BiDM: Pushing the Limit of Quantization for Diffusion Models

Xingyu Zheng[1] Xianglong Liu[1†] Yichen Bian[1] Xudong Ma[1] Yulun Zhang[2]
Jiakai Wang[3] Jinyang Guo[1] Haotong Qin[4]

[1]Beihang University [2]Shanghai Jiao Tong University
[3]Zhongguancun Laboratory [4]ETH Zürich

Paper: https://neurips.cc/virtual/2024/poster/93620
Code: https://github.com/Xingyu-Zheng/BiDM
(star is welcome)

1 Introduction: BERT Binarization

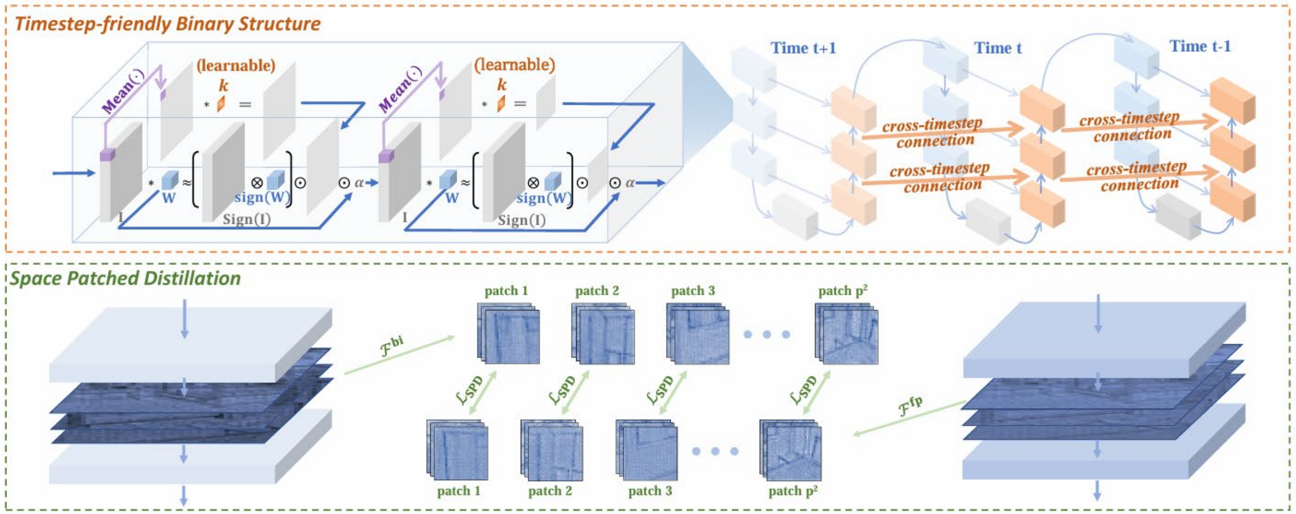
- **Large Pre-trained Diffusion models**
  - Diffusion models (DMs) have garnered impressive attention and applications in various fields, such as image, speech and video
  - it still suffers expensive FP32 parameters and operations
- **Network Binarization**
  - compression by binarizing parameters
  - accelerating by applying bitwise operations

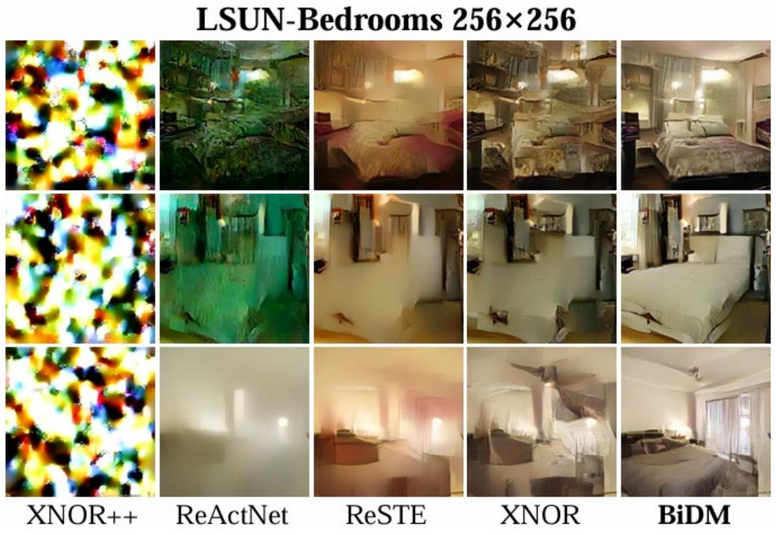
$$Q_{\mathrm{x}}(\mathbf{x}) = \alpha\, \mathbf{B}_{\mathbf{x}}$$

