DDTR: Diffusion Denoising Trace Recovery

Maximilian Matyash[1], **Avigdor Gal**[1], **Arik Senderovich**[2]

[1]Technion–Israel Institute of Technology [2]York University

maximilian.m@campus.technion.ac.il, avigal@technion.ac.il, sariks@yorku.ca

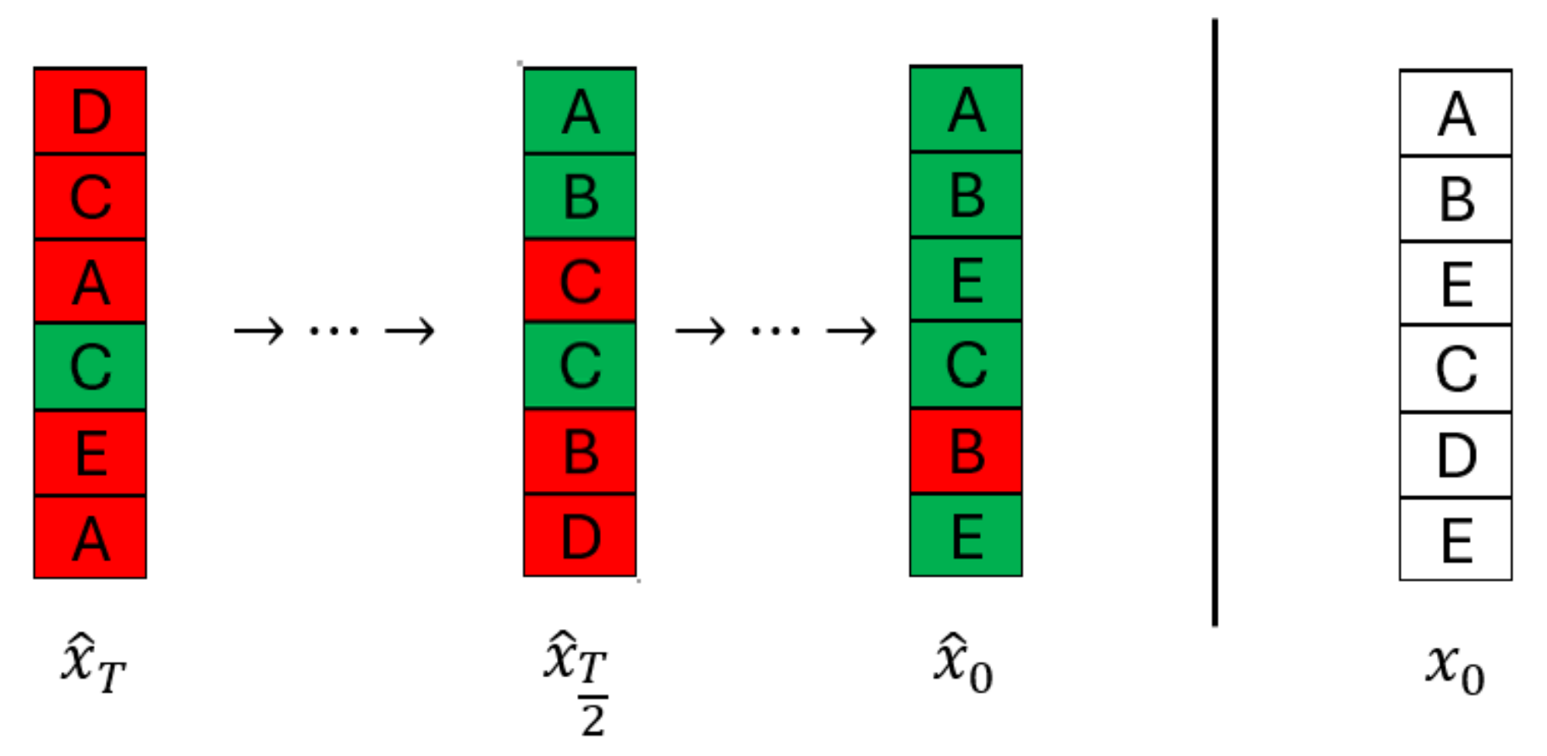
1. Introduction

The goal of process mining is to discover, analyze, and optimize real-world processes. To that end, process mining relies on a process log, a dataset of process executions, which contains a collection of traces where each trace is a sequence of categorical actions.

Process logs are assumed to record events in a deterministic fashion. However, modern means of recording processes have emerged, which result in stochastically-known process logs. In this type of logs, traces are presented as sequences of **probability distributions** instead of deterministic classes. Computer vision models trained to recognize human actions instead of clerks manually documenting their work are a good example to consider for real-life sources of stochastically-known logs.

