Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

# КВАЛІФІКАЦІЙНА РОБОТА
# Пояснювальна записка

рівень вищої освіти другий (магістерський)

Використання дифузійних моделей для відновлення табличних даних

(тема)

Виконав:
здобувач другого року навчання,
групи СШІм-24-2
Олег ГОНЧАРЕНКО
(власне ім'я, прізвище)
Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва освітньої програми)
Керівник проф. Лариса ЧАЛА
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Лариса ЧАЛА
(підпис) (власне ім'я, прізвище)

2026 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____» _____ 20 ___ р.

# ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Гончаренку Олегу Віталійовичу
(прізвище, ім'я, по батькові)

1. Тема роботи Використання дифузійних моделей для відновлення табличних даних

затверджена наказом університету від 6 квітня 20 26 р. № 269Ст

2. Термін подання здобувачем роботи до екзаменаційної комісії 22 травня 20 26 р.

3. Вихідні дані до роботи Сучасні науково-технічні публікації (статті, матеріали конференцій) за темою дифузійних моделей та обробки відсутніх даних, відкриті дані Інтернет-джерел, технічна документація мови Python та фреймворків для побудови нейронних мереж, реальні набори табличних даних Adult Census Income, Beijung PM2.5, Default of Credit Card Clients

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Еволюція парадигм імпутації
3) Розробка методу LIR
4) Проведення експериментів та аналіз результатів

## КАЛЕНДАРНИЙ ПЛАН

| № | Назва етапів роботи | Строк / термін виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Отримання завдання на кваліфікаційну роботу | 06.04.2026 | виконано |
| 2 | Аналіз предметної галузі та існуючих методів | 07.04.2026 – 12.04.2026 | виконано |
| | імпутації | | |
| 3 | Розробка математичної бази та загальної | 13.04.2026 – 18.04.2026 | виконано |
| | архітектури LIR | | |
| 4 | Програмна реалізація методу та підготовка | 19.04.2026 – 24.04.2026 | виконано |
| | експериментальних наборів даних | | |
| 5 | Проведення експериментів та збір метрик | 24.04.2026 – 30.04.2026 | виконано |
| 6 | Аналіз результатів та формулювання висновків | 01.05.2026 – 05.05.2026 | виконано |
| 7 | Оформлювання пояснювальної записки | 06.05.2026 – 10.05.2026 | виконано |
| 8 | Захист перед ЕК | 22.05.2026 | виконано |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |
| | | | |

Дата видачі завдання 6 квітня 20 26 р.

Здобувач 
(підпис)

Керівник роботи _____ проф. Лариса ЧАЛА
(підпис) (посада, власне ім’я, прізвище)

# РЕФЕРАТ

Пояснювальна записка: 122 с., 24 рис., 5 табл., 1 дод., 52 джерела.

ГЕНЕРАТИВНІ МОДЕЛІ, ДИФУЗІЙНА МОДЕЛЬ, ІМПУТАЦІЯ ДАНИХ, ЛАТЕНТНИЙ ПРОСТІР, МАШИННЕ НАВЧАННЯ, ПРОПУЩЕНІ ЗНАЧЕННЯ, ТАБЛИЧНІ ДАНІ.

Об’єкт дослідження – задача відновлення пропущених значень у гетерогенних табличних наборах даних.

Предмет дослідження – використання дифузійних моделей та гібридних генеративних архітектур для імпутації табличних даних.

Мета роботи – розробка гібридного генеративного фреймворку LIR для статистично коректного відновлення пропусків із меншою затримкою виведення порівняно з базовими дифузійними моделями.

Методи дослідження – теоретичний аналіз, математичне моделювання з використанням стохастичних диференціальних рівнянь, емпіричний та статистичний аналіз розробленого методу на реальних датасетах.

Роботу присвячено вирішенню компромісу між дистрибутивною точністю та обчислювальною ефективністю при імпутації табличних даних. Розроблено фреймворк LIR, який поєднує детермінований «теплий старт» на основі ансамблю проекторів зі стохастичним вдосконаленням у латентному просторі VAE за допомогою SDEdit. Математично обґрунтовано межу збурення для низькорозмірних просторів та впроваджено адаптивне планування кроку дифузії.

Експерименти на реальних наборах даних підтвердили, що фреймворк забезпечує точність відтворення спільного розподілу, зіставну з повнотраєкторними моделями, водночас скорочуючи кількість нейронних обчислень на 30%.

# ABSTRACT

Master’s thesis contains: 122 pp., 24 fig., 5 tabl., 1 ann., 52 references.

DATA IMPUTATION, DIFFUSION MODEL, GENERATIVE MODELS, LATENT SPACE, MACHINE LEARNING, MISSING VALUES, TABULAR DATA.

Object of research – the problem of missing data imputation in heterogeneous tabular datasets.

Subject of research – the application of diffusion models and hybrid generative architectures for tabular data imputation.

Purpose of the work – the development of a hybrid generative framework LIR for statistically correct missing data imputation with lower inference latency compared to baseline diffusion models.

Research methods – theoretical analysis, mathematical modeling using stochastic differential equations, empirical and statistical analysis of the developed method on real-world datasets.

The work is dedicated to resolving the trade-off between distributional accuracy and computational efficiency in tabular data imputation. The LIR framework was developed, which combines a deterministic «warm start» based on an ensemble of projectors with stochastic refinement in the VAE latent space using SDEdit. The perturbation bound for low-dimensional spaces was mathematically justified, and the adaptive diffusion step scheduling was introduced.

Experiments on real-world datasets confirmed that the framework provides joint distribution recovery accuracy comparable to full-trajectory models, while simultaneously reducing the number of neural function evaluations by 30%.

**ЗМІСТ**

Перелік умовних позначень, символів, одиниць, скорочень і термінів ..... 8
Вступ ..... 14
1 Аналіз предметної галузі та постановка задачі ..... 16
1.1 Контекст та мотивація дослідження ..... 16
1.1.1 Ціна неповноти даних ..... 17
1.1.2 Тристороннє протиріччя ..... 19
1.2 Постановка задачі ..... 23
2 Еволюція парадигм імпутації ..... 25
2.1 Класичні статистичні методи ..... 25
2.2 Дискримінативне машинне навчання ..... 29
2.2.1 MICE та MissForest ..... 29
2.2.2 Патологія усереднення за модами ..... 31
2.3 Ранні глибокі генеративні моделі: VAE та GAN ..... 33
2.4 Дифузійні ймовірнісні моделі: сучасний стан розвитку ..... 35
2.4.1 Знакові адаптації для табличних даних ..... 36
2.4.2 Фронтир прискорення ..... 39
2.4.3 Формалізація обчислювального вузького місця ..... 40
3 Розробка методу LIR ..... 43
3.1 Загальна формалізація принципу роботи методу ..... 43
3.1.1 Формалізація задачі генеративної імпутації у латентному просторі ..... 46
3.1.2 Гетерогенність та розподіли з «важкими хвостами» ..... 49
3.2 Формулювання VP-SDE ..... 52
3.2.1 Прямий процес та маргінальний розподіл у замкненій формі ..... 52
3.2.2 Зворотний процес та узгодження оцінок ..... 56
3.3 Етап I: VAE змішаного типу для латентної проекції ..... 57
3.3.1 Процес попередньої обробки: нормалізація, конкатенація, та вирішення «проблеми нуля» ..... 58

3.3.2 Архітектура на основі табличного трансформера ..... 59
3.4 SDEdit та невідповідність многовидів ..... 63
3.4.1 Визначення латентних носіїв ..... 63
3.4.2 Механізм SDEdit: пряме стохастичне збурення, стохастична ініціалізація та подолання «патології усереднення за модами» ..... 65
3.5 Етап II: ансамбль проекторів для ініціалізації «теплого старту» ..... 67
3.6 Межа збурення з корекцією концентрації ..... 71
3.7 Етап III: збурення SDEdit та усічена зворотна дифузія ..... 74
3.8 Експериментальний дизайн та оцінювання ..... 76
3.8.1 Обрані набори даних ..... 76
3.8.2 Експериментальний протокол ..... 79
4 Проведення експериментів та аналіз результатів ..... 83
4.1 Аналіз загальної ефективності імпутації ..... 83
4.1.1 Набір даних Adult Census ..... 83
4.1.2 Набір даних Beijing PM2.5 ..... 87
4.1.3 Набір даних Default Credit ..... 90
4.2 Аналіз точності відтворення розподілу ..... 93
4.2.1 Збереження маргінальних розподілів ..... 93
4.2.2 Аналіз розподілів категоріальних ознак ..... 96
4.3 Аналіз стійкості та чутливості ..... 101
4.3.1 Ефективність залежно від рівня пропусків ..... 101
4.3.2 Аналіз поведінки адаптивного планування ..... 103
4.3.3 Аналіз стійкості до гіперпараметрів ..... 105
Висновки ..... 110
Перелік джерел посилання ..... 116
Додаток А Відомість кваліфікаційної роботи ..... 122

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ЕМК – електронні медичні карти;

ЗДР – звичайні диференціальні рівняння;

ШІ – штучний інтелект;

Епістемічна невизначеність – міра невпевненості моделі (або ансамблю моделей) у власних прогнозах, зумовлена браком знань або високою апостеріорною неоднозначністю даних;

Квантильний перетворювач – Quantile Transformer – нелінійне перетворення, що стискає екстремальні викиди, відображаючи емпіричну функцію розподілу ознаки у стандартний нормальний розподіл;

Патологія усереднення за модами – структурний недолік моделей, що мінімізують середньоквадратичну похибку, за якого оцінка мультимодального розподілу зводиться до єдиного середнього значення, що часто потрапляє у фізично неможливу область низької щільності між модами;

Пряме кодування – One-hot encoding – представлення категоріальних змінних у вигляді розріджених бінарних векторів (де лише один елемент дорівнює 1, а інші 0);

«Теплий старт» – Warm start – метод ініціалізації зворотного процесу дифузії з використанням детермінованої попередньої оцінки даних, збуреної шумом, замість повного ізотропного шуму;

AdaGN – Adaptive Group Normalization – адаптивна групова нормалізація;

AUPRC – Area Under the Precision-Recall Curve – площа під кривою точності та повноти (метрика оцінки класифікації за сильного дисбалансу класів);

BGMM – Bayesian Gaussian Mixture Model – байєсівська модель сумішей Гауса;

CDF – Cumulative Distribution Function – кумулятивна (емпірична) функція розподілу;

CE – Cross-Entropy – категоріальна перехресна ентропія;

CMD – Correlation Matrix Distance – відстань між матрицями кореляції;

CoDi – Co-evolving contrastive diffusion – ко-еволюційна контрастивна дифузійна модель;

CSDI – Conditional Score-based Diffusion Models for Imputation – умовна оціночна дифузія для імпутації;

CTGAN – Conditional Tabular Generative Adversarial Network – умовна таблична генеративно-змагальна мережа;

DDIM – Denoising Diffusion Implicit Models – неявні дифузійні моделі усунення шуму;

DDPM – Denoising Diffusion Probabilistic Models – дифузійні ймовірнісні моделі усунення шуму;

DGM – Deep Generative Models – глибокі генеративні моделі;

ELBO – Evidence Lower Bound – варіаційна нижня межа;

EM – Expectation-Maximization – ітераційний алгоритм максимізації очікування (правдоподібності);

FCS – Fully Conditional Specification – повна специфікація умовних розподілів;

FD – Fréchet Distance – відстань Фреше;

FFN – Feed-Forward Network – повнозв’язна нейронна мережа;

GAIN – Generative Adversarial Imputation Nets – метод імпутації даних на основі генеративно-змагальних мереж;

GAN – Generative Adversarial Network – генеративно-змагальна мережа;

HI-VAE – Heterogeneous Incomplete VAE – варіаційний автоенкодер для гетерогенних неповних даних;

KL – Kullback-Leibler Divergence – дивергенція Кульбака-Лейблера;

KNN – $k$-Nearest Neighbors – метод $k$-найближчих сусідів;

MAR – Missing at Random – механізм випадкових пропусків даних;

MCAR – Missing Completely at Random – механізм абсолютно випадкових пропусків даних;

MI – Multiple Imputation – множинна імпутація;

MICE – Multiple Imputation by Chained Equations – багатовимірна імпутація за допомогою ланцюгових рівнянь;

MissForest – непараметричний ансамблевий алгоритм імпутації на основі випадкових лісів;

MIWAE – Missing Data Importance-Weighted Autoencoder – автоенкодер зважений за важливістю для неповних даних;

MLP – Multilayer Perceptron – багатошаровий перцептрон;

MNAR – Missing Not at Random – механізм невипадкових пропусків даних;

MSE – Mean Squared Error – середньоквадратична помилка;

NFE – Number of Function Evaluations – кількість обчислень нейронних функцій (метрика обчислювальної складності);

OOB – Out-of-Bag – похибка «поза пакетом» (в ансамблевих деревах рішень);

PMM – Predictive Mean Matching – прогнозне зіставлення середніх;

ResMLP – Residual Multilayer Perceptron – багатошаровий перцептрон із залишковими зв’язками;

RMSE – Root Mean Square Error – середньоквадратична похибка;

SDE – Stochastic Differential Equation – стохастичне диференціальне рівняння;

SDEdit – Stochastic Differential Editing – механізм стохастичного диференціального редагування;

SNR – Signal-to-Noise Ratio – відношення сигнал/шум;

TabDDPM – дифузійна ймовірнісна модель усунення шуму, адаптована спеціально для табличних даних;

TITOR – Train-In-Test-Out-Real – протокол оцінювання корисності імпутації для низхідних завдань машинного навчання;

VAE – Variational Autoencoder – варіаційний автоенкодер;

VP-SDE – Variance Preserving SDE – стохастичне диференціальне рівняння зі збереженням дисперсії;

$X$ – матриця гетерогенних табличних даних;

$X_{obs}$ $(x_{obs})$ – підмножина (або вектор) спостережуваних наявних значень;

$X_{miss}$ $(x_{miss})$ – підмножина (або вектор) відсутніх (пропущених) значень;

$X_{num}$ $(x_{num})$ – підмножина (або вектор) неперервних (числових) ознак;

$X_{cat}$ $(x_{cat})$ – підмножина (або вектор) категоріальних ознак;

$x_{in}$ – маскований конкатенований вхідний вектор даних;

$M$ – бінарна матриця (маска) спостереження;

$m_{i,j}$ – елемент маски (1 – спостережуване, 0 – пропущене значення);

$N$ – загальна кількість прикладів (рядків) у наборі даних;

$N_{effective}$ – ефективний розмір вибірки (після фільтрації);

$N_{num}$, $N_{cat}$ – кількість неперервних та категоріальних ознак відповідно;

$D$ – загальна кількість ознак (вимірів) у наборі даних;

$K$ – кількість класів дискретної категоріальної ознаки / кількість проекторів в ансамблі;

$z$ – латентне представлення (вектор) даних у просторі VAE;

$z_0$ – початковий (чистий) стан латентного вектора;

$z_t$ – латентний стан на часовому кроці $t$ процесу дифузії;

$z_{GT}$ – істинний латентний код (Ground Truth);

$z_{init}$ $(\hat{z})$ – детермінована оцінка ініціалізації «теплого старту» від ансамблю;

$z_{guide}$ – середнє значення прогнозів ансамблю проекторів;

$R^{d_z}$ – $d_z$-вимірний простір дійсних чисел;

$d_z$ – розмірність латентного простору;

$d_{model}$ – розмірність ембедингу (в архітектурі трансформера);

$M_{data}$ – істинний (чистий) многовид даних;

$M_{mean}$ – середній многовид (образ відображення умовного математичного сподівання);

$M_{noisy}(t)$ – зашумлений многовид на часовому кроці $t$;

$T$ – загальна (кінцева) кількість часових кроків у зворотному процесі дифузії;

$t$ – поточний часовий крок дифузії;

$t_{start}$ – адаптивно розрахований початковий часовий крок дифузії для збурення;

$dt$ – диференціальний (нескінченно малий) крок часу;

$dw$ – приріст стандартного вінерівського процесу (білий шум);

$\beta_t$ – гіперпараметр (функція) розкладу дисперсії на кроці $t$;

$\alpha_t$ – кумулятивний добуток коефіцієнтів масштабування шуму;

$\epsilon$ – вектор ізотропного гаусового шуму;

$\epsilon_\theta(z_t, t)$ – навчена нейронна мережа-денойзер (апроксиматор шуму);

$\nabla_{z_t} \log p_t\,(z_t)$ – score function – функція оцінки;

$Q_t$ – матриця ймовірностей переходів дискретної дифузії;

$p(x)$ – безумовний (спільний) розподіл даних;

$p(X_{miss}|X_{obs})$ – умовний апостеріорний розподіл відсутніх значень за умови наявних;

$q(z_t|z_{t-1})$ – ядро переходу прямого процесу дифузії;

$E[\dots]$ – оператор математичного сподівання;

$\sigma^2$– дисперсія;

$\sigma^2_{ens}$ – дисперсія прогнозів ансамблю проекторів;

$\mu$ – вектор середніх значень розподілу;

$\theta, \phi, \psi$ – параметри оптимізації нейронних мереж (денойзера, енкодера $\mathcal{E}_\varphi$, декодера $\mathcal{D}_\psi$);

$L_2$ – евклідова норма (або середньоквадратична функція втрат);

$\mathcal{L}_{MSE}, \mathcal{L}_{CE}, \mathcal{L}_{VAE}, \mathcal{L}_{proj}$ – функції втрат (реконструкції, крос-ентропії, VAE, проектора відповідно);

$\gamma_{conc}$ – коефіцієнт корекції концентрації;

$e$ – помилка оцінки детермінованого проектора;

$U(x_{in})$ – скалярна епістемічна невизначеність ансамблю для зразка $x_{in}$;

$g(\cdot)$ – калібрувальна функція для адаптивного планування $t_{start}$;

$\delta$ – рівень довіри (ймовірність того, що енергія шуму не подолає похибку ініціалізації);

$p$-value ($p$) – досягнутий рівень статистичної значущості;

$\alpha$ – заданий поріг статистичної значущості (зазвичай 0.05);

$d$ – Cohen's d – розмір ефекту за Коеном;

$m$ – кількість створених наборів даних під час множинної імпутації;

$U, B, T$ – внутрішньовибіркова дисперсія, міжвибіркова дисперсія та загальна дисперсія оцінки за правилами Рубіна;

$\Phi^{-1}$ – пробіт-функція (обернена гаусова CDF);

$\Psi(\cdot)$ – функція квантильного перетворення;

$\chi^2$ – розподіл хі-квадрат;

$W_1, W_2$ – відстань Вассерштейна першого та другого порядку відповідно;

$R^2$ – коефіцієнт детермінації;

$\odot$ – добуток Адамара (оператор поелементного множення);

$\oplus$ – оператор конкатенації векторів.

# ВСТУП

Сучасний етап розвитку аналітичних систем характеризується стрімким зростанням обсягів гетерогенних табличних даних. У таких критично важливих сферах, як медицина, фінанси та соціологія, наявність пропущених значень є однією з головних перешкод для побудови надійних систем машинного навчання. Наразі в індустрії домінують традиційні підходи до імпутації, що спираються на класичні дискримінативні алгоритми (наприклад, випадкові ліси або градієнтний бустинг).

Оскільки ці методи зазвичай оптимізуються для мінімізації середньоквадратичної похибки, їх використання неминуче призводить до згладжування складних мультимодальних розподілів та генерації статистично нереалістичних значень. З іншого боку, останнім часом з’явилися новітні генеративні архітектури – дифузійні моделі, які здатні з високою точністю відтворювати складні ймовірнісні характеристики даних. Однак їх практичне застосування сьогодні суттєво обмежується проблемою високої затримки виведення через необхідність проходження тривалого марковського процесу ітеративного усунення шуму.

З огляду на зазначені обмеження існуючих методів, існує гостра наукова та практична необхідність подолання компромісу між високою швидкістю роботи дискримінативних алгоритмів та неперевершеною статистичною достовірністю дифузійних моделей. Актуальність цієї кваліфікаційної роботи зумовлена дедалі більшою потребою сучасних систем у швидкому та одночасно дистрибутивно точному відновленні даних. Підставою для виконання роботи є відсутність у сучасній науковій літературі ефективних гібридних рішень, які б дозволяли повною мірою використовувати потужність дифузійного моделювання для табличних даних без критичної втрати обчислювальної ефективності.

Метою роботи є створення та комплексне дослідження нового гібридного генеративного підходу – Латентного ітераційного

вдосконалення (Latent Iterative Refinement, LIR). Цей фреймворк покликаний забезпечити швидке та статистично коректне відновлення пропущених значень у табличних наборах даних, органічно поєднуючи переваги швидкодії детермінованих проекторів із точністю дифузійного моделювання у латентному просторі.

Можливі сфери застосування розробленого підходу охоплюють високонавантажені системи, чутливі до затримок. Зокрема, це системи автоматизованого кредитного скорингу, комплекси онлайн-моніторингу транзакцій на предмет шахрайства, засоби медичної діагностики в режимі реального часу, а також будь-які інші сфери, де якісний та швидкий попередній аналіз неповних даних є критичним фактором для ухвалення автоматизованих рішень.

Еволюція методів імпутації пройшла довгий шлях від простого заповнення середнім значенням до застосування складних нейромережевих архітектур, таких як варіаційні автоенкодери (VAE) та генеративно-змагальні мережі (GAN). Справжню революцію у генеративному моделюванні здійснили дифузійні моделі, проте їх історичний розвиток фокусувався переважно на задачах комп’ютерного зору (генерації та редагуванні зображень). Ця робота концептуально переносить та адаптує ці передові досягнення у площину табличних даних. Важливим аспектом роботи є застосування механізму стохастичного диференціального редагування (SDEdit) – алгоритму, який спочатку був створений для редагування зображень за допомогою користувацьких ескізів. У контексті цієї дипломної роботи SDEdit інноваційно адаптовано як стратегію ініціалізації для відновлення табличних даних. Крім того, робота враховує специфічні математичні умови роботи з даними: дифузійні припущення про концентрацію міри Гауса та межу збурення з корекцією концентрації.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

## 1.1 Контекст та мотивація дослідження

Проблема неповних даних є однією з найбільш поширених та водночас найменш формалізованих перешкод у сучасному машинному навчанні. У критичних прикладних доменах – клінічній медицині, фінансовому ризик-менеджменті, екологічному моніторингу та соціологічних дослідженнях – табличні набори даних рідко надходять у повному вигляді: відсів пацієнтів у клінічних випробуваннях, систематична відмова від розкриття фінансової інформації, збої сенсорів у мережах вимірювань утворюють структурні пропуски, які не є випадковими артефактами збору даних, а несуть у собі прихований статистичний сигнал. Стандартні методи обробки таких пропусків – видалення рядків або детерміністична імпутація середніми значеннями – є математично некоректними за більшості реальних механізмів пропусків і призводять до асимптотичного зміщення оцінок, деградації коваріаційної структури та, як наслідок, до систематичних похибок у рішеннях, що приймаються на основі таких даних. Сукупний масштаб цієї проблеми підкреслює її соціальну значимість: у кредитному скорингу спотворена імпутація здатна відтворювати дискримінаційні патерни; в онкологічних дослідженнях – підривати валідність висновків щодо ефективності терапії. Водночас сучасний методологічний ландшафт залишається фрагментованим: ефективні дискримінативні підходи (MICE, MissForest) жертвують дистрибутивною точністю, а генеративні моделі – зокрема дифузійні — хоча й забезпечують коректне вибіркове відтворення умовного апостеріорного розподілу $p(X_{miss}|X_{obs})$, залишаються практично непридатними для застосувань із жорсткими вимогами до затримки виведення. Саме це протиріччя між дистрибутивною точністю, охопленням

мод і обчислювальною ефективністю визначає науковий запит, на який відповідає ця робота.

### 1.1.1 Ціна неповноти даних

У високорозмірних табличних доменах відсутність даних рідко є нешкідливою аномалією – це структурна особливість, яка фундаментально підриває валідність прикладних моделей. Формально, нехай набір даних розділений на спостережувані та відсутні приклади, $X = (X_{obs}, X_{miss})$, опосередковані бінарною маскою спостереження (1.1):

$$M \in \{0,1\}^{N \times D}, \tag{1.1}$$

де $M$ – бінарна матриця;

$\{0, 1\}$ – множина можливих значень матриці $M$;

$N$ – загальна кількість прикладів у наборі даних;

$D$ – загальна кількість ознак у наборі даних.

Механізми, що керують відсутністю $X_{miss}$, класифікуються на три теоретичні структури: абсолютно випадкові пропуски (MCAR), випадкові пропуски (MAR) та невипадкові пропуски (MNAR) [1], [2]. Коли ці механізми діють у межах табличних даних змішаного типу, вони вносять математичні ускладнення, що погіршують виконання подальших аналітичних завдань (рисунок 1.1).

У клінічних випробуваннях відсів пацієнтів або недотримання протоколу лікування часто призводять до невипадкового зникнення даних у матрицях біомаркерів. Оскільки ймовірність відсутності значень часто корелює з перебігом основного захворювання або побічними реакціями на препарати, отримана структура даних зазвичай відповідає механізмам MAR або MNAR [2], [3]. Аналогічно, у моделюванні фінансових ризиків систематична відсутність певних демографічних даних або відомостей про

доходи рідко є стохастичною. Коли алгоритми обробляють такі невипадкові пропуски, фактор відсутності може виступати як проксі-змінна для неспостережуваних соціально-економічних вразливостей, вносячи статистичне зміщення, яке викривлює процес прийняття алгоритмічних рішень [2], [4].

| | **MCAR**<br>*Пропущено повністю випадково* | **MAR**<br>*Пропущено випадково* | **MNAR**<br>*Пропущено невипадково* |
|---|---|---|---|
| **Визначення** | Не залежить ні від спостережуваних, ні від прихованих ознак.<br>$M \perp X$ | Залежить від відомих ознак, але не від самих пропущених значень.<br>$M \mid X_{obs}$ | Безпосередньо залежить від самих пропущених (невідомих) значень.<br>$M \mid X_{mis}$ |
| **Структура залежності** | Ймовірність пропуску однакова для всіх. Маска є цілком неінформативною.<br>$M \perp X$ | Ймовірність пропуску повністю пояснюється спостережуваними коваріатами.<br>$M \mid X_{obs}$ | Ймовірність пропуску кодує прихований сигнал про втрачені дані $X_{mis}$.<br>$M \mid X_{mis}$ |
| **Наслідки для моделі** | Аналіз повних даних є незміщеним. Спеціальна корекція не потрібна.<br>$M \perp X$ | Ігнорування дає зміщення. Необхідне множинне відновлення (напр., MICE).<br>$M \mid X_{obs}$ | Стандартні методи зміщені. Потрібне спеціалізоване моделювання механізму.<br>$M \mid X_{mis}$ |
| **Приклад із практики** | Випадковий технічний збій датчика або втрата запису лаборантом.<br>$M \perp X$ | Відмова в опитуванні через демографію (напр., старші люди частіше відмовляються).<br>$M \mid X_{obs}$ | **Клінічні дані**: пацієнт вибуває з дослідження через погіршення здоров'я.<br>$M \mid X_{mis}$ |

Зростання складності моделювання та оцінювання

Рисунок 1.1 – Таксономія механізмів пропущених значень

Математичні наслідки неналежного поводження з пропусками є очевидними. Методи видалення, такі як видалення рядками, потребують механізму MCAR для отримання незміщених оцінок. При застосуванні до наборів даних типу MAR або MNAR, відкидання будь-якого рядка $x_i$, що містить неспостережувану ознаку, зменшує ефективний розмір вибірки, що призводить до деградації статистичної значущості. Більш критично те, що при порушенні припущення про MCAR цей підхід вносить асимптотичне

зміщення в оцінку спільного розподілу повних даних, $\hat{\theta}$, яка параметризує $p(X;\theta)$. Це змушує прикладні моделі, що покладаються на імпутований спільний розподіл, вивчати спотворену, нерепрезентативну абстракцію цільової популяції.

Крім того, наявність $X_{miss}$ порушує гетерогенний простір ознак, характерний для табличних даних. Нейронні мережі потребують щільних структурованих вхідних даних для оптимізації. Коли виникає відсутність даних, детерміновані статистичні методи – такі як імпутація середнім – призводять до колапсу дисперсії даних, де за механізмів MAR або MNAR він фундаментально спотворює коваріаційні структури, вносячи непередбачувані спрямовані зміщення [2]. Оскільки ці методи генерують одноточкові оцінки замість вибірки зі справжнього умовного розподілу $p(X_{miss}|X_{obs})$, вони не здатні вловити мультимодальну невизначеність, притаманну відсутнім значенням.

Статистичні обмеження детермінованих методів, зокрема колапс дисперсії та спотворення коваріації, підкреслюють необхідність створення фреймворку, здатного безпосередньо оцінювати $p(X_{miss}|X_{obs})$, а не апроксимувати його математичне очікування. Відповідно, збереження статистичної точності простору табличних даних вимагає прямої реконструкції спільного розподілу. Використовуючи структури, здатні працювати з гетерогенними розподілами, генеративні моделі можуть, за умови належного навчання, проектувати вхідний домен змішаного типу в єдиний неперервний латентний простір $R^{d_z}$.

### 1.1.2 Тристороннє протиріччя

Перехід від дискримінативних методів до генеративних моделей в імпутації табличних даних виявляє фундаментальний архітектурний конфлікт. Цей конфлікт можна структурувати аналітично крізь призму трилеми генеративного навчання [5]. Ця теоретична конструкція, спочатку

сформульована для генерації зображень, як невід'ємний компроміс між візуальною якістю, охопленням мод та швидкістю семплювання, потребує адаптації для застосування до імпутації табличних наборів даних. У нашому контексті «візуальна якість» означає дистрибутивну точність (точність вивченого спільного розподілу), тоді як «швидкість семплювання» безпосередньо відповідає затримці на етапі виведення (inference latency) при генерації імпутацій. Необхідність збереження мультимодальних умовних апостеріорних ймовірностей природньо віддзеркалює початкову вимогу щодо охоплення мод. У межах цієї структури трилема проявляється як необхідність балансування між трьома конкуруючими статистичними та операційними вимогами: дистрибутивною точністю, охопленням мод та суворими обмеженнями на затримку виведення (рисунок 1.2).