$$\mathbf{B}_{\mathbf{x}} = \operatorname{sign}(\mathbf{x}) = \begin{cases} -1, & \text{if } x \geq 0 \\ 1, & \text{otherwise} \end{cases}$$

$$z = Q_{\mathrm{w}}(\mathbf{w})^{\top} Q_{\mathrm{a}}(\mathbf{a}) = \alpha_{\mathrm{w}}\alpha_{\mathrm{a}}(\mathbf{B}_{\mathbf{w}} \otimes \mathbf{B}_{\mathbf{a}})$$

1 Introduction: Overview





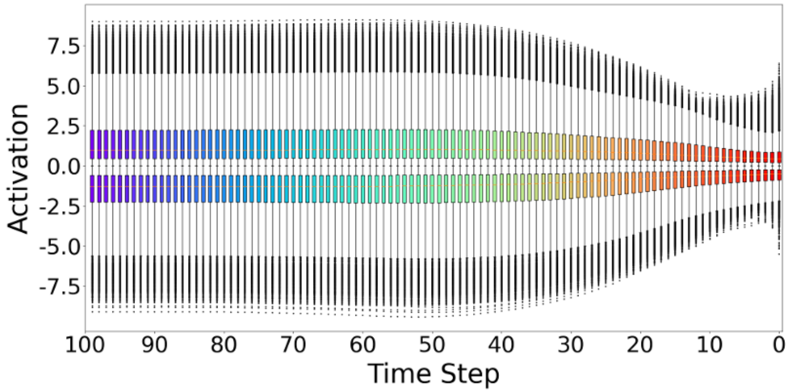
- **Main Contribution**
  - the first full binarization approaches to diffusion models;
  - identify the challenges that make existing binarization methods difficult to transfer to binarize DMs, expecially their activation;
  - achieve impressive 52.7× and 28.0× saving on FLOPs and size.

2 The Rise of BiDM: Bottlenecks of Binarized DMs

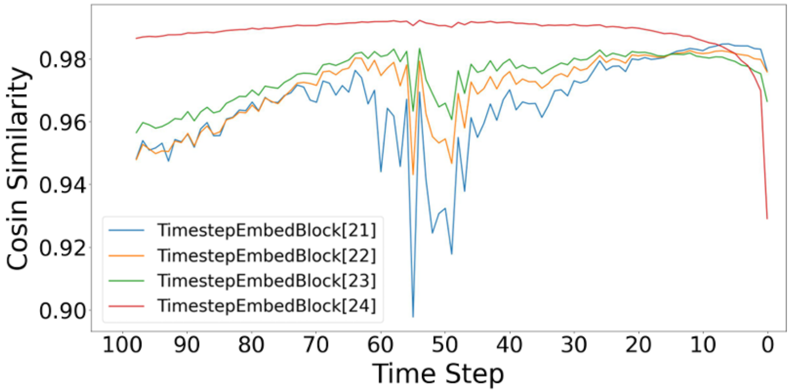
- **Binarized DMs Architecture**
  - **Architecture perspective.** As generative models, DMs have rich intermediate representations closely related to timesteps and highly dynamic activation ranges, which are both very limited in information when binarized weights and activations are used.
- **Distillation for Binarized DMs**
  - **Optimization perspective.** Generative models like DMs are typically required to output complete images, but the highly discrete parameter and feature space make it particularly difficult for binarized DMs to match the ground truth during training.

2 The Rise of BiDM: Timestep-friendly Binary Structure (TBS)

- **From a temporal perspective**
  - **Observation 1**: The activation range varies significantly across long-term timesteps, but the activation features are similar in short-term neighbouring timesteps.



