ICML
International Conference On Machine Learning
40 Years

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis

Chaejeong Lee*, Jayoung Kim*, Noseong Park
Yonsei University
{chaejeong_lee, jayoung.kim, noseong}@yonsei.ac.kr

YONSEI UNIVERSITY

BigDyL
[bɪg di:l]
Big Data Analytics Laboratory
Yonsei University

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

ICML 40 Years International Conference On Machine Learning

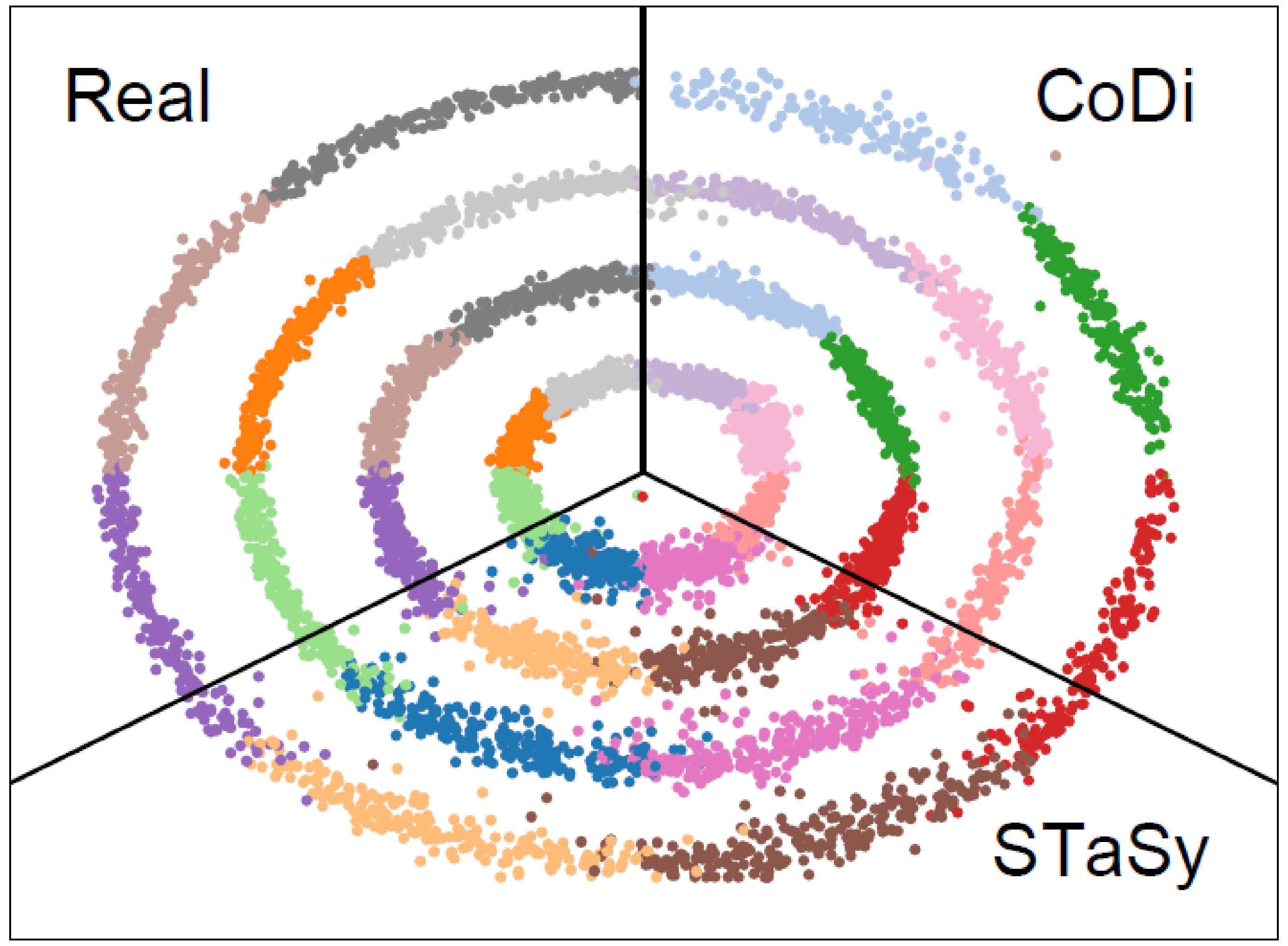
Motivation

Challenges of modeling tabular data

1) Tabular data consists of **mixed data types.**
2) **Pre/post-processing methods** impact the performance of tabular data synthesis.

