УДК 004.8

ПРОЄКТУВАННЯ ТА ОЦІНЮВАННЯ ЕФЕКТИВНОСТІ ДИФУЗІЙНИХ МОДЕЛЕЙ В ЗАВДАННЯХ ГЕНЕРАЦІЇ ЗОБРАЖЕНЬ

Цепочко М. Г., Безкоровайний В. В.

Харківський національний університет радіоелектроніки, Харків

У сучасній парадигмі штучного інтелекту моделі дифузійного типу стали провідним інструментом для генерації фотореалістичних зображень із текстових описів. Їхня перевага полягає у здатності апроксимувати складні розподіли даних у багатовимірних просторах без ризику колапсу моделі, який часто спостерігається у генеративно-змагальних мережах. Основна ідея дифузійних моделей полягає у поступовому перетворенні даних у шум і відновленні їх назад шляхом моделювання стохастичного зворотного процесу, який керується параметризованою нейронною мережею.

Таким чином, замість прямого навчання моделі відтворювати реальний розподіл даних, вона навчається усувати шум, що дозволяє досягти стабільної динаміки оптимізації та високої якості реконструкції [1].

Дифузійна модель описується двома основними процесами: прямим, який додає шум до зображення, та зворотним, який реконструює зображення з цього шуму. Прямий процес формалізується наступним чином [2]:

$$q(x_t|x_{t-1}) = \mathcal{N}\left(x_t; \sqrt{1-\beta_t}\, x_{t-1}, \beta_t I\right),$$

де β_t – вказує на кожному кроці компроміс між інформацією, яку потрібно зберегти з попереднього кроку, та шумом, який необхідно додати, а I – одинична матриця. Цей параметр β_t зазвичай збільшується з кожним кроком, щоб рівномірно додавати шум протягом усього процесу дифузії.

Відповідно, зворотний процес апроксимується нейронною мережею:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}\left(x_{t-1}; \mu_0(x_t,t), \Sigma_\theta(x_t,t)\right),$$

де $\mu_0(x_t,t)$ та $\Sigma_\theta(x_t,t)$ – параметри середнього та коваріації, що прогнозуються на кожному кроці нейронною мережею. У практичних реалізаціях часто використовується припущення про діагональність Σ_θ, що

спрощує обчислення та стабілізує градієнтне оновлення.

Оптимізація моделі здійснюється через мінімізацію очікуваної різниці між істинним шумом, доданим на етапі прямої дифузії, та шумом, який прогнозує нейронна мережа. Функція втрат формулюється як [3]:

$$L(\theta) = \mathbb{E}_{x_0,\epsilon,t}[\|\epsilon - \epsilon_0(x_t, t)\|^2],$$

де x_0 – початкове зображення; ϵ – істинний шум, згенерований з гаусівського розподілу $N(0, I)$; $\epsilon_0(x_t, t)$ – шум, передбачений моделлю, а очікування $\mathbb{E}$ береться за всіма часовими кроками t, вибраними випадково.

З архітектурного погляду дифузійні моделі реалізуються переважно на базі двох підходів: U-Net і Diffusion Transformer (DiT).

U-Net характеризується наявністю симетричної енкодер-декодерної структури зі скіп-зв'язками, що забезпечує збереження просторових ознак і деталізації на високих рівнях роздільності [4]. У той час як DiT інтегрує механізм самоуваги (Self-Attention), який дозволяє моделі враховувати глобальні контекстуальні залежності між елементами зображення.

Архітектурні блоки дифузійної моделі представлені в таблиці 1.1.

Оцінювання ефективності дифузійних моделей здійснюється за комплексом кількісних показників.

Основною метрикою є відстань початку Фреше (FID), яка визначає відстань між розподілами ознак реальних і згенерованих зображень [5]:

$$ID = \|\mu_r - \mu_g\|_2^2 + Tr\left(\Sigma_r + \Sigma_g - 2 * \sqrt{\Sigma_r * \Sigma_g}\right),$$

де μ_r та Σ_r – середнє значення і коваріаційна матриця для реальних даних, а μ_g та Σ_g – відповідні статистики для синтетичних. Менше значення FID свідчить про вищу схожість між розподілами, а отже, і про кращу візуальну достовірність. Додатково використовуються показники індексу інцепції (IS), що характеризує різноманітність зображень, та оцінка узгодженості CLIP, яка визначає семантичну відповідність між текстовим описом і згенерованим результатом.

Концептуальна логіка показника IS ґрунтується на припущенні про

наявність класифікаційної моделі, що оцінює умовний розподіл $p(y|x)$, де y – це категорія, а x – згенероване зображення. ля генеративної системи вважається оптимальним, коли кожне сформоване зображення однозначно відноситься до певного класу, тобто характеризується малою умовною ентропією $H(y|x)$.

