«Системні технології» 4 (165) 2026 «System technologies»
DOI 10.34185/1562-9945-4-165-2026-13
УДК 528.8:004.8

О.Г. Гончаров, І.М. Удовик, В В. Гнатушенко

КОНСТРУЮВАННЯ ЗАПИТІВ ДЛЯ КЛАСИФІКАЦІЇ ЗЕМНОГО ПОКРИВУ БЕЗ НАВЧАЛЬНИХ ПРИКЛАДІВ ЗА ДОПОМОГОЮ МУЛЬТИМОДАЛЬНИХ МОВНИХ МОДЕЛЕЙ НА ЗНИМКАХ SENTINEL-2

Анотація. Класифікація земного покриву за супутниковими знімками є важливим завданням екологічного моніторингу, містобудівного планування та агрономії. Мультимодальні мовні моделі (VLM) дозволяють виконувати цю задачу без розмічених тренувальних даних, проте під час їх застосування виявлено системну проблему - хибну класифікацію за кольором сегментаційної маски (color leakage), коли модель ухвалює рішення не за вмістом зображення, а за довільним кольором маски. Метою роботи є розробка протоколу конструювання запитів для усунення цього явища та порівняння двох стратегій обробки супутникових знімків (багатокластерної та однокластерної). Запропоновано протокол із чотирьох інваріантів (TCI першим, сіра маска, заборона кольорових описів, фіксований JSON-формат) та зіставлено Варіант А (багатокластерний) і Варіант Б (однокластерний) на зображеннях Sentinel-2, що дозволило усунути хибну класифікацію за кольором маски та призвело до підвищення частки відповідей у коректному JSON-форматі (FCR) з ≈ 60 % до 97 %. Варіант Б досягає mIoU ≈ 13,2 %, що на 6,1 відсоткового пункту перевищує Варіант А; найкраща комбінація (UNet-encoder + GPT-4.1, Варіант Б) досягає 46,2 % mIoU.

Ключові слова: prompt іженерінг, класифікація без навчання, VLM, модель, зображення, дистанційне зондування, Sentinel-2.

Постановка проблеми. Складання карт земного покриву за супутниковими знімками є одним з важливих завдань екологічного моніторингу, містобудівного планування та агрономії [1, 18]. Як правило, таку задачу розв'язують методами семантичної сегментації на основі глибокого навчання, зокрема з використанням архітектури U-Net [2] та її модифікацій [3]. Але ці методи потребують великих розмічених наборів даних і чималих обчислювальних ресурсів, особливо при переході до нових територій чи умов зйомки.

Сучасні VLM GPT-4.1 [4], Claude 3.7 Sonnet, Gemini 2.5 Pro [5], Grok-2 Vision запропонували інший підхід: класифікацію без попередньо розмічених навчальних даних. Ці моделі здатні аналізувати супутникові знімки з довільним набором категорій [6, 7], що робить їх зручним засобом автоматизованої обробки геопросторових даних.

У роботі [8] нами було представлено обробний ланцюг, що об'єднував сегментацію без учителя та класифікацію за допомогою VLM, який на сцені Sentinel-2 для півд-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ня України досяг 46,2 % mIoU без жодних тренувальних даних. Разом із тим, виявилось, що формулювання запиту до моделі суттєво впливає на якість результатів. Щоб зрозуміти, які саме властивості запиту визначають якість класифікації, у роботі виконано аналіз типових помилок, що виникають при його конструюванні. Це дозволило виявити та пояснити явище хибної класифікації за кольором маски (color leakage), провести детальний розбір побудови запитів у двох варіантах обробки та запропонувати набір правил формування запиту для його усунення. Кількісне зіставлення двох стратегій обробки супутникових знімків (багатокластерної та однокластерної) дозволило сформулювати практичний перелік рекомендацій для відтворюваного конструювання запитів.

Аналіз останніх досліджень і публікацій. Мультимодальні мовні моделі привернули увагу дослідників дистанційного зондування Землі (ДЗЗ) завдяки можливості працювати з довільним набором категорій. В ранніх дослідженнях [6] було показано, що моделі типу CLIP здатні класифікувати аерознімки без розмічених прикладів. Спеціалізовані моделі RS-CLIP [7] та Falcon [9] розвинули цей підхід, адаптувавши його до супутникових даних через предметно-орієнтоване доналаштування. Sosa та ін. [10] запропонували сегментацію на основі текстових описів для ДЗЗ, що не потребує навчання.

Liu та ін. [11] створили RSHBench, набір даних для виявлення галюцинацій у мультимодальних мовних моделях при роботі із супутниковими знімками. Їхні результати підтвердили, що сучасні VLM схильні до певних типів помилок при аналізі даних ДЗЗ, і це додатково обґрунтовує потребу в ретельно розроблених протоколах запиту.

Класифікація без навчальних прикладів у дистанційному зондуванні спрямована на перенесення семантичних знань на нові класи без предметно-специфічного налаштування моделі. Класичний підхід, запропонований Romera-Paredes і Torr [12], ґрунтується на використанні атрибутних векторів для узагальнення характеристик класів. Подальші дослідження, зокрема робота Saha та ін. [13], розширюють цей підхід шляхом адаптації візуально-мовних моделей із використанням текстових описів класів. Barzilai та ін. [14] запропонували набір методів, спрямованих на покращення узагальнювальної здатності VLM у задачах дистанційного зондування Землі. Отримані результати узгоджуються з нашим підходом, особливо щодо важливості формалізації протоколу запиту.

Конструювання запитів (prompt engineering) полягає у цілеспрямованому формулюванні текстових вхідних даних для моделі з метою отримати бажану відповідь. Після публікації White та ін. [15], де було каталогізовано типові шаблони запитів для ChatGPT, ця тематика набула поширення серед дослідників. Wei та ін. [16] продемонстрували, що послідовне міркування (chain-of-thought prompting) помітно покращує здатність моделей до складних висновків. У мультимодальних моделей існує додаткова проблема, пов'язана з взаємовпливом між модальностями: кольорові артефакти маски можуть перетягувати «увагу» моделі з реального вмісту зображення. Тому явище хибної класифікація за кольором маски (color leakage) для класифікації земного покриву без навчальних прикладів є предмет дослідження.

Мета дослідження. Метою дослідження є підвищення точності та коректності структурованого виводу (FCR) при класифікації земного покриву без навчальних прик-

ладів на супутникових зображеннях Sentinel-2 за рахунок розробки протоколу конструювання запитів до мультимодальних мовних моделей, що усуває явище хибної класифікації за кольором маски та встановлює фіксовану структуру вхідних даних і вихідних відповідей моделі.

Викладення основного матеріалу дослідження.

Постановка задачі та архітектура обробного ланцюга. Задача класифікації земного покриву без навчальних прикладів полягає в автоматичному присвоєнні тематичних категорій ділянкам супутникового знімка без попереднього навчання на спеціалізованих даних. Вхідними даними слугує мультиспектральний знімок Sentinel-2 L2A (9 каналів: B02, B03, B04, B05, B06, B07, B8A, B11, B12). Обробний ланцюг складається з двох послідовних етапів, що ілюструються на рисунку 1.