The task of **trace recovery** centers around yielding the most accurate sequence of deterministic actions, with respect to what happened in reality, given a stochastically-known sequence. In most cases, the naïve approach of selecting the action with the highest probability can be improved on by considering the global context of the action in the whole sequence and the process model which governs the behavior of the system which produces the log.

In this work we present a state of the art method for trace recovery using a diffusion based approach. Our method is generative in nature as we train a diffusion model to sample solutions conditioned on an input stochastically-known sequence. We provide a relative accuracy improvement of 5%-25% over the previous state of the art methods on both real-world and synthetic scenarios.



2. Background

Guided Diffusion Models are a vast family of generative models which allow sampling from a target distribution by first creating a sequence of gradually noisy datapoints via Gaussian mixture. A **denoiser** is then trained to revert the noise, which allows to sample from the original distribution given random Gaussian noise as input.

Guidance is a technique that allows to control the generative process and sample specific regions of the learned distribution. Furthermore, guided diffusion can be used to solve inverse problems by directly sampling the solutions space instead of approximating the inverse function.

Process Discovery algorithms generate a graph model of an underlying process, typically a bipartite graph called a **Petri Net**. The process model is a world model of sorts that aims to explain the behavior of the underlying system which produced the event log.

3. Diffusion Denoising Trace Recovery (DDTR)

Our approach mirrors that of image deblurring, where the diffusion denoiser iteratively ”cleans” trace distributions, resulting in an accurate sequence of actions.

The forward and backward diffusion paths are defined respectively as:

$$\vec{x}_t = \sqrt{\alpha_t}\vec{x}_{t-1} + \sqrt{1-\alpha_t}\varepsilon;\ \varepsilon \sim \mathcal{N}(\vec{0}, \mathbf{I})$$

$$\begin{cases} \hat{x}_{t-1} = \frac{\sqrt{\bar{\alpha}_{t-1}}}{1-\bar{\alpha}_t}\hat{x}_0(\hat{x}_t, t, y; \theta) + \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\hat{x}_t + \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}(1-\alpha_t)z \\ z \sim \mathcal{N}(\vec{0}, \mathbf{I}) \text{ if } t > 0 \text{ else } z = \vec{0} \end{cases}$$

The guided denoiser $\hat{x}_0(\hat{x}_t, t, y; \theta)$ generates the action trace $\hat{x}_0$ given the stochastically known trace y as guidance and optionally a process model embedding as additional guidance.
We use a combined loss to optimize both trace and process model reconstruction.