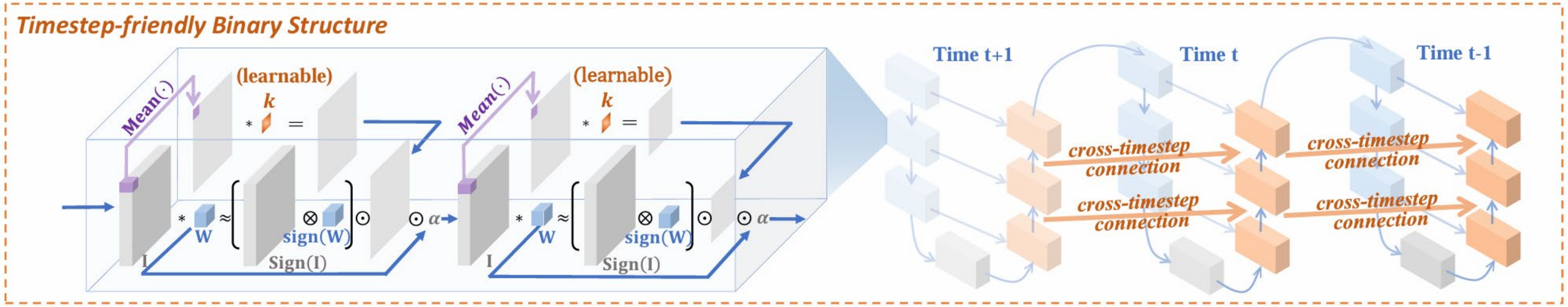
(a) Activation Range



(b) Activation Features

2 The Rise of BiDM: Timestep-friendly Binary Structure (TBS)

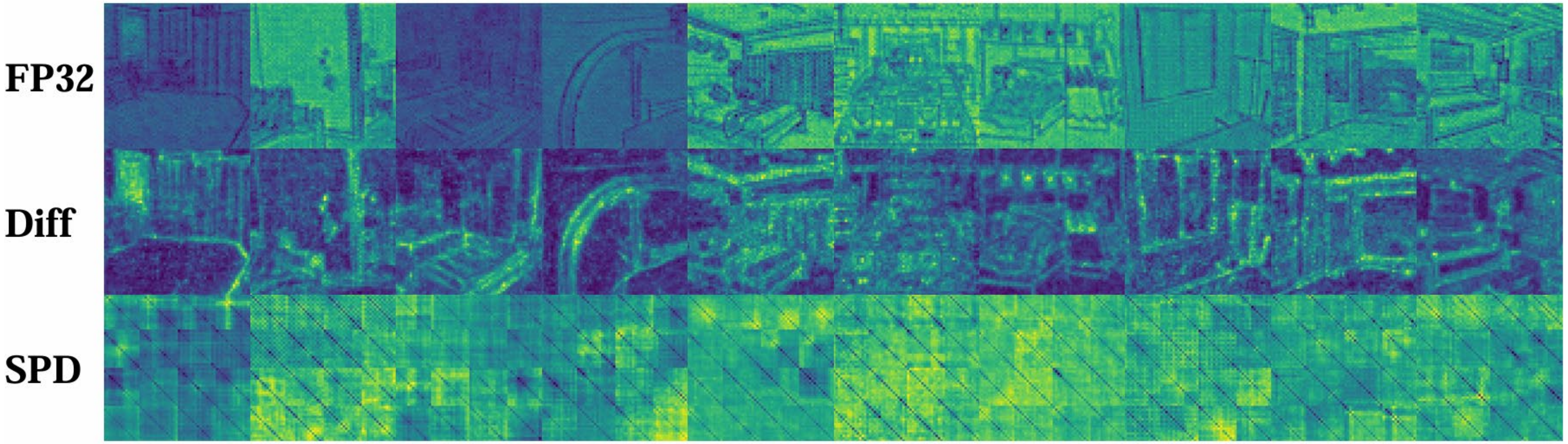
- **TBS to address the highly timestep-correlated activation features**



- Using the binary activation operator proposed by XNOR-Net and making the tiny convolution k learnable can effectively adapt to the dynamic range variations in activations.
- Adding cross-timestep connections between different output blocks can leverage the similarity of features across adjacent timesteps, enhancing the information representation of the binary outputs.