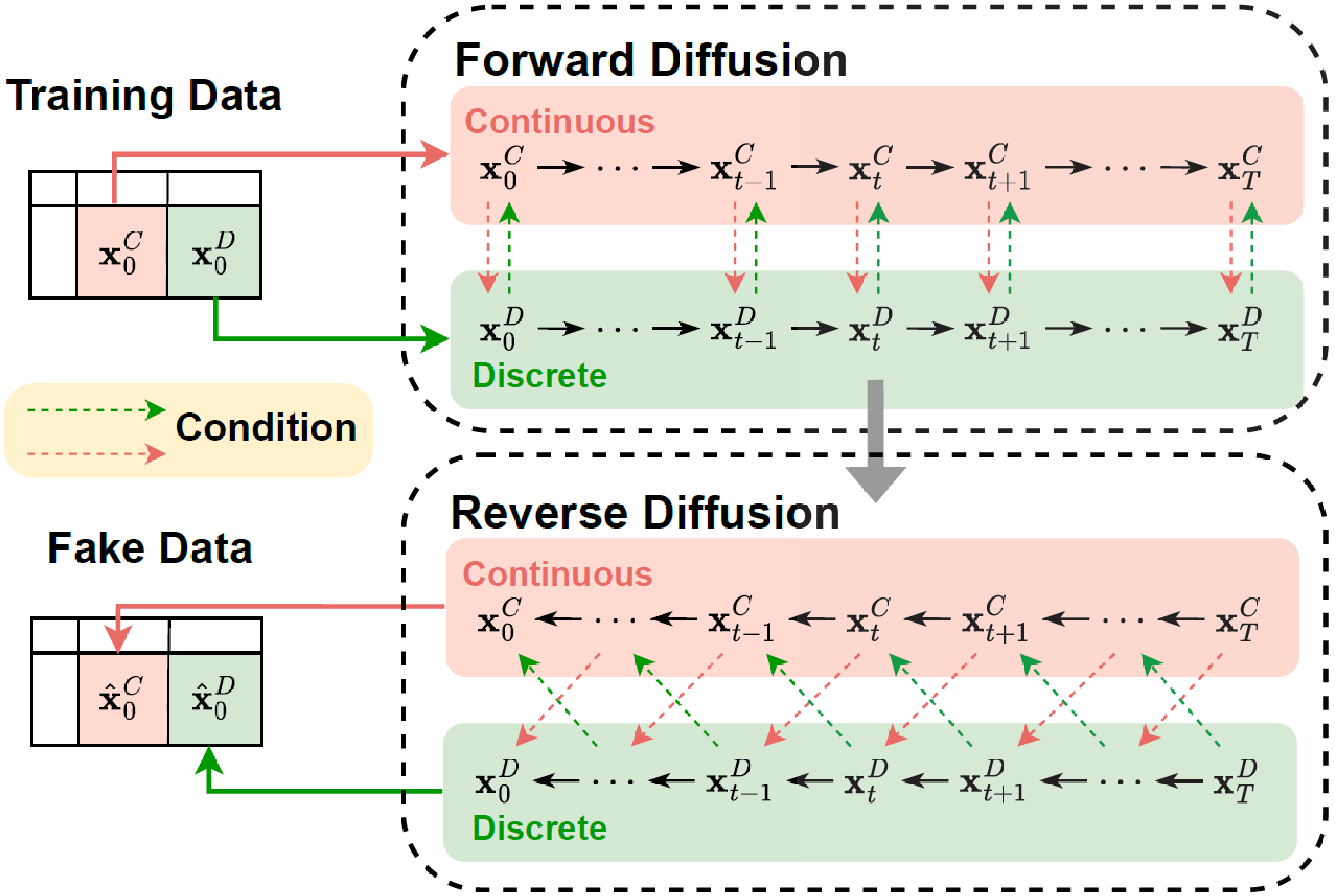
Continuous (X, Y) **Discrete** (Circle, Color)