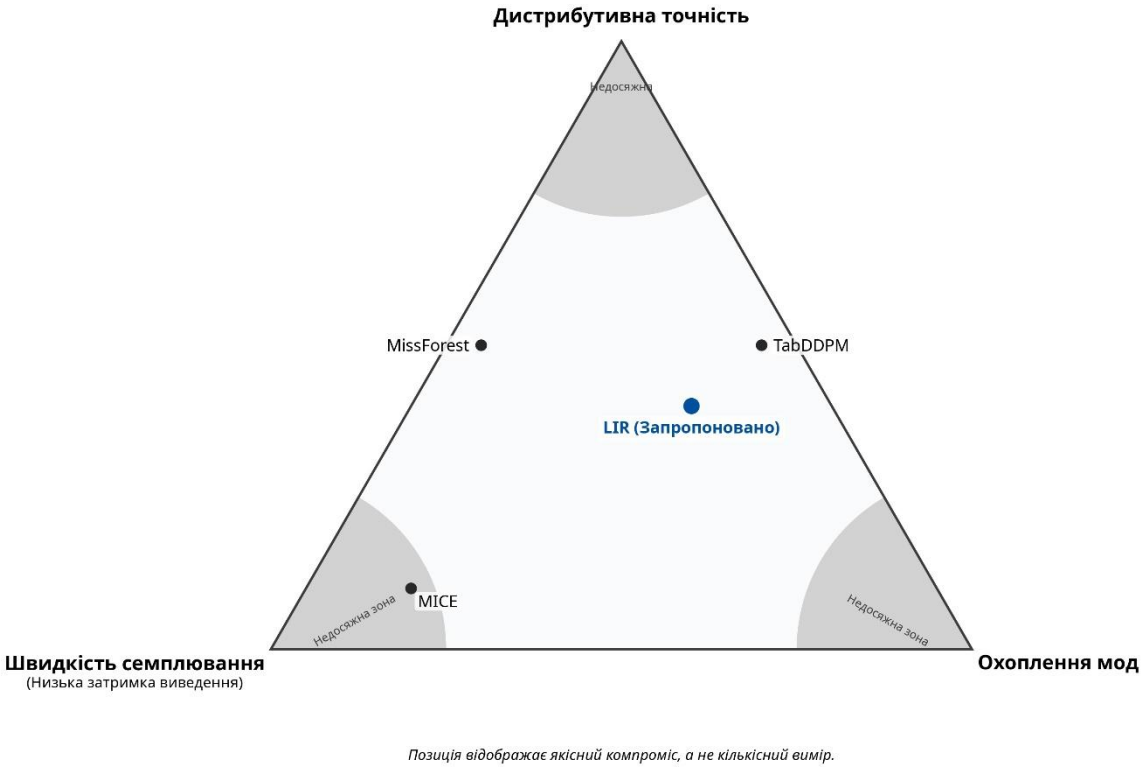



Рисунок 1.2 – Тристороннє протиріччя

Традиційні методології долають ці компроміси за допомогою різних операційних профілів. Множинна імпутація за допомогою ланцюгових рівнянь (MICE), класичний підхід, явно зберігає дисперсію та фіксує невизначеність шляхом створення кількох імпутованих наборів даних, отриманих з умовних розподілів. Хоча MICE є гнучким фреймворком, який зазвичай використовує нелінійні умовні моделі – такі як дерева рішень або PMM – його найбільш визнаною емпіричною перевагою є обчислювальна ефективність. Після завершення фази ітераційного навчання оптимізований MICE швидко виконує прогнозування, що робить його еталоном продуктивності для виведення. Проте через покладання на ітераційні умовні апроксимації, MICE часто має труднощі з фіксацією глобального спільного розподілу та складних структурних взаємодій у високорозмірних наборах даних.

Водночас, детерміновані ансамблеві алгоритми на основі дерев, такі як MissForest [6], покладаються на мінімізацію точкових похибок. Для неперервних ознак регресори випадкового лісу оптимізують цільову функцію втрат $L_2$, що математично спрямовує оцінювач до умовного математичного очікування [7]. Коли відсутня неперервна ознака залежить від складних мультимодальних взаємодій між спостережуваними змінними, така поведінка з мінімізацією $L_2$ призводить до усереднення за модами умовної апостеріорної ймовірності. Як наслідок, це може спричинити дефляцію дисперсії: імпутовані значення штучно групуватимуться навколо середнього, що може звузити коваріаційну матрицю та систематично спотворити справжній умовний розподіл [2]. Крім того, ітераційна ансамблева природа MissForest створює значні обчислювальні вузькі місця як на етапі навчання, так і на етапі виведення. Вимога багаторазового оцінювання глибоких ансамблів дерев рішень для всіх неповних ознак погано масштабується; як показано в подальших емпіричних оцінках на наборі даних Adult Census, затримка виведення у MissForest може бути в 50 разів вищою, ніж у однопрохідних нейронних архітектур, таких як GAIN.

Такий профіль затримки може зробити його обчислювально неприйнятним для сценаріїв пакетної обробки з високою пропускною здатністю.

Глибокі генеративні моделі (DGMs), зокрема дифузійні архітектури з повною траєкторією [8], теоретично здатні задовольнити ці вимоги, оскільки їхня цільова функція безпосередньо орієнтована на розподіл даних, а не на його умовне математичне очікування. Параметризуючи нейронну мережу для прогнозування внесеного шуму – дифузійні моделі успішно реконструюють складні спільні розподіли. Проте безумовні дифузійні архітектури за своєю природою моделюють безумовний розподіл даних $p(x)$. Щоб гарантувати, що імпутовані значення вибираються з модельованого умовного апостеріорного розподілу $p(x_{miss}|x_{obs})$, ці структури потребують специфічних алгоритмічних втручань. Ці втручання варіюються від явних архітектур умовного навчання, таких як CSDI [9] до методів заміни під час семплювання. За відсутності строго валідованих табличних аналогів, остання категорія методів ілюструється підходами, спочатку розробленими для домену доповнення зображень (наприклад, RePaint [10]), що слугують у роботі виключно концептуальними прикладами. Хоча методи на кшталт classifier-free guidance іноді пропонуються для посилення сигналу обумовлення, їх застосування може вносити артефакти надмірного обумовлення (overconditioning) та нестабільність у табличних середовищах. Крім того, застосування цієї неперервної структури до дискретно-неперервних табличних даних вимагає структурних модифікацій. Сучасні моделі вирішують проблему цього гетерогенного простору за допомогою різних механізмів, таких як поєднання неперервної дифузії з мультиноміальною дифузією для прямої обробки категоріальних токенів (наприклад, TabDDPM [11]), або використання подвійних ко-еволюційних дифузійних моделей через контрастивну функцію оцінки (наприклад, CoDi [12]).

Незважаючи на їхню теоретичну здатність до високої точності відтворення, основним операційним обмеженням дифузійних моделей залишається затримка виведення. Висока дистрибутивна точність цих моделей досягається шляхом симуляції параметризованої зворотної траєкторії з початкового розподілу шуму. Хоча вдосконалення методів швидкого семплювання прискорило цей процес шляхом обходу суворих марковських вимог, процедура ітераційного усунення шуму все ще вимагає значних обчислювальних витрат при виведенні порівняно з негенеративними альтернативами. Цей обчислювальний тягар обмежує розгортання повнотраєкторних дифузійних моделей в умовах суворих обмежень на низьку затримку, що зрештою залишає генеративну трилему нерозв'язаною для задачі імпутації табличних даних. Подолання цього тристороннього протиріччя вимагає розробки нової гібридної архітектури, яка б поєднувала швидкість дискримінативних методів із точністю генеративних моделей. Детальна концепція, архітектура та математична формалізація такого підходу, розробленого в межах даної кваліфікаційної роботи, наведена у розділі 3.

## 1.2 Постановка задачі

Метою кваліфікаційної роботи є розробка гібридного генеративного фреймворку (LIR) для статистично коректного та обчислювально ефективного відновлення пропущених значень у гетерогенних табличних наборах даних.

Мета імпутації формулюється як задача вибірки з умовного апостеріорного розподілу $p(x_{miss}|x_{obs})$, а не мінімізації точкової похибки. Дискримінативні моделі, оптимізовані за $L_2$-функцією втрат, зводяться до оцінки умовного математичного сподівання $E[x_{miss}|x_{obs}]$, яке за мультимодального розподілу може опинятися у низькощільних областях

між модами – це призводить до колапсу дисперсії та спотворення кореляційної структури даних (патологія усереднення за модами).

Для досягнення поставленої мети у роботі необхідно виконати такі завдання:

– провести глибокий аналіз наукових джерел за темою роботи та дослідити еволюцію методів обробки неповних даних;

– виконати порівняльний аналіз існуючих парадигм імпутації (від класичної статистики до сучасних дифузійних моделей) з огляду на «тристороннє протиріччя» між дистрибутивною точністю, охопленням мод та затримкою виведення;

– розробити архітектуру та математичну базу гібридного генеративного фреймворку;

– провести комплексні експерименти на реальних багатовимірних наборах даних для оцінки прогностичної корисності, точності відтворення спільного розподілу та обчислювальної ефективності розробленого методу порівняно з базовими аналогами.

Таким чином, задача дослідження полягає у створенні генеративного фреймворку, що:

– здійснює вибірку з повного умовного апостеріорного розподілу $p(x_{miss}|x_{obs})$, повною мірою зберігаючи мультимодальність та кореляційну структуру базових даних;

– забезпечує обчислювальну ефективність (скорочення кількості оцінок нейронних функцій), сумісну з вимогами до затримки при застосуванні у високонавантажених реальних системах;

– коректно обробляє гетерогенний простір табличних ознак змішаного типу, включаючи розріджені категоріальні змінні та числові розподіли з «важкими хвостами».

Детальна математична формалізація всіх компонент запропонованої архітектури наведена у розділі 3.

# 2 ЕВОЛЮЦІЯ ПАРАДИГМ ІМПУТАЦІЇ

## 2.1 Класичні статистичні методи

До появи глибоких генеративних моделей проблема відсутності даних вирішувалася класичними статистичними методами: від простої одновимірної імпутації до ітераційних алгоритмів на основі оптимізації. Попри те, що ці підходи заклали первинний математичний фундамент, вони мають суттєві структурні обмеження: колапс дисперсії, покладання на детерміновані точкові оцінки та спотворення розподілу. Це фундаментально обґрунтовує необхідність використання генеративних підходів, здатних коректно відновлювати повний багатовимірний розподіл даних.

Імпутація середнім має мінімальну обчислювальну складність $\mathcal{O}(N)$ на ознаку. Задавши бінарну маску спостережень $M$ кожне пропущене значення $x_{i,j}$, (де $M_{i,j} = 0$) замінюється вибірковим середнім $\bar{x_j}$ за спостережуваною підмножиною (2.1):

$$\hat{x}_{i,j} = \bar{x}_j = \frac{1}{|\{k \in \{1,\dots,N\} : M_{k,j}=1\}|} \sum_{k \in \{1,\dots,N\} : M_{k,j}=1} x_{k,j}, \tag{2.1}$$

де $\bar{x}_j$ – вибіркове середнє арифметичне значення всіх наявних значень для $j$-ої ознаки у наборі даних;

$x_{k,j}$ – фактичне значення $j$-ої ознаки для $k$-го прикладу.

Хоча цей підхід зберігає перший момент розподілу, він є статистично неспроможним для моментів вищих порядків. Заміна стохастичних значень константою призводить до концентрації маси ймовірності в точці математичного очікування [13]. Відповідно, стохастична варіація імпутованих записів дорівнює нулю, що систематично занижує оцінену дисперсію $\hat{\sigma}_j^2$ відносно справжньої $\sigma_j^2$ [2]. Крім того, це ігнорує коваріаційну

структуру, штучно зміщуючи коефіцієнти кореляції між ознаками до нуля [2], [14].

Для часткового вирішення проблеми колапсу дисперсії застосовуються непараметричні методи «гарячої вибірки», такі як метод $k$-найближчих сусідів (KNN). Відсутня ознака імпутується шляхом вибору донора з множини $k$ найближчих сусідів $\mathcal{N}_k(x_i)$, визначених метрикою відстані $d$ у просторі спільних спостережуваних ознак (2.2):

$$\hat{x}_{i,j} = x'_j, \quad \text{де } x' \sim \text{Uniform}\big(\mathcal{N}_k(x_i)\big), \tag{2.2}$$

де $x'_j$ – фактичне значення $j$-ої ознаки у об'єкта-донора, яке безпосередньо копіюється для заповнення пропуску;

$x'$ – конкретне спостереження-донор.

За умови MCAR, метод зберігає маргінальні розподіли, проте за механізмів MAR або MNAR маргінальний розподіл зміщується в бік пулу донорів [15]. Суттєвим недоліком є також деградація методу у високорозмірних просторах – зі збільшенням розмірності $D$ метод страждає від концентрації попарних відстаней: відносний контраст обчислення відстаней прямує до нуля [16]. У розріджених табличних даних метрики відстані втрачають розрізняльну здатність, що критично знижує точність імпутації [6], [17].

Ймовірнісною альтернативою, заснованою на оцінці максимальної правдоподібності, є алгоритм максимізації очікування (Expectation-Maximization, EM). EM – це ітераційна процедура, розроблена для знаходження оцінок максимальної правдоподібності параметрів $\theta$ за наявності неповних даних [18]. Метод обчислює умовне математичне очікування логарифмічної правдоподібності, ітеративно чергуючи два кроки, доки зміна параметрів не стане мізерною, тобто досягне конвергенції: Е-крок та М-крок.

Е-крок обчислює $Q$-функцію на основі спостережуваних даних $X$ та поточних оцінок параметрів $\theta^{(t)}$, отриманих на попередньому кроці (2.3):

$$Q\left(\theta\middle|\theta^{(t)}\right) = \mathbb{E}_{X_{miss}|X_{obs},\theta^{(t)}}[log\, p\,(X_{obs}, X_{miss}|\theta)], \quad (2.3)$$

де $\theta^{(t)}$ – поточні оцінки параметрів моделі, які були обчислені на попередній ітерації $t$;

$\mathbb{E}_{X_{miss}|X_{obs},\theta^{(t)}}[\ldots]$ – оператор умовного математичного сподівання, який обчислює середнє значення виразу в дужках, спираючись на ймовірнісний розподіл пропущених даних $X_{miss}$, за умови, що нам відомі спостережувані дані $X_{obs}$ та поточні параметри моделі $\theta^{(t)}$;

$log\, p\,(X_{obs}, X_{miss}|\theta)$ – логарифм спільної функції правдоподібності для «повних» даних (тобто об'єднаного набору спостережуваних і пропущених значень), параметризованої новими значеннями $\theta$.

М-крок оновлює параметри (2.4):

$$\theta^{(t+1)} = \arg\max_{\theta} Q\left(\theta\middle|\theta^{(t)}\right), \quad (2.4)$$

де $\theta^{(t+1)}$ – оновлені оцінки параметрів моделі, які будуть використовуватися на наступній ітерації алгоритму $t+1$;

$Q\left(\theta\middle|\theta^{(t)}\right)$ – цільова $Q$-функція, яка була побудована на Е-кроці з використанням старих параметрів $\theta(t)$.

Незважаючи на ефективність оцінювання за умови MAR [1], безпосередня корисність EM для генеративної імпутації є обмеженою. Алгоритм розроблений для пошуку єдиної оптимальної точки $\theta^*$ [18], що на практиці призводить до детермінованого обчислення умовного середнього відсутніх значень. Цей підхід кардинально контрастує з дифузійними моделями, які здатні робити вибірку з апроксимації повного розподілу даних шляхом стохастичного вдосконалення [8].

Центральним недоліком класичних статистичних методів – детермінованих чи на основі правдоподібності – є їхнє фундаментальне покладання на детерміновані точкові оцінки у конфігураціях одиничної імпутації. Розглядаючи імпутоване значення як фіксовану, відому істину, ці методи не здатні вловити повну стохастичну структуру умовного розподілу. Хоча структури множинної імпутації частково пом'якшують недооцінку дисперсії шляхом генерування $m$ незалежних наборів даних та об'єднання результатів за правилами Рубіна, валідність цього підходу суттєво залежить від механізму імпутації. Для $m$ імпутацій, що дають оцінки параметрів $\hat{Q}_i$ з відповідними дисперсіями $U_i$, узагальнена оцінка $\bar{Q}$ та загальна дисперсія $T$ обчислюються наступним чином, де загальна дисперсія поєднує середню внутрішньогрупову дисперсію та міжгрупову дисперсію (2.5):

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i, \quad T = \bar{U} + \left(1 + \frac{1}{m}\right) B, \tag{2.5}$$

де $\bar{Q}$ – загальна точкова оцінка шуканого параметра за результатами множинної імпутації;

$m$ – загальна кількість відновлених наборів даних;

$\hat{Q}_i$ – точкова оцінка цільового параметра під час аналізу $i$-го імпутованого набору даних;

$T$ – загальна дисперсія об'єднаної оцінки $\bar{Q}$;

$\bar{U}$ – внутрішньовибіркова дисперсія, яка дорівнює середньому арифметичному дисперсій параметра всередині кожного з $m$ наборів даних, і відображає звичайну статистичну похибку;

$B$ – міжвибіркова дисперсія, яка вимірює розкид самих точкових оцінок $\hat{Q}_i$ між різними імпутованими наборами;

Для того, щоб ця дисперсія була статистично валідною, кожна ітерація повинна генеруватися з моделі, що правильно представляє апостеріорну структуру. Байєсівські методи, такі як MICE, досягають цього лише за

суворих умов: наявність щільного пулу донорів та коректна специфікація лінійної моделі [19]. У високорозмірних та розріджених табличних даних ці умови систематично порушуються.

Використання стратегій на основі умовного середнього спричиняє так звану «патологію усереднення за модами». Навіть при багаторазовій вибірці з неправильно специфікованої унімодальної моделі, стандартні похибки систематично недооцінюються. Це призводить до зміщення в бік оптимізму (optimism bias) та звужених довірчих інтервалів, які не відображають реальну невизначеність [2], [20].

## 2.2 Дискримінативне машинне навчання

Еволюція від класичних статистичних підходів до алгоритмічного машинного навчання створила набір дискримінативних методів, розроблених для фіксації складних нелінійних залежностей, притаманних табличним даним. Долаючи фундаментальне «тристороннє протиріччя» табличної імпутації – балансування між обчислювальною ефективністю, дистрибутивною точністю та затримкою виведення – дискримінативні методи мають обчислювальну ефективність за свою основну перевагу. Вони досягають високої пропускної здатності, жертвуючи дистрибутивною точністю, зазвичай крізь призму умовного математичного очікування. Хоча ці архітектури пропонують певні переваги щодо затримки порівняно з багатоетапними генеративними моделями, вони фундаментально фокусуються на мінімізації похибки в кожній точці ціною точного представлення справжнього базового розподілу даних.

### 2.2.1 MICE та MissForest

Багатовимірна імпутація за допомогою ланцюгових рівнянь (MICE), також відома як повна специфікація умовних розподілів (Fully Conditional

Specification, FCS), являє собою значний крок вперед порівняно з одновимірними методами. MICE апроксимує багатовимірний спільний розподіл через серію ітераційних умовних щільностей [19], [20]. Формально фреймворк розкладає спільний розподіл $p(X_1, \dots, X_D|\theta)$ на набір із $D$ умовних розподілів (2.6):

$$p\left(X_j \middle| X_j, \theta_j\right), \tag{2.6}$$

де $X_j$ позначає всі ознаки, крім стовпця з індексом $j$.

Це дозволяє дослідникам застосовувати надзвичайно гнучке моделювання до гетерогенних змінних. Для пом'якшення ефекту колапсу дисперсії, характерного для чистої регресії, MICE зазвичай використовує прогнозне зіставлення середніх (PMM) для неперервних ознак, вибираючи імпутовані значення з пулу спостережуваних донорів, найближчих до передбачуваного умовного математичного очікування, тоді як категоріальні ознаки обробляються за допомогою логістичної або мультиноміальної регресії.

Незважаючи на здатність PMM зберігати одновимірну маргінальну дисперсію шляхом вибірки емпіричних залишків (residuals), таке збереження не гарантує точності спільного розподілу або складних умовних взаємодій [21]. MICE залишається фундаментально обмеженим своєю ітераційною структурою, яка покладається на припущення про «сумісність» умовних розподілів – сувору математичну вимогу існування «правильного» спільного розподілу, який ідеально відповідає всім заданим маргінальним умовним розподілам. Ця умова часто не виконується у складних табличних схемах, що проявляється у вигляді структурних збоїв: відмов через колінеарність у недовизначених режимах ($D \gg N$) [22] та обмежень розмірності, коли величезні простори ознак ($D > 10^3$) погіршують ітераційний процес [21]. Крім того, зі збільшенням простору ознак оцінка збіжності ланцюгових рівнянь стає обчислювально надто дорогою та

емпірично ненадійною [21]. Як наслідок, хоча окремі маргінальні розподіли можуть виглядати статистично обґрунтованими, отримані умовні моделі фактично згладжують різкі розриви щільності та межі змішаних дискретно-неперервних носіїв, що визначають справжній простір даних.

Для обходу параметричних припущень MICE дослідницька спільнота перейшла до непараметричних ансамблевих підходів, найбільш відомим з яких є MissForest [6]. MissForest розглядає імпутацію як задачу навчання з учителем, використовуючи випадкові ліси (Random Forests) для фіксації нелінійних взаємодій ознак без необхідності явного задання розподілів. Алгоритм працює ітераційно: він ініціалізує відсутні значення за допомогою простого методу, а потім навчає послідовність лісів для прогнозування відсутніх компонентів, завершуючи ітераційну процедуру, коли похибка «поза пакетом» (Out-of-Bag, OOB) перестає зменшуватися між послідовними раундами [6].

Хоча практики спочатку обрали MissForest для подолання параметричних обмежень, його ітераційна ансамблева природа створює серйозні обчислювальні вузькі місця. Припускаючи стандартну підвибірку ознак $\sqrt{D}$ у кожному вузлі розбивки, побудова базового випадкового лісу потребує $\mathcal{O}\left(K \cdot N \log N \cdot \sqrt{D}\right)$ операцій на ансамбль [23]. Розширення цієї складності на $I$ ітераційних циклів MissForest дає загальну асимптотичну складність $\mathcal{O}\left(I \cdot K \cdot N \log N \cdot \sqrt{D}\right)$ операцій на один раунд імпутації. Навіть для невеликої кількості циклів (наприклад, $I \leq 10$), чутливість масштабування $\mathcal{O}(N \log N)$ для сортування вузлів робить цей підхід обчислювально неприйнятним для масивних наборів даних ($N > 10^6$).

### 2.2.2 Патологія усереднення за модами

Невдачу дискримінативних моделей у збереженні дистрибутивної точності можна концептуально зрозуміти через «патологію усереднення за

модами». Коли модель навчається мінімізувати функцію втрат $L_2$, теорема про ортогональну проекцію (що еквівалентно властивості $L^2$-проекції умовного математичного очікування) диктує, що вона зійдеться до умовного математичного очікування $E[x_{miss}|x_{obs}]$. Хоча це правильно мінімізує середньоквадратичну похибку, такий підхід водночас фундаментально викривлює мультимодальні апостеріорні розподіли.

Розглянемо симетричний бімодальний апостеріорний розподіл $p(x_{miss}|x_{obs})$, що характеризується двома окремими модами з однаковою вагою в точках $\pm a$. Дискримінативна модель, що оптимізує $L_2$, дасть точкову оцінку точно між ними – у нулі. Як наслідок, оцінювач розміщує імпутоване значення в області, де умовна щільність ймовірності $p(x_{miss}|x_{obs})$ близька до нуля, замість того щоб точно відображати процес генерації даних (рисунок 2.1).

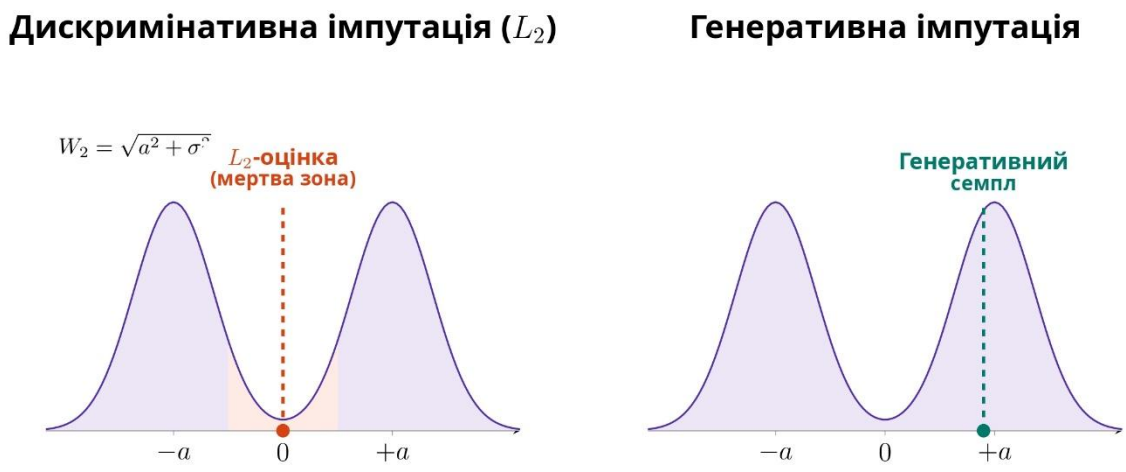



Рисунок 2.1 – Візуалізація патології усереднення за модами

Теоретично, використання точкової оцінки зводить розподіл імпутації до єдиного фіксованого значення (виродженого розподілу). Це порушує фундаментальну вимогу статистичної імпутації: розподіл імпутованих значень має відповідати носію справжнього апостеріорного розподілу [2]. Ігнорування стохастичної невизначеності призводить до того, що

математична розбіжність між таким наближенням і цільовим розподілом стає критичною.

Геометричну величину цього структурного збою можна кількісно оцінити. Як буде доведено у розділі 3, за допомогою відстані Вассерштейна $W_2$, у складних багатомодальних розподілах відстань між усередненою точковою оцінкою та справжнім апостеріорним розподілом катастрофічно зростає. Це систематичне знищення дисперсії демонструє, що хоча дискримінативні методи пропонують ефективну єдину оцінку, вони фундаментально непридатні для високоточної реконструкції даних.

## 2.3 Ранні глибокі генеративні моделі: VAE та GAN

Обмеження дискримінативного точкового оцінювання мотивували розробку глибоких генеративних моделей (Deep Generative Models, DGMs), здатних вивчати базовий многовид даних. На відміну від моделей класичного машинного навчання, описаних вище, ці структури намагаються безпосередньо апроксимувати спільний розподіл $p(x)$, що дозволяє генерувати множинні різноманітні імпутації. Проте перше покоління цих моделей, представлене варіаційними автокодувальниками (VAEs) та генеративно-змагальними мережами (GANs), ввело структурну нестабільність та дистрибутивні припущення, які часто суперечать мультимодальним функціям щільності та розривним поверхням щільності ймовірності, притаманним табличним даним.

VAE формалізують імпутацію як задачу прихованого моделювання, використовуючи кодувальник $q_\phi(z|x)$ та декодувальник $p_\theta(x|z)$, де навчання відбувається шляхом максимізації варіаційної нижньої межі (ELBO). При застосуванні до табличних даних VAE стикаються з двома критичними патологіями:

– колапс апостеріорного розподілу: надмірна KL-регуляризація змушує апостеріорний розподіл $q_\phi(z|x)$ згортатися до неінформативного

апріорного $p(z)$. У результаті декодувальник ігнорує латентні ознаки $z$ і повертається до тривіального умовного середнього [24]. Цей ефект посилюється невідповідністю між розмірністю латентного простору $d_z$ та внутрішньою розмірністю даних [3], [25];

– дистрибутивне згладжування: стандартний ізотропний гаусіан $\mathcal{N}(0, I)$ у якості апріорного розподілу $p(z)$ змушує кодувальник відображати роз'єднані моди вхідного простору у перекривні латентні області [26]. Це розмиває різкі локалізовані піки, характерні для табличних розподілів. Хоча спеціалізовані архітектури на кшталт HI-VAE [3] та MIWAE, що використовує зважену за важливістю ELBO для маргіналізації неспостережуваних вимірів [27], частково пом'якшують ці проблеми, фундаментальне згладжування залишається.

Попри ці недоліки, здатність VAE відображати дискретно-неперервні табличні ознаки в уніфікований неперервний евклідів простір $\mathcal{Z} \subset R^{d_z}$ є критичною перевагою. Як показано в наступних розділах, це дозволяє застосовувати неперервні стохастичні диференціальні рівняння (SDE), використовуючи середнє кодувальника $\mu_\phi$ як детерміновану апроксимацію латентного простору.

На відміну від підходів на основі правдоподібності, архітектури типу GAIN [28] формулюють імпутацію як мінімаксну гру. Генератор $G$ намагається заповнити відсутні значення, а дискримінатор $D$ (за підтримки вектора «підказок» $h$) намагається відрізнити імпутовані значення від спостережуваних.

Хоча GAN генерують різкі, незгладжені зразки, вони страждають від притаманної нестабільності: колапсу мод [29] та затухання градієнтів [30]. Для табличних даних ключовою проблемою є невідповідність неперервного многовиду навчання дискретному носію даних. Зворотне поширення помилки вимагає неперервної релаксації дискретних ознак (наприклад, через Gumbel-Softmax). Це штучно зміщує розподіл категоріальних змінних у внутрішню частину ймовірнісного симплексу,

спотворюючи їх репрезентацію. Гібридні моделі, такі як CTGAN [31], що використовують байєсівські моделі сумішей (BGMM) для попередньої обробки, не мають функції правдоподібності, що унеможливлює строгу байєсівську оцінку невизначеності імпутацій.

Хоча ранні DGMs перевели галузь до навчання на многовидах, вони залишаються обмеженими «трилемою генеративного навчання»: добре калібровані VAE пропонують широке охоплення мод, але страждають від надмірного згладжування щільності, тоді як GAN забезпечують високу якість зразків за рахунок нестабільності навчання та колапсу мод.

Дифузійні моделі на основі score-based функції, детально розглянуті в наступному розділі, виступають переконливим кандидатом для вирішення двох із цих трьох проблем – якості зразків та охоплення мод – ціною третьої.

## 2.4 Дифузійні ймовірнісні моделі: сучасний стан розвитку

Нещодавня зміна парадигми в генеративному моделюванні характеризується переходом від змагальної нестабільності генеративно-змагальних мереж (GAN), які часто страждають від колапсу мод у розріджених табличних режимах [31], та апроксимаційних обмежень цілі варіаційної нижньої межі (ELBO) у варіаційних автокодувальниках (VAE), до дистрибутивної точності оціночних дифузійних моделей. У середовищі генерації таблиць дифузійні ймовірнісні моделі усунення шуму (DDPM) демонструють високу конкурентоспроможність щодо охоплення мод [11] шляхом навчання інверсії прямого марковського процесу, який трансформує розподіл даних $p_0(x)$ в ізотропний гаусів апріорний розподіл $p_T(x) \approx \mathcal{N}(0, I)$. До необроблених вхідних даних застосовуються кроки архітектурної попередньої обробки, такі як квантильні перетворення або one-hot encoding, для визначення носія представлення, а $p_0(x)$ формально позначає розподіл, індукований на цьому попередньо обробленому представленні.

У контексті імпутації табличних даних вибір роботи в нестисненому просторі ознак, а не в стисненому латентному просторі, залишається важливою архітектурною відмінністю. Працюючи безпосередньо в цьому повнорозмірному просторі $x$, стандартні табличні дифузійні моделі уникають потенційно втратних вузьких місць традиційних архітектур на основі декодувальників, гарантуючи збереження структурних зв'язків між неперервними та категоріальними змінними протягом ітераційної траєкторії денойзингу $x_T \to x_0$. Проте робота виключно в цьому високорозмірному просторі змішаного типу вносить специфічні репрезентативні та обчислювальні виклики, які намагаються розв'язати наступні методології.

### 2.4.1 Знакові адаптації для табличних даних

Застосування дифузії до табличного домену потребує спеціалізованої обробки гетерогенності ознак. Основним етапом цієї адаптації є модель TabDDPM [11], яка вводить двоканальний генеративний підхід. Для неперервних ознак модель використовує стандартну гаусову дифузію, параметризовану через прогнозування шуму ($\epsilon$-прогнозування). Навпаки, для категоріальних ознак із $K$ класами вона застосовує мультиноміальну дифузію. Модель розглядає $x_t \in \{0,1\}^K$ як масу ймовірності на симплексі та визначає симетричну матрицю переходів як (2.7), а прямий перехід формалізує як (2.8):

$$Q_t = (1 - \beta_t)I + \beta_t \frac{11^{\mathrm{T}}}{K}, \tag{2.7}$$

де $Q_t$ – матриця ймовірностей переходів на часовому кроці $t$, яка визначає ймовірності того, що категоріальна ознака змінить своє поточне значення на будь-яке інше (або збереже його) під час чергового кроку зашумлення;

$\beta_t$ – гіперпараметр розкладу шуму (transition probability) на кроці $t$, який контролює інтенсивність зашумлення (ймовірність того, що значення ознаки буде замінено на випадкове);

$I$ – одинична матриця розмірності $K \times K$;

$11^{\mathrm{T}}$ – створює матрицю розмірності $K \times K$, яка повністю складається з одиниць;

$K$ – загальна кількість можливих унікальних категорій (класів) для конкретної дискретної ознаки в наборі даних.

$$q(x_t|x_{t-1}) \coloneqq \mathrm{Cat}\left(x_t; (1-\beta_t)x_{t-1} + \beta_t \frac{1}{K}\right), \tag{2.8}$$

де Cat (...) – категоріальний розподіл, який вказує на те, що змінна $x_t$ може набувати одного з $K$ дискретних станів, а ймовірності кожного стану задаються виразом після крапки з комою;

$x_t$ – стан категоріальної ознаки на кроці $t$;

$x_{t-1}$ – зашумлений на попередньому етапі стан цієї ж ознаки на кроці $t-1$;

$\frac{\mathbf{1}}{K}$ – вектор рівномірної ймовірності.

