https://doi.org/10.31891/2219-9365-2025-81-18
УДК 004.89:004.932

ШАПТАЛА Роман
Національний технічний університет України “Київський політехнічний інститут імені Ігоря Сікорського”
https://orcid.org/0000-0002-4367-5775
e-mail: r.shaptala@gmail.com
ЯКОВЕНКО Ярослава
Кременчуцький національний університет імені Михайла Остроградського
https://orcid.org/0000-0001-5042-2701
e-mail: yaroslavayakovenko@gmail.com

АНАЛІЗ МОДЕЛЕЙ ГЕНЕРАЦІЇ ЗОБРАЖЕНЬ З ТЕКСТОВИМИ ЕЛЕМЕНТАМИ

У статті розглядається проблема генерації зображень з інтегрованим текстовим контентом, що є надзвичайно актуальним завданням для сучасних технологій штучного інтелекту. Незважаючи на значні досягнення у генерації зображень за допомогою дифузійних моделей, точне відтворення тексту залишається викликом через складність збереження коректної послідовності символів та розташування текстових елементів. Метою дослідження є оцінка здатності чотирьох сучасних моделей (DALL-E, Flux, RecraftV3 та TextDiffuser-2) генерувати якісний текст при різній довжині вхідного запиту, а також виявлення критичних точок, після яких якість текстових елементів на згенерованих зображеннях значно погіршується.

Для експериментальної частини було сформовано набір текстових запитів, що охоплюють довжину від 1 до 15 слів, з використанням простих слів, коротких фраз та складніших речень. Кожен запит оброблявся десять разів кожною з моделей, що дозволило отримати репрезентативну вибірку результатів. Аналіз отриманих зображень дозволив виділити критичні точки – довжини текстів після яких моделі перестають генерувати коректний текст, а також класифікувати типові помилки на згенерованих зображеннях.

Отримані результати свідчать про суттєві відмінності між моделями: RecraftV3 показала найвищу стабільність, зберігаючи коректність тексту до 14 слів, тоді як DALL-E-3 та Flux-1-Pro демонстрували погіршення якості вже після 5 слів. TextDiffuser-2 відзначилась високою часткою помилок, що обмежує її застосування у завданнях, де точність є критичною. Результати дослідження мають практичну цінність для подальшого вдосконалення алгоритмів генерації зображень, зокрема в контексті рекламних технологій, дизайну та автоматизованого створення візуального контенту.

Ключові слова: моделі генерації зображень, дифузійні моделі, DALLE, Flux, RecraftV3, TextDiffuser-2.

SHAPTALA Roman
National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”
YAKOVENKO Yaroslava
Kremenchuk Mykhailo Ostrohradskyi National University

ANALYSIS OF IMAGE GENERATION MODELS WITH TEXTUAL ELEMENTS

This article addresses the problem of generating images with integrated textual content, which is a highly relevant challenge for modern artificial intelligence technologies. Despite significant advancements in image generation using diffusion models, accurately rendering text remains a challenge due to the complexity of maintaining correct character sequences and textual elements placement. The study aims to evaluate the ability of four modern models (DALL-E, Flux, RecraftV3, and TextDiffuser-2) to generate high-quality text with varying input lengths and identify critical points at which the quality of textual elements on generated images significantly deteriorates.

For the experimental part, a set of text prompts was created, ranging from 1 to 15 words, including simple words, short phrases, and more complex sentences. Each prompt was processed ten times by each model, providing a representative sample of results. The analysis of the generated images allowed for identifying critical points—the text lengths at which the models fail to produce correct text—and classifying typical errors in the generated images.

The results indicate significant differences between the models: RecraftV3 demonstrated the highest stability, maintaining text accuracy up to 14 words, while DALL-E-3 and Flux-1-Pro showed quality degradation after 5 words. TextDiffuser-2 exhibited a high error rate, limiting its use in tasks where accuracy is critical. The study’s findings have practical value for further improving image generation algorithms, particularly in advertising, design, and automated visual content creation.

Keywords: image generation models, diffusion models, DALLE, Flux, RecraftV3, TextDiffuser-2.

1. ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ’ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

1.1 Мотивація дослідження. В епоху штучного інтелекту генеративні моделі відіграють дедалі важливішу роль у створенні візуального контенту. Особливо актуальним стає завдання генерації зображень з інтегрованими текстовими елементами, адже точне відображення тексту критично важливе для низки практичних застосувань – від рекламних кампаній і брендингу до автоматизованого створення інформаційних візуалізацій. Проблема виникає через те, що навіть сучасні алгоритми, здатні генерувати високоякісні зображення, часто стикаються з труднощами у відтворенні текстових деталей, що негативно впливає на їх сприйняття кінцевими користувачами.

З іншого боку, збільшення попиту на автоматизацію креативних процесів і швидке виробництво візуального контенту стимулюють розвиток нових технологій у галузі генерації зображень. В умовах, коли текстові елементи мають важливе значення для передачі інформації, критично важливо зрозуміти, як саме впливають параметри запиту (наприклад, кількість слів) на якість генерованого тексту. Це дослідження спрямоване на визначення та класифікацію існуючих недоліків генеративних систем при генерації зображень з текстовим наповненням, що є ключовим завданням як для науковців, так і практиків у сфері дизайну, реклами та інформаційних технологій.

1.2 Постановка проблеми. Попри значний прогрес у розвитку генеративних моделей, проблема коректного відтворення тексту у зображеннях залишається відкритою. При збільшенні складності текстового запиту спостерігається різке погіршення якості відтворення текстових елементів: моделі можуть допускати друкарські помилки, повторювати окремі слова або генерувати нерозбірливий текст. Це обмежує застосування генеративних моделей у випадках, де точність і зрозумілість текстової інформації на зображенні є критично важливими. Також важливим аспектом задачі вибору генеративної моделі є її надійність, адже у практичних застосуваннях важливо розуміти з якою ймовірністю згенерується проблемне зображення, що дозволяє оцінити ризики впровадження технології у продукт.

Отже, дане дослідження фокусується на пошуці “критичних точок” – тих порогових значень кількості слів у вхідному запиті, після яких якість тексту значно знижується. Виявлення цих точок дозволить не тільки зрозуміти межі можливостей поточних моделей, але й окреслити напрямки для їх подальшого вдосконалення. Адже для низки практичних завдань (наприклад, при створенні рекламних матеріалів або інформаційних панелей) недопустимі випадки з появою суттєвих помилок, що можуть негативно вплинути на сприйняття бренду чи інформації.

1.3 Наукова новизна. У цьому дослідженні ми вперше проводимо систематичний аналіз генеративних моделей DALL-E, Flux, TextDiffuser та RecraftV3 з точки зору відтворення тексту, визначаючи критичні точки, після яких починають виникати текстові помилки. Завдяки використанню експертної оцінки було проведено порівняльний аналіз, який дозволив чітко окреслити межі ефективності кожної з моделей.

Іншим важливим внеском даної роботи є класифікація помилок, що виникають при генерації тексту. Виявлено кілька основних типів неточностей: друкарські помилки, відсутність текстових елементів, повтори слів та нерозбірливий текст. Такий підхід дозволяє не лише порівняти моделі за якістю генерованих зображень, але й глибше зрозуміти механізми їх роботи та слабкі місця. Результати дослідження відкривають нові перспективи для оптимізації алгоритмів генерації тексту в зображеннях, що може сприяти як подальшому розвитку теорії, так і практичному впровадженню вдосконалених рішень у різних галузях.

2. АНАЛІЗ ДОСЛІДЖЕНЬ ТА ПУБЛІКАЦІЙ

Сучасні генеративні моделі, зокрема, варіанти генеративних змагальних мереж (GAN), варіаційних автокодувальників (VAE) та дифузійних моделей, за останні роки зазнали значного розвитку. Завдяки впровадженню алгоритмів на основі дифузійних процесів, таких як DALL-E [1], Stable Diffusion [2], Imagen [3], досягнуто вражаючих результатів у задачах генерації зображення на основі текстового опису. Основна відмінність сучасних підходів полягає в здатності моделі генерувати деталізовані й реалістичні зображення на основі складних описів, що дозволяє знаходити застосування як у креативних індустріях, так і в автоматизації ряду завдань.

