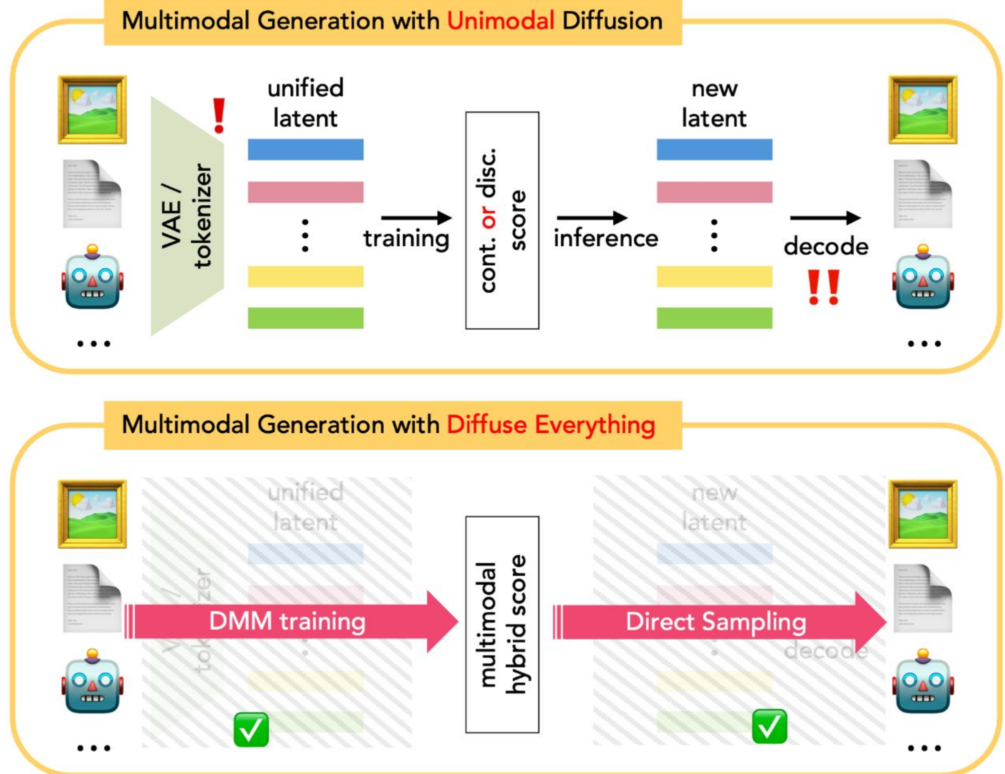
Diffuse Everything : Multimodal Diffusion Models on Arbitrary State Spaces

Kevin Rojas*[1] Yuchen Zhu*[1] Sichen Zhu[1] Felix X.-F. Ye[2] Molei Tao[1]

[1] Georgia Institute of Technology [2] SUNY Albany

PAPER CODE POST

TL;DR



! Representation unification is challenging and expensive. Good encoders are hard to train.

!! Decoding capped generation accuracy. Bad decoders creates artifacts.

✅ Diffuse Everything have minimum encoders, fewer modules, and more parameter efficiency

✅ Diffuse Everything introduce new design space for multimodal diffusion generation

Motivation

- ➢ Joint generation of multimodal data is **IMPORTANT**!
- ➢ Diffusion models are **SOTA** for many types of unimodal data, with fantastic conditional generation capability.
- ➢ Good tokenizers and VAEs are **CHALLENGING** to train.

Goal

Develop **NEW** diffusion models to generate multimodal data in their **native state space**, **bypassing** the need for tokenizers/VAEs/encoders.

Key Findings

1. Multimodal diffusions can be built by **combining** unimodal diffusions and be trained by learning scores of **joint distribution**.
2. Training multimodal diffusion models is **provably** as simple as jointly optimizing **sum** of unimodal diffusion model losses.
3. Decoupled time enables **any-to-any** generation in one model, and a new guidance scheme named noisy guidance.
4. Multimodal diffusion on native state spaces are much more **parameter-efficient**.