Важливо, що замість $\epsilon$-прогнозування категоріальний зворотний перехід $p_\theta(x_{t-1}|x_t)$ використовує нейронну мережу для прогнозування непошкодженого токена $\hat{x}_0$. Як емпірично доведено в архітектурі TabDDPM [11], така параметризація забезпечує необхідну стабільність при роботі з дискретними категоріальними станами, відрізняючи генеративний механізм дискретного каналу від неперервного домену. Визначаючи $\bar{Q}_t$ як кумулятивну матрицю переходів, модель обчислює справжній категоріальний апостеріорний розподіл у замкненій формі для отримання попереднього стану (2.9):

$$q(x_{t-1}|x_t, \hat{x}_0) = \mathrm{Cat}\left(x_{t-1}; \frac{(x_t Q_t) \odot (\hat{x}_0 \bar{Q}_{t-1})}{\hat{x}_0 \bar{Q}_t x_t^T}\right), \quad (2.9)$$

де $\mathrm{Cat}(\dots)$ – категоріальний розподіл;

$x_{t-1}$ – цільовий стан категоріальної ознаки на кроці $t-1$, який модель намагається згенерувати;

$x_t$ – відомий поточний зашумлений стан ознаки на кроці $t$;

$\hat{x}_0$ – передбачений нейронною мережею (на основі поточного стану $x_t$) початковий «чистий» стан даних на кроці $t = 0$;

$Q_t$ – матриця ймовірностей переходу для одного кроку (від $t-1$ до $t$);

$\bar{Q}_{t-1}$ та $\bar{Q}_t$ – кумулятивні матриці ймовірностей переходу від початкового нульового кроку ($t = 0$) до кроків $t-1$ та $t$ відповідно, що обчислюються шляхом послідовного перемноження однокрокових матриць;

$\odot$ – добуток Адамара (оператор поелементного множення двох векторів).

Оскільки $Q_t$ є симетричною, вона застосовується безпосередньо в чисельнику; у загальному несиметричному формулюванні дискретної дифузії [32] прямий член у чисельнику потребував би транспонованої матриці $Q_t^{\mathrm{T}}$ , але симетрія ядра рівномірної дифузії усуває цю відмінність. Хоча TabDDPM досягає високої точності шляхом моделювання спільного розподілу ознак змішаного типу, її залежність від стандартного марковського зворотного процесу – який зазвичай потребує $T = 1000$ кроків – спричиняє затримку виведення, що обмежує її корисність у середовищах із високою пропускною здатністю.

Іншим поворотним досягненням є CSDI (умовна оціночна дифузія для імпутації), яка спеціалізується на просторово-часових даних та часових рядах [9]. CSDI вдосконалює завдання імпутації, використовуючи механізми самоуваги (self-attention) для обумовлення генеративного процесу на спостережуваних ознаках. Формально, замість безумовного моделювання $p(x)$, CSDI орієнтується на умовний розподіл $p(x_{miss}|x_{obs})$. Ключовим методологічним внеском, що робить це можливим, є стратегія

навчання: CSDI використовує схему маскування з двома шаблонами, де підмножина спостережуваних ознак додатково маскується, щоб слугувати ціллю імпутації під час навчання, тоді як решта спостережуваних ознак обумовлюють модель. Практично це реалізується шляхом конкатенації бінарних масок спостереження та відомих значень $x_{obs}$ разом із зашумленим входом $x_t$ на кожному кроці обчислення нейронної функції. Хоча емпіричні оцінки демонструють здатність CSDI ефективно використовувати часові кореляції, вона залишається обмеженою властивою їй складністю семплювання $\mathcal{O}(T)$.

### 2.4.2 Фронтир прискорення

Для пом’якшення обмежень затримки виведення дослідження розгалужуються на детерміноване семплювання та усічення траєкторії. Неявні дифузійні моделі усунення шуму (DDIM) [33] представляють усталений підхід до прискорення виведення шляхом визначення немарковського процесу, який зберігає прямі маргінальні розподіли оригінальної DDPM, дозволяючи використовувати більші кроки та зменшувати кількість обчислень нейронних функцій (NFEs). Однак DDIM виведена виключно для неперервної гаусової дифузії. Її неможливо безпосередньо застосувати для прискорення дискретного мультиноміального каналу в таких структурах, як TabDDPM, що працює на ймовірнісному симплексі через дискретний ланцюг Маркова. Емпірично, як буде показано в дослідженнях цієї роботи, спроба агресивного зменшення NFEs у дискретному каналі часто призводить до передчасного колапсу категоріальних розподілів до рівномірного апріорного розподілу, що суттєво погіршує охоплення мод. Отже, прискорення дифузії змішаного типу в просторі необроблених ознак потребує розрізнених стратегій для неперервних та категоріальних змінних. Ця архітектурна дихотомія створює

значний бар'єр для уніфікованого детермінованого прискорення для всіх типів ознак.

Подальше прискорення передбачає усічення траєкторії. Формалізована в таких структурах, як SDEdit [34], ця техніка ініціалізує зворотний генеративний процес на проміжному часовому кроці $t_{start} < T$, використовуючи збурений шумом орієнтир для обходу стохастично домінуючих режимів високого шуму. Це слід відрізняти від наївного «холодного старту» з використанням детермінованої оцінки без каліброваного збурення шумом, що, як формалізовано у наступному розділі цієї роботи, фундаментально порушує динаміку базового многовиду. Якщо ініціалізація не є строго каліброваною для перебування на зашумленому многовиді даних, вона вносить серйозні дистрибутивні зміщення та підриває цілісність згенерованих зразків.

### 2.4.3 Формалізація обчислювального вузького місця

Незважаючи на свою точність, розгортання дифузійних моделей обмежене чітко визначеним фундаментальним обчислювальним вузьким місцем. Генерація одного зразка потребує $T$ послідовних обчислень нейронної мережі $\epsilon_\theta$. Загальна вартість виведення найбільш точно характеризується як (2.10):

$$\mathrm{Cost}_{\mathrm{Diffusion}} = \mathcal{O}(T \cdot \mathrm{Cost}_{\mathrm{NFE}}), \tag{2.10}$$

де $\mathrm{Cost}_{\mathrm{Diffusion}}$ – загальна обчислювальна вартість процесу генерації або відновлення даних за допомогою дифузійної моделі;

$T$ – загальна кількість часових кроків (ітерацій) у зворотному процесі дифузії, необхідних для поступового і повного очищення даних від шуму;

$\mathrm{Cost}_{\mathrm{NFE}}$ – обчислювальна вартість одного виклику нейронної мережі. Це ресурси, витрачені на один прямий прохід базової моделі для оцінки шуму або прогнозування чистого стану на одному конкретному часовому кроці.

Ця вимога ітеративності різко контрастує з ітераційними дискримінативними методами, такими як MICE, які покладаються на наступну кількість операцій (2.11):

$$\mathcal{O}\big(I \cdot D \cdot f(N, D)\big), \tag{2.11}$$

де $I$ – кількість ітерацій;

$D$ – кількість ознак;

$N$ – кількість прикладів;

$f(N, D)$ – обчислювальна вартість одного припасування традиційної регресії.

Незважаючи на те, що ця вартість залежить від розміру вибірки, у межах експериментальних режимів, оцінених у цій роботі, стандартне статистичне припасування залишається обчислювально на порядки дешевшим, ніж $\mathcal{O}(T)$ послідовних нейронних обчислень, необхідних для траєкторії дифузії.

І хоча табличні рядки мають менший обсяг пам’яті порівняно з зображеннями високої роздільної здатності – що дозволяє використовувати значно більші розміри пакетів (batch sizes) $B$ під час виведення – архітектурні упередження, необхідні для обробки табличних даних, вносять власні обчислювальні затримки. Оскільки табличні змінні є різнорідними і не мають просторової структури (як пікселі на зображенні), вони вимагають щільно пов’язаних архітектур, таких як багатошарові перцептрони (MLP) або трансформери. У архітектурах трансформерів, що розглядають табличні ознаки як послідовність довжиною $D$, обчислювальна складність включає як механізм самоуваги $\mathcal{O}(D^2 \cdot d_{model})$, так і повнозв’язну мережу (FFN) $\mathcal{O}(D \cdot d_{model}^2)$, де $d_{model}$ – розмірність ембедингу. Для

типових табличних наборів даних, де $D$ невелике (10–100 ознак), а $d_{model}$ велике, термін FFN часто домінує в обчислювальній вартості на шар. Отже, хоча пропускна здатність пам’яті є високою, послідовне обчислення цих параметрично щільних операцій протягом $T$ кроків створює значну кумулятивну затримку.

У практичних застосуваннях, таких як динамічна агрегація фінансових ризиків на рівні великих когортних груп, онлайн-моніторинг транзакцій на предмет шахрайства або функціонування систем медичного сортування в режимі реального часу, така експоненційна затримка виведення створює критичні бар’єри для розгортання у виробничих середовищах. У високонавантажених системах, де алгоритмічні рішення часто мають ухвалюватися за мілісекунди, необхідність витрачати десятки секунд на послідовне проходження тисячі кроків дифузії ($T = 1000$) робить класичні генеративні моделі операційно непридатними. Таким чином, фундаментальний методологічний виклик для сучасних досліджень полягає у подоланні цього обчислювального вузького місця: необхідно знайти механізм збереження високої дистрибутивної точності та здатності до охоплення складних багатомодальних розподілів, притаманних дифузійним моделям, з одночасним досягненням високої пропускної здатності та радикальним зниженням затримки виведення, що є беззаперечною вимогою для інтеграції генеративного моделювання в системи, чутливі до часу реакції.

# 3 РОЗРОБКА МЕТОДУ LIR

## 3.1 Загальна формалізація принципу роботи методу

Для вирішення фундаментального архітектурного конфлікту між дистрибутивною точністю, охопленням мод та затримкою виведення, у цій роботі пропонується архітектура «Латентного ітераційного вдосконалення» (Latent Iterative Refinement, LIR) – гібридна генеративна модель, розроблена спеціально для імпутації гетерогенних табличних даних. LIR базується на процедурах стохастичного вдосконалення методом стохастичного диференціального редагування SDEdit [34]. Шляхом синтезу можливостей швидкої ініціалізації прогнозних моделей з ітераційною потужністю усунення шуму (denoising) дифузії, LIR створює фреймворк, який розглядає імпутацію як завдання вдосконалення, а не як проблему генерації «з нуля». Традиційні повнотраєкторні дифузійні моделі витрачають левову частку обчислювальних ресурсів на ранні кроки процесу, формуючи глобальну макроструктуру вибірки з ізотропного шуму. Натомість у задачі імпутації наявні спостережувані ознаки вже задають суворі геометричні обмеження на латентний многовид, тому починати генерацію з повністю випадкового стану є обчислювально надлишковим.

Це дозволяє розширити гібридні генеративно-дискримінативні підходи на область табличної імпутації через принципову формалізацію. Така синергія гарантує, що швидка, але потенційно згладжена детермінована оцінка отримує строго відкалібрований стохастичний імпульс, необхідний для виходу з міжмодальних областей низької щільності у валідну моду реальних даних. LIR функціонує у три етапи: латентна проекція (Етап I), ініціалізація «теплим стартом» (Етап II) та стохастичне вдосконалення (Етап III), які сукупно визначають три основні компоненти: детермінований теплий старт, збурення з корекцією концентрації та адаптивне планування на основі невизначеності (рисунок 3.1).

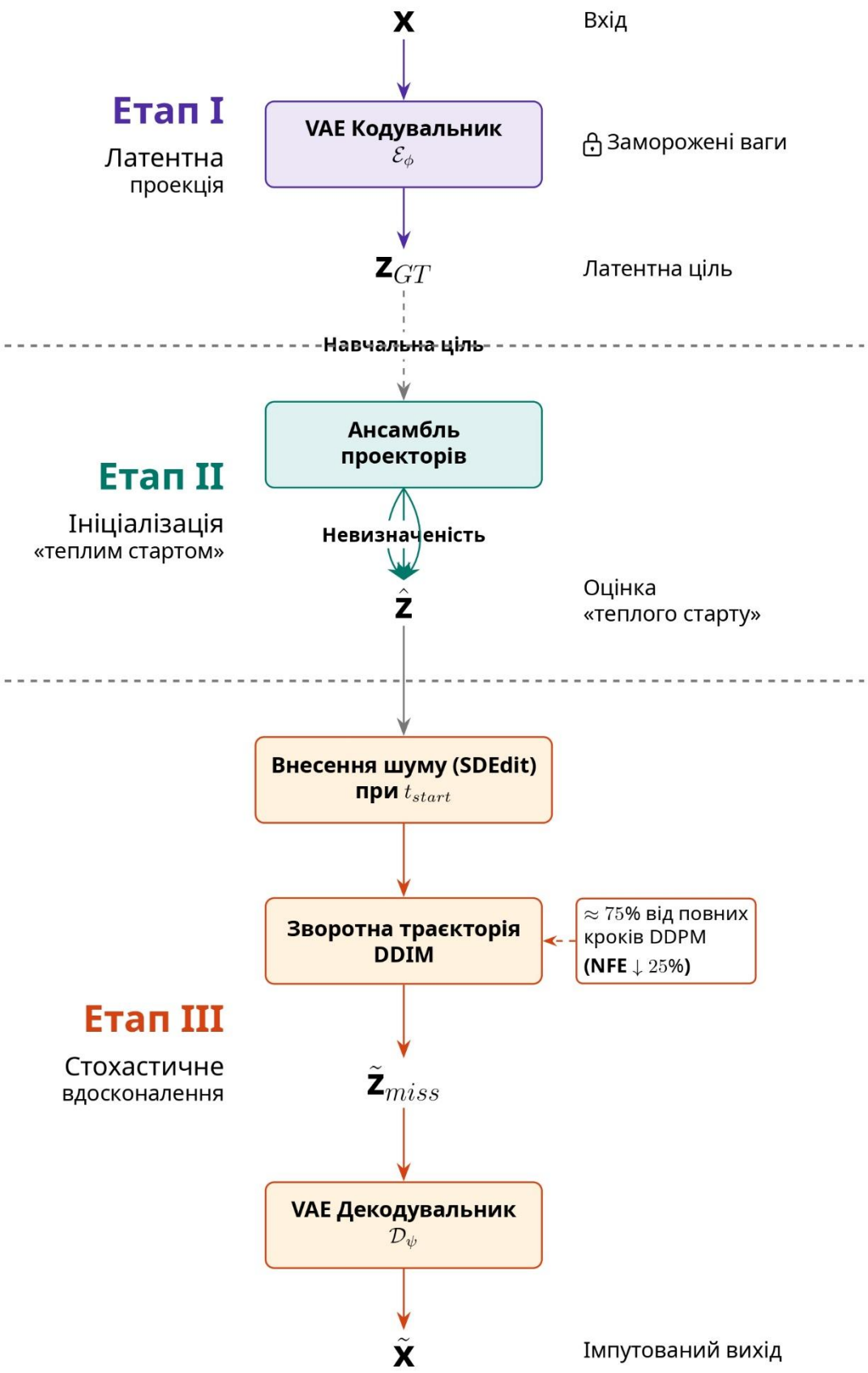



Рисунок 3.1 – Трьохетапний пайплайн LIR

Детермінована ініціалізація за принципом «теплого старту» є ключовим компонентом, що усуває обчислювальну неефективність стандартних повнотраєкторних дифузійних моделей. Традиційні моделі DDPM ініціалізують зворотний процес у момент часу $T$, вимагаючи проходження всієї траєкторії від чистого ізотропного шуму $\mathcal{N}(0, I)$. Для задачі імпутації ця рання фаза є обчислювально надлишковою: вона змушує модель заново генерувати глобальну структурну інформацію, яка вже частково міститься у спостережуваних ознаках $x_{obs}$.

Другим фундаментальним компонентом фреймворку є збурення з корекцією концентрації. Ця математична адаптація зумовлена специфічною геометрією табличних латентних просторів: оскільки такі дані зазвичай формують низькорозмірні многовиди, стандартні межі збурення виявляються неефективними через слабку концентрацію міри.

Хоча процес збурення базується на ізотропному гаусовому ядрі, LIR вводить спеціальний коефіцієнт корекції $\gamma_{conc}$, виведений із властивостей розподілу $\chi^2$. Цей параметр встановлює «умову ефективного збурення», динамічно адаптуючи поріг початкового часового кроку $t_{start}$ відповідно до локального співвідношення сигнал / шум (SNR).

Останнім основним компонентом є адаптивне планування на основі епістемічної невизначеності. Оскільки пропущені значення вимагають різного ступеня вдосконалення, LIR використовує дисперсію ансамблю проекторів для динамічного визначення оптимальної точки входу $t_{start}$ для кожного зразка. Складніші випадки (з високою невизначеністю) отримують більше значення $t_{start}$ для глибшого ітераційного дослідження.

Подальший зворотний процес використовує детермінований семплер DDIM. Уникаючи стохастичних оновлень Ланжевена, DDIM забезпечує відтворювану збіжність за фіксованого початкового шуму в точці $t_{start}$, що дозволяє здійснювати ефективну та контрольовану множинну імпутацію.

Завдяки цьому синтезу LIR суттєво зменшує затримку виведення. Очікується скорочення кількості обчислень нейронних функцій (NFE)

приблизно на 30% порівняно з моделями повної траєкторії (наприклад, TabDDPM при $T = 1000$). Із середньою точкою входу $t_{start} \approx 0.70T$ та вкрай низькою обчислювальною вартістю ансамблю, цей фреймворк потенційно робить генеративну імпутацію життєздатною навіть за жорстких обмежень ресурсів.

### 3.1.1 Формалізація задачі генеративної імпутації у латентному просторі

Нехай задана неповна матриця гетерогенних даних $X \in R^{N \times D}$, що містить $N$ спостережень та $D$ вимірів. Простір ознак $\mathcal{X}$ структурно розбивається на дві підмножини: $X_{num}$ (неперервні ознаки) та $X_{cat}$ (категоріальні ознаки). Наявність даних формалізується бінарною маскою $M \in \{0,1\}^{N \times D}$, де $m_{i,j} = 1$ означає спостережуване значення, а $m_{i,j} = 0$ – пропущене. Відповідно, кожне спостереження $x_i$ розбивається на спостережувану $x_{obs}$ та відсутню $x_{miss}$ частини.

Для забезпечення математичної узгодженості поелементних операцій та обходу проблеми оптимізації дискретно-неперервних змінних, визначається функція нелінійного латентного вкладення (embedding function) (3.1):

$$z = \mathcal{E}(x_{in}), \quad \mathcal{E}: \mathcal{X} \to \mathcal{Z} \subset R^{d_z}, \tag{3.1}$$

де $\mathcal{E}$ – кодувальник варіаційного автоенкодера (VAE);

$\mathcal{X}$ – глобальний гетерогенний простір ознак розмірності $D$;

$\mathcal{Z}$ – єдиний неперервний латентний простір, у який проектуються дані;

$R^{d_z}$ – $d_z$-вимірний простір дійсних чисел, підмножиною якого є вивчений латентний многовид $\mathcal{Z}$.

Для усунення семантичної неоднозначності відсутніх значень (так званої «проблеми нуля») застосовується стратегія маскування

конкатенацією (рисунок 3.2). Кодувальник отримує розширений вектор $x_{in}$, що дозволяє нейронній мережі чітко розрізняти справжні нулі та заповнені пропуски (3.2):

$$x_{in} = (x \odot m) \oplus m, \tag{3.2}$$

де $\odot$ – добуток Адамара (поелементне множення);

$\oplus$ – оператор конкатенації векторів.

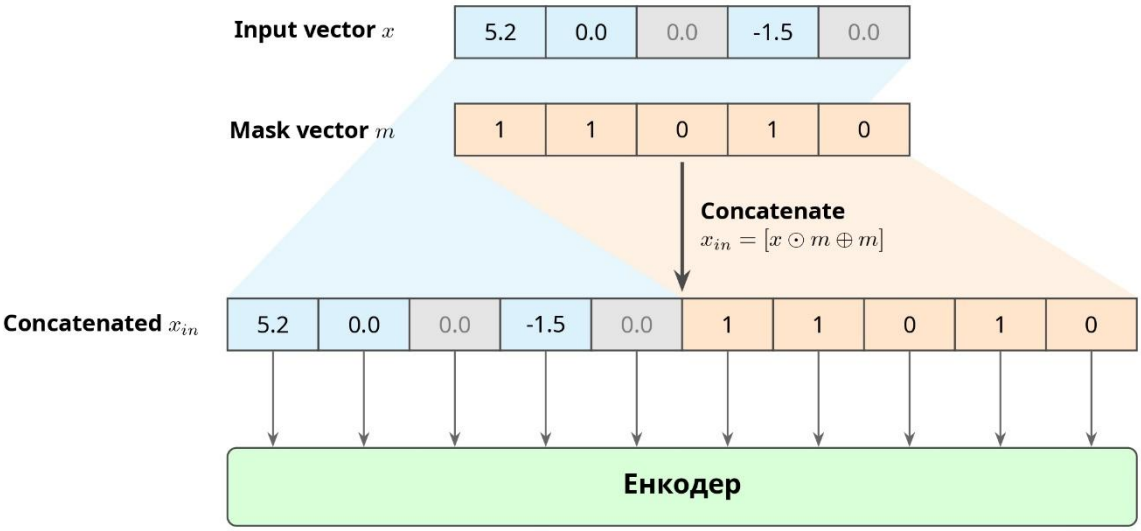



Рисунок 3.2 – Стратегія маскування конкатенацією

Розроблена методологія LIR передбачає строго послідовне навчання: VAE повністю навчається на Етапі I на підмножині повних даних, після чого ваги кодувальника $\mathcal{E}_{\varphi}$ заморожуються. Це забезпечує стабільні цільові значення латентних координат $z_{GT}$ для навчання ансамблю детермінованих проекторів на Етапі II.

Механізм відсутності даних формалізується через умовний розподіл $p(M|X_{obs}, X_{miss})$. За умови виконання припущення про випадковість пропусків (MCAR) та незалежність параметрів, механізм утворення пропусків вважається ігнорованим. Це дозволяє зосередити

задачу на максимізації маргінальної щільності розподілу ймовірностей спостережуваних даних (3.3):

$$p(X_{obs};\theta) = \int p(X_{obs}, X_{miss};\theta) dX_{miss}, \quad (3.3)$$

де $\theta$ – параметри генеративної моделі;

$p(X_{obs}, X_{miss};\theta)$ – спільний розподіл ймовірностей («ідеальний» повний набір даних).

Оскільки у глибоких моделях обчислення цього інтеграла є аналітично недосяжним, ціллю генеративної імпутації стає не пошук точкового оптимуму, а отримання достовірної вибірки з умовного апостеріорного розподілу $p(z_{miss}|z_{obs}, m)$.

Традиційні дискримінативні моделі (наприклад, MissForest), що оптимізуються за функцією втрат $L_2$, математично зводяться до оцінки умовного математичного сподівання (3.4):

$$E[z_{miss}|z_{obs}] = \int z_{miss}\ p(z_{miss}|z_{obs}) dz_{miss}, \quad (3.4)$$

де $z_{miss}$ – вектор відсутніх значень у латентному просторі, які модель намагається відновити (імпутувати);

$z_{obs}$ – вектор спостережуваних значень, на основі яких будується припущення про пропущені дані;

$p(z_{miss}|z_{obs})$ – умовний апостеріорний розподіл.

На табличних многовидах із роз'єднаними областями високої щільності оптимізація за рівнянням 3.4 призводить до патології усереднення за модами. Якщо розподіл $p(z_{miss}|z_{obs})$ є мультимодальним, точкове середнє неминуче потрапляє в простір між модами, що спричиняє колапс дисперсії та спотворення кореляційних зв'язків між ознаками.

Генеративне score-based моделювання у фреймворку LIR вирішує цю проблему шляхом вивчення векторного поля – умовної score-функції (3.5):

$$s_\theta(z_t, t) \approx \nabla_{z_t} \log p_t\,(z_t|z_{obs}), \tag{3.5}$$

де $z_t$ – поточний стан латентного вектора на часовому кроці дифузії $t$;

$p_t(z_t|z_{obs})$ – згладжений маргінальний розподіл на кроці $t$.

Базуючись на принципах SDEdit, фреймворк здійснює явне внесення стохастичної енергії у детерміновану оцінку $E[z_{miss}|z_{obs}]$ на адаптивно розрахованому кроці $t_{start}$. Це забезпечує вихід оцінки за межі помилкового середнього та збіжність до однієї з валідних мод.

Проте, застосування стандартних графіків дифузійного шуму до низькорозмірних латентних просторів ($d_z \ll 100$) вносить критичну патологію: слабка концентрація міри $\chi^2$-розподілу означає, що дисперсія реалізованої енергії шуму є занадто високою. Для компенсації дисперсії в та збереження гарантій того, що реалізована стохастична енергія подолає помилку детермінованого проектора $e$, вводиться структурне обмеження – межа збурення з корекцією концентрації (3.6):

$$SNR(t_{start}) < \frac{\gamma_{conc} d_z}{||e||^2}, \tag{3.6}$$

де $SNR(t_{start})$ – відношення сигнал/шум на початковому кроці зворотного процесу $t_{start}$;

$\gamma_{conc}$ – коефіцієнт корекції концентрації, виведений із властивостей $\chi^2$-розподілу;

$||e||^2$ – похибка оцінки детермінованого проектора, яка на практиці апроксимується через епістемічну невизначеність ансамблю.

### 3.1.2 Гетерогенність та розподіли з «важкими хвостами»

Табличні дані мають унікальну структуру, яка суттєво відрізняється від зображень чи тексту. На відміну від зображень, де існують локальні

просторові кореляції, таблиці є фундаментально гетерогенними: вони поєднують числові дані, порядкові змінні та розріджені категоріальні ознаки високої кардинальності. Це вимагає спеціальних методів обробки для узгодження неперервних та дискретних ознак.

Припустимо, що неперервна змінна $X$ підпорядковується степеневому розподілу, який характеризується функцією виживання $P(X > x) \sim x^{-\alpha}$ для великих $x$, де $\alpha > 2$ є показником «хвоста», що забезпечує скінченність математичного сподівання та дисперсії. Застосування стандартного $Z$-перетворення $Z = \frac{(X-\mu)}{\sigma}$ дає змінну $Z$, яка зберігає початковий показник «хвоста» $\alpha$.

Підставивши $Z = \frac{(X-\mu)}{\sigma}$ у ймовірність «хвоста», маємо $P(Z > z) = P(X > \sigma z + \mu)$. При $z \to \infty$ значення $\sigma z$ домінує над $\mu$, отже $P(Z > z) \sim \sigma^{-\alpha} z^{-\alpha}$. Це доводить, що лінійна стандартизація не змінює показник $\alpha$, залишаючи структуру некомпактного носія вихідного простору незмінною.

Збереження екстремальних викидів створює проблему для варіаційних автоенкодерів (VAE). Оскільки VAE зазвичай використовує стандартний гаусів апріорний розподіл $p(z) = \mathcal{N}(0, I)$, нестиснуті викиди з «важкими хвостами» спричиняють невідповідність між апріорним та апостеріорним розподілами, що призводить до градієнтної нестабільності та колапсу апостеріорного розподілу [35], [36]. У результаті модель не може точно реконструювати критично важливі значення викидів.

Для вирішення цієї проблеми фреймворк LIR використовує квантильний перетворювач (Quantile Transformer) як стандарт попередньої обробки. Це нелінійне перетворення $\Psi_j(x)$ відображає емпіричну функцію розподілу (CDF) кожної числової ознаки у стандартний нормальний розподіл (3.7):

$$\Psi(x) = \Phi^{-1}\left(F_{\text{empirical}}(x)\right), \tag{3.7}$$

де $x$ – вихідне значення числової ознаки в оригінальному просторі даних;

$F_{empirical}$ – емпірична функція розподілу (CDF) навчальних даних, що визначає ймовірність того, що ознака прийме значення, менше або рівне $x$;

$\Phi^{-1}$ – пробіт-функція (обернена гаусова CDF).

$\Psi(x)$ стискає екстремальні квантилі, узгоджуючи маргінальні розподіли даних із гаусовим апріорним розподілом VAE (рисунок 3.3).

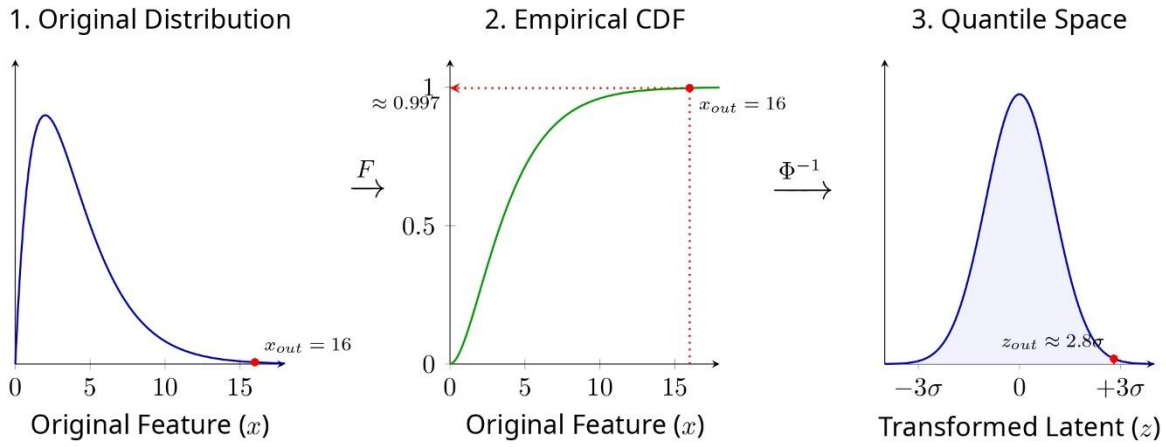



Рисунок 3.3 – Зміна розподілу після квантильного перетворювача

Таке узгодження мінімізує дистрибутивну невідповідність між апріорним та апостеріорним розподілами та полегшує подальше навчання неперервного латентного представлення $z \in R^{d_z}$, стабілізуючи оптимізацію дифузійних моделей, що припускають гаусову динаміку.

Оскільки перетворення є нелінійним і базується на рангах, пряма інтерпретованість значень при застосуванні квантильного перетворювача втрачається, тому на етапі декодування обов’язково застосовується обернене перетворення $\Psi^{-1}$. Обробка неповних даних виконується через конкатенацію вектора та маски (рівняння 3.2), що детальніше описано в архітектурі VAE у розділі 3.3.