2 The Rise of BiDM: Space Patched Distillation (SPD)

- **From a spatial perspective**
  - **Observation 2**: Conventional distillation struggles to guide fully binarized DMs to align with full-precision DMs, while the features of DM exhibit locality in space during the generation task.



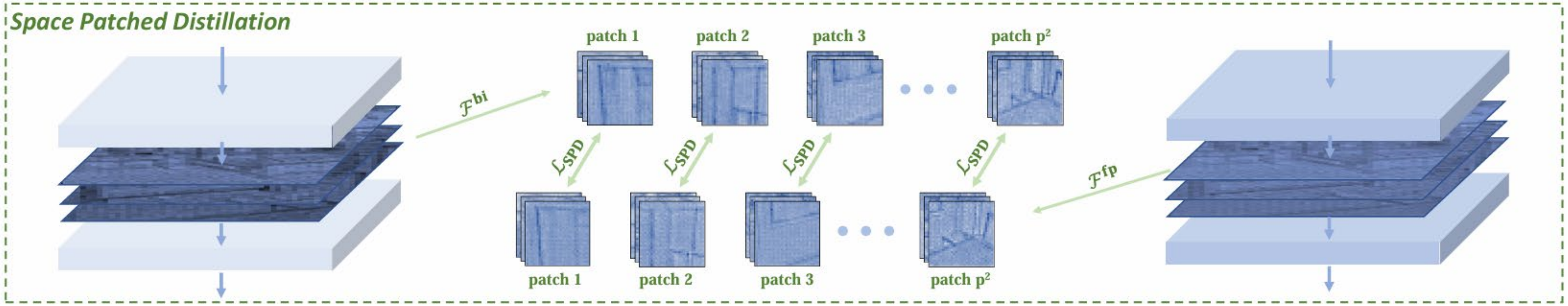
2 The Rise of BiDM: Space Patched Distillation (SPD)

- **SPD for Accurate Optimization**

$$\mathcal{P}^{\text{fp}}_{i,j} = \mathcal{F}^{\text{fp}}_{[:,:,i:i+w/p,j:j+h/p]}, \quad \mathcal{P}^{\text{bi}}_{i,j} = \mathcal{F}^{\text{bi}}_{[:,:,i:i+w/p,j:j+h/p]}$$

$$\mathcal{A}^{\text{fp}}_{i,j} = \mathcal{P}^{\text{fp}}_{i,j}{\mathcal{P}^{\text{fp}}_{i,j}}^{T}, \quad \mathcal{A}^{\text{bi}}_{i,j} = \mathcal{P}^{\text{bi}}_{i,j}{\mathcal{P}^{\text{bi}}_{i,j}}^{T}$$

$$\mathcal{L}^{m}_{\text{SPD}} = \frac{1}{p^2}\sum_{i=0}^{p-1}\sum_{j=0}^{p-1}\left\| \frac{\mathcal{A}^{\text{fp}}_{i,j}}{\|\mathcal{A}^{\text{fp}}_{i,j}\|_2} - \frac{\mathcal{A}^{\text{bi}}_{i,j}}{\|\mathcal{A}^{\text{bi}}_{i,j}\|_2} \right\|_2$$



- SPD divides intermediate features into patches and computes attention within each patch to leverage spatial locality in image generation tasks, enabling more locally targeted optimization alignment.

