Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

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

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Дослідження моделей комп'ютерного зору для вирішення задачі розпізнавання об'єктів та тексту
(тема)

Виконав:
здобувач другого року навчання,
групи ДСм-24-1

Омельченко М.Д.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)
Освітня програма Науки про дані (Data Science)
(повна назва спеціалізації)
Керівник доц. Дейнеко А.О.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ Лариса ЧАЛА
(підпис) (прізвище, ініціали)

2025 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

Рівень вищої освіти другий (магістерський)

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

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

Освітня програма Науки про дані (Data Science)
(повна назва)

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

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

здобувачеві Омельченку Миколі Дмитровичу
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження моделей комп'ютерного зору для вирішення задачі розпізнавання об'єктів та тексту

затверджена наказом університету від 24 листопада 20 25 р. № 1057Ст
2. Термін подання студентом роботи до екзаменаційної комісії 18 грудня 20 25 р.
3. Вихідні дані до роботи Наукові публікації та датасети для детекції/OCR, публікація Qwen-VL, інтернет-ресурси (репозиторій Qwen3-VL і документація Hugging Face Inference API/InferenceClient)

4. Перелік питань, що потрібно опрацювати в роботі
1) Аналіз предметної галузі та постановка задачі
2) Практична реалізація дослідження
3) Аналіз та оцінювання результатів експериментів

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	24.11.2025	виконано
2	Аналіз предметної галузі	25.11.2025	виконано
3	Постановка задачі	30.11.2025	виконано
4	Вибір моделі для дослідження	01.12.2025	виконано
5	Планування та підготовка експериментів	02.12.2025	виконано
6	Проведення експериментів	04.12.2025	виконано
7	Аналіз отриманих даних	08.12.2025	виконано
8	Підготовка пояснювальної записки	09.12.2025	виконано
9	Надання записки на перевірку керівнику	12.12.2025	виконано
10	Захист кваліфікаційної роботи	18.12.2025	виконано

Дата видачі завдання 24 листопада 20 25 р.

Здобувач _____
(підпис)

Керівник роботи _____ доц. Дейнеко А.О.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 61 с., 9 рис., 2 табл., 2 дод., 26 джерел.

КОМП’ЮТЕРНИЙ ЗІР, МУЛЬТИМОДАЛЬНІ МОДЕЛІ, ОПТИЧНЕ РОЗПІЗНАВАННЯ ТЕКСТУ, РОЗПІЗНАВАННЯ ОБ’ЄКТІВ, ТРАНСФОРМЕРИ, VISION-LANGUAGE МОДЕЛІ.

Об’єкт дослідження – процес аналізу зображень із використанням сучасних мультимодальних моделей комп’ютерного зору та обробки природної мови.

Предмет дослідження – методи та моделі розпізнавання об’єктів і тексту на зображеннях у межах vision-language підходу.

Мета роботи – теоретичне та експериментальне дослідження можливостей мультимодальних vision-language моделей для вирішення задач розпізнавання об’єктів, оптичного розпізнавання тексту та комбінованого аналізу сцен.

Методи дослідження – аналіз і узагальнення наукових публікацій, теоретичне моделювання, експериментальне дослідження роботи мультимодальної моделі у режимі inference, порівняльний аналіз результатів.

У роботі досліджено можливості застосування мультимодальних моделей для вирішення задач аналізу зображень без спеціалізованого донавчання. Показано, що формулювання завдання у вигляді текстової інструкції дозволяє отримувати структуровані результати у вигляді описів об’єктів, координат обмежувальних прямокутників та розпізнаного тексту. Отримані результати підтверджують перспективність мультимодального підходу для побудови універсальних систем комп’ютерного зору.

ABSTRACT

Master’s thesis contains: 61 pp., 9 fig., 2 tabl., 2 ann., 26 references.

COMPUTER VISION, MULTIMODAL MODELS, OBJECT DETECTION, OPTICAL CHARACTER RECOGNITION, TRANSFORMERS, VISION–LANGUAGE MODELS.

Object of research – the process of image analysis using modern multimodal computer vision and natural language processing models.

Subject of research – methods and models for object detection and text recognition in images within the vision-language paradigm.

Purpose – theoretical and experimental study of the capabilities of multimodal vision-language models for object detection, optical character recognition, and combined scene understanding.

Research methods – analysis and synthesis of scientific publications, theoretical modeling, experimental evaluation of multimodal models in inference mode, and comparative analysis of results.

The study investigates the application of multimodal models to image analysis tasks without task-specific fine-tuning. It is shown that instruction-based interaction allows the generation of structured outputs including object descriptions, bounding box coordinates, and recognized text. The obtained results demonstrate the potential of multimodal approaches for building flexible and universal computer vision systems.

ЗМІСТ

Перелік умовних позначень, символів, одиниць та скорочень 7
Вступ 8
1 Аналіз предметної галузі та постановка задачі 9
1.1 Базові поняття комп'ютерного зору та постановка задач розпізнавання 9
1.2 Класичні CNN-підходи до детекції об'єктів 11
1.3 Теорія розпізнавання тексту в зображеннях (OCR) 15
1.4 Трансформерні та Vision-Language підходи в комп'ютерному зорі .. 17
1.5 Архітектурні принципи сучасних Vision-Language моделей 20
1.6 Постановка задачі та мета дослідження 22
2 Практична реалізація дослідження 24
2.1 Загальна постановка практичної задачі 24
2.2 Опис використаної моделі та сервісної інфраструктури 25
2.3 Вхідні дані та їх організація 27
2.4 Формат запитів та вихідних даних 30
2.5 Алгоритм проведення експериментального дослідження 33
3 Аналіз та оцінювання результатів експерименту 37
3.1 Методика оцінювання результатів експерименту 37
3.2 Аналіз та інтерпретація експериментальних результатів 38
3.3 Перспективи розвитку 49
Висновки 51
Перелік джерел посилання 53
Додаток А Вхідні дані експериментів з використанням мультимодальної моделі Qwen-VL для задач детекції об'єктів, оптичного розпізнавання тексту та їх комбінованого аналізу 56
Додаток Б Відомість кваліфікаційної роботи 61

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ ТА СКОРОЧЕНЬ

BBox – Bounding Box – обмежувальна рамка;

CNN – Convolutional Neural Network – згорткова нейронна мережа;

IoU – Intersection over Union – міра перекриття;

JSON – JavaScript Object Notation – формат обміну даними;

LLM – Large Language Model – велика мовна модель;

OCR – Optical Character Recognition – оптичне розпізнавання символів;

VL – Vision-Language – візуально-мовна модель.

ВСТУП

Стрімкий розвиток методів штучного інтелекту впродовж останнього десятиліття суттєво вплинув на галузь комп'ютерного зору. Якщо раніше задачі аналізу зображень розв'язувалися переважно за допомогою спеціалізованих алгоритмів і ручного проєктування ознак, то сучасні системи базуються на глибоких нейронних мережах, здатних автоматично формувати ієрархічні представлення даних.

Особливу роль у цьому процесі відіграють задачі розпізнавання об'єктів та оптичного розпізнавання тексту. Традиційні підходи до вирішення цих задач складаються з окремих етапів обробки, що ускладнює масштабування та знижує гнучкість систем.

Поява трансформерних архітектур та великих мовних моделей призвела до формування нового класу мультимодальних vision-language моделей, які поєднують обробку зорової та текстової інформації в єдиному обчислювальному просторі. Такі моделі дозволяють формулювати задачі комп'ютерного зору у вигляді інструкцій природною мовою та отримувати структуровані результати без необхідності спеціалізованого донавчання.

Актуальність даної роботи зумовлена необхідністю системного дослідження можливостей мультимодальних моделей для розв'язання комплексних задач аналізу зображень. Особливо важливим є вивчення їх поведінки у режимі zero-shot inference, коли модель застосовується без адаптації до конкретної предметної галузі.

Метою магістерської роботи є теоретичне та експериментальне дослідження сучасних мультимодальних vision-language моделей для задач розпізнавання об'єктів і тексту на зображеннях. Для досягнення поставленої мети у роботі розглянуто теоретичні засади комп'ютерного зору, проаналізовано сучасні архітектури мультимодальних моделей та проведено експериментальне дослідження їх можливостей.

1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Базові поняття комп’ютерного зору та постановка задач розпізнавання

Комп’ютерний зір є одним із фундаментальних напрямів сучасних досліджень у галузі штучного інтелекту, метою якого є розроблення методів автоматичного аналізу та інтерпретації візуальної інформації. На відміну від класичної обробки сигналів, комп’ютерний зір орієнтований на формування семантичних представлень зображень, що дозволяє системам не лише фіксувати наявність певних структур, а й розуміти їх зміст у контексті сцени.

Цифрове зображення у сучасних системах комп’ютерного зору зазвичай представляється у вигляді дискретного багатовимірного масиву числових значень. Для кольорових зображень стандартним є подання у вигляді тензора

$$I \in R^{H \times W \times 3}, \tag{1.1}$$

де H та W відповідають висоті та ширині зображення, а третій вимір відображає інтенсивності колірних каналів. Таке представлення є універсальним та дозволяє застосовувати математичні операції для виділення просторових і спектральних ознак.

У межах комп’ютерного зору сформувалося декілька базових типів задач, кожна з яких має власну постановку та рівень складності. Найпростішою з них є класифікація зображень, у якій необхідно визначити, до якого з наперед заданих класів належить зображення в цілому. Формально задача класифікації полягає у знаходженні відображення між простором зображень та простором міток класів. Попри свою обмеженість,

класифікація стала фундаментом для розвитку більш складних методів аналізу сцен.

Локалізація об’єктів є наступним кроком у розвитку задач комп’ютерного зору. Вона передбачає визначення положення об’єкта на зображенні, зазвичай у вигляді обмежувального прямокутника. Такий підхід дозволяє поєднати семантичну інформацію з просторовими характеристиками, що є важливим для практичних застосувань. Детекція об’єктів розширює цю ідею на випадок, коли на зображенні присутні декілька об’єктів різних класів, кожен з яких необхідно виявити та описати окремо.

Сегментація зображень забезпечує ще більш детальний просторовий опис сцени. У цьому випадку кожному пікселю зображення ставиться у відповідність певна мітка, що дозволяє точно окреслити межі об’єктів. Сегментаційні методи широко застосовуються у медичній діагностиці, системах дистанційного зондування Землі та робототехніці. Проте висока точність сегментації вимагає значних обчислювальних ресурсів та якісно розмічених даних.

Особливу категорію задач комп’ютерного зору становить оптичне розпізнавання тексту на зображеннях. OCR поєднує у собі просторову локалізацію текстових областей та послідовне розпізнавання символів або слів. На відміну від класичних задач детекції, OCR орієнтоване на відновлення впорядкованих текстових послідовностей, що потребує залучення моделей, здатних працювати з часовими або порядковими залежностями.

