VAE + DDPM

* Introduction.

DiffuseVAE는 VAE와 DDPM을 합쳐, image blur 와 low-dim의 표현 력에 대해 보완한다.

stage 1 : VAE 훈련, stage 2 : diffusion 훈련

Contribution :

1. 새로운 architecture 제시

   ↳ DDPM은 refiner. ⇒ 효과적인 latent space,

2. low-dim latent space에서 controllable synthesis

3. sampling speed up

4. Generalization to auxiliary task.

5. SOTA

* Background.

- VAE

$$\mathcal{L}(\theta, \phi) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - \mathcal{D}_{KL}[q_\phi(z|x) \| p(z)]$$

- DDPM.

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$$

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

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

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mu_\theta(x_t, t), \Sigma_\theta(x_t, t))$$

* DiffuseVAE

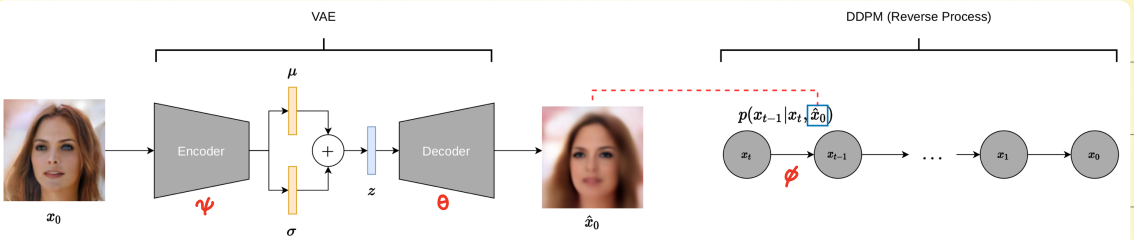


Figure 2: Proposed DiffuseVAE generative process under the simplifying design choices discussed in Section 3.2. In this setting, the VAE encoder takes the original image x_0 as input. The DDPM reverse process in the second stage is conditioned on the VAE reconstruction obtained in the first stage as discussed in Section 3.2.

- Training Objective

high-res image와 x_0, condition signal y, latent z 나온 뒤에,

$$p(x_{0:T}, y, z) = p(z)p_\theta(y|z)p_\phi(x_{0:T}|y, z)$$

$p(x_{1:T}, z|y, x_0)$는 intractable. → $q(x_{1:T}, z|y, x_0)$

$$q(x_{1:T}, z|y, x_0) = q_\psi(z|y, x_0)q(x_{1:T}|y, z, x_0)$$

log-likelihood에서,

$$\log p(x_0, y) = \log \int p(x_{0:T}, y, z) dx_{1:T} dz$$

ELBO를 적용.

$$\log p(x_0, y) \geq \underbrace{\mathbb{E}_{q_\psi(z|y,x_0)}[p_\theta(y|z)] - \mathcal{D}_{KL}(q_\psi(z|y, x_0)||p(z))}_{\mathcal{L}_{\text{VAE}}} + \mathbb{E}_{z\sim q(z|y,x_0)}\left[\underbrace{\mathbb{E}_{q(x_{1:T}|y,z,x_0)}\left[\frac{p_\phi(x_{0:T}|y, z)}{q(x_{1:T}|y, z, x_0)}\right]}_{\mathcal{L}_{\text{DDPM}}}\right] \tag{9}$$

↳ Appendix B

- Simplifying design choices.

Unconditional의 경우. model 간단화

1. $P_\theta(x_{0:T}|z)$
2. eq 9 에서 2단계로 학습, stage 1은 VAE, 2는 DDPM
3. $\hat{x}$ 에서 reverse process.
   - $\hat{x}_0$가 아닌, 다른 samples로 학습 가능
   - → fig 2.

- VAE parameterization.

2단계 훈련이 flexibility 라므로 stage 1은 유연하게 구성가능. ($\hat{x}_0$만 생성하면 됨)

- DDPM parameterization.