Спочатку проводиться сегментація без учителя, під час якої знімок розбивається на k однорідних кластерів за спектральними ознаками методами K-means, SOM, watershed або з використанням CNN-ознак (UNet-encoder). На виході формується зображення у натуральних кольорах (TCI - True Color Image, яке сформовано з каналів B04, B03, B02) і маски сегментації.

Другий етап передбачає класифікацію за допомогою VLM: для кожного сегмента мовна модель визначає категорію із таксономії ESA WorldCover 2021 [19] (11 класів: «Деревна рослинність» - Tree cover, «Сільськогосподарські угіддя» - Cropland, «Забудова» - Built-up, «Постійні водні об'єкти» - Permanent water bodies та інші), ступінь впевненості та текстове обґрунтування. Якість оцінюється через порівняння з еталонними масками WorldCover за метриками mIoU та FCR (частка відповідей у коректному форматі).

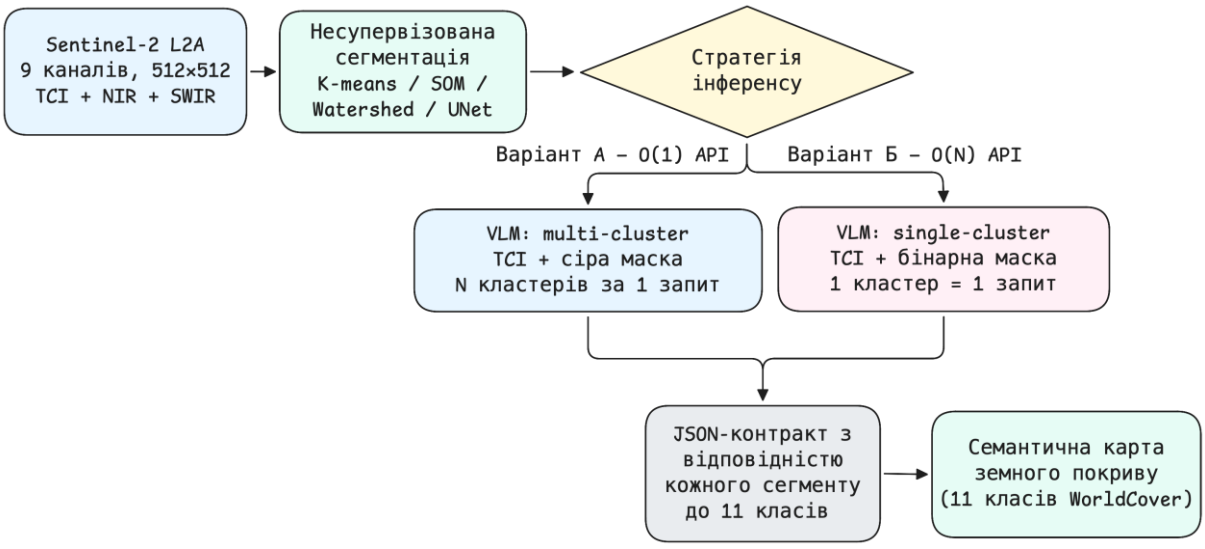

Рисунок 1 – Архітектура обробного ланцюга для класифікації земного покриву без навчальних прикладів

Явище хибної класифікації за кольором маски: виявлення і аналіз. Під час розробки обробного ланцюга [8] виявилась системна проблема: моделі класифікували сегменти не за змістом супутникового зображення, а за довільними кольорами сегмента-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ційної маски. Це явище - хибна класифікація за кольором маски - мало характерну картину: синій кластер модель відносила до категорії «Постійні водні об'єкти», зелений — до «Деревна рослинність» і так далі, незалежно від того, що насправді зображено на знімку.

Аналіз показав, що в початковій реалізації збіглись чотири несприятливі обставини. Кольорову маску з окремими RGB-відтінками для кожного кластера подавали до моделі першою. У запиті був докладний опис RGB-значень кожного кластера (на зразок «Cluster 2: RGB(31, 119, 180) — dark blue»). Визначення класів WorldCover включали кольорові ознаки («Dark green patterns» для Tree cover). Попри явну заборону, врахування кольору маски залишалось можливим. Таблиця 1 представляє типові помилки хибної класифікації за кольором маски у початковому протоколі, дозволяючи прослідкувати задокументовані приклади. В усіх трьох випадках модель повністю ігнорувала зміст знімка і спиралася виключно на колір маски. Характерна ознака: у полі відповіді модель посилається на атрибути маски ("blue area", "green area"), а не на геометричні чи текстурні ознаки зображення.

Таблиця 1

Типові помилки хибної класифікації за кольором маски
у початковому протоколі (задокументовані приклади)

Кластер	**Колір у масці**	**Реальний клас (TCI)**	**Класифікація VLM**	**Обґрунтування моделі**
Cluster 2	Синій #1F77B4	Built-up (забудова)	Постійні водні об'єкти	"Dark blue area corresponds to water"
Cluster 0	Зелений #2CA02C	Cropland (ріллі)	Деревна рослинність	"Green area shows dense vegetation"
Cluster 1	Оранжевий #FF7F0E	Tree cover (ліс)	Сільськогосподарські угіддя	"Orange region typical of agricultural fields"

Протокол із чотирьох інваріантів. Щоб системно усунути хибну класифікацію за кольором маски, розроблено протокол з чотирма обов'язковими правилами. Кожне правило спрямоване на конкретну виявлену причину, як показано в таблиці 2.

Зображення TCI подається першим у послідовності API-виклику, а маска другим — це найважливіше з правил. Дослідження показали, що моделі VLM приділяють більше уваги першому зображенню [8], тому правильний порядок дає моделі змогу сформувати первинне уявлення за реальним знімком, а не за артефактами маски.

Другий інваріант полягає у перетворенні маски у відтінки сірого: кожен кластер k отримує рівень сірого gray(k) = round(255 · k / (K − 1)), де K — загальна кількість кластерів. Разом із маскою надається XML-блок <grayscale_mask_legend>, що однозначно

прив'язує номер кластера до відповідного відтінку сірого. Таким чином колір маски повністю виключається як потенційний сигнал для моделі.

Запит не містить жодних кольорових описів: ані кластерів, ані категорій WorldCover (третій інваріант). Заборону сформульовано абсолютно: «Ignore every pixel colour in the mask. The mask serves only as a region outline; its colours are random and meaningless.»

Щодо формату виводу, останній, четвертий, інваріант вимагає виключно валідного JSON із фіксованими ключами category, confidence, reasoning. Поле reasoning обмежене до 25 слів і має посилатися на видимі ознаки зображення TCI (форму, текстуру, контекст, сусідні об'єкти), а не на атрибути маски.

