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 перший (бакалаврський)

Автоматизоване генерування ілюстрацій на основі сценографів, отриманих з тексту
(тема)

Виконав:
здобувач четвертого року навчання,
групи ІТШІ-21-4
Єва Величко
(власне ім'я, прізвище)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо–професійна
Освітня програма Штучний інтелект
(повна назва освітньої програми)
Керівник проф. Нонна Кулішова
(посада, власне ім'я, прізвище)

Допускається до захисту

Завідувач кафедри ШІ _____ Олег ЗОЛОТУХІН
(підпис) (власне ім'я, прізвище)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти перший (бакалаврський)

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

Тип програми освітньо–професійна

Освітня програма Штучний інтелект
(повна назва)

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

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Величко Єві Олексіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Автоматизоване генерування ілюстрацій на основі сценографів, отриманих з тексту

затверджена наказом університету від 19 травня 20 25 р. № 378Ст
2. Термін подання студентом роботи до екзаменаційної комісії 18 червня 20 25 р.
3. Вихідні дані до роботи Науково технічні публікації, дані Інтернет-джерел та наукових проєктів щодо розробки та дослідження глибоких нейронних мереж, набори даних для задач генерації сценографів та зображень, документація Python та бібліотек

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Теоретичні дослідження
3) Практичні дослідження
4) Аналіз результатів дослідження

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Строк / терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	07.04.2025	виконано
2	Аналіз предметної галузі та постановка задачі	08.04.2025–12.04.2025	виконано
3	Дослідження попередніх робіт	13.04.2025–16.04.2025	виконано
4	Аналіз методів вирішення поставленої задачі	17.04.2025–21.04.2025	виконано
5	Аналіз та обробка наборів даних	22.04.2025–25.04.2025	виконано
6	Розробка базової архітектури	26.04.2025–29.04.2025	виконано
7	Реалізація генеративної системи.	30.04.2025–14.05.2025	виконано
8	Написання пояснювальної записки	15.05.2025–25.05.2025	виконано
9	Попередній захист	26.05.2025–17.06.2025	виконано
10	Захист перед ЕК	18.06.2025	виконано

Дата видачі завдання 19 травня 2025 р.

Здобувач _____
(підпис)

Керівник роботи _____ проф. Нонна Кулішова
(підпис) (посада, власне ім’я, прізвище)

РЕФЕРАТ

Пояснювальна записка: 76 с., 19 рис., 3 табл., 1 дод., 35 джерел.

ГЕНЕРАЦІЯ ЗОБРАЖЕНЬ, ГРАФ СЦЕНИ, КОМП'ЮТЕРНИЙ ЗІР, МОДЕЛЬ ТЕКСТ-ЗОБРАЖЕННЯ, ПРОМІЖНА СТРУКТУРА ДАНИХ, СЕМАНТИЧНЕ ПРЕДСТАВЛЕННЯ, ШТУЧНИЙ ІНТЕЛЕКТ.

Об’єкт дослідження – процес автоматизованої генерації зображень на основі текстових описів.

Предмет дослідження – методи та алгоритми побудови сценографів з тексту і перетворення їх у зображення за допомогою нейронних мереж.

Мета роботи – розробити та дослідити технологію автоматизованого генерування ілюстрацій на основі сценографів, отриманих із текстових описів, з урахуванням семантичної та просторової відповідності.

Методи дослідження – аналіз наукових джерел у сфері генерації зображень і обробки природної мови, систематизація підходів генерації зображень, експериментальний підбір конфігурацій, візуалізація структур даних, емпіричне тестування результатів.

У цій кваліфікаційній роботі досліджується метод автоматизованої генерації ілюстрацій на основі сценографів, побудованих із текстових описів. Розглянуто технологію, яка поєднує трансформерні моделі для виділення об’єктів та їхніх відношень з генеративними нейронними мережами для синтезу відповідних зображень. Реалізовано архітектуру системи, що складається з модуля побудови семантичного графа, попередньої обробки сценографа та генератора зображень. Проведено експерименти на відкритих наборах даних, результати яких підтвердили ефективність розглянутої технології у структурній відповідності синтезованих ілюстрацій.

ABSTRACT

Bachelor’s thesis contains: 76 pp., 19 fig., 3 tabl., 1 ann., 35 references.

ARTIFICIAL INTELLIGENCE, COMPUTER VISION, IMAGE GENERATION, INTERMEDIATE DATA STRUCTURE, SCENE GRAPH, SEMANTIC REPRESENTATION, TEXT-TO-IMAGE MODEL.

Object of study – the process of automated image generation based on textual descriptions.

Subject of study – methods and algorithms for building scene graphs from text and converting them into images using neural networks.

Purpose of the work – development and investigation of a technology of automated illustration generation based on scene graphs derived from textual descriptions, taking into account semantic and spatial correspondence.

Methods of research – analysis of scientific sources in the field of image generation and natural language processing, systematisation of image generation approaches, experimental selection of configurations, visualisation of data structures, empirical testing of results.

This qualification thesis investigates a method for automated generation of illustrations based on scene graphs constructed from textual descriptions. A technology is proposed that combines transformer models to extract objects and their relationships with generative neural networks to synthesise the corresponding images. The architecture of the system, consisting of a semantic graph construction module, a scene graph preprocessing module and an image generator, is implemented. Experiments on open datasets were conducted, the results of which confirmed the effectiveness of the proposed technology in the structural correspondence of the synthesised illustrations.

ЗМІСТ

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ВАК – варіаційний автоенкодувальник;

ОПМ – обробка природної мови;

РНМ – рекурентна нейронна мережа;

ШІ – штучний інтелект;

AMR – Abstract Meaning Representation – абстрактне представлення значення;

BLEU – Bilingual Evaluation Understudy – двомовне оцінювання під час навчання;

CIDEr – Consensus–based Image Description Evaluation – оцінка опису зображень на основі консенсусу;

FID – Frechet Inception Distance – початкова відстань Фреше;

GAN – Generative Adversarial Network – генеративна змагальна мережа;

GRAF – Generative Radiance Fields for Text-to-Image Synthesis – генеративні поля випромінювання для синтезу текст-зображення;

InterACT – Interactive Text-to-Image Composition – інтерактивна композиція текст-зображення;

IS – Inception Score – початкова оцінка;

LS-GAN – Least Squares GAN – метод найменших квадратів GAN;

METEOR – Metric for Evaluation of Translation with Explicit Ordering – метрика для оцінювання перекладу з явним упорядкуванням;

MS-COCO – Microsoft Common Objects in Context – загальні об'єкти Microsoft у контексті;

SG2IM – Scene Graph to Image Model – модель перетворення графа сцени в зображення;

T5 – Text-to-Text Transfer Transformer – трансформер перетворення тексту в текст.

ВСТУП

Один із перспективних напрямів сучасного штучного інтелекту – автоматизоване генерування зображень на основі тексту – стрімко розвивається впродовж останніх років. Ця задача передбачає інтеграцію методів обробки природної мови, комп'ютерного зору та генеративного моделювання, відкриваючи широкі можливості для застосування в освітніх технологіях, дизайні, літературі та індустрії розваг.

Важливою проблемою в межах цієї галузі є забезпечення семантичної відповідності між вхідним текстовим описом та згенерованим зображенням. Незважаючи на успіхи у створенні візуально достовірних зображень, більшість сучасних наскрізних систем демонструють обмеження у відтворенні логічних і просторових взаємозв'язків між об'єктами, особливо у випадках з багатокомпонентними сценами. Зокрема, моделі часто помиляються у розміщенні об'єктів, збереженні їхніх атрибутів і взаємодій, що знижує точність фінального результату.

Ці обмеження пов'язані передусім із відсутністю проміжного представлення сцени: у підходах прямого текст-піксельного перетворення система працює як «чорна скринька» без пояснення логіки, за якою інтерпретовано структуру опису. У цьому контексті виникає потреба у структурованому проміжному рівні – графі сцен, який забезпечує формалізоване представлення об'єктів, їхніх атрибутів і взаємозв'язків. Такий підхід дозволяє підвищити керованість та пояснюваність системи, забезпечити точне позиціонування об'єктів та мінімізувати ризик втрати ключових елементів тексту в процесі синтезу зображення. Водночас візуалізація сценографа дає змогу здійснювати верифікацію на проміжному етапі генерації, що критично для застосувань з підвищеними вимогами до точності.

Зважаючи на переваги використання сценографів, метою даної кваліфікаційної роботи є розробка та дослідження технології

автоматизованого генерування ілюстрацій з тексту на їх основі. Основним завданням є досягнення високої семантичної відповідності між текстом і зображенням, тобто точного позиціонування об’єктів відповідно до логіки сцени. Передбачається також, що розроблена система буде гнучкою до розширення функціональності, зокрема додання можливості редагування сценографів перед фінальним синтезом зображень.

Практична значущість роботи полягає в можливості використання розглянутої технології у створенні осмисленого ілюстрованого контенту для креативних і професійних застосувань. Таким чином, дослідження спрямоване на розв’язання актуального наукового завдання, забезпечуючи як теоретичну новизну, так і практичну цінність роботи.

1 РОГЛЯД ЛІТЕРАТУРИ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Ілюстрації в контексті комп’ютерного зору

Розвиток комп’ютерного зору як однієї з ключових підгалузей штучного інтелекту (ШІ) відкрив нові можливості не лише для аналізу зображень, а й для їх автоматичного створення. Візуалізація результатів обчислень, генерація сцен, ілюстрування тексту – усі ці задачі вимагають здатності моделі до формування осмислених зображень на основі вхідних даних. У цьому контексті ілюстрації постають як синтетичні зображення, що створюються алгоритмічно для передання семантичної, структурної або просторової інформації.

Ілюстрації в системах комп’ютерного зору виконують низку важливих функцій: від покращення інтерпретації даних до створення контенту у творчих та прикладних галузях. Особливої актуальності ця тема набуває у завданнях генерації зображень з тексту, де ілюстрація є не просто графічним відображенням, а результатом глибокого розуміння змісту сцени та її логічної структури.

Ілюстрації відіграють ключову роль у сучасних системах штучного інтелекту, особливо в задачах, де необхідне узгодження між текстовим описом та візуальним представленням інформації. На відміну від класичних зображень, які переважно фіксують вже наявні об’єкти або сцени, ілюстрації в ШІ формуються динамічно – як результат обробки текстових даних, логічних структур, графів або семантичних зв’язків.

У межах автоматизованої генерації зображень ілюстрація стає мостом між мовою і візуальним сприйняттям, забезпечуючи трансформацію абстрактної інформації у зрозумілу форму. Це критично важливо в таких сферах, як освітній контент, автоматизоване створення коміксів або візуалізація описових сцен у відеоіграх та літературі. Саме завдяки

ілюстраціям системи ШІ можуть не лише обробляти інформацію, а й презентувати її у формі, зручній для людини.

Крім того, ілюстрації сприяють покращенню контролю над генерацією – вони дозволяють формалізувати структуру сцени, встановити просторові відносини між об’єктами, задати атрибути та зв’язки, що в сукупності підвищує точність і передбачуваність результату. Відтак, ілюстрації не лише візуалізують дані, але й виступають інструментом семантичної інтерпретації, формуючи основу для пояснюваних, адаптивних та інтуїтивних інтерфейсів між людиною і машиною.

У цьому контексті постає необхідність у такому представленні, яке не лише зберігає ключові об’єкти сцени, а й чітко відображає семантичні відносини між ними. Саме цю роль виконують графи сцен – структуровані графові моделі, у яких вузли відповідають об’єктам та їхнім атрибутам, а ребра – просторовим, функціональним чи логічним зв’язкам між ними [1].

Використання графів сцен дозволяє перейти від поверхневого візуального опису до глибинного семантичного моделювання, забезпечуючи машині можливість не лише «бачити», а й розуміти структуру зображення [2]. Завдяки цьому підходу можна формалізувати сцени, що генеруються, зробити процес побудови ілюстрацій більш контрольованим, відтворюваним та адаптивним до змісту тексту.

Таким чином, графи сцен виступають ключовим посередником між текстовими описами та візуальним результатом, що забезпечує більш точну відповідність між бажаною смисловою структурою і згенерованим зображенням.

1.2 Графи сцен як структуроване представлення даних у задачі генерації зображення з тексту

У сучасних системах генерації зображень на основі тексту важливою вимогою є не лише створення візуально достовірного результату, але й

забезпечення семантичної відповідності між вхідним описом та згенерованою сценою [3]. Традиційні дифузійні або трансформерні моделі, що здійснюють безпосереднє перетворення тексту у зображення, часто не гарантують точного відтворення логіки опису або просторових відносин між об’єктами. У зв’язку з цим усе більшої актуальності набуває використання сценографів як проміжного, структурованого рівня представлення даних.

Сценограф або граф сцени – це орієнтований граф (рисунок 1.1), у якому вузли відповідають об’єктам сцени (наприклад, «стіл», «собака»), їхнім атрибутам («дерев’яний», «білий») або діям, а дуги – семантичним чи просторовим зв’язкам між ними («на», «біля», «тримати») [1]. Така структура дозволяє формалізувати зміст сцени до етапу візуалізації, забезпечуючи чітке відображення логічних зв’язків, що істотно знижує ймовірність появи помилок або композиційно некоректних об’єктів.

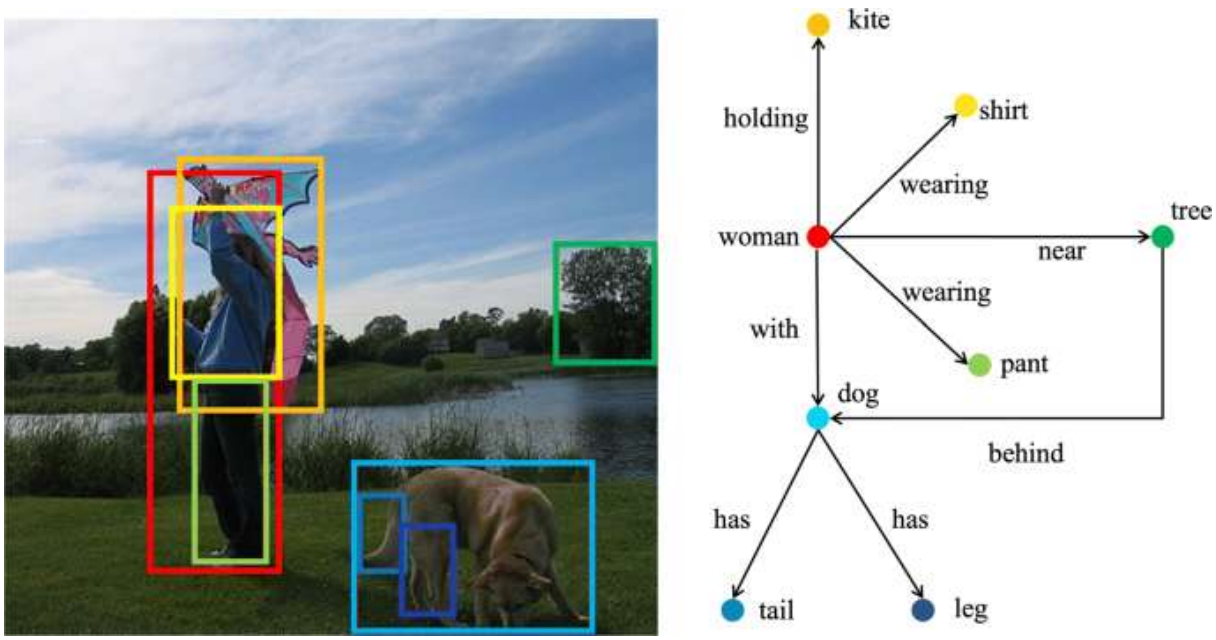

Рисунок 1.1 – Приклад графу сцени

Реалізація підходу із використанням сценографів зазвичай включає два етапи:

– парсинг тексту з побудовою сценографа за допомогою лінгвістичних чи трансформерних моделей (наприклад, парсинг на основі залежностей [4] або BERT [5], [6], BLIP);

– генерація зображення на основі сценографа за допомогою спеціалізованої моделі, що підтримує структурований вхід (наприклад, SG2IM (Scene Graph to Image Model), GRAF (Generative Radiance Fields for Text-to-Image Synthesis)).

З практичної точки зору цей підхід має низку важливих переваг:

– керованість та модульність: структура графа дає змогу змінювати об’єкти, відношення чи атрибути сцени без потреби в повторному опрацюванні всього тексту [7], що забезпечує гнучкість системи;

– семантична точність: завдяки структурованому підходу вдається досягти точнішої відповідності між змістом опису і візуальним результатом, з урахуванням просторових і логічних зв’язків;

– інтерпретованість: сценографи є прозорими з точки зору логіки побудови, що дозволяє аналізувати або вручну коригувати зміст сцени до її генерації;

– повторне використання: один і той самий граф може бути візуалізований у різних стилях, що розширює можливості повторного застосування та адаптації системи.