Проте, незважаючи на значні досягнення у сфері створення зображень за текстовим описом, відтворення текстових елементів всередині самих зображень залишається проблематичним. Оскільки початкові дослідження зосереджувались переважно на передачі змісту текстового запису через образи отримало значно менше уваги в науковій літературі. Таким чином, попри успіхи у більш широкій задачі генерації зображень на основі текстового опису, більш вузька задача коректного відтворення тексту в зображеннях досі вимагає глибокого аналізу та інноваційних підходів.

За останні роки виникло декілька підходів, спрямованих на вирішення проблеми генерації текстових елементів у зображеннях. Так у дослідженні [4] автори пропонують схему генерації тексту на зображеннях, орієнтовану на багатомовність. Метод Diff-Text використовує попередньо згенеровані ескізи як інформаційні пріори для генеративної моделі, а також локалізований механізм уваги для корекції положення текстових елементів. Запропоновані модифікації, зокрема введення контрастивних підказок на рівні зображення, сприяють досягненню високої точності у розпізнаванні тексту та природній інтеграції тексту в загальний фон зображення. Недоліком такого підходу є те, що текст повинен бути початково розміщений на ескізі (інформаційний пріор), а модель генерує лише візуальні елементи навколо з мінімальними змінами стилю. Враховуючи обмеженість інструментів для генерації самих ескізів, практична цінність підходу значно зменшується.

Іншим підходом спрямованим на вирішення проблеми генерації текстових елементів є [5]. Автори цього дослідження демонструють метод, заснований на інтеграції великих мовних моделей для планування розташування тексту. Завдяки налаштуванню мовної моделі для автоматичної генерації ключових слів і підтримці модифікації розмітки через інтерактивний чат, дане рішення забезпечує більш гнучкий та

автоматизований підхід до рендерингу тексту. Важливою особливістю є використання позиційного кодування на рівні рядків, що дозволяє уникнути надмірно жорсткого керування символами, забезпечуючи тим самим більшу різноманітність стилів відображення тексту.

Одна з найсучасніших моделей Recraft V3 [6] була розроблена на основі вищезазначеного підходу, а також ControlNet архітектури [7]. ControlNet дозволяє додати умовне керування до великих дифузійних моделей. Такий підхід є релевантним для завдання рендерингу тексту, оскільки дозволяє моделі орієнтуватися на просторову розмітку, забезпечуючи тим самим більш структуроване та коректне відображення текстових елементів. Recraft V3 виділяється завдяки здатності генерувати зображення з великими текстовими блоками. Основна ідея полягає у використанні додаткового вхідного сигналу у вигляді малюнка текстового розташування, який забезпечує моделі більш точну інформацію про структуру тексту. Важливою складовою даного підходу є розробка власної системи розпізнавання тексту для формування високоякісних текстових структур, а також вирішення проблем, пов'язаних з правильним порядком розташування слів та оптимізацією формату виводу.

Сучасні публікації демонструють значне зростання інтересу до цієї проблематики. Однак, у наявній літературі бракує систематичного аналізу, який би охоплював як залежність якості тексту від кількості слів у запиті, так і класифікацію типових помилок (друкарські помилки, відсутність тексту, повтори слів, нерозбірливий текст). Саме ці проблеми мотивували наше дослідження, спрямоване на порівняння різних моделей з точки зору їх здатності відтворювати текст в зображеннях, виявлення критичних точок та аналіз типів помилок, що виникають.

3. ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Головною метою цього дослідження є оцінка якості генерації текстових елементів у зображеннях за допомогою сучасних дифузійних моделей, зокрема DALL-E, Flux, TextDiffuser та Recraft V3. Незважаючи на суттєві покращення в області генерації зображень на основі текстового запиту, точність та розбірливість тексту, що генерується моделями, залишаються відкритими питаннями. Наша робота спрямована на виявлення критичних обмежень цих моделей, аналіз їхньої здатності коректно відображати текстові блоки різної довжини та класифікацію характерних помилок.

