BinaryDM: Accurate Weight Binarization for Efficient Diffusion Models

Xingyu Zheng[1] Xianglong Liu[1†] Haotong Qin[2] Xudong Ma[1] Mingyuan Zhang[3]
Haojie Hao[1] Jiakai Wang[4] Jinyang Guo[1] Michele Magno[1]

[1]Beihang University [2]ETH Zürich [3]Nanyang Technological University
[4]Zhongguancun Laboratory [5]Xi'an Jiaotong University

Paper: https://iclr.cc/virtual/2025/poster/29258
Code: https://github.com/Xingyu-Zheng/BinaryDM
(star is welcome)

1 Introduction: Diffusion Binarization

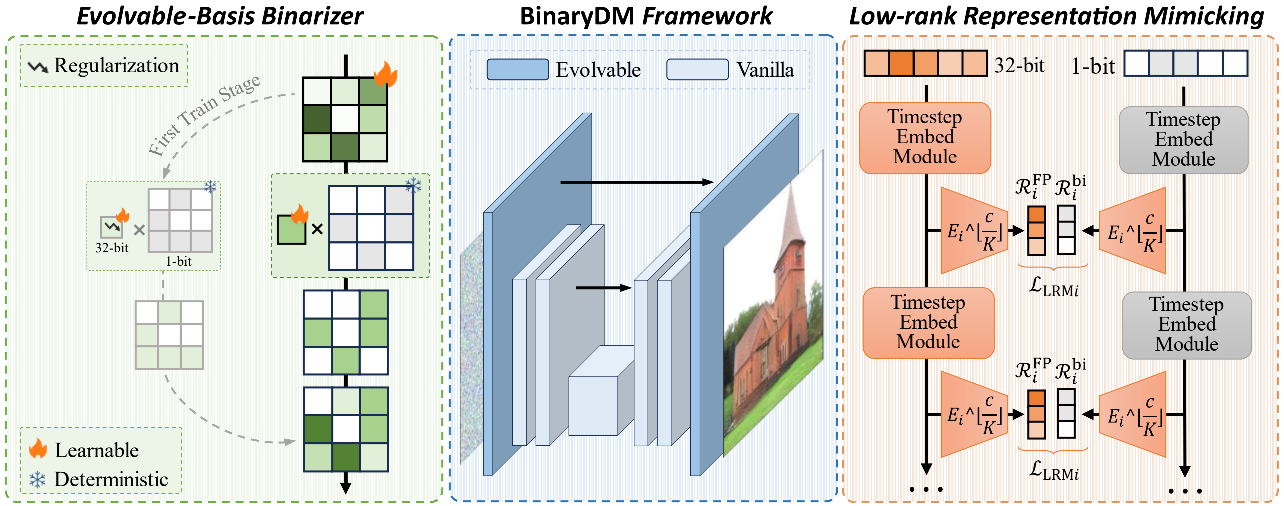
- **Large Pre-trained Diffusion models**
  - Diffusion models (DMs) have garnered impressive attention and applications in various fields, such as image, speech and video
  - it still suffers expensive FP32 parameters and operations
- **Network Binarization**
  - compression by binarizing parameters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

