Neural Residual Diffusion Models for Deep Scalable Vision Generation

The Thirty-Eighth Annual Conference on Neural Information Processing Systems

Zhiyuan Ma

(mzyth@tsinghua.edu.cn)

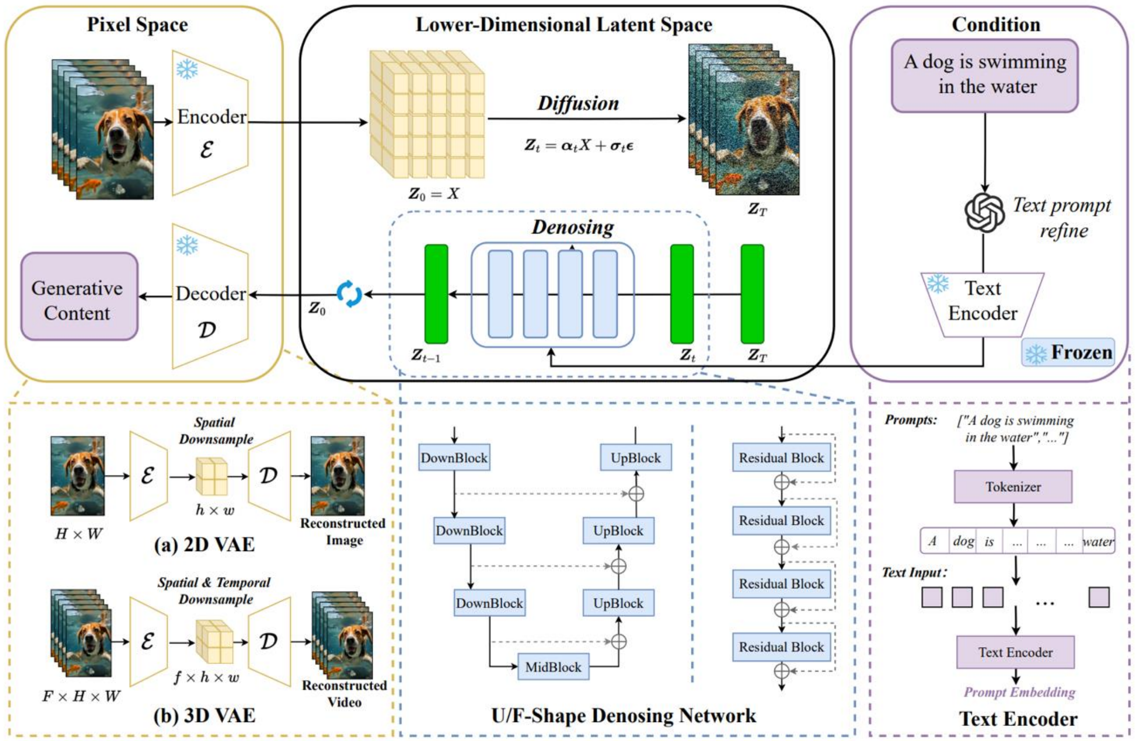
Tsinghua University, Beijing, China

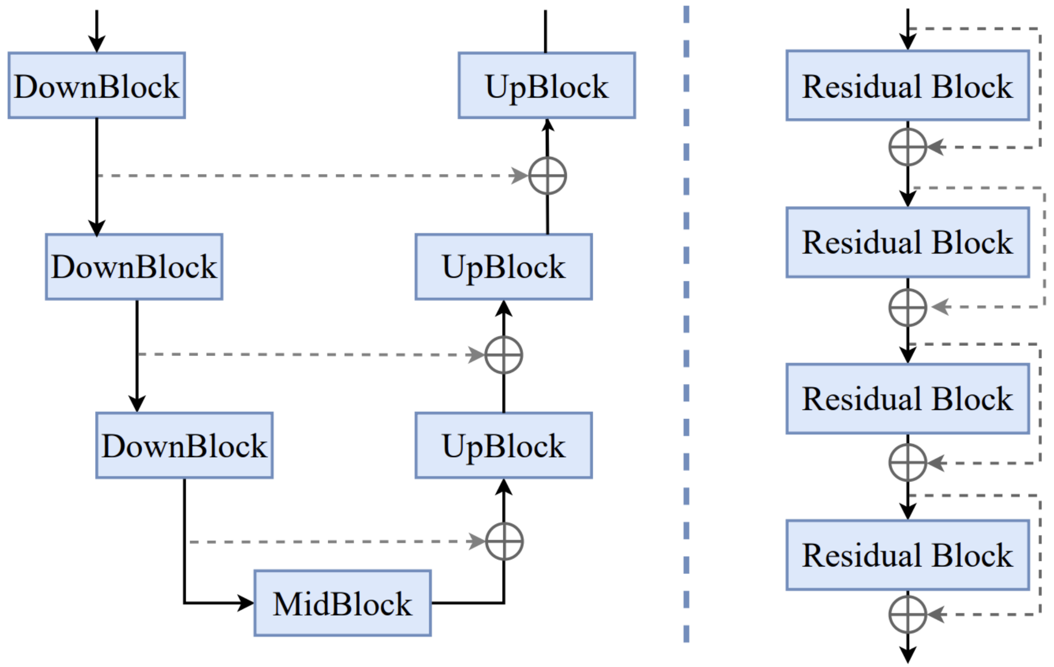
December, 2024

CONTENTS

CONTENTS

Deep Generative Diffusion Networks



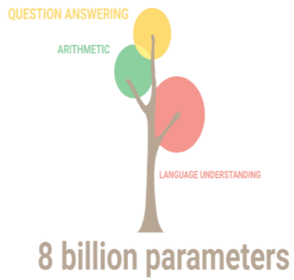


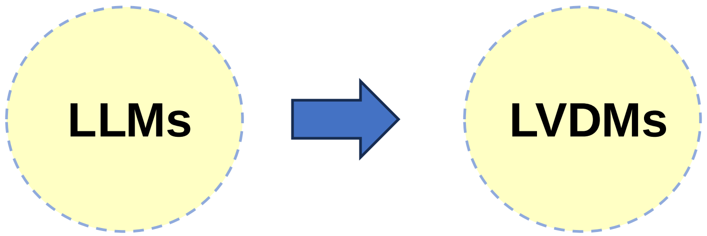
U/F-Shape Denosing Network

- ✓ The mainstream denoising backbones: U-Net、Transformer、U-ViT、DiT… (U-shaped / F-shaped)