Водночас існують певні обмеження, а саме побудова сценографа з вільного тексту є складним завданням через неоднозначність природної мови [1], особливо у випадках, коли йдеться про стилістичні, емоційні або метафоричні описи. Крім того, зі збільшенням кількості об’єктів граф може ускладнюватися, що призводить до зростання обчислювальних витрат і складності оптимізації.

В межах даної роботи застосування сценографів розглядається як доцільне та ефективне рішення для структурованого переходу від тексту до зображення. Вони забезпечують підвищену точність, керованість і

прозорість генерації, що є критично важливими для побудови систем автоматизованого ілюстрування описових сцен, особливо у контексті навчальних, літературних та інтерактивних застосунків.

1.3 Основні напрямки в класифікації підходів до генерації зображень на основі тексту

У сучасних системах генерації зображень на основі текстових описів ключову роль відіграють мультимодальні архітектури, здатні об’єднувати мовну і візуальну інформацію в єдиний генеративний процес. Основна мета таких систем – створити зображення, яке точно відображає зміст вхідного тексту не лише на рівні об’єктів, а й з урахуванням просторових, логічних та семантичних зв’язків між ними [8].

Просте пряме перетворення тексту у піксельну форму, без додаткової інтерпретації структури сцени, є надзвичайно складним, особливо у випадках складних описів, типових для датасетів, як-от Microsoft Common Objects in Context (MS-COCO) (рисунок 1.2) [9].

Рисунок 1.2 – Приклад опису зображення з датасету MS-COCO

У подібних ситуаціях моделі часто демонструють труднощі у правильному розміщенні об’єктів, збереженні їх атрибутів та взаємозв’язків. У зв’язку з цим виникає необхідність у впровадженні проміжного представлення, яке б забезпечувало збереження смислової структури опису перед етапом безпосередньої візуалізації. Такими представленнями можуть бути семантичні макети [10], сцени [11] або графи [12], [13], які покращують інтерпретованість, контрольованість і відповідність результату початковому тексту.

1.3.1 Пряме текст-піксельне перетворення

Один із базових підходів до генерації зображення з тексту ґрунтується на безпосередньому перетворенні мовного опису у піксельну форму без побудови проміжного семантичного представлення сцени [8], [14]. Такі методи умовно називають наскрізними (англ. end-to-end) моделями прямої генерації. Їх архітектура зазвичай поєднує мовну енкодерну частину (часто на основі РНМ або трансформерів) з глибокою нейронною мережею-декодером, що відповідає за побудову зображення у вигляді піксельного масиву.

Класичними прикладами такого підходу є StackGAN [15] та AttnGAN [16], які започаткували використання генеративних змагальних мереж (англ. generative adversarial networks, GANs) [17] у задачах перетворення тексту в зображення. У StackGAN текстовий опис перетворюється у векторне представлення, яке потім передається у два послідовних генеративних блоки: перший формує зображення низької роздільності (64×64), другий уточнює його до більш високої якості (256×256). AttnGAN доповнює цей підхід механізмом уваги [18], що дозволяє моделі на кожному етапі генерації фокусуватись на відповідних частинах опису (рисунок 1.3).

Рисунок 1.3 – Архітектура AttnGAN

Попри значний прорив у якості зображень, моделі прямого перетворення виявляють систематичні обмеження у випадках складних, багатокомпонентних або структурно насичених сцен, як-от ті, що зустрічаються в MS-COCO. Основна проблема полягає у слабкому узгодженні між просторовими відношеннями в тексті та розміщенням об’єктів на зображенні. Наприклад, опис типу «собака лежить під столом, а поруч стоїть миска» часто призводить до помилкової генерації, де об’єкти розташовані випадково або зовсім не відображені. Це свідчить про недостатню здатність моделі підтримувати просторову когерентність.

Крім того, відсутність явного представлення взаємозв’язків між об’єктами призводить до семантичної невизначеності: хоча зображення може бути візуально привабливим, воно не гарантує відповідності змісту тексту. У деяких випадках моделі «галюцинують» об’єкти, яких не було в описі, або ігнорують ключові сутності [13], що є критичним недоліком у застосуваннях, де точність має вирішальне значення (освітній контент, медичні ілюстрації, юридична візуалізація).

Ще одним обмеженням є низька керованість результатом. Оскільки модель діє як «чорна скринька», користувач не має змоги впливати на структуру або компоненти сцени без повторного формулювання тексту, що унеможливлює інтерактивну генерацію або редагування.

Незважаючи на зазначені недоліки, прямі текст-піксельні моделі заклали фундамент для подальших розробок у сфері синтезу зображень з тексту та продемонстрували потенціал генеративних архітектур. Вони залишаються актуальними в задачах, що не потребують структурної деталізації сцени, а також як базові моделі для попереднього тренування або генерації високоякісного фону. Проте у випадках, де важлива структура сцени, точність семантики і прозорість генерації, постає потреба в більш формалізованих проміжних представленнях, що забезпечують логічну відповідність між мовним описом та візуальним результатом.

1.3.2 Проміжне семантичне представлення сцени

Однією з найбільш перспективних стратегій у задачі генерації зображення на основі тексту є використання проміжного семантичного представлення сцени перед безпосередньою візуалізацією. На відміну від підходів прямого текст-піксельного перетворення, які намагаються синтезувати зображення безпосередньо з тексту, структуровані методи передбачають побудову логічної або семантичної моделі сцени – так званого макету або сценографа [13]. Це дозволяє моделі поетапно формувати зображення, зберігаючи при цьому зв’язність і смислову відповідність між описом та результатом.

Проміжне представлення зазвичай має форму розмітки об’єктів, карти форм об’єктів або графу відношень між об’єктами, також зазначеного як графу сцени. Такий підхід дозволяє розділити загальну задачу генерації на послідовні підзадачі:

– генерація макету – локалізація об’єктів за допомогою координат або обмежувальних рамок на основі текстового опису ;

– генерація форм/масок – уточнення форми, розміру та контурів об’єктів, визначених у макеті;

– синтез зображення – побудова піксельного зображення на основі семантичного плану (рисунок 1.4).

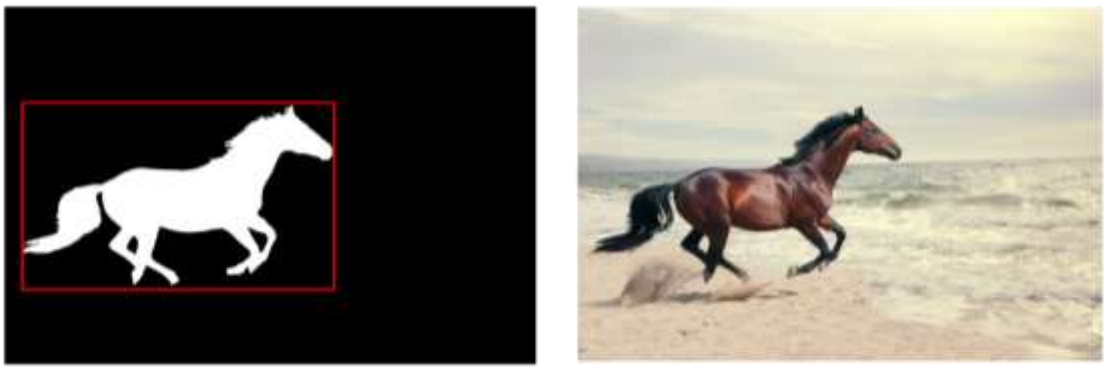

Рисунок 1.4 – Синтезоване на основі макету (зліва) зображення (справа)

Цей підхід забезпечує вищий ступінь контрольованості над результатом і дозволяє системі більш точно відтворювати не лише наявність об’єктів, а й їхні просторові відносини та взаємодії.

У дослідженні B. Zhao et al. (2018) [11] представлено систему Layout2Image, яка генерує зображення на основі попередньо заданого макету сцени. На першому етапі модель створює позиції об’єктів (у вигляді прямокутників), а далі генерує зображення відповідно до цієї структури. У системі SG2IM автори пропонують використовувати сценограф як вхід до моделі генерації. Граф сцени, який включає вузли (об’єкти) та ребра (відношення), перетворюється у відповідне зображення через модулі згортки просторового графа та інференсу розташування об'єктів.

Інший приклад – GRAF, який дозволяє опрацьовувати структуру сцени з урахуванням глибинної інформації і стилю зображення. У роботі Ashual & Wolf (2019) [19] представлено систему InterACT (Interactive Text-

to-Image Composition), яка надає можливість інтерактивно змінювати компоненти сцени (об’єкти, атрибути, положення) з використанням явного представлення кожного елемента.

Основними перевагами проміжного семантичного представлення є:

– контрольованість – користувач або система може змінювати розмітку об’єктів без повної регенерації тексту;

– семантична точність – логічна структура опису зберігається та реалізується в зображенні;

– пояснюваність – кожен етап генерації є інтерпретованим і доступним для перевірки.

Недоліком такого підходу є збільшення складності системи та необхідність створення або анотації проміжних представлень, що потребує додаткових ресурсів або окремого навчання.

У контексті даної роботи, використання проміжного представлення є обґрунтованим рішенням для забезпечення логічної цілісності, інтерпретованості та керованості процесу генерації ілюстрацій, особливо в застосуваннях, що потребують точного відтворення смислу тексту.

Варто також зазначити, що використання проміжних семантичних представлень створює передумови для багатомодульної архітектури, яка дозволяє ізолювати помилки на кожному етапі та забезпечити гнучке налаштування окремих компонентів. Наприклад, помилки в генерації макету можуть бути проаналізовані окремо від помилок у фінальному синтезі зображення, що значно полегшує відлагодження та інтерпретацію результатів. Це робить такі системи придатними для складних застосувань, де потрібно гарантувати пояснюваність і стабільність результатів.

З точки зору користувацької взаємодії, проміжне представлення відкриває шлях до інтерактивної генерації. Користувач може змінювати положення, розмір або атрибути об’єктів без необхідності переформульовувати текстовий опис. Це особливо актуально у творчих

сферах (дизайн, сторітелінг, освітні платформи), де важлива гнучкість і точність візуального втілення ідеї.

Крім цього, сучасні підходи починають поєднувати переваги структурованого представлення з генеративною силою великих моделей – зокрема, через інтеграцію сценографів і текстових промптів у мультимодальні трансформери [20]. Таке поєднання дозволяє зберігати як семантичну точність, так і естетичну варіативність, відкриваючи нові горизонти у сфері генерації зображень на основі природної мови.

Таким чином, концепція проміжного семантичного представлення є не лише технічно ефективною, але й методологічно важливою. Вона надає моделі здатність до розуміння структури описаної сцени, чітко формалізує етапи генерації та забезпечує взаємозв’язок між мовним і візуальним доменами на концептуальному рівні. У межах даної роботи цей підхід розглядається як один із ключових механізмів досягнення високого рівня відповідності, керованості та пояснюваності у процесі автоматизованої генерації ілюстрацій.

1.3.3 Графові підходи на основі сценографів

Графові підходи, що ґрунтуються на використанні сценографів (англ. scene graphs), становлять один із найбільш структурованих і формалізованих напрямів у системах генерації зображення на основі тексту. Сценограф – це орієнтований граф, у якому вузли відповідають об’єктам сцени (наприклад, «людина», «стіл»), а ребра – відношенням між ними («сидить на», «поруч із», «тримати») [1]. Така модель дозволяє явно представити семантичну структуру опису, що значно полегшує контроль над генерацією та забезпечує збереження логічної цілісності сцени.

Однією з основних переваг графових підходів є здатність поєднувати різні типи інформації про об’єкти, зокрема:

– класи об’єктів (що саме зображено);

– розташування (де об’єкт має з’явитися на сцені);

– атрибути (наприклад, колір, форма, розмір);

– взаємозв’язки між об’єктами (наприклад, хто з ким взаємодіє або який об’єкт є частиною іншого).

Це дозволяє отримувати не лише візуально реалістичні зображення, але й сцени, що є семантично узгодженими з текстовим описом. На відміну від лінійних моделей, які можуть втрачати структурні залежності, графова модель забезпечує повноцінне представлення складних взаємодій між елементами.

Серед систем, що реалізують цей підхід, однією з найбільш відомих є SG2IM. У цій архітектурі граф сцени, побудований на основі текстового опису, проходить кілька етапів обробки: спочатку відбувається генерація об’єктного макету (англ. bounding boxes), далі – формування масок об’єктів і, нарешті, – генерація зображення на основі сцени. Архітектура включає модулі для обробки просторової інформації, зокрема графові згорткові шари [21], що дозволяє ефективно враховувати відношення між об’єктами.

Іншим прикладом є GRAF, що поєднує сценографи з тривимірним представленням сцени. Ця система забезпечує високий ступінь узгодження між текстом і зображенням, а також дозволяє враховувати глибину, освітлення та вид з камери. Це особливо важливо в задачах фотореалістичної генерації.

Також слід відзначити InterACT. Ця система реалізує інтерактивне редагування сценографа, дозволяючи користувачу змінювати окремі об’єкти, їх положення та атрибути – з миттєвим оновленням зображення. Такий підхід відкриває нові можливості у створенні персоналізованого або адаптивного контенту.

Графові методи також добре масштабуються до великих сцен і підтримують модульність. У складних описах, де є багато об’єктів і взаємозв’язків, саме графова структура дозволяє уникнути плутанини і забезпечити чітке представлення семантики.

Ще однією важливою властивістю є інтерпретованість: сценографи легко візуалізуються, що дозволяє розробникам і користувачам аналізувати, як саме опис трансформувався у структуру сцени, а відтак – і у зображення. Це створює умови для побудови пояснюваних моделей, що є особливо актуальним у сферах, де критичною є прозорість прийняття рішень.

У підсумку, графові підходи демонструють високу ефективність у задачах, що вимагають точної відповідності між описом і зображенням, контролю над кожним компонентом сцени та можливості редагування. Вони забезпечують концептуальну чіткість, структурну цілісність і відкривають перспективи для побудови гнучких мультимодальних систем генерації контенту. У межах даної роботи графи сцен виступають як ключовий механізм забезпечення семантичного контролю в процесі автоматизованої генерації ілюстрацій на основі тексту.

1.3.4 Мультимодальні моделі з комбінованими просторовими і зовнішніми атрибутами

Сучасні мультимодальні системи генерації зображень з тексту дедалі частіше орієнтуються на поєднання семантичного структурування сцени із гнучким відображенням візуальних атрибутів об’єктів. Одним із перспективних напрямів є розділення вхідної інформації на просторове представлення (англ. layout embedding) та зовнішній вигляд (англ. appearance embedding). Такий підхід забезпечує декомпозицію сцени на логічні та візуальні складові, дозволяючи окремо контролювати положення, форму та стилізацію кожного об’єкта.

Ключовою перевагою подібної архітектури є гнучкість у керуванні кожним елементом сцени. Користувач або модель можуть змінювати просторові характеристики (наприклад, координати обмежувальних рамок), не зачіпаючи зовнішні ознаки об’єкта, і навпаки – редагувати вигляд об’єкта (колір, текстуру, стиль), зберігаючи його логічну позицію в межах

сцени. Це відкриває нові можливості для інтерактивного редагування, персоналізації візуального контенту, а також для навчання моделей, що зберігають семантичну цілісність навіть після локальних змін.

Однією з найбільш показових реалізацій цієї концепції є система InterACT [19]. У цій роботі модель отримує сцену у вигляді явної множини об’єктів, кожен з яких описаний через два вектори: вектор просторового представлення, який кодує позицію об’єкта (координати, розмір), та вектор зовнішнього вигляду, який містить його зовнішні характеристики (візуальний стиль, форма, колір). Обидва вектори проходять через окремі нейронні підмоделі, після чого результат інтегрується у фінальний вектор сцени, який передається генератору зображень.

Особливістю InterACT є також підтримка ітеративного оновлення сцени. Користувач може змінювати окремі атрибути одного або кількох об’єктів, не порушуючи при цьому цілісної структури сцени. Наприклад, змінити «зелений м’яч» на «червоний м’яч», зберігаючи його розташування під столом, без потреби генерувати всю сцену з нуля. Така властивість є особливо важливою у випадках, коли генерація зображення відбувається в діалоговому або інтерактивному режимі – наприклад, у дизайні інтерфейсів, системах асистивної візуалізації або креативних застосунках.