$$\gamma\text{CE}(x_0, \hat{x}_0) + (1-\gamma)\text{CE}(F, \hat{F})$$

3.1 Model Architectures

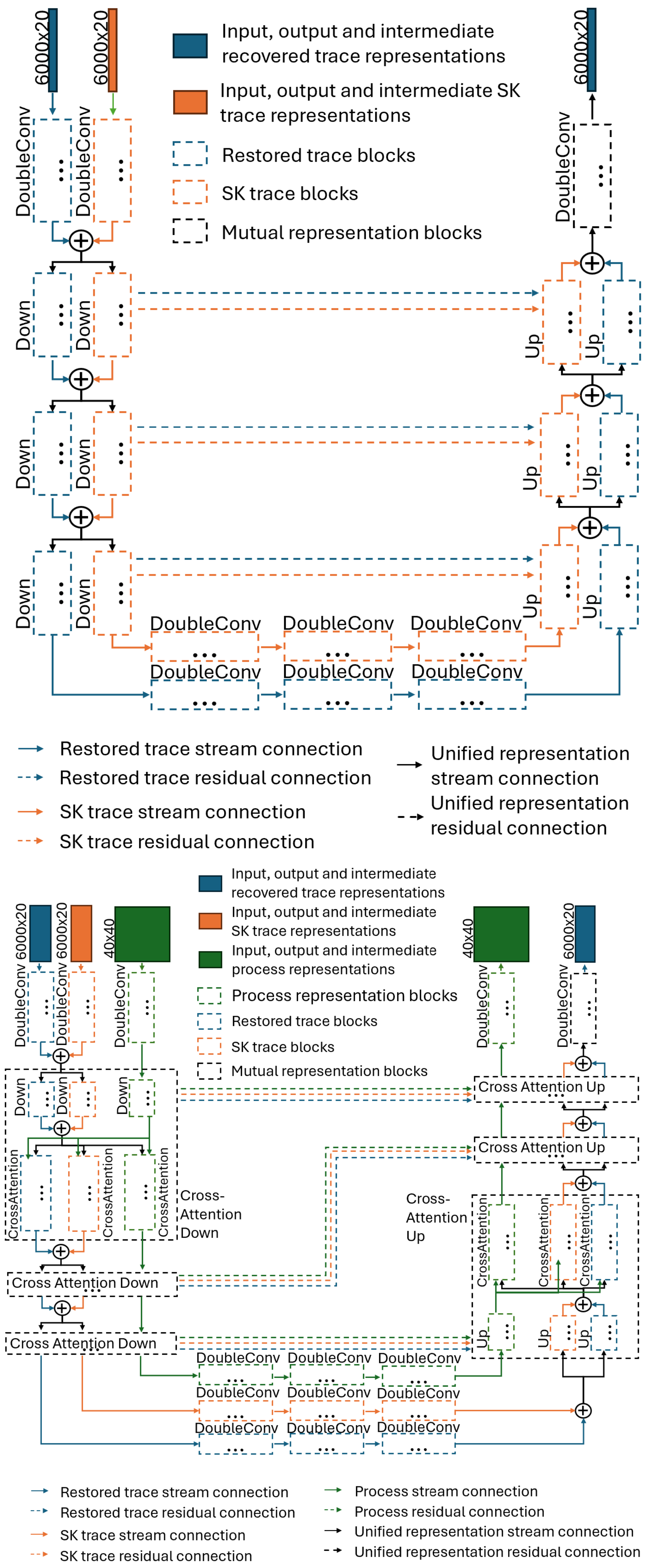

6000x20
6000x20
6000x20
DoubleConv
DoubleConv
DoubleConv
Input, output and intermediate recovered trace representations
Input, output and intermediate SK trace representations
Restored trace blocks
SK trace blocks
Mutual representation blocks
Down
Up
DoubleConv
Restored trace stream connection
Restored trace residual connection
SK trace stream connection
SK trace residual connection
Unified representation stream connection
Unified representation residual connection
Input, output and intermediate recovered trace representations
Input, output and intermediate SK trace representations
Input, output and intermediate process representations
Process representation blocks
Restored trace blocks
SK trace blocks
Mutual representation blocks
40x40
CrossAttention
Cross-Attention Down
Cross-Attention Up
Cross Attention Down
Cross Attention Up
Process stream connection
Process residual connection

Figure 1: Top: U-net denoiser for trace recovery that does not integrate process information. Bottom: U-net denoiser that integrates and reconstructs the process adjacency matrix given an internal embedding.