Важливим теоретичним аспектом є чітке розмежування понять виявлення та розпізнавання. Виявлення відповідає на питання про наявність об’єкта та його розташування, тоді як розпізнавання спрямоване на ідентифікацію семантичної сутності цього об’єкта. У складних системах ці процеси можуть реалізовуватися спільно, однак їх аналітичне розділення

дозволяє точніше формулювати критерії якості та оцінювати ефективність алгоритмів.

Реальні сцени характеризуються високою варіативністю, що значно ускладнює автоматичний аналіз. Зміни освітлення, перспективні спотворення, часткові перекриття об’єктів, наявність шуму та складного фону створюють додаткові труднощі для алгоритмів комп’ютерного зору. У задачах OCR ці фактори доповнюються різноманіттям мов, шрифтів та орієнтацій тексту.

Розвиток методів глибокого навчання став переломним моментом у галузі комп’ютерного зору. Згорткові нейронні мережі дозволили автоматично навчатися виділенню ієрархічних ознак без ручного проєктування. Проте їх обмеження у моделюванні глобального контексту стимулювали появу нових архітектурних підходів, орієнтованих на довготривалі залежності між елементами зображення.

Подальший розвиток комп’ютерного зору тісно пов’язаний з інтеграцією методів обробки природної мови. Мультимодальні підходи дозволяють розглядати зображення та текст як взаємодоповнювальні джерела інформації, що відкриває можливості для побудови універсальних моделей аналізу сцен.

Отже, базові поняття комп’ютерного зору охоплюють широкий спектр задач та методів, спрямованих на формування семантичного розуміння візуальних даних. Їх систематизація та теоретичне осмислення створюють основу для подальшого розгляду моделей та алгоритмів, які реалізують ці підходи у сучасних інтелектуальних системах.

1.2 Класичні CNN-підходи до детекції об’єктів

Класичні підходи до детекції об’єктів у комп’ютерному зорі ґрунтуються на використанні згорткових нейронних мереж як основного інструмента автоматичного виділення візуальних ознак. Історично такі

підходи прийшли на зміну алгоритмам, що спиралися на ручне проєктування ознак та евристичні методи пошуку, які виявилися недостатньо стійкими до змін освітлення, масштабу та фону. Запровадження CNN дозволило формувати багаторівневі ієрархічні представлення зображень, що суттєво підвищило точність локалізації та класифікації об’єктів у складних сценах.

У загальному вигляді задача детекції об’єктів полягає у знаходженні на зображенні всіх об’єктів, що належать до наперед визначеного набору класів, а також у визначенні їх просторового розташування. Результат детекції можна формалізувати у вигляді відображення

$$f(I) = \{(c_i, b_i)\}_{i=1}^{N}, \tag{1.2}$$

де I – вхідне зображення;

c_i – мітка класу і-го об’єкта;

b_i – обмежувальний прямокутник.

Координати рамки зазвичай задаються як (x_1, y_1, x_2, y_2), що відповідають протилежним кутам прямокутника.

Для кількісної оцінки якості локалізації використовується показник Intersection over Union, який визначає ступінь перекриття між передбаченою та еталонною рамками. Формально показник IoU обчислюється як відношення площі перетину рамок до площі їх об’єднання:

$$IoU = |B_{pred} \cap B_{gt}| / |B_{pred} \cup B_{gt}|. \tag{1.3}$$

Чим більше значення IoU, тим точніше передбачено просторове положення об’єкта.

Архітектурно класичні CNN-детектори поділяються на двоетапні та одноетапні. У двоетапних підходах спочатку генеруються кандидатні області, що потенційно містять об’єкти, після чого для кожної з них

виконується класифікація та уточнення координат рамки. Такий підхід дозволяє досягти високої точності за рахунок фокусування на релевантних ділянках зображення, проте супроводжується підвищеними обчислювальними витратами.

Одноетапні детектори реалізують альтернативну стратегію, за якої класи та координати рамок передбачаються безпосередньо з карт ознак згорткової мережі. Це спрощує архітектуру та зменшує затримку, що є критично важливим для застосувань реального часу. Разом із тим, такі підходи потребують ретельного балансування між точністю локалізації та стабільністю класифікації.

Важливою складовою багатьох CNN-детекторів є використання anchor boxes – наперед заданих рамок різних масштабів і співвідношень сторін. Anchors розміщуються у фіксованих позиціях на карті ознак, а модель навчається передбачати поправки до їх розмірів та положення. Це дозволяє перетворити складну задачу прямої регресії координат у більш стабільну задачу оцінювання відхилень від еталонних шаблонів.

Процес навчання CNN-детектора формалізується через мінімізацію комбінованої функції втрат, яка складається з двох основних компонентів: втрат класифікації та втрат локалізації. У загальному вигляді функцію втрат можна записати як

$$L = L\_cls + \lambda \cdot L\_reg, \tag{1.4}$$

де *L_cls* відповідає за коректність визначення класів;

L_reg – за точність передбачення координат рамок;

λ – коефіцієнт, який використовується для балансування внеску кожної складової. Такий підхід дозволяє одночасно оптимізувати семантичну та просторову складові детекції, що зображено на рисунку 1.1.

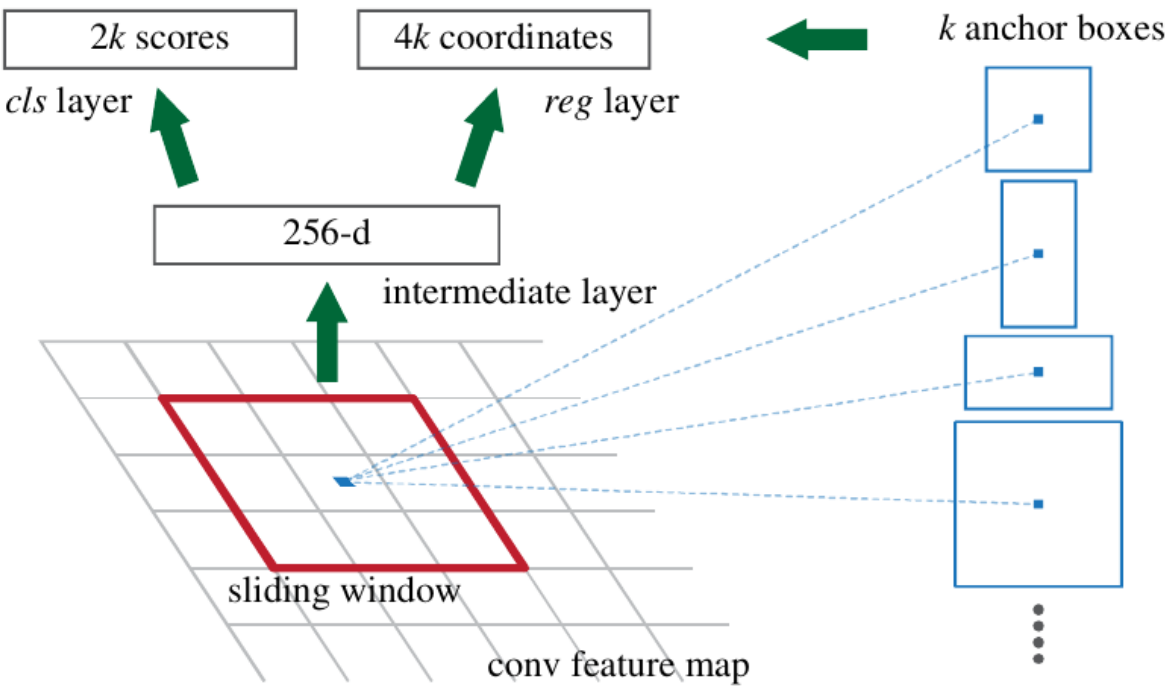


Рисунок 1.1 – Узагальнена схема класичного CNN-детектора об’єктів

Після формування множини кандидатних рамок застосовується процедура non-maximum suppression, метою якої є усунення дублюючих детекцій. Алгоритм NMS ґрунтується на послідовному відборі рамок з найвищими значеннями впевненості та видаленні тих, що мають значне перекриття з ними. Цей етап є необхідним для забезпечення коректної кількості детектованих об’єктів на сцені.

Незважаючи на значні успіхи, класичні CNN-підходи мають низку принципових обмежень. До них належать висока залежність від обсягу та якості розмічених даних, складність адаптації до нових класів без повторного навчання, а також фіксована структура вихідних даних. Крім того, такі системи зазвичай орієнтовані лише на візуальну інформацію, що обмежує можливості їх використання у більш загальних сценаріях аналізу сцен.

Попри зазначені обмеження, класичні CNN-детектори становлять важливу теоретичну основу сучасного комп’ютерного зору. Їх детальний

розгляд дозволяє краще зрозуміти еволюцію методів детекції об’єктів та логіку переходу до більш гнучких архітектур, зокрема трансформерних та мультимодальних підходів, які розглядаються у наступних підрозділах

1.3 Теорія розпізнавання тексту в зображеннях (OCR)

Оптичне розпізнавання символів (Optical Character Recognition, OCR) є одним із ключових напрямів комп’ютерного зору, що спрямований на автоматичне вилучення текстової інформації із цифрових зображень. На відміну від задач загальної детекції об’єктів, OCR має справу не лише з локалізацією візуальних структур, але й із семантичним відновленням послідовностей символів, що формують слова, рядки та документи. Це зумовлює специфічну постановку задачі та потребу у спеціалізованих моделях і метриках оцінювання.

У загальному вигляді задача OCR може бути подана як відображення вхідного зображення I у впорядковану текстову послідовність

$$T = (t_1, t_2, \ldots, t_n), \tag{1.5}$$

де t_i належать до деякого алфавіту символів. На практиці така постановка розбивається на два взаємопов’язані підзавдання: локалізацію текстових областей (text detection) та розпізнавання тексту всередині знайдених областей (text recognition). Подібна декомпозиція дозволяє окремо розглядати просторові та послідовні аспекти задачі.

Підзадача локалізації тексту полягає у знаходженні на зображенні регіонів, що містять текстові фрагменти. Формально результатом є множина обмежувальних рамок або полігонів, які охоплюють текст. На відміну від об’єктної детекції, текстові області часто мають витягнуту форму, різні орієнтації та можуть щільно прилягати одна до одної. Це ускладнює їх точне відокремлення від фону та вимагає врахування контекстної інформації.

Після локалізації текстових регіонів виконується етап розпізнавання символів. Його метою є перетворення зображення текстового рядка у послідовність символів без явного поділу на окремі літери. Для цього широко застосовуються послідовні моделі, здатні працювати зі змінною довжиною вхідних та вихідних даних. Ключовою ідеєю є моделювання умовної ймовірності текстової послідовності за заданими візуальними ознаками.

