



Sistem Lost and Found Berbasis Vision-Language Retrieval dengan Explainable AI untuk Lingkungan Smart Campus

An Explainable Vision-Language Retrieval System for Lost and Found in Smart Campus Environment

Enrico Samuel Djimesha, Regi Fernando Najoan*, Argha Orion Silitonga, Wilsen Grivin Mokodaser, George Morris William Tangka

Fakultas Ilmu Komputer, Informatika, Universitas Klabat, Minahasa Utara, Indonesia

Jl. Arnold Mononutu, Airmadidi Bawah, Kec. Airmadidi, Kabupaten Minahasa Utara, Sulawesi Utara, Indonesia

Email: ¹enrico.djimesha@unklab.ac.id, ^{2,*}reginajoan@unklab.ac.id, ³argha@unklab.ac.id, ⁴wilsenm@unklab.ac.id, ⁵gtangka@unklab.ac.id

Email Penulis Korespondensi: reginajoan@unklab.ac.id

Received: 12/08/2026, Revised: 22/08/2026, Accepted: 07/10/2026, Published: 11/10/2026

Abstrak—Bukti empiris mengenai retrieval gabungan gambar dan teks pada sistem Lost and Found dalam skenario kampus masih terbatas, sedangkan kemampuan CLIP zero-shot membedakan atribut dan identitas barang belum diketahui secara memadai. Penelitian ini mengevaluasi CLIP ViT-B/32 pada koleksi citra barang dari delapan kategori. Evaluasi multimodal terkontrol menggunakan 404 citra, 85 kelompok relevansi berbasis metadata sebagai proxy ground truth, dan 237 kueri leave-one-out. Validasi image-only dilakukan secara terpisah menggunakan 387 citra dari 79 Object ID berbasis provenance setelah penghapusan duplikat dan quality control. Setiap Object ID merepresentasikan beberapa citra dari satu barang fisik yang sama. Pada validasi Object ID, CLIP memperoleh Recall@1 0,674, Recall@10 0,935, dan mAP 0,419 dengan 95% object-cluster bootstrap CI 0,369–0,469. Pada evaluasi multimodal, image-only menghasilkan mAP 0,624, text-only 0,409, dan image+text mencapai mAP terbaik 0,732 pada $\alpha=0,5$, yang dipilih pada empat dari lima fold validasi silang kelompok. Ablasi menunjukkan single fused-index mengungguli three-index dengan mAP 0,715 berbanding 0,692, sehingga three-index tidak diklaim sebagai peningkatan akurasi. Untuk kueri Bahasa Indonesia, input langsung ke English CLIP menghasilkan mAP 0,163, dibandingkan 0,363 menggunakan multilingual CLIP dan 0,409 melalui penerjemahan. Attention Rollout hanya diposisikan sebagai visualisasi pendukung karena deletion-curve AUC 0,871 belum dibandingkan dengan random-occlusion baseline. Hasil penelitian menunjukkan manfaat fusi gambar-teks pada evaluasi terkontrol, tetapi belum membuktikan performa pada deskripsi bebas pengguna atau kesiapan produksi.

Kata Kunci: Vision-Language Retrieval; Explainable AI; CLIP; Barang Hilang dan Temuan; Smart Campus

Abstract—Evidence for combined image-text retrieval in campus lost-and-found systems remains limited, and zero-shot CLIP's ability to distinguish item attributes and identities is uncertain. This study evaluates CLIP ViT-B/32 on an item-image collection covering eight categories. The controlled multimodal evaluation uses 404 images, 85 metadata-defined relevance groups as proxy ground truth, and 237 leave-one-out queries. Separate image-only validation uses 387 images from 79 provenance-based Object IDs after duplicate removal and quality control; each Object ID contains multiple images of the same physical item. In Object-ID validation, CLIP achieved Recall@1 of 0.674, Recall@10 of 0.935, and mAP of 0.419 (95% object-cluster bootstrap CI: 0.369–0.469). In multimodal evaluation, image-only achieved mAP 0.624, text-only 0.409, and image+text reached the best mAP of 0.732 at $\alpha=0.5$, selected in four of five group cross-validation folds. Ablation showed that the single fused-index outperformed the three-index configuration (mAP 0.715 versus 0.692); therefore, three-index is not claimed as an accuracy improvement. Direct Indonesian input to English CLIP achieved mAP 0.163, compared with 0.363 using multilingual CLIP and 0.409 through translation. Attention Rollout is positioned only as a supporting visualization because its deletion-curve AUC of 0.871 lacks a random-occlusion baseline. These results indicate the benefit of image-text fusion under controlled evaluation but do not establish performance for free-form user descriptions or production readiness.

Keywords: Vision-Language Retrieval; Explainable AI; CLIP; Lost and Found; Smart Campus

1. PENDAHULUAN

Layanan Lost and Found kampus memerlukan pencocokan barang ketika informasi kueri tidak lengkap dan kondisi foto berbeda dari foto pendaftaran. Tantangan utamanya bukan sekadar digitalisasi pencatatan, melainkan membedakan objek yang tampak serupa berdasarkan kombinasi bentuk, warna, merek, goresan, aksesoris, lokasi, dan deskripsi pengguna. Karena pengguna dapat memiliki foto, ingatan tekstual, atau keduanya, sistem retrieval perlu dievaluasi per modalitas pada data yang mendekati penggunaan kampus [1].

CLIP menyediakan ruang embedding bersama untuk gambar dan teks, tetapi pretraining web-scale tidak menjamin sensitivitas terhadap atribut pembeda yang halus pada barang kampus. Penggunaan zero-shot juga menimbulkan risiko domain shift, terutama untuk foto dengan pencahayaan buruk, sudut berbeda, latar ramai, serta deskripsi lokal. Penelitian ini karena itu memosisikan CLIP ViT-B/32 sebagai baseline vision-language yang dievaluasi, bukan sebagai model yang telah terbukti siap produksi.

Kajian mutakhir menempatkan vision-language pre-training sebagai kerangka untuk memetakan citra dan bahasa ke representasi bersama, tetapi juga menekankan persoalan granularitas, domain shift, dan biaya adaptasi [2], [3]. Pada retrieval visual, survei deep learning menunjukkan bahwa kualitas fitur dan protokol evaluasi sangat

menentukan kemampuan membedakan instance yang serupa [4]. Penelitian image-text retrieval kemudian mengembangkan asymmetric bi-encoder [5], semantic completion and filtration [6], konsistensi kepentingan semantik lintas modalitas [7], external-space attention [8], keyword-guided pre-screening [9], serta penyelarasan lokal-global [10]. Temuan tersebut menunjukkan bahwa representasi bersama saja belum menjamin retrieval yang kuat pada atribut objek yang halus.

Keterbatasan CLIP standar pada penelitian ini terletak pada perbedaan antara pengenalan kategori dan pencocokan instance. CLIP-Adapter memperlihatkan bahwa adaptor fitur dapat meningkatkan transfer ke tugas spesifik [11], sedangkan semantic-aware fine-tuning meningkatkan kemampuan few-shot CLIP [12]. Studi atribut pejalan kaki menunjukkan manfaat prompt vision-language untuk ciri yang lebih terperinci [13]. Penyempurnaan semantik [14], adaptasi untuk geo-localization [15], dan penalaran interaksi lintas modalitas [16] juga menunjukkan bahwa domain dan bentuk interaksi kueri memengaruhi hasil. Karena penelitian ini menggunakan CLIP zero-shot tanpa fine-tuning, evaluasi harus dibaca sebagai pengukuran baseline transfer pada koleksi yang digunakan, bukan bukti kecukupan model untuk seluruh atribut barang kampus.