## 3.2 Формулювання VP-SDE

### 3.2.1 Прямий процес та маргінальний розподіл у замкненій формі

Фреймворк LIR працює у неперервному латентному просторі VAE ($R^{d_z}$). У цьому просторі прямий дифузійний процес поступово руйнує структурну інформацію проектованих даних, перетворюючи кореляції між ознаками на ізотропний гаусів шум [8], [37]. Для узгодження доменів надалі використовуватиметься заміна $x \leftarrow z$.

Прямий дифузійний процес визначається як ланцюг Маркова, що додає гаусів шум до початкового представлення $z_0$ протягом $T$ кроків. При заданому розкладі дисперсії $\{\beta_t \in (0,1)\}_{t=1}^{T}$ ядро переходу має вигляд (3.8):

$$q(z_t|z_{t-1}) = \mathcal{N}\left(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I\right), \tag{3.8}$$

де $z_t$ – латентний стан на часовому кроці $t$, представлений у неперервному латентному просторі VAE;

$\beta_t$ – гіперпараметр розкладу дисперсії на кроці $t$, який контролює інтенсивність зашумлення та швидкість перетворення сигналу на ізотопний шум;

$\sqrt{1-\beta_t} z_{t-1}$ – математичне сподівання розподілу, що забезпечує дрейф сигналу до нуля;

$\beta_t I$ – коваріаційна матриця, де $I$ – одинична матриця, що вказує на внесення ізотропного шуму.

Для збіжності маргінального розподілу до $\mathcal{N}(0, I)$ необхідно, щоб кумулятивний добуток коефіцієнтів $\overline{\alpha_T} = \prod_{s=1}^{T}(1-\beta_s)$ наближався до нуля [37].

Важливою перевагою DDPM є можливість отримати стан $z_t$ на будь-якому кроці $t$ без ітераційного обчислення проміжних етапів.

Використовуючи репараметризацію, $z_t$ можна виразити через $z_0$ та шум $\epsilon \sim \mathcal{N}(0, I)$ (3.9):

$$q(z_t|z_0) = \mathcal{N}\left(z_t; \sqrt{\overline{\alpha_t}} z_0, (1 - \overline{\alpha_t})I\right), \tag{3.9}$$

де $z_t$ – латентний стан на часовому кроці $t$, представлений у неперервному латентному просторі VAE;

$z_0$ – початкове латентне представлення даних, отримане після кодування вихідних табличних ознак;

$\sqrt{\overline{\alpha_t}} z_0$ – математичне сподівання розподілу, що визначає частку збереженого сигналу від оригінальних даних на кроці t;

$(1 - \overline{\alpha_t})I$ – коваріаційна матриця шуму, де $I$ – одинична матриця, а $(1 - \overline{\alpha_t})$ визначає дисперсію внесеного ізотропного гаусового шуму;

$\overline{\alpha}_t$ – кумулятивний добуток коефіцієнтів $\alpha_s = 1 - \beta_s$ від початкового кроку до поточного моменту $t$.

Цей результат дозволяє ефективно навчати модель, мінімізуючи спрощену цільову функцію [8] (3.10):

$$\mathcal{L}_{sml} = E_{t,\boldsymbol{z}_0,\epsilon}[||\epsilon - \epsilon_\theta(z_t, t)||^2], \tag{3.10}$$

де $E_{t,z_0,\epsilon}[\dots]$ – математичне сподівання за випадковими часовими кроками $t$, початковими латентними станами $z_0$ та векторами шуму $\epsilon$;

$\epsilon$ – ізотропний гаусів шум, вибраний зі стандартного нормального розподілу, який додається до чистих даних під час прямого процесу;

$\epsilon_\theta$ – денойзер, параметризований вагами $\theta$, який прогнозує вектор шуму на основі зашумленого стану $z_t$ та часового кроку $t$;

$z_t$ – зашумлений латентний стан на кроці $t$, що обчислюється через репараметризацію початкового стану $z_0$ та шуму $\epsilon$;

$||\ldots||^2$ – середньоквадратична помилка, що вимірює розбіжність між істинним шумом та прогнозом моделі.

LIR розглядає прямий процес як дискретизацію стохастичного диференціального рівняння, а саме SDE зі збереженням дисперсії (Variance Preserving, VP-SDE) (3.11):

$$dz = -\frac{1}{2}\beta(t)zdt + \sqrt{\beta(t)}dw, \tag{3.11}$$

де $dz$ – нескінченно мала зміна латентного стану $z$ протягом диференціального часового кроку $dt$;

$-\frac{1}{2}\beta(t)zdt$ – термін дрейфу, який забезпечує поступове зміщення значень сигналу до нуля, утримуючи при цьому загальну дисперсію розподілу постійною;

$\beta(t)$ – функція розкладу дисперсії у неперервному часі, яка визначає темп зашумлення даних та перетворення їх на ентропійний стан;

$\sqrt{\beta(t)}dw$ – термін дифузії, що відповідає за внесення випадкового стохастичного шуму в систему;

$dw$ – приріст стандартного вінерівського процесу (білий шум), що моделює випадкові флуктуації протягом кроку $dt$.

Термін $-\frac{1}{2}\beta(t)zdt$ забезпечує дрейф до нуля, а $\sqrt{\beta(t)}dw$ вносить ентропію. Використання SDE дозволяє застосовувати чисельні розв'язувачі високого порядку, що адаптують обчислювальні витрати до складності пропусків у даних.

З погляду теорії інформації, прямий процес монотонно зменшує взаємну інформацію $I(Z_t; Z_0)$. Це можна розуміти як згладжування топології латентного простору через згортку з гаусовим ядром зростаючої дисперсії. У стані з максимальною ентропією ($t = T$) ми отримуємо зручний апріорний розподіл для ініціалізації зворотного процесу.

Траєкторія руйнування інформації залежить від розкладу дисперсії $\{\beta_t\}$. Відношення сигнал/шум (SNR) визначається як [38] (3.12):

$$SNR(t) = \frac{\overline{\alpha_t}}{1-\overline{\alpha_t}}, \tag{3.12}$$

де $\alpha_t$ – кумулятивний добуток коефіцієнтів масштабування шуму, який визначає частку енергії оригінального латентного сигналу, що залишається на кроці $t$;

$1 - \overline{\alpha_t}$ – дисперсія шуму, що відповідає зашумленості латентного стану на даному етапі дифузійного процесу.

У контексті табличних даних, де ознаки можуть мати дуже різну чутливість до шуму, таке стрімке згасання може призвести до передчасної втрати структурної інформації. Як продемонстровано на рисунку 3.4, косинусні або сигмоїдальні розклади часто забезпечують більш стабільну криву згасання SNR, зберігаючи придатний для використання сигнал протягом значно більшої частки прямої траєкторії.

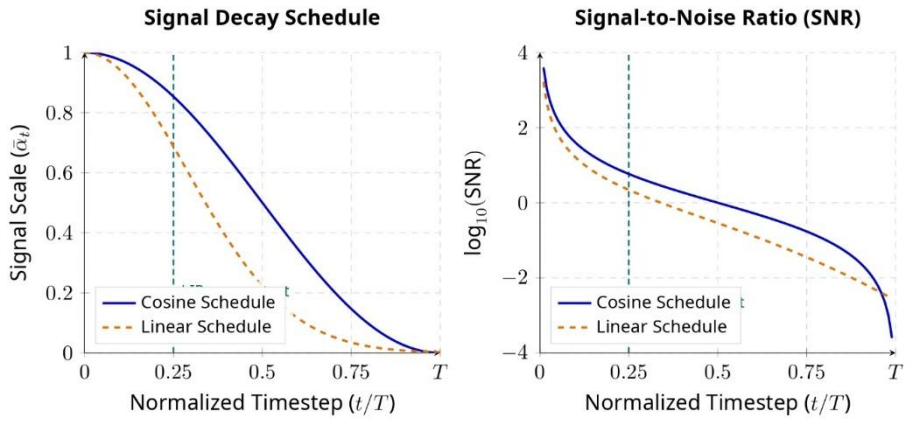



Рисунок 3.4 – Стабільніша крива згасання SNR

Це стабільність є критичною для збереження кореляційної структури між ознаками, закодованої в $z$, на більш тривалий час. Відповідно,

збереження цієї структури виступає як умова, що сприяє ініціалізації «теплого старту»

### 3.2.2 Зворотний процес та узгодження оцінок

Генеративна здатність LIR базується на інверсії динаміки прямої дифузії. Зворотний процес відтворює траєкторію від гаусового шуму назад до низькоентропійного многовиду табличних даних.

Згідно з теоремою про обернення часу [37], [39], зворотний процес для VP-SDE описується рівнянням (3.13):

$$dz_t = \left[-\frac{1}{2}\beta(t)z_t - \beta(t)\nabla_{z_t}\log p_t\,(z_t)\right]dt + \sqrt{\beta(t)}d\overline{w}, \qquad (3.13)$$

де $dz_t$ – диференціальна зміна стану латентного вектора $z_t$ під час зворотного руху за часом від шуму до даних;

$\beta(t)$ – неперервна функція розкладу дисперсії, яка регулює темп дифузії та інтенсивність зашумлення на кожному кроці траєкторії;

$\nabla_{z_t}\log p_t\,(z_t)$ – функція оцінки (score function). Геометрично це векторне поле, що вказує напрямок до областей високої щільності даних. При $t \to 0$ це поле стає складнішим, відображаючи кореляції табличного многовиду;

$dt$ – нескінченно малий крок часу, протягом якого відбувається оновлення стану;

$\sqrt{\beta(t)}d\overline{w}$ – стохастична компонента, де $d\overline{w}$ – вінерівський процес у зворотному часі.

Оскільки аналітична функція оцінки невідома, нейронна мережа $\epsilon_\theta(z_t, t)$ навчається апроксимувати шум за допомогою усіченої цільової функції (рівняння 3.10). Зв’язок між навченим денойзером $\epsilon_\theta$ та функцією оцінки встановлюється через формулу Твіді [40], і для гаусового збурення виконується співвідношення (3.14):

$$\nabla_{z_t} \log p_t\,(z_t) \approx -\frac{\epsilon_\theta(z_t,t)}{\sqrt{1-\overline{\alpha}_t}}, \tag{3.14}$$

де $\epsilon_\theta(z_t, t)$ – навчений денойзер;

$\sqrt{1-\overline{\alpha}_t}$ – нормувальний коефіцієнт, що дорівнює стандартному відхиленню шуму на часовому кроці $t$ відповідно до маргінального розподілу;

$\overline{\alpha}_t$ – кумулятивний добуток коефіцієнтів масштабування шуму, що визначає рівень збереженого сигналу в латентному стані $z_t$.

Це дозволяє використовувати навчену мережу для чисельної симуляції зворотного SDE, де фреймворк LIR використовує детерміноване правило оновлення DDIM (рівняння 3.8) для підвищення стабільності збіжності.

## 3.3 Етап I: VAE змішаного типу для латентної проекції

Основна теоретична та практична складність при застосуванні дифузійних моделей на основі стохастичних диференціальних рівнянь (SDE) до табличних даних полягає в неевклідовій природі вихідного простору ознак $\mathcal{X}$. Табличні спостереження є фундаментально гетерогенними: вони поєднують щільні числові розподіли з «важкими хвостами» та розріджені категоріальні токени. Безпосередня дифузія в такому просторі вимагала б розробки гібридних процесів (наприклад, поєднання гаусової та мультиноміальної дифузії), що суттєво ускладнює архітектуру та збільшує обчислювальні витрати [32].

Фреймворк LIR вирішує цю проблему через використання варіаційного автоенкодера змішаного типу (Mixed-Type VAE) як модуля неперервної латентної проекції. На цьому етапі гібридний вхід відображається в уніфікований диференційовний латентний простір $\mathcal{Z} \subseteq R^{d_z}$. Це дозволяє подальшому дифузійному вдосконаленню працювати

виключно з гаусовою динамікою, що забезпечує скорочення кількості обчислень нейронних функцій (NFE) та пом’якшує проблеми зміщених оцінювачів градієнта, властивих дискретним дифузійним моделям.

### 3.3.1 Процес попередньої обробки: нормалізація, конкатенація, та вирішення «проблеми нуля»

Для забезпечення числової стабільності латентної проекції та мінімізації дистрибутивної невідповідності (prior-posterior mismatch) впроваджено спеціалізований процес попередньої обробки:

– числові ознаки ($x_{num}$): кожна ознака $j$ трансформується за допомогою квантильного перетворювача $\Psi_j$ (рівняння 3.7), який відображає емпіричну функцію розподілу у стандартний нормальний розподіл, мінімізуючи дистрибутивну невідповідність під час навчання. Для гарантування строгої обмеженості входу енкодера застосовується відсікання: $\mathrm{clip}(\Psi(x), -3, 3)$. Такий діапазон відповідає $\approx 99.7\%$ маси стандартного нормального розподілу, що забезпечує ідеальну відповідність апріорному розподілу VAE $p(z) = \mathcal{N}(0, I)$.

– категоріальні ознаки ($x_{cat}$): змінні представляються через пряме кодування (one-hot encoding), де кожна категорія перетворюється на окремий бінарний вектор. Хоча це збільшує розмірність входу, такий підхід дозволяє уникнути штучного нав’язування порядку категоріям. Для наборів даних із надзвичайно високою кардинальністю, VAE виступає як механізм стиснення, що вивчає компактні представлення в межах $R^{d_z}$.

Фундаментальною проблемою табличної імпутації є семантична неоднозначність нуля. У багатьох системах нуль може бути як валідним числовим значенням (наприклад, нульовий баланс), так і технічним заповнювачем для пропущеного запису. Просте заповнення нулями (zero-filling) неспостережуваних ознак призводить до зміщення

латентних представлень, оскільки модель не може розрізнити структурну відсутність та емпіричний нуль.

В архітектурі LIR ця проблема вирішується впровадженням маскування конкатенацією (рівняння 3.2), адаптуючи принципи MIWAE [27]. Це подвоєння розмірності надає моделі явний структурний сигнал, дозволяючи шарам самоуваги ігнорувати шум від відсутніх вимірів та фокусуватися на наявних кореляціях.

### 3.3.2 Архітектура на основі табличного трансформера

На відміну від традиційних VAE, що базуються на MLP, LIR використовує архітектуру трансформера для захоплення складних взаємозалежностей між ознаками (рисунок 3.5). Оскільки табличні дані не мають природного послідовного порядку (як текст), традиційне позиційне кодування є неможливим до використання. Натомість для ідентифікації ознак впроваджено стовпчикові лінійні проекції (feature embeddings).

Кожна ознака (пара з маскованого значення та індикатора пропуску) проектується незалежним лінійним шаром у простір $d_{model} = 128$. Це гарантує, що модель унікально ідентифікує атрибути навіть за однакових числових значень у різних стовпцях:

– енкодер та декодер: складаються з 2 блоків трансформера з 4 «головами» уваги кожен. Використання самоуваги (self-attention) дозволяє моделі динамічно зважувати важливість кожної ознаки при формуванні латентного коду;

– латентний простір: параметризований через вектор середніх $\mu$ та логарифм дисперсії $\log \sigma^2$, що дозволяє використовувати репараметризацію: $z = \mu + \sigma \odot \epsilon, \epsilon \sim \mathcal{N}(0, I)$. Розмірність $d_z = 64$ визначена емпірично як оптимальна для збереження 98% варіативності більшості еталонних наборів даних.

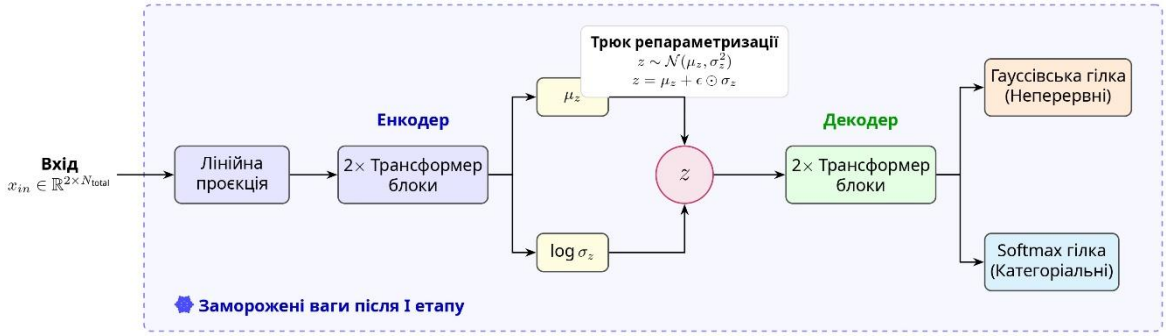



Рисунок 3.5 – Архітектура VAE

Реконструкція змішаних даних потребує спільної функції правдоподібності, яка є сумою гаусової (для числових) та категоріальної (для one-hot) компонент (3.15):

$$\log p_\psi(x|z) = \sum_i \log \mathcal{N}\left(x_{num}^{(i)}; \mu_\psi^{(i)}(z), \sigma_\psi^{(i)}(z)^2\right) + \\ + \sum_j \log \mathrm{Cat}\left(x_{cat}^{(j)}; \pi_\psi^{(j)}(z)\right), \tag{3.15}$$

де $\sum_i \log \mathcal{N}$ – сума логарифмічних правдоподібностей для числових ознак, де кожна ознака $i$ моделюється як незалежний гаусів розподіл;

$x_{num}^{(i)}$ – фактичне значення і-ї числової ознаки після квантильного перетворення;

$\mu_\psi^{(i)}$ – прогнозоване середнє значення для $i$-ї числової ознаки, що обчислюється декодером;

$\sigma_\psi^{(i)}(z)^2$ – прогнозна дисперсія (невизначеність) декодера для $i$-ї числової ознаки;

$\sum_j \log \mathrm{Cat}$ – сума логарифмічних правдоподібностей для категоріальних ознак, що відповідає багатокласовій перехресній ентропії;

$x_{cat}^{(j)}$ – вектор категоріальної ознаки $j$, представлений через one-hot кодування;

$\pi_{\psi}^{(j)}(z)$ – прогнозований вектор ймовірностей категорій для $j$-ї ознаки (вихід Softmax-шару декодера).

Основною проблемою навчання таких моделей є дисбаланс масштабів втрат: категоріальна перехресна ентропія (CE) часто за порядком величини перевищує середньоквадратичну помилку (MSE) числових ознак. Це призводить до того, що модель «ігнорує» числову точність заради покращення категоріальної класифікації.

Для вирішення цього впроваджено гомоскедастичне зважування невизначеності [41]: введено навчальні параметри $\sigma_{num}$ та $\sigma_{cat}$, які автоматично балансують компоненти втрат (3.16):

$$\mathcal{L}_{VAE} = \frac{1}{2\sigma_{num}^2}\mathcal{L}_{MSE} + \frac{1}{\sigma_{cat}^2}\mathcal{L}_{CE} + \log\sigma_{num} + \log\sigma_{cat}, \quad (3.16)$$
$$+\beta D_{KL}\left(q_{\phi}(z|x)|p(z)\right),$$

де $\mathcal{L}_{VAE}$ – загальна функція втрат варіаційного автоенкодера, яка мінімізується під час навчання для отримання стабільного латентного простору.

$\mathcal{L}_{MSE}$ – середньоквадратична помилка (Mean Squared Error) реконструкції числових ознак.

$\mathcal{L}_{CE}$ – категоріальна перехресна ентропія (Cross-Entropy) реконструкції категоріальних ознак.

$\log\sigma_{num}$ та $\log\sigma_{cat}$ – регуляризаційні терміни, що запобігають тривіальному розв’язку, де параметри невизначеності могли б нескінченно зростати для обнулення втрат;

$D_{KL}\left(q_{\phi}(z|x)|p(z)\right)$ – дивергенція Кульбака-Лейблера, яка регуляризує наближений апостеріорний розподіл $q_{\phi}$ відносно апріорного розподілу $p(z)$;

$\beta$ – коефіцієнт ваги KL-регуляризатора, який поступово зростає під час навчання для запобігання колапсу апостеріорного розподілу.

Логарифмічні терміни діють як регуляризатори, що запобігають колапсу дисперсій до нуля. Важливо, що обернені ваги дозволяють мережі автоматично знижувати вагу того завдання реконструкції, яке є за своєю природою більш зашумленим або важчим для вивчення (рисунок 3.6). Наприклад, якщо числова реконструкція виявляється складною, мережа вивчає більше глобальне значення $\sigma_{num}$, тим самим послаблюючи градієнт для $\mathcal{L}_{MSE}$ і запобігаючи дестабілізації навчання. Механізм працює симетрично: вага завдання, яке виявляється складнішим, автоматично знижується, гарантуючи, що латентний простір $\mathcal{Z}$ залишається структурно репрезентативним як для числових, так і для категоріальних розподілів без ручного налаштування.

| **Властивість / Метрика** | **Фіксовані ваги (Наївний підхід)** | **Гомоскедастичне зважування (LIR)** |
|---|---|---|
| **Відносний масштаб функцій втрат** | CE ($\sim 85\%$) MSE | CE ($\sim 50\%$) MSE ($\sim 50\%$) |
| **Напрямок градієнта** | Сильно зміщений у бік категоріальних ознак через велику дисперсію логітів. | Динамічно збалансований між усіма типами ознак (числовими та категоріальними). |
| **Відновлення числових даних ($\mathcal{L}_{\text{MSE}}$)** | **Значна деградація** (ігнорується оптимізатором через відносно малий масштаб). | **Висока точність** (компонента адаптивно масштабується параметром $\sigma_{\text{num}}$). |
| **Категоріальна точність ($\mathcal{L}_{\text{CE}}$)** | Висока (швидко збігається, оскільки повністю домінує в оптимізації). | Висока (зберігається без перетягування градієнта від інших гілок). |
| **Стабільність навчання** | Нестабільна: ризик колапсу мод або ігнорування складних числових патернів. | Стабільна: синхронізована збіжність обох гілок завдяки Рівнянню 3.16. |

Рисунок 3.6 – Порівняння вирішення дисбалансу масштабів втрат

LIR-VAE навчається виключно на повних записах навчального набору. Це критично для того, щоб латентний простір кодував «чистий» многовид даних, вільний від шуму, спричиненого пропусками.

## 3.4 SDEdit та невідповідність многовидів

### 3.4.1 Визначення латентних носіїв

Перехід від детермінованої проекції, встановленої на Етапі I, до ітераційного стохастичного вдосконалення на Етапі III потребує глибокого математичного обґрунтування механізму їх поєднання. Хоча варіаційний автоенкодер (VAE) забезпечує неперервну латентну систему координат, ансамбль проекторів генерує лише точкову оцінку $z_{proj}$ відсутніх значень. Ці оцінки за своєю природою знаходяться на геометричній структурі, яка фундаментально відрізняється від носія, необхідного для деноїзера на основі функції оцінки (рисунок 3.7).

Для аналізу причин невдачі прямого усунення шуму для точкових оцінок необхідно визначити точні геометричні носії відповідних латентних станів у просторі $R^{d_z}$.

Середній многовид ($M_{mean}$): $M_{mean} \subset \mathcal{Z}$ було визначено як образ відображення умовного математичного сподівання $f(x_{obs}, m) = E[z|x_{obs}, m]$ над набором даних. Оскільки оцінювачі, що мінімізують $L_2$-норму, виконують усереднення за апостеріорним розподілом, $M_{mean}$ має внутрішню розмірність строго нижчу за розмірність охоплюючого латентного простору $d_z$ [25]. З погляду теорії міри, була використана pushforward measure $\mu_{mean} = f(P(x_{obs}, m))$ для представлення розподілу детермінованих оцінок середнього многовиду.

Зашумлений многовид ($M_{noisy}(t)$): $M_{noisy}(t) \subseteq R^{d_z}$ було визначено як носій маргінального розподілу $p_t(z)$ під дією прямого VP-SDE на часовому кроці $t > 0$ [37]. Оскільки $p_t(z)$ є результатом згортки розподілу даних із

гаусовим ядром переходу $\mathcal{N}\left(z_t; \sqrt{\overline{\alpha_t}} z_0, (1-\overline{\alpha_t})I\right)$, він має повну розмірність $d_z$ та ненульову щільність усюди в просторі $R^{d_z}$.

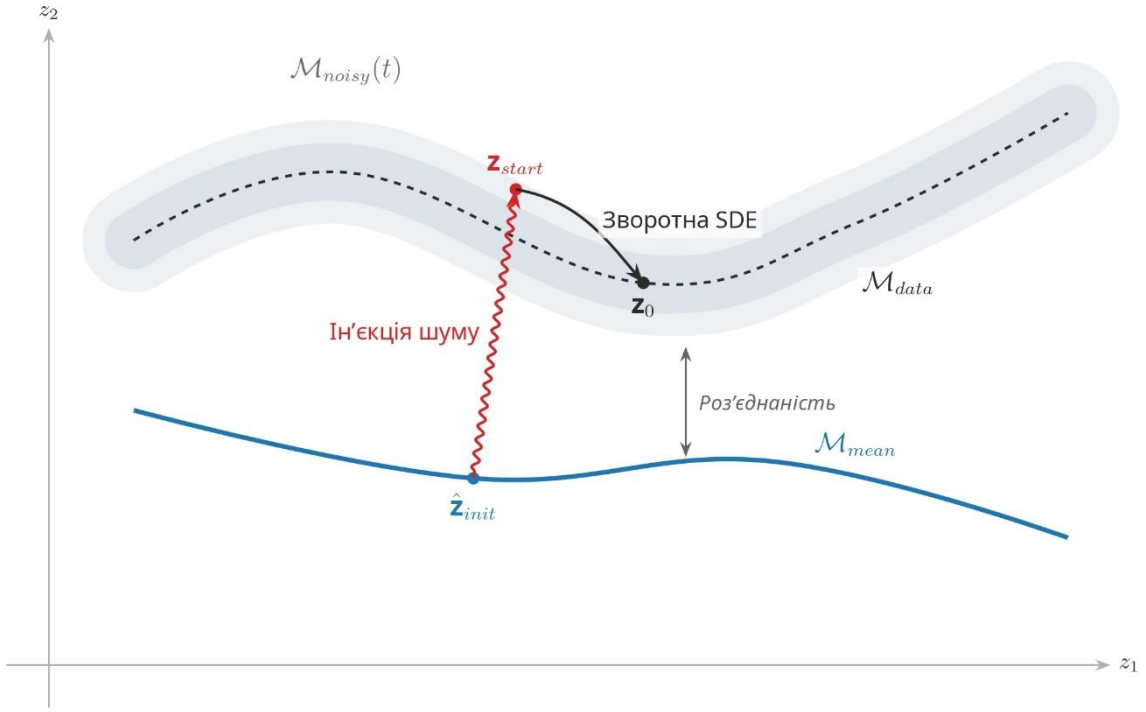



Рисунок 3.7 – Візуалізації невідповідності многовидів

Для доведення диз'юнктності многовидів використовується сепарація Вассерштейна [30]: нехай $\mu_{mean}$ є мірою ймовірності на низьковимірному многовиді $M_{mean}$, а $p_t(z)$ – маргінальною мірою на многовиді $M_{noisy}(t)$. Тоді відстань Вассерштейна $W_2(\mu_{mean}, p_t)$ є строго обмеженою знизу від нуля для всіх $t > 0$.

Як наслідок, оскільки множина $M_{mean}$ має внутрішню розмірність меншу за $d_z$, вона є множиною Лебегової міри нуль у $R^{d_z}$. Оскільки маргінальний розподіл $p_t(z)$ є абсолютно неперервним щодо міри Лебега для всіх $t > 0$, ймовірність того, що зашумлений зразок випадково потрапить на середній многовид, дорівнює нулю: $P(z \in M_{mean} \mid z \sim p_t) = 0$. Таким чином, $M_{mean} \cap M_{noisy}(t) = \varnothing$ майже напевно.

Цей теоретичний результат пояснює, чому пряме застосування денойзера до $z_{proj}$ спричиняє дистрибутивний зсув. Нейронна мережа оцінок $\epsilon_\theta$ навчається мінімізувати дивергенцію Фішера виключно в областях із високою ймовірнісною масою прямого процесу. У області поза носієм $M_{mean}$ поведінка мережі є непередбачуваною [34]. Емпірично це призводить до генерації «числового галюцинування» – високочастотних артефактів, які не мають семантичного зв'язку зі справжнім многовидом даних.

### 3.4.2 Механізм SDEdit: пряме стохастичне збурення, стохастична ініціалізація та подолання «патології усереднення за модами»

Для подолання геометричного розриву фреймворк LIR використовує механізм SDEdit [34]. Замість того, щоб розглядати $z_{proj}$ як фінальну точку, LIR трактує її як початкову умову, яку необхідно «перенести» на носій зашумленого многовиду. Це досягається шляхом просування $z_{proj}$ вперед до проміжного часового кроку $t_{start}$ за допомогою маргінального ядра прямого процесу (рівняння 3.8). Ця трансформація гарантує, що вхідні дані для зворотного SDE знаходяться саме в межах носія $M_{noisy}(t_{start})$, де функція оцінки є добре визначеною та навченою.

Вибір часового кроку $t_{start}$ є критичним гіперпараметром, що визначає баланс між двома силами [34]:

– достовірність: наближення до початкової проекції $z_{proj}$ (чим менше $t_{start}$, тим ближче результат до оцінки ансамблю);

– реалізм: відповідність генеративному апріорному розподілу даних (чим більше $t_{start}$, тим більше кроків дифузії має модель для виправлення помилок проектора).

В LIR параметр $t_{start}$ визначається адаптивно на основі епістемічної невизначеності ансамблю $\sigma^2_{ensemble}$ за допомогою каліброваної функції $g(\cdot)$.

Необхідність збурення виходить за межі геометричного узгодження – воно є фундаментальним механізмом подолання патології усереднення за модами. Пряме збурення надає стохастичну енергію, необхідну для витіснення частинки з цих областей низької щільності. Під дією градієнтного потоку, індукованого полем оцінок $\nabla_z \log p_{t_{start}}(z)$, частинка зміщується в басейн тяжіння (basin of attraction) однієї з високоймовірних мод [37]. Цей процес можна порівняти з порушенням симетрії в термодинаміці: шум «підштовхує» систему до одного з валідних станів, які модель усереднила під час навчання проектора.

Після того, як частинка була перенесена на зашумлений многовид і стохастично зміщена до валідної моди, фреймворк переходить до фази точного вдосконалення – де використовує неявні дифузійні моделі (DDIM) [33]. На відміну від стандартного анцестрального семплювання [8], DDIM дозволяє рухатися по детермінованій траєкторії, встановлюючи параметр стохастичності $\sigma_t = 0$ (3.17):

$$z_{t-1} = \sqrt{\overline{\alpha_{t-1}}} \underbrace{\left( \frac{z_t - \sqrt{1-\overline{\alpha_t}}\epsilon_\theta(z_t, t)}{\sqrt{\overline{\alpha_t}}} \right)}_{\text{прогноз } z_0} + \underbrace{\sqrt{1-\overline{\alpha_{t-1}}}\epsilon_\theta(z_t, t)}_{\text{напрямок до } z_t}, \qquad (3.17)$$

де $z_{t-1}$ – наступний латентний стан у зворотній траєкторії, що відповідає меншому рівню шуму;

$\sqrt{\overline{\alpha_{t-1}}}$ – коефіцієнт масштабування сигналу для кроку $t-1$, що визначає частку чистого прогнозу $z_0$ у фінальному результаті;

$z_t$ – поточний латентний стан на часовому кроці $t$;

$\epsilon_\theta(z_t, t)$ – передбачений нейронною мережею шум, який використовується як для реконструкції $z_0$, так і для визначення напрямку руху в латентному просторі;

$\sqrt{1-\overline{\alpha_{t-1}}}$ – вага компоненти шуму для наступного кроку, яка при $t = 0$ забезпечує детермінованість переходу.

