УДК 004.932:81’25

# АНАЛІЗ ПИТАННЯ ПЕРЕКЛАДУ ТЕКСТУ НА ЗОБРАЖЕННЯХ НА ОСНОВІ МУЛЬТИМОДАЛЬНИХ ВЕЛИКИХ МОВНИХ МОДЕЛЕЙ

Лазарєва Д.Є.
e-mail: daryna.lazarieva@nure.ua
Науковий керівник – канд. техн. наук, доц. Яковлева О.В.
Харківський національний університет радіоелектроніки, каф. ІНФ
м. Харків, Україна

Traditional image translation typically relies on sequential OCR and machine translation, which often fail to account for visual layouts and context. This research evaluates multimodal large language models in simultaneously processing textual and visual data to ensure accurate, context-aware translations while preserving document topology. The study demonstrates that these models automate 80–90% of visual content localisation by maintaining structural integrity across various formats. Despite these advancements, human verification remains necessary for low-quality images to prevent errors. Future work in this field will focus on technologies like visual font transfer to achieve seamless, design-consistent localisation.

У сучасному світі потреба в швидкому та точному перекладі візуальної інформації (інфографіки, технічної документації, коміксів тощо) постійно зростає. Традиційні методи перекладу тексту на зображеннях складаються з послідовного використання класичних алгоритмів комп’ютерного зору, систем оптичного розпізнавання символів (англ. Optical Character Recognition, OCR) та окремих моделей машинного перекладу і мають певні обмеження [1, 2]. Системи часто не враховують візуальне розташування тексту на зображенні, що призводить до втрати контексту, помилок у розпізнанні шрифтів, стилів та структури макетів.

Останні роки активно розвиваються мультимодальні великі мовні моделі (англ. Multimodal Large Language Models, MLLMs), які здатні одночасно обробляти текстову та візуальну інформацію. Це відкриває багато нових можливостей для швидкого, автоматичного перекладу тексту, враховуючи контекст та структуру документа.

Метою даної роботи є порівняльний аналіз ефективності провідних MLLMs, таких як GPT-4 Vision [3], Claude 3 [4] та Gemini [5]. Дослідження зосереджено на оцінюванні точності контекстуального перекладу, збереженні смислової цілісності макета та здатності моделей до просторового логічного висновку.

Для проведення аналізу було сформовано власний датасет, що складається з трьох категорій зображень: «Технічні схеми», «Інфографіка» та «Художні макети», які містять складні візуальні структури та специфічний контекст (рис. 1). Технічні схеми дозволяють оцінити здатність

моделей до точного розпізнання дрібних деталей, тоді як художні макети допомагають перевірити якість лінгвістичної адаптації та розуміння культурного контексту.

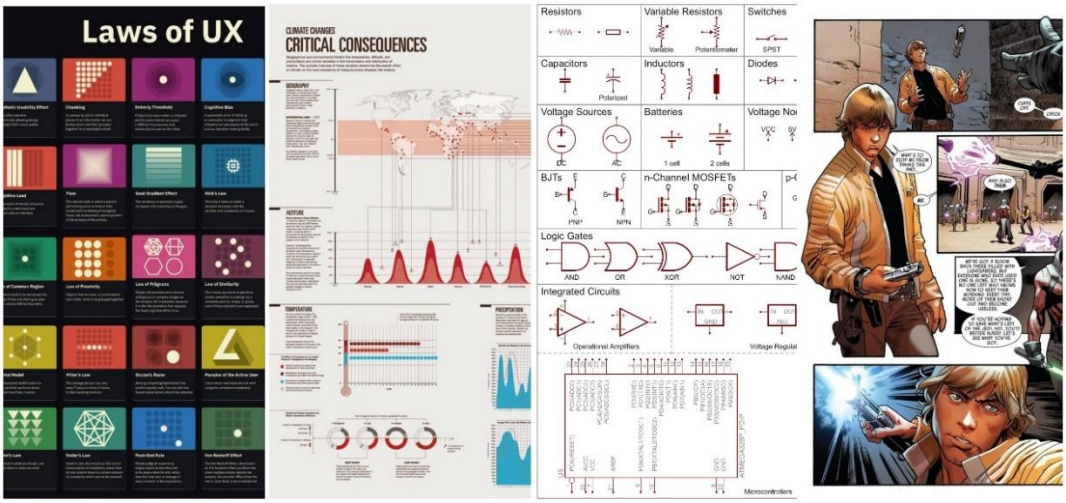


Рисунок 1 – Приклад зображень з датасету

Для кількісної оцінки та порівняння обраних мультимодальних моделей було проведено серію експериментів. Кожне зображення з категорій проходило обробку з ідентичним системним промптом для верифікації здатності моделей до контекстуального перекладу та збереження топології документа. Оцінювання базувалося на чотирьох ключових критеріях: точність OCR, просторова логіка, контекстуальність перекладу та цілісність структури. Результати проведеного дослідження візуалізовано на рисунку 2.