Approach

MODALITY		DYNAMIC	OBJECTIVE
image	Continuous	$\mathrm{d}X_t = f(X_t,t)\mathrm{d}t + g(t)\mathrm{d}W_t$	$\mathbb{E}\left[\frac{1}{2}g^2(t)\lVert s_\theta^X - \nabla \log p_t(X_t\mid x_0)\rVert^2\right]$
text	Discrete	$Y_t \sim \mathrm{CTMC}(Q_t)$	$\mathbb{E}\left[\sum_{y\neq Y_t} Q_t(Y_t,y)\left((s_\theta^Y)_y - \frac{p_t(y\mid y_0)}{p_t(Y_t\mid y_0)}\log(s_\theta^Y)_y\right)\right]$
image+text	!!NEW!! Continuous-Discrete	$\begin{cases}\mathrm{d}X_t = f(X_t,t)\mathrm{d}t + g(t)\mathrm{d}W_t\\ Y_s \sim \mathrm{CTMC}(Q_s)\end{cases}$	$\mathbb{E}\left[\frac{1}{2}g^2(t)\lVert s_\theta^X - \nabla \log p_t(X_t\mid x_0)\rVert^2\right] + \mathbb{E}\left[\sum_{y\neq Y_s} Q_s(Y_s,y)\left((s_\theta^Y)_y - \frac{p_s(y\mid y_0)}{p_s(Y_s\mid y_0)}\log(s_\theta^Y)_y\right)\right]$
mixed modalities	!!NEW!! Diffuse Everything	$\begin{cases}\boldsymbol{X}_t = (X^1_{t_1},\dots,X^i_{t_i},\dots,X^n_{t_n})\\ X^i_{t_i} \sim \text{Markov process with generator } \mathcal{L}_{X^i}\end{cases}$	$\mathbb{E}\left[\sum_{i=1}^n \frac{\mathcal{L}_{X^i}(p_{t\mid 0}/\beta_\theta)(\boldsymbol{X}_t,\boldsymbol{t})}{(p_{t\mid 0}/\beta_\theta)(\boldsymbol{X}_t,\boldsymbol{t})} - \mathcal{L}_{X^i}\log(p_{t\mid 0}/\beta_\theta)(\boldsymbol{X}_t,\boldsymbol{t})\right]$

Methodology

Multimodal Gen. with Denoising Markov Models

Generative modeling with dynamic:

$$\boldsymbol{X}_t = (X^1_{t^1},\dots,X^i_{t^i},\dots,X^n_{t^n}),\ 0\le t^1,\dots,t^n \le T$$
$$(X^1_0,\dots,X^i_0,\dots,X^n_0) \sim p_{\text{data}}(\boldsymbol{x})$$

… requires only learning score of ↳ **Theorem 1**

$$p(\boldsymbol{x}_t,\boldsymbol{t}) = \mathbb{P}(X^1_{t^1} = (\boldsymbol{x}_t)_1,\dots,X^n_{t^n} = (\boldsymbol{x}_t)_n)$$

Training with Generalized Score Matching

Equivalent ↳ **Theorem 2**

$$\mathcal{I}_{\text{GESM}} = \mathop{\mathbb{E}}_{\boldsymbol{t},\boldsymbol{x}_t\sim p(\cdot,\boldsymbol{t})}\left[\sum_{i=0}^n \frac{\mathcal{L}_{X^i}(p/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})}{(p/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})} - \mathcal{L}_{X^i}\log(p/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})\right]$$
$$\mathcal{I}_{\text{GDSM}} = \mathop{\mathbb{E}}_{\boldsymbol{t},p_0,p_{t\mid 0}}\left[\sum_{i=1}^n \frac{\mathcal{L}_{X^i}(p_{t\mid 0}/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})}{(p_{t\mid 0}/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})} - \mathcal{L}_{X^i}\log(p_{t\mid 0}/\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})\right]$$
$$\mathcal{I}_{\text{GISM}} = \mathop{\mathbb{E}}_{\boldsymbol{t},p_t}\left[\sum_{i=1}^n \frac{\mathcal{L}^*_{X^i}\beta_\theta(\boldsymbol{x}_t,\boldsymbol{t})}{\beta_\theta(\boldsymbol{x}_t,\boldsymbol{t})} - \mathcal{L}^*_{X^i}\log(\beta_\theta)(\boldsymbol{x}_t,\boldsymbol{t})\right]$$