Penelitian Lost and Found di lingkungan perguruan tinggi masih terbatas. Tan dan Chong mengembangkan sistem Lost and Found kampus yang berfokus pada pengelolaan informasi [1], sedangkan CampusTrace menawarkan aplikasi pencatatan dan penelusuran barang [17]. LostNet menggunakan MobileNetV2, Convolutional Block Attention Module, dan perceptual hashing untuk pencocokan visual [18]. Ketiganya belum mengevaluasi secara bersamaan image-only, text-only, fusi image+text, sensitivitas Bahasa Indonesia, dan pilihan strategi indeks. Di luar domain Lost and Found, studi pengguna retrieval multimodal berbasis CLIP menunjukkan bahwa penggabungan modalitas dan rancangan antarmuka perlu dievaluasi bersama [19]. Pada aspek interpretabilitas, tinjauan XAI untuk computer vision menegaskan bahwa visualisasi yang tampak masuk akal belum otomatis menjadi penjelasan yang faithful [20]. Attention Rollout dapat membantu inspeksi wilayah visual yang relatif penting, tetapi validasi yang lebih ketat memerlukan perbandingan dengan random occlusion pada gambar dan jumlah patch yang sama. Karena baseline tersebut belum tersedia, Attention Rollout dalam studi ini diperlakukan sebagai visualisasi pendukung, bukan bukti kausal mengenai alasan model memberi peringkat tertentu.

Berdasarkan penelitian yang ditinjau, belum ditemukan studi yang mengevaluasi secara bersamaan retrieval image-only, text-only, dan image+text pada foto Lost and Found kampus, perbandingan three-index dan single fused-index, sensitivitas terhadap Bahasa Indonesia, serta keterbatasan validasi heatmap attention. Penelitian ini menanggapi ruang tersebut melalui evaluasi multimodal terkontrol menggunakan koleksi citra barang untuk skenario kampus dan teks berbasis metadata, sweep bobot fusi, ablasi indeks, serta pengujian tiga opsi pemrosesan kueri Bahasa Indonesia.

Penelitian ini mengembangkan kerangka vision-language retrieval untuk pencarian barang hilang dan temuan dalam lingkungan smart campus. CLIP ViT-B/32 digunakan sebagai dual encoder untuk memetakan gambar dan teks ke ruang embedding bersama. Sistem menyediakan indeks gambar, teks, dan gabungan sesuai modalitas kueri, sedangkan kueri gambar-teks dibentuk melalui late fusion. Heatmap Attention Rollout dan skor atribut ditampilkan hanya sebagai informasi pendukung pada hasil retrieval.

Kontribusi empiris penelitian ini mencakup lima hal: (1) validasi image-only dengan label relevansi Object ID berbasis provenance yang dikonfirmasi berdasarkan proses pengumpulan data dan pemeriksaan duplikat; (2) evaluasi retrieval multimodal terkontrol menggunakan koleksi citra barang untuk skenario kampus dan teks claimant-style berbasis metadata; (3) perbandingan single fused-index dan three-index; (4) analisis sensitivitas bobot fusi melalui sweep α dan group cross-validation; serta (5) analisis sensitivitas kueri Bahasa Indonesia menggunakan English CLIP, multilingual CLIP, dan terjemahan. Attention Rollout dan zero-shot attribute scoring digunakan sebagai komponen pendukung antarmuka. Attention Rollout diposisikan sebagai supporting visualization tanpa klaim validasi faithfulness, bukan sebagai kontribusi explainability empiris. Dalam judul, istilah Explainable AI merujuk pada integrasi kedua komponen penjelasan tersebut sebagai fungsi sistem, bukan pada klaim bahwa faithfulness penjelasan telah tervalidasi.

Penelitian ini bertujuan mengukur batas kemampuan CLIP ViT-B/32 zero-shot untuk retrieval barang kampus melalui empat pertanyaan: (1) bagaimana performa tiap modalitas dan fusi gambar-teks; (2) berapa bobot fusi yang didukung data; (3) apakah three-index mengungguli single fused-index; dan (4) bagaimana dampak Bahasa Indonesia serta degradasi foto. Teks pada evaluasi multimodal dibangkitkan secara terkontrol dari metadata dan belum mewakili variasi deskripsi pengguna alami.

2. METODOLOGI PENELITIAN

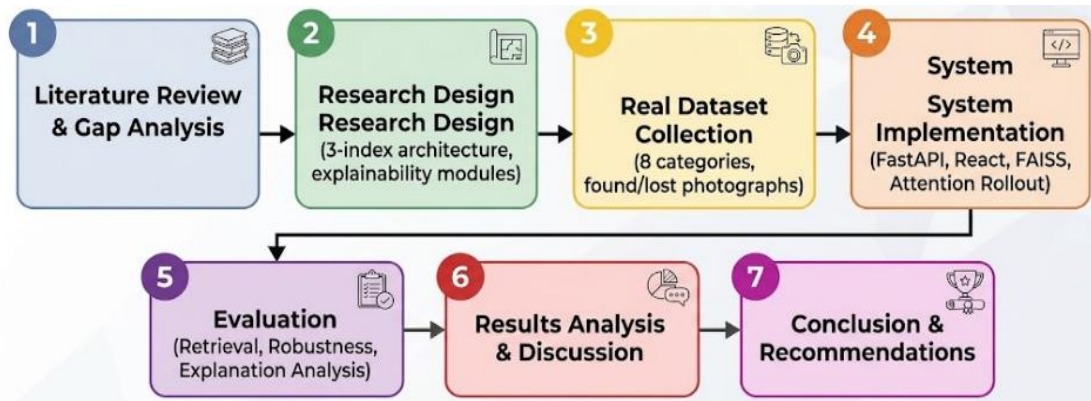
2.1 Objek Penelitian

Objek penelitian meliputi data, komputasi, dan evaluasi. Objek data terdiri atas dua koleksi terpisah: 404 citra dengan kelompok relevansi berbasis metadata untuk evaluasi utama dan 387 citra dari 79 Object ID berbasis provenance setelah quality control terhadap 394 berkas sumber. Setiap subfolder Object ID memuat citra dari satu barang fisik yang sama; kedua koleksi tidak digabungkan. Objek komputasionalnya adalah vision-language retrieval berbasis CLIP ViT-B/32 zero-shot, sedangkan evaluasi mencakup pemeringkatan kueri gambar, teks, dan

gambar-tekst secara terkontrol. Attention Rollout dan zero-shot attribute scoring hanya digunakan sebagai komponen pendukung yang menampilkan heatmap dan atribut semantik. Unit analisis tetap kualitas pemeringkatan terhadap gallery, bukan keputusan kepemilikan atau faithfulness penjelasan.

2.2 Tahapan Penelitian

Penelitian ini dilakukan melalui beberapa tahapan yang mencakup pengumpulan data, pembentukan representasi multimodal, proses retrieval, pembangkitan penjelasan, integrasi sistem, dan evaluasi. Alur penelitian secara keseluruhan ditunjukkan pada Gambar 1.



Gambar 1. Diagram alur penelitian

Gambar 1 memperlihatkan tujuh tahap yang dilaksanakan secara berurutan. Tahap pertama dan kedua merumuskan masalah, research gap, objek penelitian, serta rancangan eksperimen. Tahap ketiga menyiapkan dua koleksi evaluasi, snapshot proxy metadata dan koleksi Object ID, beserta quality control masing-masing. Tahap keempat membangun embedding CLIP, indeks retrieval, mekanisme fusi, dan modul explainability. Tahap kelima mengevaluasi modalitas gambar, teks, dan gambar-tekst, bobot fusi, rancangan indeks, opsi Bahasa Indonesia, robustness tersimulasi, serta sensitivitas heatmap. Tahap keenam membahas hasil dan keterbatasan, kemudian tahap ketujuh merumuskan kesimpulan dan rekomendasi pengembangan.

a. Pengumpulan data dan konstruksi dataset.

