ULST Timisoara
Multidisciplinary Conference on Sustainable Development
21-22 May 2026
2026 MULTIDISCIPLINARY CONFERENCE ON SUSTAINABLE DEVELOPMENT

FROM PHOTOS TO RECOMMENDATIONS: CLIP-BASED VISUAL RETRIEVAL AND HYBRID ATTRIBUTE VERIFICATION FOR TOURISM SEARCH

Florin Daniel MILITARU[1], Cosmin Alin POPESCU[2], Ramona CIOLAC[1], Sebastian MOISA[1], Flavius Ionatan GAVRILĂ[3], Gabriela POPESCU[1]

[1] *University of Life Sciences "King Mihai I" from Timişoara, Faculty of Management and Rural Tourism, Romania*
[2 1] *University of Life Sciences "King Mihai I" from Timişoara, Faculty of Agriculture, Romania*
[3] *IOSUD [1] University of Life Sciences "King Mihai I" from Timişoara, Romania*
Corresponding author: gabrielapopescu@usvt.ro

Abstract: *Tourism search systems rely almost exclusively on text descriptions, which routinely omit what travellers most want to see — a fireplace, animals in the yard, a mountain through the window. TourismVisual inverts this priority: photographs are the primary evidence, and natural-language queries are matched directly against them. The prototype builds CLIP-based embeddings over destination photo collections, retrieves candidates via FAISS vector similarity, then applies hybrid attribute verification — whole-image CLIP scoring combined with open-vocabulary object detection (Grounding DINO) — to confirm or deny features such as pool, horse, traditional food or mountain view. Final ranking weights semantic similarity (0.60), attribute confirmation (0.35) and brightness (0.05); results are presented with evidence photographs and explicit per-attribute verdicts.*

Key words: *multimodal retrieval, tourism search, visual recommendation, open-vocabulary detection, CLIP*

Introduction

Listing descriptions on accommodation platforms struggle to convey what most matters to travellers — especially in rural and agro-tourism contexts, where the product is an experience rather than a commodity service.

Photographs make immediately clear what text usually cannot: animals in the yard, a wooden interior, a clay oven, the surrounding landscape. Photo-accompanied evidence has been shown to increase booking confidence under uncertainty.

CLIP-based retrieval enables matching natural-language queries directly against photo collections. TourismVisual is designed for the constrained rural setting: a few dozen guesthouses, minimal descriptions, no booking history — but a reasonable set of photographs.

Materials and Methods

Web application; operators upload photos per destination (no metadata or written descriptions required). Pipeline:

- Indexing: CLIP image embeddings + metadata (checksum, brightness), cached for fast retrieval.
- Retrieval: query encoded by a multilingual text encoder; FAISS flat inner-product search returns the top-K images.
- Bilingual intent map (~60 RO/EN concepts: pool, fireplace, horse, mountain view, traditional food…) translates the query into visual targets.
- Hybrid verification: whole-image CLIP scoring for scene attributes (cosy, traditional food, mountain view); Grounding DINO open-vocabulary detection for objects (horses, cows, playground).
- Aggregate score = 0.60·semantic + 0.35·detection + 0.05·brightness.

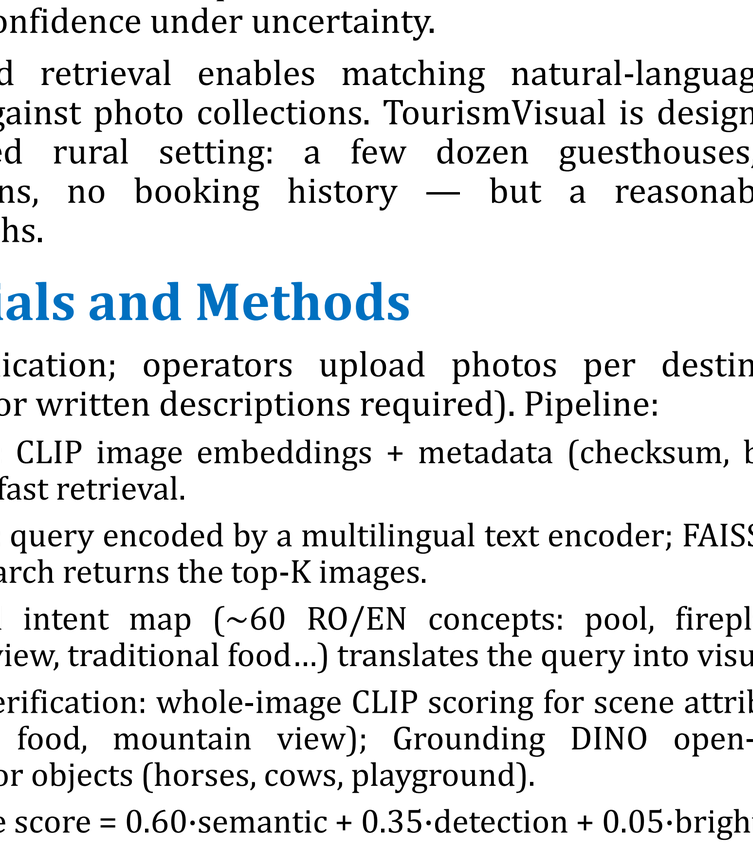

Figure 1. *Two-stage retrieval pipeline: semantic search (FAISS over CLIP) + hybrid attribute verification.*

Results and Discussion

- Top-ranked destinations actually show what the user asked for — pools, farm animals, mountain landscapes, traditional interiors — visible to the user as evidence.
- Pool query example (Fig. 2): destination SINAIA ranked #1, semantic similarity 0.28, detection 1.00, aggregate 0.48, attribute "pool" visually confirmed.
- Results are organised around destinations, not single images: a poorly lit photograph has less effect when the score aggregates the best-matching images of a property.
- Confirmed vs. unconfirmed attributes shown explicitly — prevents the system from overpromising and gives travellers actionable information.
- Two-stage architecture handles both mood-based queries ("somewhere cosy with a fireplace and a forest view") and constraint-checking ("pool AND horses").
- Free-text natural language input — no rigid filters or category hierarchy; rural attributes are rarely captured in standard filter sets anyway.