Крім підвищеної варіативності результатів, даний підхід сприяє структурному моделюванню складних сцен, які складаються з багатьох об’єктів із унікальними характеристиками. Замість глобального латентного простору, де кожен новий опис потенційно змінює всю сцену, комбінована модель дозволяє локально оновлювати вміст, підвищуючи керованість і пояснюваність системи.

Недоліком таких архітектур є зростання обчислювальної складності: кожен об’єкт має окрему репрезентацію, що вимагає більших ресурсів для обробки та тренування. Крім того, побудова вбудовувань просторового

представлення та зовнішнього вигляду потребує точних анотацій, або ж додаткового модуля автоматичної екстракції атрибутів.

Попри ці виклики, комбіновані мультимодальні підходи становлять важливий крок до побудови гнучких, редагованих та інтерпретованих систем генерації зображень. Вони дозволяють реалізовувати індивідуалізовану генерацію, адаптацію до контексту користувача і збереження логічної структури сцени при динамічних змінах. У межах даного дослідження такі моделі є перспективним напрямом розвитку інтелектуальних ілюстративних систем, які оперують як семантикою, так і візуальною стилістикою зображення.

Отже, сучасна динаміка розвитку методів генерації зображень з тексту свідчить про поступовий відхід від підходів прямої генерації на користь структурованих та напівструктурованих архітектур, які забезпечують модульність, інтерпретованість та контрольованість генеративного процесу. Хоча моделі типу «перетворення тексту на пікселі» залишаються ефективними для генерації простих сцен, вони демонструють суттєві обмеження у відтворенні складних семантичних структур, особливо в умовах насичених текстових описів, відповідно до таблиці 1.1.

Таблиця 1.1 – Порівняльний аналіз підходів до генерації зображень

Підхід	Текст-піксельне перетворення	Проміжне семантичне представлення	Графові підходи	Мультимодальні моделі
Опис	Безпосередній синтез зображення з тексту	Побудова макету або семантичної карти перед візуалізацією	Використання графів сцен	Поєднання просторових та візуальних атрибутів об’єктів
Приклади моделей	StackGAN, AttnGAN	Layout2Image, SG2IM	SG2IM, GRAF, InterACT	InterACT, GRAF

Продовження таблиці 1.1

Контроль генерації	Низький, результати важко передбачити	Середній, через проміжне представлення	Високий, можливість редагування окремих об'єктів	Дуже високий, незалежне керування положенням та виглядом
Відповідність опису	Обмежена, помилки в логіці сцени	Точні просторові відносини	Висока точність розташування	Дуже висока через розділення структури та стилю
Пояснюваність	Низька («чорна скринька»)	Середня через проміжну розмітку	Висока, можливість візуалізації графа	Висока через розділення атрибутів
Складність реалізації	Відносно низька	Середня, потребує додаткової розмітки	Висока, потребує обробки графів	Висока, потребує двох незалежних репрезентацій
Можливість редагування	Обмежена, вимагає повторної генерації	Редагування макету перед генерацією	Гнучке редагування об’єктів, зв’язків та атрибутів	Незалежне редагування положення та вигляду об’єктів
Обчислювальні витрати	Середні	Високі	Високі	Дуже високі

Підходи, що передбачають використання проміжного представлення – таких як семантичні макети, сценографи або комбіновані мультимодальні репрезентації – дозволяють формалізувати зміст сцени до етапу візуалізації. Це відкриває можливості для редагування структури сцени, підвищення точності локалізації об’єктів, відображення атрибутів та просторових зв’язків, що особливо актуально в системах

ілюстрування [22], [23], візуалізації даних та мультимедійного сторітелінгу [24].

Перспективи подальшого розвитку лежать у напрямку уніфікації мовних, графових і візуальних репрезентацій, що дозволить створювати більш гнучкі, пояснювані та адаптивні генеративні системи. Інтеграція структурованого знання про сцену з потужними генеративними архітектурами відкриває нові горизонти в автоматизованому створенні ілюстрацій, де машинна модель здатна не лише «бачити» зображення, а й «розуміти» його логіку та зміст. У контексті даного дослідження такі підходи становлять концептуальну основу для реалізації інтелектуальної системи, що здатна формувати узгоджене візуальне представлення на основі описової текстової інформації.

1.4 Постановка задачі дослідження

У межах даного дослідження розглядається задача автоматизованої генерації ілюстрацій на основі текстового опису, що передбачає перетворення природномовного тексту на зображення із збереженням смислової відповідності. Метою дослідження є розробка та оцінка комбінованого підходу, який поєднує мовний аналіз, побудову структурованого представлення сцени у вигляді графа та подальшу генерацію зображення за допомогою моделі, заснованої на графовому описі. Такий підхід спрямований на досягнення вищого рівня зрозумілості, керованості та логічної узгодженості між текстовим описом і результатом у вигляді зображення.

Більшість сучасних моделей перетворення тексту на зображення, зокрема моделі глибокого навчання на основі змагального або стохастичного навчання, хоч і забезпечують високу якість візуального результату, виявляють суттєві обмеження при роботі зі складними описами, які містять кілька об’єктів, просторові розміщення та взаємозв’язки між

об’єктами. Такі моделі, зазвичай, працюють як «чорні скриньки», де неможливо передбачити точне розташування елементів сцени або перевірити відповідність зображення змісту опису. Це створює серйозні обмеження у використанні подібних систем у сферах, де важлива точна інтерпретація інформації: освіта, візуалізація знань, науково–популярні ілюстрації тощо.

Аналіз існуючих підходів свідчить, що:

– моделі прямого перетворення ігнорують внутрішню структуру сцени, що призводить до втрати логіки зв’язків між об’єктами;

– методи зі структурованим представленням (наприклад, макети або семантичні карти) вимагають складного налаштування або ручного створення структури сцени;

– моделі, що працюють із графами, забезпечують високу точність, але потребують окремих засобів для автоматичного побудування графа за текстом.

У цьому контексті доцільним є комбінований трирівневий підхід:

– текст у граф сцени, який формується автоматично на основі мовної моделі Flan-T5 [25];

– граф сцени у формат для моделі генерації, де граф перетворюється спеціалізованими функціями у придатну структуру;

– формалізоване представлення у зображення, що генерується з урахуванням об’єктів, їх розміщення та відношень.

Розглянутий підхід дозволяє формалізувати зміст сцени, зберігати логіку опису, а також забезпечити керований та прозорий процес генерації. Завдяки можливості поетапної обробки, система може бути адаптована до редагування сцен і забезпечити високу відповідність результату до початкового тексту, на відміну від повністю автоматичних методів, що не враховують структуру.

2 ТЕХНОЛОГІЯ АВТОМАТИЗОВАНОГО ГЕНЕРУВАННЯ ІЛЮСТРАЦІЙ НА ОСНОВІ СЦЕНОГРАФІВ, ОТРИМАНИХ З ТЕКСТУ

2.1 Загальна архітектура системи

Розглянута технологія покликана забезпечити керовану генерацію зображень на основі текстового опису через проміжне формалізоване представлення у вигляді графа сцени. Вона повинна дозволити створювати зображення, які не тільки точно відображають описані об'єкти, але й правильно інтерпретують їх характеристики та просторові відносини. Для цього з тексту автоматично вилучається інформація про об’єкти та відношення, яка перетворюється на сценограф, на основі якого синтезується фінальне зображення.

Розглянута гібридна технологія об'єднує переваги підходів прямого синтезу зображення з тексту, структурованого представлення або графу, усуваючи недоліки кожного з них.

Відсутність чіткої логічної структури тексту в підході перетворення тексту на зображення компенсується використанням графа сцени, що забезпечує точне відтворення смислу природньомовного опису, а також прозорість і контрольованість процесу.

Складне ручне налаштування, необхідне в підході перетворення структурованого представлення на зображення замінюється автоматичною генерацією макетів розташування, обмежувальних рамок та масок форм за допомогою ЗНМ. Зазначений підхід є менш креативним, оскільки вимагає від користувача самостійно визначати структуру сцени. Натомість, гібридний підхід забезпечує створення нових унікальних композицій.

Необхідність попередньої побудови графа та висока чутливість до формату вхідних даних у підході перетворення графу на зображення в межах розглянутої технології усувається за рахунок автоматизованого

вилучення графа сцени з тексту. При цьому зберігається можливість повторного використання побудованих графів сцен.

Крім зазначених переваг, технологія дозволяє редагувати генерацію на різних її етапах: як шляхом корекції вхідного текстового опису, так і проміжного графу сцени. Останній підхід особливо ефективний: він дає змогу тонко настроювати окремі елементи без перегенерації всього зображення. Доступні такі види модифікацій:

а) редагування просторового представлення:

– просторові зміни: переміщення, масштабування та корекція форми об'єктів із збереженням логіки сцени;

– оновлення взаємодій. Наприклад, зміна положення частин тіла персонажів, напрямку погляду тощо;

– додавання/видалення елементів без впливу на інші готові частини композиції;

б) візуальні налаштування: корекція кольорів, текстур та інших атрибутів зовнішнього вигляду окремих компонентів сцени.

Це забезпечує ідеальний баланс між оригінальністю синтезованих зображень та їх придатністю до цілеспрямованих користувацьких уточнень.

Отже, використання трирівневої технології перетворення тексту у граф і в зображення забезпечує: семантичну точність та контрольованість генерації, пояснюваність процесу, покращену узгодженість сцени, адаптивність до складних описів, інтерактивність, скорочення обчислювальних помилок завдяки багатомодальності, широке застосування в професійних сферах.

Для реалізації цієї технології передбачено три автономні етапи (рисунок 2.1):

– аналіз тексту та побудова графа сцени мовною моделлю Flan-T5. Трансформер аналізує вхідний текст, виділяючи ключові об'єкти, їх атрибути та взаємозв'язки. Результатом цього етапу є граф сцени у формі списку відношень між об’єктами та їхніми атрибутами;

– оптимізація графа, під час якої спеціалізований трансформатор приводить граф до нормалізованого словникового формату, який очікує модель генерації зображень;

– генерація фінального зображення моделлю SG2IM на основі оптимізованого графа з відтворенням всіх об'єктів з їхніми атрибутами та просторовими відношеннями. Модель враховує як глобальну композицію, так і локальні деталі кожної складової.

Модульна архітектура додатково дозволяє легко знаходити недоліки та вдосконалювати окремі компоненти системи.

Рисунок 2.1 – Архітектура розглянутої технології

Під час етапу екстракції сценографів із тексту необхідно зазначити певні обмеження, що дозволяють зменшити складність аналізу, проте можуть впливати на повноту та деталізацію отриманих сценографів.

По-перше, не враховуються кількісні та якісні характеристики об’єктів, за винятком їхнього розміру. Це означає, що атрибути, які описують кольори, форми чи інші якісні характеристики, не будуть зберігатися при екстракції. Наприклад, у реченні «Два хлопчики сидять на червоній лавці» числівник «два» і прикметник «червона» не будуть

враховані як атрибути «хлопчиків» та «лавки». Однак, якщо текст містить інформацію про розміри об'єктів, наприклад, «велика лавка», ці розміри будуть збережені в сценографі.

Окрім цього, система обмежується лише прямими відносинами між об’єктами та не виводить симетричні зв’язки. Наприклад, у реченні «Собака знаходиться поруч із дівчинкою» буде зафіксований факт, що «собака» знаходиться «поруч із дівчинкою», але зворотній зв’язок «дівчинка поруч із собакою» не буде доданий.

По-третє, буде передбачена екстракція лише просторових відносин між об'єктами на основі фіксованого словника (наприклад, «поруч», «над», «під», «перед», «за»), а характеристики дій об’єктів ігноруватимуться.

Зрештою, в рамках даної роботи буде прийняте до уваги обмежене редагування отриманих ілюстрацій користувачем. Тобто, у процесі генерації неможливо буде змінити орієнтацію об'єктів, додати або змінити їхні текстури, кольори чи інші візуальні характеристики.

Навіть з урахуванням зазначених обмежень, головною перевагою гібридного підходу залишається поєднання семантичної точності (завдяки структурі графа) та творчої варіативності (через нейромережеву генерацію). Далі детально розглянемо роботу окремих модулів в рамках розглянутої технології.

2.2 Модуль семантичного аналізу тексту

Модуль семантичного аналізу тексту призначений для перетворення текстових описів сцен у графи, які можуть бути використані для подальшої формалізації та генерації зображень. Основне завдання модуля – екстракція об’єктів, їхніх атрибутів і взаємозв’язків. Для досягнення цього, вхідний текстовий опис в рамках модуля проходить два етапи обробки (рисунок 2.2):

– нормалізація тексту – приведення до нижнього регістру, лематизація та токенізація;

– екстракція семантичних компонентів – виявлення об'єктів, атрибутів та відносин між об'єктами за допомогою навченого трансформера. Отримується вихід у вигляді списку відношень графа, який ще потребує подальшої формалізації для синтезу зображення.

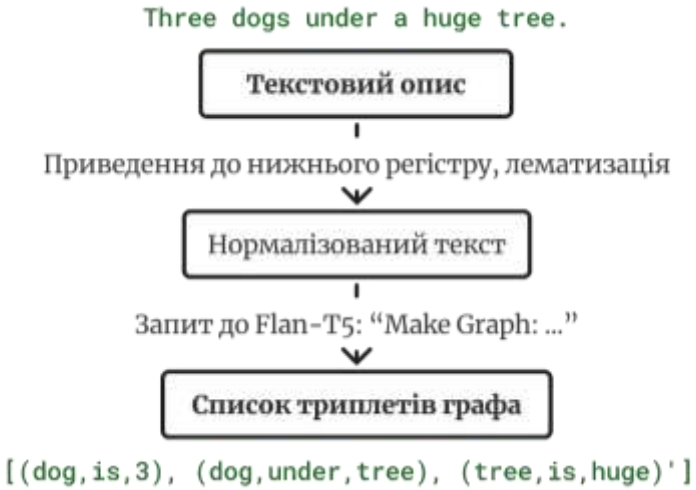

Рисунок 2.2 – Архітектура модуля семантичного аналізу тексту

Синтаксичний аналіз тексту для побудови графів може включати різноманітні підходи. Класичні методи обробки природної мови (ОПМ) спираються на проміжні представлення, наприклад, графи залежностей [4] або Абстрактне Представлення Значення (англ. Abstract Meaning Representation, AMR) [26]. Ці проміжні структури конвертуються у сценографи за допомогою детермінованих правил або моделей машинного навчання.

Парсинг на основі залежностей (англ. dependency parsing). Одним із підходів до створення сценографів є використання графів залежностей, які моделюють синтаксичні зв’язки. В таких графах вузли представляють слова в реченні, а ребра – їхні граматичні відносини (наприклад, підмет, присудок, додаток). Цей підхід популярний через свою простоту та доступність, проте він не завжди відображає семантичну суть тексту на високому рівні абстракції.

Наприклад, SPICE-Parser [27] перетворює графи залежностей, отримані з описів, у сценографи за допомогою наперед визначених граматичних правил. Цей метод має два основні недоліки. По-перше, жорстко задані правила можуть накопичувати помилки при перетворенні графів залежностей у сценографи. Він також обмежений у розумінні складних або неоднозначних описів, оскільки не враховує глибоку семантику. По-третє, нестача або неточності у навчальних даних знижують узагальнюючі здатності моделі на нових сценаріях.

Абстрактне семантичне представлення – це альтернативне графове семантичне представлення тексту, яке моделює сенс речення у форматі «хто що робить і кому». Воно подає речення як спрямований ациклічний граф, де вузли – це семантичні концепти, а ребра – їхні відношення. Цей метод фокусується на витягненні семантичних концепцій із тексту, одночасно абстрагуючись від синтаксичних представлень. Таким чином, реченням, які мають подібне значення, повинні бути надані однакове представлення AMR, навіть якщо вони сформульовані по-різному.

Задля цього алгоритм спрощує речення, виокремлюючи головні елементи: дію, виконавця та об'єкт. Наприклад, речення «Кіт переслідував мишу» та «Миша була переслідувана котом» отримають однаковий граф, де кіт – агент (:arg0), переслідувати – дія, а миша – об'єкт (*:arg1*). Методами є статистичні або нейронні AMR-парсери.

AMR вважається більш придатним для побудови сценографів, оскільки враховує семантику і дозволяє будувати більш узагальнені графи. Проте, оскільки підхід базується на стандартизованих фреймах, речення із різним контекстом можуть отримувати однакові графи, через що втрачаються специфічні деталі.

