Latent-to-data Cascaded Diffusion Models for Unconditional Time Series Generation

Lifeng Shen[1,2], Kai Syun Hou[2], Weiyu Chen[2], James T. Kwok[2]

[1] Chongqing University of Posts and Telecommunications (CQUPT); [2] Hong Kong University of Science and Technology (HKUST)

ICLR 2026

Background

- Synthetic time series generation (TSG) is crucial for applications privacy preservation, data augmentation, and anomaly detection.
- A challenge in TSG lies in modeling the multi-modal distributions of time series, which requires simultaneously capturing diverse high-level representation distributions and preserving local temporal fidelity.
- Existing diffusion models, however, are constrained by their single-space focus: latent-space models capture latent distributions but often compromise local fidelity, while data-space models preserve local details in the data space but struggle to learn high-level representations essential for multi-modal time series.

	data	latent	cascaded
TimeGrad	✓	✗	✗
CSDI	✓	✗	✗
TimeDiff	✓	✗	✗
TSDE	✓	✗	✗
TimeLDM	✗	✓	✗
LDT	✗	✓	✗
DiffusionTS	✓	✗	✗
L2D-Diff (proposed)	✓	✓	✓

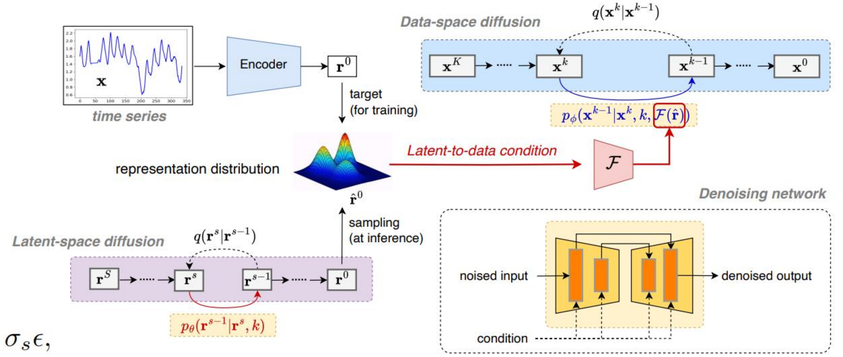
Contributions: we propose L2D-Diff, a latent-to-data diffusion framework that integrates the strengths of latent-space modeling and data-space refinement to overcome the limitations of unconditional generation.

L2D-Diff: Latent-to-Data Cascaded Diffusion

- L2D-Diff is simple yet effective. As a cascaded diffusion model, it bridges the latent and data diffusion processes, transforming an unconditional time series generation problem into a conditional one.
- It proceeds in two complementary stages:

  i) Latent-space coarse generation: A latent diffusion model captures representation distributions by representation learning.

  ii) Data-space refinement: A subsequent denoising process integrates the global latent codes into the data space, enabling fine-grained temporal precision and ensuring consistency with the original data distribution.



$$\hat{\mathbf{r}}^{s-1} = \frac{\sqrt{\alpha_s}(1-\bar{\alpha}_{s-1})}{1-\bar{\alpha}_s}\mathbf{r}^s + \frac{\sqrt{\bar{\alpha}_{s-1}}(1-\alpha_s)}{1-\bar{\alpha}_s}\mathbf{r}_\phi(\mathbf{r}^s, s) + \sigma_s\epsilon,$$

$$\hat{\mathbf{x}}^{k-1} = \frac{\sqrt{\alpha_k}(1-\bar{\alpha}_{k-1})}{1-\bar{\alpha}_k}\mathbf{x}^k + \frac{\sqrt{\bar{\alpha}_{k-1}}(1-\alpha_k)}{1-\bar{\alpha}_k}\mathbf{x}_\theta(\mathbf{x}^k, k, \mathcal{F}(\mathbf{c})) + \sigma_k\epsilon.$$

Here, $\mathbf{c} = \hat{\mathbf{r}}^0$ is generated representations to guide the data denoising process.

denoising losses

$$\mathcal{L}_\mathbf{r} = \mathbb{E}_{\mathbf{r}^0,\epsilon,k}\|\mathbf{r}^0 - \mathbf{r}_\phi(\mathbf{r}^k, k)\|^2$$