Dataset dikumpulkan dari foto barang hilang dan temuan yang mewakili delapan kategori objek. Pengumpulan awal menargetkan sekitar 50 foto per kategori, sehingga ditargetkan sekitar 400 foto secara keseluruhan. Foto-foto tersebut diorganisasikan untuk merepresentasikan pasangan found/lost dari objek fisik yang sama sehingga dapat digunakan sebagai dasar penentuan ground truth pada proses evaluasi.

b. Encoding multimodal.

CLIP ViT-B/32 [21] [22] digunakan sebagai dual encoder untuk menghasilkan representasi gambar dan teks. Setiap input dikonversi menjadi embedding berdimensi 512 yang kemudian dinormalisasi menggunakan normalisasi L2 sehingga representasi visual dan tekstual dapat dibandingkan dalam ruang embedding yang sama.

c. Query-aware three-index retrieval.

Sistem membangun tiga indeks FAISS yang masing-masing digunakan untuk kueri berbasis gambar, teks, dan kombinasi gambar-tekst. Pemisahan indeks dilakukan agar setiap jenis kueri dibandingkan dengan representasi yang sesuai dengan modalitasnya.

d. Pembangkitan explainability.

Setiap hasil retrieval diproses melalui dua mekanisme penjelasan. Attention Rollout digunakan untuk menghasilkan heatmap visual, sedangkan zero-shot attribute scoring digunakan untuk memberikan informasi tekstual berdasarkan atribut semantik hasil tersebut.

e. Integrasi sistem.

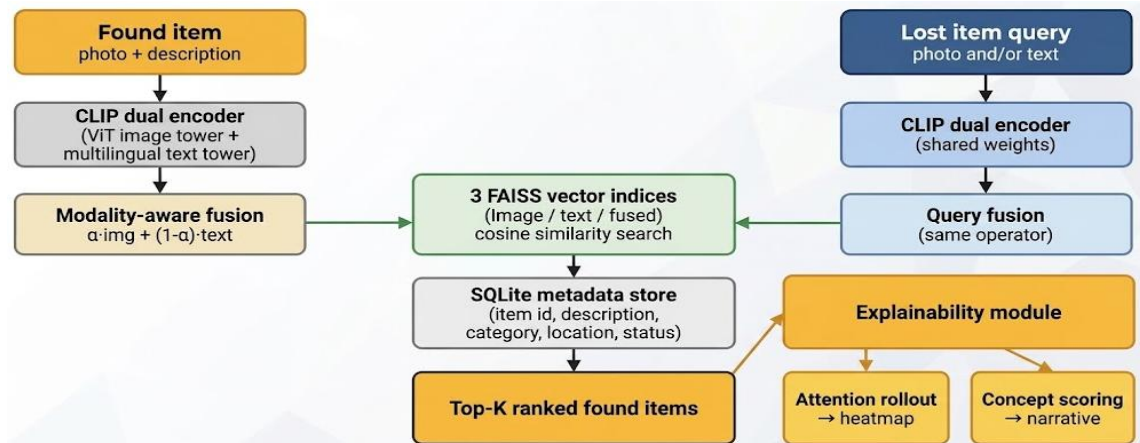
Komponen backend, frontend, penyimpanan metadata, retrieval, dan explainability diintegrasikan melalui satu pipeline utama. Pendekatan ini digunakan agar logika pemrosesan yang dipakai pada aplikasi sama dengan yang digunakan dalam proses evaluasi.

f. Evaluasi.

Performa sistem dievaluasi menggunakan Recall@K, mean Average Precision (mAP), pengujian robustness pada beberapa kondisi degradasi foto, serta deletion-curve analysis untuk mengamati perilaku penjelasan visual.

2.3 Arsitektur dan Encoding Multimodal

Arsitektur sistem terdiri atas dua jalur utama, yaitu proses pendaftaran barang temuan dan proses pencarian barang hilang. Kedua jalur menggunakan encoder yang sama sehingga barang yang tersimpan maupun kueri pengguna direpresentasikan dalam ruang embedding yang konsisten. Alur pemrosesan tersebut ditunjukkan pada Gambar 2.



Gambar 2. Pipeline end-to-end

Setiap barang yang tersimpan memiliki tiga representasi: embedding gambar, embedding teks, dan embedding gabungan. Indeks dipilih berdasarkan jenis kueri. Kueri gambar diarahkan ke indeks gambar, kueri teks ke indeks teks, sedangkan kueri gambar-teks ke indeks gabungan. Informasi seperti deskripsi, kategori, warna, lokasi, dan status barang disimpan pada SQLite. Setelah pencarian selesai, hasil diteruskan ke modul explainability sebelum ditampilkan kepada pengguna.

CLIP ViT-B/32 menghasilkan embedding gambar dan teks berdimensi 512 yang dinormalisasi L2. Model utama digunakan secara zero-shot karena jumlah kelompok relevansi berbasis metadata yang memiliki pasangan hanya 85; fine-tuning tanpa pemisahan objek yang ketat berisiko menyebabkan kebocoran identitas antara training dan testing. Pilihan ini menjadikan eksperimen sebagai evaluasi baseline transfer, bukan bukti bahwa fine-tuning tidak diperlukan. Untuk penggunaan Bahasa Indonesia, eksperimen tambahan membandingkan kueri langsung pada encoder CLIP berbahasa Inggris, multilingual CLIP, dan terjemahan Bahasa Indonesia ke bahasa Inggris.

$$v = \text{normalize}(\alpha v_{\text{image}} + (1 - \alpha) v_{\text{text}}) \quad (1)$$

Implementasi awal menggunakan $\alpha=0,6$. Sweep $\alpha \in \{0,0; 0,2; 0,4; 0,5; 0,6; 0,8; 1,0\}$ kemudian dilakukan, dengan seluruh citra dari satu kelompok relevansi berbasis metadata ditempatkan pada fold yang sama untuk mencegah kebocoran pasangan. Hasil menunjukkan bahwa $\alpha=0,5$ merupakan konfigurasi terbaik pada sweep dan pilihan yang paling sering dipilih pada five-fold group cross-validation, sedangkan $\alpha=0,6$ memberikan performa yang sangat dekat. Karena itu, $\alpha=0,5$ digunakan sebagai konfigurasi empiris utama untuk pelaporan evaluasi terkontrol.

Pada rancangan awal, seluruh barang disimpan sebagai satu representasi fused. Cara ini menimbulkan ketidakseimbangan ketika kueri hanya berupa gambar karena embedding gambar harus dibandingkan dengan representasi yang juga mengandung informasi teks. Untuk menghindari kondisi tersebut, three-index memisahkan representasi gambar, teks, dan gabungan. Kueri gambar dibandingkan dengan embedding gambar, kueri teks dengan embedding teks, sedangkan kueri multimodal dengan embedding gabungan. Pemisahan ini menjaga konsistensi representasi ketika dua gambar identik dibandingkan melalui indeks gambar. Pengaruhnya terhadap performa retrieval tetap diuji melalui eksperimen ablasi pada Bagian 3.2.

2.4 Modul Explainability

Embedding setiap barang temuan disimpan dalam FAISS IndexFlatIP. Karena semua vektor telah dinormalisasi, inner product yang dihasilkan setara dengan cosine similarity. Jika FAISS tidak tersedia, sistem menggunakan pencarian brute-force berbasis NumPy sebagai fallback.

