УДК 004.8, 004.92

АНАЛІЗ АРХІТЕКТУР ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ ОПТИМІЗАЦІЇ ПРОЦЕСУ 3D-ВІЗУАЛІЗАЦІЇ

Куций В. В.
vladkytsiy@gmail.com
Черкаський державний фаховий бізнес-коледж
Марченко С. В.
м.Черкаси, Україна

Сучасна індустрія комп’ютерної графіки та створення цифрового контенту (Digital Content Creation) перебуває на етапі фундаментальної технологічної трансформації [1; 2]. Ця зміна характеризується поступовим переходом від класичних, детермінованих алгоритмів розрахунку освітлення до імовірнісних методів генеративного штучного інтелекту. Традиційним промисловим стандартом фотореалістичної візуалізації сьогодні є метод трасування шляхів (Path Tracing), який базується на чисельному розв’язанні рівняння рендерингу [3]. Попри те, що фізична коректність цього підходу гарантує високий реалізм, його обчислювальна складність зростає експоненційно зі збільшенням кількості джерел світла та деталізації геометричної складної сцени. Для отримання зображення без шуму необхідно розрахувати тисячі світлових шляхів для кожного пікселя. Емпіричні дослідження продуктивності сучасних рендер-рушіїв свідчать, що візуалізація одного кадру у роздільній здатності 4K на споживчому обладнанні може тривати від кількох десятків хвилин до кількох годин. Окрім часових затрат, традиційний метод висуває жорсткі вимоги до апаратного забезпечення користувача, зокрема до обсягу відеопам’яті (VRAM). Спроба відрендерити сцену з використанням високоякісних PBR-матеріалів часто призводить до переповнення пам’яті графічного процесора та переходу в режим вивантаження даних в оперативну пам’ять (Out-of-core rendering), що додатково знижує швидкість роботи у десятки разів. Це створює нагальну потребу в розробці та впровадженні нових архітектурних рішень, здатних оптимізувати процес візуалізації.

Проаналізувати еволюцію інженерних підходів та існуючі архітектури генеративного штучного інтелекту в контексті їх застосування для оптимізації процесів 3D-візуалізації, а також обґрунтувати вибір оптимальної технології для подолання поточних апаратних обмежень користувачів.

Поняття нейронного рендерингу охоплює широкий спектр методів, що еволюціонували від простих інструментів пост-обробки до комплексних генеративних систем. Першим успішним етапом впровадження нейромереж у виробничі пайплайни стала технологія інтелектуального знешумлення (AI-Denoising). Алгоритми цього класу використовують згорткові автоенкодери та допоміжні геометричні буфери (карти кольору Albedo та карти нормалей Normal Pass) для відокремлення корисного візуального сигналу від стохастичного шуму [5]. Хоча цей метод дозволяє прискорити візуалізацію у кілька разів, він залишається суто інструментом реконструкції, який не здатний генерувати нові деталі чи оптимізувати роботу з важкими текстурами.

Наступним, значно потужнішим етапом стала поява латентних дифузійних моделей (Latent Diffusion Models, LDM), яскравими представниками яких є сімейство Stable Diffusion та розроблені на їх базі плагіни для 3D-редакторів. Фундаментальна ідея цього підходу полягає у перенесенні ітеративного процесу знешумлення з піксельного простору у стиснутий векторний (латентний) простір за допомогою варіаційного автоенкодера (VAE) [4]. Проте, при спробі інтеграції дифузійних моделей у професійні процеси 3D-візуалізації виявляється низка критичних архітектурних недоліків. Насамперед, такі системи створюють новий апаратний бар’єр: для стабільної генерації зображень потрібні відеокарти з обсягом пам’яті не менше 12–14 ГБ. У разі нестачі ресурсів система використовує повільну оперативну пам’ять, що нівелює будь-який виграш у швидкості. Крім того, традиційні LDM мають суттєві обмеження контексту через використання текстового енкодера CLIP, який обрізає вхідну послідовність після 77 токенів, унеможливлюючи точний опис складної сцени. Найбільш вразливим місцем таких архітектур є явище просторової агнозії (Spatial Agnosia). Оскільки модель навчалася переважно на двовимірних зображеннях, вона не має

глибинного розуміння тривимірних відношень, що призводить до некоректної інтерпретації взаємного розташування об’єктів та проблем із масштабуванням генерації до нативної роздільної здатності 4K.