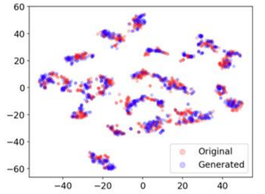
$$\mathcal{L}_\mathbf{x} = \mathbb{E}_{\mathbf{x}^0,\epsilon,k}\|\mathbf{x}^0 - \mathbf{x}_\theta(\mathbf{x}^k, k)\|^2$$

Experiments

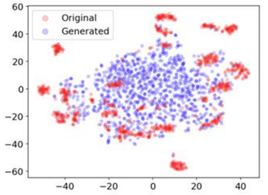
Contextual-FID	Stock	Energy	ETTh	Riverflow	Two Patterns	ECG5000	Medical Images	Arabic Digits	Atrial Fibrillation	Japanese Vowels	Character Trajectories
L2D-Diff	0.31	0.53	**0.45**	0.32	**0.21**	**0.11**	**0.08**	1.29	1.15	**0.46**	**0.28**
Diffusion-TS	0.49	0.82	4.75	1.24	1.69	1.95	3.10	1.66	2.39	1.93	3.57
TSDE	3.90	4.13	-	**0.26**	2.83	1723.0	0.85	2.60	-	3.97	4.70
TimeLDM	6.17	3.51	9.52	1.01	1.40	0.76	0.88	5.99	5.48	0.99	2.00
EDDPM	2.31	2.89	10.76	28.29	6.72	1.11	1.01	5.40	4.63	1.40	3.56
FourierDiffusion	**0.21**	0.48	3.38	3.54	1.16	0.32	0.41	**1.26**	**1.14**	0.49	3.58
ImagenTime	4.23	2.22	7.72	0.50	4.82	6.38	2.99	2.98	1.66	1.08	12.02
FourierFlow	1.15	**0.38**	3.17	1.843	1.21	0.98	1.52	2.84	2.37	0.74	5.07
TimeFlow	0.41	0.85	3.19	2.177	1.17	0.20	0.65	8.40	173.12	-	3.45
TimeGAN	0.88	0.87	20.32	2.00	2.26	3.88	0.70	4.73	6.63	1.30	3.97
GTGAN	0.70	2.55	26.60	3.23	25.53	3.39	2.82	16.23	3.23	2.24	10.01
KoVAE	0.48	1.17	6.78	1.72	8.82	1.17	0.80	2.46	2.89	3.85	6.54
TimeVQVAE	2.45	6.05	8.40	0.74	5.06	4.20	2.93	8.17	3.77	4.62	3.98
LS4	5.85	10.97	23.47	3.47	15.81	24.21	31.81	14.45	8.15	11.34	24.67
VAE	4.41	7.16	35.65	1.67	28.16	3.42	2.62	15.70	7.66	6.88	10.02

	type	training (ms/sample)	inference (ms/sample)	# of trainable parameters
L2D-Diff	data + latent	0.52	3.47	2.2M
Diffusion-TS	data	14.28	5.10	25M
TSDE	data	2.10	5.05	1.3M
TimeLDM	latent	0.51	4.85	1.9M
EDDPM	latent	0.51	3.80	1.9M
FourierDiffusion	frequency	0.36	9.66	1.6M
ImagenTime	frequency	1.22	2.82	1.1M

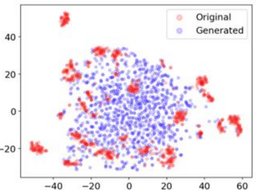
- The proposed L2D-Diff achieves superior overall performance with an average rank of 1.45, significantly outperforming all the baselines.
- This not only reduces complexity but also ensures efficient generation of time series, particularly when dealing with complex or multimodal distributions.



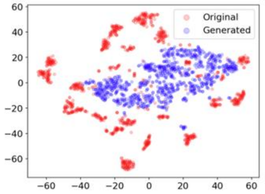
(a) L2D-Diff (*proposed*).



(b) FourierDiffusion.



(c) FourierFlow.



(d) Diffusion-TS.

Visualization of the t-SNE embeddings of data (not representations) shows that the synthetic time series generated by L2D-Diff have distributions more closely aligned with the real multi-modal data, effectively capturing the intrinsic class-specific patterns of the dataset.