Penjelasan visual dihasilkan menggunakan Attention Rollout [23]. Metode ini menggabungkan bobot self-attention dari beberapa layer transformer untuk memperkirakan kontribusi setiap patch terhadap representasi visual akhir. Secara umum, matriks attention dari setiap layer dirata-ratakan pada seluruh attention head, kemudian dikombinasikan dengan matriks identitas untuk mempertimbangkan residual connection. Matriks tersebut selanjutnya dinormalisasi dan dikomposisikan antar-layer sebagaimana dinyatakan pada Persamaan (2):

$$\text{Rollout} = \tilde{A}^{(L)} \cdot \tilde{A}^{(L-1)} \dots \tilde{A}_\lambda \tilde{A}^{(1)} = 0.5A^{(1)} + 0.5I \quad (2)$$

Hasil akhir divisualisasikan dalam bentuk heatmap untuk menunjukkan bagian gambar yang memperoleh relevansi relatif lebih tinggi terhadap representasi model. Selain penjelasan visual, sistem menghasilkan penjelasan tekstual melalui zero-shot attribute scoring. Sejumlah atribut yang mudah dipahami, seperti warna, jenis objek, dan material, direpresentasikan dalam ruang embedding yang sama. Kemiripan antara representasi hasil pencarian dan atribut tersebut digunakan sebagai informasi semantik tambahan pada hasil retrieval. Pendekatan ini sejalan dengan upaya interpretasi embedding CLIP melalui analisis kontribusi semantik pada representasi teks [13]



2.5 Dataset dan Pembentukan Label Relevansi

Benchmark sintetis 100 pasangan hanya digunakan sebagai sanity check pipeline, bukan sebagai bukti performa lapangan. Evaluasi multimodal, bahasa, bobot fusi, dan ablasi menggunakan snapshot 404 citra dari delapan kategori (tas ransel, kacamata, laptop, smartphone, payung, dompet, jam tangan, dan botol minum), dengan 24–59 citra per kategori. Pada protokol leave-one-out, teks claimant-style dalam bahasa Inggris dan Indonesia dibangkitkan dari kategori, warna, dan lokasi. Pasangan kategori-deskripsi identik membentuk 85 kelompok relevansi berbasis metadata dan 237 kueri. Kelompok relevansi berbasis metadata digunakan sebagai proxy ground truth, sedangkan teks claimant-style dibangkitkan secara terkontrol dari atribut objek.

Validasi image-only menggunakan koleksi Object ID terpisah, bukan turunan dari snapshot 404 citra. Penulis mengonfirmasi berdasarkan provenance pengumpulan bahwa setiap subfolder memuat beberapa citra dari barang fisik yang sama. Quality control terhadap 394 berkas mengecualikan satu duplikat eksak dalam satu Object ID dan enam duplikat lintas-Object ID, serta menggabungkan dua folder payung yang merepresentasikan objek sama. Koleksi akhir terdiri atas 387 citra, 79 Object ID dengan 3–9 citra per objek, dan 1.578 pasangan relevan query-gallery berarah. Dalam evaluasi leave-one-out, citra kueri dikeluarkan dari gallery dan citra lain dengan Object ID yang sama ditetapkan sebagai relevan.

2.6 Metrik Evaluasi dan Protokol Robustness

Performa diukur menggunakan Recall@K, mAP, dan MRR. Recall@K menyatakan proporsi kueri yang memiliki sedikitnya satu item relevan pada $K \in \{1, 5, 10\}$ hasil teratas. mAP merupakan rata-rata Average Precision seluruh kueri berdasarkan posisi setiap item relevan, sedangkan MRR mengukur kebalikan peringkat item relevan pertama. Ketidakpastian dihitung melalui 10.000 cluster-bootstrap replicates dengan kelompok relevansi sebagai unit resampling, sementara perbandingan arsitektur menggunakan paired group-level randomisation test dengan 100.000 pengacakan.

CLIP ViT-B/32 dievaluasi secara zero-shot tanpa fine-tuning. Pada protokol leave-one-out, setiap dari 237 citra digunakan bergiliran sebagai kueri terhadap 404 kandidat; citra kueri dikeluarkan untuk mencegah self-match, sedangkan citra lain dalam kelompok relevansi yang sama dianggap relevan. Protokol identik digunakan untuk image-only, text-only, dan image+text agar perbandingan bersifat berpasangan. Pemilihan bobot fusi menggunakan five-fold group cross-validation: α dipilih dari kelompok training dan diuji pada kelompok held-out. Confidence interval mengukur ketidakpastian sampling, tetapi tidak menghilangkan potensi bias proxy ground truth.

Deletion-curve AUC mengukur perubahan cosine similarity ketika patch dengan relevansi Attention Rollout tertinggi dioklusi bertahap berdasarkan konsep deletion/insertion [24]. Tanpa random-occlusion baseline, metrik ini hanya menunjukkan sensitivitas, bukan faithfulness. Robustness diuji terhadap Gaussian blur, pencahayaan buruk, perubahan sudut melalui perspective warp dan rotasi, latar ramai, serta kombinasi beberapa gangguan; citra tanpa degradasi digunakan sebagai acuan.

2.7 Implementasi Sistem

Sistem dibangun secara modular dengan satu pipeline yang menghubungkan encoding, retrieval, dan explainability. Pipeline yang sama digunakan oleh backend aplikasi dan komponen evaluasi sehingga logika inti tetap konsisten antara penggunaan sistem dan pengujian. Komponen teknologi yang digunakan pada setiap lapisan sistem dirangkum pada Tabel 1.

Tabel 1. Technology stack sistem

Layer	Teknologi	Fungsi
Vision-language encoder	CLIP ViT-B/32 [21] [22]; multilingual option [25]	Menghasilkan embedding gambar dan teks
Retrieval index	FAISS IndexFlatIP [26]	Pencarian nearest-neighbor berdasarkan modalitas kueri
Metadata store	SQLite	Menyimpan deskripsi, kategori, lokasi, dan status
Visual explainability	Attention Rollout [23]	Menghasilkan heatmap relevansi visual
Textual explainability	Zero-shot attribute scoring (CLIP)	Menghasilkan informasi atribut semantik
Backend API	FastAPI + Uvicorn	Menyediakan endpoint sistem
Frontend client	React + Vite	Antar muka search / report

3. HASIL DAN PEMBAHASAN

3.1 Hasil Retrieval pada Dataset

Evaluasi retrieval awal dilakukan menggunakan snapshot proxy sebagaimana dijelaskan pada Bagian 2.5. Backbone CLIP ViT-B/32 digunakan secara zero-shot, sedangkan proses evaluasi mengikuti protokol leave-one-

out image retrieval. Ringkasan performa pada berbagai metrik retrieval dan hasil analisis deletion-curve disajikan pada Tabel 2.

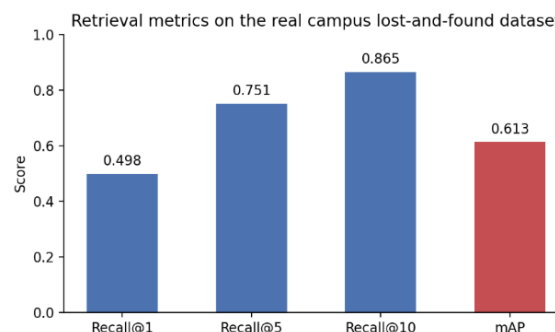
Tabel 2. Hasil evaluasi retrieval dan deletion-curve

Metrik	Nilai
Recall@1	0.498
Recall@5	0.751
Recall@10	0.865
Mean Reciprocal Rank (MRR)	0.613
mAP	0.624
Deletion-curve AUC	0.871

Tabel 2 menunjukkan Recall@1 sebesar 0,498, yang berarti barang yang sesuai berada pada peringkat pertama untuk sekitar setengah dari seluruh kueri. Ketika jumlah hasil yang diperiksa diperluas, Recall meningkat menjadi 0,751 pada lima hasil teratas dan 0,865 pada sepuluh hasil teratas. Dengan kata lain, meskipun hasil yang tepat tidak selalu muncul di posisi pertama, sebagian besar kueri masih menemukan barang yang sesuai dalam daftar yang relatif singkat. Temuan ini mendukung antarmuka yang menampilkan beberapa hasil pencarian, bukan hanya satu hasil dengan skor tertinggi. Audit implementasi menemukan bahwa nilai 0,613 pada evaluasi awal merupakan Mean Reciprocal Rank (MRR), karena perhitungan hanya menggunakan kebalikan peringkat citra relevan pertama. Nilai tersebut tidak lagi disebut mAP. Dengan definisi Average Precision atas seluruh citra relevan pada pasangan query-gallery yang sama, evaluasi image-only menghasilkan mAP 0,624. Evaluasi image+text terkontrol mencapai mAP tertinggi 0,732 pada $\alpha=0,5$ menggunakan snapshot citra barang dan teks claimant-style berbasis metadata.