Крім того, анотація AMR-графів потребує значних ресурсів і знань, достатніх для покриття всіх можливих варіантів текстів. До зазначених недоліків варто додати складність реалізації AMR-парсингу, через що

більшість досліджень у сфері генерації сценографу з тексту поки що уникають його використання.

Конвертація тексту в проміжні графи, а потім у сценограф створює додаткові етапи, де накопичуються помилки, що погіршує кінцеві результати. Таким чином, використання проміжних представлень в задачі генерації сценографу з тексту часто призводить до низької продуктивності в подальших завданнях.

Альтернативним та більш сучасним підходом до задачі вилучення сценографу з тексту є наскрізне перетворення тексту в сценографи за допомогою моделей «кодувальник-декодувальник», що використовують обидві частини архітектури трансформерів. Вони дозволяють безпосередньо генерувати графи з тексту, що мінімізує накопичення помилок.

Трансформери є ідеальними для задач, пов’язаних з генеруванням нових речень чи структур. Це включає узагальнення, переклад та генерацію графів, що відповідають текстовим описам. Основна перевага таких моделей полягає в їхній універсальності: вони можуть бути налаштовані для вирішення багатьох задач ОПМ, таких як вилучення сценографів з тексту.

Розглянемо принцип роботи моделі T5 (Text-to-Text Transfer Transformer) [28], яка є класичним представником таких моделей та «прабатьком» Flan-T5, використаного у модулі семантичного аналізу. Вона складається з двох частин: кодувальника та декодувальника, що робить її універсальною для різних завдань. На відміну від моделей BERT чи GPT, які мають лише один із цих компонентів, T5 здатна не тільки розуміти текст, а й генерувати структуровані дані.

Кодувальник в T5 перетворює вхідний текст на послідовність вбудованих токенів, яка передається на декодувальник. Він, у свою чергу, за допомогою маскованої самоуваги генерує вихідний текст. Таким чином, T5 приблизно еквівалентна оригінальному трансформеру, запропонованому Vaswani ed al. (2017) [18], за винятком видалення зміщення норми шару,

розміщення нормалізації шару за межами залишкового шляху та використання іншої схеми вбудовування позиції.

Ця модель є потужною та універсальною, здатною працювати з великими наборами текстових даних. Однак її можна вдосконалити для специфічних задач, як-от генерація графів, шляхом донавчання на спеціалізованих наборах даних.

Наприклад, для адаптації моделі T5 до задачі генерації графів був розроблений VG-T5 (Visual Genome T5) [29]. Тонке настроювання на парах «текст-сценограф» дозволило моделі точніше відображати реляційні структури та генерувати графи, які відображають відносини між об’єктами в тексті. Наприклад, текст «чоловік стоїть з дитиною на схилі» може бути перетворений на граф, де кожен факт представлений окремим ребром або відношенням між об’єктами (наприклад, «дитина, на, схилі»).

Для підвищення точності використовуються промпти із префіксами: до вхідного тексту додається маркер «Make graph:», що дозволяє моделі чітко ідентифікувати завдання генерації графа. Отримані графи кодуються у вигляді послідовностей із спеціальними маркерами (SEP, EOF) для розділення елементів.

Модель Flan-T5-XL (2022) [25] є іншою удосконаленою версією T5, яка була спочатку навчена на контрольних точках T5 XL, а потім додатково тонко налаштована на широкому інструкційному наборі даних FLAN. Це зробило модель універсальною для роботи над різними завданнями, включаючи генерацію графів.

Окрім цього, модель Flan-T5 проходить донавчання на наборі даних FACTUAL [30] (таблиця 2.1), зосередженому на парсингу саме сценографів. Це покращує здатність моделі розпізнавати складні реляційні відносини та генерувати графи сцен на основі текстових даних.

Таким чином, використання моделей «послідовність до послідовності», таких як Flan-T5, є ефективним підходом для екстракції семантичних компонентів із тексту в модулі семантичного аналізу.

Таблиця 2.1 – Приклад записів із навчального датасету FACTUAL

image_id	region_id	caption	scene_graph
2365262	2416695	people sitting in bleachers	(people, v:sit in, bleachers)
2326684	4258207	water is coming out of a hydrant	(water, v:come out of, hydrant)
2408309	300332	there is a car on the train track	(car, on, train track)

Для синтезу графа сцени модулю подається детальний текстовий опис. В результаті попередньої обробки тексту, його нормалізований варіант подається на вхід моделі Flan-T5 із префіксом «Make graph:». Для створення сценографу на основі введеного промпту в моделі застосовується метод променевого пошуку (англ. beam search), що дозволяє генерувати кілька варіантів результатів на кожному кроці, зберігаючи найкращий. Це підвищує точність і коректність створеного сценографа. В результаті генерації, модуль надає список триплетів графа сцени. Наприклад, отримавши на вхід текст «Three dogs under a huge tree», модуль повертає [(dog, is, 3), (dog, under, tree), (tree ,is ,huge)].

Модуль семантичного аналізу тексту дозволяє ефективно перетворювати текстові описи на структуровані семантичні графи, використовуючи сучасні моделі обробки природної мови (Flan-T5). Це усуває недоліки традиційних підходів, спрощує процес генерації структурованих представлень і відкриває нові можливості для автоматичної обробки тексту.

2.3 Попередня обробка сценографа

Одним із найбільш перспективних напрямків у використанні моделей типу T5 є перехід від прямої генерації зображень з тексту до використання семантичних графів як проміжних представлень. Ці графи стають основою для аналізу текстової інформації та розбивають процес генерації зображень

на кілька етапів: аналіз тексту, створення графа, побудову сцени та її візуалізацію.

Серед інших переваг семантичних графів є їх масштабованість. Графи легко можна доповнювати новими об'єктами чи взаємозв'язками. Також полегшується створення нових варіацій сцени, наприклад: зміна розташування об'єктів, їх атрибутів чи оточення.

Семантичний граф є способом представлення знань у формі графової моделі. Він подає інформацію у вигляді вузлів і ребр, а також структурує таким чином, щоб її можна було використовувати для аналізу [31]. Основний фокус в семантичному графі на об'єктах та логіці їх взаємодії.

Натомість, сценограф – це наступний, конкретизований варіант семантичного графа, створений для візуальної генерації сцени. Він описує розташування, форму та розміри об'єктів у двовимірному чи тривимірному просторі, роблячи фокус на просторових характеристиках сцени. Конкретизація семантичного графу виконується за рахунок надання об'єктам в сцені додаткових атрибутів (позиція, розмір, орієнтація). Ребра тепер відповідають саме за просторові взаємозв'язки (наприклад, «над», «зліва», «всередині»). Сценограф додає геометричні деталі, тим самим конкретизуючи семантичний граф для побудови візуальної сцени.

Щоб побудувати сценограф на основі семантичного графу, необхідно, що кожен об'єкт (наприклад, «діти» або «м'яч») отримав свої координати та атрибути, і, що не менш важливо, відносини. Останні тепер мають враховувати не дії і інші семантичні особливості, а саме взаємне розташування об'єктів (наприклад, що стоїть вище – діти чи м’яч) і їх глобальні позиції на сцені. Отримана семантична розмітка визначить структуру сцени і детальну інформацію про неї, зокрема категорії об’єктів, їх розташування, розміри тощо. Таким чином, подальша генерація зображення буде обумовлена прогнозованим макетом і дозволить досягти осмислених результатів.

Модуль попередньої обробки сценографа виконує перетворення вихідного семантичного графа у формат, який може бути використаний моделлю SG2IM. Основне його завдання полягає у стандартизації структури, отриманої від модуля семантичного аналізу тексту для забезпечення її сумісності з моделлю генерації зображень. Для цього з графу вилучаються вузли (об'єкти), атрибути (характеристики об’єктів) та ребра (відносини між об'єктами). Важливою частиною цієї роботи є класифікація типів відносин на просторові та атрибутивні. В результаті, це дозволить моделі SG2IM зрозуміти, як об'єкти повинні бути організовані та виглядати в фінальній ілюстрації.

На цьому етапі варто заздалегідь вирішити такі питання:

– які типи відносин будуть включені в граф (лише просторові («собака під столом»), або також семантичні («людина тримає м’яч»), ієрархічні («будинок має двері»));

– чи матимуть об’єкти атрибути, які впливають на їх візуальне представлення, розмір чи кількість.

На основі відповідей на ці питання, здійснюється екстракція ключових елементів із вхідного графу. Об'єкти визначаються на основі наявних у списку триплетів об’єктів, якщо ті присутні у словнику класів. При цьому кожен об’єкт отримує унікальний індекс, який використовується для його ідентифікації у взаємозв’язках та атрибутах. На початку кожному об'єкту задається стандартне значення його розміру в шкалі від 1 до 10, яке визначається зі словника відносних розмірів різних класів об'єктів. За наявності відповідної інформації у відношеннях об’єкта, атрибут його розміру може бути модифікований. Наприклад, у відношенні «кішка велика», згенерованому трансформером, система розпізнає «велика» як описовий атрибут об'єкта «кішка» і фіксує його розмір як великий. Тоді чисельне значення атрибуту розміру об’єкта збільшується за попередньо визначеними правилами.

Окрім атрибутів, важливу роль відіграють відношення між об’єктами. Для покращення їх розпізнавання використовується попередня лематизація. У межах дослідження зв'язки між об'єктами будуються на основі просторових відносин, таких як «поруч з» або «на ліво від». Вони включаються у граф як трійки: [індекс_суб'єкта, відношення, індекс_об'єкта]. Відношення, що не стосуються геометричного розташування, фільтруються. Це дозволяє побудувати уніфікований граф, який відображає лише релевантні просторові взаємозв’язки між об’єктами. На основі виявлених відношень також автоматично генерується грубе положення об'єктів на сітці розміром 5×5.

Крім цього, може знадобитися злиття однакових об'єктів для спрощення структури графа. Деякі підходи також включать зміну типу відносин або атрибутів об'єктів залежно від ситуації, в якій їх знайдено. Наприклад, «яблуко на столі» може отримати додаткові атрибути, такі як «поруч із ножем», якщо також присутнє відношення «ніж на столі». Втім, в межах даного дослідження здійснюється оптимізація структури графа, тому логічний вивід відношень не передбачено.

Якщо у тексті присутні кількісні відносини, наприклад, «три яблука у тарілці», вони несуть важливу інформацію про кількість об'єктів одного типу у сцені. Існує кілька підходів до їх обробки. Перший – це групування, коли у граф додається один вузол із зазначенням типу об'єкта («яблуко») та відповідним атрибутом кількості (наприклад, кількість: 3). Другий спосіб це реплікація вузлів. Такий підхід дозволить відобразити розташування та атрибути кожного об'єкта окремо (наприклад: «яблуко1», «яблуко2», «яблуко3»). В рамках дослідження кількісні особливості вирішено ігнорувати, тому кількісні атрибути видаляються і всі об’єкти вважаються наявними у єдиному екземплярі.

Фінальним етапом сценограф зберігається у форматі, підходящому для передачі на модуль генерації зображення. Це може бути JSON, матриці суміжності (дозволять компактно представити зв'язки у графі, але не такі

гнучкі для зберігання складних наборів атрибутів) або через API для графів (бібліотеки на кшталт NetworkX, що підтримують оптимізовану роботу з великими графами). В нашому випадку використовується перший варіант, оскільки він забезпечує зручне представлення і є сумісним із SG2IM (рисунок 2.3).

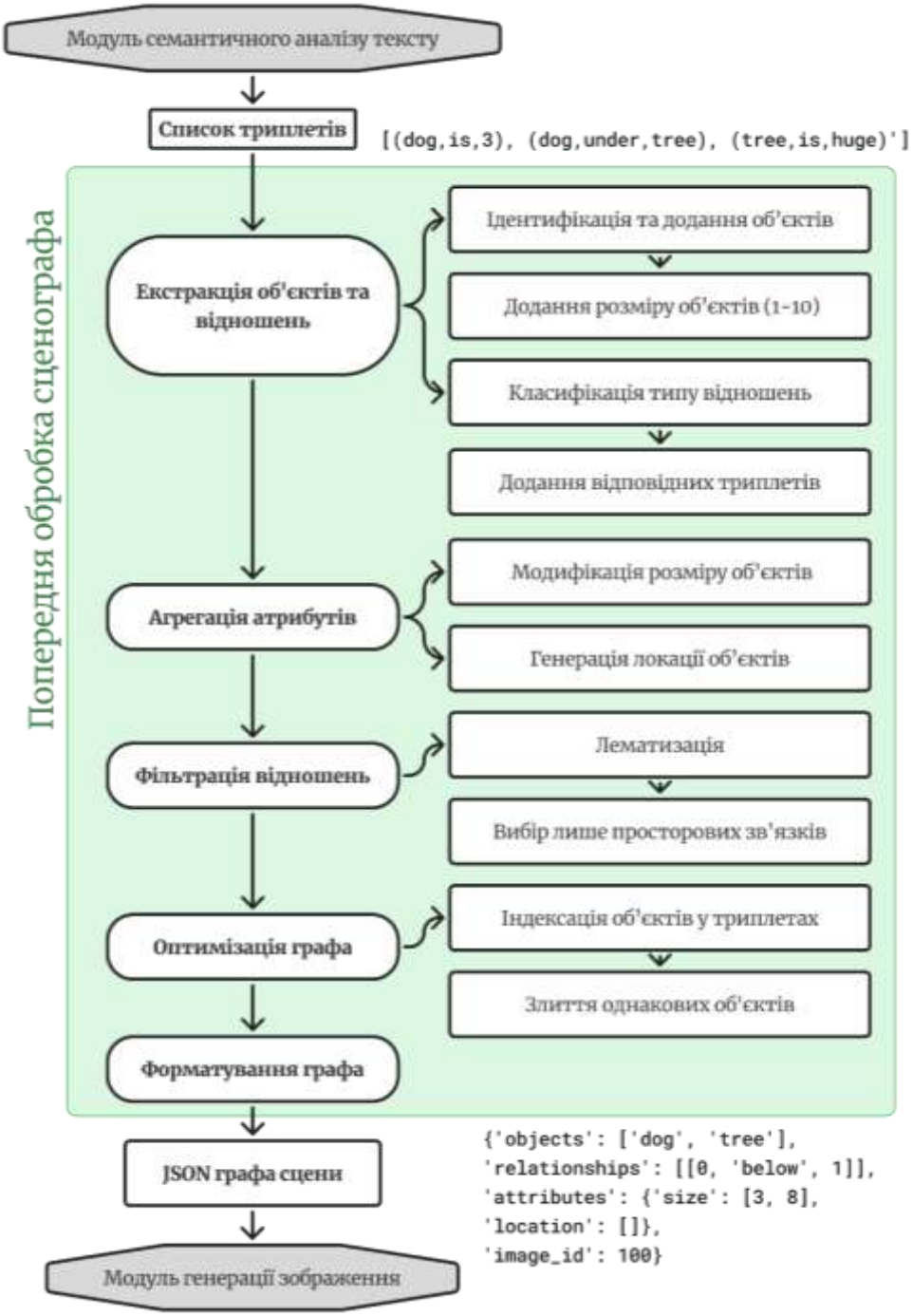

Рисунок 2.3 – Архітектура модуля попередньої обробки сценографа

В результаті трансформації модулем, сценограф має таку структуру:

– об’єкти (англ. «Objects») – список всіх об'єктів у сцені;

– відношення (англ. «Relationships») – список взаємозв'язків між об'єктами [індекс_суб'єкта, відношення, індекс_об'єкта];

– атрибути (англ. «Attributes») – зберігає атрибути, а саме розмір та місце розташування об'єктів;

– ознаки (англ. «Features») – додаткові характеристики зовнішності, які можуть бути отримані з зображень;

– ідентифікатор зображення (англ. «Image_id») – унікальний ідентифікатор зображення, що може бути прив'язаний до конкретного графа.

Важливим аспектом є наявність атрибута місця розташування, який використовується для подальшої генерації координат об'єктів у сцені.

Описаний процес екстракції є основою для подальшого використання сценографа для генерації ілюстрацій. Від рішень на цьому етапі залежить якість та зміст майбутніх зображень. Можливо налаштувати фокус своєї моделі на різних аспектах зображень, збагатити чи спростити передачу властивостей об’єктів з тексту тощо.

2.4 Модуль генерації зображення

В рамках даного дослідження отриманий на попередньому етапі JSON графа сцени подається на вхід модулю синтезу зображення. В цьому контексті визначальним кроком стає вибір ефективної генеративної моделі. Сучасні методи генерації зображень на основі графів умовно поділяються на кілька категорій, кожна з яких має свої переваги та обмеження [32].