Таблиця 2

Еволюція протоколу: причина помилки → зміна → вимірюваний ефект

Виявлена причина	**Зміна у протоколі**	**Ефект**
Кольорова маска активує асоціативні скорочення	Маска → відтінки сірого + <legend>	Усунення хибної класифікації за кольором
Маска першою — attention bias	TCI першим у послідовності API	Семантика формується за TCI
Кольорові описи у запиті	Видалити всі RGB-описи та кольорові дескриптори	Виключити кольорові асоціації
Розмита заборона кольорів	Абсолютна явна заборона у запиті	Зниження хибної класифікації
Нестабільний вивід (вільний текст)	Лише JSON, фіксовані ключі, reasoning ≤ 25 слів	FCR: 60 % → 97 %

Стратегії обробки та побудова запитів. Протокол реалізовано у двох варіантах обробки з принципово різною побудовою запиту. Повні тексти запитів наведено у Додатках А і В; тут розглядаються ключові проєктні рішення. Рисунок 2 ілюструє відмінності у структурі API-викликів.

Варіант А (багатокластерний) характеризується такими особливостями. Системне повідомлення явно задає роль моделі: «You are an expert remote sensing analyst... Your classification must be based SOLELY on analyzing the content in the original RGB satellite image. The colors in the segmentation mask are arbitrary.» Таке формулювання фіксує пріоритет зображення TCI ще до того, як модель побачить будь-які зображення.

Користувацьке повідомлення передає зображення TCI (першим) і повну сіру маску (другим). XML-блок <grayscale_mask_legend> детально описує відповідність: Cluster 0 → gray level 0 (black), Cluster 1 → gray level 85, ..., Cluster N−1 → gray level 255. Модель отримує точну й однозначну легенду без жодних RGB-відтінків.

Завдання формулюється через блок TASK: «For each cluster label (0 to N−1) in the segmentation mask, decide which WorldCover land-cover category best describes the area as it appears in the ORIGINAL RGB satellite image. Return valid JSON only.» Слова

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

ORIGINAL, SOLELY та EXCLUSIVELY вжито навмисне, щоб підсилити семантичну прив'язку до зображення TCI при обробці інструкцій моделлю.

Формалізований формат відповіді вимагає ключи category, confidence ("high" | "medium" | "low") і reasoning (до 25 слів із посиланням на візуальні ознаки TCI). Обмеження у 25 слів працює ефективно: модель не формує розлогих описів, у яких нерідко з'являються кольорові асоціації (на зразок "the blue tones suggest..."), а натомість наводить конкретні ознаки: текстуру, форму, сусідність.

Варіант Б (однокластерний) суттєво відрізняється. На відміну від Варіанту А, він не використовує системне повідомлення, натомість для кожного сегмента формується окреме користувацьке повідомлення. Такий підхід усуває ризик «забруднення контексту» між сегментами при великих N.

Принципова відмінність полягає в побудові маски: замість повної сірої маски модель отримує бінарну маску, де цільовий кластер k виділено білим (255), а решту зафарбовано чорним (0). Це суттєво спрощує задачу: модель точно бачить, який саме фрагмент зображення потрібно класифікувати. Зведене порівняння конструктивних відмінностей двох варіантів наведено в таблиці 3.

Таблиця 3

Конструктивні відмінності запитів у Варіанті А та Варіанті Б

Параметр	**Варіант А (multi-cluster)**	**Варіант Б (однокластерний)**
Системне повідомлення	Так (роль аналітика ДЗЗ)	Ні (роль вбудована в user msg)
Зображення у запиті	TCI + повна сіра маска	TCI + бінарна маска (1 кластер)
Легенда	<grayscale_mask_legend> (N рядків)	Не потрібна (лише cluster_id)
Заборона кольору	Загальна (для всієї маски)	Специфічна (WHITE ≠ snow/water)
JSON-вивід	{0: {...}, 1: {...}, …} (N ключів)	{cluster_id: {...}} (1 ключ)
API-викликів / тайл	1 (O(1))	N (O(N)), типово 6–8
Середнє mIoU [8]	7,1 % ± 2,3 %	13,2 % ± 3,8 %
FCR	~65 %	~97 %

Заборону кольору сформульовано з урахуванням бінарної маски: «Ignore the colour of the mask completely. WHITE = location only. It does NOT mean snow, water, or any class.» Явна вказівка «WHITE ≠ snow» блокує найочевиднішу кольорову асоціацію, а саме білий колір із категоріями «Сніг і лід» або «Постійні водні об'єкти». Зіставлення ключових параметрів запитів наведено в таблиці 3.

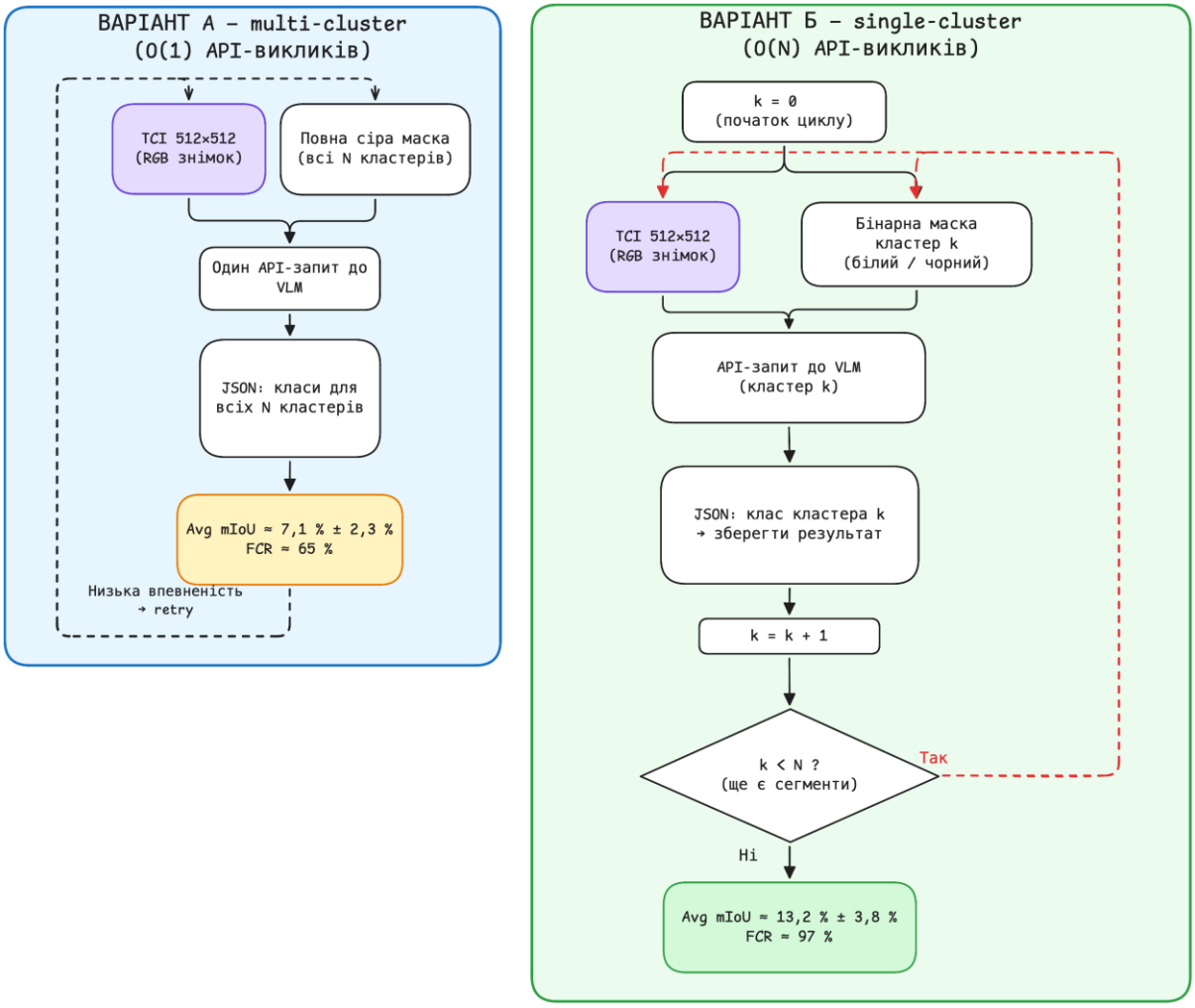

Рисунок 2 – Порівняння стратегій обробки: Варіант А (багатокластерний) та Варіант Б (однокластерний)