☐ Representative Generative Diffusion Models

Methods	Year	Organization	Backbone	VAE	Text Encoder	# Params
ADM [4]	2021	OpenAI	Unet	None	-	554M
CDM [157]	2021	Google			-	-
DALL-E 2 [6]	2022	OpenAI			CLIP	6.5B
Imagen [5]	2022	Google			T5-XXL	3B
LDM [33]	2022	LMU Munich		2D VAE	CLIP ViT-L	400M+55M(VAE)
SD1.5 [33]	2022	LMU Munich			CLIP ViT-L	860M
SD2.0 [33]	2022	LMU Munich			OpenCLIP ViT-H	865M
SDXL [8]	2023	Stability AI			CLIP ViT-L & OpenCLIP ViT-bigG	2.6B
Playground-v2.5 [158]	2024	Playground			CLIP	-
UViT [72]	2022	Tsinghua University	Transformer	2D VAE	CLIP ViT-L	501M+84M(VAE)
DiT [73]	2022	UC Berkeley			CLIP ViT-L	675M+84M(VAE)
PixArt-α [81]	2023	Huawei Noah's Ark Lab			T5-XXL	600M
FiT [74]	2024	Shanghai AI Lab			CLIP ViT-L	-
SiT [75]	2024	New York University			CLIP ViT-L	675M
Latte [79]	2024	Shanghai AI Lab			T5-XXL	673.68M
Hunyuan-DiT [159]	2024	Tencent Hunyuan			mCLIP & mT5-XL	1.5B
LuminaT2X [160]	2024	Shanghai AI Lab			LLama2-7B	7B
Kolors [161]	2024	Kuaishou			ChatGLM3-6B-Base	2.6B
SD3.0 [80]	2024	Stability AI			CLIP ViT-L & OpenCLIP ViT-bigG & T5-XXL	8B
Flux.1 [162]	2024	BlackForestLabs			CLIP ViT-L & OpenCLIP ViT-bigG & T5-XXL	12B
Sora [163]	2024	OpenAI		3D VAE	-	-
Open-Sora [164]	2024	Hpcaitech			T5-XXL	1.2B
Open-Sora-Plan [165]	2024	Peking University			T5 & mT5	-
EasyAnimate [166]	2024	Alibaba Group			mCLIP & mT5-XL	1.5B
CogvideoX [82]	2024	Zhipu AI			T5-XXL	2B/5B
Moive Gen [83]	2024	Meta		TAE	MetaCLIP & UL2 & ByT5	30B

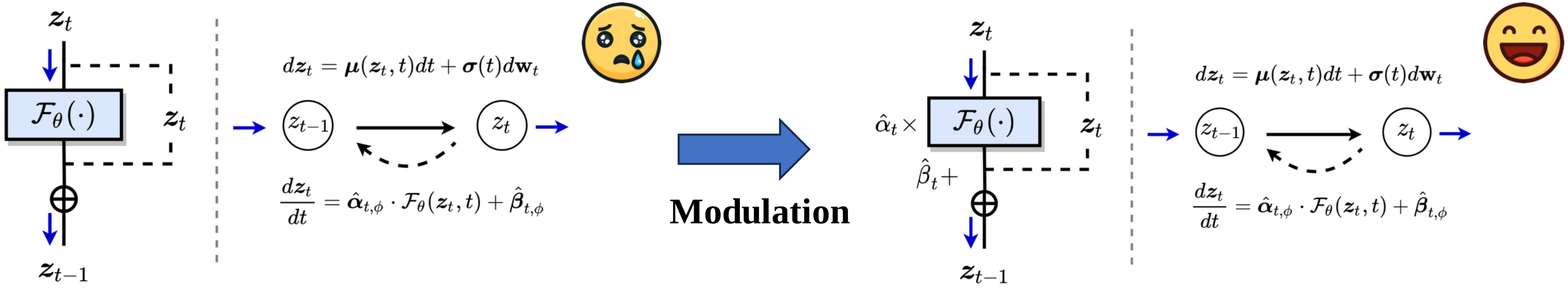




✓ The core of generative intelligence emergence: <u>Scaling Law with increasingly deep stacked networks</u>

CONTENTS

◻ The essential principle of how the generative denoising network works ?



➢ **Core:**

The optimization direction of neural network $\mathcal{F}_\theta(z_t, t)$ = The direction of inverted diffusion of the data z_t

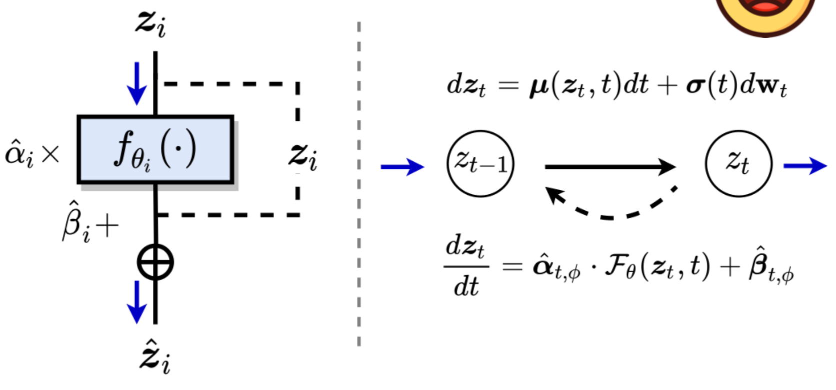
➢ **Issue:**

1. Asymmetry (coupling) of network predictions: $\mathcal{F}_\theta(z_t, t)$  $\hat{\alpha}_{t,\phi} \cdot \mathcal{F}_\theta(z_t, t) + \hat{\beta}_{t,\phi}$ **Unbalanced (One Lay**
2. Training architecture is difficult to scale: $\mathcal{F}_\theta(z_t, t)$  $\{f_{\theta_1}, \cdots, f_{\theta_i}, \cdots, f_{\theta_L}\}$ **Unstable (Deep Layer**

