

# Novel Tabular Generative Metrics Scaling with High-Cardinality Categorical Features

Assignee Research

June 9, 2026

## Abstract

This report synthesises findings from 4 peer-reviewed papers addressing the following research question: How do novel tabular generative metrics scale computationally with high-cardinality categorical features compared to baseline metrics on large-scale datasets. 9 claims were extracted from source literature; 1 was independently verified against retrieved documents. An automated multi-reviewer quality assessment produced a score of 4.2/10. This report is a machine-generated literature synthesis and does not constitute original research.

## 1 Introduction

This paper examines: ciDATGAN: Conditional Inputs for Tabular GANs. Research question: How do novel tabular generative metrics scale computationally with high-cardinality categorical features compared to baseline metrics on large-scale datasets?.

## 2 Methodology

Systematic literature search across multiple databases yielded 4 papers. Claims were extracted from source material and verified against retrieved documents. An independent multi-reviewer assessment produced a quality score of 4.2/10.

## 3 Results

4 papers retrieved. 9 claims extracted; 1 independently verified. Quality review score: 4.2/10.

## 4 Limitations

This report is a machine-generated literature synthesis and does not constitute original research. Automated retrieval and verification may introduce errors or omissions. Review scores reflect automated assessment, not human peer review. Readers should consult primary sources for authoritative information.

## 5 Extracted Claims

Claim	Verified	Confidence
ciDATGAN generates data that are representative and correspond to the original data.	×	0.05
ciDATGAN can correct bias using unbiased conditional inputs.	×	0.08
ciDATGAN can be used to generate large synthetic populations by learning from a feeder dataset and efficiently generating	×	0.14
ciDATGAN and DATGAN were trained on the LPMC dataset five times, generating five different synthetic datasets for each m	×	0.10
The LPMC dataset consists of 18 variables and 16,904 trips.	×	0.02
The conditional inputs for ciDATGAN are the variables age, female, and hh_region.	×	0.07
The SRMSE metric was used to assess the results, with values reported as 0.103, 0.095, 0.1, 0.08, 0.28, 0.23, 0.24, 0.25	×	0.02
ciDATGAN takes inspiration from image completion for conditional inputs, unlike latent conditionality.	✓	0.18
ciDATGAN can be used for removing bias in a dataset using unbiased variables as conditional inputs or combining informat	×	0.12

## References

- <http://arxiv.org/abs/2107.02521v3>
- <http://arxiv.org/abs/2409.01628v1>
- <http://arxiv.org/abs/2210.02404v1>