  - Formulation 1.

    $\hat{x}$ 란 z가 forward process 에서 independent 라고 가정.

    - $q(x_{1:T}|z, x_0) \simeq q(x_{1:T}|x_0)$

    reverse process transition의 경우 VAE recon에만 conditionally dependent 라고 가정.

    - $p(x_{0:T}|z) \approx p(x_{0:T}|\hat{x}_0)$

    x_t 에서 x_{t-1}을 얻음

  - Formulation 2.

    forward process transition은 VAE recon이 conditionally dependent 라고 가정.

    - $q(x_{1:T}|z, x_0) \simeq q(x_{1:T}|\hat{x}_0, x_0)$

    reverse process transition의 경우 VAE recon에만 conditionally dependent 라고 가정.

    - $p(x_{0:T}|z) \approx p(x_{0:T}|\hat{x}_0)$

VAE의 reconstruction은 forward process transition을 포함하도록 설계할 수 있다.

$$q(x_1|x_0,\hat{x}_0) = \mathcal{N}(\sqrt{1-\beta_1}x_0 + \hat{x}_0, \beta_1 I) \tag{10}$$
$$q(x_t|x_{t-1},\hat{x}_0) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1} + (1-\sqrt{1-\beta_t})\hat{x}_0, \beta_t I) \quad \text{where} \quad t>1$$

forward conditional marginal은

$q(x_t|x_0,\hat{x}_0) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x_0 + \hat{x}_0, (1-\bar{\alpha}_t)I)$ → Appendix B.2.

↳ 잘 schedule 된 β_t 일 때, $\bar{\alpha}_T \approx 0$, → $q(x_T|x_0,\hat{x}_0) \approx \mathcal{N}(\hat{x}_0, I)$

↳ $\mathcal{N}(\hat{x}_0, I)$가 x_T가 됨을 의미.

↳ DDPM에서 몇 개 수정해서 사용함.

↳ Appendix C

↳ Z를 사용하는 DDPM의 $\hat{x}_0$을 추가하는 방법도 있지만 여기서는 ×.

* Related Work.

- Unconditional DDPM
- Conditional DDPM
- VAE based model

Hierarchical VAE는 posterior collapse가 일어날 수 있다.

↳ gradient skipping, spectral normalization 이 사용되기도 한다.

↳ high-fidelity를 위해선 large latent가 필요함.

DiffuseVAE의 경우, 단일 latent만 필요함. (GAN과 비슷)

VAEBM과 LSGM도, NVAE를 사용했지만, 이는 low-dim latent code가 복잡하다.

* Experiments.

Hyperparam은 Appendix D.

- Generator-refiner framework.

Fig 3는 DiffuseVAE의 stage1이 formulation의 stage2로 적용.

(a) Formulation-1 (b) Formulation-2

Figure 3: Illustration of the generator-refiner framework in DiffuseVAE. The VAE generated samples (Bottom row) are refined by the Stage-2 DDPM model.

↳ 둘다 성능 좋음.

- Controllable synthesis via low-dimensional DiffuseVAE latents.

VAE의 latent인 z_{VAE}와 x_T에 대해 discuss

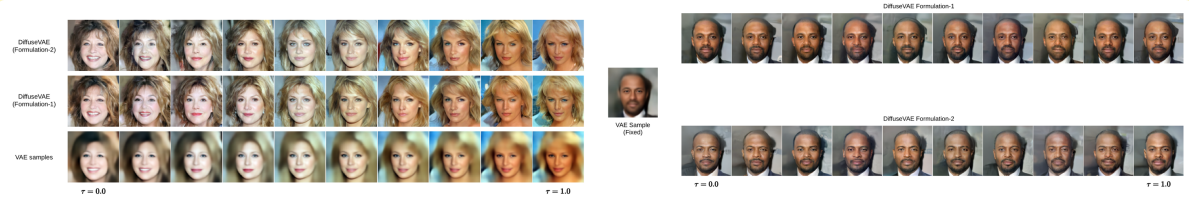
Interpolation.

