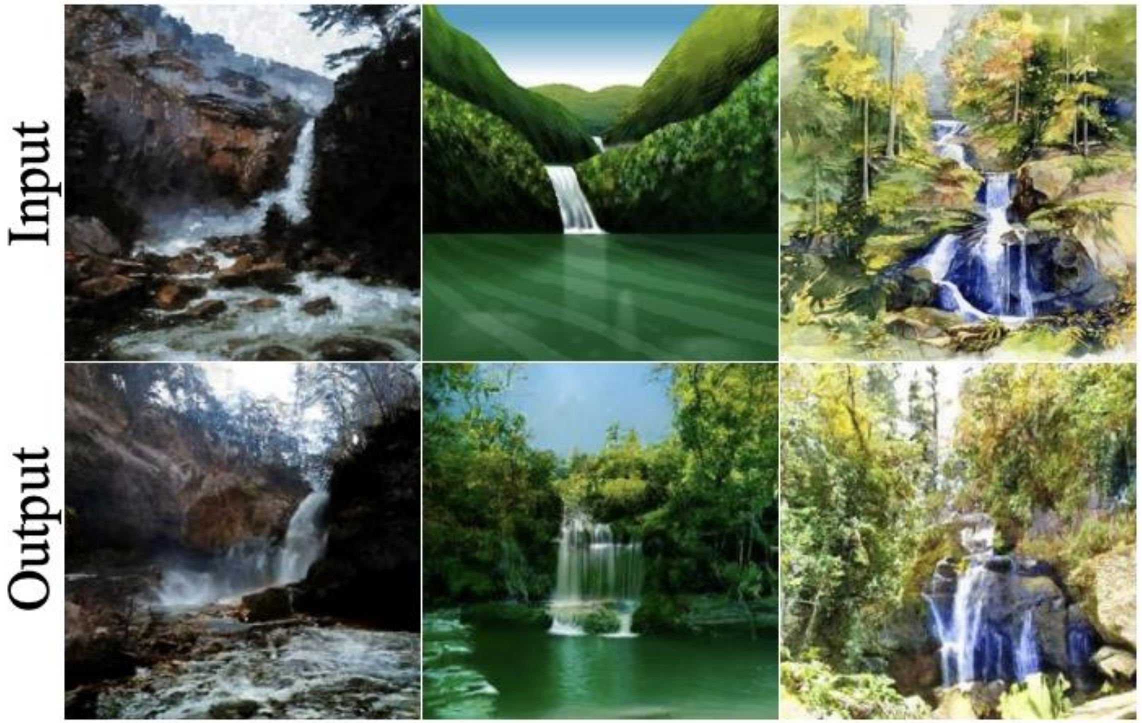
  Moreover, our method generate diverse samples.



Experiments

- **Paint-to-Image.**

  Phot-realistic images generated from unnatural images (oil painting, water color, clip art)



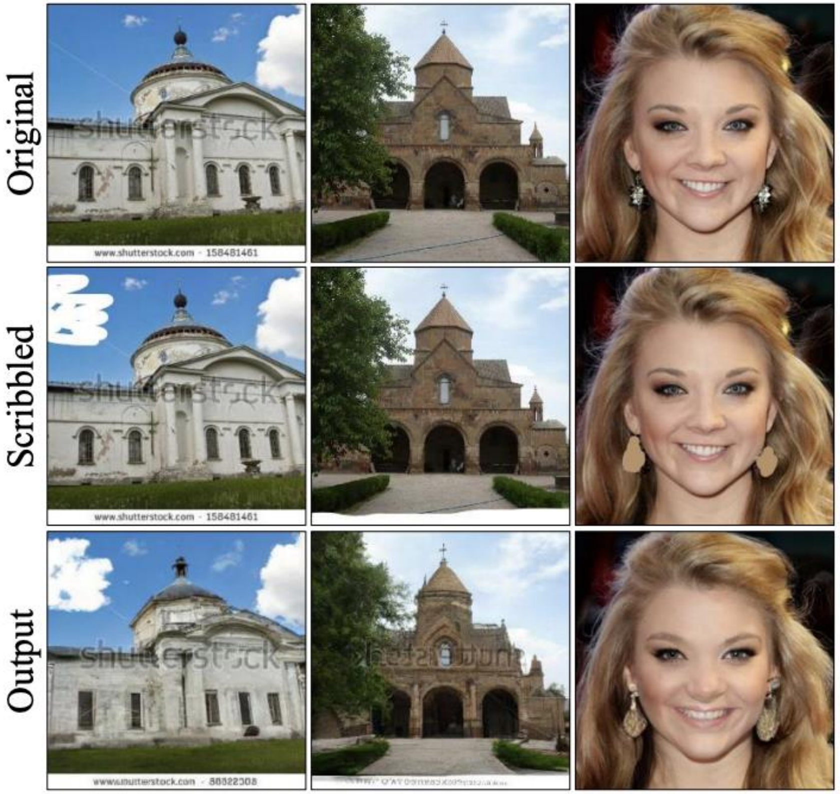
Experiments

- **Editing with scribbles.**

  Row1: cloud generated from white scribble at top-left corner.