LostNet [18] berfokus pada pencocokan visual berbasis citra, sedangkan sistem pada penelitian ini mendukung kueri gambar, teks, kombinasi gambar-teks, serta informasi penjelasan. Karena dataset dan protokol evaluasi kedua penelitian berbeda, nilainya tidak dibandingkan secara langsung. Perbandingan hanya digunakan untuk menunjukkan perbedaan cakupan fungsi dan skenario evaluasi. Pada snapshot 404 citra, evaluasi menggunakan protokol leave-one-out image retrieval: citra kueri dikeluarkan dari gallery dan citra lain dalam kelompok relevansi berbasis metadata yang sama diperlakukan sebagai relevan. Sebanyak 237 citra dari 85 kelompok relevansi memenuhi syarat sebagai kueri, sedangkan 167 citra lainnya tidak memiliki pasangan metadata identik. Recall@K bernilai satu apabila sedikitnya satu citra relevan muncul pada K hasil teratas. Average Precision dihitung berdasarkan seluruh citra relevan pada setiap kueri dan dirata-ratakan menjadi mAP, menghasilkan mAP image-only sebesar 0,624. Selain mAP, MRR digunakan untuk mengukur posisi citra relevan pertama. Pada validasi Object ID, seluruh 387 citra digunakan bergiliran sebagai kueri terhadap 386 kandidat gallery. Aturan Average Precision yang sama diterapkan pada seluruh citra dengan Object ID identik, dan interval kepercayaan 95 % dihitung melalui 10.000 cluster-bootstrap dengan Object ID sebagai unit resampling. Protokol Object ID dimaksudkan sebagai validasi independen terhadap perilaku retrieval image-only, bukan sebagai pengganti atau versi yang dapat dibandingkan langsung dengan benchmark berbasis metadata.

Untuk memberikan gambaran visual mengenai kinerja retrieval, nilai Recall@1, Recall@5, Recall@10, dan mAP divisualisasikan pada Gambar 3. Visualisasi ini memperlihatkan peningkatan kemampuan menemukan citra relevan seiring bertambahnya jumlah kandidat yang dipertimbangkan.



Gambar 3. Metrik retrieval pada snapshot proxy berbasis metadata.

Sebagai pemeriksaan tambahan, self-match sanity check dilakukan dengan menggunakan foto identik sebagai kueri terhadap representasi gambar yang tersimpan. Pengujian menghasilkan cosine similarity 1,0000 dan menempatkan item yang sesuai pada peringkat pertama (Top-1 match), sehingga memperoleh status PASS. Hasil ini menunjukkan bahwa routing kueri gambar ke indeks gambar mempertahankan representasi yang sama untuk input identik tanpa dipengaruhi komponen teks.

Audit label relevansi dimulai dari 394 file foto pada delapan kategori. Setiap subfolder sumber dibuat dari beberapa citra satu barang fisik yang sama, sehingga label Object ID didasarkan pada provenance pengumpulan.

Sebanyak tujuh berkas dikeluarkan karena merupakan duplikat eksak: satu duplikat di dalam Object ID yang sama dan enam kemunculan duplikat lintas-Object ID. Setelah penggabungan dua folder payung yang merepresentasikan objek sama, data evaluasi berisi 387 citra dari 79 Object ID. Setiap objek memiliki 3–9 citra, dan aturan relevansi menghasilkan 1.578 pasangan query-gallery berarah. Seluruh 387 citra digunakan sebagai kueri secara bergiliran, dengan 386 kandidat gallery setelah self-match dikeluarkan.

Tabel 3. Ringkasan evaluasi dengan relevansi Object ID berbasis provenance

Protokol	Data	R@1	R@5	R@10	MRR	mAP
Object ID	79 objek; 387 kueri	0,674	0,848	0,935	0,757	0,419

Pada Tabel 3 dapat dilihat bahwa CLIP ViT-B/32 zero-shot menghasilkan Recall@1 0,674, Recall@5 0,848, Recall@10 0,935, dan mAP 0,419 dengan 95% object-cluster bootstrap CI 0,369–0,469. mAP per kategori tertinggi terdapat pada backpack (0,630) dan watch (0,607), sedangkan nilai terendah terdapat pada umbrella (0,224) dan handphone (0,255). Recall@1 menunjukkan bahwa sedikitnya satu tampilan objek yang benar sering muncul di peringkat awal, sedangkan mAP menunjukkan bahwa seluruh tampilan relevan belum selalu terkumpul pada posisi atas. Protokol Object ID ini merupakan validasi independen perilaku retrieval image-only, bukan pengganti atau versi yang dapat dibandingkan langsung dengan benchmark metadata yang menghasilkan mAP image-only 0,624

3.2 Ablasi Arsitektur Three-Index

Eksperimen ablasi membandingkan konfigurasi single fused-index dan query-aware three-index menggunakan data yang sama, yaitu 85 kelompok relevansi berbasis metadata dan 237 kueri. Pada single fused-index, representasi gambar dan teks digabungkan dalam satu indeks, sedangkan pada three-index, kueri gambar diarahkan ke indeks gambar (image routing).

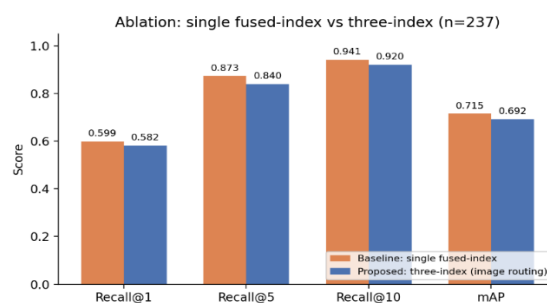
Tabel 4. Ablasi single fused-index dan three-index pada kueri berbasis gambar

Configuration	Recall@1	Recall@5	Recall@10	mAP
single fused-index	0.599	0.873	0.941	0.715
three-index (image routing)	0.582	0.840	0.920	0.692

Berdasarkan Tabel 4 dan Gambar 4, single fused-index menghasilkan performa retrieval yang lebih tinggi dibandingkan three-index pada seluruh metrik yang dievaluasi. Nilai mAP masing-masing sebesar 0,715 dan 0,692, sedangkan Recall@1, Recall@5, dan Recall@10 juga menunjukkan pola yang sama.

Analisis sensitivitas dengan mempertahankan seluruh 404 kandidat menunjukkan hasil yang konsisten, dengan mAP sebesar 0,655 pada single fused-index dan 0,624 pada three-index. Selisih sebesar 0,031 didukung oleh interval kepercayaan cluster-bootstrap 95% sebesar 0,013–0,063 dan paired group-level randomisation test ($p=0,0025$). Hasil ini menunjukkan bahwa pada data yang diuji, pemisahan indeks belum memberikan peningkatan terhadap akurasi retrieval. Data yang tersedia belum memungkinkan penentuan penyebab perbedaan kinerja kedua konfigurasi secara langsung. Salah satu kemungkinan adalah bahwa penggabungan representasi gambar dan teks pada fused-index dapat memanfaatkan informasi metadata yang membantu ketika citra dalam kelompok yang sama memiliki perbedaan sudut atau kondisi visual. Namun, dugaan tersebut belum diuji melalui ablasi khusus sehingga tidak dapat dianggap sebagai kesimpulan kausal.