1. Varying z_{VAE}, fixed x_T

z_{VAE}에서 2개의 점을 잡고, interpolation들을 봅니다.

2. Fixed z_{VAE}, Varying x_T

하나의 z_{VAE}로부터 $p(x_T)$를 2개 뽑아서 x_T^1, x_T 사이를 interpolation.



(a) DiffuseVAE samples generated from formulations-1 and 2 by interpolating in the VAE latent space.

(b) DiffuseVAE samples generated from formulations-1 and 2 by interpolating in the DDPM latent space with a fixed VAE reconstruction.

Figure 4: Interpolation in DiffuseVAE. τ denotes the linear interpolation factor. (Best viewed with zoom-in)

↳ z_{VAE}는 전체 content가 바뀌었고, z_T는 detail만 바뀜

Controllable Generation.

z_{VAE}를 interpolation을 통해 머리색, 성별 등을 제어.

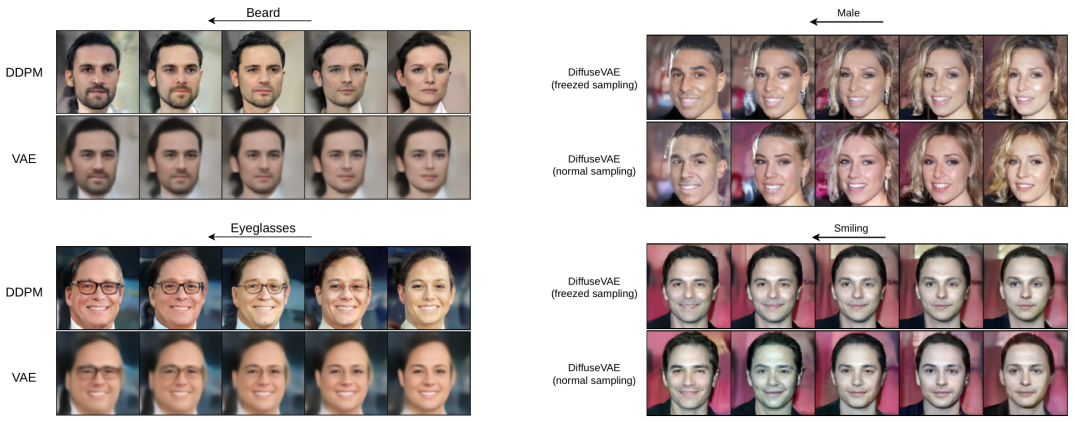


Figure 5: (a) Interpolations in the VAE latent space along meaningful directions for controllable synthesis of DiffuseVAE (Formulation-1) generated samples. (b) Fixing the DDPM stochasticity improves controllable synthesis interpolations using DiffuseVAE. (Best viewed with zoom-in)

↳ DDPM은 오직 refine만.

Handling the DDPM stochasticity

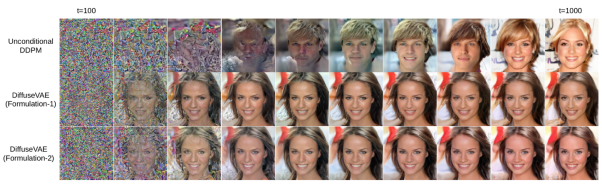
Diversity는 VAE의 latent code에서 발생된다.

하지만, DDPM덕분에 원하지 않아도 다른 특징으로 refine될 수 있다.

따라서 x_T sampling 중 제한

- Sampling speedups with DiffuseVAE

DDPM보다 적은 sampling이 필요함



(a) Comparison between samples generated from the unconditional DDPM and DiffuseVAE (formulations 1 and 2) vs the number of sampling steps.