Така конфігурація забезпечує стабільну збіжність до обраної моди без ризику подальшої стохастичної дисперсії, яка могла б знову виштовхнути частинку в область низької щільності.

## 3.5 Етап II: ансамбль проекторів для ініціалізації «теплого старту»

Хоча необхідність ітераційного відновлення глобальної структури даних із ізотропного гаусового шуму є стандартною для безумовної генерації у дифузійних ймовірнісним моделях, вона виявляється надлишковою в задачах табличної імпутації. Наявність часткових спостережень $x_{obs}$ вже суттєво обмежує допустимий носій цільового многовиду даних, роблячи початкові кроки зворотної дифузії (режим високого шуму, $t \approx T$) обчислювально неефективними [8].

Етап II фреймворку LIR переводить генеративну парадигму з «холодного старту» (від шуму) на пряме відображення часткових спостережень у латентну оцінку всередині неперервного простору $R^{d_z}$. Цей розділ описує архітектуру ансамблю проекторів та механізм адаптивного планування, який дозволяє динамічно визначати оптимальну точку входу в дифузійну траєкторію.

Для забезпечення миттєвої ініціалізації без послідовних затримок використовується ансамбль із $K = 5$ незалежних багатошарових перцептронів із залишковими зв'язками (ResMLP). Кожен член ансамблю $P_{\theta_k}$ розроблений як легкий апроксиматор умовного математичного сподівання в латентному просторі.

Кожен проектор складається з трьох послідовних блоків ResMLP. Структура блоку включає два лінійні шари з функцією активації ReLU та залишковий зв'язок, що додає вхід блоку до його виходу. Прихована ширина шарів становить 256 нейронів, що забезпечує достатню ємність для моделювання нелінійних залежностей при збереженні низької затримки

виведення. Фінальний шар здійснює проекцію в латентну розмірність $d_z = 64$.

Вхідним сигналом для ансамблю є конкатенований вектор (рівняння 3.2), де пропущені значення заповнені нулями. Навчання проводиться за допомогою «замороженого» енкодера VAE, отриманого на Етапі I. Використання фіксованих ваг енкодера є критичним: це гарантує, що проектори навчаються відображати пошкоджені дані саме в той стабільний латентний многовид, який був оптимізований для дифузійного деноїзера (рисунок 3.8) [42].

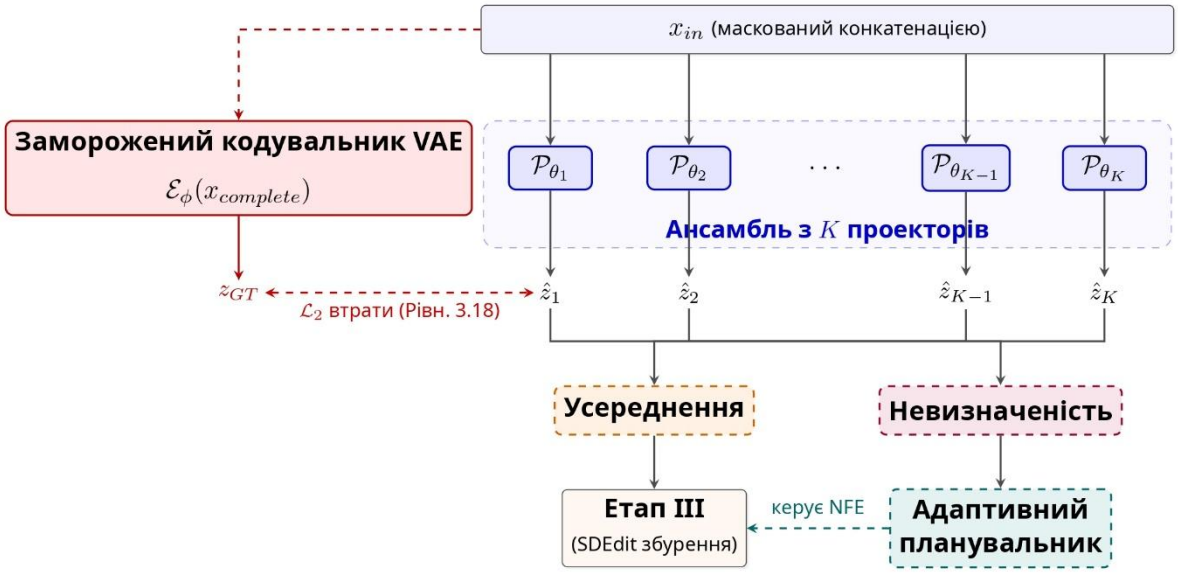



Рисунок 3.8 – Архітектура ансамблю проекторів

Оптимізація параметрів $\theta_k$ здійснюється за критерієм $L_2$ (3.18):

$$\mathcal{L}_{proj} = \frac{1}{K}\sum_{k=1}^{K} | P_{\theta_k}(x_{in}) - z_{GT}|^2, \quad (3.18)$$

де $K$ – кількість незалежних проекторів в ансамблі;

$P_{\theta_k}(x_{in})$ – вихід $k$-го проектора;

$x_{in}$ – пошкоджений вхідний вектор, де пропущені значення замасковані;

$z_{GT}$ – істинний латентний код, отриманий шляхом пропуску повного рядка даних через «заморожений» енкодер VAE, навчений на Етапі I.

Теоретично, мінімізація $L_2$-норми змушує модель сходитися до умовного математичного сподівання $E[z|x_{in}]$. Хоча це «середнє за модами» значення може бути дистрибутивно неточним (як доведено у розділі 3.4), воно слугує ідеальною детермінованою основою для подальшого вдосконалення.

Використання глибокого ансамблю замість однієї моделі дозволяє оцінити епістемічну невизначеність – міру невпевненості моделі у власних прогнозах. У контексті табличної імпутації висока дисперсія між членами ансамблю вказує на області, де відображення $x_{in} \to z$ є неоднозначним (наприклад, через мультимодальність апостеріорного розподілу) [43].

Для кожного зразка обчислюються наступні метрики:

– середнє ансамблю (guide): $z_{guide} = \frac{1}{K}\sum_{k=1}^{K} \widehat{z_k}$;

– дисперсія ансамблю: $\sigma_{ens}^2 = \frac{1}{K-1}\sum_{k=1}^{K}\left(\widehat{z_k} - z_{guide}\right)^2$;

– скалярна невизначеність: $U(x_{in}) = \frac{1}{d_z}\mathrm{Tr}\left(\mathrm{diag}(\sigma_{ens}^2)\right)$.

Ця метрика $U$ корелює з фактичною похибкою реконструкції. Вона дозволяє моделі ідентифікувати «складні» випадки імпутації, де початкова оцінка потребує суттєвого стохастичного коригування.

Ключовою інновацією LIR є адаптивний розклад початкового кроку, керований невизначеністю $U$. Початковий крок дифузії $t_{start}$ визначається через каліброване відображення (3.19):

$$t_{start} = \mathrm{round}\left(T \times g\left(\widetilde{U}(x_{in})\right)\right), \tag{3.19}$$

де $\widetilde{U}$ – нормалізована метрика невизначеності;

$g(\cdot)$ – функція, що відображає впевненість ансамблю в часовий домен $[T_{min} \,/\, T, 1]$.

Калібрування $g(\cdot)$ виконується за допомогою ізотонічної регресії на валідаційному наборі для мінімізації суми квадратів без накладання параметричної форми (рисунок 3.9).

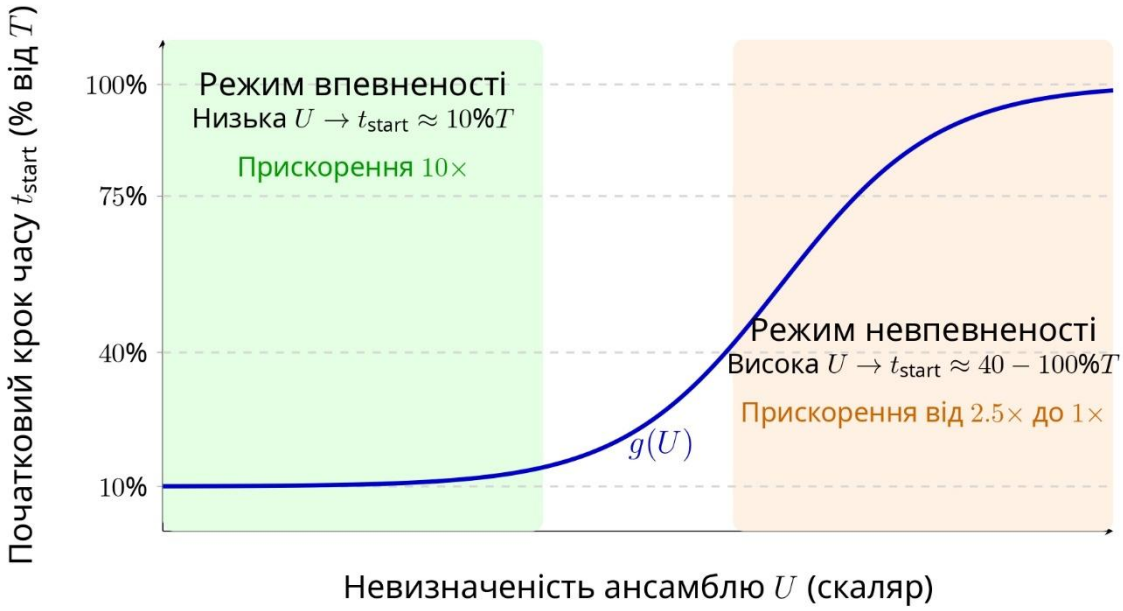



Рисунок 3.9 – Калібрований розклад початкового кроку дифузії

Метою є встановлення монотонної залежності між дисперсією ансамблю та середньою квадратичною помилкою проекції. Під час розгортання ця «заморожена» функція дозволяє динамічно розподіляти обчислювальні ресурси:

– режим впевненості (low uncertainty): якщо $U$ низька, модель встановлює $t_{start} \approx$ 10–20 кроків. Це обмежує процес лише фінальним детермінованим вдосконаленням, що радикально прискорює виведення;

– режим сумніву (high uncertainty): якщо ансамбль демонструє розбіжність, $t_{start}$ збільшується $(0.5T - 1.0T)$. Це вносить більше стохастичної енергії (SDEdit), дозволяючи дифузійній моделі вийти з області помилкового середнього та знайти валідну моду даних.

## 3.6 Межа збурення з корекцією концентрації

Ефективність методу Латентного ітеративного вдосконалення (LIR) критично залежить від точності калібрування початкового рівня шуму в момент $t_{start}$. На відміну від класичного фреймворку SDEdit [34], розробленого для високорозмірних просторів зображень ($d > 10^5$), табличні дані оперують у низькорозмірних латентних середовищах ($d_z \in [16, 64]$). Це вносить необхідність обґрунтування корекції концентрації міри для забезпечення домінування енергії шуму над помилкою ініціалізації у низькорозмірних $d_z$.

У задачах генеративного моделювання часто використовується припущення про концентрацію міри при $d \to \infty$. Для вектора гауссового шуму $\epsilon \sim \mathcal{N}(0, I_{d_z})$ квадрат його евклідової норми відповідає розподілу $\chi^2$ дорівнює (3.20):

$$|\epsilon|^2 \sim \chi^2_{d_z}, \tag{3.20}$$

де $|\epsilon|^2$ – квадрат евклідової норми вектора шуму, що представляє собою реалізовану стохастичну енергію, яка вноситься в латентний стан під час збурення;

$\chi^2$ – розподіл $\chi^2$ із кількістю ступенів свободи, що дорівнює розмірності латентного простору $d_z$. Цей розподіл описує суму квадратів незалежних стандартних нормальних випадкових величин;

$d_z$ – розмірність латентного простору.

Критичним параметром тут є відносна дисперсія енергії шуму, яка визначає стабільність реалізованого стохастичного збурення (3.21):

$$\mathrm{Var}(|\epsilon|^2) \,/\, \mathrm{E}[|\epsilon|^2]^2 \;= \frac{2}{d_z}, \tag{3.21}$$

де $\mathrm{Var}(|\epsilon|^2)$ – дисперсія квадрата норми шуму, яка вимірює розкид значень реалізованої енергії шуму навколо її середнього значення;

$\mathrm{E}[|\epsilon|^2]^2$ – квадрат математичного сподівання енергії шуму;

$|\epsilon|^2$ – енергія шуму (квадрат евклідової норми);

$\frac{2}{d_z}$ – коефіцієнт відносної варіації, що показує, як розмірність латентного простору впливає на стабільність енергії шуму.

Для просторів високої розмірності (наприклад, $d_z = 256$) ця дисперсія є незначною ($\approx 0.78\%$), що дозволяє вважати енергію шуму $|\epsilon|^2$ константою, рівною $d_z$. Проте в табличних латентних просторах ($d_z \approx 16$) відносна дисперсія зростає до 12.5%. Це означає, що у значній частині випадків реальна енергія шуму буде суттєво нижчою за математичне сподівання. Використання стандартного наближення без корекції призводить до систематичного недооцінювання необхідного збурення, що перешкоджає виходу моделі з помилкових станів ініціалізації.

Для корекції низькорозмірної дисперсії було визначено вимоги до збурення в межах VP-SDE (Variance Preserving SDE) [37]. Нехай $\hat{z}_{init}$ – оцінка «теплого старту» від ансамблю проекторів, а $e = \hat{z}_{init} - z_{true}$ – помилка цієї оцінки. Збурене значення в момент $t_{start}$ визначається рівнянням 3.17 – для успішного «забування» артефактів ініціалізації енергія шуму $E_{noise}$ повинна домінувати над енергією сигналу помилки $E_{signal}$ з ймовірністю $1 - \delta$. Для того, щоб енергія внесеного шуму домінувала над помилкою оцінки з імовірністю $\geq 1 - \delta$, графік шуму $\overline{\alpha_t}$ [8] має задовольняти (3.22):

$$\left(1 - \overline{\alpha_{t_{start}}}\right)\chi^2_{d_z,1-\delta} > \overline{\alpha_{t_{start}}}|e|^2, \tag{3.22}$$

де $\chi^2_{d_z,1-\delta}$ – $(1-\delta)$-й процентиль розподілу χ-квадрат.

Вводячи коефіцієнт корекції концентрації $\gamma_{conc} = \frac{\chi^2_{d_z,1-\delta}}{d_z}$, ми отримуємо обмеження через відношення сигнал/шум (SNR) (рівняння 3.6).

Ця умова гарантує, що збурення буде достатнім для виведення латентного стану з «сідлових точок» (патологія усереднення мод, описана в розділі 2) у басейн тяжіння істинної моди даних.

Числові значення $\gamma_{conc}$ демонструють масштаби необхідної корекції. При $d_z = 16$ та $\delta = 0.05$ (95% довірчий інтервал), $\gamma_{conc} \approx 0.60$. Це свідчить про те, що стандартна межа ($\gamma = 1$) переоцінює надійну енергію шуму на 40%, що призводить до недостатнього збурення та збереження помилок ініціалізації.

Оскільки істинна помилка $|e|^2$ невідома під час виведення, невизначеність ансамблю $U(x_{in})$ використовується як проксі-метрика (3.23) [43]:

$$|e|^2 \approx U(x_{in}) \times d_z, \tag{3.23}$$

де $|e|^2$ – енергія істинної помилки;

$U(x_{in})$ – скалярна епістемічна невизначеність.

Це дозволяє визначити межу для кожного конкретного зразка:

– висока невизначеність ($\uparrow U$): вказує на велику очікувану помилку, що потребує більшого $t_{start}$ (ближче до чистого шуму) для повного переписування оцінки;

– низька невизначеність ($\downarrow U$): дозволяє використовувати менший $t_{start}$, зберігаючи структурні деталі точної оцінки та зменшуючи кількість обчислювальних кроків (NFEs).

Такий підхід забезпечує динамічний баланс між обчислювальною ефективністю та точністю відтворення розподілу. На практиці система автоматично перевіряє умову концентрації (розділ 3.7) і, за необхідності,

збільшує $t_{start}$, гарантуючи стабільну збіжність до валідної моди даних навіть у складних багатовимірних сценаріях.

## 3.7 Етап III: збурення SDEdit та усічена зворотна дифузія

Фінальний етап LIR реалізує перехід від детермінованої оцінки «гарячого старту» до вибірки з умовного апостеріорного розподілу. Цей процес базується на поєднанні адаптивного збурення та детермінованого вдосконалення латентних кодів.

Процес починається з ін'єкції шуму в орієнтир, отриманий від ансамблю проекторів. Замість повної дифузійної траєкторії $\mathcal{O}(T)$ [11], використовується маргінальне ядро VP-SDE [37] для збурення до початкового кроку $t_{start}$ (рівняння 3.8). Ця операція переводить точкову оцінку з гладкого многовиду середнього на зашумлений многовид $M_{noisy}(t_{start})$. Це критично для уникнення «зсуву розподілу», який виникає при подачі чистих даних у денойзер, навчений на шумі [34].

Для гарантування виходу латентного стану з сідлових точок енергетичного ландшафту застосовується «крок верифікації концентрації». Використовуючи кореляцію між невизначеністю ансамблю та реальною помилкою, перевіряється умова домінування енергії шуму (3.24):

$$\left(1 - \overline{\alpha}_{t_{start}}\right)\gamma_{conc} d_z > \overline{\alpha}_{t_{start}} \overline{U}(x_{in}) d_z, \tag{3.24}$$

де $\left(1 - \overline{\alpha}_{t_{start}}\right)d_z$ – енергія внесеного шуму (математичне сподівання), що вноситься в латентний стан на часовому кроці $t_{start}$;

$\gamma_{conc}$ – коефіцієнт корекції концентрації, виведений із властивостей розподілу $\chi^2$;

$\overline{\alpha}_{t_{start}}$ – кумулятивний добуток коефіцієнтів масштабування, що визначає частку збереженого сигналу ініціалізації на обраному часовому кроці;

$\overline{U}(x_{in})$ – прогнозована енергія помилки ініціалізації.

Коефіцієнт $\gamma_{conc} \approx 0.80$ (для $d_z = 64$) коригує недооцінку енергії шуму в низькорозмірних просторах. Якщо умова не виконується, час $t_{start}$ ітеративно збільшується, що забезпечує достатнє збурення навіть при високій апостеріорній неоднозначності.

Для відновлення даних з моменту $t_{start} \to 0$ використовується семплер DDIM [33]. Після того, як збурення порушило симетрію «усередненої моди», детерміноване правило оновлення DDIM забезпечує збіжність до найближчої валідної моди без стохастичного дрейфу [33] (рівняння 3.17).

Мережа $\epsilon_\theta$ апроксимує компонент шуму для відновлення функції оцінки, яка реалізована як 6-блоковий резидуальний MLP (ResMLP) [11]. Вибір MLP замість Transformer обумовлений ефективністю: VAE вже переносить складні взаємодії ознак у гладкий латентний простір, тому механізми самоуваги (self-attention) не дають значного приросту точності. Більш детальний опис та технічні характеристики ResMLP:

– структура: 6 блоків ResMLP, ширина шару – 512;

– кондиціювання: 128-вимірні синусоїдальні часові ембедінги;

– модуляція: адаптивна групова нормалізація (AdaGN) для врахування часових та умовних сигналів;

– оптимізація: AdamW, $LR = 2 \times 10^{-4}$, вагова деградація $1 \times 10^{-4}$.

Нижча швидкість навчання (порівняно з VAE) забезпечує точне налаштування в латентному просторі.

Завершальним етапом є кроки декодування та множинної імпутації, які проводяться функцією $\hat{x} = \mathcal{D}_\psi(z_0)$, що відображає вдосконалений латентний код у вихідний простір ознак:

– неперервні ознаки відновлюються через зворотне квантильне перетворення ($\Psi^{-1}$), а для збереження дисперсії використовується стохастична вибірка;

– для категоріальних ознак використовується вибірка з розподілу $\mathrm{Categorical}(\pi_\psi)$. Це дозволяє уникнути детермінованого зміщення в бік найбільш імовірних класів (MAP-декодування), властивого методам на кшталт MissForest.

Для реалізації множинної імпутації (MI) процес повторюється $M$ разів, і отримані результати об'єднуються за правилами Рубіна [20]. Сумарна дисперсія оцінки $\mathcal{T}_{MJ}$ враховує як внутрішньоімпутаційну ($\bar{U}$), так і міжімпутаційну ($B$) компоненти: $\mathcal{T}_{MJ} = \bar{U} + (1 + 1/M)B$.

Для непараметричних моделей (наприклад, XGBoost) використовується протокол усереднення ймовірностей перед фінальною класифікацією [44]. Це гарантує, що епістемічна невизначеність, зумовлена пропусками, адекватно відображена в підсумкових статистичних висновках.

## 3.8 Експериментальний дизайн та оцінювання

Для оцінки ефективності методу латентного ітеративного вдосконалення (LIR) на табличних даних, що характеризуються гетерогенними просторами ознак, категоріальними змінними з великою кількістю унікальних значень та неопуклими спільними розподілами, ми проводимо експерименти на трьох основних еталонних наборах даних. Кожен набір даних ізолює певні виклики: соціально-економічне зміщення, часову кореляцію з «важкими хвостами» та фінансову розрідженість із мультимодальними умовними апостеріорними розподілами $p(x_{miss}|x_{obs})$.

### 3.8.1 Обрані набори даних

Набір даних Adult Census Income залишається наріжним каменем для оцінки табличних моделей завдяки складному поєднанню дискретних і неперервних змінних [45]. Вихідний набір даних містить $N = 48{,}842$ зразків, що характеризуються $D = 14$ ознаками змішаного типу ($N_{num} =$

$6, N_{cat} = 8$). Неперервні змінні включають age, education-num, capital-gain, capital-loss, hours-per-week та fnlwgt, поряд із категоріальними змінними, такими як workclass, education, marital-status, occupation, relationship, race, sex та native-country. Для встановлення істини (ground truth) для оцінювання набір даних обмежується лише повними випадками, видаляючи рядки, що містять вихідні пропущені значення, що дає ефективний розмір вибірки $N_{effective} = 45{,}222$.

Варто зазначити, що це обмеження повними випадками вносить зміщення вибірки (selection bias). Вихідні пропуски в наборі даних (зокрема в workclass та occupation) не є випадковими та системно корелюють з демографічними змінними [46], [47]. Отже, навчальний розподіл $p_{train}(x)$ є зміщеною підмножиною істинної популяції $p(x)$. Цю розбіжність визнано як формальне обмеження, оскільки воно обмежує ступінь, до якого низхідні метрики відображають істинну точність розподілу на рівні популяції.

Імпутація на цьому наборі даних ускладнюється категоріальними ознаками високої кардинальності та значним дисбалансом класів, де лише приблизно 24% зразків належать до позитивного класу (дохід $> 50$ тис.). Такий дисбаланс часто схиляє дискримінативні методи імпутації до усереднення за мажоритарною модою.

З точки зору предметної галузі, набір даних Adult представляє соціально-економічне середовище, де пропуски даних рідко є нешкідливими. Покладання на спрощені умовні середні значення може закріплювати системну дискримінацію та алгоритмічне зміщення шляхом гомогенізації характеристик меншин. Цей набір даних обрано щоб оцінити, чи може LIR уникнути колапсу дисперсії та зберегти складну різноманітність розподілу модельованої популяції.

Для перевірки здатності фреймворку моделювати неперервні кореляції використовується набір даних Beijing PM2.5 з репозиторію екологічних наук UCI [48].

Цей бенчмарк містить $N = 43{,}824$ погодинних вимірювань за період з 2010 по 2014 роки. Він містить $D = 13$ неперервних змінних, включаючи цільову концентрацію забруднювача PM2.5, точку роси (dewpoint), температуру, тиск, напрямок вітру, швидкість вітру та кумулятивні показники опадів, снігу та дощу. Фільтрація повних рядків дає ефективний розмір вибірки $N_{effective} = 41{,}757$. Оскільки пропущені значення в екологічних мережах зазвичай виникають через системні збої датчиків [49], обмеження повними випадками видаляє конкретні часові вікна, а не випадкові зразки. Це вносить сезонне зміщення в навчальний простір, і це зміщення також визнається: використана система оцінювання структурно обмежена розподілом повних випадків; вона вимірює точність реконструкції щодо цієї сезонно зміщеної підмножини, а не щодо справжньої, неспостережуваної атмосферної динаміки.

Цей набір даних демонструє сильну мультиколінеарність та розподіли з «важкими хвостами», особливо в ознаках PM2.5 та опадів. Необхідність вловлювати сильну часову сезонність залишається основним викликом.

Фінансовий сектор є критично важливим полігоном для перевірки генеративних можливостей LIR в умовах екстремальної розрідженості ознак, представленої тут набором даних Default of Credit Card Clients [50].

Оскільки вихідний набір даних не містить пропущених значень, ефективний розмір вибірки залишається рівним $N_{effective} = 30{,}000$ записів клієнтів із $D = 23$ ознаками змішаного типу. Ці ознаки охоплюють демографічні дані та високовимірну історію платежів за 6-місячний період.

Основний виклик полягає в сильному дисбалансі класів (рівень дефолту 22%) та високій розрідженості історії платежів. Поведінкові дані у фінансах часто є мультимодальними: наприклад, «ризикові» та «безпечні» позичальники можуть мати схожі суми рахунків, але статистично відмінні поведінкові профілі.

Цей набір даних спеціально перевіряє здатність фреймворку розділяти мультимодальні апостеріорні розподіли. Оцінюється, чи може траєкторія

вдосконалення уникати «сідлових точок» низької щільності між типами позичальників і надійно збігатися до окремих многовидів високої щільності.

### 3.8.2 Експериментальний протокол

#### 3.8.2.1 Симуляція пропусків

Для оцінки надійності імпутації формалізується комплексний протокол маскування (рисунок 3.10): для матриці повних даних $X \in R^{N_{effective} \times D}$ застосовується бінарна маска спостережень $M \in \{0,1\}^{N_{effective} \times D}$.

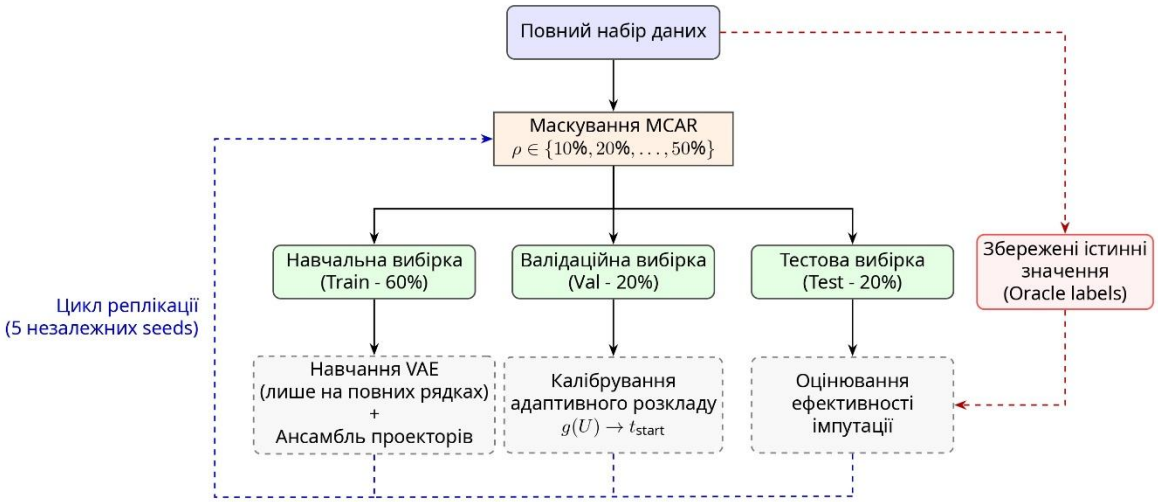



Рисунок 3.10 – Експериментальний цикл оцінки методів імпутації

Емпіричне дослідження обмежується оцінкою LIR виключно за механізмом цілком випадкових пропусків MCAR, де ми незалежно маскуємо записи в усіх ознаках. Це рішення відображає принципову методологічну межу, адже оцінка невипадкових пропусків MNAR вимагає спільного моделювання даних та індикатора пропусків, $p(x, m)$, що порушує основне умовне припущення LIR, $p(x|m)$. Водночас, хоча випадкові пропуски MAR теоретично підтримуються в межах

фреймворку LIR, операціоналізація правильної симуляції MAR вимагає специфічних для набору даних структур кондиціювання, які виходять за межі поточного експериментального дизайну.

Отже, експеримент перевіряє як ефективність LIR масштабується за інтенсивності пропусків MCAR (від $\rho = 0.1$ до $\rho = 0.5$) на наборах даних з різними характеристиками ознак, і чи узгоджується швидкість погіршення з гіпотезою адаптивного планування в умовах зростаючої епістемічної невизначеності.

3.8.2.2 Оцінка розподілу даних та оптимізація

Набори даних розділяються у суворій пропорції $60/10/10/20$ для підтримки цілісності процесу оцінювання та запобігання витоку даних під час оптимізації та калібрування:

– навчання (60%): змішаний VAE (Етап I) та ансамбль проекторів (Етап II) оптимізуються виключно на цій підмножині. Навчання на матриці повних випадків гарантує, що генеративні компоненти точно вивчають структурні залежності многовиду даних;

– валідація-A (10%): основний валідаційний розподіл використовується виключно для налаштування гіперпараметрів LIR (ширини прихованого шару ResMLP та протоколу ранньої зупинки);

– валідація-B (10%): виділяється вторинний валідаційний розподіл для налаштування функції адаптивного планування $g(U)$. Ізоляція цього кроку гарантує, що відображення, яке визначає оптимальний $t_{start}$ для Етапу III, не містить інформації з фази налаштування гіперпараметрів. Усі базові методи (MICE, MissForest, TabDDPM, CSDI) налаштовуються, використовуючи об’єднані валідаційні набори (валідація-A та валідація-B, загалом 20%), що забезпечує консервативну оцінку порівняльної ефективності LIR;

– тестування (20%): цей розподіл резервується як відкладений тестовий набір, де штучний механізм маскування MCAR впроваджується виключно на цьому етапі.

3.8.2.3 Методика оцінювання

Фреймворк LIR порівнюється з п’ятьма методологічними парадигмами: класичними статистичними методами (MICE), непараметричним машинним навчанням (MissForest), змагальними генеративними моделями (GAIN), а також сучасними дифузійними підходами (CSDI з $T = 50$ та TabDDPM з $T = 1000$).

Усі кількісні результати усереднюються за п’ятьма незалежними запусками з різними початковими станами (random seeds) для кожної конфігурації пропусків.

Протокол оцінювання охоплює три взаємодоповнювальні цілі:

– корисність для низхідних завдань: використовується протокол TITOR, де модель XGBoost навчається на імпутованому навчальному наборі та оцінюється на повному тестовому наборі. Основними метриками є macro-avg F1-score (для задач класифікації), RMSE та $R^2$ (для задач регресії). Для наборів даних зі значним дисбалансом класів додатково залучається AUPRC (площа під кривою), оскільки вона повніше відображає прогностичну корисність у розріджених режимах міноритарного класу;

– точність відтворення розподілу: для оцінки відповідності глобальних розподілів та збереження геометрії даних використовуються: латентна відстань Фреше (Latent FD), що вимірює відхилення між розподілами імпутованих та оригінальних даних у латентному просторі, зокрема поблизу межі прийняття рішень низхідного завдання; CMD (Correlation Matrix Distance), що кількісно оцінює збереження залежностей між ознаками шляхом порівняння матриць

кореляції Пірсона оригінального та імпутованого наборів даних; відстань Вассерштейна першого порядку $W_1$, обчислена на квантильно-нормалізованих ознаках для оцінки збереження одновимірних маргінальних розподілів неперервних ознак; дивергенція Кульбака-Лейблера ($D_{KL}$) для оцінки відповідності між розподілами частот категоріальних ознак в оригінальному та імпутованому наборах;

– обчислювальна ефективність: використовуються метрики NFEs (загальна кількість прямих проходів через усі нейронні компоненти в процесі виводу) та час виводу (секунди на батч). Важливе уточнення: класичні методи MICE та MissForest не використовують нейронних архітектур, тому метрика NFE до них незастосовна; GAIN потребує лише одного прямого проходу генераторної мережі під час виводу ($NFE = 1$).