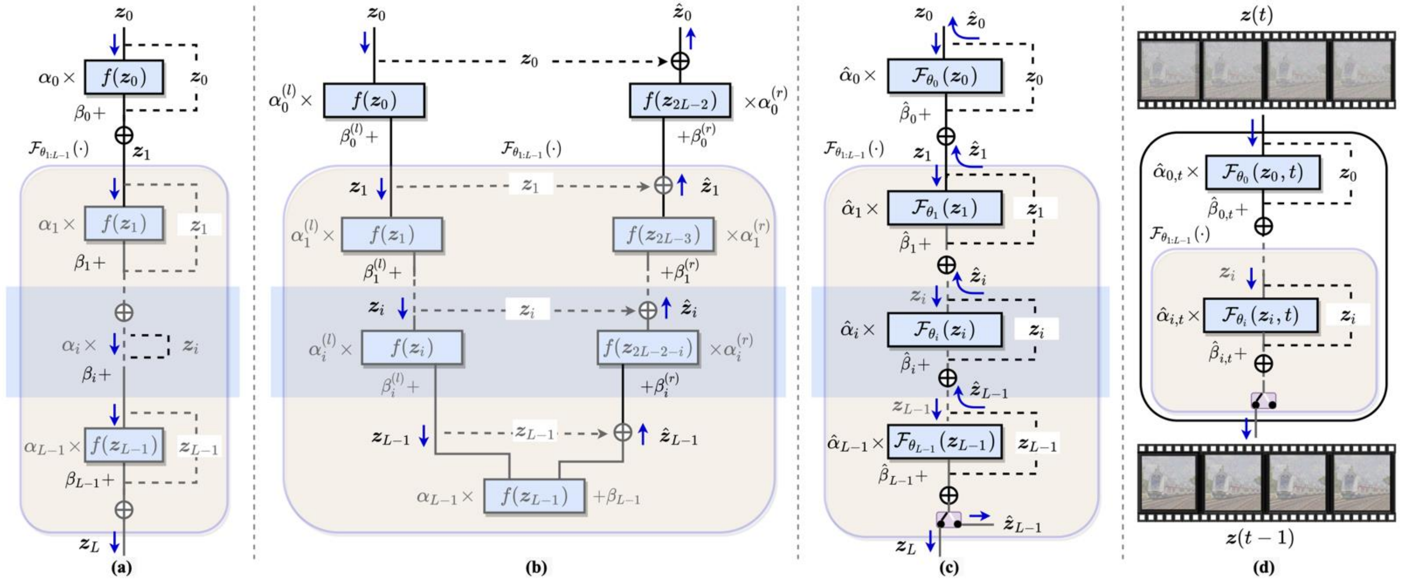
CONTENTS

Methods

Neural Residual Diffusion Models



(a) MRS-Unit $\mathtt{block}_i$ (b) Neural Residual Denoising Models $\mathcal{F}_\theta$



➢ **Gating-Residual Mechanism**

$$\begin{cases} \hat{z}_i = z_{i+1} = z_i + [\alpha_i \cdot f_{\theta_i}(z_i) + \beta_i]. \\ \hat{z}_i = \underbrace{\alpha_i^{(l)} \cdot f_{\theta_i^{(l)}}(z_i) + \beta_i^{(l)}}_{\text{read-in branch}} \hookrightarrow \underbrace{z_i + \alpha_i^{(r)} \cdot f_{\theta_i^{(r)}}(z_{2L-2-i}) + \beta_i^{(r)}}_{\text{read-out branch}} = z_i + \hat{\alpha}_i \cdot \mathcal{F}_{\theta_i}(z_i) + \hat{\beta}_i. \end{cases}$$

(Discrete Form) → $\frac{z_{i+\delta} - z_i}{\delta} = \hat{z}_i - z_i = \hat{\alpha}_i \cdot \mathcal{F}_{\theta_i}(z_i) + \hat{\beta}_i \Longrightarrow \frac{dz_t}{dt} = \hat{\alpha}_\phi \cdot \mathcal{F}_{\theta_t}(z_t) + \hat{\beta}_\phi$

(U-shaped)

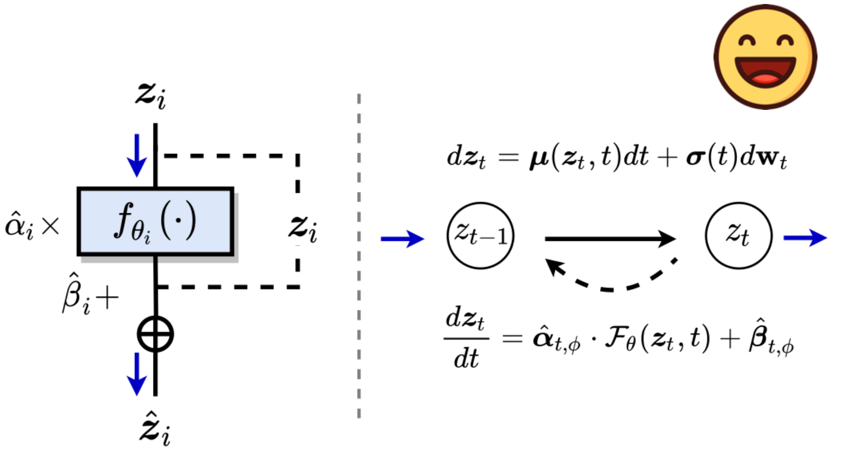
(Continuous Form)