Генеративні змагальні мережі (англ. generative adversarial networks, GANs) [17], зокрема сценографічні мережі (англ. scene graph GANs) [33], використовують сценографи для створення векторних представлень, які потім трансформуються в зображення через генератор. Дискримінатор

оцінює ці зображення, щоб забезпечити їх реалістичність. Хоча GAN здатні генерувати високоякісні зображення, процес їх навчання часто є нестабільним.

Інша категорія – дифузійні моделі, працюють за принципом поступового видалення шуму, керуючись структурою графа. Такий підхід забезпечує високу деталізацію, але вимагає значного часу генерації та обчислювальних ресурсів, не передбачених умовами даного дослідження.

Варіаційні автокодувальники (ВАК, англ. variational autoencoder) трансформують вхідні дані в ймовірнісні розподіли, що дозволяє генерувати нові варіанти зображень та робить їх особливо цінними для творчих завдань. Традиційні ВАК адаптуються у графові ВАК, де вхідними даними виступають не пікселі зображень, а складні структуровані об'єкти з їх атрибутами та взаємозв'язками. Однак, порівняно з GAN або дифузійними моделями, ВАК можуть створювати менш деталізовані зображення.

Для подолання цих обмежень найбільш потужним варіантом є архітектура ВАК-GAN [34], яка поєднує переваги обох підходів. У цій моделі ВАК відповідає за ефективне кодування складних графових структур, яке потім використовується GAN-мережею для забезпечення вищої реалістичності фінальних зображень. Яскравим прикладом успішної реалізації цього підходу є модель SG2IM.

Модель SG2IM – це нейронна мережа, яка перетворює сценографи на точні реалістичні зображення. Вхідний JSON описує об'єкти, їх властивості та взаємозв'язки, що дозволяє моделі генерувати зображення з урахуванням контексту. Для цього модель включає кілька компонентів: графову ЗНМ [35], звичайні ЗНМ, мультиплексор та резидуальну мережу для фінальної генерації. Розглянемо ці елементи в деталях.

Графова згорткова мережа (G) перетворює сценограф на вектори розташування об'єктів. Обробка здійснюється за механізмом дифузії інформації, що дозволяє враховувати локальні та глобальні залежності. Вхідними даними для мережі є: інформація про вершини графа (клас об'єкта

та його грубе розташування на сітці 5×5) та інформація про ребра (кодують відношення між об'єктами через вивчені вкладення). Графова ЗНМ комбінує ці дані для створення згортки графа, що складається з п’яти шарів. Вони дозволяють краще розуміти структуру сцени мережами генерації.

ЗНМ генерації масок (M) перетворює вкладення розташування об'єктів на бінарні маски, додаючи випадкові вектори для варіативності. Вона складається з шести згорткових шарів і функції активації сигмоїди.

3-шарова ЗНМ генерації рамок (B) працює паралельно з M, генеруючи рамки для об'єктів на основі їх вкладень, та додає нелінійність через функцію ReLU.

ЗНМ генерації зовнішнього вигляду (A) перетворює дані про клас об’єкта в компактне векторне представлення. Вона включає три згорткових шари, шар глобального усереднення та три повнозв’язних шари.

Мультиплексор (T) – це фіксована функція, яка об'єднує маски, рамки та вектори зовнішнього вигляду в багатовимірний тензор для подальшої обробки. Це дозволяє моделі обробляти кілька вхідних даних одночасно.

Залишкова мережа-автокодувальник (англ. residual network, R) генерує фінальне зображення на основі тензора, отриманого мультиплексуванням. Вона складається з кодувальника та декодувальника, які відповідно забезпечують зменшення та збільшення розмірності зображення. Пропускові з'єднання при цьому дозволяють уникнути зникання градієнтів, прискорюють навчання і знижують складність мережі.

Ця заключна модель виконує кілька важливих функцій: забезпечує узгодженість деталей (стилю, форми, кольору) між об'єктами та їх згенерованими відповідниками, очищає зображення від артефактів і додає специфічні деталі (освітлення, колір тощо), що не були очевидні у вхідному графі. Саме завдяки автокодувальнику SG2IM створює більш реалістичні зображення та точніше відображає просторові та семантичні відношення.

З огляду на архітектуру, процес роботи SG2IM поділяється на кілька етапів (рисунок 2.4). Вони включають перетворення сценографа у векторне

представлення, де об'єкти і їх відношення кодуються для подальшої обробки. Сценограф передається до мережі G для створення вбудовувань розташування об'єктів, а також додається випадковий вектор для варіативності масок. Мережа M обчислює вбудовування маски, а мережа B одночасно генерує обмежувальну рамку. Мережа A створює зовнішній вигляд об'єктів на основі їх належності до класів. Генерація зовнішнього вигляду здійснюється з використанням змагального підходу, формуючи два тензори з обмежувальними рамками та масками для навчання. Мультиплексор T об'єднує інформацію про зовнішній вигляд, клас і розташування об'єктів, а автокодувальник R генерує на його основі фінальне зображення.

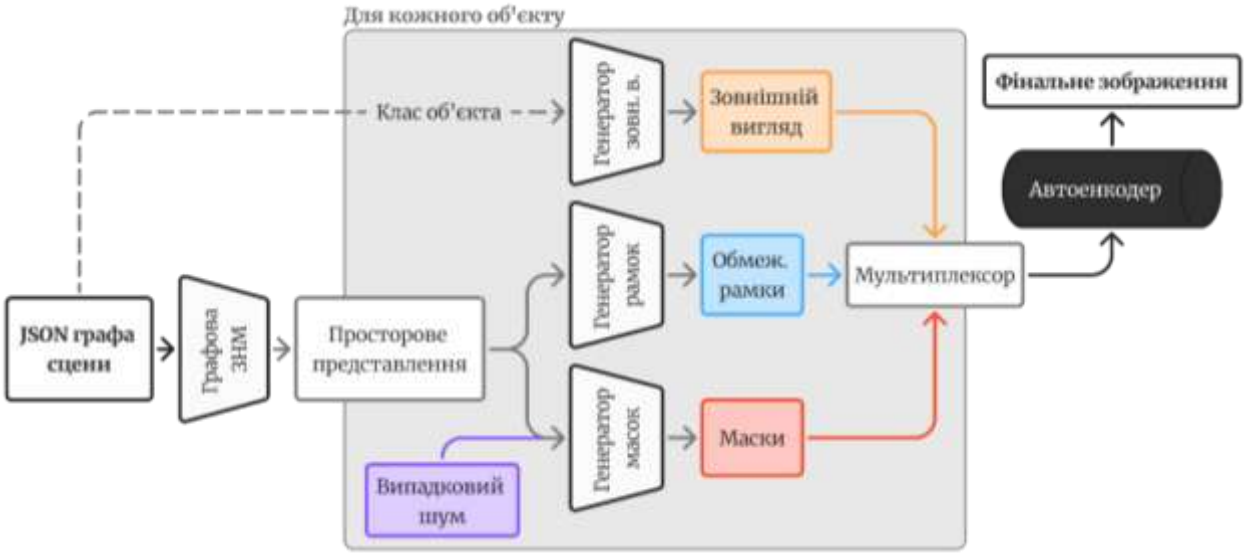

Рисунок 2.4 – Архітектура SG2IM

Перевагою SG2IM є ієрархічна обробка сцени: модель спочатку аналізує об'єкти та їх атрибути, потім враховує їх просторові відношення, і, нарешті, інтегрує глобальний контекст для створення узгодженої композиції. Механізми уваги точно відображають зв'язки між об'єктами, а каскадна генерація покращує деталізацію від грубих макетів до фінальних зображень. Конкурентне навчання в цьому контексті допомагає досягти

реалістичності зображень на всіх етапах обробки: створенні масок об’єктів, генерації зображень окремих об’єктів та оцінки зображення загалом.

Перетворення графів у зображення потребує використання кількох нейронних мереж і системи функцій втрат, яка враховує їхні ролі. Процес включає навчання п’яти мереж (без дискримінаторів) і два векторних представлення (для класів об’єктів та інформації про ребра графа). З огляду на це, оптимізаційна функція для SG2IM містить сім зважених доданків:

– втрата реконструкції (LRec) оцінює точність відновлення зображення за допомогою метрики L1;

– втрата для обмежувальних рамок (Lbox) покращує розташування об'єктів, використовуючи середнє квадратичне відхилення (MSE) між згенерованими і реальними рамками;

– перцептуальна втрата (Lperceptual) оцінює якість згенерованого зображення, враховуючи як високорівневі, так і низькорівневі особливості, подібно до людського сприйняття. Вона порівнює активації різних шарів мережі VGG між згенерованим і реальним зображенням. Обчислюється як середнє L1-значення;

– втрати для масок (LD-mask) і зображень (LD-image) допомагають реалістичніше відображати об’єкти та їх форми;

– втрата для об'єктів (LD-object) сприяє коректному відтворенню елементів сцени;

– втрати узгодженості ознак (LFM-mask, LFM-image) допомагають поліпшити узгодженість між згенерованими масками та зображеннями об’єктів. Основою їх роботи є порівняння активацій у різних шарах дискримінаторів на основі відстані L1.

В роботі SG2IM кожна з цих функцій сприяє балансу між точністю відтворення об'єктів, їх взаємним розташуванням і загальною візуальною якістю. При цьому для оцінки масок втрата не передбачена, адже їх генерація включає елемент стохастичності.

Змагальні мережі в SG2IM використовують три дискримінатори, що навчаються в змагальному режимі. Для кожного з них визначено власну функцію втрат залежно від завдання.

Дискримінатор масок (Dmask) оцінює реалістичність масок об'єктів, враховуючи їхній клас. Використовує функцію втрат найменших квадратів LS-GAN (Least Squares GAN). Головна ідея LS-GAN полягає в тому, що дискримінатор оцінює не бінарну ймовірність реальності зображення, а ступінь наближення згенерованих зразків до реальних. Такий підхід сприяє уникненню проблеми зникаючих градієнтів. Функція втрат дискримінатора LD-mask складається з двох частин: Lreal максимізує ймовірність правильної класифікації реальних масок, а Lfake мінімізує ймовірність того, що згенеровані маски будуть класифіковані як справжні.

Дискримінатор зображення (Dimage) оцінює загальну реалістичність зображення та його відповідність істинному макету сцени. Його функція втрат включає: Lreal, яка оцінює, наскільки добре дискримінатор відрізняє реальні зображення; Lfake-image, що змушує дискримінатор класифікувати згенеровані зображення як фейкові; Lfake-layout, що оцінює відповідність згенерованого макету реальному, та Lalt-appearance, що запобігає ситуаціям, коли зображення має правдоподібний макет, але неправильний вигляд. Щоб підвищити точність дискримінатора для об'єктів різної деталізації, LS-GAN застосовується як на повнорозмірних зображеннях, так і їх зменшених копіях.

Дискримінатор об'єктів (Dobject) аналізує окремі об'єкти у зображенні на основі їхніх вирізок, створених за допомогою обмежувальних рамок. Його втрата складається з Lreal, яка максимізує ймовірність правильної класифікації реальних об'єктів, та Lfake, що змушує дискримінатор зменшувати оцінку згенерованих об'єктів.

Для забезпечення коректної роботи цієї системи, навчання моделі SG2IM має проходити з акцентом на стабільність та ефективність роботи її

складових, і для різних масштабів вихідних зображень. З огляду на це, критичним є підбір ефективних гіперпараметрів.

Серед інших оптимізаторів для навчання мереж обрано Adam. Він поєднує переваги моменту (як у SGD) і адаптивного регулювання швидкості навчання (як у RMSprop), що покращує збіжність оптимізації. В рамках дослідження Adam використовується із параметром beta1=0.5. Ціною сповільнення навчання, таке значення допомагає стабілізувати градієнти під час оновлення ваг.

Навчання триває 1 мільйон ітерацій, що є достатнім для досягнення збіжності. Для всіх компонентів, окрім LD-mask, встановлено розмір кроку навчання на рівні 1e-4. Для LD-Mask використовувався менший крок 1e-5, що забезпечило додаткову стабілізацію роботи мережі масок, оскільки її завдання є порівняно складною задачею.

Також використано різні розміри батчів в залежності від роздільної здатності зображень. Така конфігурація покликана допомогти уникнути перевантаження пам’яті та гарантувати, що модель встигатиме обробляти численні деталі зображень. Кожне зображення при цьому може містити до 8 об'єктів.

В рамках дослідження модель SG2IM чудово справляється із передачею логіки та розташування сцени, проте на даному етапі їй бракує реалістичності синтезу окремих об’єктів. Це пов’язано із браком обчислювальної потужності для достатнього навчання генератора зовнішнього вигляду об’єктів. Незважаючи на ці обмеження, логіка сцени передається точно та узгоджено з вхідним сценографом, що є важливим досягненням у сфері генерації зображень.

3 ПРАКТИЧНА РЕАЛІЗАЦІЯ СИСТЕМИ ГЕНЕРАЦІЇ ІЛЮСТРАЦІЙ НА ОСНОВІ ГРАФІВ СЦЕН

3.1 Використані інструменти та середовище розробки

Реалізація системи генерації ілюстрацій на основі графів сцен потребує використання гнучких, продуктивних та масштабованих інструментів, здатних забезпечити обробку природної мови, побудову графових структур і генерацію зображень за допомогою методів глибокого навчання. Такі інструменти мають підтримувати роботу з великими обсягами даних, дозволяти швидку побудову прототипів моделей, а також бути сумісними з сучасними бібліотеками машинного навчання.

У цьому дослідженні для розробки системи було обрано мову програмування Python, яка є стандартом у сфері ШІ завдяки своїй простоті та багатому екосистемному середовищу. Основним фреймворком для реалізації нейронних мереж обрано PyTorch – один із провідних інструментів у сфері глибокого навчання. Його динамічний обчислювальний граф, зручний API, підтримка GPU та широкий вибір бібліотек (зокрема torchvision і transformers) роблять його ідеальним для мультимодальних задач, що поєднують зображення й текст. Крім того, велика кількість відкритих моделей, активна спільнота та простота розробки кастомних компонентів забезпечують гнучкість у проєктуванні архітектури системи.

Наступним важливим компонентом розробки стали засоби для перетворення текстових описів у структуровані графи сцен. У розробленій системі використовуються кілька потужних інструментів для обробки тексту та екстракції інформації, серед яких NLTK та Transformers. окрема, NLTK забезпечує лематизацію слів та базову обробку тексту, тоді як бібліотека Transformers використовується для семантичного аналізу,

зокрема за допомогою трансформерних моделей, таких як Flan-T5, для генерації тексту.

Для побудови графових структур, а також їхньої візуалізації у вигляді ілюстрацій, у системі застосовано низку допоміжних інструментів. Зокрема, бібліотека Graphviz забезпечує візуальне подання зв’язків між об’єктами сцени. Також активно використовуються модулі matplotlib.pyplot, imageio та matplotlib.patches.Rectangle для побудови, збереження та виведення графічних зображень, що ілюструють просторові взаємозв’язки, маски та обмежувальні рамки об’єктів у сцені. Модулі os та tempfile забезпечують керування тимчасовими файлами під час побудови і візуалізації результатів.

Таким чином, комбінація описаних інструментів в рамках дослідження дозволила реалізувати ефективну систему, здатну на основі текстових описів будувати графи сцен і генерувати відповідні ілюстрації.

3.2 Опис та підготовка набору даних

Для тренування та оцінювання системи було використано набір даних MS-COCO 2017 у поєднанні з його розширенням COCO-Stuff. MS-COCO – один із найпоширеніших датасетів для задач комп’ютерного зору, що містить приблизно 164 000 зображень, кожне з яких анотоване за допомогою:

– координат об'єктів (bounding boxes);

– масок сегментації;

– класів об’єктів (всього 80 основних категорій);

– п’яти текстових описів, створених вручну для кожного зображення, які передають основний зміст сцени.

Щоб забезпечити більш повне моделювання просторових взаємозв’язків, було використано COCO-Stuff – розширення, яке додає 91 категорію контекстуальних об’єктів (небо, трава, дорога, стіна тощо) до класів з основного COCO. Загалом, це формує 171 унікальний

клас (80 об’єктів + 91 фоновий елемент), що дозволяє моделі опрацьовувати не лише активні об’єкти, а й фон сцени – важливий аспект при генерації повноцінних ілюстрацій.

Для зменшення складності й покращення інтерпретованості моделі в рамках дослідження було проведено фільтрацію даних за допомогою інструменту з відкритим кодом FiftyOne:

– залишено лише зображення з менше ніж 8 об’єктами;

– враховано лише зразки з повними описами та анотаціями.