Conclusions

- Image-first retrieval with hybrid attribute verification produces more transparent and more useful recommendations for rural and experiential tourism than text-based alternatives.
- Especially relevant for rural guesthouses, agro-tourism farms and mountain cabins — where the experience is visible but rarely described in text.
- Limitations: depends on the photos operators choose to upload; atmospheric qualities (authenticity, cosiness) are hard to verify; threshold calibration remains an open problem.
- Future work: user studies, comparison vs. text-based retrieval, expanded vocabulary tailored to Romanian rural tourism.

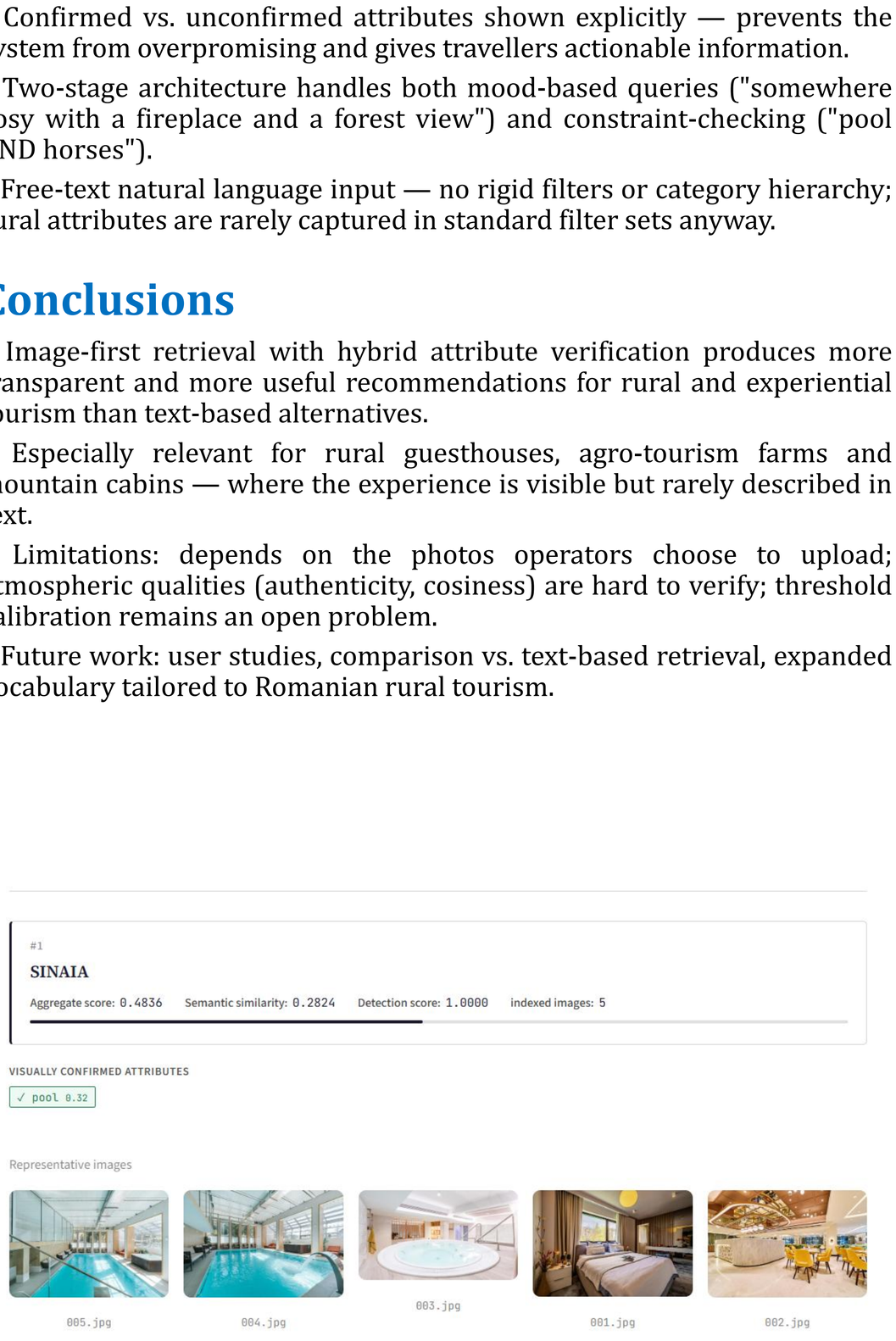

Figure 2. *TourismVisual output for the query "accommodation with a swimming pool" — top-ranked destination with aggregate, semantic and detection scores, visually confirmed attribute, and evidence images.*