За результатами тестування було виявлено високі показники точності перекладу тексту у всіх досліджуваних моделей. Найбільш стабільні результати продемонструвала модель GPT-4 Vision, яка показала найвищі значення точності OCR та просторової, особливо при обробці технічних схем та інфографіки. Модель Gemini продемонструвала гарні результати у завданнях контекстуального перекладу, у той час як Claude 3 показала збалансовані результати за більшістю критеріїв, однак дещо поступається іншим моделям у задачах просторової логіки.

MLLMs досягли рівня, коли вони досить добре розуміють візуальний контент. Сьогоднішні MLLMs можуть змінювати результат перекладу залежно від контексту, добре працюють зі складними макетами та зберігають логічну структуру документа. Але проблеми все ще існують: моделі можуть помилятися, якщо є багато дрібних деталей, і можуть видавати недостовірну інформацію у випадку, коли зображення поганої якості. Зараз ця технологія дозволяє автоматизувати аналіз приблизно 80–90% візуального контенту, проте для критично важливих матеріалів все ще необхідна перевірка.

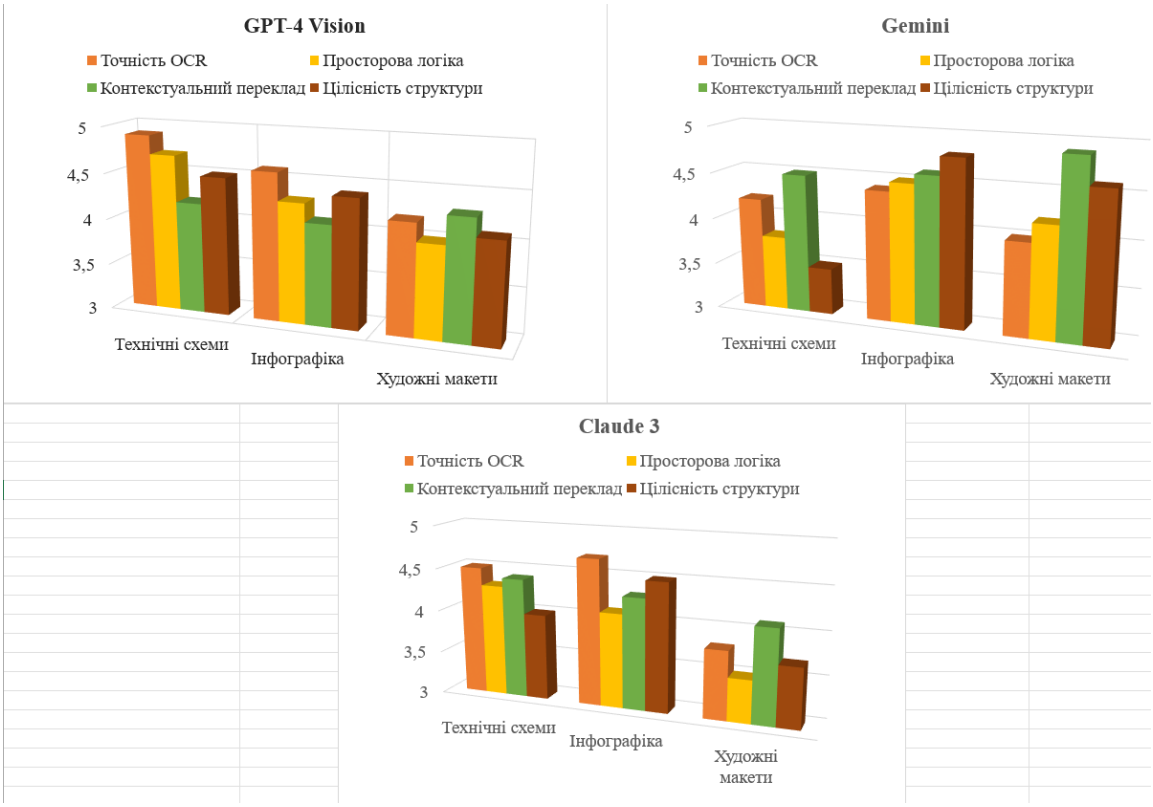


Рисунок 2 – Порівняльна характеристика моделей GPT-4 Vision, Gemini та Claude 3

У майбутньому перспективним напрямом є впровадження технології Visual Font Transfer, яка дозволить автоматично відтворювати оригінальні шрифти під час перекладу. Це зробить локалізацію більш органічною та візуально узгодженою з початковим дизайном.

Список використаних джерел:

1. Kim G., Hong T., Yim M. *OCR-Free Document Understanding Transformer (Donut).* – ECCV, 2022. – URL: https://arxiv.org/abs/2111.15664

2. Gorokhovatskyi V., Tvoroshenko I., Yakovleva O., and Hudáková M. Feature space quantisation and application of metrics in structural methods of image recognition, IEEE Access, 2026, vol. 14, pp. 17812-17824, doi: 10.1109/ACCESS.2026.3659757.

3. OpenAI. *GPT-4(Vision) System Card.* – 2023. – URL: https://openai.com/research/gpt-4v-system-card

4. Anthropic. *Claude 3 Model Card.* – 2024. – URL: https://www.anthropic.com/news/claude-3-family

5. Team Google DeepMind. *Gemini: A Family of Highly Capable Multimodal Models.* – arXiv:2312.11805, 2023. – URL: https://arxiv.org/abs/2312.11805