Ці структури перетворюються у вхід до моделі, яка генерує зображення, дотримуючись описаної у тексті сцени. Додатково побудовано індексований словник об’єктів (з 171 класу) разом з їхніми відносними розмірами, а також словник просторових відношень («on», «next to», «behind», «under», тощо), що враховуються при формуванні графів сцен модулем семантичного аналізу.

Окрім навчання генераторів, в рамках дослідження набір даних MS-COCO використовувався також для оцінки здатності моделі розуміти текст і перетворювати його у візуальний зміст. Зокрема, на вхід модулю семантичного аналізу подаються текстові описи з COCO, що відповідають таким критеріям:

– семантична повнота: наявність щонайменше одного об’єкта з дією або розташуванням;

– наявність відношень: опис має включати позицію або взаємозв’язок між об’єктами;

– однозначність: уникнення метафор та абстракцій, що не мають візуального втілення.

Приклади таких описів із датасету COCO: «a young person sitting on the floor with a book near a dog»; «a basket of apples and a box of bananas next to each other»; «brush, books, bag, and an umbrella laid out on the bed»; «a green, orange and white train in a train station».

На основі цих описів модуль семантичного аналізу виділяє об’єкти, атрибути й зв’язки між ними, формуючи граф сцени, який слугує входом до моделі генерації зображення.

3.3 Побудова графа сцени на основі текстового опису та генерація зображення

3.3.1 Реалізація семантичного аналізу тексту та формалізації графу

Для перетворення текстового опису у структуроване графове представлення, придатне для подальшої генерації зображень, в рамках дослідження реалізовано клас SceneGraphParser. Він об'єднує всі етапи: від попередньої обробки тексту до генерації та формалізації сценографа.

Система використовує потужні інструменти для обробки природної мови, а саме: NLTK та Transformers. Зокрема, NLTK забезпечує лематизацію слів за допомогою класу WordNetLemmatizer, що дозволяє зводити слова до базової форми (наприклад, «running» скорочується до «run») й забезпечує нормалізацію тексту. Крім того, текст приводиться до нижнього регістру, що усуває варіативність написання.

Після попередньої обробки трансформерна модель генерує текстове представлення сценографа, що містить список об'єктів, їхні атрибути та взаємозв’язки. На відміну від класичних методів, які покладаються на графові нейронні мережі або жорсткі правила парсингу, система використовує flan-t5-large-VG-factual-sg. Ця модель, імпортована через бібліотеку Transformers, спеціалізована на генерації графів сцени з описів завдяки попередньому тренуванню на датасетах Visual Genome та FACTUAL.

Це дає змогу: уникнути необхідності в жорсткій попередній обробці тексту; враховувати глобальний контекст опису; отримувати точні графи з

варіативних формулювань текстів; легко масштабувати систему на нові домени завдяки підтримці трансферного навчання.

Таким чином, клас SceneGraphParser реалізує повний цикл обробки тексту та побудови сцени:

– ініціалізація моделі, що включає завантаження токенізатора та трансформера, встановлення словника об’єктів, а також підготовку лематизованих версій їх назв;

– метод парсингу відповідає за основну генерацію графа. Він приймає текстові описи, проводить їх попередню обробку, формує запити до трансформера у форматі «Make Graph: <опис>", та виконує променевий пошук, генеруючи текстове подання графа;

– вбудована функція graph_string_to_object перетворює отриманий текст графа у словник з об’єктами, атрибутами та відношеннями.

Цей процес формалізації включає наступні етапи:

– ідентифікація об’єктів – кожен об’єкт, згаданий у текстовому графі, перевіряється на відповідність словнику. Якщо лема слова наявна в словнику, об’єкт додається до графа. Потім для кожного об’єкта створюється унікальний індекс, який використовується для ідентифікації об’єкта у графі, а також при встановленні його атрибутів і зв’язків;

– визначення атрибутів – спочатку кожному об’єкту присвоюється стандартне числове значення розміру згідно зі словником класів. За наявності в триплетах об’єкта атрибута розміру, це значення модифікується;

– витяг відношень – в текстовому графі взаємозв’язки між об’єктами визначаються з триплетів виду [суб'єкт, предикат, об'єкт]. Предикати очищуються від малозначущих слів (наприклад, «on the» змінюється на «on»), після чого нормалізуються і порівнюються з доступними у словнику просторовими відношеннями. Відповідні пари додаються до графа у вигляді [індекс_суб'єкта, відношення, індекс_об'єкта].

– генерація просторового розташування – окрім базової структури, система формує умовну просторову розкладку об’єктів на сітці розміром 5x5, що імітує сцену. Кожному об’єкту випадково призначається початкова позиція в межах центральної частини сітки. Потім позиції коригуються відповідно до змісту предикатів. Наприклад, якщо зазначено «apple above plate», то об’єкт «apple» розміщується над «plate».

Таким чином, модель не лише інтерпретує текст у вигляді графа, а й формує його семантичну просторову структуру, придатну для подальшого аналізу або візуалізації модулем генерації зображень. Наприклад, за допомогою бібліотеки GraphViz можна побудувати графічне зображення сценографа для контролю генерації на проміжному етапі.

3.3.2 Реалізація модуля генерації зображення

Фінальним кроком після побудови графа сцени є перетворення цього представлення на ілюстрацію. Для цього використовується модуль генерації зображення, що базується на глибоких нейронних мережах і включає низку взаємопов’язаних компонентів.

Генератор – основна частина моделі, яка відповідає за створення зображень зі сцени, описаної графом.

Дискримінатори. Для підвищення якості згенерованих зображень модуль включає кілька дискримінаторів, які виконують специфічні функції:

– дискримінатор зображень оцінює правдоподібність зображень, визначаючи їх реальність чи згенерованість;

– дискримінатор об'єктів фокусується на виявленні правильності об'єктів сцени на згенерованих зображеннях;

– дискримінатор масок оцінює правильність масок, що описують форму об'єктів на зображеннях.

Реалізовано класи AcDiscriminator (оцінка реалізму та класифікація об’єктів), AcCropDiscriminator (робота з фрагментами зображень), а також

MultiscaleMaskDiscriminator і NLayerMaskDiscriminator, які забезпечують багаторівневу перевірку зображень та масок на різних масштабах, підвищуючи точність оцінки.

Функції втрат. У навчанні моделі використовуються різноманітні функції втрат для генератора та дискримінаторів. Зокрема, бінарна крос-ентропія вимірює різницю між передбаченими та істинними мітками, дозволяючи моделі оцінювати, наскільки правдоподібно вона генерує зображення. Також застосовуються функції втрат для оцінки точності об'єктів і їхніх масок, а також для текстурної правдоподібності класів через VGG–функції. Останні використовують попередньо навчену модель VGG19, щоб порівняти згенеровані зображення об’єктів з реальними.

Першим етапом навчання модуля є ініціалізація його основних компонентів: генератора та кількох спеціалізованих дискримінаторів, що відповідають за обробку зображень, об'єктів і масок. Далі відбувається їх паралельне навчання з використанням оптимізатора Adam. Завдяки поєднанню цих компонентів із різноманітними функціями втрат, модуль здатен ефективно генерувати реалістичні зображення, які узгоджуються із заданими сценами.

В рамках дослідження генеративні моделі відіграють ключову роль у перетворенні графових представлень сцени в реалістичні зображення, базуючись на ЗНМ та механізмах обробки графових даних (рисунок 3.1).

Рисунок 3.1 – Пайплайн генерації зображень на основі сценографу

Для інтеграції генеративних мереж в рамках задачі генерації зображень з описів реалізовано основний клас ImageGenerator. Його функціонал включає приймання сценографа, прогнозування меж і масок

об'єктів, побудову карти розміщення та синтез фінального зображення. Генеративний процес в цьому контексті включає наступні етапи:

– кодування вхідного сценографа – об'єкти графа кодуються за допомогою вбудовувань об’єктів, а відносини між ними – через вбудовування предикатів. Нарешті, ГЗМ (об’єкт класу GraphTripleConvNet, який базується на GraphTripleConv) обробляє ці зв'язки для отримання компактних представлень;

– генерація меж об'єктів та масок – векторні представлення об'єктів подаються на багатошарову мережу генерації рамок (box_net), що прогнозує координати меж. Генератор масок (mask_net) використовує ці вектори разом із випадковим шумом, щоб створити бінарні карти масок для кожного об'єкта. Обидва модулі реалізовані за допомогою згорткових шарів, зокрема torch.nn.ConvTranspose2d для підвищення роздільної здатності;

– кодування зовнішнього вигляду об'єктів, де об’єкт класу AppearanceEncoder обробляє вирізані зображення об'єктів через згорткову нейромережу, витягуючи візуальні ознаки. Ці ознаки потім поєднуються з відповідними векторними представленнями, отриманими з графа. Кодувальник побудований на основі torchvision.models з використанням глобального усереднення (GlobalAvgPool);

– формування розміщення сцени – об'єкти, їхні маски та межі використовуються для створення карти розміщення;

– генерація фінального зображення – генеративна мережа (об’єкт класу GlobalGenerator) отримує карту розміщення та синтезує зображення сцени. Ця модель включає каскад залишкових блоків та шарів підвищення роздільної здатності, в поєднанні з torch.nn.BatchNorm2d для стабілізації процесу навчання.

Таким чином, модель використовує графову обробку для розпізнавання та розміщення об'єктів, а потім нейронні мережі для генерації текстур і фінального вигляду зображення. На технічному рівні, генератори виконують багаторівневе згорткове перетворення, починаючи зі зменшення

розмірності, обробки у залишкових блоках і подальшого відновлення вихідного розміру зображення. Використання залишкових блоків допомагає зберегти контекстну інформацію, а ГЗМ забезпечують якісну генерацію зображень.

3.4 Аналіз результатів генерації зображень

Метою експерименту в рамках дослідження є оцінка якості згенерованих на основі текстових описів зображень, та перевірка ефективності використання проміжного етапу вилучення сценографів. Дослідження зосереджене на порівнянні результатів роботи спеціалізованої моделі з іншими генеративними системами та визначенні основних характеристик згенерованих ілюстрацій.

Етапи експерименту передбачають складання тестового набору, який містить текстові описи, що служать основою для синтезу зображень. Далі для вилучення сценографів із текстів використовується розроблений модуль семантичного аналізу тексту та формалізації графа SceneGraphParser. Результати його роботи візуалізуються за допомогою інструменту GraphViz, що дає можливість наочно оцінити точність та повноту структури сценографів. На основі отриманого графа модуль ImageGenerator генерує фінальні зображення.

Кількісна оцінка якості згенерованих зображень проводиться за допомогою кількох метрик, що дозволяють перевірити різні аспекти результатів генерації. Зокрема, Inception Score (IS) оцінює як якість, так і різноманітність згенерованих зображень, використовуючи попередньо натреновану модель Inception-V3. Якщо розподіл міток класів об’єктів (рисунок 3.2) на зображенні має низьку ентропію, це свідчить про високу впізнаваність і точність. IS добре корелює з людським судженням, тому є корисною метрикою для оцінки якості.

Рисунок 3.2 – Приклад розпізнання об’єктів моделлю Inception-V3

Frechet Inception Distance (FID) вимірює схожість між згенерованими та реальними зображеннями, базуючись на активаціях передостаннього шару моделі Inception-V3. Менший FID вказує на високу якість згенерованих зображень.

Важливим аспектом є також оцінка відповідності згенерованого зображення текстовому опису методом зворотної генерації опису (англ. captioning). Якщо зображення є реалістичним і точно відображає вхідний текст, модель генерації підписів повинна відновити оригінальний опис. Для цього використовується модель, натренована на MS-COCO. Оцінка здійснюється за допомогою стандартних метрик порівняння тексту: BLEU (Bilingual Evaluation Understudy), METEOR (Metric for Evaluation of Translation with Explicit Ordering) і CIDEr (Consensus-based Image Description Evaluation). Вони дозволяють виміряти ступінь семантичного перекриття між початковим та згенерованим на основі зображення описами, тим самим опосередковано оцінюючи відповідність генерації вихідному тексту.

Одним із важливих аспектів експерименту була візуальна оцінка точності вилучених сценографів. Візуалізації, створені за допомогою інструменту GraphViz (рисунок 3.3), підтвердили високу точність і повноту

побудови графів, що засвідчило правильну ідентифікацію об’єктів і їхніх взаємодій модулем SceneGraphParser.

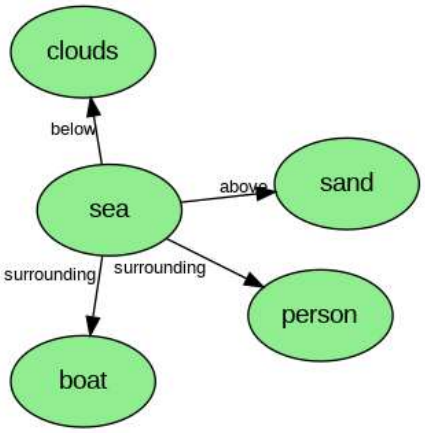

Рисунок 3.3 – Приклад візуалізації графа сцени за допомогою GraphViz

У рамках експерименту було здійснено генерацію зображень через сценографи, побудовані за описами з датасету MS-COCO (рисунок 3.4).

Рисунок 3.4 – Оригінальне зображення та згенерований за його описом відповідник

Згенеровані зображення демонструють високий рівень відповідності просторовій організації об’єктів, що вказує на коректну інтерпретацію

просторових відношень модулем ImageGenerator. Це особливо помітно у складних композиціях з декількома об'єктами та зв’язками (рисунок 3.5).

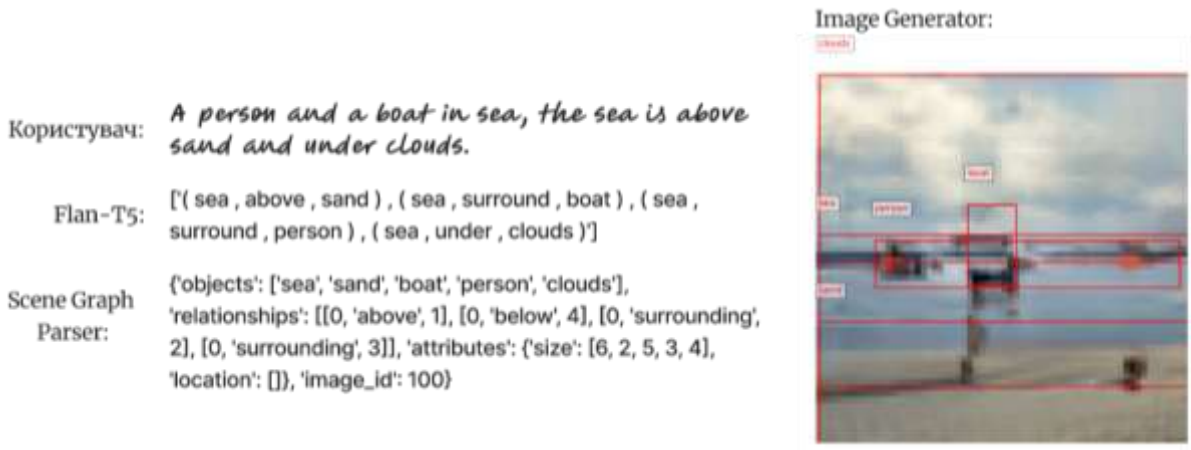

Рисунок 3.5 – Приклад покрокового синтезу зображення моделлю

Незважаючи на загальну коректність просторового розташування об’єктів, окремі згенеровані елементи мають сильно узагальнений або викривлений зовнішній вигляд, що може бути зумовлено недостатньою деталізацією сценографа чи обмеженнями кодувальника зовнішності. Водночас сегментаційні маски демонструють задовільну точність (рисунок 3.6).