Gating Residual Modulation

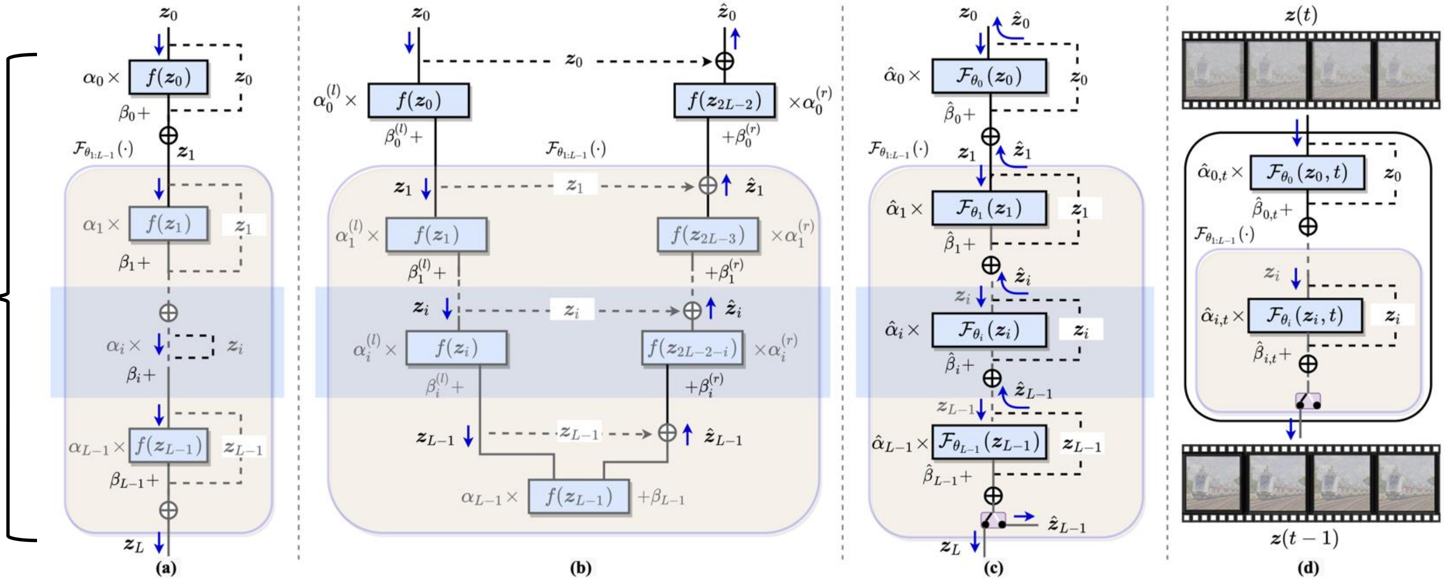
Methods

3.2

☐ Neural Residual Diffusion Models



(a) MRS-Unit $\mathtt{block}_i$ (b) Neural Residual Denoising Models $\mathcal{F}_\theta$

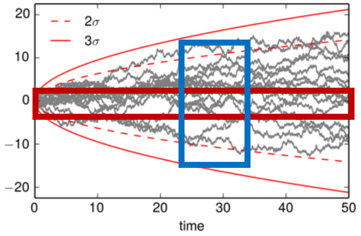


➢ **Denoising Dynamics Parameterization**

$$dz_t = \mu(z_t, t)dt + \sigma(t)dw_t \Longrightarrow \frac{dz_t}{dt} = \mu(z_t, t) + \sigma(t) \cdot \epsilon_t;$$ **(Noise-Adding SDE)**

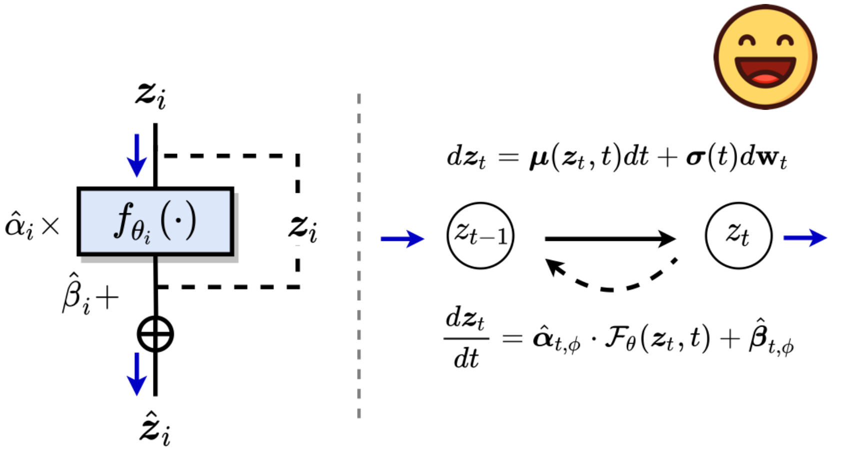
$$\frac{dz_t}{dt} = \boxed{\mu(z_t, t)} \boxed{-\frac{1}{2}\sigma(t)^2 \cdot} \left[\nabla_z \log p_t(z_t)\right] = \boxed{\hat{\alpha}_{t,\phi}} \cdot \mathcal{F}_\theta(z_t, t) + \boxed{\hat{\beta}_{t,\phi}},$$ **(Denoising-ODE)**

$-\frac{1}{2}\sigma(t)^2 \rightarrow \hat{\alpha}_{t,\phi}$ **Variance Parameterization**