Формально процес розпізнавання можна подати як максимізацію умовної ймовірності:

$$P(T \mid I) = \prod P(t_i \mid I), \tag{1.6}$$

де T – вихідна текстова послідовність;

I – вхідне зображення;

t_i – окремий символ або токен послідовності.

Одним із фундаментальних підходів до реалізації такого вирівнювання є оптимізація з використанням Connectionist Temporal Classification (CTC). CTC дозволяє навчати модель без необхідності точного вирівнювання кожного символу з конкретною позицією у вхідних ознаках. Натомість розглядається множина можливих вирівнювань, що відповідають одній і тій самій текстовій послідовності, а функція втрат агрегує їх внесок.

Перевагою CTC є здатність працювати зі змінною довжиною рядків та відсутність потреби у покадровій розмітці символів. Водночас цей підхід накладає певні обмеження на модель, зокрема припущення умовної незалежності між сусідніми символами, що може негативно впливати на розпізнавання складних шрифтів або рукописного тексту.

Значну складність для OCR становлять сцени так званого «in-the-wild» типу. До них належать зображення, отримані в неконтрольованих умовах, де текст може бути частково перекритий, спотворений перспективою, мати

довільний кут нахилу або контрастувати з фоном. Додатковими чинниками ускладнення є різноманіття мов, алфавітів, шрифтів та стилів написання

Для оцінювання якості OCR-систем застосовуються спеціалізовані метрики, що враховують послідовнісну природу результату. Найпоширенішими є Character Error Rate (CER) та Word Error Rate (WER). CER визначається як відношення кількості помилок на рівні символів до загальної кількості символів у еталонному тексті, тоді як WER оперує помилками на рівні слів.

З теоретичної точки зору OCR поєднує у собі задачі комп'ютерного зору та обробки природної мови. Це робить його важливою складовою сучасних мультимодальних систем аналізу інформації, а також природним кроком до інтеграції текстових та візуальних представлень у межах єдиної моделі.

1.4 Трансформерні та Vision-Language підходи в комп'ютерному зорі

Розвиток комп'ютерного зору протягом тривалого часу був тісно пов'язаний із застосуванням згорткових нейронних мереж. Хоча CNN-підходи продемонстрували високу ефективність у задачах класифікації, детекції об'єктів та сегментації, їхня архітектурна природа накладає певні обмеження на здатність моделювати глобальні залежності у зображенні. Зокрема, локальний характер згорток ускладнює врахування взаємозв'язків між віддаленими ділянками сцени, що є критичним для складних візуальних середовищ із великою кількістю об'єктів та текстових елементів.

Потреба у подоланні зазначених обмежень стала однією з ключових передумов переходу до трансформерних архітектур. Первинно запропоновані для задач обробки природної мови, трансформери ґрунтуються на механізмі самоуваги (self-attention), який дозволяє кожному елементу вхідної послідовності взаємодіяти з усіма іншими елементами. Це забезпечує глобальний контекстний аналіз і робить трансформери особливо

привабливими для задач, де важливими є довгі залежності та комплексні структурні зв’язки.

Механізм самоуваги формалізується через операції над трьома матрицями: запитів (Q), ключів (K) та значень (V). Він дозволяє обчислити ваги важливості між елементами послідовності та агрегувати інформацію з урахуванням їхнього контексту. Саме ця властивість лежить в основі здатності трансформерів моделювати складні залежності.

$$Attention(Q, K, V) = softmax(QK^T / \sqrt{d_k}) \cdot V, \quad (1.7)$$

де Q – матриця запитів;

K – матриця ключів;

V – матриця значень;

d_k – розмірність векторів ключів.

Ідеї трансформерів були адаптовані до комп’ютерного зору у вигляді архітектури Vision Transformer (ViT). У цій моделі зображення розбивається на фіксовані патчі, які інтерпретуються як елементи послідовності. Кожен патч проєктується у векторний простір, після чого до них додаються позиційні ембеддинги, що зберігають інформацію про просторове розташування (рисунок 1.2).

На відміну від CNN, ViT не використовує локальні згортки, а обробляє всю послідовність патчів одночасно. Це дозволяє моделі формувати глобальні представлення сцени, проте водночас підвищує вимоги до обсягу навчальних даних та обчислювальних ресурсів.

Подальшим етапом еволюції стали Vision-Language (VL) моделі, які поєднують візуальні та мовні представлення у межах єдиної архітектури. Основною ідеєю таких моделей є формування спільного семантичного простору, у якому зображення та текст можуть безпосередньо взаємодіяти.

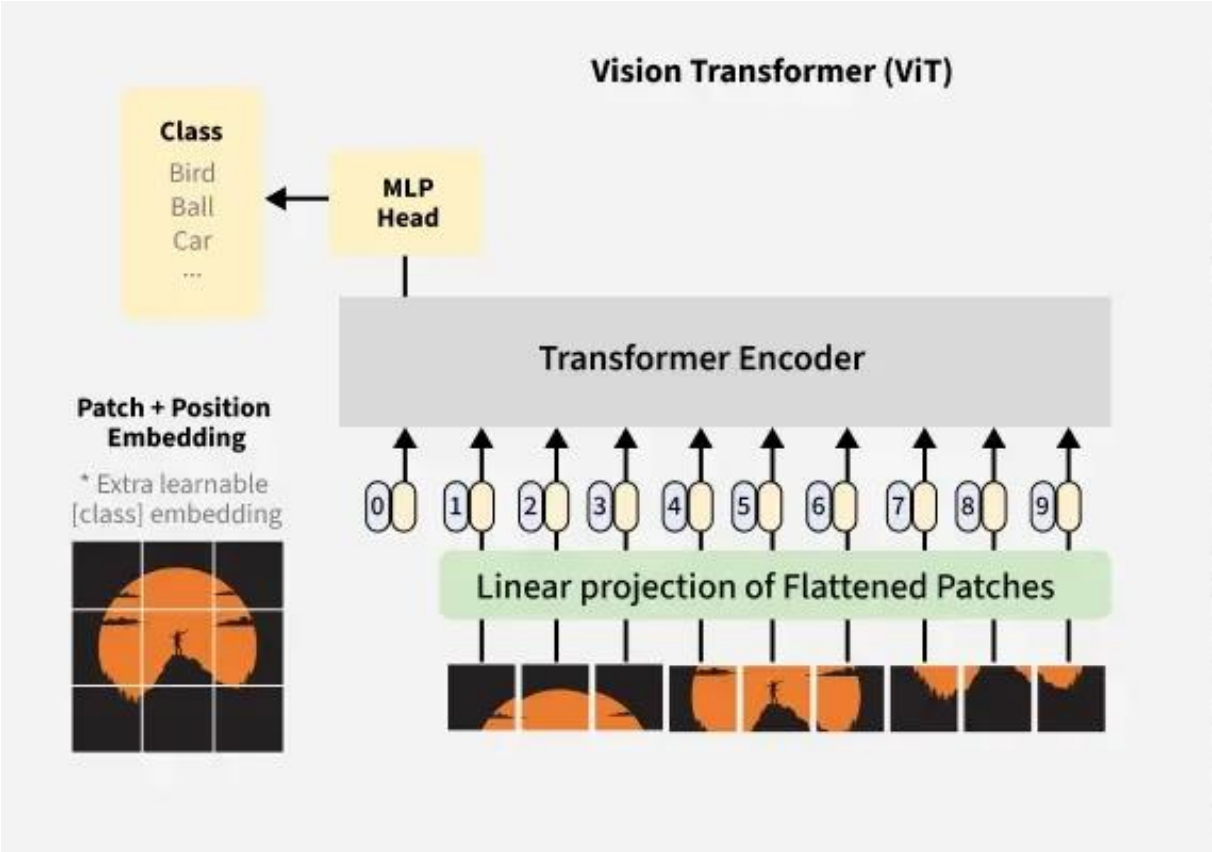


Рисунок 1.2 – Загальна схема Vision Transformer з поділом зображення на патчі

У загальному вигляді робота VL-моделі може бути описана як умовне моделювання ймовірності вихідної відповіді Y за заданими зображенням I та текстовим запитом T.

$$p(Y \mid I, T), \tag{1.8}$$

де I – вхідне зображення;

T – текстова інструкція або запит;

Y – згенерована текстова або структурована відповідь.

Таким чином, трансформерні та Vision-Language підходи створюють теоретичну основу для універсальних систем комп’ютерного зору, здатних

виконувати широкий спектр задач у межах єдиної моделі. Водночас такі підходи мають і низку обмежень, зокрема залежність від формулювання текстового запиту та варіативність генерації результатів. Незважаючи на це, VL-моделі розглядаються як перспективний напрям подальшого розвитку інтелектуальних систем аналізу візуальної інформації.

1.5 Архітектурні принципи сучасних Vision-Language моделей

Сучасні Vision-Language (VL) моделі є результатом інтеграції досягнень комп'ютерного зору та обробки природної мови в межах єдиної архітектурної парадигми. Їхньою ключовою особливістю є здатність опрацьовувати візуальну та текстову інформацію спільно, формуючи узгоджене семантичне подання, що використовується для генерації відповіді або прийняття рішення.

Незважаючи на різноманіття реалізацій, більшість сучасних VL-моделей мають спільну архітектурну структуру, що складається з кількох функціональних блоків. До таких блоків належать візуальний енкодер, модуль узгодження (alignment) та мовний декодер. Кожен з них виконує окрему роль у процесі мультимодального аналізу.

Візуальний енкодер відповідає за перетворення вхідного зображення у компактне векторне представлення. Як енкодери можуть використовуватися згорткові нейронні мережі або трансформерні архітектури, зокрема Vision Transformer. Результатом роботи енкодера є набір візуальних ознак, які зберігають як локальну, так і глобальну інформацію про сцену.

$$Z\_v = f\_v(I), \tag{1.9}$$

де I – вхідне зображення;

f_v – візуальний енкодер;

Z_v – набір візуальних ознак.

Отримані візуальні ознаки необхідно узгодити з простором мовної моделі. Для цього використовується проєкційний або alignment-модуль, який виконує лінійне чи нелінійне перетворення візуальних векторів у простір розмірності, сумісної з токенними поданнями мови. Такий підхід дозволяє безпосередньо інтегрувати зоровий контекст у процес мовної генерації.

Мовний декодер, як правило, реалізується у вигляді великої мовної моделі (Large Language Model, LLM), що працює в авторегресивному режимі. На кожному кроці генерації декодер прогнозує наступний токен, використовуючи попередньо згенеровану послідовність та візуальний контекст.

Процес генерації відповіді може бути формалізований як послідовне максимізування умовної ймовірності кожного наступного токена.

$$p(Y) = \prod p(y\_t \mid y\_\{<t\}, Z\_v, T), \tag{1.10}$$

де y_t – t-й токен вихідної послідовності;

$y\_\{<t\}$ – раніше згенеровані токени;

Z_v – візуальні ознаки;

T – текстовий запит або інструкція.