Secara keseluruhan, Tabel 3 dan Gambar 4 menunjukkan adanya trade-off desain, bukan keunggulan three-index. Pada data yang diuji, single fused-index memberikan kinerja retrieval yang lebih tinggi dan karena itu digunakan sebagai konfigurasi utama, sedangkan three-index dapat dipertahankan sebagai alternatif routing untuk kebutuhan pemisahan modalitas.



Gambar 4. Perbandingan performa single fused-index dan arsitektur three-index.

3.3 Evaluasi Multimodal pada Benchmark Sintetis

Evaluasi multimodal dilakukan untuk membandingkan kontribusi informasi visual dan tekstual terhadap kinerja retrieval. Evaluasi menggunakan 237 kueri dan snapshot 404 citra sebagai gallery, dengan proxy ground truth

berdasarkan kelompok relevansi berbasis metadata. Hasil perbandingan berbagai konfigurasi multimodal berdasarkan nilai mAP disajikan pada Tabel 5.

Tabel 5. Evaluasi multimodal pada snapshot proxy (237 kueri; 85 kelompok)

Konfigurasi	mAP
Image-only	0,624
Text-only (Inggris)	0,409
Image+text, $\alpha=0,5$	0,732
Image+text, $\alpha=0,6$	0,727
Indonesia langsung, English CLIP	0,163
Indonesia, multilingual CLIP	0,363
Indonesia→Inggris, CLIP	0,409

Hasil menunjukkan bahwa image-only memperoleh mAP sebesar 0,624, sedangkan text-only bahasa Inggris memperoleh 0,409. Fusi image+text dengan $\alpha=0,5$ menghasilkan mAP tertinggi sebesar 0,732, diikuti $\alpha=0,6$ sebesar 0,727. Dengan demikian, konfigurasi terbaik meningkatkan mAP sebesar 0,108 dibandingkan image-only. Namun, hasil berbasis templat claimant-style perlu ditafsirkan secara hati-hati karena dapat memberikan estimasi performa yang lebih optimistis dibandingkan deskripsi bebas pengguna.

Sweep bobot menghasilkan mAP 0,409; 0,459; 0,646; 0,732; 0,727; 0,678; dan 0,624 masing-masing untuk $\alpha=0,0$; 0,2; 0,4; 0,5; 0,6; 0,8; dan 1,0. Five-fold group cross-validation memilih $\alpha=0,5$ pada empat fold dan $\alpha=0,6$ pada satu fold, dengan pooled held-out mAP 0,722. Hasil ini menunjukkan bahwa $\alpha=0,5$ merupakan konfigurasi terbaik pada sweep dan pilihan yang paling sering dipilih pada group cross-validation, sedangkan $\alpha=0,6$ memberikan performa yang sangat dekat.

Pada kueri terkontrol dalam dataset ini, Bahasa Indonesia yang dimasukkan langsung ke English CLIP menghasilkan mAP 0,163. Multilingual CLIP meningkatkan mAP menjadi 0,363, sedangkan terjemahan ke bahasa Inggris mencapai 0,409, sama dengan kueri Inggris paralel. Hasil ini menunjukkan bahwa pada kondisi pengujian tersebut, terjemahan memberikan mAP tertinggi di antara tiga opsi pemrosesan Bahasa Indonesia yang dibandingkan.

3.4 Evaluasi Robustness pada Kondisi Foto yang Realistik

Selain evaluasi pada foto asli, dilakukan pengujian robustness untuk mengevaluasi ketahanan retrieval terhadap berbagai bentuk degradasi visual yang mungkin terjadi pada penggunaan nyata. Pengujian mencakup Gaussian blur, pencahayaan buruk, perubahan sudut kamera, latar belakang ramai, serta kondisi gabungan yang mengombinasikan beberapa gangguan. Setiap kondisi diuji menggunakan 237 kueri yang sama dengan evaluasi utama. Hasil evaluasi pada berbagai kondisi foto disajikan pada Tabel 6.

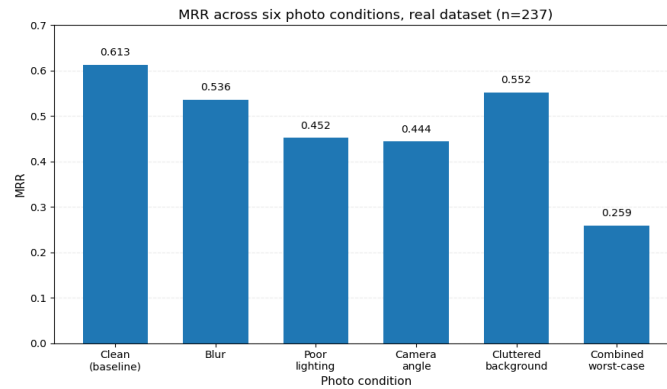
Tabel 6. Hasil retrieval pada berbagai kondisi foto

Condition	Recall@1	Recall@5	Recall@10	MRR
Clean (baseline)	0.498	0.751	0.865	0.613
Blur	0.430	0.679	0.772	0.536
Poor lighting	0.354	0.599	0.705	0.452
Camera angle	0.342	0.591	0.679	0.444
Cluttered background	0.443	0.709	0.806	0.552
Combined worst-case	0.177	0.367	0.494	0.259

Tabel 6 menunjukkan bahwa seluruh gangguan visual menyebabkan penurunan kinerja dibandingkan kondisi foto tanpa degradasi. Pada gangguan tunggal, Recall@10 masih berada pada rentang 0,679–0,806, yang menunjukkan bahwa sebagian besar citra relevan masih dapat ditemukan dalam sepuluh hasil teratas meskipun kualitas visual mengalami penurunan. Pada kondisi gabungan, Recall@10 menurun menjadi 0,494 dan MRR menjadi 0,259, menunjukkan bahwa kombinasi beberapa gangguan memberikan dampak yang lebih besar terhadap kualitas pemeringkatan.

Di antara gangguan tunggal, perubahan sudut kamera menghasilkan penurunan MRR terbesar, dari 0,613 menjadi 0,444 atau sekitar 27,6%. Pencahayaan buruk menunjukkan penurunan yang relatif serupa, dengan MRR sebesar 0,452 atau sekitar 26,3% lebih rendah dibandingkan kondisi awal. Sementara itu, latar belakang ramai memberikan dampak paling kecil, dengan MRR sebesar 0,552 atau sekitar 9,9% lebih rendah dari kondisi tanpa degradasi.

Hasil tersebut menunjukkan bahwa perubahan sudut pengambilan gambar dan kondisi pencahayaan lebih berpengaruh terhadap kinerja retrieval dibandingkan keramaian latar belakang pada dataset yang digunakan. Dalam penggunaan sistem, kualitas foto dapat dijaga dengan memastikan objek terlihat jelas dan memperoleh pencahayaan yang memadai. Penurunan yang lebih besar pada kondisi gabungan juga menunjukkan bahwa keberadaan beberapa gangguan visual secara bersamaan dapat semakin menurunkan kualitas representasi dan akurasi pemeringkatan hasil retrieval.

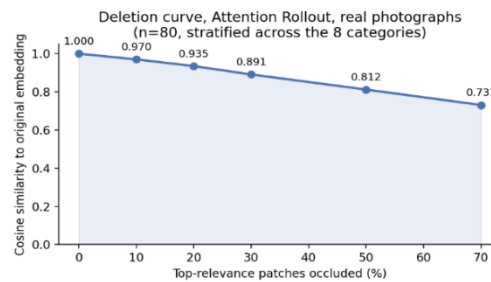


Gambar 5. Perbandingan MRR pada enam kondisi foto.