X	Y	Circle	Color
0.8	1.6	Circle1	Blue
-0.5	2.7	Circle2	Gray



CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

Proposed Method



Co-evolving Conditional Diffusion Models

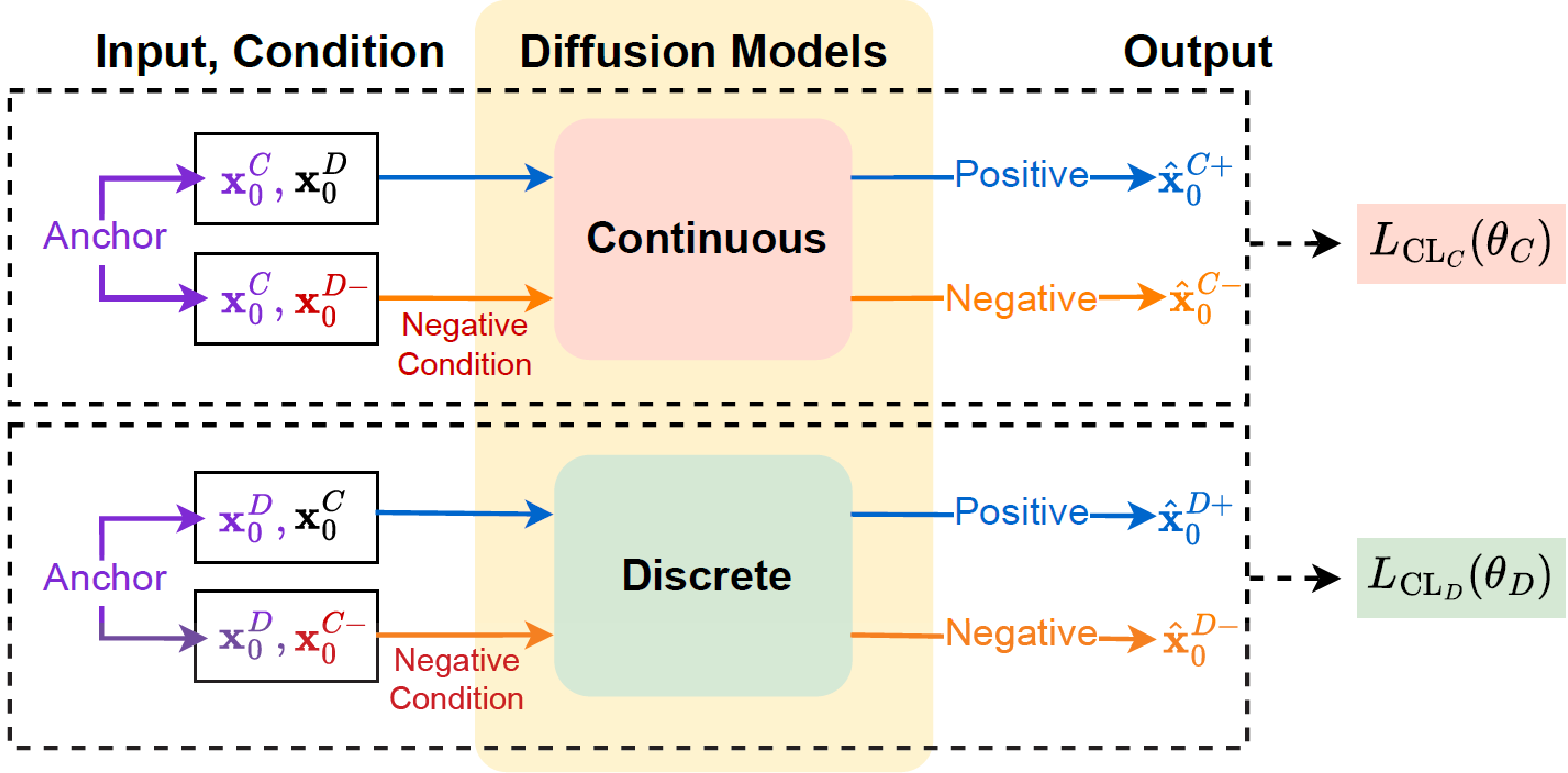
1) Forward Diffusion
   - The pair $(\mathbf{x}_0^C, \mathbf{x}_0^D)$ are simultaneously **perturbed in each space conditioned on each other.**