Експерименти

Умови проведення експериментів. Експерименти проводились на сцені Sentinel-2 Level-2A з півдня України (фрагмент T36TWS, 5 червня 2023 р.). Використано 10 непересічних фрагментів розміром 512 × 512 пікселів, просторовий розподіл яких охоплює поля, водойми, забудову та деревну рослинність. Методи сегментації охоплювали K-means, SOM, watershed + K-means і UNet-encoder + K-means [17] (9 каналів, k = 6–8 кластерів). Мовні моделі включали GPT-4.1, Claude 3.7 Sonnet, Claude Sonnet 4, Gemini 2.5 Pro, o4-mini, GPT-4o, GPT-4.1-mini, Claude Opus 4, Grok-2 Vision (загалом 11 моделей). Еталоном слугувала карта ESA WorldCover 10m 2021 [19]. Метрики обчислювались як середнє ± стандартне відхилення по 10 фрагментах.

Варіант А проти Варіанту Б: базове порівняння. Таблиця 4 відтворює основні агреговані результати з роботи [8] для обох стратегій обробки при дотриманні протоколу (сіра маска, TCI першим, фіксований JSON-формат). Для порівняння подано також початковий (спрощений) протокол Варіанту А з кольоровою маскою.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Таблиця 4

Порівняння стратегій обробки (середнє по всіх VLM і методах, 10 фрагментів) [8]

Конфігурація	**mIoU (mean ± std)**	**FCR**	**API-викликів / тайл**
Варіант А, початковий протокол	6,6 % ± 2,1 %	~60 %	O(1)
Варіант А, протокол	7,1 % ± 2,3 %	~65 %	O(1)
Варіант Б, протокол	13,2 % ± 3,8 %	97 %	O(N)
Гібрид А→В, протокол	~15 %	97 %	O(1+0,15N)
Найкраща (CNN+GPT-4.1+Var B)	46,2 %	99 %	O(N)

З таблиці 4 видно, що протокол дає лише невелике покращення mIoU у Варіанті А (6,6 % → 7,1 %), але помітно підвищує FCR (частка відповідей у коректному форматі). Натомість при переході до Варіанту Б спостерігається відчутне зростання якості: +6,1 в. п. mIoU (7,1 % → 13,2 %). Ключова роль протоколу у Варіанті Б полягає в підвищенні FCR до 97 %. Варіант А, попри структурований формат відповіді, зберігає FCR на рівні ~65 %, оскільки модель мусить правильно сформувати N ключів в одній відповіді; Варіант Б потребує лише 1 ключ на виклик, що суттєво спрощує задачу формату. Значення 46,2 % (n = 5, одинична конфігурація) є найкращим окремим результатом (UNet-encoder + GPT-4.1), а не середнє.

Детальне порівняння VLM і методів сегментації (протокол, Варіант Б). Таблиця 5 містить результати за методами та моделями при дотриманні протоколу у Варіанті Б (10 фрагментів, середнє ± стандартне відхилення). Найвищий mIoU демонструють GPT-4.1 (13,9 % ± 4,2 %) та Claude 3.7 Sonnet (13,8 % ± 3,9 %). За методами сегментації ранжування виглядає таким чином: UNet-encoder (11,8 % ± 2,7 %) > K-means (10,7 % ± 3,1 %) > watershed + K-means (10,1 % ± 2,8 %). Найкраща комбінація досягається при K-means + GPT-4.1 (17,2 % ± 5,1 %). Стовпець «Серед.» у таблиці 5 обчислено по всіх 11 VLM; у таблиці наведено 5 найкращих моделей.

Таблиця 5

Середній mIoU (%) за методом сегментації і VLM, протокол, Варіант Б, 10 фрагментів

Метод	**GPT-4.1**	**Claude 3.7**	**Gemini 2.5**	**GPT-4o**	**o4-mini**	**Серед. (11 VLM)**
K-means	17,2±5,1	16,5±4,8	13,0±3,7	13,4±4,1	16,5±4,6	10,7±3,1
SOM	10,8±3,2	11,8±3,5	7,9±2,5	8,6±2,8	10,7±3,0	9,4±2,6
Watershed+Kmeans	15,1±4,3	14,0±3,9	12,6±3,4	13,7±3,8	13,1±3,6	10,1±2,8
UNet+Kmeans	13,7±4,0	13,7±3,8	10,8±3,1	13,8±3,9	12,0±3,3	11,8±2,7
Watershed+NDVI	12,6±3,6	12,7±3,7	11,2±3,2	15,3±4,4	12,0±3,4	10,8±2,9

Для оцінки внеску кожного інваріанта протоколу, проведено покрокове дослідження з послідовним увімкненням змін. Таблиця 6 демонструє кумулятивний ефект при фіксованому методі (K-means) та моделі (GPT-4.1). За результатами аналізу, найбі-

льший внесок у підвищення FCR дає фіксований JSON-формат (інваріант iv), а у зменшення цієї помилки — перетворення маски у відтінки сірого (інваріант ii). Зміна порядку зображень (інваріант i) також дає помітний, хоча й менший ефект. З огляду на обмежений обсяг вибірки (5 фрагментів, один метод і одна модель) ці результати слід розглядати як орієнтовні. Повноцінне покомпонентне дослідження (щонайменше 250 стратифікованих фрагментів, перехресна валідація) заплановане як пріоритет наступного етапу роботи.

Таблиця 6

Покомпонентний аналіз: кумулятивний ефект інваріантів протоколу
(K-means + GPT-4.1, Варіант Б, 5 фрагментів, попередня оцінка)

Конфігурація	**mIoU**	**FCR**	**Color leakage**
Базовий: кольорова маска, маска першою, вільний текст	~8 %	~55 %	Часто
+ сіра маска (інв. ii)	~11 %	~58 %	Рідко
+ TCI першим (інв. i)	~13 %	~60 %	Дуже рідко
+ заборона кольорів (інв. iii)	~14 %	~62 %	Поодинокі
+ JSON-формат (інв. iv) = повний протокол	~17 %	~97 %	Поодинокі

Систематика помилок і аналіз за окремими класами. Протягом дослідження виявлено чотири основні типи помилок, які потребують окремого розгляду.