3.5 Evaluasi Penjelasan Visual

Evaluasi penjelasan visual dilakukan menggunakan deletion curve pada 80 foto yang dipilih secara terstratifikasi dari delapan kategori. Patch dengan skor relevansi tertinggi berdasarkan Attention Rollout dihilangkan secara bertahap untuk mengukur perubahan cosine similarity terhadap embedding awal.

Dapat dilihat pada Gambar 6, cosine similarity menurun dari 1,000 pada kondisi awal menjadi 0,970, 0,935, 0,891, 0,812, dan 0,731 setelah masing-masing 10%, 20%, 30%, 50%, dan 70% patch relevan dihilangkan. Kurva tersebut menghasilkan AUC sebesar 0,871, yang menunjukkan adanya perubahan representasi ketika wilayah dengan relevansi tinggi dihilangkan.

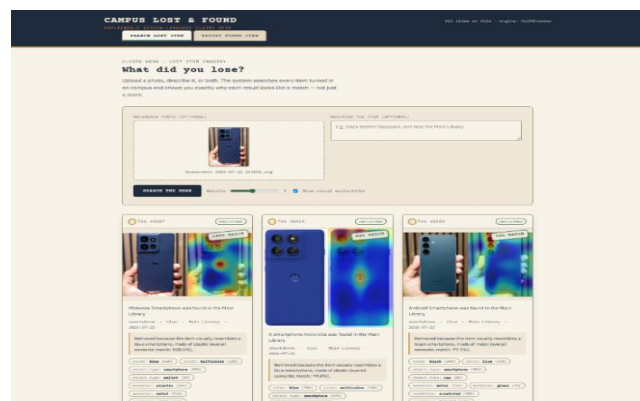


Gambar 6. Deletion curve Attention Rollout.

Hasil ini memberikan indikasi awal bahwa wilayah yang disorot Attention Rollout berkaitan dengan representasi visual model. Namun, nilai tersebut belum dapat digunakan sebagai bukti faithfulness karena belum dibandingkan dengan kontrol seperti random occlusion. Oleh karena itu, AUC 0,871 diposisikan sebagai evaluasi awal, sedangkan perbandingan dengan random occlusion diperlukan untuk memperkuat validasi penjelasan visual.

3.6 Antarmuka Pengguna

Antarmuka sistem menampilkan hasil retrieval dalam bentuk kartu yang berisi foto barang, persentase kemiripan, heatmap Attention Rollout, penjelasan singkat, serta skor atribut seperti warna, jenis objek, material, dan kondisi. Informasi tersebut membantu pengguna membandingkan beberapa hasil yang memiliki karakteristik serupa seperti terlihat pada Gambar 7.



Gambar 7. Antarmuka pencarian Claims Desk yang menampilkan hasil retrieval, heatmap, persentase kemiripan, dan skor atribut.

Gambar 7 menunjukkan antarmuka “Claims Desk” yang menerima foto referensi, deskripsi teks, atau kombinasi keduanya. Sistem menampilkan hasil berurutan berdasarkan tingkat kemiripan dan melengkapinya dengan heatmap, penjelasan tekstual, serta informasi atribut.

Penyajian beberapa hasil didukung oleh peningkatan Recall dari 0,498 pada Recall@1 menjadi 0,865 pada Recall@10 untuk kondisi awal. Pada kondisi gabungan, Recall@1 sebesar 0,177 meningkat menjadi 0,494 pada Recall@10, sehingga daftar hasil memberi ruang verifikasi yang lebih baik dibandingkan hanya menampilkan peringkat pertama.

Pada contoh pencarian smartphone, barang yang sesuai memperoleh tingkat kemiripan tertinggi, sementara objek dengan tampilan serupa muncul pada peringkat berikutnya. Skor atribut dan heatmap digunakan sebagai informasi pendukung untuk memahami hasil pemeringkatan; heatmap tidak digunakan sebagai satu-satunya dasar penentuan kecocokan.

3.7 Keterbatasan Penelitian dan Pengembangan Selanjutnya

Penelitian ini memiliki beberapa keterbatasan. Eksperimen multimodal, bobot fusi, bahasa, dan ablasi masih menggunakan 85 kelompok relevansi berbasis metadata dari 404 citra, sehingga angka pada eksperimen tersebut belum merupakan performa terhadap Object ID. Validasi image-only tambahan telah memakai 79 Object ID dan 387 citra setelah audit duplikat. Distribusi jumlah citra per objek juga terbatas pada 3–9 citra. Selain itu, data Object ID belum dilengkapi deskripsi bebas yang ditulis pengguna tanpa melihat metadata gallery, sehingga belum dapat dipakai untuk menghitung ulang hasil image+text, sensitivitas bahasa, dan ablasi fused-index pada protokol relevansi yang sama. Karena asal seluruh citra tidak diverifikasi sebagai dokumentasi pengambilan langsung di kampus, naskah secara konsisten menyebutnya sebagai koleksi citra barang untuk skenario kampus dan tidak mengklaim provenance pengambilan langsung di kampus.

Confidence interval cluster-bootstrap dan uji randomisasi pada eksperimen multimodal mengukur ketidakpastian sampling pada 85 kelompok relevansi berbasis metadata yang tersedia. Validasi image-only berbasis Object ID telah menurunkan risiko salah kelompok akibat kategori-deskripsi identik dan menggunakan Object ID sebagai unit cluster-bootstrap. Identitas fisik Object ID didasarkan pada konfirmasi penulis bahwa setiap subfolder dibentuk dari beberapa citra barang yang sama selama pengumpulan, bukan dari pengelompokan ulang berdasarkan kemiripan visual. Anotasi independen oleh dua anotator tidak dilakukan; karena itu, percentage agreement dan Cohen’s κ tidak dilaporkan. Re-anotasi independen pada penelitian lanjutan dapat memperkuat bukti reliabilitas tanpa mengubah sumber label berbasis provenance tersebut.

Pengujian cluttered background masih menggunakan komposit sehingga belum sepenuhnya mencerminkan gangguan alami. Pemilihan bobot fusi telah dievaluasi melalui sweep α dan group cross-validation, tetapi evaluasi teks tetap menggunakan kalimat berbasis templat. Pengumpulan berikutnya perlu merekam deskripsi bebas dari pengguna yang tidak melihat metadata gallery dan membagi data berdasarkan identitas objek untuk fine-tuning serta pengujian terpisah.

Random-occlusion baseline untuk Attention Rollout belum tersedia, sehingga deletion AUC hanya dilaporkan sebagai pemeriksaan sensitivitas. Pada kueri terkontrol dalam dataset ini, English CLIP langsung menghasilkan mAP lebih rendah daripada multilingual CLIP dan terjemahan; ketiga opsi tersebut masih perlu diuji pada deskripsi pengguna alami. Dataset juga kecil dan tidak seimbang, CLIP belum di-fine-tune, serta evaluasi multimodal memanfaatkan atribut berlabel yang dapat menguntungkan text retrieval. Sistem karena itu belum dinilai siap produksi.

3.8 Pembahasan

Hasil penelitian menunjukkan bahwa CLIP ViT-B/32 mampu mendukung retrieval barang pada skenario Lost and Found, dengan Recall@10 sebesar 0,935 pada validasi Object ID. Dibandingkan penelitian sebelumnya seperti Tan dan Chong [1] serta CampusTrace [17] yang berfokus pada pengelolaan dan penelusuran barang, penelitian ini memperluas fungsi retrieval dengan mendukung kueri gambar, teks, dan kombinasi keduanya. Sementara itu, LostNet [18] menggunakan pendekatan visual berbasis MobileNetV2 dan CBAM, sedangkan penelitian ini menggunakan representasi vision-language berbasis CLIP. Karena dataset dan protokol evaluasi berbeda, perbandingan nilai performa secara langsung tidak dilakukan.