Зокрема, у межах дослідження передбачається:

- Порівняння моделей за здатністю відтворювати текст у зображеннях із збереженням правильного написання, розташування та читабельності.
- Визначення критичної довжини тексту (кількості слів та знаків), при якій моделі починають допускати помилки.
- Класифікація типових помилок, що виникають у процесі генерації зображень з текстовими елементами.
- Аналіз сильних і слабких сторін кожної моделі та визначення можливих напрямків для покращення їхньої здатності до генерації тексту.

Отримані результати дозволять не лише оцінити поточний рівень розвитку генеративних моделей у контексті текстового рендерингу, але й вказати напрями для їх подальшої оптимізації. Виявлення обмежень існуючих підходів сприятиме розробці нових методів для покращення точності й узгодженості текстових елементів у зображеннях, що є важливим для широкого спектра застосувань – від рекламної індустрії до автоматизованого створення інфографіки та цифрових документів.

4. ВИКЛАД ОСНОВНОГО МАТЕРІАЛУ

4.1 Опис обраних моделей. У даному дослідженні розглянуто чотири сучасні генеративні моделі для створення зображень із текстовими елементами: DALL-E 3, Flux 1 Pro, Recraft-V3 та TextDiffuser 2.

DALL-E-3 – модель від OpenAI, яка використовує дифузійний підхід для генерації високоякісних зображень на основі текстових описів. Це одна з найпопулярніших моделей генерації зображень на основі текстового запиту. Дана модель була обрана як приклад типової дифузійної моделі, яка не розроблялась з метою отримання якісних текстових елементів [1].

Flux-1-Pro – це вдосконалена модель генерації зображень на основі архітектури FLUX-1-Dev. Вона розроблена для забезпечення вищої якості результатів, ефективності та узгодження з текстовим запитом, що робить її популярною для використання як у художніх, так і комерційних програмах [8].

TextDiffuser-2 – модель, яка використовує дифузійні методи в поєднанні з мовними моделями для більш ефективного текстового рендерингу [5].

Recraft-V3 – модель, яка позиціонується як лідер у генерації довгих текстів у зображеннях, застосовуючи спеціальні алгоритми для збереження точності текстового відтворення [6].

4.2 Підготовка експерименту. У дослідженні було сформовано серію тестових текстів англійською мовою, що варіюються за довжиною від 1 до 15 слів. Кожен тестовий текст включався у текстовий запит "Blue sky and yellow field with text: {текст}". Даний запит було обрано так як на практиці користувачі

моделей зацікавлені в певній сцені або зображенні об'єктів разом з текстовими елементами. Далі наводяться тестові тексти використані у роботі:

- "Hello"
- "Dogs bark"
- "Birds fly high"
- "The sun shines brightly"
- "Many flowers bloom in spring"
- "The old car drives very fast"
- "Happy children play in the wonderful park"
- "The friendly neighborhood cat visits us every day"
- "The curious little squirrel gathered acorns for the winter"
- "The enthusiastic crowd cheered loudly when their favorite team scored"
- "The dedicated research team made an important scientific breakthrough this year"
- "The skilled carpenter carefully crafted a beautiful wooden cabinet for the kitchen"
- "The excited students eagerly participated in the science fair project with great enthusiasm"
- "The dedicated teacher worked tirelessly to help her students achieve their academic goals daily"
- "The experienced chef carefully prepared an exquisite meal using fresh local ingredients two days ago"

На основі кожного текстового запиту було згенеровано по 10 зображень кожною з розглянутих моделей. Загалом було отримано 600 зображень для подальшого аналізу. DALL-E 3, Flux 1 Pro та Recraft-V3 були використані через відповідні офіційні прикладні програмні інтерфейси (HTTP API), а TextDiffuser 2 через веб-інтерфейс Huggingface spaces (https://huggingface.co/spaces/JingyeChen22/TextDiffuser-2).