Continuous-Discrete Diffusion Model

Forward Process

$$\begin{cases}\mathrm{d}X_t = f(X_t,t)\mathrm{d}t + g(t)\mathrm{d}B_t\\ Y_s \sim \mathrm{CTMC}(Q_s),\ (X_0,Y_0)\sim p_{\text{data}}(x,y)\end{cases}$$

Time reversal

Backward Process

Score of joint dist.

$$\begin{cases}\mathrm{d}\overleftarrow{X}_t = -f(\overleftarrow{X}_t,T-t) + g^2(T-t)\nabla\log\overleftarrow{p}_{t,s}(\overleftarrow{X}_t,\overleftarrow{Y}_s)\mathrm{d}t + g(T-t)\mathrm{d}B_t\\ \overleftarrow{Y}_s \sim \mathrm{CTMC}(\overleftarrow{Q}(\overleftarrow{X}_t,t,s)),\ (\overleftarrow{X}_0,\overleftarrow{Y}_0)\sim p(x,y,T,T)\\ \overleftarrow{p}_{t,s} = p_{T-t,T-s},\quad \overleftarrow{Q}(\overleftarrow{X}_t,t,s)_{y',y} = \frac{\overleftarrow{p}_{t,s}(\overleftarrow{X}_t,y')}{\overleftarrow{p}_{t,s}(\overleftarrow{X}_t,y)}(Q_{T-s})_{y,y'}\end{cases}$$

- A combination of diffusion SDE on Euclidean space and CTMC jump process on finite state space.
- Aiming at generating data with both continuous values and discrete values

Denoising training objective

Requires multimodal input

Unimodal conditional score

$$\mathop{\mathbb{E}}_{\substack{t,s,x_0,y_0\sim p_0\\ x_t,y_s\sim p_{t,s\mid 0}}}\Big[\underbrace{\frac{1}{2}g^2(t)\lVert s_\theta^X - \nabla\log p_t(x_t\mid x_0)\rVert^2}_{\text{Euclidean Score Matching}} + \underbrace{\sum_{y\neq y_s} Q_s(y_s,y)\left((s_\theta^Y)_y - \frac{p_s(y\mid y_0)}{p_s(y_s\mid y_0)}\log(s_\theta^Y)_y\right)}_{\text{Discrete Score Matching}}\Big]$$

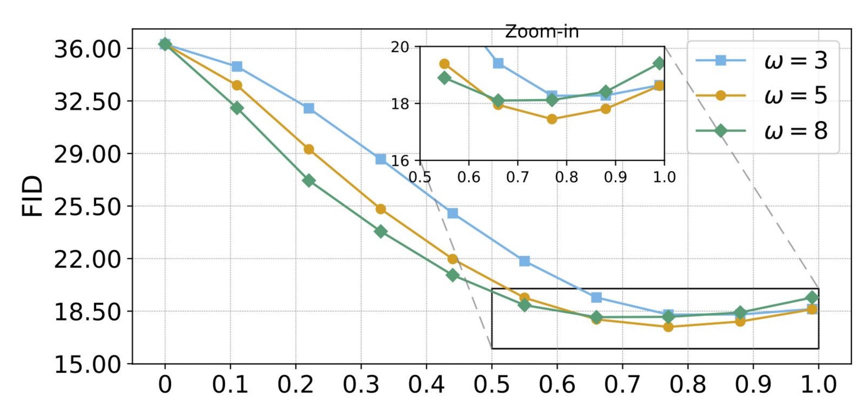
Decoupled Time Design

Decoupled time design means:

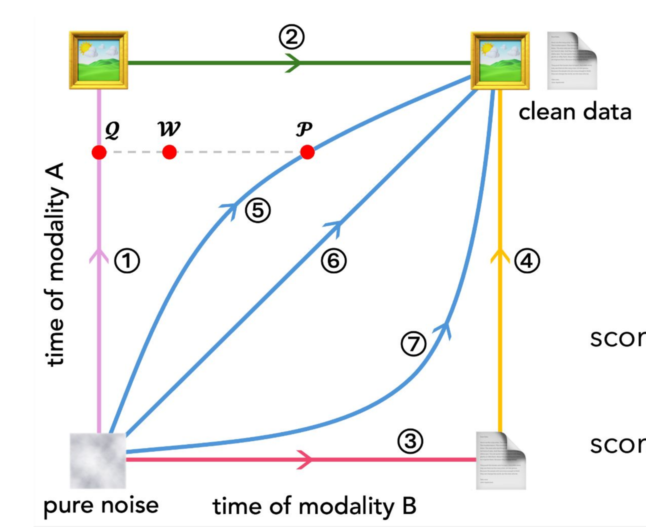
- ❑ Each modality is noised and denoised at independent pace
- ❑ Requires learning denoiser/score at more scenarios

Benefit 1 – Flexible any-to-any generation

- Given a partially noisy text Y_s, simulating only the X -backward dynamics samples from $p_{\text{data}}(x\mid Y_s,s)$
- Given a partially noisy image X_t, simulating the Y-backward dynamics samples from $p_{\text{data}}(y\mid X_t,t)$



FID on MS-COCO v.s. sigma. Optimum achieved around 0.6-0.8



①: sample 🖼
②: sample 📄 given 🖼
③: sample 📄
④: sample 🖼 given 📄
①+②, ③+④, ⑤, ⑥, ⑦: sample 🖼 📄

score $\mathcal{P}$ + score $\mathcal{Q}$ → classifier-free guidance

score $\mathcal{P}$ + score $\mathcal{W}$ → noisy guidance !!NEW!!

Benefit 2 – Better guidance scheme than CFG

- Noisy guidance: guiding with a **partial corrupted conditional** model
- Achieving better diversity-quality trade-off
- Recover CFG when s = 0, sigma = T

Noisy Guidance

$$\omega s_\theta(x_t,y_s,t,s) + (1-\omega)s_\theta(x_t,y_\sigma,t,\sigma) \qquad \sigma > s$$

Inference Algorithms

Algorithm 3 Discrete Sampler with τ-leaping

Require: N : Number of steps, ω : Guidance Strength
1: $[a,b]$: Guidance Interval, model : $s_\theta(x_t,y_s,t,s,\omega)$
2: x : a clean image condition
Ensure: $y_0 \sim p_{\text{data}}(\cdot\mid x)$
3: $y_t \leftarrow [\mathbf{M},\dots,\mathbf{M}]$
4: **for** t in times **do**
5: $\omega_t = \omega$ if $t\in[a,b]$ else 1.
6: $s_\theta^x, s_\theta^y \leftarrow s_\theta(x,y_t,0,t,\omega_t)$
7: $y_t \leftarrow \tau$-leaping$(s_\theta^y, y_t, t, \lvert \mathrm{d}t\rvert)$
8: **end for**
9: **return** y_0

Sampling discrete data only

Algorithm 4 Continuous Sampler with Heun's method

Require: N : Number of steps, ω : Guidance Strength
1: $[a,b]$: Guidance Interval, model : $s_\theta(x_t,y_s,t,s,\omega)$
2: y : A clean text condition
Ensure: $x_0 \sim p_{\text{data}}(\cdot\mid y)$
3: $x_t \leftarrow \mathcal{N}(0,I)$
4: **for** t in times **do**
5: $\omega_t = \omega$ if $t\in[a,b]$ else 1.
6: $s_\theta^x, s_\theta^y \leftarrow s_\theta(x_t,y,t,0,\omega_t)$
7: $v_{\text{old}} = f(x_t,t) - \frac{1}{2}g^2(t)s_\theta^x$
8: $\hat{x} \leftarrow x_t + v_{\text{old}}\mathrm{d}t$, $\hat{s}_\theta^x, \hat{s}_\theta^y \leftarrow s_\theta(\hat{x},y,t+\mathrm{d}t,0,\omega_t)$
9: $v_{\text{new}} = f(\hat{x}_t,t) - \frac{1}{2}g^2(t)\hat{s}_\theta^x$
10: $x_t \leftarrow x_t + \frac{1}{2}\cdot(v_{\text{old}} + v_{\text{new}})\mathrm{d}t$
11: **end for**
12: **return** x_0