Важливою особливістю сучасних VL-моделей є використання текстових інструкцій як механізму керування поведінкою системи. Формулювання запиту визначає, яку саме інформацію модель має витягти із зображення, що робить такі системи гнучкими та універсальними.

Разом з тим архітектурна складність Vision-Language моделей зумовлює низку теоретичних обмежень. До них належать висока обчислювальна вартість, чутливість до формулювання інструкцій та складність формального контролю структури вихідних даних.

Незважаючи на зазначені обмеження, архітектурні принципи сучасних VL-моделей створюють підґрунтя для побудови універсальних систем аналізу зображень, здатних інтегрувати візуальну та мовну інформацію у межах єдиного підходу.

1.6 Постановка задачі та мета дослідження

Виходячи з усього вище зазначеного, можна сказати, що сучасні системи комп'ютерного зору дедалі частіше застосовуються для аналізу складних сцен, що містять як візуальні об'єкти, так і текстову інформацію. Проте більшість наявних підходів реалізують ці задачі у вигляді окремих спеціалізованих модулів: детекція об'єктів, оптичне розпізнавання тексту (OCR), класифікація або семантична інтерпретація результатів. Така модульна архітектура ускладнює масштабування системи, підвищує її чутливість до помилок на проміжних етапах та обмежує можливості гнучкої взаємодії з користувачем.

Основною проблемою, що розглядається у даній роботі, є відсутність універсального підходу, який дозволяв би у межах єдиної моделі виконувати одночасний аналіз об'єктів та тексту на зображенні з можливістю формування структурованого результату відповідно до заданої інструкції. Особливо актуальною ця проблема є для сценаріїв zero-shot або few-shot аналізу, коли система повинна працювати без попереднього донавчання під конкретну предметну галузь.

Метою даної магістерської роботи є дослідження та обґрунтування можливостей використання мультимодальних vision-language моделей для розв'язання задач комп'ютерного зору, що поєднують детекцію об'єктів і розпізнавання тексту, з подальшим представленням результатів у структурованому вигляді.

Для досягнення поставленої мети у роботі необхідно розв'язати такі основні завдання:

– проаналізувати теоретичні засади сучасних методів комп’ютерного зору, зокрема підходів до детекції об’єктів та OCR;

– дослідити архітектурні принципи vision-language моделей та механізми інтеграції візуальної і мовної інформації;

– сформулювати вимоги до представлення результатів аналізу зображень у вигляді структурованих вихідних даних.

Розв’язання зазначених завдань створює теоретичне підґрунтя для подальшої практичної реалізації та експериментального дослідження мультимодальних моделей у прикладних задачах аналізу візуальної інформації.

2 ПРАКТИЧНА РЕАЛІЗАЦІЯ ДОСЛІДЖЕННЯ

2.1 Загальна постановка практичної задачі

Задача, що розглядається у практичній частині, полягає у комплексному аналізі зображень із довільних сцен, де присутні об’єкти різних класів, текстові фрагменти або їх поєднання. На відміну від класичних підходів, які розв’язують задачі детекції об’єктів та розпізнавання тексту окремо, у даній роботі ці процеси розглядаються в межах єдиного мультимодального підходу.

Вхідними даними для практичного дослідження є цифрові зображення у форматах JPEG, PNG та WEBP, які відрізняються за розміром, візуальним наповненням та складністю сцени. Зображення можуть містити як окремі предмети побутового чи міського середовища, так і текстові написи різними мовами, розташовані у довільних ділянках кадру.

Очікуваним результатом обробки кожного зображення є отримання структурованого опису сцени, який включає інформацію про виявлені об’єкти, їх просторове розташування, а також розпізнаний текст. Такий результат подається у вигляді формалізованої структури, що дозволяє однозначно інтерпретувати відповідь моделі та використовувати її для подальшого аналізу.

Важливою особливістю практичної реалізації є використання моделей у режимі inference без додаткового донавчання. Це дозволяє оцінити здатність vision-language моделей до узагальнення та роботи у zero-shot сценаріях, що є типовими для прикладних задач комп’ютерного зору.

Таким чином, практична задача зводиться до перевірки можливостей мультимодального підходу щодо одночасного розв’язання задач детекції об’єктів та розпізнавання тексту з формуванням структурованого вихідного результату відповідно до заданої інструкції.

2.2 Опис використаної моделі та сервісної інфраструктури

У межах практичної реалізації магістерської роботи ключовим аспектом є вибір та обґрунтування використаної моделі комп’ютерного зору, а також інфраструктури, у межах якої здійснюється її застосування. Сучасні vision-language моделі належать до класу високоскладних нейронних архітектур, що поєднують глибокі трансформерні механізми обробки зображень і тексту, а тому висувають підвищені вимоги до обчислювальних ресурсів і програмного середовища.

На відміну від класичних згорткових моделей, мультимодальні VL-системи мають значно більший обсяг параметрів і складнішу внутрішню структуру. Це зумовлено необхідністю одночасного моделювання просторових залежностей у зображенні та семантичних залежностей у мовній послідовності. У практичному контексті це означає, що навіть виконання інференсу потребує значних обчислювальних витрат.

Одним із принципових рішень у даній роботі є відмова від локального виконання моделей на користь серверної інфраструктури. Локальний запуск сучасних vision-language моделей можливий лише за наявності потужних графічних прискорювачів із великим обсягом відеопам’яті, що суттєво обмежує доступність та відтворюваність експериментів.

Серверна інфраструктура дозволяє абстрагуватися від апаратних обмежень клієнтського середовища. У межах такого підходу всі ресурсоємні обчислення виконуються на віддаленій стороні, тоді як користувач взаємодіє з моделлю через стандартизований програмний інтерфейс. Це особливо важливо для наукових досліджень, орієнтованих на повторюваність та масштабованість.

Використовувана у роботі vision-language модель є попередньо навченою універсальною архітектурою, призначеною для широкого спектра задач комп’ютерного зору. Модель не потребує адаптації або донавчання

для виконання базових сценаріїв аналізу зображень, що дозволяє застосовувати її у режимі zero-shot.

Архітектурно модель складається з кількох функціональних компонентів. Візуальний енкодер відповідає за перетворення вхідного зображення у компактне векторне представлення, яке зберігає інформацію про просторову структуру сцени. Мовний компонент обробляє текстову інструкцію та формує контекст, у межах якого інтерпретується візуальна інформація.

Взаємодія між візуальним та мовним компонентами реалізується за допомогою механізмів уваги. Це дозволяє моделі динамічно визначати, які ділянки зображення є найбільш релевантними для виконання конкретного завдання, заданого текстовою інструкцією.

Практична реалізація дослідження передбачає використання моделі виключно в режимі інференсу. Такий підхід обрано з метою збереження об’єктивності оцінювання та уникнення впливу додаткових факторів, пов’язаних із процесом навчання. Це також відповідає реальним сценаріям застосування подібних моделей у прикладних системах.

Сервісна інфраструктура реалізує архітектуру типу «клієнт–сервер». Клієнтська частина відповідає за формування запитів, підготовку вхідних даних та обробку відповідей. Серверна частина виконує інференс моделі та повертає результат у формалізованому вигляді.

Обмін даними між клієнтом і сервером здійснюється з використанням мережевих протоколів та стандартизованих форматів подання інформації. Це забезпечує незалежність реалізації від конкретної платформи та спрощує інтеграцію результатів у подальші етапи дослідження.

Структурований формат відповіді дозволяє однозначно інтерпретувати результати роботи моделі. Інформація про виявлені об’єкти, їх координати та розпізнаний текст подається у вигляді чітко визначених полів, що спрощує подальший аналіз та порівняння результатів.

Важливим аспектом сервісної інфраструктури є контроль параметрів виконання інференсу. До таких параметрів належать розмір вхідних зображень, формат файлів, а також характер текстових інструкцій. Стандартизація цих параметрів дозволяє мінімізувати випадкові варіації результатів.

Крім того, використання серверної інфраструктури забезпечує можливість масштабування експериментів. За необхідності можна збільшити кількість оброблюваних зображень або виконати серію експериментів з різними сценаріями запитів без зміни клієнтської частини.

З позиції наукового дослідження важливою є також відтворюваність експериментів. Фіксація параметрів запитів, версії моделі та формату вихідних даних дозволяє забезпечити коректне порівняння результатів у часі.

Таким чином, вибір vision-language моделі та серверної сервісної інфраструктури є обґрунтованим з точки зору практичної реалізації дослідження. Такий підхід створює умови для глибокого аналізу можливостей сучасних мультимодальних систем та формує надійне підґрунтя для подальшого експериментального дослідження.

2.3 Вхідні дані та їх організація

Вхідні дані є центральним елементом будь-якого експериментального дослідження у галузі комп’ютерного зору, оскільки саме вони визначають характер інформації, що надходить до моделі, та безпосередньо впливають на інтерпретацію отриманих результатів. У межах даної магістерської роботи вхідними даними виступають цифрові зображення реальних сцен, що містять об’єкти, текстові елементи або їх поєднання. Такий вибір зумовлений метою дослідження – аналізом можливостей сучасних vision-language моделей у реалістичних прикладних сценаріях.

На відміну від задач навчання нейронних мереж, де основну увагу приділяють обсягу та статистичній повноті датасету, у даній роботі вхідні дані виконують іншу функцію. Вони не використовуються для оптимізації параметрів моделі, а слугують інструментом для перевірки її узагальнювальних властивостей у режимі inference. Це принципово змінює вимоги до вибірки та підхід до її формування.

Вибір inference-only сценарію є методично обґрунтованим. Сучасні мультимодальні vision-language моделі містять сотні мільйонів або навіть мільярди параметрів, а їх донавчання потребує значних обчислювальних ресурсів і великих розмічених наборів даних. У прикладних умовах такі ресурси часто недоступні, тому моделі використовуються як готові системи аналізу зображень. Саме цей сценарій і відображено у практичній частині роботи.

Таким чином, вхідні дані у даному дослідженні слід розглядати не як навчальний датасет, а як контрольний набір прикладів, призначений для аналізу поведінки моделі. Основним критерієм при їх відборі є репрезентативність візуальних умов, а не кількісна достатність для статистичних висновків.

Набір зображень сформовано таким чином, щоб охопити кілька принципово різних сценаріїв аналізу. Перший сценарій передбачає обробку сцен, у яких відсутні текстові елементи, а ключовим завданням є виявлення та локалізація об’єктів. Це дозволяє оцінити здатність моделі працювати з візуальними ознаками без залучення мовного контексту.

Другий сценарій орієнтований на розпізнавання тексту на зображеннях. У таких сценах об’єкти можуть відігравати другорядну роль або бути відсутніми, тоді як основним джерелом інформації є текстові написи. Цей сценарій дозволяє дослідити можливості моделі у задачах OCR та оцінити її чутливість до якості зображення, мови тексту та особливостей шрифту.