4. Results

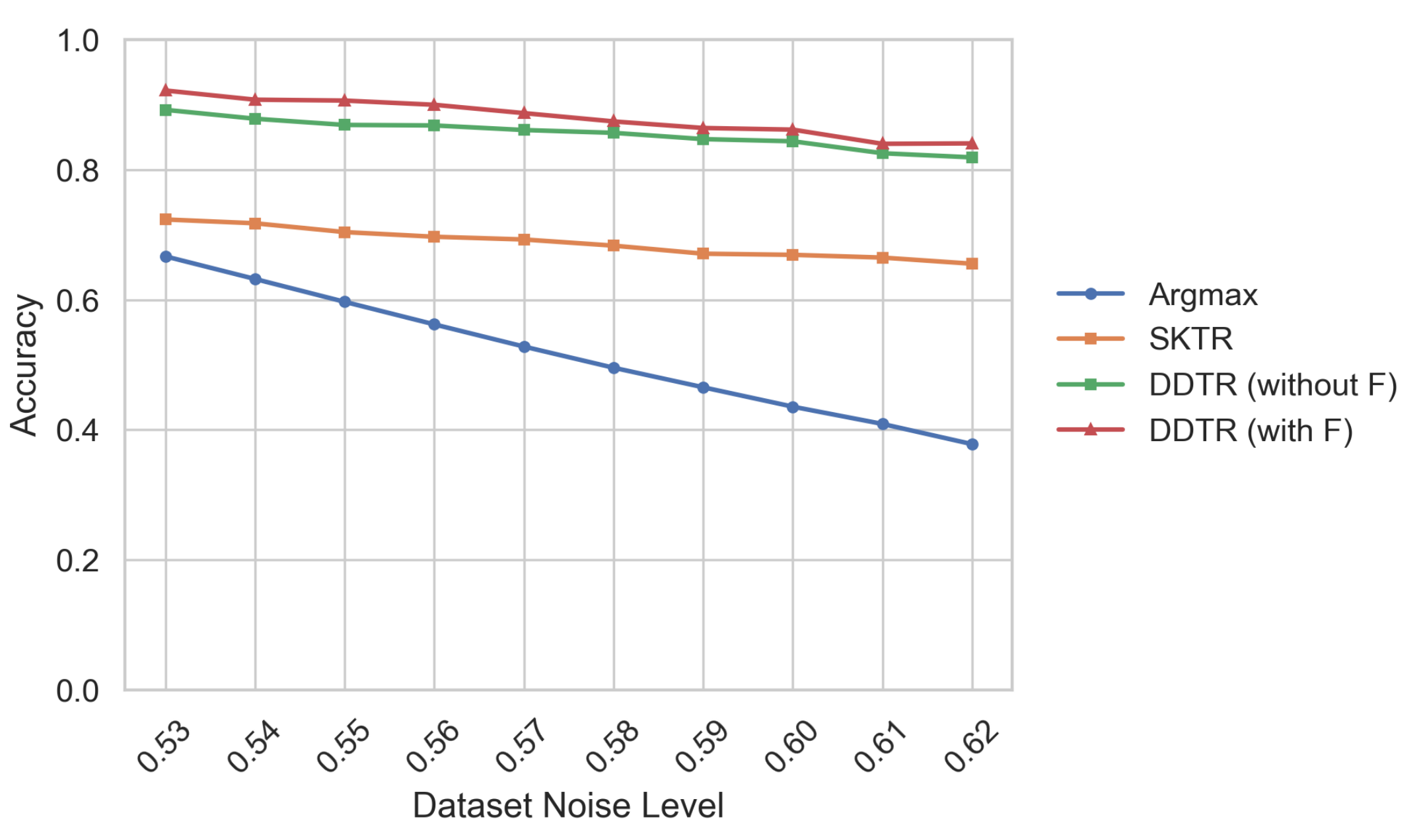
We evaluate trace recovery methods on both real-world and synthetic datasets. The real-world datasets consist of food preparation datasets: 50-Salads, Breakfast and GTEA, while synthetic datasets consisted of process mining challenge datasets: BPI12 and BPI19.

4.1 Overall performance

Dataset	Method	Accuracy	Precision	Recall
50-Salads	Argmax	0.774	0.744	0.731
	SKTR	0.89	0.83	0.81
	DDTR	**0.937**	**0.907**	**0.9**
GTEA	Argmax	0.751	0.751	0.755
	SKTR	0.79	0.819	0.782
	DDTR	**0.988**	**0.974**	**0.969**
Breakfast	Argmax	0.737	0.604	0.61
	SKTR	0.81	0.724	0.744
	DDTR	**0.931**	**0.876**	**0.864**
BPI 2012	Argmax	0.778	0.639	0.651
	SKTR	0.935	0.855	0.858
	DDTR	**0.997**	**0.994**	**0.994**
BPI 2019	Argmax	0.784	0.685	0.69
	SKTR	0.866	0.804	0.818
	DDTR	**0.982**	**0.973**	**0.976**

We use macro accuracy, precision and recall as performance metrics. We observe that DDTR achieves the highest performance across all datasets.

4.2 Robustness under increasing uncertainty



In addition to performance, we test the robustness of recovery methods by artificially increasing the entropy of the stochastically-known traces. We compare both variants of DDTR (F denotes the process model's adjacency matrix) and observe that our method retains its lead in this scenario as well.

5. Conclusion and Future Directions

This work introduced DDTR, a novel method for recovering deterministically known traces from stochastically known process logs using guided diffusion denoising models.

By reframing trace recovery as an inverse problem and leveraging the flexibility of Diffusion Denoising Probabilistic Models (DDPMs), we present a deep learning approach capable of reconstructing traces with high accuracy and robustness. The method generalizes DDPM guidance to include both probabilistic trace data and process model structure, grounding the recovery in both local uncertainty and global constraints.

DDTR outperforms prior methods across real-world and synthetic datasets, improving accuracy by up to 25%. It remains robust under increasing noise and handles traces of arbitrary length without reliance on alignment-based search. Incorporating a latent representation of the process model enhances recovery in high uncertainty scenarios, especially when event probabilities become unreliable.

Future work may extend DDTR to incorporate timestamps, resources, or natural language annotations. Another direction involves supporting concurrent or hierarchical process structures beyond sequential traces (*e.g.*, process trees and event structures). Finally, integrating DDTR into full process mining pipelines would enable end-to-end handling of uncertain event data in practical applications.

Contact Us!

(a) Full Paper

(b) Maximilian Matyash

(c) Avigdor Gal

(d) Arik Senderovich