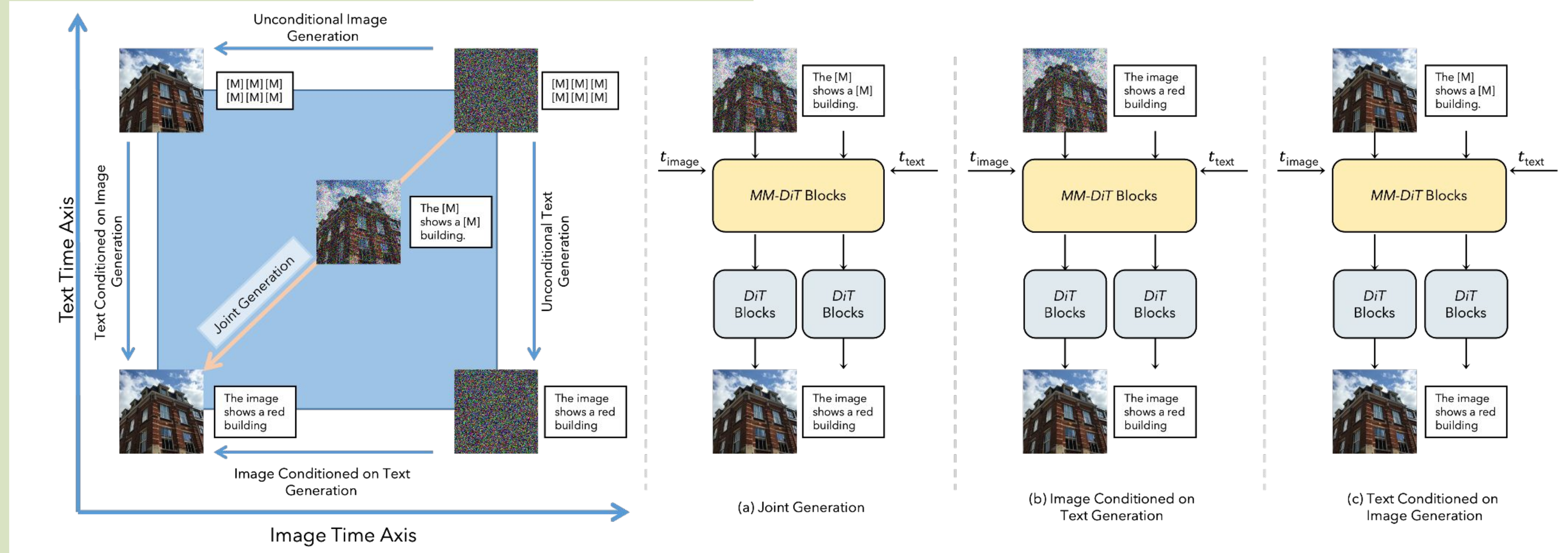
Sample continuous data only

Algorithm 5 Multimodal Sampler with τ-leaping and Heun's Method

Require: N : Number of steps, ω : Guidance Strength,
1: $[a,b]$: Guidance Interval,
2: model : $s_\theta(x_t,y_s,t,s,\omega)$
Ensure: $x_0, y_0 \sim p_{\text{data}}$
3: $x_t \leftarrow \mathcal{N}(0,I)$, $y_t \leftarrow [\mathbf{M},\dots,\mathbf{M}]$
4: **for** t in times **do**
5: $\omega_t = \omega$ if $t\in[a,b]$ else 1.
6: $s_\theta^x, s_\theta^y \leftarrow s_\theta(x_t,y_t,t,t,\omega_t)$, $v_{\text{old}} = f(x_t,t) - \frac{1}{2}g^2(t)s_\theta^x$
7: $\hat{x} \leftarrow x_t + v_{\text{old}}\mathrm{d}t$
8: $\hat{s}_\theta^x, \hat{s}_\theta^y \leftarrow s_\theta(\hat{x},y_t,t+\mathrm{d}t,t,\omega_t)$, $v_{\text{new}} = f(\hat{x}_t,t) - \frac{1}{2}g^2(t)\hat{s}_\theta^x$
9: $x_t \leftarrow x_t + \frac{1}{2}\cdot(v_{\text{old}} + v_{\text{new}})\mathrm{d}t$, $y_t \leftarrow \tau$-leaping$(s_\theta^y, y_t, t, \lvert \mathrm{d}t\rvert)$
10: **end for**
11: **return** x_0, y_0

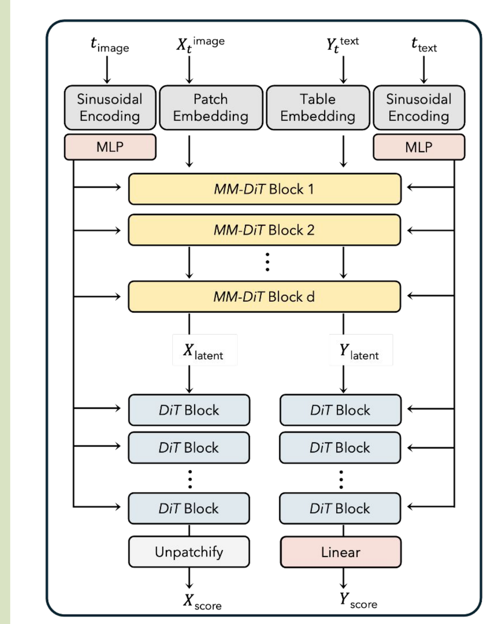
Jointly sample continuous and discrete data

Text-Image generation



Jointly generate images and its captions, and we..

- **Minimally rely on pretrained model** (except for one image VAE for dimension reduction). No T5/CLIP/ViT/etc.
- Achieve satisfactory performance while using **much smaller model**
- Design **multi-stage training** strategy to aid training of decoupled time