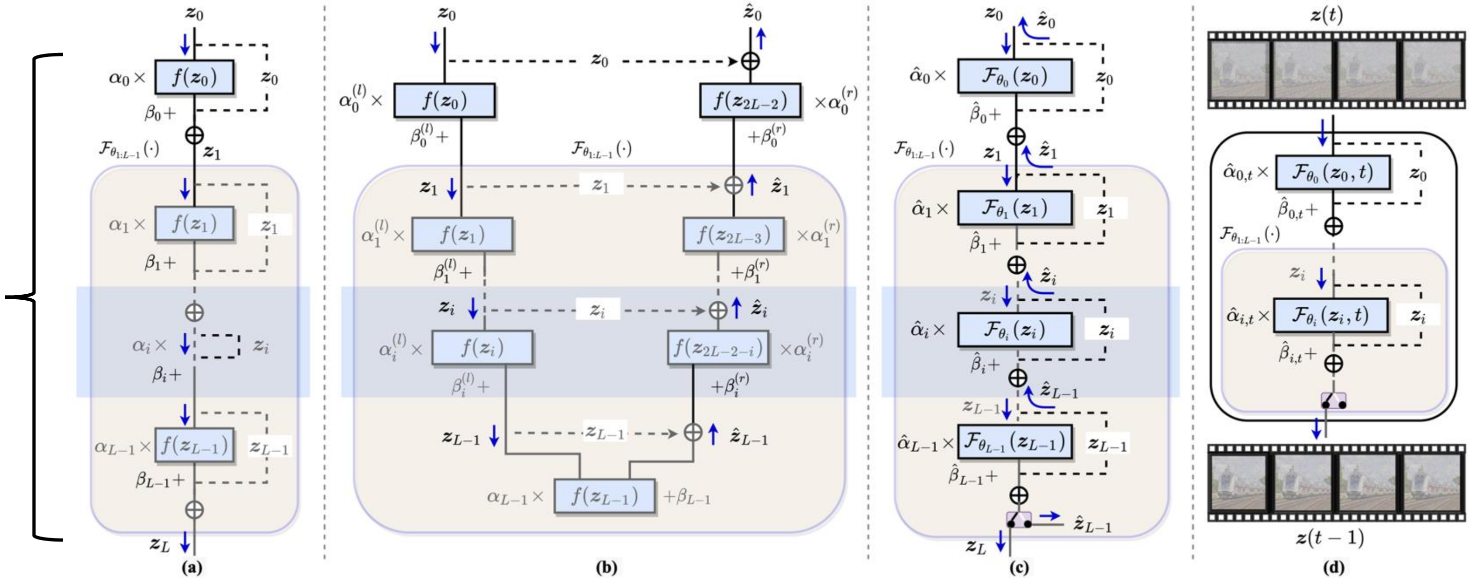


$\mu(z_t, t) \rightarrow \hat{\beta}_{t,\phi}$ **Mean Parameterization**

Neural Residual Diffusion Models



(a) MRS-Unit $\texttt{block}_i$ (b) Neural Residual Denoising Models $\mathcal{F}_\theta$



- **Residual Sensitivity Control**
  - To control the **numerical errors** in back-propagation and achieve **steadily and massively scalable training**
  - We introduce ***Residual-Sensitivity ODE***:

Define ***Residual-Sensitivity*** : $\boldsymbol{s}_t = \frac{d\mathcal{L}}{d\boldsymbol{z}_t} = \frac{d\mathcal{L}}{d\boldsymbol{z}_{t+\delta}} \cdot \frac{d\boldsymbol{z}_{t+\delta}}{d\boldsymbol{z}_t} = \boldsymbol{s}_{t+\delta} \cdot \frac{d\boldsymbol{z}_{t+\delta}}{d\boldsymbol{z}_t}.$

Residual-Sensitivity ODE

$$\boxed{\frac{d\boldsymbol{s}_t}{dt} = \lim_{\delta \to 0+} \frac{\boldsymbol{s}_{t+\delta} - \boldsymbol{s}_t}{\delta} = \lim_{\delta \to 0+} \frac{-\boldsymbol{s}_{t+\delta} \cdot \frac{\partial}{\partial \boldsymbol{z}_t}\left(\int_t^{t+\delta} f_\theta(\boldsymbol{z}_t)dt\right)}{\delta} = -\boldsymbol{s}_t \cdot \frac{\partial f_\theta(\boldsymbol{z}_t, t)}{\partial \boldsymbol{z}_t}.}$$

Methods

3.3.1

- **Residual Sensitivity Control**

  - To control the **numerical errors** in back-propagation and achieve **steadily and massively scalable training**
  - First, we define ***Residual-Sensitivity***: $\boldsymbol{s}_t = \frac{d\mathcal{L}}{d\boldsymbol{z}_t}$

$$\boldsymbol{s}_t = \frac{d\mathcal{L}}{d\boldsymbol{z}_t} = \frac{d\mathcal{L}}{d\boldsymbol{z}_{t+\delta}} \cdot \frac{d\boldsymbol{z}_{t+\delta}}{d\boldsymbol{z}_t} = \boldsymbol{s}_{t+\delta} \cdot \frac{d\boldsymbol{z}_{t+\delta}}{d\boldsymbol{z}_t}. \longrightarrow \because d\boldsymbol{z}_{t+\delta} = d\boldsymbol{z}_t + \int_t^{t+\delta} f_\theta(\boldsymbol{z}_t, t)dt. \quad \therefore \boldsymbol{s}_t = \boldsymbol{s}_{t+\delta} + \boldsymbol{s}_{t+\delta} \cdot \frac{\partial}{\partial \boldsymbol{z}_t}\left(\int_t^{t+\delta} f_\theta(\boldsymbol{z}_t, t)dt\right).$$

(Chain Rule) **(Euler Solver)**

Residual-Sensitivity ODE

$$\because \quad \boldsymbol{s}_t = \boldsymbol{s}_{t+\delta} + \boldsymbol{s}_{t+\delta} \cdot \frac{\partial}{\partial \boldsymbol{z}_t}\left(\int_t^{t+\delta} f_\theta(\boldsymbol{z}_t, t)dt\right). \quad \therefore \quad \boxed{\frac{d\boldsymbol{s}_t}{dt} = \lim_{\delta \to 0^+} \frac{\boldsymbol{s}_{t+\delta} - \boldsymbol{s}_t}{\delta} = \lim_{\delta \to 0^+} \frac{-\boldsymbol{s}_{t+\delta} \cdot \frac{\partial}{\partial \boldsymbol{z}_t}\left(\int_t^{t+\delta} f_\theta(\boldsymbol{z}_t)dt\right)}{\delta} = -\boldsymbol{s}_t \cdot \frac{\partial f_\theta(\boldsymbol{z}_t, t)}{\partial \boldsymbol{z}_t}.}$$