Для оцінювання якості моделей у контексті генерації зображень з текстовим наповненням було розмічено згенеровані зображення за наступними критеріями:

- наявність текстового елементу (так/ні)
- наявність помилок (так/ні)
- наявність друкарської помилки (так/ні)
- наявність повторів слів (так/ні)
- текст розбірливий (так/ні)

На основі розмічених даних було обчислено точність моделей при генерації зображень з текстовими елементами різної довжини за формулою:

$$Acc(m,k) = \frac{\sum_{i=1}^{n} 1_{\{E(m,k,i)=0\}}}{n}, \quad (1)$$

де m – модель, k – довжина текстового елементу, n – кількість зображень згенерованих моделлю m за запитом з довжиною текстового елементу k, $E(m,k,i)$ – наявність помилок на згенерованому зображенні i.

Наступним етапом дослідження був пошук порогу помилки – критичної довжини тексту, при якій відсоток коректного розпізнавання різко знижується. Після цього було здійснено аналіз помилок та їх класифікація.

4.3 Результати експерименту. Точність моделей при генерації зображень з текстовими елементами різної довжини зображена на рис. 1 та рис. 2. Як бачимо, критичною точкою для DALL-E 3, Flux 1 Pro та TextDiffuser 2 є 5 слів. Незважаючи на те, що кожна з цих моделей може допускати помилки з текстами меншої довжини, саме після порогу в 5 слів починається різке зростання кількості помилок. Для DALL-E 3 характерне повторення слів або повна втрата тексту після досягнення критичної довжини. Flux 1 Pro демонструє тенденцію до спотворення тексту або створення нерозбірливих символів, що особливо помітно після 9 слів. TextDiffuser-2, своєю чергою, демонструє найгірші показники серед усіх моделей, з високою часткою помилок навіть для коротких текстів, що обмежує її практичне застосування.

Модель Recraft-V3 показала найвищу стабільність: вона не допускає помилок до порогу в 14 слів і навіть після цього зберігає відносно високу точність. Цей результат свідчить про значно кращу оптимізацію алгоритму обробки текстових елементів порівняно з іншими моделями. Така ефективність може бути обумовлена як архітектурними особливостями, так і спеціалізованими підходами до роботи з текстовими компонентами.

ISSN 2219-9365

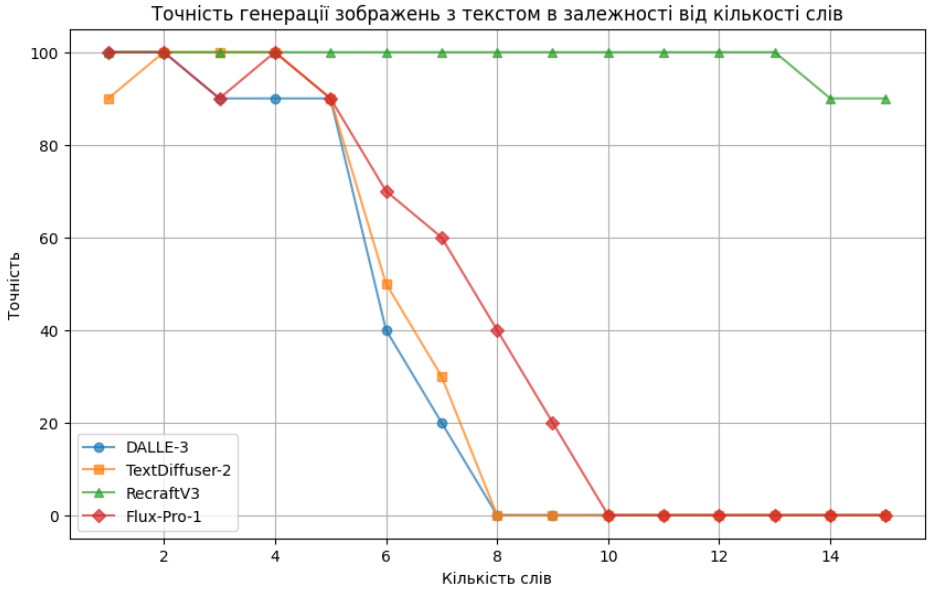


Рис. 1. Точність моделей при генерації зображень з текстовими елементами в залежності від кількості слів