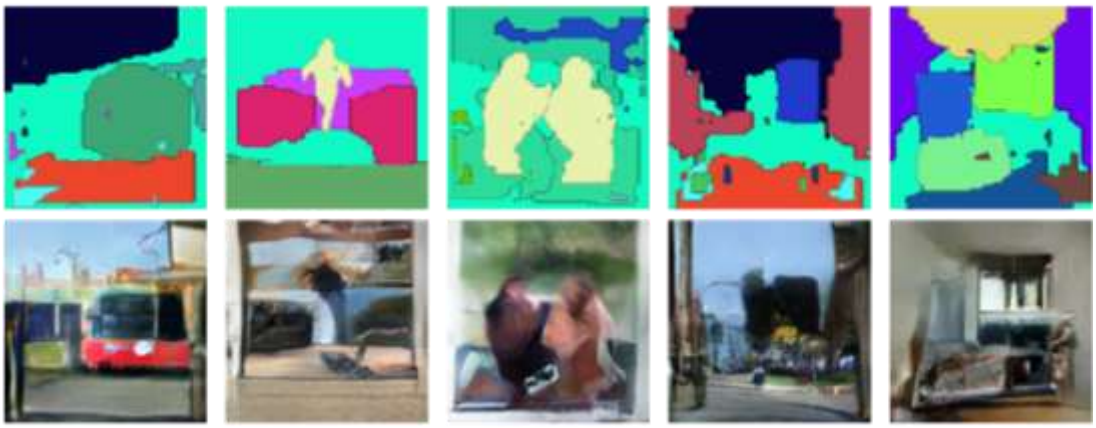

Рисунок 3.6 – Приклад масок об’єктів перед фінальною генерацією

Для кількісної оцінки якості згенерованих зображень були обчислені середні показники метрик IS та FID. IS склав 12.1, що свідчить про задовільний рівень різноманітності та впізнаваності об’єктів (рисунок 3.7). FID склав 63.6, що є характерним для моделей із структурованим проміжним представленням, але свідчить про необхідність подальшого покращення візуальної правдоподібності зображень.

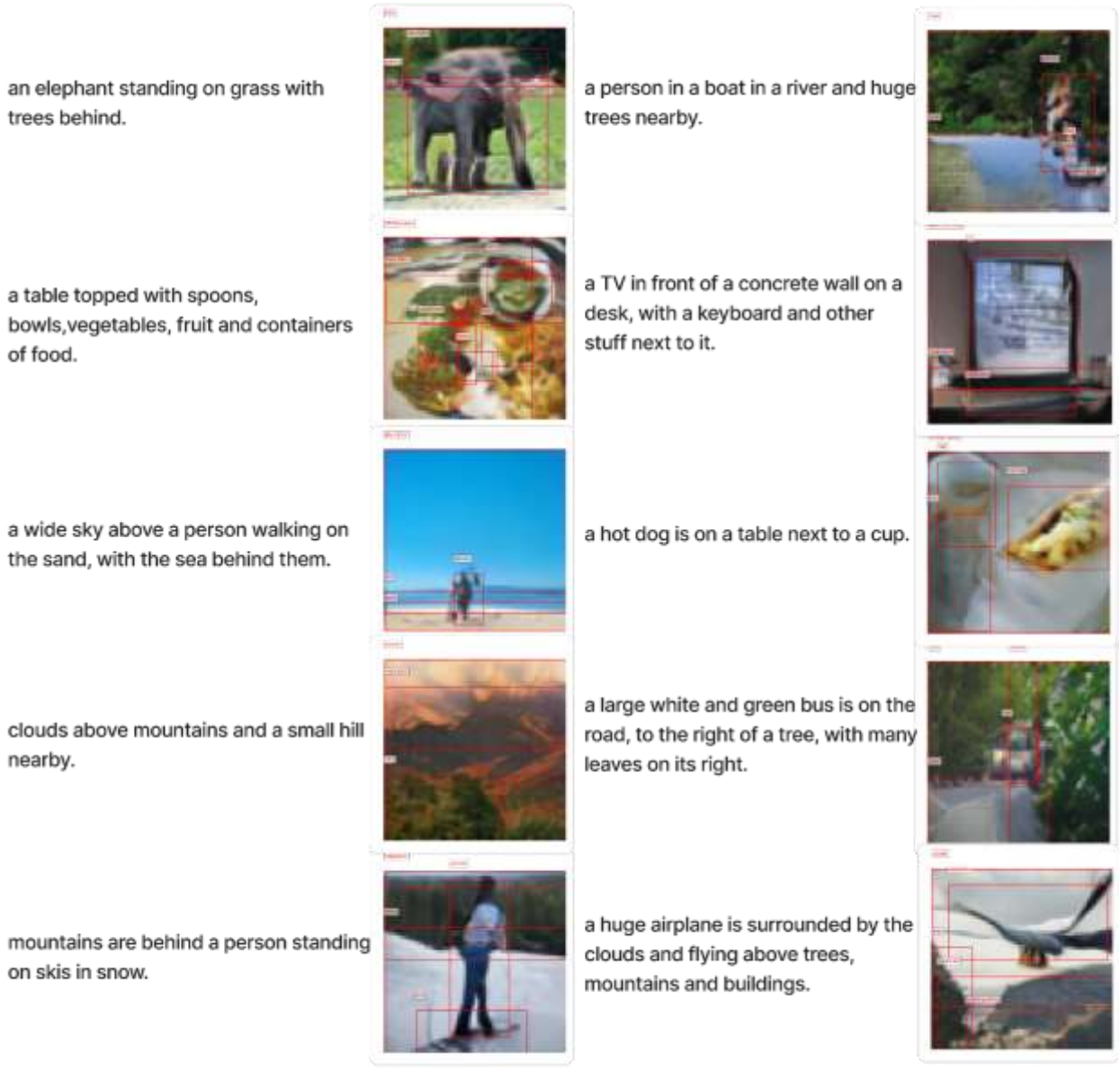

Рисунок 3.7 – Інші приклади синтезованих на основі тексту зображень

Середній час повного циклу генерації зображення склав 9.26 секунд.

Також було проведено тестування зворотної генерації підписів (рисунок 3.8). Отримані результати оцінки за метриками: BLEU-4: 0.142, METEOR: 0.165, CIDEr: 0.384. Ці значення підтверджують наявність змістовного перекриття між оригінальними та зворотно згенерованими описами, особливо щодо набору об'єктів і базових просторових відношень, однак залишається необхідність підвищення точності деталей опису.

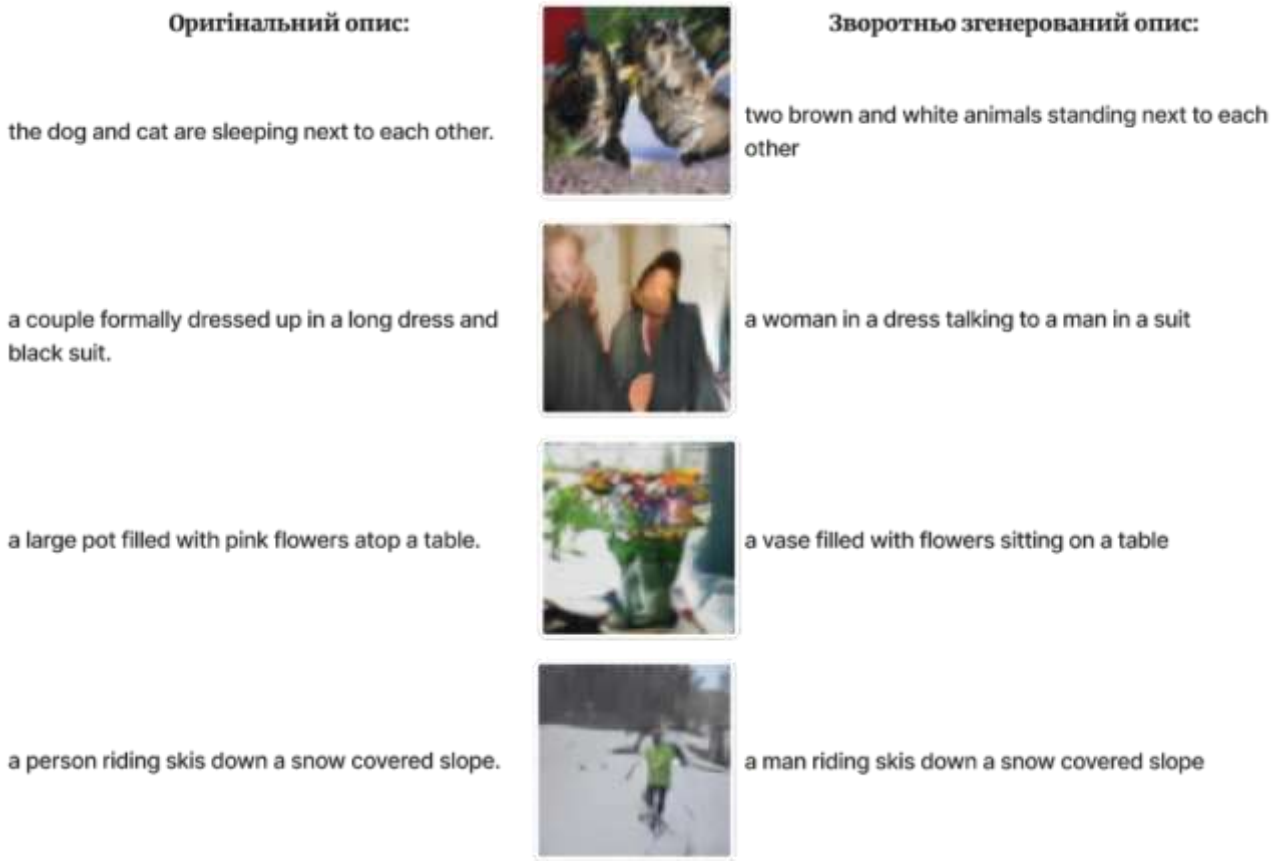

Рисунок 3.8 – Оригінальний та зворотньо згенерований описи до синтезованих розробленою моделлю зображень

Таким чином, у процесі аналізу були виявлені низка викликів, що впливають на якість генерації моделлю зображень:

– при екстракції сценографів не враховуються якісні та кількісні характеристики об’єктів (наприклад, «два», «червоний», «круглий»), за винятком розмірів. Це знижує контрольованість генерації зображень;

– фіксований словник просторових відносин не дозволяє точно передавати взаємодії об’єктів. Наприклад, дії або напрямки не враховуються, і рух об’єктів задається лише типовими позами, характерними для їх класу;

– на поточному етапі користувач не може змінювати орієнтацію, текстури, кольори чи інші візуальні характеристики об’єктів у фінальному зображенні.

З огляду на обмеження, подальший розвиток системи передбачає розширення функціональності через:

– адаптивну генерацію – врахування якісних та кількісних характеристик об’єктів при екстракції графа та побудова зображення на основі цього збагаченого графа;

– розширення словника відносин – замість фіксованого набору – використання гнучкої онтології просторових, логічних і контекстних зв’язків;

– інтерфейс інтерактивного редагування – надання користувачу можливості змінювати сценограф до або після генерації зображення – додавати або видаляти об’єкти, редагувати зв’язки, уточнювати атрибути, орієнтацію об’єктів тощо, при цьому зберігаючи логіку сцени.

Водночас система стикається з обмеженням у вигляді недостатньої реалістичності окремих згенерованих об'єктів. Підвищити візуальну якість зображень можна шляхом інтеграції більш потужного генератора на фінальному етапі, який забезпечуватиме деталізацію без втрати точності розміщення та контрольованості, гарантованих сценографом.

Таке розширення функціональності та оптимізація можуть суттєво підвищити відповідність зображення текстовому опису та відкрити нові можливості для інтерактивного застосування системи в освітньому, дизайнерському та дослідницькому середовищі.

3.5 Порівняльний аналіз

Для оцінки ефективності трирівневої системи перетворення тексту на зображення через проміжний сценограф було проведено порівняння з прямими методами текст-піксельного перетворення (AttnGAN та Stable Diffusion). Подібно до основного експерименту, обидва підходи були протестовані на наборі описів із датасету MS-COCO, а результати оцінювалися за метриками IS, FID та якістю відновлених описів**.**

Зазначимо, що в рамках дослідження не передбачено порівняння трирівневої системи з моделями, де використовується проміжне семантичне представлення сцени або генерація на основі графа. Оскільки ці моделі потребують додаткового етапу генерації структурованих представлень, це значно ускладнює порівняння за часом та іншими метриками. Крім того, це не відповідає меті експерименту, зосередженому на порівнянні здатності моделей генерувати зображення, що відповідає логіці опису.

Модель AttnGAN безпосередньо перетворює текстовий опис у зображення, фокусуючись на ключових словах і фразах за допомогою механізму уваги (рисунок 3.9).

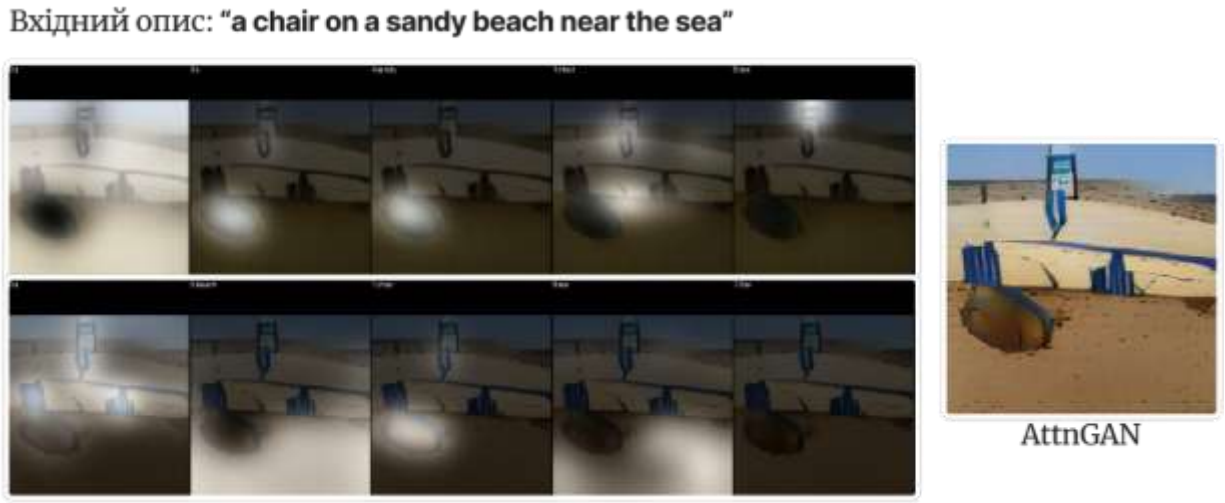

Рисунок 3.9 – Приклад використання механізму уваги моделлю AttnGAN для генерації зображення з тексту

У ході експерименту цей підхід продемонстрував здатність створювати візуально привабливі зображення з загалом впізнаваними кольорами та стилістичною відповідністю тексту. Водночас було виявлено низку обмежень: модель часто не відтворює точну форму об’єктів, має труднощі з коректним просторовим розташуванням елементів сцени та схильна до змістових суперечностей між частинами зображення.

Модель Stable Diffusion, як більш сучасний дифузійний підхід, забезпечує вищу якість зображень у плані художнього виконання. Вона часто генерує естетично привабливі та деталізовані сцени, проте схильна втрачати структуру сцени у випадках з великою кількістю об’єктів або складними взаємозв’язками між ними (рисунок 3.10).

Рисунок 3.10 – Втрата об’єктів (стіни та озера) в роботі Stable Diffusion

Обидві моделі прямого синтезу пікселів з тексту здатні враховувати атрибути об’єктів, якщо вони чітко вказані в описі. Натомість реалізована система із проміжним етапом екстракції сценографів демонструє кращу структурну узгодженість між текстом і фінальним зображенням із задовільним рівнем впізнаваності об’єктів (рисунок 3.11).