2) Reverse Diffusion
   - $p(\mathbf{x}_T^C) = \mathcal{N}(\mathbf{x}_T^C; \mathbf{0}, \mathbf{I}),\ p(\mathbf{x}_T^{D_i}) = \mathcal{C}(\mathbf{x}_T^{D_i}; 1/K_i)$
   - The noises are converted into fake samples **while being conditioned on the denoised samples at the previous time step.**

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

ICML
International Conference
On Machine Learning
40 Years

Proposed Method



Contrastive Learning

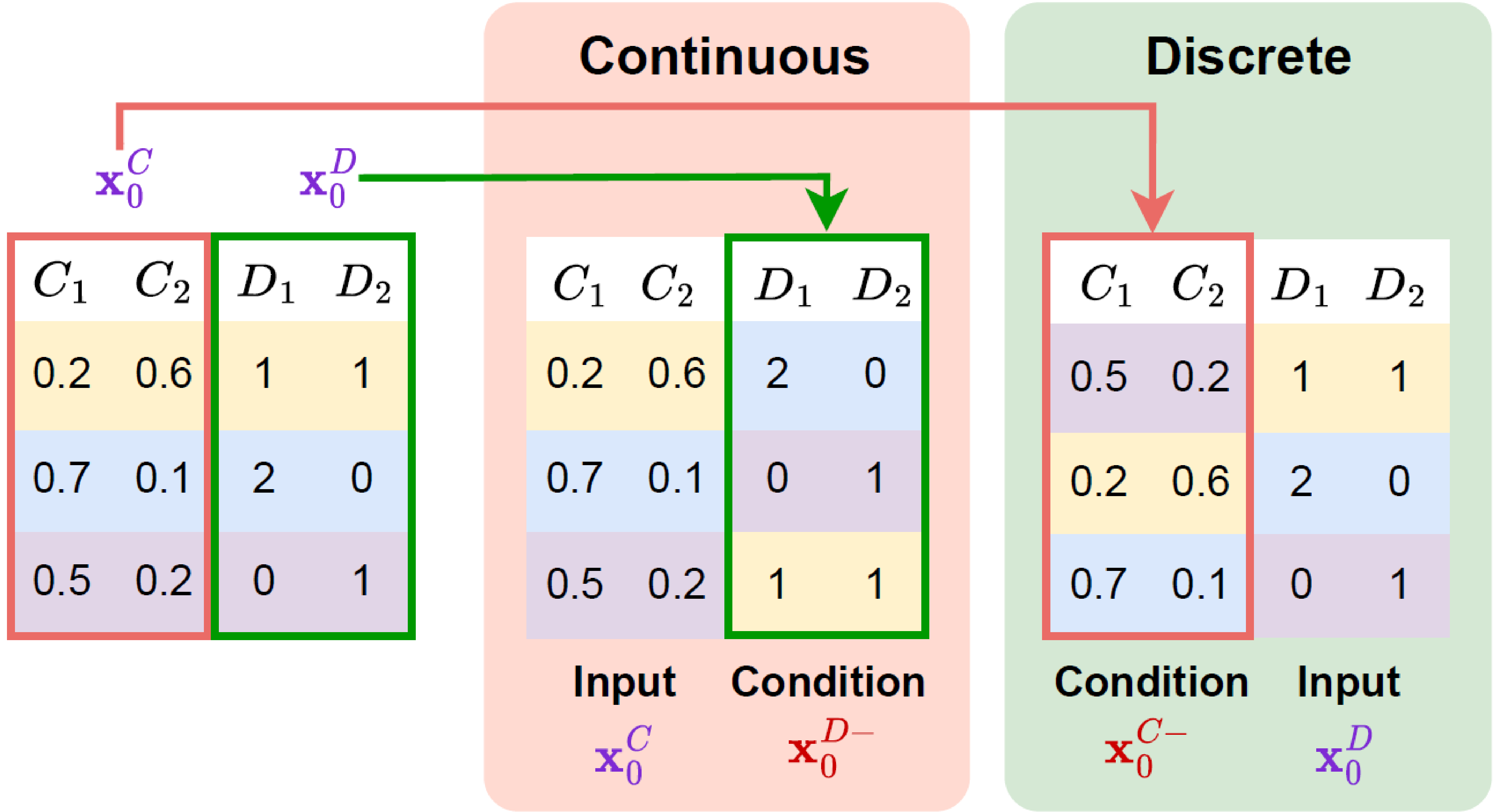
1) Triplet loss

- **Anchor**: a real sample $\mathbf{x}_0^C$
- **Positive sample**: a generated sample $\hat{\mathbf{x}}_0^{C+}$ conditioned on $\mathbf{x}_0^D$
- **Negative sample**: a generated sample $\hat{\mathbf{x}}_0^{C-}$ **with negative condition** $\mathbf{x}_0^{D-}$**.**

$$L_{\mathrm{CL}}(A, P, N) = \sum_{i=0}^{S} \left[\max \{ d(A_i, P_i) - d(A_i, N_i) + m, 0 \} \right]$$