  - Then, we further use the **Euler solver** to obtain the sensitivity $\boldsymbol{s}_{t_0}$:

$$\boldsymbol{s}_{t_0} = \boldsymbol{s}_{t_L} + \int_{t_L}^{t_0} \frac{d\boldsymbol{s}_t}{dt} dt = \boldsymbol{s}_{t_L} - \int_{t_L}^{t_0} \boldsymbol{s}_t \cdot \frac{\partial f_\theta(\boldsymbol{z}_t, t)}{\partial \boldsymbol{z}_t} dt. \longrightarrow \boldsymbol{s}_{t_L} > \boldsymbol{s}_{t_{L-1}} > \cdots > \boldsymbol{s}_{t_0}$$

(non-negativity) **gradually decaying sensitivity !**

  - Similarly, we can define ***parameter-sensitivity***: $\boldsymbol{s}_\theta = \frac{d\mathcal{L}}{d\boldsymbol{\theta}}$, we can derive:

$$\boldsymbol{s}_{\theta_0} = \boldsymbol{s}_{\theta_L} + \int_{t_L}^{t_0} \frac{d\boldsymbol{s}_\theta}{dt} dt = \boldsymbol{s}_{\theta_L} - \int_{t_L}^{t_0} \boldsymbol{s}_\theta \cdot \frac{\partial f_\theta(\boldsymbol{z}_t, t)}{\partial \theta} dt. \longrightarrow \boldsymbol{s}_{\theta_L} > \boldsymbol{s}_{\theta_{L-1}} > \cdots > \boldsymbol{s}_{\theta_0}$$

Methods

NEURAL INFORMATION PROCESSING SYSTEMS

3.3.2

- **Residual Sensitivity Control**

  - So far, we have explored the current situation of the problem:

Residual-Sensitivity ODE

$$\boxed{\frac{d\boldsymbol{s}_t}{dt} = \lim_{\delta\to 0^+} \frac{\boldsymbol{s}_{t+\delta} - \boldsymbol{s}_t}{\delta} = \lim_{\delta\to 0^+} \frac{-\boldsymbol{s}_{t+\delta}\cdot \frac{\partial}{\partial \boldsymbol{z}_t}\left(\int_t^{t+\delta} f_\theta(\boldsymbol{z}_t)dt\right)}{\delta} = -\boldsymbol{s}_t \cdot \frac{\partial f_\theta(\boldsymbol{z}_t,t)}{\partial \boldsymbol{z}_t}.}$$

$$\rightarrow \begin{cases} \boldsymbol{s}_{t_0} = \boldsymbol{s}_{t_L} + \int_{t_L}^{t_0} \frac{d\boldsymbol{s}_t}{dt} dt = \boldsymbol{s}_{t_L} - \int_{t_L}^{t_0} \boldsymbol{s}_t \cdot \frac{\partial f_\theta(\boldsymbol{z}_t,t)}{\partial \boldsymbol{z}_t} dt. \\ \boldsymbol{s}_{\theta_0} = \boldsymbol{s}_{\theta_L} + \int_{t_L}^{t_0} \frac{d\boldsymbol{s}_\theta}{dt} dt = \boldsymbol{s}_{\theta_L} - \int_{t_L}^{t_0} \boldsymbol{s}_\theta \cdot \frac{\partial f_\theta(\boldsymbol{z}_t,t)}{\partial \theta} dt. \end{cases}$$

  - Subsequently, we apply **Gating-Residual** and **Mean-Variance Parameterization** to ***Residual-Sensitivity ODE*** :

Rectified Residual-Sensitivity ODE

$$\boxed{\frac{d\hat{\boldsymbol{s}}_t}{dt} = \lim_{\delta\to 0^+} \frac{\hat{\boldsymbol{s}}_{t+\delta} - \hat{\boldsymbol{s}}_t}{\delta} = -(\alpha_{t,\phi}\cdot \hat{\boldsymbol{s}}_t)\cdot \frac{\partial f_\theta(\hat{\boldsymbol{z}}_t,t)}{\partial \hat{\boldsymbol{z}}_t} - (\beta_{t,\phi}\cdot \hat{\boldsymbol{s}}_t).}$$

$$\begin{aligned} \hat{\boldsymbol{s}}_{t_0} &= \hat{\boldsymbol{s}}_{t_L} + \int^{t_0} \frac{d\hat{\boldsymbol{s}}_t}{dt} dt \\ &= \hat{\boldsymbol{s}}_{t_L} - \int_{t_L}^{t_0} \left[(\alpha_{t,\phi}\cdot \hat{\boldsymbol{s}}_t)\cdot \frac{\partial f_\theta(\hat{\boldsymbol{z}}_t,t)}{\partial \hat{\boldsymbol{z}}_t} + (\beta_{t,\phi}\cdot \hat{\boldsymbol{s}}_t)\right] dt. \end{aligned}$$

  - Eventually, we can supervise it to achieve ***Residual Sensitivity Control*** via:

$$\mathcal{L}_s = ||\mathcal{F}_\theta(\boldsymbol{z}_t,t) - \nabla_z \log p_t(\boldsymbol{z}_t)||_2^2 + \gamma \cdot \sum_L ||\alpha_{t,\phi}\cdot \frac{\partial f_\theta(\hat{\boldsymbol{z}}_t,t)}{\partial \hat{\boldsymbol{z}}_t} - \beta_{t,\phi}||_2^2$$ **(Rectified Term)**

CONTENTS

Experiment

4.1

◻ Experiments on Image Synthesis with Deep Scalable Spatial Learning

Architecture	Method	Scalability	Class-to-Image			Text-to-Image		
			FID↓	sFID↓	IS↑	FID↓	sFID↓	IS↑
GAN	BigGAN-deep [56]	✗	6.95	7.36	171.4	-	-	-
	StyleGAN-XL [57]	✗	2.30	4.02	265.12	-	-	-
U-shaped	ADM [58]	✔	10.94	6.02	100.98	-	-	-
	ADM-U	✔	7.49	5.13	127.49	-	-	-
	ADM-G	✔	4.59	5.25	186.70	-	-	-
	LDM-8 [30]	✔	15.51	-	79.03	16.64	11.32	64.50
	LDM-8-G	✔	7.76	-	209.52	9.35	10.02	125.73
	LDM-4	✔	10.56	-	103.49	12.37	11.58	94.65
	LDM-4-G	✔	3.60	-	247.67	3.78	5.89	182.53
F-shaped	DiT-XL/2 [59]	✔	9.62	6.85	121.50	8.53	5.47	144.26
	DiT-XL/2-G	✔	2.27	4.60	278.24	3.53	5.48	175.63
	Latte-XL [60]	✔	2.35	5.17	224.75	2.74	5.35	195.03
Unified	**Neural-RDM-U (Ours)**	✔✔	3.47	5.08	256.55	**2.25**	**4.36**	**235.35**
	Neural-RDM-F (Ours)	✔✔	**2.12**	**3.75**	**295.32**	2.46	5.65	206.32