Таблиця 1.1 – Функціональні блоки архітектури дифузійної моделі

Блок архітектури	Основне призначення	Типові реалізації та модулі	Ключовий внесок у модель
Блок попередньої обробки	Перетворення зображення або зашумленого латента у компактний простір ознак	Нормалізація, лінійні проєкції, згорткові шари	Забезпечує коректне входження сигналу в модель, вирівнює масштаб і структуру даних
Латентний автоенкодер (VAE)	Стиснення зображення в латентний простір та подальше відновлення	Енкодер VAE, Декодер VAE, стохастичний семплінг	Дозволяє працювати з компактними латентами замість пікселів, знижує обчислювальні витрати
Енкодерний блок (Downsampling)	Поступове зниження просторової роздільності та виділення високорівневих ознак	ResNet-подібні модулі, Downsampling-операції (stride=2), згортки	Формує абстрактні ознаки: глобальні форми, контури та структурні патерни
Контекстуальний блок (центральна частина)	Узгодження інформації між різними масштабами та частинами зображення	Глобальні інтеграційні модулі, крос-модальні блоки, агрегатори контексту	Створює цілісне семантичне уявлення сцени, критичне для складних композицій

Продовження таблиці 1.1

Декодерний блок (Upsampling)	Відновлення просторової структури та інтеграція локальних ознак	Upsampling (nearest/linear), ResNet-модулі відновлення, згортки	Повертає дрібні деталі, забезпечує плавність контурів
Блок архітектури	Основне призначення	Типові реалізації та модулі	Ключовий внесок у модель
Skip-з'єднання між енкодером і декодером	Передавання локальних ознак високої роздільності до декодера	Симетричні пропускні з'єднання	Зберігають локальну інформацію, запобігають втраті деталей
Вихідний декодер зображення	Перетворення латентів у фінальне зображення	Декодер VAE або згорткова мережа	Відтворює зображення з високою якістю та узгодженими текстурами

Одночасно бажано, щоб у сукупності всіх згенерованих зразків розподіл класів $p(y)$ був максимально збалансованим, що проявляється у високій ентропії $H(y|x)$.

Формально IS визначається як експонента від середнього значення дивергенції Кульбака-Лейблера між $p(y|x)$ та $p(y)$ [6]:

$$IS = xp\left(\mathbb{E}_{x\sim p_{data}}\left[KL\left(p(y|x)||p(y)\right)\right]\right),$$

де $p(y) = \int p(y|x)p_{data}(x)\,dx$ – це середній розподіл класів у згенерованих зображеннях. Чим більшим є показник IS, тим вищою вважається якість зображень. Водночас цей критерій має недоліки: він не оцінює відповідність реальному розподілу даних і залежить від класифікатора.

Отже, проєктування дифузійних моделей передбачає не лише інженерну оптимізацію архітектури, а й глибоке статистичне моделювання параметрів шуму, часової дискретизації та апроксимаційної стабільності.

Подальший розвиток дифузійних систем передбачає інтеграцію механізмів навчання з підкріпленням, гібридних трансформерів і динамічних

варіаційних процедур, що відкриває шлях до формування інтелектуальних мультимодальних моделей нового покоління, здатних до автономного візуального синтезу, контекстного дизайну та самонавчання в реальному часі з урахуванням множини показників якості [7–8].

Література:

1. O. Dalmaz, B. Saglam, G. Elmas, M. Mirza and T. Çukur. Denoising Diffusion Adversarial Models for Unconditional Medical Image Generation. 2023 *Signal Processing and Communications Applications Conference*. pp. 1-5. URL: https://ieeexplore.ieee.org/document/10223912 (дата звернення: 10.11.2025).
2. H. Nam, J. Park, J. Choi and S. L. Kim. Sequential Semantic Generative Communication for Progressive Text-to-Image Generation. *IEEE International Conference on Communication, and Networking*. 2023, pp. 91-94. URL: https://ieeexplore.ieee.org/document/10287475 (дата звернення: 10.11.2025).
3. High-Resolution Image Synthesis with Latent Diffusion Models (A.K.A. LDM & Stable Diffusion). URL: https://ommer-lab.com/research/latent-diffusion-models/ (дата звернення: 10.11.2025).
4. Comprehensive exploration of diffusion models in image generation: a survey. URL: https://link.springer.com/article/10.1007/s10462-025-11110-3 (дата звернення: 10.11.2025).
5. J. Mao and X. Wang. Training-Free Location-Aware Text-to-Image Synthesis. International Conference on Image Processing. 2023, pp. 995-999. URL: https://ieeexplore.ieee.org/document/10222616 (дата звернення: 10.11.2025).
6. Advances in diffusion models for image data augmentation: a review of methods, models, evaluation metrics and future research directions. Artif Intell Rev 58, 112 (2025). DOI: https://doi.org/10.1007/s10462-025-11116-x.
7. Beskorovainyi V. Combined method of ranking options in project decision support systems // Innovative Technologies and Scientific Solutions for Industries. 2020. No 4 (14). P. 13–20. URL: http://journals.uran.ua/itssi/article/view/ITSSI.2020.14.013 (дата звернення:

10.11.2025)

8. Bezkorovainyi V., Kolesnyk L., Gopejenko V., Kosenko V. The method of ranking effective project solutions in conditions of incomplete certainty // Advanced Information Systems, 2024. Vol. 8. No 2. P. 27–38. URL: http://ais.khpi.edu.ua/article/view/305462/297067 (дата звернення: 10.11.2025).