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

Proposed Method



Contrastive Learning

2) How to define **negative conditions**

- The negative samples are generated **by corrupting the inter-correlation between the continuous and discrete variable sets.**

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

ICML
International Conference On Machine Learning
40 Years

Experiments

Experimental Results

1. Sampling quality

METHODS	BINARY		MULTI-CLASS		REGRESSION	
	BINARY F1	AUROC	MACRO F1	AUROC	R^2	RMSE
MEDGAN	0.1523	0.5464	0.1537	0.5015	-INF	INF
VEEGAN	0.2591	0.5520	0.1206	0.5082	-INF	INF
CTGAN	0.3432	0.6745	0.2355	0.5546	-INF	INF
TVAE	0.3188	0.6867	0.2361	0.5974	-INF	INF
TABLEGAN	0.4078	0.7480	0.2715	0.6072	-0.0704	1.0015
OCT-GAN	0.3814	0.7350	0.3314	0.6434	-0.0868	1.0210
RNODE	0.3208	0.6651	0.3692	0.7037	-0.3037	1.1270
STASY	0.4559	0.7961	0.6078	0.7997	-1.3200	1.2227
CODI	**0.4726**	**0.8106**	**0.6221**	**0.8026**	**0.4794**	**0.6477**

2. Diversity

METHODS	COVERAGE
MEDGAN	0.0155
VEEGAN	0.0019
CTGAN	0.3834
TVAE	0.3903
TABLEGAN	0.5759
OCT-GAN	0.2547
RNODE	0.3841
STASY	0.5771
CODI	**0.6931**

3. Time

METHODS	RUNTIME
MEDGAN	0.0200
VEEGAN	0.0169
CTGAN	0.1260
TVAE	**0.0140**
TABLEGAN	0.0224
OCT-GAN	0.6008
RNODE	103.1449
STASY	4.6417
CODI	0.5187

- Compared to other models, CoDi can sample **in reliable runtime with high quality and diversity**.

CoDi: Co-evolving Contrastive Diffusion Models for Mixed-type Tabular Synthesis, ICML 2023

ICML 40 Years International Conference On Machine Learning

Experiments

Contrastive Learning

- Ablation study on the efficacy of **contrastive learning.**

Datasets	CoDi w/o CL		CoDi	
	F1 (R^2)	Coverage	F1 (R^2)	Coverage
Bank	0.527±0.032	**0.699±0.003**	**0.566±0.014**	0.687±0.002
Heart	**0.886±0.043**	0.879±0.017	0.872±0.039	**0.949±0.012**
Seismic	0.210±0.064	**0.380±0.016**	**0.305±0.040**	0.359±0.005
Stroke	0.129±0.036	0.651±0.020	**0.147±0.016**	**0.919±0.008**
CMC	0.484±0.024	0.932±0.011	**0.503±0.008**	**0.934±0.015**
Customer	0.350±0.008	0.789±0.019	**0.352±0.015**	**0.833±0.021**
Faults	0.705±0.047	**0.272±0.016**	**0.715±0.046**	0.270±0.017
Obesity	0.912±0.038	**0.777±0.018**	**0.919±0.034**	0.742±0.015
Absent	(-0.026±0.036)	0.801±0.009	(**0.095±0.022**)	**0.843±0.023**
Drug	(0.748±0.074)	0.813±0.013	(**0.768±0.049**)	**0.827±0.046**
Insurance	(0.531±0.308)	0.218±0.028	(**0.575±0.398**)	**0.262±0.020**

ICML
International Conference On Machine Learning
40 Years

Thank you

chaejeong_lee@yonsei.ac.kr
https://github.com/ChaejeongLee/CoDi

YONSEI UNIVERSITY

BigDyL
[bɪg di:l]
Big Data Analytics Laboratory
Yonsei University