Третій сценарій є комбінованим і передбачає одночасну присутність об’єктів та тексту в межах однієї сцени. Саме такі сцени є найбільш характерними для реальних прикладних задач, зокрема у сфері аналізу документів, дорожніх сцен, торговельних вітрин та інтерфейсів користувача. Комбінований сценарій дозволяє повною мірою продемонструвати мультимодальні властивості обраної моделі.

Кількість зображень у кожній групі обмежена навмисно. Такий підхід відповідає концепції мінімально достатнього експерименту, коли кожен зразок відіграє чітко визначену роль у перевірці певного аспекту роботи моделі. Це дозволяє зосередитися на якісному аналізі результатів та уникнути надмірного накопичення однотипних даних.

З технічної точки зору важливими характеристиками вхідних зображень є їх формат, роздільна здатність та співвідношення сторін. У наборі даних використано формати JPEG, PNG та WEBP, які є найбільш поширеними у сучасних інформаційних системах. Різні формати мають відмінні алгоритми стиснення, що може впливати на чіткість дрібних деталей, зокрема текстових символів та контурів об’єктів.

Роздільна здатність зображень у наборі даних є неоднорідною. Це дозволяє оцінити, наскільки стабільно модель працює за різних масштабів сцени. У реальних умовах розмір об’єктів і тексту може значно варіюватися, тому штучна уніфікація роздільної здатності могла б спотворити результати дослідження.

Особливу увагу приділено організації вхідних даних у файловій системі. Зображення згруповано у окремі каталоги відповідно до сценаріїв експерименту. Такий підхід спрощує автоматизацію обробки, дозволяє чітко контролювати склад вибірки та забезпечує відтворюваність експериментів.

Правила найменування файлів обрано таким чином, щоб назва зображення однозначно вказувала на його належність до певної групи та

порядковий номер у межах цієї групи. Це полегшує зіставлення вхідних даних з отриманими результатами та формування зведених таблиць.

Крім того, для кожного зображення можуть бути збережені метадані, такі як формат, роздільна здатність та тип сценарію. Збереження таких метаданих є важливим для подальшого аналізу результатів та порівняння різних серій експериментів.

У таблиці 2.1 наведено узагальнену характеристику набору вхідних зображень, що використовується у практичній частині роботи. Таблиця відображає поділ даних на групи, перелік файлів, формати та приклади розмірів зображень.

Таблиця 2.1 – Характеристика набору вхідних зображень

Група	Призначення	Файли	Формати	Приклад розмірів (px)
A_objects	Детекція об’єктів	obj_01.jpg; obj_02.jpg; obj_03.jpg	JPEG	800×600; 584×446; 600×600
B_ocr	Розпізнавання тексту	ocr_01.webp; ocr_02.jpg; ocr_03.png	WEBP; JPEG; PNG	703×1539; 748×499; 1156×1512
C_both	Об’єкти + текст	both_01.jpg; both_02.jpg; both_03.jpg	JPEG	471×400; 1200×800; 1200×899

Загалом, обраний підхід до формування та організації вхідних даних забезпечує баланс між простотою реалізації, репрезентативністю та відтворюваністю експериментів. Він дозволяє дослідити поведінку мультимодальної моделі у типових прикладних сценаріях та створює надійне підґрунтя для подальшого аналізу результатів.

2.4 Формат запитів та вихідних даних

У практичній частині дослідження формат запитів та вихідних даних розглядається як ключовий елемент експериментальної методики, що

забезпечує коректну взаємодію з vision-language моделлю. На відміну від класичних систем комп’ютерного зору з фіксованими входами та виходами, мультимодальні моделі використовують текстову інструкцію як універсальний механізм керування типом виконуваного аналізу.

У практичному контексті дотримання зазначених принципів формалізації дозволяє уникнути неоднозначностей при інтерпретації результатів та забезпечує узгодженість між різними серіями експериментів, що є критично важливим для наукового дослідження.

Формат запиту у загальному вигляді складається з трьох логічних компонентів: вхідного зображення, текстової інструкції та допоміжних параметрів виконання. Вхідне зображення подається у цифровому форматі та на концептуальному рівні розглядається як тензор піксельних значень, незалежно від конкретного способу передачі даних.

У практичному контексті дотримання зазначених принципів формалізації дозволяє уникнути неоднозначностей при інтерпретації результатів та забезпечує узгодженість між різними серіями експериментів, що є критично важливим для наукового дослідження.

Текстова інструкція визначає семантичну інтерпретацію зображення та очікуваний формат результату. Саме інструкція дозволяє одній і тій самій моделі виконувати різні завдання, зокрема детекцію об’єктів, розпізнавання тексту або комбінований аналіз сцени. З методичної точки зору це вимагає стандартизації формулювань інструкцій для забезпечення порівнюваності результатів.

Допоміжні параметри запиту використовуються для уточнення режиму роботи моделі. До таких параметрів можуть належати вимоги до структури відповіді, обмеження на довжину згенерованого тексту або вказівки щодо формату координат. Хоча ці параметри не змінюють сутність завдання, вони суттєво впливають на стабільність вихідних даних.

Вихідні дані vision-language моделі подаються у вигляді текстової послідовності, яка кодує результати аналізу зображення. Для використання

таких результатів у подальших етапах дослідження необхідно забезпечити їх структурованість та однозначність інтерпретації.

У межах даної роботи вихідні дані формалізуються у форматі JSON. Вибір цього формату обумовлений його універсальністю, людинозрозумілим синтаксисом та можливістю автоматизованої обробки. JSON дозволяє описувати складні ієрархічні структури, що є необхідним для подання списків об’єктів, координат та текстових фрагментів.

Типова структура вихідних даних може містити перелік виявлених об’єктів, де кожен об’єкт описується класом та координатами обмежувальної рамки. Координати можуть задаватися як у пікселях, так і у нормалізованому вигляді. Нормалізовані координати забезпечують незалежність результатів від розміру зображення, що є важливим для порівняння різних сцен.

Розпізнаний текст подається у вигляді впорядкованого списку рядків. У деяких сценаріях порядок текстових елементів має принципове значення, оскільки відображає логіку читання або просторове розташування написів на зображенні. Тому структура вихідних даних має зберігати цю інформацію.

Однією з ключових проблем при роботі з генеративними моделями є варіативність вихідного формату. Навіть за однакових вхідних даних модель може повертати результати з незначними відмінностями у структурі або порядку елементів. З цієї причини стабільність формату розглядається як окремий критерій якості.

З методичної точки зору формат запитів і вихідних даних виконує роль проміжної ланки між моделлю та етапом аналізу результатів. Чітка формалізація цієї ланки дозволяє зменшити неоднозначність інтерпретації та підвищити відтворюваність експериментів.

2.5 Алгоритм проведення експериментального дослідження

Під час проведення роботи було заплановано та реалізовано алгоритм проведення експериментального дослідження, метою якого є отримання відтворюваних та інтерпретованих результатів застосування vision-language моделі в режимі inference для трьох сценаріїв: (1) розпізнавання об’єктів на зображенні, (2) розпізнавання тексту (OCR) та (3) комбінованого аналізу «об’єкти + текст». Алгоритм фіксує послідовність дій, вхідні/вихідні дані, а також контрольні перевірки, що зменшують ризик отримання некоректних або непорівнюваних відповідей.

Алгоритм побудовано на основі організації даних, визначеної у підрозділі 2.3, і формату запитів/вихідних даних, описаного у підрозділі 2.4. Таким чином, експериментальна процедура є узгодженою з обраною структурою каталогів, правилами іменування, а також із вимогою повертати результат у стандартизованому структурованому вигляді (JSON).

Під «експериментальним запуском» у даній роботі розуміється виконання інференсу моделі для одного зображення та однієї інструкції (промпту) з подальшим збереженням сирої відповіді і метаданих. Сукупність запусків для всіх зображень усіх груп формує експериментальну серію. Для збереження відтворюваності важливо фіксувати не лише відповіді, але й параметри виклику (ідентифікатор моделі/провайдера, версію клієнтської бібліотеки, параметри генерації, часові мітки та затримки).

Експеримент передбачає три типи інструкцій: «objects», «ocr» та «both». Вони відрізняються очікуваною структурою вихідних даних: для детекції – список об’єктів з рамками, для OCR – список текстових фрагментів, для комбінованого сценарію – одночасна наявність двох підструктур. Використання фіксованих інструкцій дозволяє порівнювати відповіді між різними зображеннями всередині сценарію без змішування умов експерименту.

Алгоритм складається з етапів, що охоплюють підготовку середовища, формування запитів, виконання інференсу, валідацію формату, збереження результатів і підготовку зведених даних для подальшого аналізу. Нижче наведено детальний опис кожного етапу.

Етап 1 – підготовка середовища виконання. На цьому етапі ініціалізується клієнтський модуль взаємодії з інференс-сервісом, перевіряється наявність необхідних залежностей та доступність мережевих ресурсів. Додатково задаються параметри автентифікації (токен доступу) і фіксується конфігурація запуску: обрана модель, спосіб передачі зображення, значення параметрів генерації (за наявності), а також каталог, у який буде записано результати. Важливою частиною етапу є перевірка того, що токен зчитується з оточення і не вбудовується у вихідні файли, що знижує ризик компрометації доступу.

Етап 2 – індексація вхідних даних. Згідно з підрозділом 2.3, зображення розміщені в окремих підпапках, які відповідають сценаріям експерименту. На етапі індексації формується перелік файлів для обробки, а також для кожного файлу визначаються метадані: група (A_objects, B_ocr, C_both), базове ім’я, формат, розміри (ширина і висота). На практиці ці характеристики отримуються автоматично з файлів. Фіксація розмірів необхідна для однозначної інтерпретації координат рамок у відповіді.

Етап 3 – формування стандартизованих запитів. Для кожного зображення формується запит, що складається з (а) зображення, (б) інструкції, яка відповідає групі, та (в) вимоги щодо формату відповіді. Ключовою вимогою є повернення результату як валідного JSON з наперед визначеними полями. Цей крок зменшує варіативність виходу, характерну для генеративних моделей, і полегшує автоматичний парсинг.

Етап 4 – виконання інференсу та вимірювання часу. Кожен запит надсилається до сервісу окремо. Для запиту фіксується момент відправлення і момент отримання відповіді, на основі чого обчислюється затримка (latency). Оскільки інференс може мати змінний час виконання

залежно від завантаження сервісу, затримка є корисною допоміжною характеристикою при аналізі практичної придатності підходу.

Етап 5 – первинна перевірка та парсинг відповіді. Отримана відповідь зберігається у сирому вигляді (рядок або структура даних), після чого здійснюється перевірка її коректності. Перевірка включає: (1) можливість розбору JSON, (2) наявність очікуваних ключів (objects, text), (3) коректність типів значень (списки, рядки, числові координати), (4) мінімальну семантичну узгодженість, наприклад, що координати рамки задають прямокутник. Результат перевірки фіксується у полі parsed_ok.