score backbone

Table 1. Results on the **text to image conditional generation** on MS-COCO. We mark the extra encoders leveraged by each model with the corresponding sizes and types. SR: super resolution, TE: text encoder, VAE: variational autoencoder, VE: visual encoder, VQ-GAN: Vector Quantized GAN, VQ-VAE: vector-quantized variational autoencoder.

Model	FID	Number of Images	#Params	Extra Encoders
Models for Text-to-Image generation only				
DALL-E 2 (Ramesh et al., 2022)	10.39	650M	6.5B	123M (TE) + 700M (SR)
Imagen (Saharia et al., 2022)	7.27	860M	3B	4.6B (TE) + 600M (SR)
Stable Diffusion (Rombach et al., 2022)	12.63	400M	1.45B	123M (TE) + 83M (VAE)
PixArt-α XL/2 (Chen et al., 2023)	7.32	25M	600M	123M (TE) + 83M (VAE)
MMDiT-improved (Ifriqi et al., 2024)	6.79	12M	600M	123M (TE) + 83M (VAE)
Models for multimodal generation and understanding				
Show-o (Xie et al., 2024)	9.24	35M	1.3B	115M (VE) + 307M (VQ-VAE)
Transfusion (Zhou et al., 2024)	6.78	692M	7B	86M (VAE)
Chameleon (Meta, 2024)	26.7	600M	7B	307M (VQ-GAN)
JetFormer (Tschannen et al., 2024)	20.86	1B	2.75B	—
Models for multimodal generation only				
Versatile Diffusion (Xu et al., 2023)	11.10	400M	1.45B	123M (TE) + 83M (VAE) + 110M (TE)
UniD3 (Hu et al., 2022)	25.11	592K	600M	123M (TE) + 307M (VQ-GAN)
Our model	16.16	12M	481M	83M (VAE)

FID-30k evaluation on MS-COCO. Our model trained on SAM-LLaVA.