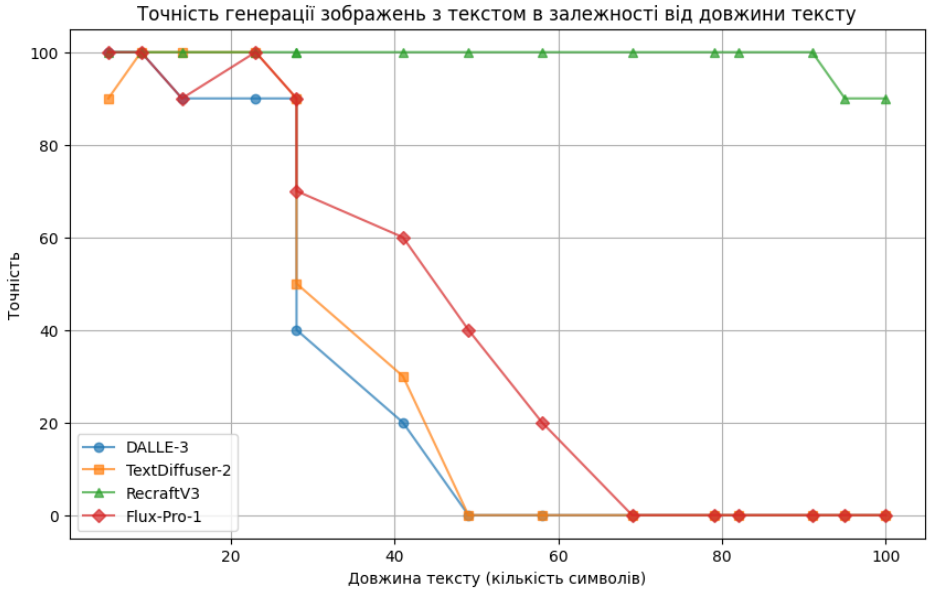


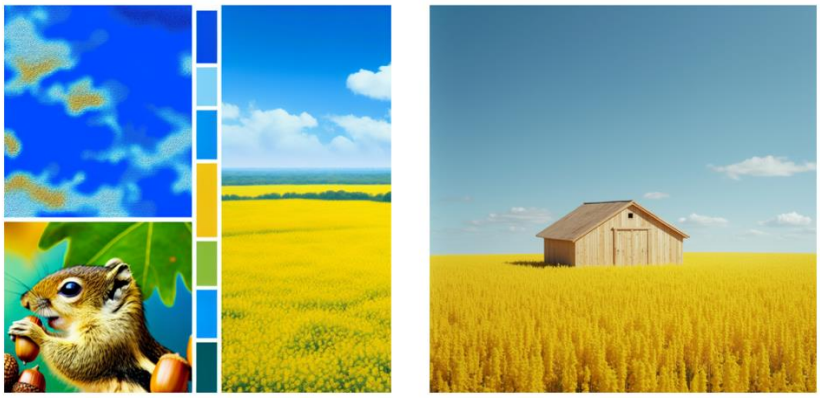
Рис. 2. Точність моделей при генерації зображень з текстовими елементами в залежності від кількості символів

Аналіз помилок при генерації дозволив виділити наступні класи помилок при генерації зображень з текстовою інформацією:

- Друкарська помилка – зайві або неочікувані символи в тексті. Найбільше зустрічаються у DALL-E, Flux та TextDiffuser-2. Приклади зображено на рис. 3.
- Відсутність тексту – згенероване зображення не включає жодних текстових елементів. Часто спостерігається у Flux після 12 слів, а також у DALL-E після 9 слів. Приклади зображено на рис. 4.
- Повтор слів – певні слова у тексті повторюються кілька разів. Часто зустрічаються разом з друкарськими помилками та на межах ліній у багаторядкових текстах. Зафіксовано у DALL-E та TextDiffuser-2. Приклади зображено на рис. 5.
- Нерозбірливий текст – шрифт, стиль написання, або кількість помилок роблять текст неможливим для прочитання. Особливо характерно для Flux моделі. Приклади зображено на рис. 6.