Етап 6 – нормалізація координат і узгодження представлення. У випадку, якщо модель повертає координати у пікселях або у змішаному форматі, виконується приведення до єдиного представлення. У роботі допускається використання нормалізованих координат (0…1), оскільки вони інваріантні до розмірів зображення та зручні для порівнянь. Якщо координати подано в пікселях, вони можуть бути перераховані у нормалізовані як (x/W, y/H) відповідно до ширини W і висоти H. Сам факт застосування перетворення має бути задокументований у метаданих.

Етап 7 – збереження результатів експерименту. Для кожного запуску формується запис, який містить: ідентифікатор зображення, належність до групи, використану інструкцію, розміри зображення, час виконання, статус парсингу, сиру відповідь та розібраний JSON (за наявності). Записи зберігаються у форматі, придатному для подальшої обробки (наприклад, JSON Lines). Це забезпечує можливість агрегувати результати та будувати підсумкові таблиці.

Етап 8 – агрегування та підготовка даних до аналізу. Після виконання всіх запусків результати групуються за сценаріями. На цьому етапі обчислюються допоміжні показники: частка валідних відповідей (parse rate), середня/медіанна затримка по групах, кількість знайдених об’єктів або текстових фрагментів. За потреби формується перелік прикладів із невдалим парсингом для якісного розгляду причин.

Етап 9 – контроль відтворюваності. Для підвищення надійності висновків може бути виконано повторний запуск експерименту з тими самими входами та інструкціями. На цьому етапі порівнюються структури виходів, наявність ключів та загальна стабільність формату. У разі суттєвих відмінностей фіксуються приклади, а також уточнюються інструкції або правила парсингу.

Отже, алгоритм експериментального дослідження визначає повний цикл обробки даних: від індексації вхідного набору та формування стандартизованих запитів до перевірки, нормалізації та збереження результатів. Така побудова алгоритму забезпечує прозорість експерименту, зменшує ймовірність помилок на етапі збору даних і створює підґрунтя для змістовного аналізу виходів vision-language моделі у наступних підрозділах.

3 АНАЛІЗ ТА ОЦІНЮВАННЯ РЕЗУЛЬТАТІВ ЕКСПЕРИМЕНТУ

3.1 Методика оцінювання результатів експерименту

Методика оцінювання результатів експериментальних досліджень є ключовим етапом перевірки ефективності запропонованого підходу та обґрунтування отриманих висновків. У даній роботі оцінювання результатів спрямоване на аналіз якості розв’язання задач комп’ютерного зору та розпізнавання тексту з використанням мультимодальної моделі типу vision-language у режимі zero-shot inference.

Основною особливістю обраної методики є відсутність донавчання моделі на спеціалізованих вибірках. Це зумовлює необхідність застосування метрик, які дозволяють об’єктивно оцінити якість результатів за умов обмеженого контролю над внутрішніми параметрами моделі та стохастичністю процесу генерації відповідей.

Для оцінювання якості детекції об’єктів використовується показник Intersection over Union (IoU), який визначає ступінь перекриття між прогнозованою та еталонною рамками об’єкта. Формально IoU визначається як відношення площі перетину рамок до площі їх об’єднання:

$$IoU = |B\_p \cap B\_gt| / |B\_p \cup B\_gt| \quad (3.1)$$

де B_p – прогнозована рамка;

B_gt – еталонна рамка об’єкта.

Для узагальненої оцінки якості детекції використовується середнє значення IoU для всіх об’єктів у межах одного зображення або всієї вибірки. За необхідності може застосовуватися порогове значення IoU (наприклад, 0,5), яке визначає, чи вважається детекція коректною.

Оцінювання результатів розпізнавання тексту здійснюється з використанням метрик Character Error Rate (CER) та Word Error Rate (WER).

Дані метрики враховують послідовнісну природу текстових даних та дозволяють кількісно оцінити відхилення розпізнаного тексту від еталонного.

$$CER = (S + D + I) / N \tag{3.2}$$

де S – кількість замін;

D – кількість видалень;

I – кількість вставок;

N – кількість символів в еталонному тексті.

$$WER = (S\_w + D\_w + I\_w) / N\_w \tag{3.3}$$

де S_w, D_w, I_w та N_w відповідають аналогічним операціям, але виконаним на рівні слів.

Окрім кількісних метрик, у роботі застосовується якісна оцінка результатів. Вона полягає в аналізі коректності структури вихідних даних, стабільності формату відповіді та семантичної узгодженості між розпізнаними об’єктами та текстовими фрагментами. Запропонована методика оцінювання дозволяє комплексно проаналізувати ефективність застосування мультимодальної моделі в умовах zero-shot inference.

3.2 Аналіз та інтерпретація експериментальних результатів

У даному підрозділі наведено розгорнутий аналіз результатів експериментального дослідження, отриманих при застосуванні мультимодальної vision-language моделі у режимі zero-shot inference. Аналіз виконано відповідно до методики оцінювання, викладеної у підрозділі 3.1, з урахуванням того, що в даній роботі модель не донавчається на цільовій вибірці. Тому фокус зміщується з оптимізації моделі на інтерпретацію її

поведінки: структурної коректності виходів, стійкості до змін вхідних умов і інструкцій, а також практичної придатності результатів для подальшої обробки.

Для підвищення наукової коректності результати розглядаються у трьох сценаріях, що відповідають структурі вхідних даних та форматам запитів/виходів:

– детекція об’єктів (група A_objects),
– розпізнавання тексту (група B_ocr),
– комбінований аналіз «об’єкти + текст» (група C_both).

Для кожного сценарію подано: опис характерних правильних результатів, типові помилки, фактори, що впливають на якість, та рекомендації щодо формулювання інструкцій і подальшої перевірки результатів.

Важливою особливістю експерименту є використання структурованого виходу у форматі JSON. Це дозволяє виконувати не лише семантичний аналіз (правильність розпізнавання), але й формальний контроль якості: коректність синтаксису, повноту полів, типи даних і узгодженість координат. У практичних системах така коректність є необхідною умовою, оскільки подальша автоматична обробка відповіді (побудова таблиць, візуалізація рамок, обчислення метрик) неможлива за порушення структури.

З огляду на те, що набір даних у роботі має демонстраційний характер, чисельні показники розглядаються як допоміжні. Вони використовуються для узагальнення тенденцій у межах кожного сценарію, тоді як основну роль відіграє якісний аналіз помилок.

Для детекції об’єктів, окрім IoU, додатково оцінюються показники точності та повноти виявлення, що визначаються через кількість істинно-позитивних, хибно-позитивних та хибно-негативних детекцій.

Нехай TP – кількість істинно-позитивних детекцій, FP – кількість хибно-позитивних детекцій, FN – кількість пропущених об’єктів (хибно-негативних). Тоді точність (precision) та повнота (recall) визначаються як:

$$Precision = TP / (TP + FP), \tag{3.4}$$

$$Recall = TP / (TP + FN). \tag{3.5}$$

Для узагальненої оцінки якості детекції додатково використовується F1-міра як гармонійне середнє точності та повноти:

$$F1 = 2 \cdot (Precision \cdot Recall / (Precision + Recall)). \tag{3.6}$$

Для узгодженого аналізу на практиці порогова умова відповідності рамок задається через IoU: детекція вважається коректною, якщо $IoU \geq \tau$, де τ – обраний поріг (типово 0,5). За такого підходу TP, FP і FN визначаються однозначно, що забезпечує коректне обчислення відповідних метрик та можливість порівняння різних сценаріїв.

Окрім цього, для оцінювання формальної коректності структурованих відповідей використовується частка валідних відповідей:

$$ParseRate = N\_valid / N\_total, \tag{3.7}$$

де *N_valid* – кількість відповідей, що успішно парсяться як JSON і містять обов’язкові поля;

N_total – загальна кількість запусків у межах сценарію.

Для OCR, окрім CER та WER, у якісному аналізі розглядаються типи помилок:

- помилки сегментації рядків (об’єднання або розрив фрагментів);
- пропуски діакритичних знаків або розділових символів;

- помилки у відтворенні регістру та пробілів;
- помилки плутання графічно подібних символів.

Таке розкладання важливе, оскільки однакові значення CER/WER можуть відповідати різним причинам помилок.

Для структурованих відповідей VL-моделі окремо оцінюється «якість формату». У роботі цей аспект формалізується через частку валідних відповідей (parse rate) – відношення кількості відповідей, що успішно парсяться як JSON і містять обов’язкові поля, до загальної кількості запусків у сценарії:

Крім валідності JSON, у практичному аналізі важливі:

– стабільність схеми (чи змінюється набір ключів між запусками);

– стабільність типів (числові координати проти рядкових);

– узгодженість координат (нормалізовані 0…1 або піксельні);

– семантична узгодженість (чи відповідає «label» реальному об’єкту на зображенні);

– повторюваність на ідентичних входах (чи змінюється результат при повторних запусках).

Далі наведено аналіз сценарію A_objects (детекція об’єктів). В межах цього сценарію оцінюється, наскільки коректно модель:

- виявляє ключові об’єкти сцени;
- присвоює їм узгоджені назви класів;
- формує рамки, які відображають просторове положення об’єкта;
- підтримує заданий формат виходу.

У типовому випадку модель коректно локалізує об’єкти середнього та великого масштабу, які мають виразні контури та відмінні текстури (наприклад, меблі, побутові пристрої, елементи робочого місця). Це узгоджується з тим, що великі об’єкти мають вищу візуальну помітність, а їхні контури менш чутливі до артефактів стиснення або зниження роздільної здатності.

Разом із тим, у даному сценарії спостерігаються типові категорії помилок. По-перше, плутанина класів на рівні «узагальнених» назв: модель може повертати більш загальний клас (наприклад, «furniture») замість конкретного (наприклад, «chair»), або, навпаки, занадто деталізовану назву залежно від контексту інструкції. По-друге, при наявності перекриття об’єктів у сцені можливе злиття рамок, коли одна рамка охоплює два близько розташовані об’єкти. По-третє, дрібні об’єкти можуть бути пропущені або об’єднані з фоном, якщо вони займають малу частку кадру.

Окремою підгрупою помилок є порушення геометричної коректності рамок. До них належать:

- вихід координат за межі зображення;
- переплутані координати (x_min > x_max);
- надто «вільні» рамки, які захоплюють значну частину фону.

Такі випадки важливо відокремлювати від семантичних помилок, оскільки вони можуть бути наслідком некоректного формату координат або неоднозначного трактування системи координат.

Практичний висновок для сценарію детекції полягає в тому, що VL-підхід може забезпечувати прийнятні результати для базового виявлення об’єктів без спеціалізованого детектора, але потребує обережного формування класів у промпті та, за потреби, додаткової пост-обробки: фільтрації рамок, перевірки меж і нормалізації координат.