Table 1: The main results for image generation on ImageNet [61] (Class-to-Image) and JourneyDB [53] (Text-to-Image) with 256 × 256 image resolution. We highlight the best value in blue, and the second-best value in green. The *Scalability* column indicates the scaling capability of the parameter scale and architecture.

- Neural-RDMs have obtained competitive and state-of-the-art results across image synthesis benchmarking.
- Benefiting from the rectification of generative dynamics, it highlights the semantics of the subject more.

"*Christina Hendricks*, dressed as a *Disney princess*, stands in full torso on a city street, ... "

"A mystical being made of *dust surrounded* by *vines*, *moss*, *ferns*, *foliage*, and *wild plants* ..."

"A handsome *fairy prince* with blue hair delicately holds out a single *white apple*, his perfect facial ..."

"A beautiful *garden* in summer adorned with *white lilies*, featuring a natural wreath and bushes ..."

"A grand *cathedral corridor* in a hypermaximalist style, adorned with intricate white marble and *blue ornate* ..."

"A *landscape* featuring a *wooden bridge* over a serene *lake* with a majestic *mountain* in the background."

"A cartoon-style *watercolor* cover illustration featuring *vibrant pink* and *cream hydrangeas* in a Disney-inspired setting, complemented by typography."

"A speckled *headscarf*, a swamp adder, and a *flag of India* depicted in an isometric illustration."

"A *castle* situated in the *mountains* with an array of very high thin *towers* adorned with numerous arrowslits."

"The *King* of Pentacles stands in a regal pose, surrounded by earthly riches, while a *mysterious UFO* hovers above, adding an element of otherworldly intrigue."

"A hyper-realistic portrait of a 17-year-old *English girl* with mismatched eyes, *blonde curly hair* adorned with flowers, holding a flute, radiating pure joy."

"*woman captain* wearing a *cocked hat* stands on a ship, gazing in awe and fear as a huge Cthulhu emerges from the water."

"A *cyberpunk man* with silver skin wearing a helmet featuring a *large glass visor,* holding a rocket launcher in a futuristic setting, depicted with hyper-realistic..."

Experiment

4.1

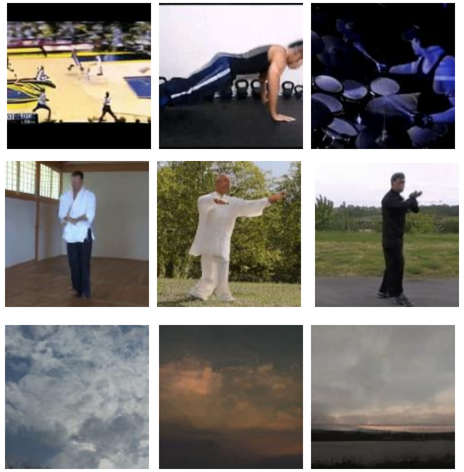
NEURAL INFORMATION PROCESSING SYSTEMS

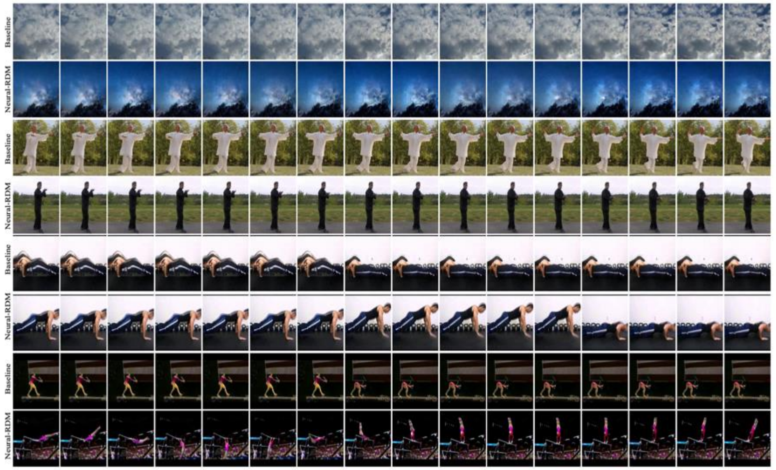
☐ Experiments on Video Generation with Deep Scalable Temporal Learning

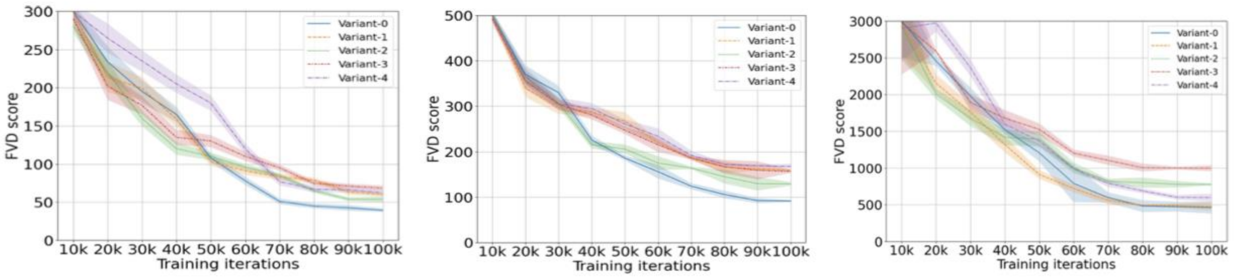
Method	Scalability	Frame Evaluation		None-to-Video		Class-to-Video
		FID↓	IS↑	SkyTimelapse (FVD↓)	Taichi-HD (FVD↓)	UCF-101 (FVD↓)
MoCoGAN [71]	✗	23.97	10.09	206.6	-	2886.9
MoCoGAN-HD [72]	✗	7.12	23.39	164.1	128.1	1729.6
DIGAN [73]	✗	19.10	23.16	83.11	156.7	1630.2
StyleGAN-V [70]	✗	9.45	23.94	79.52	-	1431.0
MoStGAN-V [74]	✗	-	-	65.30	-	1380.3
PVDM [75]	✓	29.76	60.55	75.48	540.2	1141.9
LVDM [12]	✓	-	-	95.20	99.0	**372.0**
VideoGPT [76]	✓	22.70	12.61	222.7	-	2880.6
Latte-XL [60]	✓	5.02	68.53	59.82	159.60	477.97
Neural-RDM (Ours)	✓✓	3.35	85.97	**39.89**	**91.22**	**461.03**