(b) Effect of temperature (λ) on DiffuseVAE reverse process sampling for DiffuseVAE (formulation-1)

Figure 6: Qualitative analysis of the sampling speed-ups in DiffuseVAE. The time t denotes the number of reverse process steps performed during inference. The samples were obtained at equidistant time points between t=100 and t=1000. (Best viewed with zoom-in)

VAE에서 temperature scaling이 성능 저하 시킴.

# steps	Noise schedule	DiffuseVAE (Form-1)	DiffuseVAE (Form-2)
10	Linear(1e-6, 0.8)	34.97	**33.71**
10	Linear(1e-6, 0.7)	34.51	**33.41**
25	Linear(1e-6, 0.6)	30.88	**29.61**
50	Linear(1e-6, 0.3)	29.35	**28.45**

Table 3: FID score (5k samples) comparison between DiffuseVAE formulations 1 and 2 when using continuous noise conditioning for different inference schedules. No temperature scaling was used.

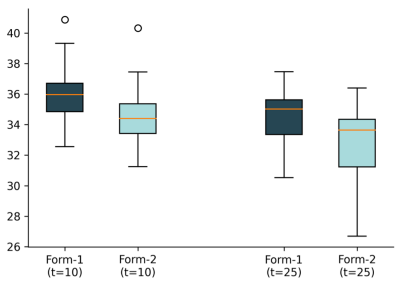


Figure 7: FID score comparison (5k samples) between DiffuseVAE formulations-1 and 2 conditioned on continuous noise for t=10 and 25 inference steps with shared DDPM latents.

- Generalization to downstream tasks.

refiner의 성능을 보기위해 styleGAN2의 데이터로 실험

기본적으로 VAE의 흐릿한 결과에 대한 refiner이기 때문에

이미 high-fidelity에는 좋지 않은 수 있다.

하지만 noise에서 샘플링이 있으므로 항상 새로운 데이터 있다.

Figure 9: DiffuseVAE (Formulation-1) generalization to auxiliary tasks like Image super-resolution (Left) and Image denoising (Right)

-SOTA 비교.

	Method	FID ↓	IS ↑
Ours	DiffuseVAE (t=1000)	8.72	8.63 ± 0.06
	DiffuseVAE (t=500)	9.72	8.53 ± 0.11
	DiffuseVAE (t=100) (Cont. Noise)	11.71	8.27 ± 0.01
	DDPM (t=1000) [21]	3.90	9.34 ± 0.07
	DDPM (t=500) [21]	14.31	9.04 ± 0.11
VAE-based methods	VAEBM [62] (w/ PC)	12.19	8.43
	DC-VAE [42]	17.90	8.2
	NVAE [56]	51.67	5.51
	NCP-VAE [1]	24.08	-
GAN-based methods	AutoGAN [6]	12.4	8.55 ± 0.1
	BigGAN [3]	14.73	9.22
	StyleGAN2 (w/o ADA) [24]	8.32	9.21 ± 0.09
	Progressive GAN [23]	-	8.8
	SNGAN [37]	21.7	8.22
	SNGAN + DDLS [7]	15.42	9.09
Score-based methods	NCSN [52]	25.32	8.87
	NCSNv2 (w/denoising) [53]	10.87	8.40 ± 0.07
	DDPM [21]	3.17	9.46 ± 0.11
	SDE (NCSN++) [54]	2.45	9.73
	SDE (DDPM++) [54]	2.78	9.64
	LSGM (FID) [57]	2.10	-

Table 4: Generative performance on unconditional CIFAR-10

	Method	FID ↓
Ours	DiffuseVAE	4.76
VAE-based methods	NCP-VAE [1]	5.25
	VAEBM [62]	5.31
	NVAE [56]	14.74
Score-based methods	NCSN [52]	25.30
	NCSNv2 [53]	10.23
GAN-based methods	QA-GAN [41]	6.42
	COCO-GAN [34]	4.0

Table 5: Generative performance on CelebA (64 x 64) (t=1000)