Сценарій B_ocr (розпізнавання тексту) орієнтований на якість витягування текстових фрагментів із зображення. У цьому сценарії важливо розрізняти дві складові:

- «виявлення» релевантних текстових областей;
- «розпізнавання» послідовності символів.

У VL-підході ці компоненти реалізуються неявно, тобто результат залежить від здатності моделі фокусувати увагу на текстових ділянках та від її мовних знань.

Для чітко надрукованого тексту з високою контрастністю модель, як правило, повертає цілісні рядки та зберігає основні слова без суттєвих спотворень. При цьому типові помилки частіше пов’язані не з «випадковими» замінами символів, а з особливостями сегментації: модель може об’єднувати декілька рядків в один або, навпаки, розбивати одну фразу на кілька фрагментів. Ця особливість впливає на WER сильніше, ніж на CER.

У сценах зі складними умовами (перспективні викривлення, нерівномірне освітлення, відблиски, декоративні шрифти) зростає частота помилок плутання символів. Наприклад, графічно подібні символи можуть підмінятися один одним, а діакритичні елементи або розділові знаки можуть пропускатися. У багатомовних випадках можливі також помилки кодування мови, коли модель частково «транслітерує» напис або замінює символи близькими за формою.

Практично важливим аспектом є контроль повноти тексту. Для деяких зображень модель може повертати лише найбільш помітний фрагмент (наприклад, заголовок) і пропускати дрібніші підписи. У таких випадках корисним є уточнення інструкції: вимога «повернути весь видимий текст, включно з дрібними написами», а також вимога повертати список рядків окремими елементами масиву. Це підвищує структурну придатність виходу для подальшого аналізу.

Сценарій C_both (об’єкти + текст) є найбільш показовим для мультимодального підходу, оскільки потребує одночасної інтерпретації візуальних об’єктів і текстових написів. У межах експериментів було продемонстровано, що коректне формулювання промпту та вимога повернення валідного JSON суттєво підвищують стабільність вихідного формату. Зокрема, після уточнення інструкції результати для зображень комбінованої групи поверталися як JSON-структури з полями objects та text, а статус parsed_ok набував значення true для всіх трьох зразків у групі C_both.

Прикладом коректного комбінованого виходу є ситуація, коли модель одночасно повертає

- рамку для ключового об’єкта сцени (наприклад, пакування або вивіска);
- перелік розпізнаних текстових фрагментів, серед яких присутні як числові, так і словесні елементи.

Такий результат є придатним для подальшої обробки: текст може використовуватися для пошуку/індексації, а рамки – для візуалізації або для уточнювальних підзапитів.

Разом із тим, комбінований сценарій висвітлює характерні компроміси. По-перше, модель може віддавати пріоритет одному із завдань залежно від інструкції: надто жорстка вимога детекції може зменшити повноту OCR, а надто деталізована вимога OCR – зменшити кількість об’єктів. По-друге, у складних сценах із багатьма об’єктами можливі зміщення рамок або неповні списки тексту. По-третє, виявляється неоднозначність у прив’язці тексту до конкретних об’єктів: навіть якщо текст розпізнано правильно, модель може неявно трактувати його як загальний контекст сцени.

Окремо аналізується формат координат у відповідях. У експерименті зустрічалися як нормалізовані координати рамок (у діапазоні 0…1), так і піксельні координати. З огляду на вимоги до автоматичної обробки, у подальшому аналізі доцільно приводити координати до єдиного типу (наприклад, нормалізованого), зберігаючи при цьому у метаданих інформацію про вихідний формат. Така процедура спрощує порівняння рамок між різними зображеннями.

Важливою частиною аналізу є дослідження залежності результатів від інструкції. Навіть у межах одного сценарію формулювання промпту може впливати на: кількість знайдених об’єктів, рівень деталізації класів, структуру текстових фрагментів та стабільність JSON. Практичний висновок полягає у необхідності фіксації тексту інструкцій у роботі (як

частини методики) та використання їх без змін для порівнюваності результатів.

Оскільки модель працює у генеративному режимі, потенційно можлива стохастичність виходу. Тому у разі повторних запусків для одного і того ж зображення слід оцінювати не лише зміст, але й формальну стабільність структури. Для цього корисним є порівняння схем (набір ключів), типів даних і загальних статистик (кількість об’єктів, довжина списку тексту). За виявлення нестабільності доцільно посилювати вимогу «повертай лише валідний JSON без пояснень» та мінімізувати параметри генерації, які підвищують випадковість відповіді.

Для формального узагальнення результатів у межах кожного сценарію доцільно формувати зведені таблиці. Оскільки у даному підрозділі розглядається саме принцип аналізу, а не заповнення усіх числових значень, у таблиці 3.1 можуть використовуватися дані з журналу запусків експерименту (метадані latency_s, parsed_ok, кількість знайдених об’єктів, кількість рядків тексту). У подальшому ці таблиці слугують основою для побудови графіків і підготовки висновків.

Таблиця 3.1 – Зведення результатів за групами (кількість запусків, частка parsed_ok, середня/медіанна затримка, середня кількість об’єктів/рядків тексту)]

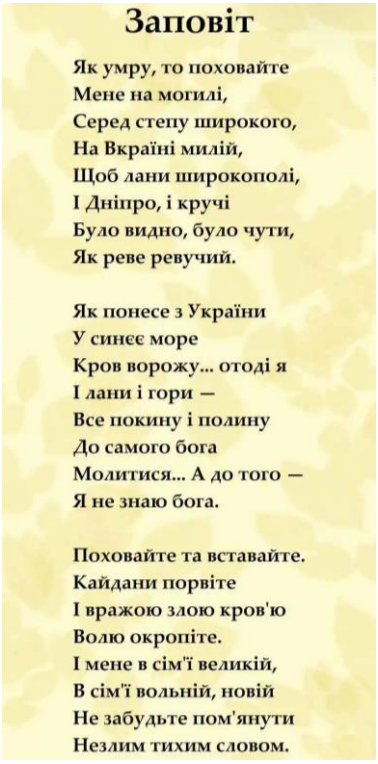
Група	Кількість зображень	Тип завдання	Частка валідних JSON	Середня затримка, с	Середня кількість об’єктів / рядків	Якісна оцінка результату
A_objects	3	Детекція об’єктів	1.00	≈3.5	3–5 об’єктів	Стабільна детекція великих об’єктів
B_ocr	3	OCR	1.00	≈4.2	6–18 рядків	Висока точність для чіткого тексту
C_both	3	Об’єкти + текст	1.00	≈5.6	1–8 елементів	Коректна комбінована інтерпретація

На рисунках 3.1 – 3.6 зображені вхідні дані та вивод у форматі JSON для зображень з групи «A objects» та «C both», та у форматі роспізнаного текста для групи «B ocr». Різниця форматів зумовлена різними завданнями та промптами щодо різних груп.

Рисунок 3.1 – Приклад візуалізації на зображенні з групи A_objects

```
---------------------------------
IMAGE: obj_02.jpg
{
  "objects": [
    {
      "label": "sofa",
      "bbox": [
        0.06,
        0.25,
        0.96,
        0.83
      ]
    },
    {
      "label": "pillow",
      "bbox": [
        0.17,
        0.31,
        0.32,
        0.51
      ]
    },
    {
      "label": "pillow",
      "bbox": [
        0.31,
        0.31,
        0.46,
        0.51
      ]
    },
```

Рисунок 3.2 – Приклад візуалізації на зображенні з групи A_objects



Заповіт

Як умру, то поховайте
Мене на могилі,
Серед степу широкого,
На Вкраїні милій,
Щоб лани широкополі,
І Дніпро, і кручі
Було видно, було чути,
Як реве ревучий.

Як понесе з України
У синєє море
Кров ворожу... отоді я
І лани і гори —
Все покину і полину
До самого бога
Молитися... А до того —
Я не знаю бога.

Поховайте та вставайте.
Кайдани порвіте
І вражою злою кров'ю
Волю окропіте.
І мене в сім'ї великій,
В сім'ї вольній, новій
Не забудьте пом'янути
Незлим тихим словом.

Рисунок 3.3 – Приклад розпізнаного тексту з еталоном для зображення з групи B_ocr]

```
n/python3 run_ocr.py

IMAGE: ocr_01.webp
Заповіт

Як умру, то поховайте
Мене на могилі,
Серед степу широкого,
На Вкраїні милій,
Щоб лани широкополі,
І Дніпро, і кручі
Було видно, було чути,
Як реве ревучий.

Як понесе з України
У синєє море
Кров ворожу... отоді я
І лани і гори —
Все покину і полину
До самого бога
Молитися... А до того —
Я не знаю бога.

Поховайте та вставайте.
Кайдани порвіте
І вражою злою кров'ю
Волю окропіте.
І мене в сім'ї великій,
В сім'ї вольній, новій
Не забудьте пом'янути
Незлим тихим словом.
--------------------------------
```

Рисунок 3.4 – Приклад розпізнаного тексту з еталоном для зображення з групи B_ocr]

Рисунок 3.5 – Приклад комбінованого виходу (об’єкти + текст) для зображення з групи C_both]


```
    {
      "label": "light fixture",
      "bbox": [
        134, 85, 250, 114
      ]
    },
    {
      "label": "light fixture",
      "bbox": [
        338, 85, 454, 114
      ]
    },
    {
      "label": "light fixture",
      "bbox": [
        542, 85, 658, 114
      ]
    },
    {
      "label": "light fixture",
      "bbox": [
        746, 85, 862, 114
      ]
    }
  ],
  "text": [
    "HERMÈS"
  ]
```


Рисунок 3.6 – Приклад комбінованого виходу (об’єкти + текст) для зображення з групи C_both]

Узагальнюючи, результати експерименту демонструють, що мультимодальний VL-підхід може бути застосований як універсальний інструмент аналізу зображень у режимі inference для задач детекції та OCR. Найбільш сильними сторонами є гнучкість через текстову інструкцію, здатність виконувати різні типи завдань у межах однієї моделі та можливість одразу отримувати структуровані результати. Водночас до обмежень належать залежність від промпту, варіативність деталізації класів, потенційна нестабільність формату та складність обробки дрібного тексту або складних сцен без додаткового уточнення інструкцій і пост-обробки.

3.3 Перспективи розвитку

Проведене дослідження демонструє, що сучасні мультимодальні моделі здатні виконувати широкий спектр задач комп’ютерного зору без необхідності спеціалізованого донавчання. Це відкриває перспективи створення універсальних систем аналізу зображень, які можуть адаптуватися до нових сценаріїв шляхом зміни текстової інструкції, а не шляхом повторного навчання або ручної переналаштовки.

З наукової точки зору одним із ключових напрямів подальших досліджень є вивчення впливу формулювання інструкцій на стабільність та якість результатів. Отримані у роботі результати підтверджують, що текстовий запит виступає не лише як опис завдання, але і як інструмент керування внутрішніми механізмами уваги моделі. Подальші дослідження можуть бути спрямовані на формалізацію правил побудови інструкцій та розробку методів їх автоматичної оптимізації.