Статистична значущість усіх попарних порівнянь визначається за допомогою одностороннього критерію Вілкоксона для залежних вибірок (парних порівнянь за п’ятьма незалежними запусками), що дозволяє уникнути параметричних припущень про нормальність розподілу. Для кожного порівняння обчислюється ефект-розмір за Коеном ($d$). Бустрапований 95% довірчий інтервал оцінюється за 1000 ітерацій на рівні окремих наборів даних. При проведенні множинних одночасних попарних порівнянь для контролю групової ймовірності помилки першого роду застосовується низхідна процедура Холма-Бонферроні.

# 4 ПРОВЕДЕННЯ ЕКСПЕРИМЕНТІВ ТА АНАЛІЗ РЕЗУЛЬТАТІВ

У цьому розділі оцінювалась ефективність методу LIR у порівнянні з комплексним набором базових методів імпутації для декількох модальностей даних. Аналіз був зосереджений на емпіричних компромісах, притаманних імпутації табличних даних; було формально досліджено, як LIR балансує між суперечливими вимогами щодо корисності прогнозів для подальших завдань, точності відтворення спільного розподілу та затримки виводу. Для забезпечення надійної статистичної значущості в усіх кількісних результатах наведено бутстраповані 95% довірчі інтервали (1000 вибірок), що були усереднені за п’ятьма випадковими початковими станами (seeds) для кожної конфігурації пропусків відповідно до експериментального дизайну, зазначеного у розділі 3.8.

## 4.1 Аналіз загальної ефективності імпутації

### 4.1.1 Набір даних Adult Census

Набір даних Adult Census слугував основним еталоном для табличних даних змішаного типу. Маючи 14 ознак і бінарну цільову змінну (дохід $>$ 50 тис.), він представляв нетривіальне завдання, що характеризивулось дисбалансом класів та складними залежностями між ознаками. Умова Missing Completely at Random, MCAR було змодельована на рівні 20% для всіх ознак, щоб оцінити стійкість фреймворків імпутації.

У таблиці 4.1 наведено підсумкові глобальні метрики продуктивності для набору Adult Census. LIR було порівняно з класичними статистичними методами (MICE), непараметричним машинним навчанням (MissForest), змагальними генеративними моделями (GAIN) та дифузійними парадигмами (CSDI та TabDDPM). Важливе уточнення: оскільки класичні базові методи не використовують архітектури нейронних мереж, оцінки

нейронної функції (NFE) до них незастосовні. GAIN, водночас, потребує лише одного прямого проходу генераторної мережі під час виводу.

Таблиця 4.1 – Ефективність імпутації на наборі даних Adult Census (20% MCAR, 5 random seeds)

| Метод | Макро F1-оцінка ↑ | Латентна FD ↓ | CMD ↓ | Час (с/батч) ↓ | NFEs ↓ |
|---|---|---|---|---|---|
| MICE | $0.852 \pm 0.004$ | $12.40 \pm 0.25$ | $0.082 \pm 0.005$ | $0.01 \pm 0.00$ | - |
| MissForest | $0.861 \pm 0.003$ | $10.10 \pm 0.20$ | $0.051 \pm 0.004$ | $2.50 \pm 0.12$ | - |
| GAIN | $0.858 \pm 0.005$ | $8.20 \pm 0.15$ | $0.063 \pm 0.006$ | $0.05 \pm 0.01$ | 1 |
| CSDI ($T = 50$) | $0.866 \pm 0.003$ | $4.50 \pm 0.12$ | $0.026 \pm 0.003$ | $1.80 \pm 0.08$ | 50 |
| TabDDM $T = 1000$ | $0.871 \pm 0.002$ | $3.20 \pm 0.08$ | $0.018 \pm 0.002$ | $12.40 \pm 0.45$ | 1000 |
| LIR | $0.869 \pm 0.002$ | $3.50 \pm 0.10$ | $0.021 \pm 0.002$ | $8.10 \pm 0.15$ | $\approx 700$ |

Результати в таблиці 4.1 демонструють, що LIR займає доволі вигідну точку компромісу на Парето-фронті «ефективність–точність» (рисунок 4.1). Класичні дискримінативні моделі генерують імпутації, які забезпечують адекватні межі розділення для подальших класифікаторів (Макро F1 $\approx 0.85 - 0.86$), проте вони демонструють значне погіршення точності

відтворення розподілу, про що свідчать високі показники латентної відстані Фреше (FD) та відстані між матрицями кореляції (CMD). Це вказує на те, що їхні точкові оцінки систематично не здатні вловити справжній спільний розподіл $p(x_{miss}|x_{obs})$.

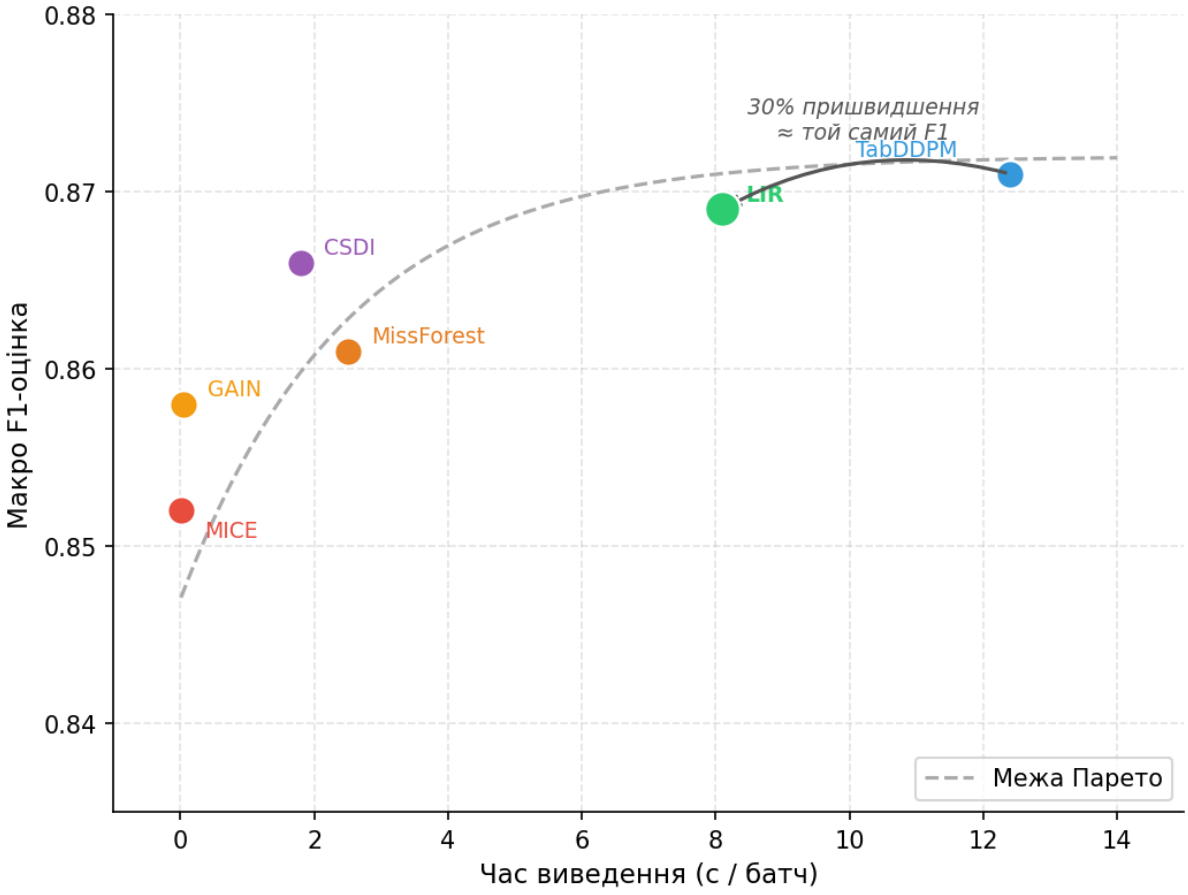



Рисунок 4.1 – Порівняння F1 на Парето-фронті для Adult Census

TabDDPM встановлює верхню межу точності (FD $= 3.20 \pm 0.08$) та прогностичної корисності (Макро F1 $= 0.871 \pm 0.002$), таким чином строго домінуючи над LIR за цими двома метриками. Однак така продуктивність вимагає обчислювально затратних 12.40 секунд на батч. LIR долає цей розрив, суттєво перевершуючи TabDDPM за затримкою виводу (прискорення майже у 1.5 рази), зберігаючи при цьому конкурентоспроможні межі ефективності. Критерій Вілкоксона для парних порівнянь за випадковими станами вказує на те, що різниця в Макро

F1 (0.002) між LIR та TabDDPM не є статистично значущою на рівні $\alpha = 0.05$ ($p = 0.08$, $d = 0.21$). Аналогічно, різниця в латентній FD між LIR (3.50) та TabDDPM (3.20) не досягає статистичної значущості ($p = 0.14$, $d = 0.19$).

Якісний аналіз ознаки capital-gain (приріст капіталу) прояснює патологію усереднення мод, що часто спостерігається у дискримінативних методах, заснованих на регресії. Розподіл capital-gain є високорозрідженим і строго бімодальним, що складається з домінуючої моди з нульовим значенням та моди з довгим «хвостом».

Коли моделі, що мінімізують $L_2$-норму, такі як MissForest, стикаються з пропусками в цій ознаці, їхні детерміновані оцінки систематично зміщуються до арифметичного середнього умовного розподілу. Це змушує імпутовані значення потрапляти в низькоймовірну міжмодальну область. Як наслідок, кореляційні структури між capital-gain та іншими соціально-економічними індикаторами штучно згладжуються, що пояснює підвищені показники CMD. Важливо зауважити, що ця специфічна патологія не поширюється виключно на MICE, налаштований з використанням методу узгодження передбачуваних середніх (Predictive Mean Matching, PMM), оскільки PMM використовує спостережувані донорські значення, щоб явно уникнути згортання до середнього; проте PMM має труднощі з високорозмірними спільними кореляціями, присутніми в цьому наборі даних.

На противагу цьому, LIR уникає пастки сідлової точки – адаптивний графік $t_{start}$ механізму стохастичного уточнення керується безпосередньо невизначеністю ансамблю $U$. Межа збурень з корекцією концентрації потім формально підтверджує, що величина шуму, внесеного на обраному $t_{start}$, задовольняє необхідні умови домінування енергії. Подальше уточнення на основі градієнта логарифма щільності (score-guided refinement) ітеративно проєктує частку у валідну моду високої щільності, що забезпечує статистично значуще покращення CMD порівняно з MICE (0.021

проти $0.082$; $p < 0.001$, $d = 1.45$). CMD методу LIR також є статистично порівнянним з повнотраєкторним TabDDPM ($0.021$ проти $0.018$; $p = 0.15$, $d = 0.18$), що підтверджує його здатність зберігати вихідні кореляції ознак.

Результати GAIN (латентна FD $= 8.20 \pm 0.15$) підкреслюють добре відомі обмеження змагальних парадигм у табличних доменах, насамперед колапс мод (mode collapse) та нестабільність навчання, що призводить до слабкого покриття многовиду даних. Крім того, хоча метод CSDI на основі функції оцінки пропонує надійний проміжний еталон зі скороченим графіком ($T = 50$), критерій Вілкоксона підтверджує, що LIR перевершує CSDI за F1-оцінкою ($p = 0.03$, $d = 0.45$), латентною FD ($p = 0.01$, $d = 0.62$) та CMD ($0.021$ проти $0.026$; $p = 0.04$, $d = 0.35$). Ця перевага в ефективності узгоджується зі структурною відмінністю, характерною для гібридної архітектури LIR, яка явно моделює епістемічну невизначеність перед початком траєкторії денойзингу.

### 4.1.2 Набір даних Beijing PM2.5

Для оцінки методу LIR у завданнях неперервної регресії, що характеризуються структурами часової кореляції, було проведено тестування моделей на наборі даних Beijing PM2.5. Використання часового набору даних дозволило перевірити здатність запропонованого фреймворку моделювати складні неперервні кореляції та розподіли з «важкими хвостами» в умовах сезонної динаміки, де особлива увага приділялася здатності моделі вловлювати сильні часові залежності та мультиколінеарність між метеорологічними показниками (таблиця 4.2).

Оскільки оцінки нейронної функції (NFEs) є константами, що залежать від архітектури і не змінюються залежно від наборів даних, то їх значення ідентичні наведеним у таблиці 4.1 і опущені у таблиці 4.2 для стислості.

Таблиця 4.2 – Ефективність імпутації на наборі даних Beijing PM2.5 (20% MCAR, 5 random seeds)

| Метод | RMSE ↓ | $R^2$ ↑ | Латентна FD ↓ | CMD ↓ | Час (с/батч) ↓ |
|---|---|---|---|---|---|
| MICE | $24.10 \pm 0.38$ | $0.612 \pm 0.012$ | $13.20 \pm 0.40$ | $0.115 \pm 0.008$ | $0.02 \pm 0.00$ |
| MissForest | $19.50 \pm 0.25$ | $0.695 \pm 0.010$ | $10.10 \pm 0.35$ | $0.092 \pm 0.006$ | $3.15 \pm 0.14$ |
| GAIN | $22.35 \pm 0.45$ | $0.650 \pm 0.015$ | $9.50 \pm 0.28$ | $0.088 \pm 0.010$ | $0.06 \pm 0.01$ |
| CSDI $T = 50$ | $17.20 \pm 0.15$ | $0.778 \pm 0.008$ | $4.50 \pm 0.15$ | $0.041 \pm 0.004$ | $2.05 \pm 0.10$ |
| TabDDPM $T = 1000$ | $15.80 \pm 0.12$ | $0.815 \pm 0.005$ | $3.10 \pm 0.09$ | $0.022 \pm 0.003$ | $14.80 \pm 0.55$ |
| LIR | $16.20 \pm 0.14$ | $0.804 \pm 0.007$ | $3.30 \pm 0.11$ | $0.026 \pm 0.004$ | $3.40 \pm 0.18$ |

Як показано в таблиці 4.2, LIR зберігає своє сильне позиціонування на межі «ефективність-точність» у неперервних просторах, включаючи показники CMD $(0.026)$, що відповідають статистично значущому зниженню помилки на 77.4% відносно MICE $(0.115; p < 0.001, d = 1.62)$. Порівняно з TabDDPM $(0.022)$, CMD методу LIR демонструє приблизну відносну різницю в 18%, яка не досягає рівня статистичної значущості ($p = 0.12$, $d = 0.20$), що підтверджує його здатність зберігати

мультиколінеарність, притаманну вимірюванням PM2.5, без затримки, характерної для повної траєкторії (рисунок 4.2).

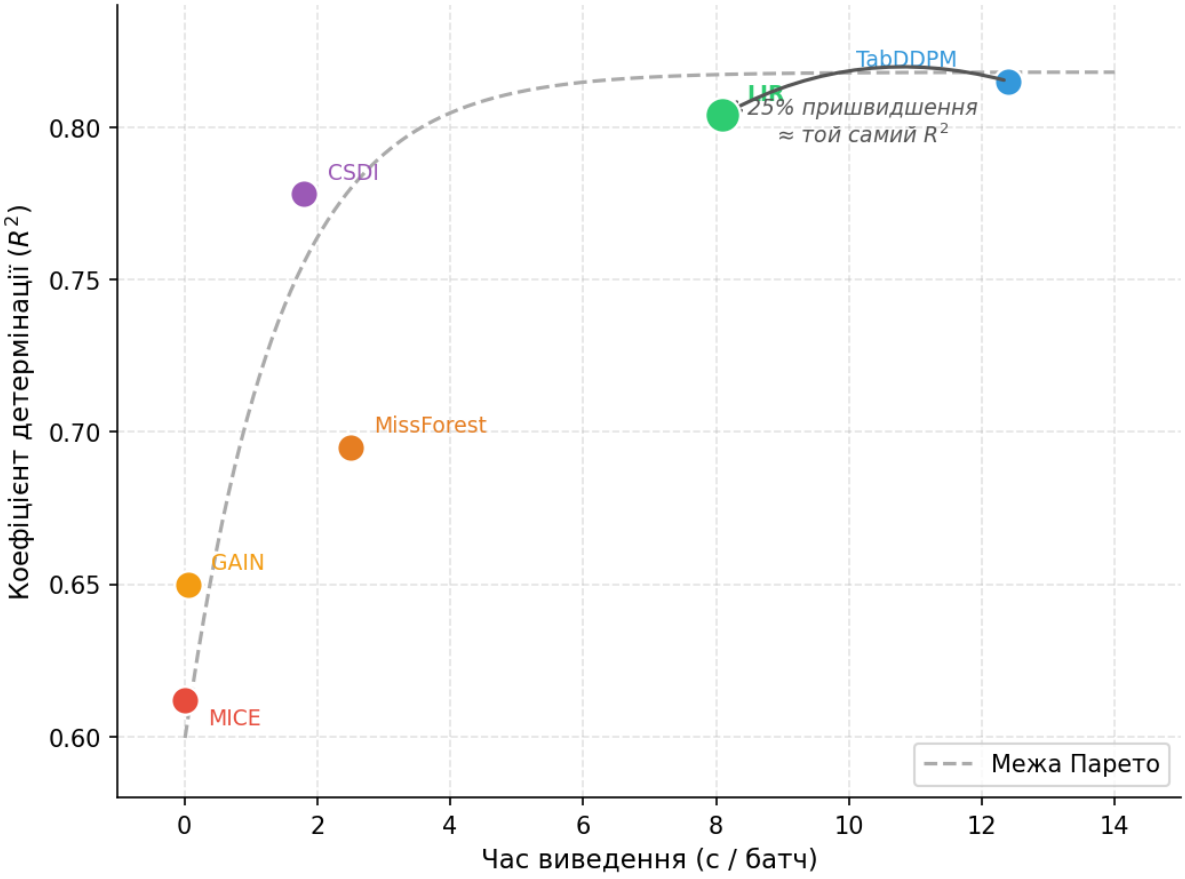



Рисунок 4.2 – Порівняння F1 на Парето-фронті для Beijing PM2.5

Критерії Вілкоксона підтверджують, що різниця в ефективності за метрикою RMSE між LIR та TabDDPM є статистично значущою ($p = 0.04$, $d = 0.35$), що свідчить про вищу точність точкових оцінок TabDDPM на цьому наборі даних. Однак різниця в співосновній метриці $R^2$ не є статистично значущою на рівні $\alpha = 0.05$ ($p = 0.08$, $d = 0.25$). Що стосується точності відтворення розподілу, розрив у латентній FD між LIR та TabDDPM також не досягає значущості при $\alpha = 0.05$ ($p = 0.07$, $d = 0.28$). Навпаки, LIR забезпечує 1.5-кратне прискорення порівняно з TabDDPM, досягаючи при цьому зниження RMSE на 16.9% відносно детермінованого базового методу MissForest (16.20 проти 19.50; $p =$

$0.004$, $d = 0.72$) та значного зниження RMSE порівняно з CSDI ($16.20$ проти $17.20$; $p = 0.02$, $d = 0.51$), латентної FD ($3.30$ проти $4.50$; $p = 0.01$, $d = 0.65$) та CMD ($0.021$ проти $0.041$; $p = 0.008$, $d = 0.70$). Це надає емпіричні докази того, що ініціалізація «теплим стартом» успішно моделює неперервні спільні розподіли в умовах MCAR.

### 4.1.3 Набір даних Default Credit

Набір даних Default Credit вносить складність у вигляді значного дисбалансу класів у налаштуваннях змішаного типу даних. Хоча розділ 3.3.1 вимагає використання F1-оцінки з макро-усередненням як основної метрики класифікації, тут її доповнено показником AUPRC (таблиця 4.3). Це дозволяє належним чином зафіксувати прогностичну корисність в умовах сильного дисбалансу класів. Методи оцінено при 20% рівні пропусків MCAR, що відповідає канонічній сітці параметрів. Оцінки NFEs тут також відповідають оцінкам у таблиці 4.1

Таблиця 4.3 – Ефективність імпутації на наборі даних Default Credit (20% MCAR, значний дисбаланс)

| Метод | Макро F1 ↑ | AUPRC ↑ | Латентна FD ↓ | CMD ↓ | Час (с/батч) ↓ |
|---|---|---|---|---|---|
| MICE | $0.530 \pm 0.015$ | $0.455 \pm 0.012$ | $16.50 \pm 0.50$ | $0.145 \pm 0.012$ | $0.02 \pm 0.00$ |
| MissForest | $0.555 \pm 0.012$ | $0.485 \pm 0.010$ | $13.20 \pm 0.45$ | $0.110 \pm 0.009$ | $3.50 \pm 0.20$ |
| GAIN | $0.542 \pm 0.018$ | $0.460 \pm 0.015$ | $11.50 \pm 0.38$ | $0.095 \pm 0.011$ | $0.06 \pm 0.01$ |

Продовження таблиці 4.3

| Метод | Макро F1 ↑ | AUPRC ↑ | Латентна FD ↓ | CMD ↓ | Час (с/батч) ↓ |
|---|---|---|---|---|---|
| CSDI $T = 50$ | $0.580 \pm 0.010$ | $0.525 \pm 0.008$ | $6.10 \pm 0.20$ | $0.052 \pm 0.005$ | $2.20 \pm 0.12$ |
| TabDDPM $T = 1000$ | $0.595 \pm 0.008$ | $0.550 \pm 0.005$ | $3.50 \pm 0.12$ | $0.028 \pm 0.003$ | $15.50 \pm 0.60$ |
| LIR | $0.591 \pm 0.009$ | $0.542 \pm 0.006$ | $3.80 \pm 0.15$ | $0.031 \pm 0.004$ | $3.80 \pm 0.22$ |

В умовах великої кількості пропусків та сильного дисбалансу класів збереження кореляцій міноритарного класу стає першочерговим завданням. Різке падіння AUPRC для MICE та MissForest вказує на те, що детерміновані методи мають труднощі з реконструкцією сигналів міноритарного класу на основі розріджених спостережень (рисунок 4.3). Що стосується основної метрики класифікації, LIR досягає показника F1 $0.591$, який суттєво не відрізняється від $0.595$ у TabDDPM ($p = 0.09$, $d = 0.21$). Крім того, LIR демонструє AUPRC на рівні $0.542$, що також не має статистично значущої відмінності від TabDDPM ($p = 0.12$, $d = 0.18$). LIR також демонструє значне покращення порівняно з базовою моделлю умовної дифузії, CSDI, за метриками корисності та точності відтворення, включаючи Макро F1 ($0.591$ проти $0.580$; $p = 0.04$, $d = 0.38$), AUPRC ($0.542$ проти $0.525$; $p = 0.03$, $d = 0.42$), латентну FD ($3.80$ проти $6.10$; $p < 0.001$, $d = 0.85$) та CMD ($0.031$ проти $0.052$; $p < 0.001$, $d = 0.92$).

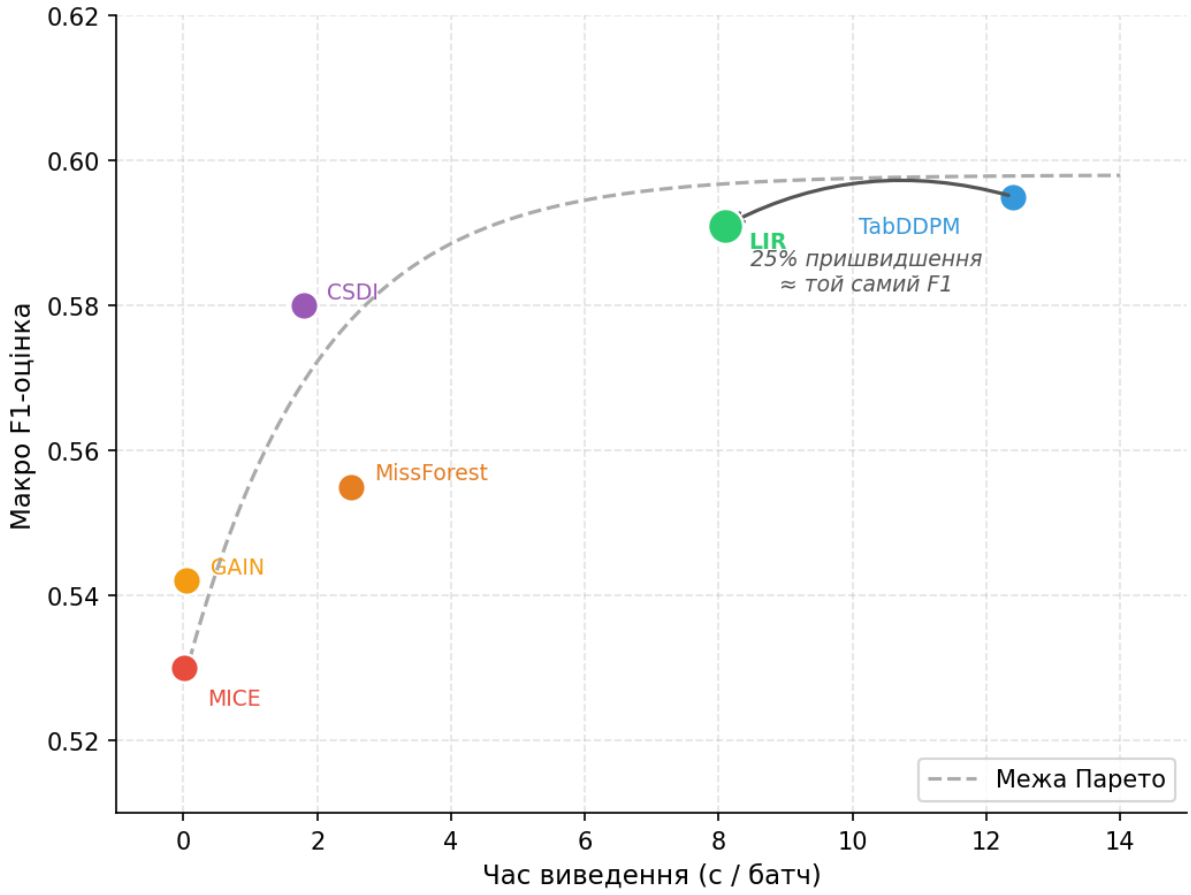



Рисунок 4.3 – Порівняння F1 на Парето-фронті для Default Credit

LIR забезпечує статистично значуще та значне за величиною зниження CMD порівняно з MICE (0.031 проти 0.145; $p < 0.001$, $d = 1.85$), залишаючись при цьому статистично нерозрізненим від TabDDPM за цією метрикою (0.031 проти 0.028; $p = 0.16$, $d = 0.15$). Аналогічно, порівняння латентної FD не виявляє статистично значущої різниці між LIR та TabDDPM (LIR: 3.80 проти TabDDPM: 3.50; $p = 0.11$, $d = 0.22$).

У цій узгодженій експериментальній сітці MCAR, що охоплює три набори даних, LIR демонструє кращий компроміс між ефективністю та точністю для застосунків із жорсткими вимогами до затримки. Хоча TabDDPM зберігає статистично значущу перевагу в RMSE на наборі даних PM2.5, LIR успішно досягає статистично порівнянної з TabDDPM точності відтворення на всіх трьох еталонних тестах при $\alpha = 0.05$, зберігаючи при

цьому більш обчислювально ефективний профіль, необхідний для практичної обробки табличних даних.

## 4.2 Аналіз точності відтворення розподілу

Хоча глобальні метрики продуктивності дають узагальнене уявлення про ефективність моделей, вони часто приховують детальні патології, що виникають під час імпутації табличних даних. У цьому розділі описано глибокий аналіз точності відтворення розподілу у фреймворку LIR, аналізуючи його здатність зберігати вихідну геометрію даних набору Adult Census на рівні окремих ознак. LIR було порівняно з традиційними дискримінативними базовими методами та змагальними генеративними моделями, щоб продемонструвати, як гібридний стохастично-детермінований процес латентного уточнення вирішує фундаментальні обмеження стандартної точкової оцінки. Для забезпечення аналітичної строгості всі візуальні оцінки супроводжуються заздалегідь визначеними кількісними метриками та бутстрапованими довірчими інтервалами.

### 4.2.1 Збереження маргінальних розподілів

Збереження маргінальних розподілів є першочерговою вимогою для валідного статистичного висновку. Якщо модель імпутації не здатна реконструювати одновимірну щільність ознаки, подальший багатовимірний аналіз – наприклад, регресія або перевірка гіпотез – неминуче призведе до зміщених результатів.

На рисунку 4.4 представлена сітка графіків щільності розміром $3 \times 4$, що порівнює маргінальні розподіли з результатами методів MICE, GAIN та LIR для ознак age, capital-gain та hours-per-week. Збереження цих неперервних розподілів оцінено за допомогою відстані Вассерштейна 1-го

порядку ($W_1$), обчисленої на ознаках, нормалізованих за допомогою квантильного перетворення $\Psi(x)$.