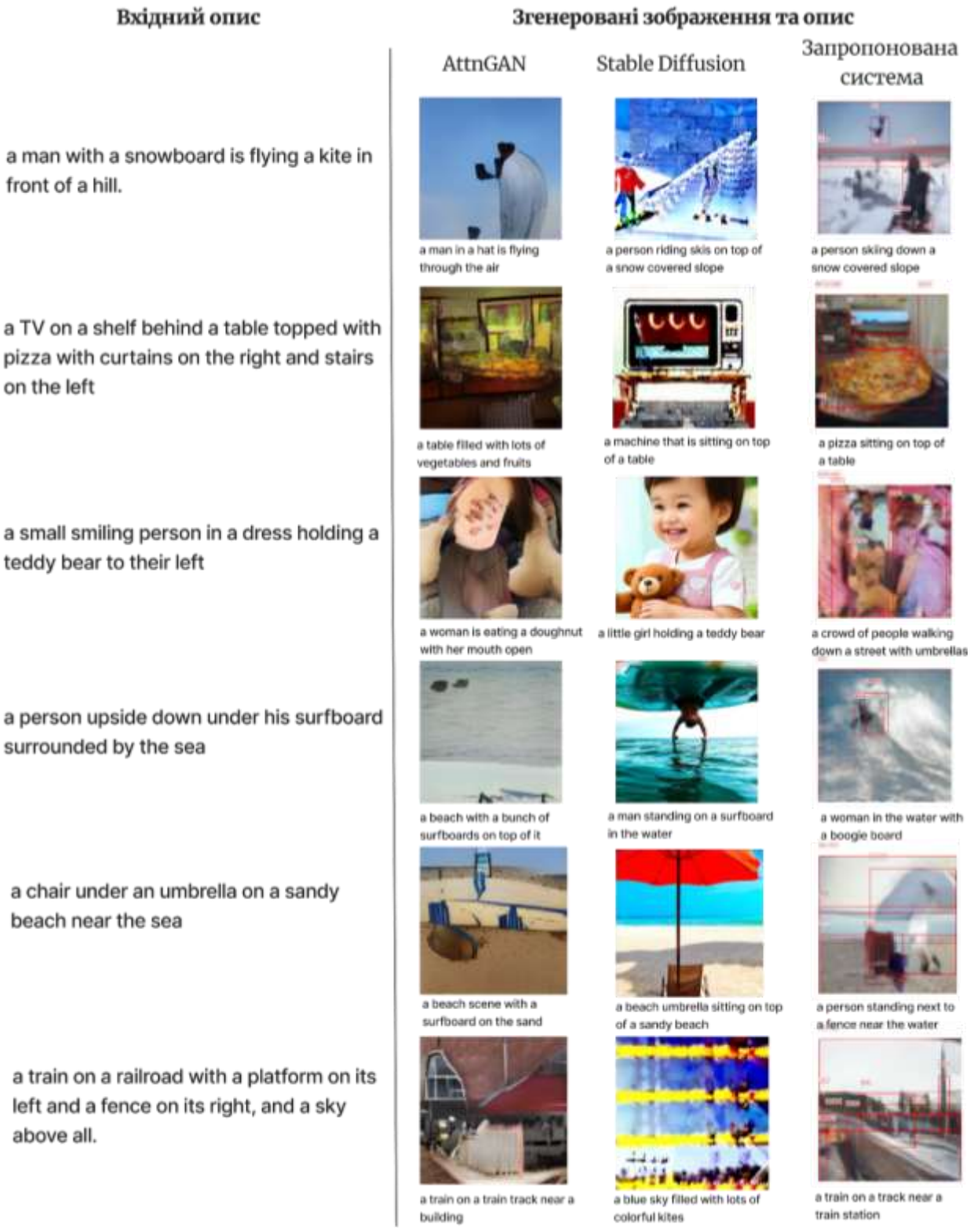

Рисунок 3.11 – Приклади синтезованих моделями зображень разом із зворотньо згенерованими описами до них

Оцінювання результатів за метриками в рамках експерименту дало наступні результати, відповідні до таблиці 3.1.

Таблиця 3.1 – Порівняльна таблиця результатів генерації зображень

Метрика / Модель	IS	FID	BLEU-4	METEOR	CIDEr	Середній час генерації (с)
AttnGAN	9.8	91.2	0.118	0.149	0.312	13.31
Stable Diffusion	24.5	12.7	0.153	0.176	0.421	5.12
Трирівневий підхід	12.1	63.6	0.142	0.165	0.384	9.26

Аналіз метрик показує, що Stable Diffusion перевершує трирівневий підхід за всіма метриками. Вона генерує більш різноманітні та візуально реалістичні зображення з кращими значеннями IS та FID. Водночас її перевага у текстовій відповідності є незначною, що свідчить про високий потенціал трирівневої архітектури, орієнтованої саме на точність інтерпретації змісту.

Модель AttnGAN поступається обом підходам як у візуальній точності, так і у відповідності опису, що робить її менш ефективною в контексті поставленої задачі.

Порівняння середнього часу генерації показує, що реалізований підхід займає більше часу (9.26 с), ніж робота Stable Diffusion (5.12 с), але значно швидший, ніж AttnGAN (13.31 с).

При цьому трирівнева технологія в рамках проведеного дослідження виявила ряд безумовних переваг:

– краще відображення просторової структури – графова модель дозволяє точно відтворювати позиції об’єктів та їх взаємозв’язки;

– підвищена контрольованість генерації – потенційно сцена може бути відредагована до етапу візуалізації, що забезпечить більшу гнучкість у налаштуванні результату;

– семантична прозорість – проміжний граф дає змогу візуально перевірити, як текстовий опис був інтерпретований системою;

– можливість модульної оптимізації – кожен етап (семантичний аналіз, побудова графа, генерація) можна вдосконалювати незалежно.

Крім того, використання більш потужного візуального генератора на останньому етапі перетворення сценографу на зображення може істотно покращити візуальну якість об'єктів, зберігаючи переваги точності та контрольованості сценографу.

Підсумовуючи, трирівнева система демонструє вищу відповідність текстовому опису в аспектах структури, змісту та логіки сцени, тоді як прямі прямі текст-піксельні моделі можуть запропонувати більш привабливі зображення з відображеними якісними атрибутами, проте ціною втрати деяких елементів сцени, просторової узгодженості та інтерпретованості процесу. Перспективним напрямом розвитку розробленої системи є модульна оптимізація окремих компонентів системи та розширення функціоналу.

ВИСНОВКИ

Дана кваліфікаційна робота присвячена технології автоматизованої генерації зображень на основі текстових описів із використанням сценографів як проміжного представлення. Метою роботи було розробити та дослідити систему, здатну генерувати ілюстрації, які забезпечують семантичну та просторову відповідність текстовим описам.

Для досягнення цієї мети було розглянуто трирівневу технологію «текст-сценограф-зображення». У першому етапі задіяний модуль SceneGraphParser, який відповідає за обробку тексту трансформерною моделлю для вилучення текстового графа сцени, а також подальшу формалізацію сценографу. На другому кроці сценограф передається до модуля ImageGenerator, який реалізує каскадну архітектуру BAK-GAN з використанням кількох спеціалізованих ЗНМ та механізмів обробки графових даних. В рамках цього етапу відбувається кодування графу, прогнозування меж і масок, формування карти розміщення, кодування зовнішнього вигляду об'єктів та генерація фінального зображення.

Реалізована система продемонструвала високу семантичну точність та структурну узгодженість між текстовим описом і зображенням. Кількісні метрики (Inception Score: 12.1, FID: 63.6, BLEU-4: 0.142) свідчать про відповідність поставленим вимогам, хоча й визначають напрями для подальшого вдосконалення, зокрема покращення візуальної реалістичності окремих об'єктів.

Порівняно зі світовими аналогами, такими як AttnGAN та Stable Diffusion, розроблений підхід виявився більш ефективним у відтворенні просторової структури сцени, проте поступився Stable Diffusion у візуальній якості. На відміну від зазначених методів, система забезпечує контрольованість та інтерпретованість процесу генерації завдяки використанню сценографів, що є її ключовою перевагою.

Методологія цього дослідження включала аналіз релевантних наукових джерел, систематизацію підходів генерації зображень, а також повний цикл розробки системи: підготовку даних, реалізацію модулів, навчання моделей, тестування та порівняння зі світовими аналогами. Робота тісно пов’язана з науково–дослідними напрямами кафедри ШІ, зокрема в галузях ОПМ, комп’ютерного зору та генеративних моделей.

Особливістю цього дослідження є його наукова новизна, оскільки вперше інтегрується використання трансформерів для вилучення з тексту сценографів з подальшою генерацією ілюстрацій на їх основі. Матеріали роботи можуть бути використані в навчальних курсах університету (наприклад, з ШІ, комп’ютерного зору та обробки даних), а також у подальших наукових розробках – зокрема для вдосконалення мультимодальних систем та інтерактивних інструментів візуалізації. Крім того, отримані результати мають практичне значення для сфер освіти, дизайну та медіа.

Для подальшого розвитку системи рекомендується розширити словник атрибутів і відношень у графах, а також вдосконалити генератор зображень для підвищення їх візуальної якості. Іншим важливим напрямом є створення повноцінного користувацького інтерфейсу, який забезпечить можливість інтерактивного редагування проміжних сценографів. Реалізація цих удосконалень сприятиме розширенню функціональних можливостей системи та підвищенню її ефективності в реальних застосуваннях.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Johnson J., Gupta A., Fei-Fei L. Image generation from scene graphs. *2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Salt Lake City, UT, 18–23 June 2018. 2018. URL: https://doi.org/10.1109/cvpr.2018.00133 (date of access: 06.05.2025).

2. A comprehensive survey of scene graphs: generation and application / X. Chang et al. *IEEE transactions on pattern analysis and machine intelligence*. 2022. P. 21. URL: https://doi.org/10.1109/tpami.2021.3137605 (date of access: 06.05.2025).

3. Generating semantically precise scene graphs from textual descriptions for improved image retrieval / S. Schuster et al. *Proceedings of the fourth workshop on vision and language*, Lisbon, Portugal. Stroudsburg, PA, USA, 2015. URL: https://doi.org/10.18653/v1/w15–2812 (date of access: 06.05.2025).

4. De Marneffe M.-C., Manning C. D. The Stanford typed dependencies representation. *Coling 2008: the workshop*, Manchester, United Kingdom, 23 August 2008. Morristown, NJ, USA, 2008. URL: https://doi.org/10.3115/1608858.1608859 (date of access: 06.05.2025).

5. Devlin J., Chang M.W., Lee K., and Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. *NAACL-HLT*, 2019. URL: https://doi.org/10.48550/arXiv.1810.04805 (date of access: 06.05.2025).

6. Hangbo B., Li D., Songhao P., and Furu W. Beit: Bert pre–training of image transformers. *International conference on learning representations*, 2021. URL: https://doi.org/10.48550/arXiv.2106.08254 (date of access: 06.05.2025).

7. Semantic image manipulation using scene graphs / H. Dhamo et al. *2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Seattle, WA, USA, 13–19 June 2020. 2020.

URL: https://doi.org/10.1109/cvpr42600.2020.00526 (date of access: 06.05.2025).

8. Ramesh A., Pavlov M., Goh G., Gray S., Voss C., Radford A., Chen M., Sutskever I. Zero-Shot Text-to-Image Generation. *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*. 2021. Vol. 139. P. 8821–8831. URL: https://proceedings.mlr.press/v139/ramesh21a.html (date of access: 06.05.2025).

9. Caesar H., Uijlings J., Ferrari V. COCO-Stuff: thing and stuff classes in context. *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.2018.00132 (date of access: 06.05.2025).

10. Inferring semantic layout for hierarchical text-to-image synthesis / S. Hong et al. *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.2018.00833 (date of access: 06.05.2025).

11. Image generation from layout / B. Zhao et al. *2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, Long Beach, CA, USA, 15–20 June 2019. 2019. URL: https://doi.org/10.1109/cvpr.2019.00878 (date of access: 06.05.2025).

12. Wu Y., Wei P., Lin L. Scene graph to image synthesis via knowledge consensus. *Proceedings of the AAAI conference on artificial intelligence*. 2023. Vol. 37, no. 3. P. 2856–2865. URL: https://doi.org/10.1609/aaai.v37i3.25387 (date of access: 06.05.2025).

13. Scene graph-grounded image generation / F. Wang et al. *Proceedings of the AAAI conference on artificial intelligence*. 2025. Vol. 39, no. 7. P. 7646–7654. URL: https://doi.org/10.1609/aaai.v39i7.32823 (date of access: 06.05.2025).

14. Glide: Towards photorealistic image generation and editing with textguided diffusion models / Nichol A. et al. 2022. URL: https://doi.org/10.48550/arXiv.2112.10741 (date of access: 06.05.2025).

15. StackGAN: text to photo-realistic image synthesis with stacked generative adversarial networks / H. Zhang et al. *2017 IEEE international conference on computer vision (ICCV)*, Venice, 22–29 October 2017. 2017. URL: https://doi.org/10.1109/iccv.2017.629 (date of access: 06.05.2025).

16. AttnGAN: fine-grained text to image generation with attentional generative adversarial networks / T. Xu et al. *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.2018.00143 (date of access: 06.05.2025).

17. Goodfellow I., Pouget-Abadie J., Mirza M. et al. Generative adversarial nets. *NIPS*, 2014. URL: https://doi.org/10.48550/arXiv.1406.2661 (date of access: 06.05.2025).

18. Vaswani A., Shazeer N., Parmar N. et al. Attention is all you need. *Adv Neural Inf Process Syst*, 2017. URL: https://doi.org/10.48550/arXiv.1706.03762 (date of access: 06.05.2025).

19. Ashual O., Wolf L. Specifying object attributes and relations in interactive scene generation. *2019 IEEE/CVF international conference on computer vision (ICCV)*, Seoul, Korea (South), 27 October – 2 November 2019. 2019. URL: https://doi.org/10.1109/iccv.2019.00466 (date of access: 06.05.2025).

20. Transformer–based image generation from scene graphs / R. Sortino et al. *Computer vision and image understanding*. 2023. P. 103721. URL: https://doi.org/10.1016/j.cviu.2023.103721 (date of access: 06.05.2025).

21. Kipf T. How powerful are graph convolutional networks?. *Thomas Kipf | Staff Research Scientist @ Google DeepMind*. URL: https://tkipf.github.io/graph–convolutional–networks/ (date of access: 06.05.2025).

22. Кулішова Н. Є., Столяров І., Цикало С. Процес прийняття рішень при дизайні ілюстрацій настільних ігор з використанням додатків генеративного штучного інтелекту. *Технологія і техніка друкарства*. 2024.

№ 1(83). С. 26–38. URL: https://doi.org/10.20535/2077–7264.1(83).2024.299490 (дата звернення: 07.05.2025).

23. Kulishova N., Sajek D. Using machine learning and generative intelligence in book cover development. *Journal of imaging*. 2025. Vol. 11, no. 2. P. 46. URL: https://doi.org/10.3390/jimaging11020046 (date of access: 07.05.2025).

24. Storytelling from an image stream using scene graphs / R. Wang et al. *Proceedings of the AAAI conference on artificial intelligence*. 2020. Vol. 34, no. 05. P. 9185–9192. URL: https://doi.org/10.1609/aaai.v34i05.6455 (date of access: 06.05.2025).

25. Chung H. W., Hou L., Longpre S. et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research.* 2024. Т. 25. № 70. С. 1–53. URL: https://doi.org/10.48550/arXiv.2210.11416 (date of access: 06.05.2025).

26. Banarescu, L., Bonial, C., Cai, S. et al. Abstract Meaning Representation for Sembanking. *Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse,* Sofia, Bulgaria, August 8–9, 2013. С. 178–186. URL: https://arxiv.org/pdf/1607.08822 (date of access: 13.04.2025).

27. SPICE: semantic propositional image caption evaluation / P. Anderson et al. *Computer vision – ECCV 2016*. Cham, 2016. P. 382–398. URL: https://doi.org/10.1007/978–3–319–46454–1_24 (date of access: 07.05.2025).

28. Raffel C., Shazeer N., Roberts A. et al. Exploring the limits of transfer learning with a unified text–to–text transformer. *Journal of Machine Learning Research.* 2020. Т. 21. № 140. С. 1–67. URL: https://doi.org/10.48550/arXiv.1910.10683 (date of access: 07.05.2025).

29. Improving scene graph classification by exploiting knowledge from texts / S. Sharifzadeh et al. *Proceedings of the AAAI conference on artificial intelligence*. 2022. Vol. 36, no. 2. P. 2189–2197. URL: https://doi.org/10.1609/aaai.v36i2.20116 (date of access: 07.05.2025).

30. FACTUAL: a benchmark for faithful and consistent textual scene graph parsing / Z. Li et al. *Findings of the association for computational linguistics: ACL 2023*, Toronto, Canada. Stroudsburg, PA, USA, 2023. URL: https://doi.org/10.18653/v1/2023.findings–acl.398 (date of access: 07.05.2025).

31. Etaiwi W., Awajan A. SemanticGraph2Vec: semantic graph embedding for text representation. *Array*. 2023. Vol. 17. P. 100276. URL: https://doi.org/10.1016/j.array.2023.100276 (date of access: 06.05.2025).

32. Generative artificial intelligence and its applications in materials science: current situation and future perspectives / Y. Liu et al. *Journal of materiomics*. 2023. URL: https://doi.org/10.1016/j.jmat.2023.05.001 (date of access: 07.05.2025).

33. Klawonn M., Heim E. Generating triples with adversarial networks for scene graph construction. *Proceedings of the AAAI conference on artificial intelligence*. 2018. Vol. 32, no. 1. URL: https://doi.org/10.1609/aaai.v32i1.12321 (date of access: 07.05.2025).

34. Larsen A. B. L., Sønderby S. K., Larochelle H., Winther O. Autoencoding beyond pixels using a learned similarity metric. *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. 2016. P. 1558–1566.

35. Kipf T.N., Welling M. Semi–Supervised Classification with Graph Convolutional Networks. *Proceedings of the 5th International Conference on Learning Representations (ICLR 2017)*. 2017. 14 p. URL: https://doi.org/10.48550/arXiv.1609.02907 (date of access: 07.05.2025).