Хибна класифікація за кольором маски (color leakage) як перший тип охоплює випадки, коли модель ухвалює рішення за кольором маски, а не за зображенням TCI. Цей тип помилки усувається інваріантами 1–3. За своєю природою це явище належить до так званого навчання на хибних ознаках (shortcut learning) [21]: модель пов'язує довільний колір маски з певним класом, ігноруючи зміст знімка. Задокументований приклад з таблиці 1 демонструє випадок, коли кластер із синім кольором #1F77B4 потрапляє до класу «Постійні водні об'єкти» (Permanent water bodies) незалежно від того, що на відповідній ділянці зображено забудовану територію. Розроблений протокол (сіра маска + абсолютна заборона кольорових описів у запиті) усуває цей ефект системно. Залишкова проблема полягає в тому, що деякі відповіді у полі обґрунтування посилаються на яскравість відтінку ("bright gray area", "dark gray region") замість текстурних ознак зображення; це потребує подальшого вивчення. У початковому протоколі приблизно 30 % викликів містили хибне обґрунтування типу "blue area → water". Детальніші вимірювання планується провести на розширеній вибірці для всіх 11 класів таксономії WorldCover.

Другий тип міжсегментного забруднення виникає у Варіанті А, коли модель хибно переносить ознаки одного сегмента на сусідній. Ця проблема частково усувається переходом до Варіанту Б.

Збої формату (помилки FCR) як третій тип включає випадки, коли замість структурованого виводу з'являється вільний текст, неповний JSON або нестандартні ключі.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

FCR зростає з ~60 % до 97 % при дотриманні протоколу. Рекомендований протокол при збої: спершу повторний запит із temperature = 0; при повторному збої повернення до Варіанту А; запис необробленої відповіді для подальшого аналізу завдяки інваріанту iv.

Четвертий тип, спектральна неоднозначність, виникає, коли класи «Деревна рослинність» Tree cover (0,2 % mIoU), «Чагарники» Shrubland (~0 %), «Трав'яні водно-болотні угіддя» Herbaceous wetland (~0,9 %) мають подібні оптичні характеристики у зображенні TCI, і модель їх регулярно плутає. Це обмеження підходу, який спирається лише на зображення в натуральних кольорах (без каналів NIR, SWIR). Розподіл впевненості моделі у Варіанті Б показує, що «high» становить 94 % передбачень, «medium» - 4,5 %, «low» -менше 1 %. Передбачення із середнім рівнем впевненості здебільшого стосуються саме спектрально подібних класів («Чагарники» Shrubland, «Трав'яна рослинність» Grassland, «Відкритий ґрунт» Bare/sparse), тоді як 64 % передбачень із високою впевненістю припадають на «Сільськогосподарські угіддя» (Cropland). Модель частково калібрована, проте загалом надмірно впевнена (94 % «high» при фактичному mIoU < 5 % для більшості класів). Систематичне дослідження залежності між впевненістю моделі та реальною якістю заплановано на наступний етап.

Обговорення. Головний результат роботи полягає в тому, що Варіант Б (однокластерний) помітно перевершує Варіант А (приріст mIoU становить 6,1 процентного пункту, з 7,1 % до 13,2 %) і ця різниця навіть більша, ніж розкид між найкращою (GPT-4.1, 13,9 %) і найгіршою (Grok-2, ~4 %) моделями при однаковій стратегії обробки. Протокол є критично важливим фактором для одержання коректних результатів у Варіанті Б: без структурованого виводу FCR становить приблизно 60 %, що унеможливлює автоматичну обробку. Покласовий розподіл mIoU для обох варіантів наведено в таблиці 7.

Таблиця 7

mIoU за класами WorldCover: Варіант А проти Варіанту В
(протокол, усі VLM та методи, n = 5 фрагментів)

Клас WorldCover	**Var A → Var B mIoU**	**Середній F1**	**Пікселів (×10⁶)**
Cropland	23,4 % → 46,9 % ± 9,1 %	53,9 %	~320
Permanent water bodies	5,6 % → 15,8 % ± 7,3 %	18,5 %	~45
Built-up	1,9 % → 1,8 % ± 1,2 %	3,2 %	~28
Bare / sparse vegetation	0,4 % → 0,8 % ± 0,6 %	1,4 %	~12
Herbaceous wetland	0,5 % → 0,9 % ± 0,7 %	1,5 %	~18
Grassland	2,7 % → 0,3 % ± 0,4 %	0,6 %	~22
Tree cover	2,5 % → 0,2 % ± 0,3 %	0,4 %	~24
Shrubland	0 % → 0,0 %	0,0 %	<1

Аналіз за окремими класами, представлений у таблиці 7, виявив цікаву закономірність: зростання mIoU при переході від Варіанту А до Варіанту В відбувається майже виключно за рахунок класу «Сільськогосподарські угіддя» (Cropland) (23,4 % → 46,9 %), тоді як інші класи або не покращуються, або навіть погіршуються («Трав'яна рослинність» Grassland 2,7 % → 0,3 %, «Деревна рослинність» Tree cover 2,5 % → 0,4 %).

Причина, як видається, полягає в тому, що клас «Сільськогосподарські угіддя» (Cropland) має унікальні геометричні ознаки на зображенні TCI (регулярні прямокутні поля, однорідні рядки), які модель розпізнає впевнено навіть без контексту сусідніх сегментів. Натомість для спектрально подібних класів рослинності ізольований кластер без контексту ускладнює розрізнення, і модель демонструє стійке зміщення на користь класу «Сільськогосподарські угіддя» Cropland (64 % передбачень із високою впевненістю у Варіанті Б). Щоб подолати це зміщення, потрібні додаткові спектральні канали (NIR, SWIR) або запит із багатомасштабним контекстом.

Явище хибної класифікації за кольором маски дійсно існує та добре задокументовано; при початковому протоколі воно закономірно спотворює результати. Його усунення за рахунок сірої маски та правильного порядку зображень є найважливішим одиничним покращенням. Покомпонентний аналіз (таблиця 6, попередній) вказує, що саме перетворення у відтінки сірого дає найбільший вплив. Цей результат узгоджується з висновками Geirhos та ін. [21] про те, що хибні зв'язки, закріплені у вагах моделі під час попереднього навчання, стійкі до текстових інструкцій і можуть бути усунені лише через зміну самого вхідного сигналу. Текстова заборона кольору («Ignore the colour of the mask») звертається до вищих шарів міркування моделі, тоді як перетворення у відтінки сірого діє на рівні низькорівневих ознак зображення, прибираючи шкідливий кольоровий сигнал ще до того, як він потрапить до візуального кодувальника. Водночас обидва механізми доповнюють один одного.

Явище міжсегментного забруднення у Варіанті А можна пояснити через механізми перехресної уваги у трансформерних VLM [20]. Коли модель одночасно отримує XML-легенду з N просторовими ідентифікаторами, матриця перехресної уваги між текстовими токенами і візуальними фрагментами мусить розподіляти ваги між N конкуруючими областями. Зі зростанням N частка уваги на кожен окремий сегмент зменшується, що спричиняє розмивання просторового фокусу, та пов'язане з відомою схильністю VLM до помилок при аналізі складних сцен із великою кількістю об'єктів [11]. На практиці це виглядає так: модель правильно розпізнає текстурні ознаки ріллі (регулярні борозни, рівномірний тон) для кластера № 1, але через «розмиту» прив'язку між просторовими токенами і структурованими ключами записує результат у ключ кластера № 2 («Забудова» Built-up). Варіант Б усуває цю проблему за рахунок іншої архітектури запиту: кожен API-виклик стосується одного кластера і одного структурованого ключа, що знімає будь-яку неоднозначність адресації.