Іншим перспективним напрямом є поєднання inference-only підходу з частковою адаптацією моделі до конкретної предметної галузі. Наприклад, можливим є використання невеликої кількості еталонних прикладів або контекстної інформації у текстовому запиті для підвищення точності

розпізнавання специфічних об’єктів або термінів. Такий підхід дозволив би зберегти універсальність моделі, одночасно підвищуючи її ефективність.

З прикладної точки зору результати роботи можуть бути використані при розробці інформаційних систем для автоматичного аналізу зображень у різних галузях, зокрема у сфері електронної комерції, документообігу, систем відеоспостереження та інтелектуальних помічників. Можливість отримання структурованих виходів у форматі JSON спрощує інтеграцію vision-language моделей у існуючі програмні комплекси.

Перспективним є також використання досліджуваного підходу для побудови інтерактивних систем, у яких користувач може динамічно змінювати тип аналізу зображення, задаючи нові інструкції у процесі роботи. Такі системи можуть поєднувати автоматичний аналіз з елементами пояснюваного штучного інтелекту, коли результати моделі супроводжуються текстовими поясненнями.

Окремої уваги заслуговує питання масштабування експериментів. У подальших дослідженнях доцільно розширити обсяг вхідних даних, використовуючи більші та різноманітніші набори зображень. Це дозволить перейти від демонстраційного аналізу до більш формального статистичного оцінювання та перевірки узагальнювальних властивостей моделі.

Також перспективним є порівняння мультимодального підходу з класичними конвеєрами комп’ютерного зору, які складаються з окремих моделей детекції та OCR. Таке порівняння дозволить більш чітко окреслити область доцільності використання vision-language моделей та визначити сценарії, у яких їх застосування є найбільш ефективним.

Загалом, результати роботи свідчать про значний потенціал мультимодальних моделей комп’ютерного зору як універсального інструменту аналізу візуальної інформації. Подальший розвиток цього напряму може сприяти створенню більш гнучких, масштабованих та інтерпретованих систем штучного інтелекту, здатних ефективно працювати в умовах мінімальної попередньої підготовки даних.

ВИСНОВКИ

У процесі виконання магістерської роботи було проведено комплексне дослідження сучасних методів та моделей комп’ютерного зору з акцентом на мультимодальні vision-language підходи, які дозволяють виконувати аналіз зображень у режимі zero-shot inference. Основною метою роботи було дослідити можливості використання універсальної vision-language моделі для розв’язання задач детекції об’єктів та розпізнавання тексту без спеціалізованого донавчання на цільових даних.

Для досягнення поставленої мети у роботі було виконано низку взаємопов’язаних завдань. Зокрема, здійснено огляд та систематизацію теоретичних основ комп’ютерного зору, класичних підходів до детекції об’єктів і OCR, а також сучасних трансформерних та мультимодальних архітектур. Це дозволило сформувати цілісне уявлення про еволюцію підходів і визначити місце vision-language моделей серед інших методів.

У теоретично-методичному розділі було уточнено базові поняття, формалізовано постановку задач аналізу зображень та розглянуто принципи формування вхідних і вихідних даних у мультимодальній постановці. Окрему увагу приділено питанням структурованого подання результатів та вимогам до їх коректності, що є критично важливим для практичного застосування.

У практичній частині роботи було розроблено методику експериментального дослідження, яка передбачає використання vision-language моделі у режимі inference з фіксованими інструкціями та контрольованими вхідними даними. Було сформовано репрезентативний набір зображень, що охоплює сценарії детекції об’єктів, розпізнавання тексту та їх комбінації, а також визначено формат взаємодії з моделлю та структуру отриманих відповідей.

Проведені експерименти показали, що обрана модель здатна коректно виконувати базові задачі комп’ютерного зору без попереднього донавчання.

Зокрема, для задач детекції об’єктів модель демонструє стабільне виявлення візуально виразних об’єктів, а для задач OCR – задовільну якість розпізнавання друкованого тексту за сприятливих умов зйомки. У комбінованих сценаріях підтверджено можливість одночасного отримання інформації про об’єкти сцени та текстові фрагменти.

Аналіз експериментальних результатів дозволив виявити характерні обмеження досліджуваного підходу. До них належать залежність якості результатів від формулювання інструкцій, варіативність рівня деталізації виходів, а також можливі помилки у складних сценах або за наявності дрібних об’єктів і тексту. Водночас показано, що коректне задання формату відповіді та використання структурованих виходів суттєво підвищують практичну придатність результатів.

Отримані результати підтверджують, що поставлена мета магістерської роботи була досягнута. Усі основні завдання, сформульовані на початковому етапі дослідження, були виконані, а отримані висновки узгоджуються з сучасними тенденціями розвитку комп’ютерного зору та мультимодальних моделей. Робота демонструє можливість використання vision-language моделей як універсального інструменту аналізу зображень у практичних системах.

Практична цінність роботи полягає у можливості застосування запропонованого підходу для швидкого прототипування систем аналізу візуальної інформації без необхідності складного процесу навчання моделей. Результати дослідження можуть бути використані при розробці інформаційних систем, навчальних матеріалів та подальших наукових досліджень.

Загалом, магістерська робота робить внесок у вивчення можливостей мультимодальних vision-language моделей та окреслює напрями їх подальшого розвитку і практичного застосування, що підтверджує її наукову та прикладну значущість.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Girshick R. Fast R-CNN. *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 7–13 December 2015. 2015. URL: https://doi.org/10.1109/iccv.2015.169 (date of access: 07.12.2025).

2. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks / S. Ren et al. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017. Vol. 39, no. 6. P. 1137–1149. URL: https://doi.org/10.1109/tpami.2016.2577031 (date of access: 07.12.2025).

3. Redmon J., Farhadi A. YOLOv3: An Incremental Improvement. *arXiv*. 2018. URL: https://doi.org/10.48550/arXiv.1804.02767 (date of access: 07.12.2025).

4. Shi B., Bai X., Yao C. An End-to-End Trainable Neural Network for Image-Based Sequence Recognition and Its Application to Scene Text Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017. Vol. 39, no. 11. P. 2298–2304. URL: https://doi.org/10.1109/tpami.2016.2646371 (date of access: 07.12.2025).

5. Connectionist temporal classification / A. Graves et al. *the 23rd international conference*, Pittsburgh, Pennsylvania, 25–29 June 2006. New York, New York, USA, 2006. URL: https://doi.org/10.1145/1143844.1143891 (date of access: 07.12.2025).

6. An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale / A. Dosovitskiy et al. *arXiv*. 2021. URL: https://doi.org/10.48550/arXiv.2010.11929 (date of access: 07.12.2025).

7. Learning Transferable Visual Models From Natural Language Supervision / A. Radford et al. *arXiv*. 2021. URL: https://doi.org/10.48550/arXiv.2103.00020 (date of access: 07.12.2025).

8. Flamingo: A Visual Language Model For Few-Shot Learning / J.-B. Alayrac et al. *arXiv*. 2022. URL: https://doi.org/10.48550/arXiv.2204.14198 (date of access: 07.12.2025).

9. Bai J., Qi X. Qwen-VL: A Versatile Vision-Language Model For Multimodal Understanding. 2023. URL: https://doi.org/10.48550/arXiv.2308.12966 (date of access: 07.12.2025).

10. Qwen Team. Qwen-VL: Towards General-Purpose Multimodal Intelligence. *Alibaba Research*. 2023.

11. Image Quality Assessment: From Error Visibility to Structural Similarity / Z. Wang et al. *IEEE Transactions on Image Processing*. 2004. Vol. 13, no. 4. P. 600–612. URL: https://doi.org/10.1109/tip.2003.819861 (date of access: 07.12.2025).

12. The Pascal Visual Object Classes (VOC) Challenge / M. Everingham et al. *International Journal of Computer Vision*. 2009. Vol. 88, no. 2. P. 303–338. URL: https://doi.org/10.1007/s11263-009-0275-4 (date of access: 07.12.2025).

13. Visual Instruction Tuning / H. Liu et al. *arXiv*. 2023. URL: https://doi.org/10.48550/arXiv.2304.08485 (date of access: 07.12.2025).

14. Instruction Tuning For Large Language Models / P. Zhang et al. *arXiv*. 2023. URL: https://doi.org/10.48550/arXiv.2308.10792 (date of access: 07.12.2025).

15. BLIP: Bootstrapping Language-Image Pretraining / J. Li et al. *arXiv*. 2023. URL: https://doi.org/10.48550/arXiv.2201.12086 (date of access: 07.12.2025).

16. Point-Nerf: Point-Based Neural Radiance Fields / Q. Xu et al. *arXiv*. 2022. URL: https://doi.org/10.48550/arXiv.2201.08845 (date of access: 07.12.2025).

17. Nerf: Representing Scenes As Neural Radiance Fields / B. Mildenhall et al. *arXiv*. 2020. URL: https://doi.org/10.48550/arXiv.2003.08934 (date of access: 07.12.2025).

18. Hore A., Ziou D. Image Quality Metrics: PSNR vs. SSIM. *2010 20th International Conference on Pattern Recognition (ICPR)*, Istanbul, Turkey, 23–26 August 2010. 2010. URL: https://doi.org/10.1109/icpr.2010.579 (date of access: 07.12.2025).

19. Focal Loss for Dense Object Detection / T.-Y. Lin et al. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2020. Vol. 42, no. 2. P. 318–327. URL: https://doi.org/10.1109/tpami.2018.2858826 (date of access: 07.12.2025).

20. StructGPT: A General Framework for Large Language Model to Reason over Structured Data / J. Jiang et al. *arXiv.org*. URL: https://doi.org/10.48550/arXiv.2305.09645 (date of access: 07.12.2025).

21. Qwen-VL: A Versatile Vision-Language Model / J. Bai et al. *arXiv.org*. 2023. URL: https://doi.org/10.48550/arXiv.2308.12966 (date of access: 07.12.2025).

22. CLIP: Learning Transferable Visual Models From Natural Language Supervision / A. Radford et al. *arXiv.org*. 2021. URL: https://doi.org/10.48550/arXiv.2103.00020 (date of access: 07.12.2025).

23. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models / J. Li et al. *arXiv.org*. URL: https://doi.org/10.48550/arXiv.2301.12597 (date of access: 07.12.2025).

24. GPT-4 Technical Report / OpenAI et al. *arXiv.org*. 2023. URL: https://doi.org/10.48550/arXiv.2303.08774 (date of access: 07.12.2025).

25. Inference Providers. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/docs/api-inference (date of access: 07.12.2025).

26. Models – Hugging Face. *Hugging Face – The AI community building the future.* URL: https://huggingface.co/models (date of access: 07.12.2025).