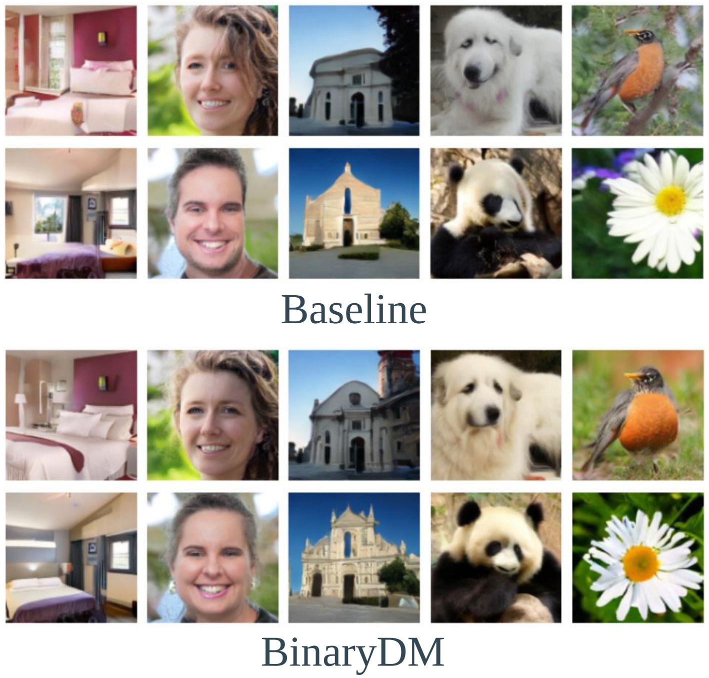
$$Q_{\mathrm{x}}(\mathbf{x}) = \alpha\, \mathbf{B}_{\mathbf{x}}$$

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

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

1 Introduction: Overview





- **Main Contribution**
  - W1A4 BinaryDM achieves as low as 7.74 FID and saves the performance from collapse (baseline FID 10.87)
  - W1A4 BinaryDM achieves impressive 15.2x OPs and 29.2x model size savings, showcasing its substantial potential for edge deployment

2 The Rise of BinaryDM: Bottlenecks of Binarized DMs

- **Binarized DMs Architecture**
  - **Representation perspective:** Weight binarization severely restricts the feature extraction capability of diffusion models, causing significant damage to information in critical representations of generative models.

- **Distillation for Binarized DMs**
  - **Optimization perspective**: Introducing discrete binarization functions in DMs poses a significant hurdle to stable convergence.

2 The Rise of BinaryDM: Evolvable-Basis Binarizer

EBB enables a smooth evolution of DMs from full-precision to accurately binarized

Learnable Multi-Basis: In the forward propagation of the first stage, EBB is defined as:

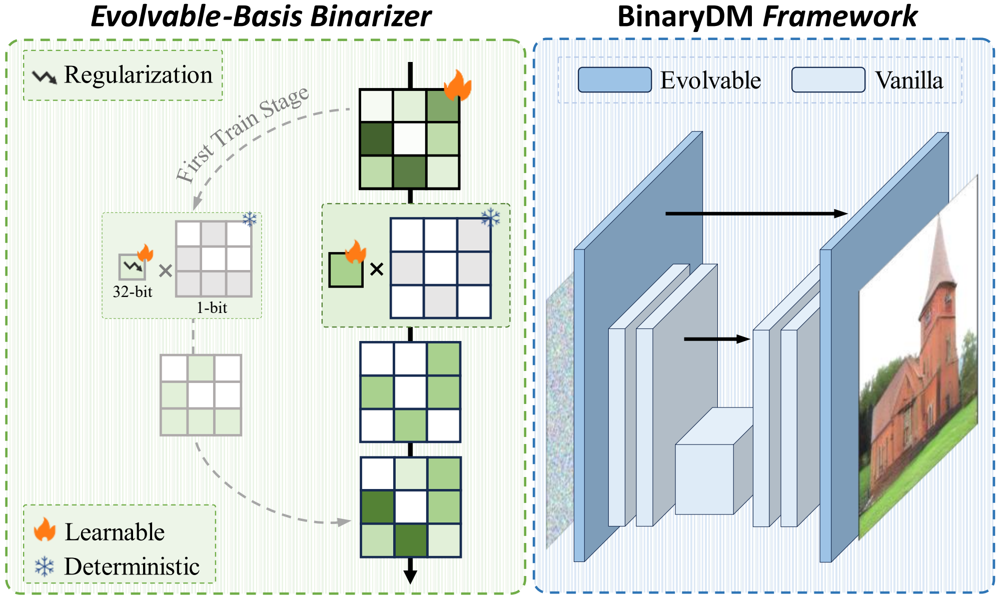
$$w_{\text{EBB}}^{\text{bi}} = \sigma_I \text{sign}(w) + \sigma_{II} \text{sign}\big(w - \sigma_1 \text{sign}(w)\big)$$