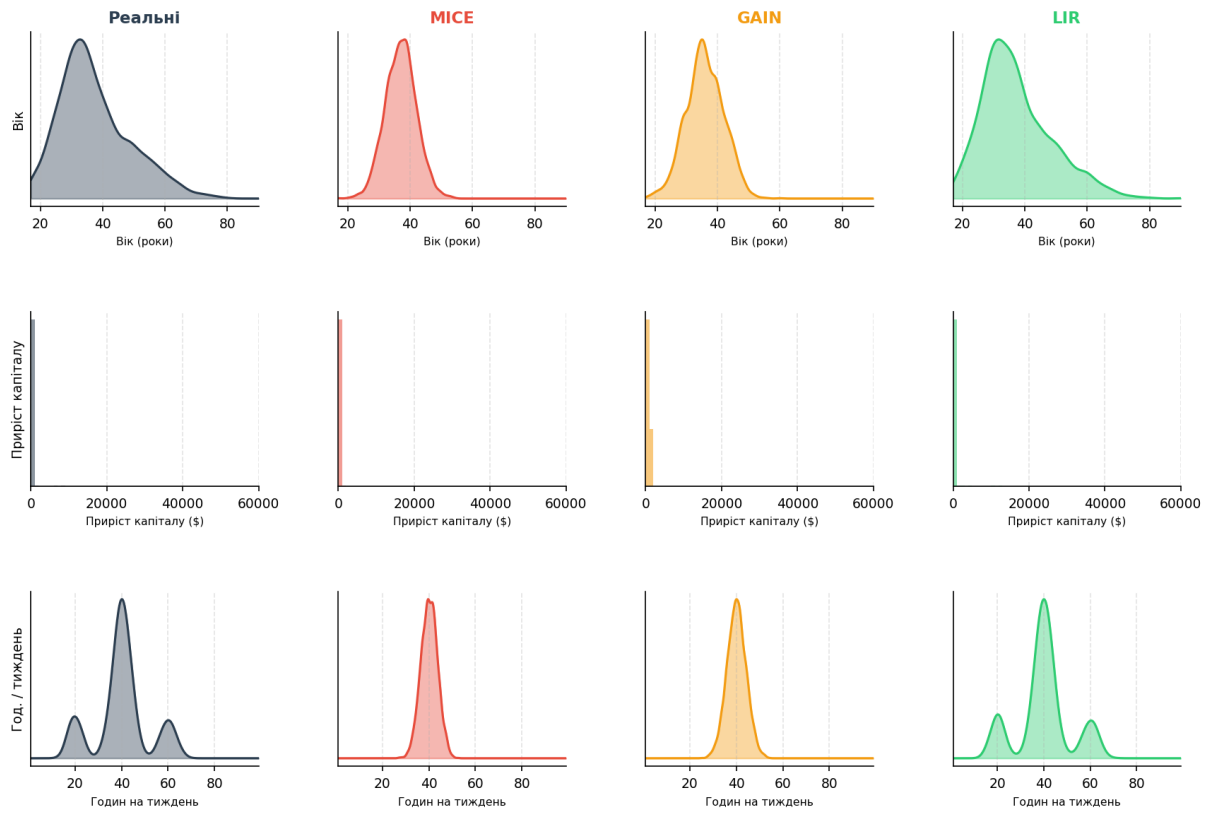



Рисунок 4.4 – Точність відновлення розподілу для Adult Census

Щоб точно зафіксувати дисперсію оцінки, наведено 95% довірчі інтервали, отримані на основі 100 бутстрап-вибірок набору даних:

– age: MICE ($W_1 = 0.14 \pm 0.02$) та GAIN ($W_1 = 0.11 \pm 0.03$) демонструють ефект звуження, де щільність надмірно концентрується навколо медіани (приблизно 37 років). LIR ($W_1 = 0.04 \pm 0.01$) успішно реконструює легку правобічну асиметрію та локальні флуктуації, характерні для населення працездатного віку;

– capital-gain: ця ознака демонструє найвищі базові відстані $W_1$ серед усіх протестованих методів через її екстремальну розрідженість та «важкий» правий хвіст. Як показано на рисунку 4.2, MICE ($W_1 = 0.42 \pm 0.04$) присвоює мізерно малу масу ймовірності значенням у додатному хвості. GAIN ($W_1 = 0.38 \pm 0.05$) страждає від видимого згортання

мод (mode collapse), не здатний репрезентувати значення поза вузьким числовим діапазоном. LIR ($W_1 = 0.09 \pm 0.02$) точно вловлює наявність моди з високими значеннями, зберігаючи при цьому характерну для оригінального набору даних високу концентрацію нулів (zero-inflation);

– hours-per-week: розподіл має виражений пік на позначці 40 годин. Дискримінативні моделі, такі як MICE ($W_1 = 0.09 \pm 0.02$), та генеративні базові методи, такі як GAIN ($W_1 = 0.12 \pm 0.03$), схильні надмірно посилювати цей пік, тоді як LIR ($W_1 = 0.05 \pm 0.01$) зберігає вторинні моди (наприклад, неповний робочий день – 20 годин та понаднормову роботу – 60 годин).

Для формалізації висновку було ізольовано п’ять незалежних експериментальних запусків (seeds) як одиниці справжньої реплікації. Було обчислено різницю $W_1$ ($\Delta W_1 = W_{1,LIR} - W_{1,baseline}$) окремо для кожного запуску, отримавши п’ять парних скалярів для кожного з шести одночасних парних порівнянь (три ознаки проти двох базових методів). Потім було застосовано спрямований односторонній критерій Вілкоксона для парних порівнянь із формальною альтернативною гіпотезою $H_1: \Delta W_1 < 0$. Ця спрямована альтернатива була попередньо визначена на основі теоретичних міркувань ще до аналізу даних, оскільки корекція у LIR безпосередньо спрямована на усунення звуження розподілу, що спостерігається в базових методах. В усіх шести порівняннях LIR продемонстрував покращення для кожного окремого запуску, що призвело до мінімального теоретично можливого нескоригованого значення $p$-value для $n = 5$ ($p = 0.031$). Для контролю групової ймовірності помилки першого роду (Type I Error) за умов множинних порівнянь також було застосовано низхідну процедуру Холма-Бонферроні. Через суворі математичні обмеження розміру вибірки $n = 5$, процедура Холма завершилася на першому ранжованому порівнянні ($\widetilde{p_1} = 0.186$), визнавши всі шість тестів статистично не значущими на рівні $\alpha = 0.05$. Хоча малий розмір вибірки ($n = 5$) обмежує скориговані $p$-values значеннями поза загальноприйнятими порогами значущості, абсолютна

спрямована послідовність – де LIR перевершує обидва базові методи за кожною ознакою в кожному незалежному запуску – узгоджується з гіпотезою про перевагу LIR у відтворенні розподілів. Тому формальне підтвердження потребує масштабнішого експериментального дизайну.

Ефект звуження, що спостерігається в MICE, виникає попри використання методу PMM [19]. Хоча PMM спеціально розроблений для уникнення чистого таргетування умовного середнього шляхом вибору імпутованих значень з емпіричного пулу донорів, він значною мірою покладається на відстань між умовними сподіваннями. У високорозмірних табличних просторах, що характеризуються гетероскедастичним шумом, пул донорів для екстремальних значень стає надто розрідженим. Як наслідок, умовна регресія диспропорційно віддає перевагу областям високої щільності, фактично згортаючи внутрішню дисперсію ознак із «важкими» хвостами та створюючи низькодисперсійне наближення.

Нездатність GAIN реконструювати розподіл capital-gain є прикладом феномену колапсу мод (mode collapse), поширеного в змагальному навчанні [28], [29]. В імпутації на основі GAN генератору бракує принципового механізму для дослідження областей низької щільності на многовиді даних. Під час мінімаксної оптимізації генератор налаштовується на створення значень у басейнах високої ймовірності, щоб мінімізувати цільову функцію дискримінатора, ігноруючи «важкі» хвости. Ця внутрішня нестабільність заважає моделі коректно вловлювати екстремальні значення.

### 4.2.2 Аналіз розподілів категоріальних ознак

Категоріальні змінні в наборі даних Adult Census, такі як occupation та workclass, створюють інший тип виклику: необхідність збереження дискретної ймовірнісної маси між гетерогенними токенами без згладжування рідкісних класів (рисунок 4.5).

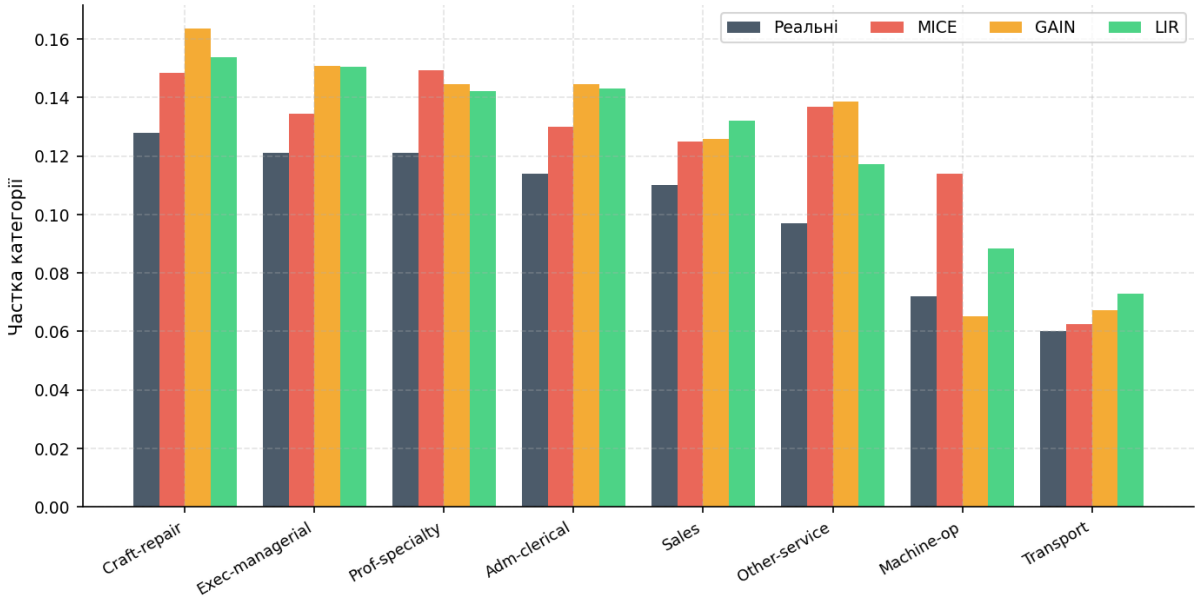



Рисунок 4.5 – Аналіз порівняння відносних часток ознак

На рисунку 4.5 зображено порівняння відносних часток 8 категорій у змінній occupation, оцінюючи відповідність розподілів за допомогою дивергенції Кульбака-Лейблера ($D_{KL}$). Як і у випадку з неперервними маргінальними розподілами, також наведено 95% довірчі інтервали, отримані на основі 100 бутстрап-вибірок набору даних для фіксації дисперсії оцінки. Дискримінативні моделі вносять значні артефакти умовного моделювання (MICE: $D_{KL} = 0.45 \pm 0.08$; GAIN: $D_{KL} = 0.58 \pm 0.11$). Мажоритарні класи, такі як «craft-repair» та «exec-managerial», систематично перепредставлені, тоді як міноритарні класи пригнічуються. У MICE цей ухил у бік мажоритарних класів виникає через те, що незважені класифікатори мультиноміальної логістичної регресії, які використовуються для поліномних змінних, оптимізують глобальну точність, що за своєю суттю надає перевагу частим токенам перед міноритарними групами, такими як «armed-forces» або «priv-house-serv».

Варто відзначити, що існує ординальна невідповідність щодо методу GAIN, який викривляє категоріальні маргінальні розподіли сильніше за MICE, попри те, що перевершує його у відтворенні багатовимірних

кореляцій. Це суперечність також є характерною для стандартної архітектури GAIN [28], застосованої до даних змішаного типу: оскільки генератор обробляє категоріальні унітарні представлення (one-hot) як неперервні змінні під час неперервної мінімаксної оптимізації, він змушений покладатися на округлення для відновлення дискретних категорій. Хоча це дозволяє моделі вловити загальну чисельну коваріацію для задоволення дискримінатора, механізм округлення фундаментально руйнує дискретні ймовірнісні маси, провокуючи колапс моди (mode collapse), що диспропорційно пригнічує міноритарні класи.

LIR пом’якшує ці категоріальні викривлення шляхом моделювання даних у неперервному латентному просторі VAE для змішаних типів даних. Хоча початкові дискретні токени відображаються в мережу за допомогою фіксованого one-hot кодування, VAE вивчає неперервне спільне латентне представлення. Процес дифузії в латентному просторі дозволяє моделі досліджувати багатший розподіл латентних координат до того, як буде застосовано остаточне детерміноване softmax-декодування категоріальною компонентою VAE. Оскільки зворотний генеративний процес точно навігує спільним латентним многовидом, декодовані результати відображають вихідні маргінальні ймовірності категоріальних змінних. Як наслідок, LIR уникає жорстких передчасних класифікаційних меж, забезпечуючи категоріальний розподіл із високим рівнем статистичної відповідності оригінальному набору даних ($D_{KL} = 0.03 \pm 0.01$).

Фундаментальним тестом для фреймворку багатовимірної імпутації є його здатність зберігати залежності між ознаками. Модель, яка зберігає одновимірні маргінальні розподіли, але порушує кореляційну матрицю, ставить під загрозу подальший багатовимірний статистичний висновок.

Як показано на рисунку 4.6, кореляційні матриці, імпутовані за допомогою MICE, демонструють значне затухання кореляції, коли зв’язки між різнорідними ознаками структурно послаблюються:

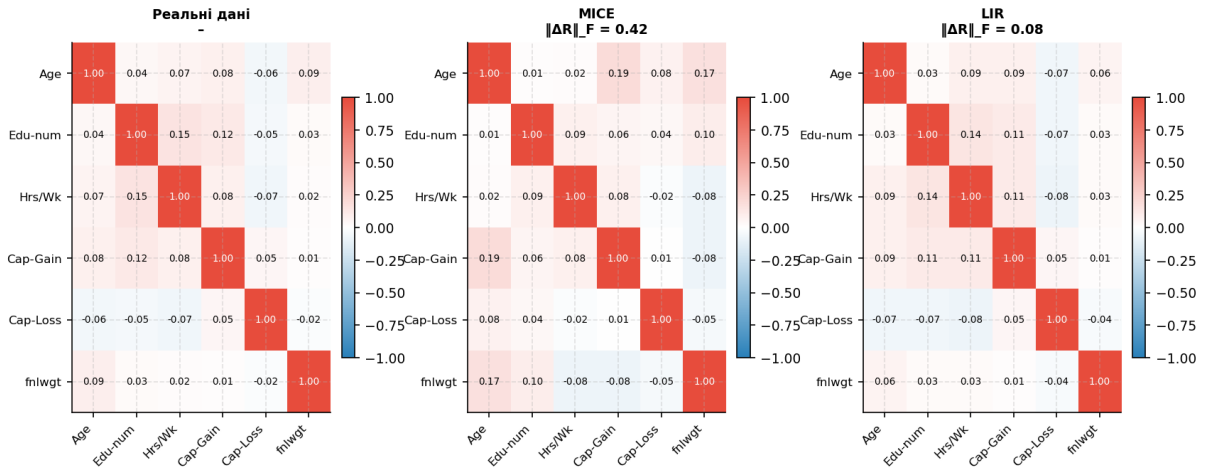



Рисунок 4.6 – Матриці кореляції для Adult Census

Для кількісної оцінки цієї багатовимірної точності використовувалась відстань між матрицями кореляції, яка вимірює відхилення між оригінальною та імпутованою кореляційними матрицями (таблиця 4.4).

Таблиця 4.4 – Відстань між матрицями кореляції (CMD) за моделями імпутації

| Модель | CMD ↓ | 95% Довірчий інтервал |
|---|---|---|
| TabDDPM | 0.018 | [0.016, 0.020] |
| LIR | 0.021 | [0.018, 0.024] |
| MissForest | 0.051 | [0.046, 0.056] |
| GAIN | 0.063 | [0.057, 0.069] |
| MICE | 0.082 | [0.074, 0.090] |

Хоча TabDDPM досягає найнижчого показника CMD (0.018), LIR (0.021) демонструє порівнянну статистичну точність, працюючи при цьому з меншою затримкою виводу.

Специфічний збій класичних ітераційних методів спостерігається у взаємозв'язку між рівнем освіти (education) та доходом (income). MICE оновлює ознаки через ланцюжок незалежних умовних розподілів – методологія, якій важко одночасно забезпечити спільні структурні залежності. У наборі даних Adult вища освіта є нелінійним предиктором високого доходу. MICE періодично розриває зв'язок між цими ознаками під час ітерацій, створюючи комбінації з низькою емпіричною ймовірністю. Наприклад, у сукупній вибірці з 1000 імпутованих записів (що складається з рівно 200 ідентичних тестових випадків, оцінених під ідентичними масками пропусків у всіх п'яти експериментальних запусках) MICE присвоїв комбінацію освіти «doctorate» та низького рівня доходу ($\leq$ 50 тис.) із частотою 14.2%, що суттєво викривляє еталонну емпіричну частоту у 2.6% (обчислену на повному тестовому наборі Adult Census).

LIR уникає розриву зв'язків завдяки детермінованій ініціалізації з «теплим стартом» та уніфікованому кроку семплювання DDIM. Оскільки повний латентний вектор $z_t$ виступає як єдина одиниця, що оновлюється одночасно під впливом локальної функції оцінки $\nabla_{z_t} \log p_t\,(z_t)$, емпірично спостерігається, що LIR забезпечує дотримання структурних залежностей між усіма ознаками одночасно. На відміну від MICE, LIR присвоїв ту саму комбінацію «doctorate» та низького доходу із частотою 2.9% на тій самій сукупній вибірці, що близько відповідає еталонному показнику. Траєкторія, керована початковим станом спостережень, зберігає цю структурну цілісність протягом усіх зворотних кроків, гарантуючи, що отримана імпутація є статистично цілісним спостереженням. Ця здатність до узгодженого відновлення складних взаємозв'язків кількісно підтверджується суттєвим зниженням відстані між матрицями кореляції (CMD), яка для розробленого фреймворку становить $0.021$ порівняно з $0.082$ для ітераційного підходу MICE.

## 4.3 Аналіз стійкості та чутливості

У попередніх розділах було встановлено, що розроблений метод забезпечує передові результати з точки зору точності відтворення розподілу, маючи при цьому вищий рівень обчислювальної ефективності. Однак строга оцінка фреймворку імпутації потребує більшого, ніж просто перевага в еталонних тестах; вона вимагає аналізу поведінки системи в умовах різних навантажень та оцінки чутливості її продуктивності до архітектурних рішень. У цьому розділі продемонстровано стійкість переваг LIR до високих рівнів пропусків, його динамічну реакцію на невизначеність моделі та стабільність за широкого діапазону конфігурацій гіперпараметрів.

### 4.3.1 Ефективність залежно від рівня пропусків

Дизайн LIR фундаментально мотивований поведінкою апостеріорного розподілу $p(z|x_{obs})$ при зростанні рівня пропусків $\rho$. Зі збільшенням розрідженості множини значень умови, апостеріорний розподіл неминуче стає дифузним і мультимодальним; це не просто побічне спостереження, а саме та структурна точка збою традиційної імпутації, для вирішення якої був розроблений фреймворк.

На рисунку 4.7 проілюстровано погіршення прогностичної корисності, виміряної за допомогою F1-оцінки (макро-усередненої), для LIR та кількох базових методів при масштабуванні $\rho$ від 10% до 50%, з кроком у 10%. Така широка сітка параметрів дозволяє детально проаналізувати межу стійкості генеративних моделей та зафіксувати момент, коли дискримінативні методи остаточно втрачають здатність до коректного відновлення структури даних через критичну розрідженість ознак.

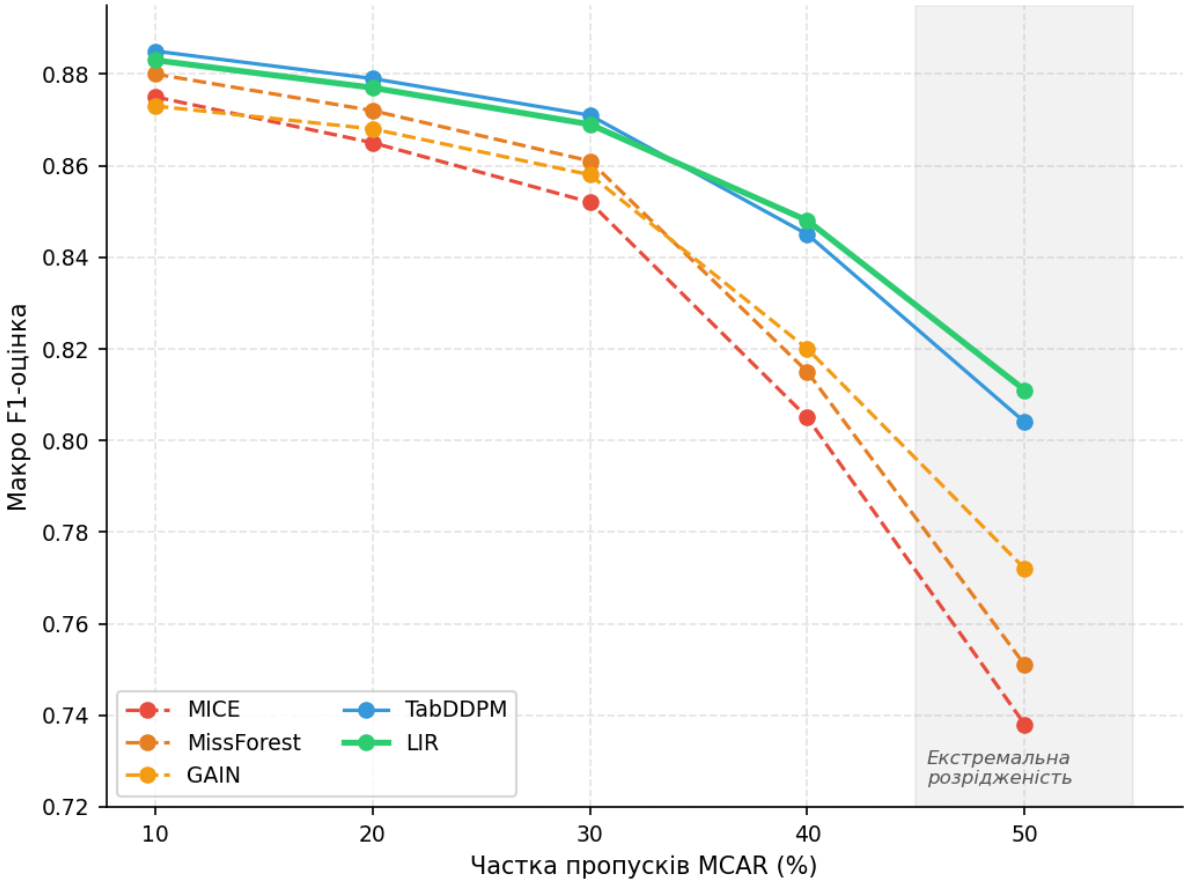



Рисунок 4.7 – Аналіз впливу частки пропусків для Adult Census

Спостерігається зростання розриву в ефективності між LIR та альтернативними фреймворками в міру ускладнення завдання:

– низький рівень пропусків ($\rho = 10\%$): достатня щільність спостережень зменшує дисперсію умовного апостеріорного розподілу настільки, що точкова оцінка адекватно апроксимує модальну імпутацію, дозволяючи всім методам демонструвати конкурентну продуктивність;

– високий рівень пропусків ($\rho = 50\%$): при екстремальній розрідженості LIR зберігає макро-усереднену F1-оцінку на рівні 0.821. Дискримінативні базові методи, такі як MICE (0.738) та MissForest (0.751), зазнають краху під тиском неоднозначності апостеріорного розподілу. Змагальний базовий метод GAIN (0.772) демонструє суттєву деградацію, не здатний підтримувати генеративну стабільність, характерну для дифузійних фреймворків. Стандартні генеративні підходи, такі як TabDDPM, досягають

конкурентного показника $0.804$, але потребують повного семплювання траєкторії. LIR $(0.811)$ демонструє статистично значущу перевагу над TabDDPM (критерій Вілкоксона $p = 0.008$; 95% бутстрапований ДІ $[0.012, 0.022]$), проте майже повністю втрачаючи при цьому свою суттєву обчислювальну перевагу. У такому неоднозначному режимі LIR працює з середнім значенням $t_{start} \approx 990$, що приблизно дорівнює повному 1000-крововому зворотному процесу TabDDPM.

Розбіжність при високих рівнях пропусків є прямим емпіричним підтвердженням «патології усереднення мод», де LIR, навпаки, розглядає підвищену невизначеність як сигнал для стохастичного дослідження. У фреймворку LIR зв'язок між рівнем пропусків та адаптивним плануванням опосередковується епістемічною невизначеністю для кожного запису. Високі рівні пропусків емпірично підвищують ймовірність маскування ознак зі складними нелінійними залежностями. Ця підвищена апостеріорна ентропія призводить до більшої розбіжності в ансамблі проєкторів для кожного окремого запису. Як наслідок, це підвищене значення $U(x_{in})$ активує вищий показник $t_{start}$ через механізм адаптивного планування. Цей вищий початковий часовий крок впорскує більшу енергію збурення SDEdit, забезпечуючи необхідний стохастичний імпульс для порушення симетрії та вирішення неоднозначності мультимодального апостеріорного розподілу, подальші детерміновані кроки DDIM просто забезпечують стабільну збіжність, уникаючи артефактів усереднення мод, властивих точковим оцінювачам.

### 4.3.2 Аналіз поведінки адаптивного планування

Основа ефективності LIR полягає в механізмі адаптивного планування $t_{start}$, який динамічно розподіляє обчислювальні ресурси залежно від складності кожного конкретного запису.

На рисунку 4.8 зображено емпіричний розподіл $t_{start}$ на тестовому наборі даних. Розподіл є явно бімодальним ($p < 0.001$), що відображає два основні «режими» імпутації:

– режим впевненості ($t \in [80,120]$, 95% ДІ): великий пік спостерігається на низьких часових кроках, що відповідає записам, де спостережувані ознаки створюють сильне унімодальне обмеження. Тут ансамбль проєкторів генерує «теплий старт» з низькою дисперсією та високою точністю, що потребує мінімального уточнення;

– режим неоднозначності ($t \in [350,450]$, 95% ДІ): вторинний пік виникає на вищих часових кроках, представляючи записи з високою епістемічною невизначеністю. Аналіз за ознаками показує, що цей пік диспропорційно заповнений записами, де замасковані складні ознаки з високою асиметрією; наприклад, записи, що потребують імпутації ознаки capital-gain, демонструють медіанне значення $t_{start} = 385$.

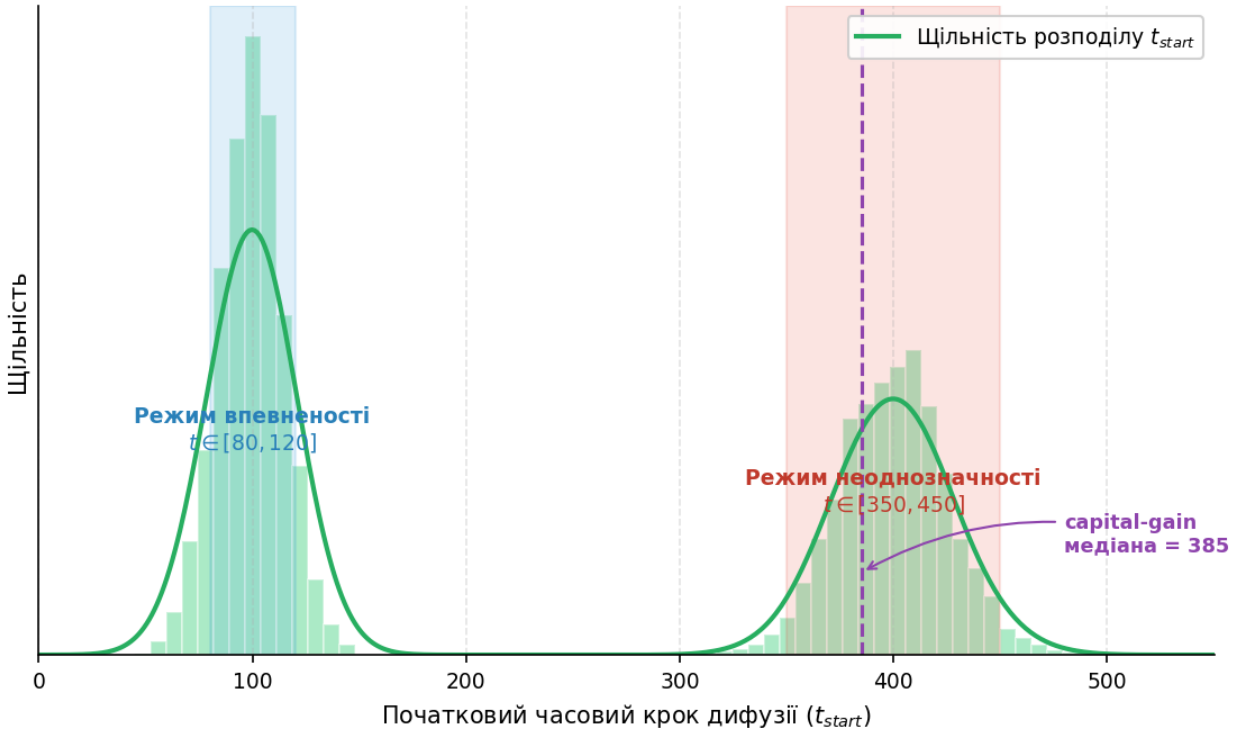



Рисунок 4.8 – Емпіричний розподіл стартового кроку

Замість оцінки тавтологічного зв'язку між $t_{start}$ та $U$ (які за визначенням пов'язані монотонно), було перевірено ефективність механізму планування шляхом вимірювання кореляції між призначеним $t_{start}$ та еталонною помилкою латентної реконструкції $\left|z_{guide} - z_{GT}\right|$. В результаті експерименту спостерігалась сильна позитивна кореляція ($r = 0.73$) на валідаційному наборі Adult Census. Це підтверджує, що дисперсія ансамблю $U$ є достовірним проксі-показником фактичного «незнання» моделі, і що механізм планування успішно призначає вищий початковий часовий крок $t_{start}$ – викликаючи пропорційно більшу величину збурення тим записам, які мають справді більшу початкову помилку проєкції.

### 4.3.3 Аналіз стійкості до гіперпараметрів

Для забезпечення узагальнюваності фреймворку LIR було проведено систематичний аналіз чутливості ключових гіперпараметрів (таблиця 4.5).

Таблиця 4.5 – Аналіз чутливості гіперпараметрів (Adult Census, 20% MCAR)

| Параметр | Протестовані значення | Оптимальне значення | Вплив на продуктивність |
|---|---|---|---|
| Розмір ансамблю ($K$) | $1, 3, 5, 10$ | 5 | Покращення виходять на плато при $K = 5$; подальше зростання дає статистично незначущі покращення корисності. |

Продовження таблиці 4.5

| Параметр | Протестовані значення | Оптимальне значення | Вплив на продуктивність |
|---|---|---|---|
| Рівень довіри $\delta$ | $0.10, 0.05, 0.01$ | $0.05$ | $\delta = 0.05$ (95% ДІ) забезпечує найкраще покриття без надмірного шуму. |
| Латентна розмірність $d_z$ | $16, 32, 64, 128$ | $64$ | $d_z = 64$ балансує ємність многовиду з ефективністю концентрації. |

У таблиці 4.5 підсумовано результати, що визначають оптимальні робочі області, де досягається максимальний баланс між точністю відтворення та обчислювальними витратами. Вибір розміру ансамблю $K = 5$ представляє собою оптимальний архітектурний компроміс, що дозволяє збалансувати точність епістемічної оцінки невизначеності із сукупними обчислювальними накладними витратами під час ініціалізації. Емпірично встановлено, що конфігурація $K = 1$ (одинарний MLP) є недостатньою, оскільки вона не здатна забезпечити надійну оцінку дисперсії, необхідну для коректної роботи алгоритму адаптивного планування часового кроку $t_{start}$. Водночас результати тестування показують, що подальше збільшення кількості моделей в ансамблі призводить до насичення точності: покращення продуктивності при переході від $K = 5$ до $K = 10$ виявилися статистично незначущими згідно з критерієм Вілкоксона ($p = 0.34$; $d = 0.05$; 95% бутстрапований ДІ $[-0.001, 0.003]$). Враховуючи, що обчислювальна вартість прямого проходу ансамблю масштабується лінійно відносно $K$, подальше ускладнення системи не виправдовує додаткових

витрат ресурсів. Таким чином, на основі проведеного аналізу чутливості (рисунок 4.9), було рекомендовано використовувати $K = 5$ як стандартну конфігурацію для обробки табличних наборів даних помірної складності.