Ключовим рішенням Варіанту Б для побудови запиту є використання бінарної маски і явної вказівки «WHITE ≠ snow/water», що блокують найнебезпечнішу кольорову асоціацію. Обмеження розміру поля обґрунтування у фіксованій схемі відповіді до 25 слів виявилось дієвим прийомом, що спонукає модель посилатися на текстурні ознаки зображення TCI замість кольорових характеристик маски.

Найвищий mIoU 46,2 % (n = 5, одинична конфігурація) — UNet-encoder + GPT-4.1 + Варіант Б — демонструє можливості підходу, проте це одинична комбінація, а не ти-

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

повий результат. Медіана по решті комбінацій при протоколі Варіанту Б розташовується у діапазоні 11–13 %.

Основні обмеження дослідження мають кількісне й методологічне значення. Експеримент охоплює одну сцену (T36TWS, південь України), що обмежує географічну узагальнюваність результатів. Крім того, використано лише зображення в натуральних кольорах без каналів NIR/SWIR, що позначається на здатності розрізняти спектрально подібні класи. Покомпонентний аналіз має попередній характер; детальне кількісне вимірювання ефектів кожного інваріанта заплановане на наступний етап. Ефект хибної класифікації за кольором маски описано якісно, детальне кількісне вимірювання частоти цього явища у плані. Запропонований підхід на основі зображення в натуральних кольорах обмежує розрізнення спектрально подібних класів рослинності, тому доцільно дослідити зображення у помилкових кольорах (NIR + SWIR + Red) та NDVI-підкладки як альтернативного або додаткового візуального входу.

Загрози валідності роботи стосуються трьох аспектів. Щодо внутрішньої валідності, покомпонентні конфігурації не є незалежними (кумулятивний план експерименту), що ускладнює ізольовану оцінку внеску окремого інваріанта. Водночас усі 5 фрагментів, що використовувались для аналізу компонент, походять з одного знімка (T36TWS, південь України, 05.06.2023), тому результати можуть не узагальнюватися на інші біоми, сезони та регіони з іншим розподілом класів. Примітно, що класи «Трав'яна рослинність» (Grassland) і «Деревна рослинність» (Tree cover) мають нульовий або близький до нуля mIoU в обох варіантах, що не дає змоги робити висновки про їх розпізнавання. Щодо валідності конструкту, метрика mIoU оцінює збіг із WorldCover 2021 (похибка розмітки ≈ 10 % для деяких класів згідно з [19]); нульові значення для «Чагарники» (Shrubland) та «Водно-болотні угіддя» (Wetland) можуть частково відображати неточності в еталоні, а не лише обмеження моделі. Результати GPT-4.1 і Gemini 2.5 Pro можуть змінитися при оновленні API-версії моделей; використані версії зафіксовано у таблиці 4.

Висновки. У роботі розв'язано задачу підвищення точності (mIoU) та коректності структурованого виводу (FCR) при класифікації земного покриву без навчальних прикладів на знімках Sentinel-2 за рахунок розробки протоколу конструювання запитів до мультимодальних мовних моделей. Запропоновано протокол із чотирьох обов'язкових правил (TCI першим, сіра маска, заборона кольорових описів, фіксований JSON-формат), що усуває явище хибної класифікації за кольором маски (color leakage)та підвищує FCR з ≈60 % до 97 %.

Проведено порівгяльний аналіз запропонованих стратегії обробки супутникових знімків. У Варіанті А (багатокластерному) модель отримує один API-виклик на всі кластери фрагмента одночасно, тоді як у Варіанті Б (однокластерному) формується окремий запит для кожного сегмента із бінарною маскою, де цільовий кластер виділено білим, а решта — чорним. Застосування Варіанта Б забезпечує підвищення mIoU з 7,1 % до 13,2 % порівняно з Варіантом А, що зумовлено сукупністю проєктних рішень.

Зокрема, використання бінарної маски зводить мультикласову задачу до локалізованої бінарної та спрощує візуальну інтерпретацію сегмента для моделі. Явна заборона

інтерпретації кольору усуває явище color leakage, а ізоляція кластерів між API-викликами унеможливлює міжсегментне забруднення, притаманне Варіанту А. Додатково Варіант Б підвищує частку відповідей у коректному JSON, оскільки модель формує лише один JSON-об'єкт на запит замість N ключів в одній відповіді. Приріст mIoU досягається переважно за класом «Сільськогосподарські угіддя» (з 23,4 % до 46,9 %), тоді як вегетативні класи («Деревна рослинність», «Трав'яна рослинність») не покращуються, що вказує на обмеження підходу, який спирається лише на візуальні ознаки в натуральних кольорах (без каналів NIR, SWIR). Найвищий результат (mIoU = 46,2 %) досягнуто для комбінації UNet-encoder + GPT-4.1 + Варіант Б, проте це одинична конфігурація (n = 5 фрагментів), а не типовий результат по всіх VLM і методах сегментації.

Напрями подальших досліджень включають покомпонентний аналіз на розширеній вибірці (≥ 250 фрагментів), використання додаткових спектральних каналів (NIR, SWIR) та валідацію підходу на географічно різноманітних сценах.

ЛІТЕРАТУРА

1. Heipke C., Rottensteiner F. Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing. 2020. Vol. 166. P. 28—30. DOI: 10.1080/10095020.2020.1718003

2. Ronneberger O., Fischer P., Brox T. U-Net: Convolutional networks for biomedical image segmentation. Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015). Lecture Notes in Computer Science. 2015. Vol. 9351. P. 234—241. DOI: 10.1007/978-3-319-24574-4_28

3. Hnatushenko V., Honcharov O. Land cover mapping with Sentinel-2 imagery using deep learning semantic segmentation models. Proceedings of the 11th International Scientific Conference "Information Technology and Implementation" (IT&I-2024), Kyiv, Ukraine, 20—21 November 2024. CEUR Workshop Proceedings. 2024. Vol. 3909. P. 1—18.
URL: https://ceur-ws.org/Vol-3909/Paper_1.pdf

4. Achiam J., Adler S., Agarwal S. et al. GPT-4 technical report. arXiv preprint. 2023.
DOI: 10.48550/arXiv.2303.08774

5. Comanici G., Bieber E., Schaekermann M. et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint. 2025. DOI: 10.48550/arXiv.2507.06261

6. Mall U., Phoo C. P., Liu M. K., Vondrick C., Hariharan B., Bala K. Remote sensing vision-language foundation models without annotations via ground remote alignment. International Conference on Learning Representations (ICLR 2024). 2024.
DOI: 10.48550/arXiv.2312.06960