Transition Strategy: In the first stage, regularization loss is applied to the higher-order learnable scaling factors, encouraging them to approach zero:

$$\mathcal{L}_{\text{EBB}} = \tau \frac{1}{N} \sum_{i=1}^{N} \sigma_{II}^{i}$$

In the second stage, all higher-order terms are removed, and the forward propagation is simplified to:

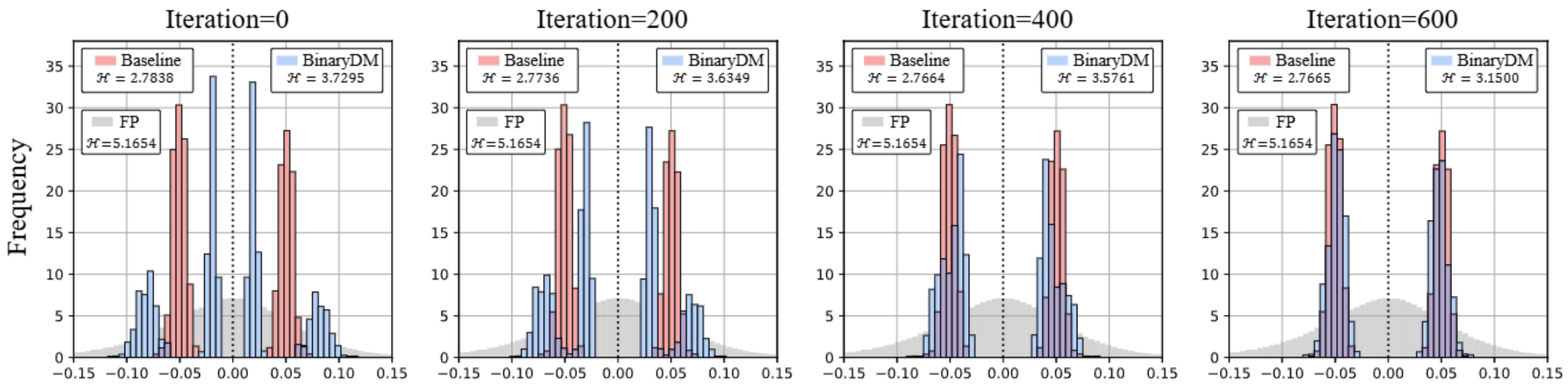
$$w^{\text{bi}} = \sigma_I \text{sign}(w)$$



Location Selection: In BinaryDM, EBB is partially applied to crucial and parameter-sparse locations of the diffusion models to reduce unnecessary evolution processes and the associated training overhead.

2 The Rise of BinaryDM: Evolvable-Basis Binarizer (EBB)

- **From the representation perspective**
  - EBB possesses a broader representation range at the early stage and then gradually transitions to a single-basis state, while the quantitative information entropy $\mathcal{H}$ further illustrates its enhanced representation capacity.



Comparison of binarized weights(channel-wise) for a convolutional layer.

2 The Rise of BinaryDM: Low-rank Representation Mimicking

LRM for Accurate Optimization

- We use principal component analysis (PCA) to project representations to low-rank space:

$$\boldsymbol{\mathcal{R}}_i^{\mathrm{FP}}\left(\boldsymbol{x}_t, t\right)=\hat{\boldsymbol{\varepsilon}}_{\theta_i}^{\mathrm{FP}}\left(\boldsymbol{x}_t, t\right) E_i^{\left\lceil\frac{c}{K}\right\rfloor}, \quad \boldsymbol{\mathcal{R}}_i^{\mathrm{bi}}\left(\boldsymbol{x}_t, t\right)=\hat{\boldsymbol{\varepsilon}}_{\theta_i^{\mathrm{bi}}}^{\mathrm{bi}}\left(\boldsymbol{x}_t, t\right) E_i^{\left\lceil\frac{c}{K}\right\rfloor}$$