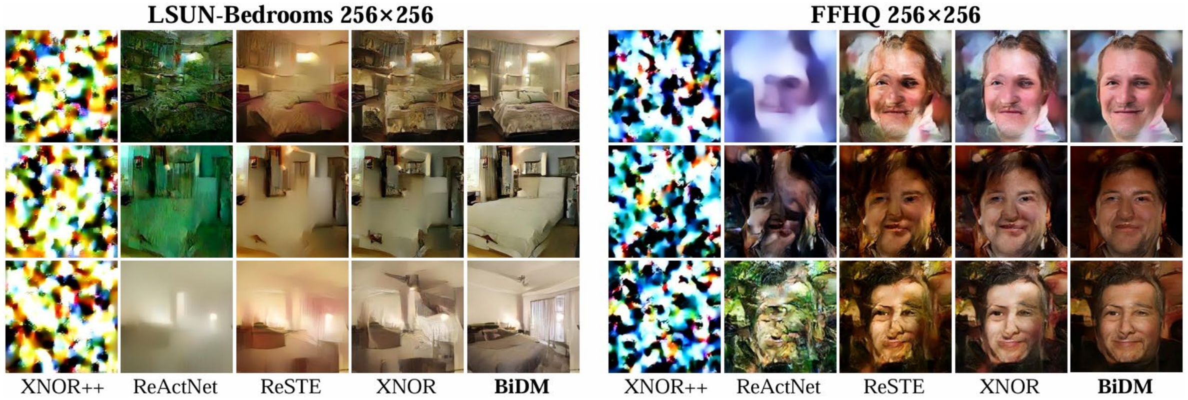
Experiments: Generation Performance

Table 2: Quantization results for LDM on LSUN-Bedrooms, LSUN-Churches and FFHQ datasets.

Model	Dataset	Method	#Bits	FID↓	sFID↓	Precision↑	Recall↑
LDM-4	LSUN-Bedrooms 256 × 256	FP	32/32	2.99	7.08	65.02	47.54
		XNOR++	1/1	319.66	184.75	0.00	0.00
		BBCU	1/1	236.07	89.66	0.59	5.66
		EfficientDM	1/1	194.45	113.24	0.99	9.20
		DoReFa	1/1	188.30	89.28	0.86	0.18
		ReActNet	1/1	154.74	61.50	4.63	9.30
		ReSTE	1/1	59.44	42.16	12.06	2.92
		XNOR	1/1	106.62	56.81	6.82	5.22
		BiDM	1/1	**22.74**	**17.91**	**33.54**	**19.90**
LDM-8	LSUN-Churches 256 × 256	FP	32/32	4.36	16.00	74.64	48.98
		XNOR++	1/1	292.48	168.65	0.02	0.00
		DoReFa	1/1	162.06	95.37	7.85	0.74
		ReActNet	1/1	56.39	54.68	45.13	2.06
		ReSTE	1/1	47.88	52.44	51.98	3.34
		XNOR	1/1	42.87	49.24	51.53	4.28
		BiDM	1/1	**29.70**	**45.14**	**55.75**	**14.80**
LDM-4	FFHQ 256 × 256	FP	32/32	4.87	6.96	74.73	50.57
		XNOR++	1/1	379.49	320.64	0.00	0.00
		DoReFa	1/1	214.06	177.63	2.09	0.00
		ReActNet	1/1	147.88	141.31	3.36	0.69
		ReSTE	1/1	144.37	97.43	4.03	0.03
		XNOR	1/1	89.37	54.04	31.31	4.11
		BiDM	1/1	**43.42**	**32.35**	**49.44**	**13.96**

Table 1: Binarization results for DDIM on CIFAR-10 datasets with 100 steps.

Model	Dataset	Method	#Bits	IS↑	FID↓	sFID↓	Precision↑
DDIM	CIFAR-10 32 × 32	FP	32/32	8.90	5.54	4.46	67.92
		XNOR++[2]	1/1	2.23	251.14	60.85	44.98
		DoReFa[78]	1/1	1.43	397.60	139.97	0.17
		ReActNet[33]	1/1	3.35	231.55	119.80	18.37
		ReSTE[62]	1/1	1.26	394.29	125.84	0.18
		XNOR[49]	1/1	4.23	113.36	27.67	46.96
		BiDM	**1/1**	**5.18**	**81.65**	**25.68**	**52.92**



NEURAL INFORMATION PROCESSING SYSTEMS

Conclusion

- **Novel Technique:** the first full binarization approaches for diffusion models.
- **Summary of Observations:** provide observations on the spatiotemporal properties of full-precision DMs to effectively guide the design of binarized DMs.
- **Good Precision:** show improvements of full DM binarization than existing methods across several mainstream Image Generation tasks.
- **High efficiency:** achieves impressive $\mathbf{52.7}\times$ computational FLOPs and $\mathbf{28.0}\times$ storage saving.

Thank you!

Paper: https://neurips.cc/virtual/2024/poster/93620
Code: https://github.com/Xingyu-Zheng/BiDM
(star is welcome)