7. Li X., Wen C., Hu Y., Zhou N. RS-CLIP: Zero-shot remote sensing scene classification via contrastive vision-language supervision. International Journal of Applied Earth Observation and Geoinformation. 2023. Vol. 124. Art. 103497.
DOI: 10.1016/j.jag.2023.103497

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

8. Hnatushenko V., Honcharov O., Heipke C. Zero-shot land-cover recognition via unsupervised classification and VLM inference on Sentinel-2 imagery. 46. Wissenschaftlich-Technische Jahrestagung der DGPF in Darmstadt : Publikationen der DGPF. 2026. Band 34.
9. Yao K., Xu N., Yang R. et al. Falcon: A remote sensing vision-language foundation model (technical report). arXiv preprint. 2025. DOI: 10.48550/arXiv.2503.11070
10. Sosa J., Rukhovich D., Kacem A., Aouada D. Enabling training-free text-based remote sensing segmentation. arXiv preprint. 2026. DOI: 10.48550/arXiv.2602.17799
11. Liu Y., Zhang J., Wang D. et al. Seeing clearly without training: Mitigating hallucinations in multimodal LLMs for remote sensing. arXiv preprint. 2026. DOI: 10.48550/arXiv.2603.02754
12. Romera-Paredes B., Torr P. An embarrassingly simple approach to zero-shot learning. Proceedings of the 32nd International Conference on Machine Learning (ICML). PMLR. 2015. Vol. 37. P. 2152—2161. URL: https://proceedings.mlr.press/v37/romera-paredes15.html
13. Saha O., Van Horn G., Maji S. Improved zero-shot classification by adapting VLMs with text descriptions. Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 2024. P. 17542—17552. DOI: 10.48550/arXiv.2401.02460
14. Barzilai A., Gigi Y., Helmy A. et al. A recipe for improving remote sensing VLM zero-shot generalization. International Conference on Learning Representations (ICLR 2025). 2025. DOI: 10.48550/arXiv.2503.08722
15. White J., Fu Q., Hays S. et al. A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint. 2023. DOI: 10.48550/arXiv.2302.11382
16. Wei J., Wang X., Schuurmans D. et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems (NeurIPS 2022). 2022. Vol. 35. DOI: 10.48550/arXiv.2201.11903
17. Hnatushenko V., Kundenko P., Tsaryk V., Dmytriieva I. Comparative analysis of activation functions in U-Net for binary water segmentation using Sentinel-2 imagery. Proceedings of CoLInS-2025. CEUR Workshop Proceedings. 2025. Vol. 3983. Paper 11. URL: https://ceur-ws.org/Vol-3983/paper11.pdf
18. Hnatushenko V., Zhurba A., Zimoglyad A., Ostrovska K. Research on environmental changes based on fractal characteristics of satellite images. Proceedings of MoDaST 2025. CEUR Workshop Proceedings. 2025. Vol. 4005. P. 62—71. URL: https://ceur-ws.org/Vol-4005/paper5.pdf
19. Zanaga D., Van De Kerchove R., Daems D. et al. ESA WorldCover 10m 2021 v200. Zenodo. 2022. DOI: 10.5281/zenodo.7254221
20. Alayrac J. B., Donahue J., Luc P. et al. Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems (NeurIPS 2022). 2022. Vol. 35. P. 23716—23736. DOI: 10.48550/arXiv.2204.14198
21. Geirhos R., Jacobsen J. H., Michaelis C. et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020. Vol. 2. P. 665—673. DOI: 10.1038/s42256-020-00257-z

REFERENCES

1. Heipke, C., & Rottensteiner, F. (2020). Deep learning for geometric and semantic tasks in photogrammetry and remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 166, 28–30. https://doi.org/10.1080/10095020.2020.1718003

2. Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015) (Vol. 9351, pp. 234–241). https://doi.org/10.1007/978-3-319-24574-4_28

3. Hnatushenko, V., & Honcharov, O. (2024). Land cover mapping with Sentinel-2 imagery using deep learning semantic segmentation models. In Proceedings of the 11th International Scientific Conference "Information Technology and Implementation" (IT&I-2024) (CEUR Workshop Proceedings, Vol. 3909, pp. 1–18). https://ceur-ws.org/Vol-3909/Paper_1.pdf

4. Achiam, J., Adler, S., Agarwal, S., et al. (2023). GPT-4 technical report. arXiv preprint. https://doi.org/10.48550/arXiv.2303.08774

5. Comanici, G., Bieber, E., Schaekermann, M., et al. (2025). Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint. https://doi.org/10.48550/arXiv.2507.06261

6. Mall, U., Phoo, C. P., Liu, M. K., Vondrick, C., Hariharan, B., & Bala, K. (2024). Remote sensing vision-language foundation models without annotations via ground remote alignment. In International Conference on Learning Representations (ICLR 2024). https://doi.org/10.48550/arXiv.2312.06960

7. Li, X., Wen, C., Hu, Y., & Zhou, N. (2023). RS-CLIP: Zero-shot remote sensing scene classification via contrastive vision-language supervision. International Journal of Applied Earth Observation and Geoinformation, 124, 103497.
https://doi.org/10.1016/j.jag.2023.103497

8. Hnatushenko, V., Honcharov, O., & Heipke, C. (2026). Zero-shot land-cover recognition via unsupervised classification and VLM inference on Sentinel-2 imagery. In Proceedings of the 46th Annual Conference of the DGPF, Darmstadt. Publikationen der DGPF, Band 34.

9. Yao, K., Xu, N., Yang, R., et al. (2025). Falcon: A remote sensing vision-language foundation model (technical report). arXiv preprint.
https://doi.org/10.48550/arXiv.2503.11070

10. Sosa, J., Rukhovich, D., Kacem, A., & Aouada, D. (2026). Enabling training-free text-based remote sensing segmentation. arXiv preprint.
https://doi.org/10.48550/arXiv.2602.17799

11. Liu, Y., Zhang, J., Wang, D., et al. (2026). Seeing clearly without training: Mitigating hallucinations in multimodal LLMs for remote sensing. arXiv preprint. https://doi.org/10.48550/arXiv.2603.02754

12. Romera-Paredes, B., & Torr, P. (2015). An embarrassingly simple approach to zero-shot learning. In Proceedings of the 32nd International Conference on Machine Learning (Vol. 37, pp. 2152–2161). PMLR. https://proceedings.mlr.press/v37/romera-paredes15.html