Mixed-type Tabular data Synthesis

Jointly generate tabular data with both categorical values (eg. age) and continuous values (eg. income), and we …

- Achieve comparable or beat previous SOTA method with a **significantly smaller model**
- Design a new score backbone based on modification of DiT for mixed-type data, which is **effective**!

Table 2. Performance on the **Trend** metric in percentage (%). Higher values indicate better performance. Best performance in **bold**. Second best in underline.

Methods	#Parameters	Adult	Default	Shoppers	Magic	Beijing	News
GOGGLE (Liu et al., 2023)	~ 5.6M	54.71	78.06	76.10	90.53	54.06	76.81
STaSy (Kim et al., 2022)	~ 10.3M	85.49±0.25	94.04±0.26	91.51±0.15	93.39±0.53	92.00±0.10	96.93±0.04
CoDi (Lee et al., 2023)	~ 25.0M	77.51±0.08	31.59±0.05	82.22±0.11	93.47±0.25	92.93±0.15	88.90±0.01
TabDDPM (Kotelnikov et al., 2023)	~ 11.7M	96.99±0.25	95.11±0.10	93.39±0.16	98.30±0.22	97.20±0.09	86.84±0.11
TABSYN (Zhang et al., 2023)	~ 10.7M	<u>98.46</u>±0.27	**97.95**±0.12	<u>97.93</u>±0.21	<u>98.94</u>±0.31	**97.76**±0.28	<u>98.56</u>±0.03
TABSYN (reproduced)	~ 10.7M	98.29±0.22	95.25 ±0.51	97.82±0.14	**99.16**±0.16	94.86±0.34	98.52±0.09
Our model	~ **64K**	**98.75**±0.09	<u>96.00</u>±1.23	**98.24**±0.13	98.85±0.42	<u>97.42</u>±0.11	**98.57**±0.16

Trend is a metric that captures pair-wise column correlation by computing Pearson correlation for numerical columns, contingency similarity for categorical columns, and contingency similarity between bucketed numerical values and categorical values.

Previous model size **10M~25M** 100-200X Reduction Ours: **64K**

Table 3. Performance on the **MLE** metric. Higher values in AUC and lower values in RMSE indicate better testing performance. Best performance in **bold**. Second best in underline.

Methods	#Parameters	Adult (AUC↑)	Default (AUC↑)	Shoppers (AUC↑)	Magic (AUC↑)	Beijing (RMSE↓)	News (RMSE↓)
GOGGLE (Liu et al., 2023)	~ 5.6M	.778±0.012	.584±0.005	.658±0.052	.654±0.024	1.090±0.025	.877±0.002
STaSy (Kim et al., 2022)	~ 10.3M	.906±0.001	.752±0.006	.914±0.005	.934±0.003	.656±0.014	.871±0.002
CoDi (Lee et al., 2023)	~ 25.0M	.871±0.006	.525±0.006	.865±0.006	.932±0.003	.818±0.021	1.21±0.005
TabDDPM (Kotelnikov et al., 2023)	~ 11.7M	.907±0.001	.758±0.004	.918±0.005	.935±0.003	.592±0.011	4.86±3.04
TABSYN (Zhang et al., 2023)	~ 10.7M	**.915** ±0.002	**.764**±0.004	<u>.920</u>±0.005	.938±0.002	<u>.582</u>±0.008	<u>.861</u>±0.027
TABSYN (reproduced)	~ 10.7M	.910±0.001	.755±0.004	.916±0.004	<u>.939</u>±0.003	.655±0.012	**.851**±0.024
Our model	~ **64K**	**.915**±0.001	**.764**±0.002	**.924**±0.003	**.941**±0.002	**.543**±0.012	.864±0.021

MLE is the testing accuracy of the classification or regression task on real data after training an XGBoost Classifier or an XGBoost Regressor on the synthetic tabular data.