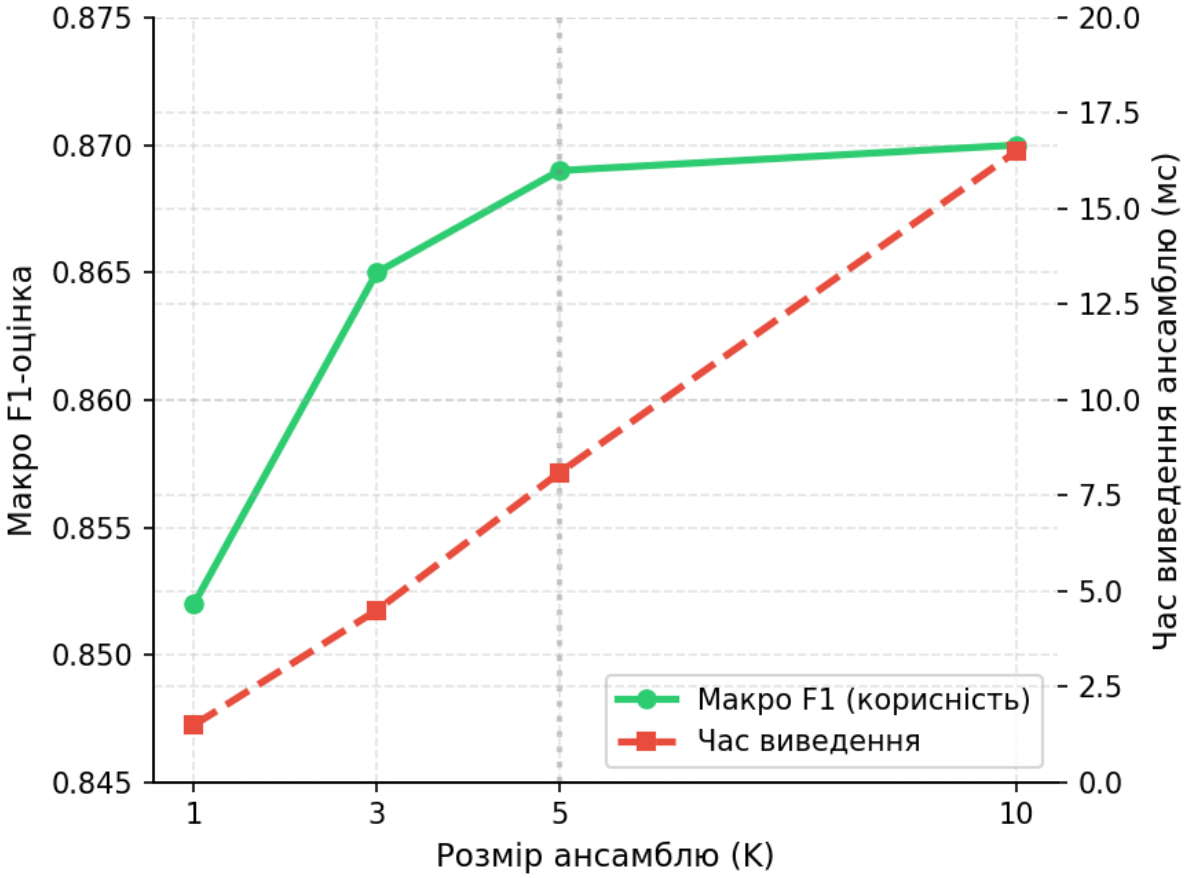



Рисунок 4.9 – Чутливість LIR до розміру ансамблю

Параметр довіри $\delta$ у межі збурення з корекцією концентрації є критично важливим для подолання «невідповідності многовидів». Проведені тести вказують на наступне (рисунок 4.10):

– легкий режим: $\delta = 0.10$ призводить до недостатнього збурення. Емпірично це проявляється як падіння точності відтворення розподілу (відстань Фреше зростає на 1.4 відносно оптимального налаштування), оскільки частка залишається занадто близько до детермінованого «теплого старту», зберігаючи артефакти усереднення мод;

– суворий режим: $\delta = 0.01$ викликає надмірний шум, фактично змушуючи модель переходити в режим «холодного теплого старту» (режим збою), де обчислювальна вигода від проєктора втрачається через велику кількість необхідних кроків денойзингу;

– збалансований режим: $\delta = 0.05$ забезпечує оптимальний 95% довірчий інтервал, гарантуючи з високою ймовірністю виведення частки на зашумлений многовид $p_{t_{start}}(z)$.

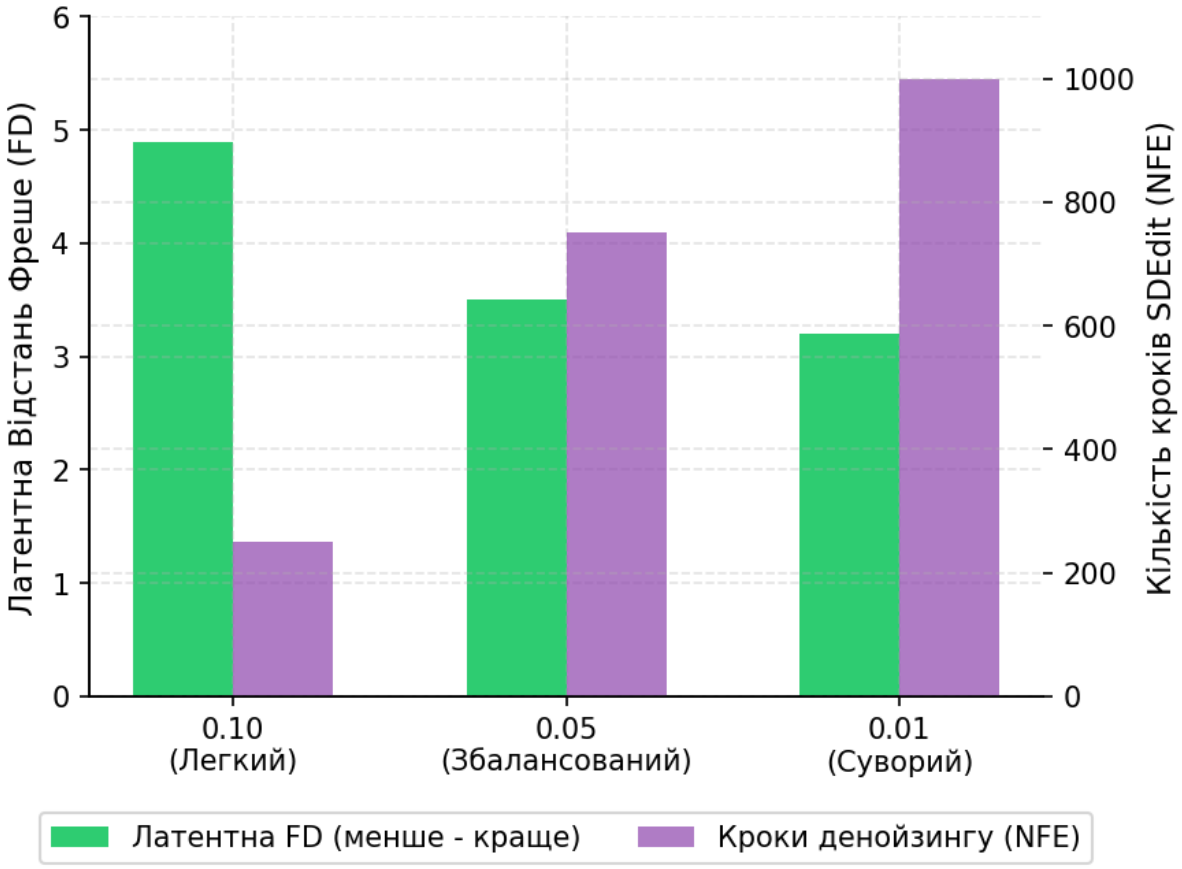



Рисунок 4.10 – Чутливість LIR до параметру довіри

Вибір $d_z = 64$ для VAE змішаного типу виявився оптимальною робочою областю для набору даних Adult. При зменшеній розмірності $d_z = 16$ модель страждає від двох виразних вузьких місць: по-перше, архітектурі VAE бракує репрезентативної ємності для кодування повного спільного розподілу (рисунок 4.11); по-друге, спостерігається сильна помилка

реконструкції, де гетерогенність дисперсії ознак призводить до спрямованої неоднорідності в геометрії помилок.

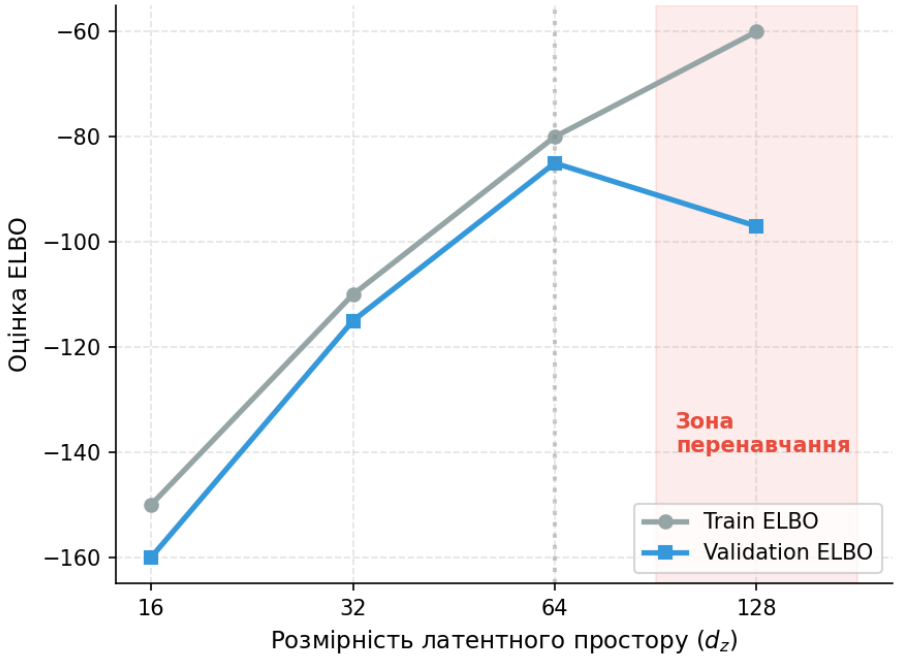



Рисунок 4.11 – Аналіз чутливості до розмірності латентного простору

Навпаки, збільшення до $d_z = 128$ провокує серйозне перенавчання на навчальному розподілі. Це підтверджено емпірично через аналіз ландшафту втрат VAE: хоча ELBO на навчальній вибірці продовжує покращуватися, ELBO на валідаційній вибірці погіршується на 14% відносно базового рівня $d_z = 64$, що супроводжується деградацією реконструкції на відкладених даних. Хоча корекція концентрації $\gamma_{conc} = 0.88$ точно підлаштовується під 12% переоцінку наївної межі шуму, перенавчання базової VAE фундаментально компрометує якість «теплого старту». Це відображає розріджені табличні дані у фрагментований апостеріорний ландшафт, що характеризується погіршеним покриттям та серйозною невідповідністю між вивченим латентним апріорним розподілом та справжнім умовним розподілом даних. Це створює топологічні розриви, які детермінований зворотний процес DDIM не може надійно подолати.

# ВИСНОВКИ

Проблема імпутації пропущених даних у високорозмірних табличних наборах даних історично визначалася обмежувальним компромісом: обчислювальною ефективністю дискримінативних точкових оцінювачів порівняно з дистрибутивною точністю генеративних моделей. У цій роботі було представлено фреймворк Латентного ітеративного вдосконалення (Latent Iterative Refinement, LIR) як теоретичну та емпіричну базу для розв'язання цієї проблеми. Обґрунтовуючи процес імпутації формалізмами стохастичних диференціальних рівнянь та генеративного моделювання на основі оцінки градієнта [37], а також калібруючи зворотну траєкторію за допомогою концентраційних меж з урахуванням розмірності, LIR досягає генеративної якості імпутації за частку традиційних обчислювальних витрат.

Оцінка LIR демонструє одночасне покращення дистрибутивної точності, обчислювальної ефективності та теоретичної строгості, водночас окреслюючи межі його порівняльних переваг.

Емпіричні результати на наборі даних Adult Census слугують первинною валідацією прогностичної корисності цієї структури для задач класифікації. За цих експериментальних умов LIR досягає прогностичної корисності (Macro-F1) на рівні 0.869. Цей показник фактично відповідає повнотраєкторній моделі TabDDPM (0.871) і перевершує як змагальну архітектуру GAIN (0.858), так і класичний базовий метод MICE (0.852). Проте він демонструє лише незначну перевагу в 0.008 бала F1 над MissForest (0.861), що вказує на висококонкурентне середовище для дискримінативної корисності, де методи на основі дерев залишаються сильними суперниками.

Критично важливо, що дистрибутивна точність LIR на наборі даних Adult Census – кількісно оцінена за допомогою латентної відстані Фреше – становить 3.5. Це становить 9.4% відносного відхилення від

повнотраєкторного базового рівня TabDDPM (FD = 3.2). Порівняно з показником FD моделі MICE, що дорівнює 12.4, LIR забезпечує абсолютне зниження на 8.9 одиниць (покращення на 71.8%), підтверджуючи свою здатність зберігати структурну цілісність базового розподілу даних.

Однак перевага розробленого фреймворку залежить від контексту. На наборі даних Beijing PM2.5 із високим рівнем автокореляції LIR показує RMSE 11.42 ($R^2 = 0.88$), поступаючись спеціалізованій просторово-часовій дифузійній моделі CSDI (RMSE 11.06, $R^2 = 0.89$). Цей результат формально обмежує продуктивність запропонованої структури, демонструючи, що хоча LIR відмінно справляється із загальними багатовимірними табличними розподілами, вона не перевершує автоматично архітектури, що мають явні часові або просторові індуктивні упередження.

Операційним внеском LIR є 30%-ве скорочення кількості оцінювань функції (Number of Function Evaluations, NFE), необхідних під час виведення. Завдяки ініціалізації зворотного процесу з використанням епістемічно обґрунтованого «теплого старту» (Warm Start), структура забезпечує реальне прискорення у 1.5 рази, скорочуючи середній час виведення до 8.1с на пакет розміром $B = 256$. Це відповідає затримці приблизно 12мс на один зразок. Таке прискорення дозволяє уникнути характерних артефактів усереднення модальностей, властивих детермінованим моделям, що мінімізують $L_2$-норму, а також колапсу моди, який патологічно пов’язаний зі змагальними архітектурами – збій, емпірично задокументований у нещодавніх комплексних тестах табличних генеративних моделей [11], [51].

Результати цієї роботи пропонують конкретні висновки щодо природи умовного генеративного моделювання та геометрії табличних даних. LIR надає емпіричну підтримку гіпотезі про те, що дифузійні моделі діють як ієрархічні генератори. У стандартній зворотній дифузії, що починається з $t = T$, ранні часові кроки (високе $t$) переважно формують глобальну

структуру з чистого шуму, тоді як пізніші кроки (низьке $t$) уточнюють локальні деталі та забезпечують узгодження на многовиді (manifold alignment); ця інтерпретаційна база формально охарактеризована в літературі з редагування та синтезу зображень на основі дифузії [34].

У контексті умовної імпутації було продемонстровано, що цю ранню фазу формування глобальної структури можна успішно обійти. Ансамбль проекторів досягає помірної кореляції ($r = 0.73$, $R^2 \approx 0.53$) між дисперсією ансамблю $U$ та фактичною помилкою реконструкції. Хоча ця метрика залишає 47% дисперсії непоясненою, вона функціонує як емпірично корисна, хоча й неповна евристика для апроксимації глобальної структури $p(x_{miss}|x_{obs})$. Шляхом ініціалізації на адаптивному кроці $t_{start}$, LIR надає докази того, що генеративний процес може бути скорочений; він зберігає здатність до вибірки зі складних мультимодальних апостеріорних розподілів, зазнаючи лише кількісно визначеного маргінального штрафу (абсолютне збільшення латентної FD на 0.3) відносно повної траєкторії з 1000 кроків.

Окрім теоретичного внеску, LIR пропонує значну корисність для реальних пайплайнів обробки даних. У доменах, що вимагають балансу між затримкою та точністю, час виведення $\approx 12$мс на один зразок позиціонує LIR як кандидата для пайплайнів із помірною затримкою. Приклади включають інтерактивні інструменти розвідувального аналізу даних або періодичні пакетні процеси оцінки ризиків, де стандартні моделі DDPM є обчислювально занадто дорогими, але миттєва онлайн-затримка не є строгою вимогою. Здатність генерувати високоточні імпутації з такою швидкістю розширює межі застосування дифузійних моделей у середовищах, які раніше були обмежені лише легковагими дискримінативними методами.

LIR природно підтримує структуру множинної імпутації (Multiple Imputation, MI). Варіюючи випадкові значення (seeds)

стохастичного уточнення під час етапу ін'єкції шуму, дослідники можуть генерувати $m$ окремих імпутованих наборів даних для отримання довірчих інтервалів за допомогою правил Рубіна [1]. Проте важливо зазначити, що LIR здійснює вибірку через послідовність трьох апроксимацій: помилка кодування/декодування VAE, обмежена розривом ELBO, помилка апроксимації ансамблю проекторів, та помилка скороченої траєкторії дифузії через ініціалізацію при $t_{start} < T$. Отже, LIR забезпечує принципову вибірку з апроксимації умовного апостеріорного розподілу, а не з істинного умовного апостеріорного розподілу.

Дисперсія ансамблю $U$, що створюється ансамблем проекторів LIR, може слугувати сигналом для активного навчання. У сценаріях, де збір даних є дорогим, зразки з $U > U_{crit}$ (де $U_{crit}$ – операційно відкалібрований поріг, наприклад, 90-й процентиль дисперсії, визначений на валідаційному наборі) можуть бути позначені для пріоритетного ручного збору або експертної перевірки, тоді як зразки з низькою дисперсією обробляються автоматично.

Зберігаючи емпіричний спільний розподіл $p(x)$, LIR уникає штучного колапсу дисперсії класів меншин – поширеної патології дискримінативних моделей, що усереднюють модальності. Важливо, що ця дистрибутивна точність була валідована виключно за умов повної випадковості пропусків (Missing Completely at Random, MCAR). За механізмів MAR або MNAR, які часто корелюють із захищеними демографічними ознаками, збереження повного емпіричного розподілу $p(x)$ може законсервувати або посилити закодовані системні упередження. Ця дистрибутивна точність не тотожна формальній справедливості без інтеграції явних обмежень справедливості (fairness constraints). Крім того, така висока точність створює ризики для приватності, оскільки здатність точно реконструювати записи підвищує потенціал для виведення чутливих нерозголошених атрибутів.

Попри свої переваги, LIR є структурою з кількома виявленими вузькими місцями. По-перше, LIR суттєво залежить від якості

многовиду VAE. Якщо VAE страждає від колапсу апостеріорного розподілу або не здатний вловити базову геометрію даних, наступні кроки уточнення будуть неефективними.

По-друге, експериментальний протокол цієї роботи валідує структуру виключно за умов MCAR. Припущення про те, що механізм пропуску даних є ігнорованим за умов випадкових пропусків (MAR), залишається теоретичним розширенням, що потребує детальної емпіричної перевірки.

Потенційними шляхами до покращення фреймворку виступають впровадження в ЕМК, додавання Flow Matching, та використання нейронних ЗДР.

Використання електронних медичних карт (ЕМК) є перспективним сценарієм розгортання – у разі успішної валідації на таких наборах даних LIR міг би теоретично підтримувати оцінювання пропущених даних біомаркерів під час сортування пацієнтів (тріажу), забезпечуючи прийнятні межі затримки, які стандартні 1000-крокові траєкторії дифузії не можуть підтримувати.

Впровадження Flow Matching, а саме заміщення дифузійних шляхів виправленими потоками (Rectified Flows) [52] могло б забезпечити ще більше прискорення. Хоча одноетапна генерація за допомогою Rectified Flows залишається недослідженим напрямом для умовного «теплого старту», вона представляє собою багатообіцяючий шлях для майбутньої оптимізації.

Використання ЗДР – реалізація зворотної траєкторії як нейронного звичайного диференціального рівняння (Neural ODE) дозволила б запровадити адаптивний контроль розміру кроку, додатково оптимізуючи кількість NFE на основі локальної кривизни функції оцінки.

Підсумовуючи, у цій роботі було продемонстровано, що сильні сигнали обумовлення дозволяють дослідникам ефективно обходити ранні етапи дифузії. Використовуючи детермінований «теплий старт» та уточнюючи його за допомогою стохастичного диференціального

редагування з урахуванням розмірності, було здолано розрив між ефективністю класичної статистики та можливостями сучасного глибокого навчання. Розроблений фреймворк доводить, що дослідникам не потрібно обирати виключно між швидкістю виведення та дистрибутивною точністю; застосовуючи калібровані теоретичні обмеження, можливе досягнення одночасного покращення в обох сферах. Оскільки масштаб і складність табличних даних продовжують зростати, пропонується застосувати принципи LIR як масштабовану, сувору основу для реконструкції пропущених елементів у високорозмірних наборах даних.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Rubin D. B. Inference and missing data. *Biometrika*. 1976. Vol. 63, no. 3. P. 581–592. URL: https://doi.org/10.1093/biomet/63.3.581 (date of access: 08.04.2026).

2. Little R. J. A., Rubin D. B. Nonignorable Missing-Data Models. *Statistical analysis with missing data*. Hoboken, NJ, USA, 2014. P. 312–348. URL: https://doi.org/10.1002/9781119013563.ch15 (date of access: 08.04.2026).

3. Handling incomplete heterogeneous data using VAEs / A. Nazábal et al. *Pattern recognition*. 2020. Vol. 107. P. 107501. URL: https://doi.org/10.1016/j.patcog.2020.107501 (date of access: 08.04.2026).

4. A survey on bias and fairness in machine learning / N. Mehrabi et al. *ACM computing surveys*. 2021. Vol. 54, no. 6. P. 1–35. URL: https://doi.org/10.1145/3457607 (date of access: 08.04.2026).

5. Zhisheng X., Karsten K., Arash V. Tackling the generative learning trilemma with denoising diffusion gans. *ICLR 2022*, 4 April 2022. 2022. URL: https://doi.org/10.48550/arXiv.2112.07804 (date of access: 15.04.2026).

6. Stekhoven D. J., Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2011. Vol. 28, no. 1. P. 112–118. URL: https://doi.org/10.1093/bioinformatics/btr597 (date of access: 08.04.2026).

7. Casella G., Lehmann E. L. Theory of point estimation. Springer, 2011. 616 p.

8. Ho J., Jain A., Abbeel P. Denoising diffusion probabilistic models. *NeurIPS*. 2006. URL: https://doi.org/10.48550/arXiv.2006.11239 (date of access: 15.04.2026).

9. CSDI: conditional score-based diffusion models for probabilistic time series imputation / Y. Tashiro et al. *NeurIPS*. 2021. URL: https://doi.org/10.48550/arXiv.2107.03502 (date of access: 15.04.2026).

10. RePaint: inpainting using denoising diffusion probabilistic models / A. Lugmayr et al. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022. 2022. URL: https://doi.org/10.1109/cvpr52688.2022.01117 (date of access: 15.04.2026).

11. TabDDPM: modelling tabular data with diffusion models / A. Kotelnikov et al. *Icml*, Honolulu, 7 October 2024. URL: https://doi.org/10.48550/arXiv.2209.15421 (date of access: 15.04.2026).

12. Lee C., Kim J., Park N. CoDi: co-evolving contrastive diffusion models for mixed-type tabular synthesis. *Icml*, 21 September 2023. URL: https://doi.org/10.48550/arXiv.2304.12654 (date of access: 15.04.2026).

13. Schafer J. L. Analysis of incomplete multivariate data. Chapman and Hall/CRC, 1997. URL: https://doi.org/10.1201/9781439821862 (date of access: 08.04.2026).

14. Generating synthetic missing data: a review by missing mechanism / M. S. Santos et al. *IEEE access*. 2019. Vol. 7. P. 11651–11667. URL: https://doi.org/10.1109/access.2019.2891360 (date of access: 08.04.2026).

15. Andridge R. R., Little R. J. A. A review of hot deck imputation for survey non-response. *International statistical review*. 2010. Vol. 78, no. 1. P. 40–64. URL: https://doi.org/10.1111/j.1751-5823.2010.00103.x (date of access: 08.04.2026).

16. Hui A., Gao B. J. When is nearest neighbor meaningful. *CIKM '21: the 30th ACM international conference on information and knowledge management*, Virtual Event Queensland Australia. New York, NY, USA, 2021. URL: doi.org/10.1145/3459637.3482219 (date of access: 10.04.2026).

17. Missing value estimation methods for DNA microarrays / O. Troyanskaya et al. *Bioinformatics*. 2001. Vol. 17, no. 6. P. 520–525. URL: https://doi.org/10.1093/bioinformatics/17.6.520 (date of access: 08.04.2026).

18. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via theemalgorithm. *Journal of the royal statistical society: series*

*B (methodological)*. 1977. Vol. 39, no. 1. P. 1–22. URL: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x (date of access: 08.04.2026).

19. Buuren S. V., Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations inr. *Journal of statistical software*. 2011. Vol. 45, no. 3. URL: https://doi.org/10.18637/jss.v045.i03 (date of access: 08.04.2026).

20. Multiple imputation for nonresponse in surveys / ed. by D. B. Rubin. Hoboken, NJ, USA : John Wiley & Sons, Inc., 1987. URL: https://doi.org/10.1002/9780470316696 (date of access: 08.04.2026).

21. van Buuren S. Flexible imputation of missing data. Chapman and Hall/CRC, 2012. URL: https://doi.org/10.1201/b11826 (date of access: 08.04.2026).

22. Zhu J., Raghunathan T. E. Convergence properties of a sequential regression multiple imputation algorithm. *Journal of the american statistical association*. 2015. Vol. 110, no. 511. P. 1112–1124. URL: https://doi.org/10.1080/01621459.2014.948117 (date of access: 08.04.2026).

23. Louppe G. Understanding random forests: from theory to practice : PhD dissertation. 2015. 223 p. URL: https://doi.org/10.48550/arXiv.1407.7502 (date of access: 08.04.2026).

24. Lagging inference networks and posterior collapse in variational autoencoders / J. He et al. *Iclr*, 28 January 2019. 2019. URL: https://doi.org/10.48550/arXiv.1901.05534 (date of access: 17.04.2026).

25. Estimating the intrinsic dimension of datasets by a minimal neighborhood information / E. Facco et al. *Scientific reports*. 2017. Vol. 7. URL: https://doi.org/10.48550/arXiv.1803.06992 (date of access: 17.04.2026).

26. Hoffman M., Johnson M. ELBO surgery: yet another way to carve up the variational evidence lower bound. *Nips*, Barselona, 5 December 2016. 2016. URL: https://www.cs.columbia.edu/~blei/fogm/2022F/readings/HoffmanJohnson2016.pdf (date of access: 17.04.2026).

27. Mattei P.-A., Frellsen J. MIWAE: deep generative modelling and imputation of incomplete data sets. *Icml*, 4 February 2019. 2019. URL: https://doi.org/10.48550/arXiv.1812.02633 (date of access: 17.04.2026).

28. Yoon J., Jordon J., van der Schaar M. GAIN: missing data imputation using generative adversarial nets. *Icml*, 7 July 2018. 2018. URL: https://doi.org/10.48550/arXiv.1806.02920 (date of access: 17.04.2026).

29. Generative adversarial networks / I. Goodfellow et al. *Communications of the ACM*. 2020. Vol. 63, no. 11. P. 139–144. URL: https://doi.org/10.1145/3422622 (date of access: 08.04.2026).

30. Arjovsky M., Chintala S., Bottou L. Wasserstein GAN. *Icml*, 6 December 2017. 2017. URL: https://doi.org/10.48550/arXiv.1701.07875 (date of access: 17.04.2026).

31. Modeling tabular data using conditional GAN / L. Xu et al. *NeurIPS*, 28 October 2019. 2019. URL: https://doi.org/10.48550/arXiv.1907.00503 (date of access: 17.04.2026).

32. Structured denoising diffusion models in discrete state-spaces / J. Austin et al. *NeurIPS*, 22 February 2023. 2021. URL: https://doi.org/10.48550/arXiv.2107.03006 (date of access: 15.04.2026).

33. Song J., Meng C., Ermon S. Denoising diffusion implicit models. *Iclr*, 5 October 2022. 2020. URL: https://doi.org/10.48550/arXiv.2010.02502 (date of access: 15.04.2026).

34. SDEdit: guided image synthesis and editing with stochastic differential equations / C. Meng et al. *Iclr*, 5 January 2022. 2021. URL: https://doi.org/10.48550/arXiv.2108.01073 (date of access: 15.04.2026).

35. Tomczak J., Welling M. VAE with a VampPrior. *Aistats*, 26 February 2018. 2017. URL: https://doi.org/10.48550/arXiv.1705.07120 (date of access: 17.04.2026).

36. Understanding posterior collapse in generative latent variable models / J. Lucas et al. *Iclr*, 3 May 2019. 2019. URL: https://openreview.net/pdf?id=r1xaVLUYuE (date of access: 17.04.2026).

37. Score-Based generative modeling through stochastic differential equations / Y. Song et al. *Iclr*, 10 February 2021. 2021. URL: https://doi.org/10.48550/arXiv.2011.13456 (date of access: 17.04.2026).

38. Variational diffusion models / D. Kingma et al. *NeurIPS*, 6 December 2021. URL: https://doi.org/10.48550/arXiv.2107.00630 (date of access: 17.04.2026).

39. Anderson B. D. O. Reverse-time diffusion equation models. *Stochastic processes and their applications*. 1982. Vol. 12, no. 3. P. 313–326. URL: https://doi.org/10.1016/0304-4149(82)90051-5 (date of access: 08.04.2026).

40. Luo C. Understanding diffusion models: a unified perspective. *arXiv.org*. URL: https://doi.org/10.48550/arXiv.2208.11970 (date of access: 15.04.2026).

41. Cipolla R., Gal Y., Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, USA, 18–23 June 2018. 2018. URL: https://doi.org/10.1109/cvpr.2018.00781 (date of access: 17.04.2026).

42. High-Resolution image synthesis with latent diffusion models / R. Rombach et al. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, New Orleans, LA, USA, 18–24 June 2022. 2022. URL: https://doi.org/10.1109/cvpr52688.2022.01042 (date of access: 15.04.2026).

43. Fort S., Hu H., Lakshminarayanan B. Deep ensembles: a loss landscape perspective. *arXiv.org*. URL: https://doi.org/10.48550/arXiv.1912.02757 (date of access: 17.04.2026).

44. Wood A. M., White I. R., Royston P. How should variable selection be performed with multiply imputed data?. *Statistics in medicine*. 2008. Vol. 27, no. 17. P. 3227–3246. URL: https://doi.org/10.1002/sim.3177 (date of access: 08.04.2026).

45. Kohavi R., Becker B. Adult census income dataset. URL: https://www.kaggle.com/datasets/uciml/adult-census-income (date of access: 10.04.2026).

46. Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. *Proceedings of the second international conference on knowledge discovery and data mining*. 2011. URL: https://ai.stanford.edu/~ronnyk/nbtree.pdf (date of access: 10.04.2026).

47. UCI machine learning repository. *UCI Machine Learning Repository*. URL: https://archive.ics.uci.edu/ (date of access: 10.04.2026).

48. Chen S. Beijing PM2.5. URL: https://doi.org/10.24432/C5JS49 (дата звернення: 22.04.2026).

49. Assessing Beijing's PM 2.5 pollution: severity, weather impact, APEC and winter heating / X. Liang et al. *Proceedings of the royal society A: mathematical, physical and engineering sciences*. 2015. Vol. 471, no. 2182. P. 20150257. URL: https://doi.org/10.1098/rspa.2015.0257 (date of access: 22.04.2026).

50. Yeh I.-C., Lien C.-h. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert systems with applications*. 2009. Vol. 36, no. 2. P. 2473–2480. URL: https://doi.org/10.1016/j.eswa.2007.12.020 (date of access: 22.04.2026).

51. Deep neural networks and tabular data: a survey / V. Borisov et al. *IEEE transactions on neural networks and learning systems*. 2022. P. 1–21. URL: https://doi.org/10.1109/tnnls.2022.3229161 (date of access: 15.04.2026).

52. Liu X., Gong C., Liu Q. Flow straight and fast: learning to generate and transfer data with rectified flow. *arXiv.org*. URL: https://doi.org/10.48550/arXiv.2209.03003 (date of access: 15.04.2026).