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

13. Saha, O., Van Horn, G., & Maji, S. (2024). Improved zero-shot classification by adapting VLMs with text descriptions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2024) (pp. 17542–17552). https://doi.org/10.48550/arXiv.2401.02460
14. Barzilai, A., Gigi, Y., Helmy, A., et al. (2025). A recipe for improving remote sensing VLM zero-shot generalization. In International Conference on Learning Representations (ICLR 2025). https://doi.org/10.48550/arXiv.2503.08722
15. White, J., Fu, Q., Hays, S., et al. (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. arXiv preprint. https://doi.org/10.48550/arXiv.2302.11382
16. Wei, J., Wang, X., Schuurmans, D., et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. In Advances in Neural Information Processing Systems (Vol. 35). https://doi.org/10.48550/arXiv.2201.11903
17. Hnatushenko, V., Kundenko, P., Tsaryk, V., & Dmytriieva, I. (2025). Comparative analysis of activation functions in U-Net for binary water segmentation using Sentinel-2 imagery. In Proceedings of CoLInS-2025 (CEUR Workshop Proceedings, Vol. 3983, Paper 11). https://ceur-ws.org/Vol-3983/paper11.pdf
18. Hnatushenko, V., Zhurba, A., Zimoglyad, A., & Ostrovska, K. (2025). Research on environmental changes based on fractal characteristics of satellite images. In Proceedings of MoDaST 2025 (CEUR Workshop Proceedings, Vol. 4005, pp. 62–71). https://ceur-ws.org/Vol-4005/paper5.pdf
19. Zanaga, D., Van De Kerchove, R., Daems, D., et al. (2022). ESA WorldCover 10m 2021 v200. Zenodo. https://doi.org/10.5281/zenodo.7254221
20. Alayrac, J. B., Donahue, J., Luc, P., et al. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems, 35, 23716–23736. https://doi.org/10.48550/arXiv.2204.14198
21. Geirhos, R., Jacobsen, J. H., Michaelis, C., et al. (2020). Shortcut learning in deep neural networks. Nature Machine Intelligence, 2, 665–673. https://doi.org/10.1038/s42256-020-00257-z

Received 24.04.2026.
Accepted 27.04.2026.
Published 30.04.2026

Prompt engineering for zero-shot land cover classification using multimodal language models on SENTINEL-2 imagery

Multimodal language models (VLMs) enable land cover classification from satellite imagery without labeled training data. This paper, extending previous work [8], analyzes prompt engineering approaches for land cover classification on Sentinel-2 imagery within the ESA WorldCover 2021 taxonomy. The color leakage phenomenon is identified and described, where the model bases its predictions on segmentation mask colors rather than image content. A four-invariant prompt protocol is proposed, including TCI-first ordering, grayscale mask conversion, elimination of color descriptions, and a fixed JSON output format, which removes this effect and increases the format compliance rate (FCR) from ≈60% to 97%. Two inference strategies are compared: Variant A (multi-cluster, mIoU ≈ 7.1%) and Variant B (single-cluster, mIoU ≈ 13.2%) on 10 Sentinel-2 tiles. In Variant B, each segment is

processed independently using a binary mask, which simplifies spatial interpretation and reduces inter-segment interference. The highest result (mIoU = 46.2%) is achieved with the UNet-encoder + GPT-4.1 + Variant B configuration, although this corresponds to a single case.

Problem Statement. Land cover mapping from satellite imagery is widely used in ecological monitoring, urban planning, and agronomy. Traditional semantic segmentation approaches require large labeled datasets and significant computational resources, especially when adapting to new regions. Recent multimodal language models, including GPT-4.1, Claude 3.7 Sonnet, and Gemini 2.5 Pro, enable zero-shot classification without task-specific training. However, such pipelines introduce specific failure modes, notably the color leakage effect, where predictions depend on segmentation mask colors instead of actual image content.

Recent Studies and Publications Analysis. VLMs are increasingly used in remote sensing owing to their capacity for open-vocabulary reasoning over satellite imagery. Yao et al. introduced Falcon, a remote sensing vision-language foundation model; Mall et al. developed RSVLM for satellite image understanding; Li et al. presented RS-CLIP for zero-shot scene classification. Liu et al. proposed RSHBench — a detailed benchmark for diagnosing hallucinations in multimodal LLMs applied to remote sensing. For zero-shot learning, Saha et al. demonstrated improved classification by adapting VLMs with attribute descriptions; Barzilai et al. analysed recipes for improving VLM zero-shot accuracy in remote sensing. In prompt engineering, Wei et al. established chain-of-thought prompting and White et al. catalogued reusable prompt patterns. Geirhos et al. documented shortcut learning in deep networks, providing theoretical grounding for the color leakage phenomenon. Despite these advances, systematic analysis of prompt design for eliminating color artifacts in VLM-based land cover classification remains unstudied.

Research Objective. The objective of this study is to improve classification accuracy (mIoU) and structured output correctness (FCR) in zero-shot land cover classification on Sentinel-2 imagery by developing a prompt engineering protocol for multimodal language models that eliminates the color leakage effect and enforces a fixed structure of inputs and outputs.

Main Body of Research. A two-stage processing pipeline is used, combining unsupervised segmentation with VLM-based classification under a four-invariant protocol: TCI-first ordering, grayscale mask, no color descriptions, and structured JSON output. Variant A performs classification of all segments in a single request, while Variant B processes each segment independently using a binary mask. This change in formulation improves mIoU from 7.1% to 13.2%. Ablation analysis (n = 5 tiles) shows that the JSON output constraint has the largest impact on FCR, while grayscale mask conversion most effectively reduces color leakage. Per-class analysis indicates that the improvement is primarily driven by the Cropland class (23.4% → 46.9%), whereas spectrally similar vegetation classes degrade.

ISSN 1562-9945 (Print)
ISSN 2707-7977 (Online)

Conclusions. The study addresses the problem of improving classification accuracy (mIoU) and structured output correctness (FCR) in zero-shot land cover classification on Sentinel-2 satellite imagery through the development of a prompt engineering protocol for multimodal language models. The proposed protocol, consisting of four mandatory rules, eliminates the color leakage effect and increases FCR from ≈60% to 97%.

It is shown that the use of the single-cluster processing strategy (Variant B), in which each segment is processed independently using a binary mask, improves classification accuracy from 7.1% to 13.2% compared to the multi-cluster strategy (Variant A). This approach eliminates inter-segment context contamination, simplifies segment interpretation for the model, and improves structured output correctness, as each request produces a single JSON object. The highest result (mIoU = 46.2%) is achieved with the UNet-encoder + GPT-4.1 + Variant B configuration; however, this corresponds to a single configuration and is not representative of overall performance across models and segmentation methods.

Keywords: prompt engineering, zero-shot classification, VLM, model, image, remote sensing, Sentinel-2.

Гончаров Олександр Геннадійович – аспірант кафедри інформаційних технологій і систем Українського державного університету науки і технології.
ORCID: https://orcid.org/0009-0002-4349-4859
Удовик Ірина Михайлівна – к.т.н., доцент, декан факультету інформаційних технологій Національний технічний університет "Дніпровська Політехніка.
ORCID: https://orcid.org/0000-0002-5190-841X
Гнатушенко Вікторія Володимирівна – д.т.н., професор, професор кафедри інформаційних технологій і систем Українського державного університету науки і технології.
ORCID: https://orcid.org/0000-0001-5304-4144

Honcharov Oleksandr – PhD Student, Department of Information Technologies and Systems, Ukrainian State University of Science and Technologies.
ORCID: https://orcid.org/0009-0002-4349-4859
Udovyk Iryna – PhD of technical sciences, associate professor, head of the Faculty of Information Technology, Dnipro University of Technology.
ORCID: https://orcid.org/0000-0002-5190-841X
Hnatushenko Viktoriia – doctor of engineering sciences, professor, professor of Department of Information Technologies and Systems, Ukrainian State University of Science and Technologies.
ORCID: https://orcid.org/0000-0001-5304-4144