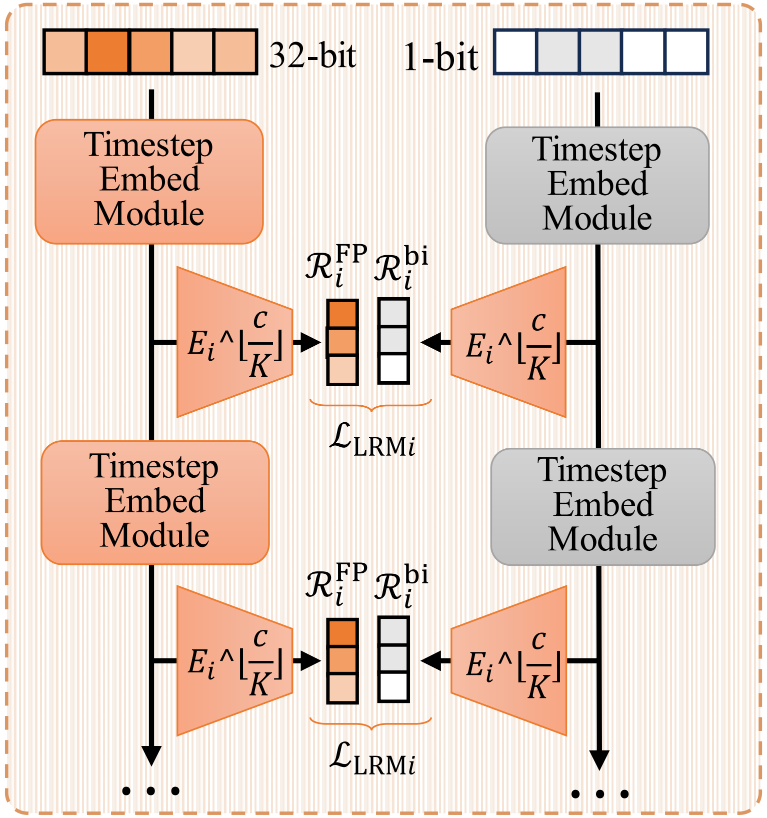
- We construct a mean squared error (MSE) loss between the i-th module of low-rank representations between full-precision and binarized DMs:

$$\mathcal{L}_{\mathrm{LRM}i}=\left\|\boldsymbol{\mathcal{R}}_i^{\mathrm{FP}}-\boldsymbol{\mathcal{R}}_i^{\mathrm{bi}}\right\|$$

- The total loss function:

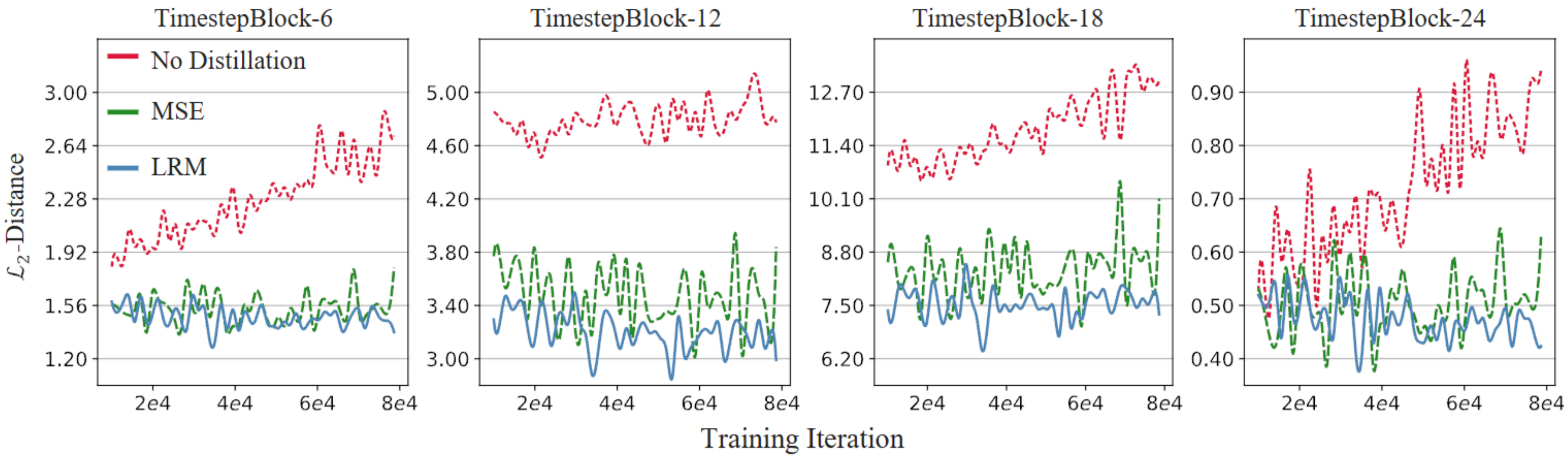
$$\mathcal{L}_{\text{total}}=\mathcal{L}_{\text{simple}}+\mathcal{L}_{\text{EBB}}+\lambda \frac{1}{M} \sum_{i=1}^M \mathcal{L}_{\mathrm{LRM}i}$$

Low-rank Representation Mimicking



2 The Rise of BinaryDM: Low-rank Representation Mimicking

- **From the optimization perspective**
  - LRM enables binarized DMs to mimic the representation of full-precision counterparts, improving the optimization process by introducing additional supervision.



Experiments: Generation Performance

Table 2: Results for LDM on multiple datasets in unconditional generation by DDIM with 100 steps.

Model	Dataset	Method	#Bits	Size(MB)	FID↓	sFID↓	Precision↑	Recall↑
LDM-4	LSUN-Bedrooms 256 × 256	FP	32/32	1045.4	3.09	7.08	65.82	45.36
		LSQ	2/32	69.8	7.49	12.79	64.02	37.60
		Baseline	1/32	35.8	8.43	13.11	65.45	29.88
		BinaryDM	1/32	35.8	**6.99**	**12.15**	**67.51**	**36.80**
		Q-Diffusion	2/8	69.8	62.01	33.56	16.48	14.12
		LSQ	2/8	69.8	6.48	11.66	62.55	38.92
		Baseline	1/8	35.8	9.37	12.10	64.36	30.76
		BinaryDM	1/8	35.8	**6.51**	**11.67**	**65.80**	**35.28**
		Q-Diffusion	4/4	134.9	427.46	277.22	0.00	0.00
		EfficientDM	4/4	134.9	10.60	-	-	-
		LSQ	2/4	69.8	12.95	12.79	55.97	34.30
		Baseline	1/4	35.8	10.87	15.46	64.05	26.50
		TDQ	1/4	35.8	11.28	12.80	55.14	27.32
		ReActNet	1/4	35.8	10.23	13.02	61.43	29.68
		Q-DM	1/4	35.8	9.99	11.96	57.62	29.30
		INSTA-BNN	1/4	35.8	9.42	12.39	60.05	31.08
		BI-DiffSR	1/4	35.8	8.58	11.81	62.61	30.86
		BinaryDM	1/4	35.8	**7.74**	**10.80**	**64.71**	**32.98**
LDM-8	LSUN-Churches 256 × 256	FP	32/32	1125.2	4.82	17.66	75.18	46.80
		LSQ	2/32	74.1	8.16	19.87	74.98	35.76
		Baseline	1/32	38.1	9.91	17.94	74.89	26.88
		BinaryDM	1/32	38.1	**8.14**	**17.44**	**75.51**	**34.56**
		Q-Diffusion	2/8	74.1	201.23	238.70	2.39	8.60
		LSQ	2/8	74.1	8.11	19.25	77.04	34.98
		Baseline	1/8	38.1	10.94	16.95	74.30	25.66
		BinaryDM	1/8	38.1	**8.63**	**15.13**	**77.74**	**33.48**
		EfficientDM	4/4	144.2	14.34	-	-	-
		Q-Diffusion	4/4	144.2	198.35	184.43	5.48	0.12
		LSQ	2/4	74.1	10.00	19.08	74.93	25.80
		Baseline	1/4	38.1	12.98	21.55	70.78	25.30
		BinaryDM	1/4	38.1	**9.91**	**18.04**	**73.72**	**29.96**
LDM-4	FFHQ 256 × 256	FP	32/32	1045.4	6.64	14.16	76.88	50.82
		Q-Diffusion	4/32	134.9	11.60	10.30	-	-
		Baseline	1/32	35.8	10.49	11.56	72.64	39.62
		BinaryDM	1/32	35.8	**8.70**	**9.68**	**73.92**	**42.22**
		Q-Diffusion	8/8	265.0	10.87	10.01	-	-
		Q-Diffusion	4/8	134.9	11.45	9.06	-	-
		Baseline	1/8	35.8	10.79	10.77	73.20	41.70
		BinaryDM	1/8	35.8	**9.58**	**10.74**	**74.48**	**41.75**
		Baseline	1/4	35.8	15.07	12.48	74.34	35.12
		BinaryDM	1/4	35.8	**12.34**	**11.18**	**74.83**	**38.09**