DALLE-3 TextDiffuser-2 RecraftV3 Flux-Pro-1

Рис. 3. Приклади помилок класу “друкарська помилка”



DALLE-3 Flux-Pro-1

Рис. 4. Приклади помилок класу “відсутність тексту”

DALLE-3 TextDiffuser-2

Рис. 5. Приклади помилок класу “повтор слів”

Flux-Pro-1 Flux-Pro-1

Рис. 6. Приклади помилок класу “нерозбірливий текст”

З проведених експериментів видно, що якість генерації текстових зображень значно варіюється залежно від моделі. Recraft-V3 виявилась найбільш стабільною, тоді як TextDiffuser-2 має суттєві обмеження у відтворенні тексту. DALL-E та Flux показали змішані результати, але загалом демонструють зниження якості при збільшенні довжини тексту.

Відмінності у якості порівнюваних моделей можна пояснити наступними факторами: архітектурні особливості та спеціалізація, різниця у навчальних даних та задачах тренування, використання методів умовного керування та інтеграції текстових компонентів, та обмеження дифузійних процесів у відтворенні тексту.

У контексті архітектурних особливостей та спеціалізації RecraftV3 демонструє найкращу точність завдяки спеціалізованим архітектурним компонентам, які орієнтовані на генерацію довгих текстових блоків. Послідовний механізм з генерації текстового шаблону та генерації зображення на основі шаблону забезпечує збереження послідовності символів та коректного розташування тексту, що дозволяє моделі працювати стабільно до 14 слів. DALL-E і Flux застосовують більш універсальні підходи до генерації зображень, де текст виступає як один із елементів композиції. Дані моделі оптимізовані для створення універсальних зображень, а не для точного відтворення високо деталізованого тексту. TextDiffuser-2 має суттєві проблеми з відтворенням тексту навіть для коротких запитів, що може свідчити про відсутність спеціалізованих механізмів для роботи з текстовим контентом у його архітектурі, а також гіршу якість генерації універсальних зображень, ніж DALL-E і Flux.

Різниця у навчальних даних та задачах тренування полягає у тому, що RecraftV3 була навчена на спеціалізованих наборах даних, де велика увага приділялася точності рендерингу тексту, зокрема шляхом використання високоякісних текстових шаблонів, отриманих за допомогою власної моделі розпізнавання тексту. Це дозволило RecraftV3 «вивчити» точне відтворення послідовностей символів та уникати типових помилок. Інші моделі, такі як DALL-E та Flux, були оптимізовані для більш універсальних завдань генерації зображень на основі запиту, де текстові елементи не є основним акцентом. В результаті, вони демонструють гіршу якість відтворення тексту при збільшенні його довжини. TextDiffuser-2 була обмежена в навчальних даних, які містять текстову інформацію, або не містила достатньої кількості прикладів для генерації точного тексту, що призводить до низької точності вже на низьких значеннях довжини тексту.

RecraftV3 використовує методи умовного керування, які дозволяють точно інтегрувати текстовий компонент у загальну композицію зображення. Це включає використання додаткових вхідних сигналів – текстових, що знижує ймовірність появи помилок. У випадку з DALL-E та Flux, текст генерується як додатковий елемент, якому доводиться конкурувати з іншими елементами зображення за обчислювальними ресурсами та місцем на зображенні. Це призводить до того, що при збільшенні довжини тексту точність його відтворення знижується, оскільки модель не отримує достатньої інформації для підтримки високої деталізації. Попри те, що TextDiffuser-2 має механізм умовного керування для текстового компонента, генерація початкового текстового шаблону має не високу якість, що пояснює появу помилок.