Сучасним вирішенням цих проблем є перехід до архітектури мультимодальних великих мовних моделей (Multimodal LLM) на базі трансформерів, таких як Google Gemini. Фундаментальна відмінність цього підходу полягає у концепції уніфікованого сприйняття. Замість використання ізольованих потоків для тексту та зображень, мультимодальні трансформери розбивають вхідну візуальну інформацію (наприклад, карту глибини) на сітку фіксованих фрагментів (патчів). Кожен патч проходить через шар лінійної проєкції та отримує позиційне кодування, перетворюючись на послідовність візуальних токенів, які обробляються спільно з текстовими інструкціями механізмом глобальної само-уваги (Self-Attention) [6; 7].

Завдяки такій архітектурі система здобуває емерджентну властивість просторового інтелекту. Модель здатна виконувати складні логічні операції з геометрією, коректно інтерпретуючи градієнти яскравості на карті глибини як метричну інформацію про відстань. Це дозволяє генерувати складні оптичні ефекти, обчислювати явища оклюзії та вибудовувати логічні зв’язки освітлення ще до початку візуалізації (Visual Chain-of-Thought). Більше того, трансформери оперують послідовностями токенів довільної довжини, що знімає жорстку прив’язку до роздільної здатності і дозволяє здійснювати генерацію у нативному форматі 4K. Перенесення цих обчислювальних процесів у хмарне середовище повністю усуває проблему локального «вузького місця» відеопам’яті, демократизуючи доступ до високоякісного рендерингу [8].

Проведений аналіз архітектурних підходів демонструє, що традиційні методи візуалізації та локальні дифузійні моделі вичерпують свій потенціал через надмірні вимоги до апаратного забезпечення та нездатність до глибокого розуміння тривимірного простору. Інтеграція хмарних мультимодальних трансформерів у конвеєр 3D-моделювання є найбільш перспективним напрямком. Використання таких моделей дозволяє розгорнути клієнт-серверну

архітектуру, яка ефективно інтерпретує структурні дані сцени (карти глибини) та забезпечує генерацію фізично коректних зображень високої роздільної здатності, мінімізуючи навантаження на локальні обчислювальні ресурси користувача.

Список використаних джерел:

1. 3D Rendering Market Size, Share, and Trends 2025 to 2034. URL: https://www.precedenceresearch.com/3d-rendering-market (дата звернення: 12.03.2026).
2. 3D Rendering Market (2025–2033). URL: https://www.grandviewresearch.com/industry-analysis/3d-rendering-market-report (дата звернення: 12.03.2026).
3. Kajiya J. T. The rendering equation // *ACM SIGGRAPH Computer Graphics*. 1986. Vol. 20, No. 4. P. 143–150. DOI: https://doi.org/10.1145/15886.15902.
4. Rombach R., Blattmann A., Lorenz D., Esser P., Ommer B. High-resolution image synthesis with latent diffusion models. arXiv. 2022. URL: https://arxiv.org/abs/2112.10752 (дата звернення: 12.03.2026).
5. Chaitanya C. R. A., Kaplanyan A. S., Schied C., Salvi M., Lefohn A., Nowrouzezahrai D., Aila T. Interactive reconstruction of Monte Carlo image sequences using a recurrent denoising autoencoder. ACM Transactions on Graphics. 2017. Vol. 36, No. 4. P. 1–12. DOI: https://doi.org/10.1145/3072959.3073601.
6. Dosovitskiy A. et al. An image is worth 16×16 words: transformers for image recognition at scale. arXiv. 2020. URL: https://arxiv.org/abs/2010.11929 (дата звернення: 12.03.2026).
7. Wu B. et al. Visual transformers: token-based image representation and processing for computer vision. arXiv. URL: https://arxiv.org/abs/2006.03677 (дата звернення: 12.03.2026).
8. Gemini Team. Gemini: a family of highly capable multimodal models. arXiv. 2023. URL: https://arxiv.org/abs/2312.11805 (дата звернення: 12.03.2026).