Table 4: Ablation results on LSUN-Bedrooms 256 × 256.

Method	#Bits	FID↓	sFID↓	Prec.↑	Recall↑
FP	32/32	3.09	7.08	65.82	45.36
Vanilla	1/32	8.43	13.11	65.45	29.88
+EBB	1/32	7.39	12.34	65.98	35.84
+LRM	1/32	**6.99**	**12.15**	**67.51**	**36.80**

Table 5: Inference efficiency of our proposed BinaryDM of LDM-4 on LSUN-Bedrooms 256 × 256

Model	Method	#Bits	Size(MB)	OPs$_{(\times 10^9)}$	FID↓
LDM-4	Full-Precision	4/4	1045.4	96.0	3.09
	Q-Diffusion	4/4	134.9	24.3	427.46
	EfficientDM	4/4	134.9	24.3	10.60
	LSQ	2/4	69.8	12.3	12.95
	BinaryDM	1/4	**35.8**	**6.3**	**7.74**

Table 6: Training time-cost of BinaryDM compared to the advanced PTQ method.

Dataset	Method	#Bits	Size(MB)	Time(h)	FID↓
LSUN-Bedrooms	Q-Diffusion	4/4	134.9	13.7	427.46
	BinaryDM	1/4	**35.8**	**11.3**	**13.93**
LSUN-Churches	Q-Diffusion	4/4	144.2	10.9	198.35
	BinaryDM	1/4	**38.1**	**9.0**	**15.11**

Baseline

BinaryDM

Conclusion

- From the representation perspective, we present an Evolvable-Basis Binarizer (EBB) to enable a smooth evolution of DMs from full-precision to accurately binarized. EBB enhances information representation in the initial stage through the flexible combination of multiple binary bases and applies regularization to evolve into efficient single-basis binarization.
- From the optimization perspective, a Low-rank Representation Mimicking (LRM) is applied to assist the optimization of binarized DMs. The LRM mimics the representations of full-precision DMs in low-rank space, alleviating the direction ambiguity of the optimization process caused by fine-grained alignment.
- **W1A4** BinaryDM achieves as low as **7.74 FID** and saves the performance from collapse (baseline FID 10.87), achieving impressive **15.2x OPs** and **29.2x** model size savings, showcasing its substantial potential for edge deployment.

Thank you!

Paper: https://iclr.cc/virtual/2025/poster/29258
Code: https://github.com/Xingyu-Zheng/BinaryDM
(star is welcome)