Дифузійні моделі, за своєю природою, орієнтовані на генерацію універсальних візуальних елементів і можуть мати труднощі з високо деталізованими структурами, такими як текст. Генерація тексту вимагає високої точності розташування символів та збереження їх послідовності, що є складним завданням для моделей, оптимізованих під універсальну генерацію зображень. Моделі, які не мають додаткових механізмів для контролю генерації текстових елементів, стикаються з труднощами при відтворенні тексту, що проявляється у формі друкарських помилок, відсутності текстових елементів чи повторів слів.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

У даній статті було проведено комплексне дослідження можливостей сучасних генеративних моделей (DALL-E, Flux, RecraftV3 та TextDiffuser-2) у задачі генерації зображень із інтегрованими текстовими елементами. Було сформовано набір текстових запитів з різною довжиною (від 1 до 15 слів) та проведено генерацію зображень кожною моделлю з повторенням експерименту 10 разів для кожного запиту.

Основний аналіз результатів засвідчує, що моделі показують суттєві відмінності у здатності відтворювати текст. Зокрема, модель RecraftV3 продемонструвала найвищу стабільність, зберігаючи якість тексту до 14 слів, що свідчить про її спеціалізовану оптимізацію для генерації довгих текстових блоків. У свою чергу, DALL-E та Flux показали зниження якості при збільшенні кількості слів – критична точка для обох моделей настає вже після 5 слів. Найбільш проблемною виявилася модель TextDiffuser-2, яка допускала помилки навіть при мінімальній довжині тексту, що свідчить про обмеження її архітектури та недостатню кількість навчальних даних для роботи з текстовим контентом.

Аналіз отриманих результатів дозволив класифікувати типові помилки генерації: друкарські помилки, відсутність тексту, повторення слів та нерозбірливість відображення. Висунуто кілька гіпотез, що пояснюють отримані результати, зокрема: особливості архітектурних рішень, різниця у підходах до навчання, використання або відсутність спеціалізованих механізмів умовного керування текстовим компонентом та специфіка дифузійного процесу у відтворенні високо деталізованої текстової інформації.

Перспективи подальших досліджень включають розширення набору моделей для порівняння, вдосконалення методів умовного керування текстом, інтеграцію спеціалізованих компонентів для роботи з текстовими даними, а також збільшення обсягу та якості навчальних даних, що стосуються текстових елементів. Подальший аналіз сприятиме розробці нових алгоритмів, які дозволять досягти більш високої точності та стабільності у генерації зображень з текстовими елементами, що є надзвичайно актуальним завданням для застосувань у рекламі, дизайні та автоматизованому створенні контенту.

Література

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.
2. Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
3. Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." Advances in neural information processing systems 35 (2022): 36479-36494.
4. Zhang, Lingjun, et al. "Brush your text: Synthesize any scene text on images via diffusion model." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 7. 2024.
5. Chen, Jingye, et al. "Textdiffuser-2: Unleashing the power of language models for text rendering." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
6. “How To Create SOTA Image Generation with Text: Recraft’s ML Team Insights.” RecraftAI Blog, 7 Nov. 2024, www.recraft.ai/blog/how-to-create-sota-image-generation-with-text-recrafts-ml-team-insights.
7. Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
8. Black Forest Labs. FLUX. 2024, https://github.com/black-forest-labs/flux.

References

1. Ramesh, Aditya, et al. "Hierarchical text-conditional image generation with clip latents." arXiv preprint arXiv:2204.06125 1.2 (2022): 3.
2. Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.
3. Saharia, Chitwan, et al. "Photorealistic text-to-image diffusion models with deep language understanding." Advances in neural information processing systems 35 (2022): 36479-36494.
4. Zhang, Lingjun, et al. "Brush your text: Synthesize any scene text on images via diffusion model." Proceedings of the AAAI Conference on Artificial Intelligence. Vol. 38. No. 7. 2024.
5. Chen, Jingye, et al. "Textdiffuser-2: Unleashing the power of language models for text rendering." European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024.
6. “How To Create SOTA Image Generation with Text: Recraft’s ML Team Insights.” RecraftAI Blog, 7 Nov. 2024, www.recraft.ai/blog/how-to-create-sota-image-generation-with-text-recrafts-ml-team-insights.
7. Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
8. Black Forest Labs. FLUX. 2024, https://github.com/black-forest-labs/flux.