Table 2: The main results for video generation on the SkyTimelapse [62], Taichi-HD [63] and UCF-101 [64] with 256 × 256 resolution of each frame. We highlight the best value in blue, and the second-best value in green.

- Neural-RDMs (flow-shaped version) basically achieves the best results, except for the second-best results in class-to-video evaluation.
- Compare with the baseline, Neural-RDM maintains temporal coherence and consistency, resulting in smoother and more dynamic video frames.



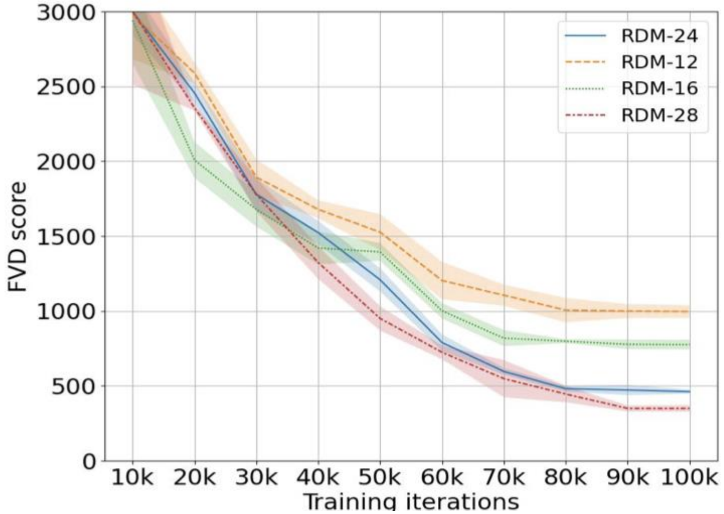




(a) (b) (c)

Figure 6: **(a)**, **(b)**, and **(c)** respectively illustrate the performance of the five residual structures variant models across the SkyTimelapsee [62], Taichi-HD[63], and UCF-101 [64].

◻ Comparison Experiments of Gating Residual Variants and Deep Scalability



- As the number of training steps increases, almost all variants can converge effectively, but only Variant-0 (Our approach) achieves the best FVD scores.
- As the depth of residual units increases, the performance of the model can be further improved, which further highlights the deep scalability advantage of Neural-RDM.

① *Variant-0 (Ours)*: $z_{i+1} = z_i + \alpha f(z_i) + \beta$;

② *Variant-1 (AdaLN [77])*: $z_{i+1} = z_i + f(\alpha z_i + \beta)$;

③ *Variant-2*: $z_{i+1} = \alpha z_i + f(z_i) + \beta$;

④ *Variant-3 (ResNet [78])*: $z_{i+1} = z_i + f(z_i)$;

⑤ *Variant-4 (ReZero [79])*: $z_{i+1} = z_i + \alpha f(z_i)$.

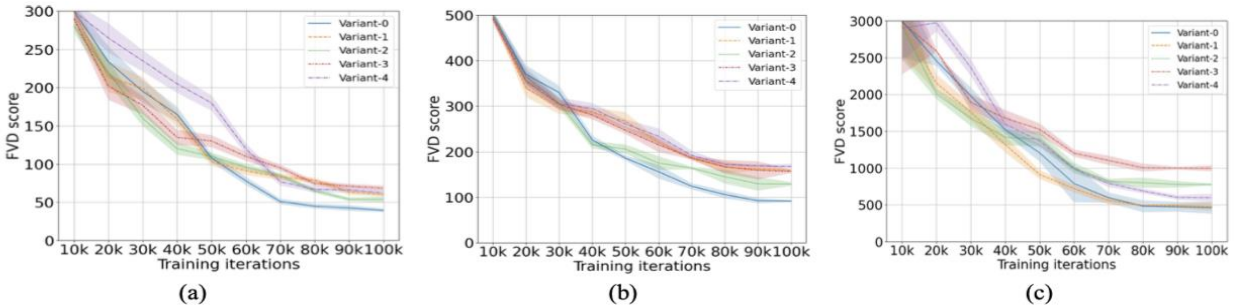


Figure 6: **(a)**, **(b)**, and **(c)** respectively illustrate the performance of the five residual structures variant models across the SkyTimelapsee [62], Taichi-HD[63], and UCF-101 [64].

CONTENTS

Conclusion

05

NEURAL INFORMATION PROCESSING SYSTEMS

- ✓ **Propose a unified neural residual diffusion models framework**

We practically unify u-shaped and flow-shaped stacking networks and to propose a unified and deep scalable neural residual diffusion model framework.

- ✓ **Parameterize the mean-variance scheduler for excellent dynamics consistency**

Moreover, we theoretically parameterize the previous human-designed mean-variance scheduler and demonstrate excellent dynamics consistency.

- ✓ **Adequate and extensive experiments and analyses**

Experimental results on various generative tasks show that Neural-RDM obtains the best results, and extensive experiments also demonstrate the advantages in improving the fidelity, consistency of generated content and supporting large-scale scalable training.

NeurIPS 2024

Thanks for your listening

Zhiyuan Ma

(https://ponymzy.github.io)

Department of Electronic Engineering,

Tsinghua University, Beijing, China

mzyth@tsinghua.edu.cn