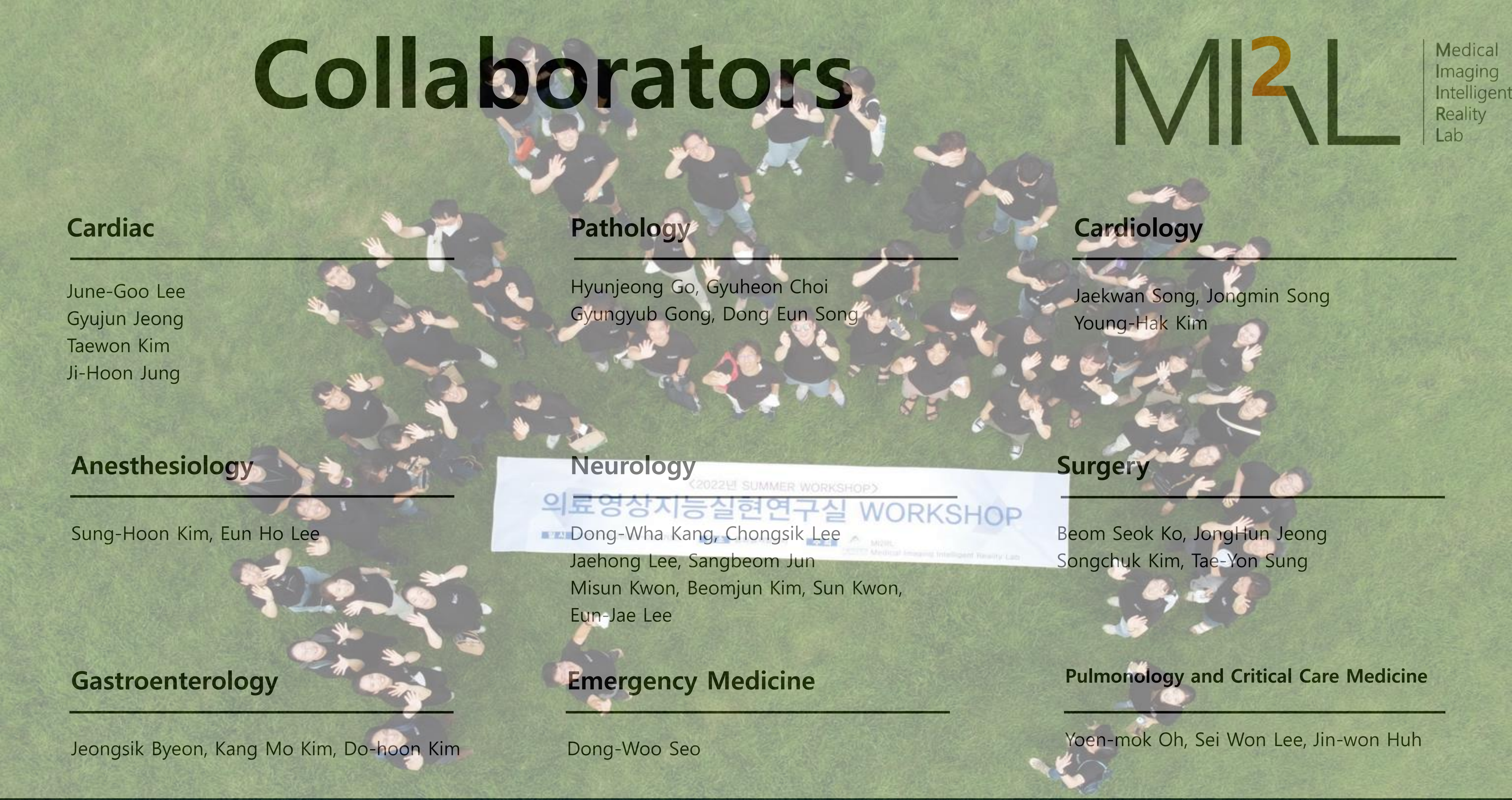
  Row2: watermark generated from white scribble at the bottom.

  Row3: beige earring generated from scribble at top of silver earring.



Conclusion

- We proposed a **learning-free method** of conditioning the generation process of unconditional DDPM.
- Further, **downsampling factors** and the **conditioning range** provide **user controllability** over this method.
- By refining each transition with given reference, we **enable sampling** from the space of plausible images.
- We demonstrated that a single unconditional DDPM **can be leveraged to various tasks** without any additional learning and models.



Collaborators

MI2RL
Medical Imaging Intelligent Reality Lab

Cardiac

June-Goo Lee
Gyujun Jeong
Taewon Kim
Ji-Hoon Jung

Pathology

Hyunjeong Go, Gyuheon Choi
Gyungyub Gong, Dong Eun Song

Cardiology

Jaekwan Song, Jongmin Song
Young-Hak Kim

Anesthesiology

Sung-Hoon Kim, Eun Ho Lee

Neurology

Dong-Wha Kang, Chongsik Lee
Jaehong Lee, Sangbeom Jun
Misun Kwon, Beomjun Kim, Sun Kwon,
Eun-Jae Lee

Surgery

Beom Seok Ko, JongHun Jeong
Songchuk Kim, Tae-Yon Sung

Gastroenterology

Jeongsik Byeon, Kang Mo Kim, Do-hoon Kim

Emergency Medicine

Dong-Woo Seo

Pulmonology and Critical Care Medicine

Yoen-mok Oh, Sei Won Lee, Jin-won Huh