Pada evaluasi multimodal, fusi image+text menghasilkan mAP 0,732, lebih tinggi dibandingkan image-only 0,624 dan text-only 0,409. Temuan ini menunjukkan bahwa informasi visual dan tekstual dapat saling melengkapi dalam retrieval, sejalan dengan penelitian vision-language retrieval yang menekankan manfaat integrasi kedua modalitas [5]–[10]. Namun, three-index justru menghasilkan mAP 0,692, lebih rendah dibandingkan single fused-index sebesar 0,715. Hasil ini menunjukkan bahwa pemisahan indeks tidak selalu meningkatkan akurasi dan perlu dibuktikan melalui evaluasi empiris.

Pada aspek robustness dan explainability, perubahan sudut kamera dan pencahayaan memberikan penurunan performa terbesar, sedangkan Attention Rollout menghasilkan deletion-curve AUC 0,871. Temuan tersebut menunjukkan bahwa kondisi visual masih menjadi faktor penting dalam retrieval, sementara Attention Rollout dapat digunakan sebagai visualisasi pendukung. Namun, tanpa random-occlusion baseline, hasil tersebut belum cukup untuk menyatakan faithfulness penjelasan, sebagaimana ditekankan dalam kajian XAI [20].

4. KESIMPULAN

Penelitian ini mengevaluasi CLIP ViT-B/32 secara zero-shot untuk image retrieval pada skenario Lost and Found. Validasi image-only berbasis Object ID terhadap 387 citra dari 79 objek menghasilkan Recall@1 sebesar 0,674, Recall@5 sebesar 0,848, Recall@10 sebesar 0,935, dan mAP sebesar 0,419 dengan 95% object-cluster bootstrap confidence interval sebesar 0,369–0,469, yang menunjukkan bahwa model dapat menemukan citra relevan pada peringkat awal, meskipun belum konsisten mengumpulkan seluruh tampilan relevan pada posisi teratas. Pada benchmark metadata, mAP image-only mencapai 0,624 setelah penerapan definisi Average Precision yang konsisten, sedangkan nilai 0,613 pada evaluasi sebelumnya dikategorikan sebagai MRR. Evaluasi multimodal berbasis proxy menghasilkan mAP sebesar 0,732 pada $\alpha=0,5$ dan 0,727 pada $\alpha=0,6$, lebih tinggi dibandingkan image-only, tetapi hasil tersebut belum dapat digeneralisasi ke deskripsi bebas pengguna karena teks dibangkitkan dari metadata dan koleksi evaluasinya berbeda dari dataset Object ID. Single fused-index juga menunjukkan kinerja lebih tinggi dibandingkan three-index pada protokol yang diuji. Selain itu, pemrosesan Bahasa Indonesia menunjukkan mAP sebesar 0,163 untuk kueri langsung pada English CLIP, meningkat menjadi 0,363 pada multilingual CLIP dan 0,409 dengan pendekatan terjemahan. Secara keseluruhan, penelitian ini memberikan validasi empiris terhadap image-only retrieval, evaluasi multimodal terkontrol, perbandingan arsitektur indeks, analisis bobot fusi, dan strategi pemrosesan Bahasa Indonesia, sementara re-anotasi independen serta evaluasi dengan deskripsi bebas pengguna masih diperlukan untuk memperkuat validitas hasil sebelum implementasi produksi.

5. PERNYATAAN PENULIS

5.1 Kontribusi Penulis

Konseptualisasi: Enrico Samuel Djimesha, Regi Fernando Najoan, dan Argha Orion Silitonga;
Metodologi: Enrico Samuel Djimesha, Regi Fernando Najoan, dan Argha Orion Silitonga;
Perangkat Lunak: Enrico Samuel Djimesha dan Argha Orion Silitonga;
Validasi: Enrico Samuel Djimesha, Wilsen Grivin Mokodaser, dan George Morris William Tangka;
Analisis Formal: Regi Fernando Najoan, Enrico Samuel Djimesha, dan Argha Orion Silitonga;
Investigasi: Enrico Samuel Djimesha, Wilsen Grivin Mokodaser, dan George Morris William Tangka;
Sumber Daya: Regi Fernando Najoan, Wilsen Grivin Mokodaser, dan George Morris William Tangka;
Kurasi Data: Enrico Samuel Djimesha, Regi Fernando Najoan, dan George Morris William Tangka;
Penulisan Draf Awal: Enrico Samuel Djimesha dan Regi Fernando Najoan;
Penulisan, Peninjauan, dan Penyuntingan: Enrico Samuel Djimesha, Regi Fernando Najoan, Argha Orion Silitonga, Wilsen Grivin Mokodaser, dan George Morris William Tangka;
Visualisasi: Argha Orion Silitonga dan Enrico Samuel Djimesha;
Seluruh penulis telah membaca dan menyetujui versi manuskrip yang telah diterbitkan.

5.2 Ketersediaan Data

Data yang disajikan dalam penelitian ini tersedia berdasarkan permintaan kepada penulis korespondensi.

5.3 Pendanaan

Penelitian dan penulisan manuskrip ini tidak menerima pendanaan khusus dari lembaga pemerintah, organisasi komersial, maupun organisasi nirlaba. Seluruh kegiatan penelitian dilakukan menggunakan sumber daya yang tersedia bagi penulis.

5.4 Pernyataan Dewan Peninjau Institusional

Tidak Berlaku (Not applicable).

5.5 Pernyataan Persetujuan Berdasarkan Informasi

Tidak Berlaku (Not applicable).

5.6 Pernyataan Konflik Kepentingan

Para penulis menyatakan bahwa mereka tidak memiliki kepentingan keuangan yang bersaing atau hubungan pribadi yang diketahui yang dapat dianggap memengaruhi pekerjaan yang dilaporkan dalam artikel ini.

5.7 Penggunaan AI Generatif dan Teknologi Berbantuan AI dalam Proses Penulisan

Para penulis menggunakan AI Generatif untuk meningkatkan kejelasan penulisan dalam artikel ini. Para penulis telah meninjau dan menyunting konten yang dihasilkan dengan bantuan AI serta bertanggung jawab penuh atas versi akhir publikasi.



REFERENCES

- [1] S. Y. Tan and C. R. Chong, “An Effective Lost and Found System in University Campus,” *Journal of Information System and Technology Management*, vol. 8, no. 32, pp. 99–112, 2023, doi: 10.35631/JISTM.832007.
- [2] J. Zhang, J. Huang, S. Jin, and S. Lu, “Vision-Language Models for Vision Tasks: A Survey,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 8, pp. 5625–5644, 2024, doi: 10.1109/TPAMI.2024.3369699.
- [3] F.-L. Chen et al., “VLP: A Survey on Vision-Language Pre-Training,” *Machine Intelligence Research*, vol. 20, no. 1, pp. 38–56, 2023, doi: 10.1007/s11633-022-1369-5.
- [4] S. R. Dubey, “A Decade Survey of Content-Based Image Retrieval Using Deep Learning,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 5, pp. 2687–2704, 2022, doi: 10.1109/TCSVT.2021.3080920.
- [5] W. Xiong, H. Liu, S. Mi, and Y. Zhang, “Asymmetric Bi-Encoder for Image-Text Retrieval,” *Multimed. Syst.*, vol. 29, no. 6, pp. 3805–3818, 2023, doi: 10.1007/s00530-023-01162-2.
- [6] S. Yang et al., “Semantic Completion and Filtration for Image-Text Retrieval,” *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 19, no. 4, pp. 1–20, 2023, doi: 10.1145/3572844.
- [7] Z. Liu, F. Chen, J. Xu, W. Pei, and G. Lu, “Image-Text Retrieval With Cross-Modal Semantic Importance Consistency,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 5, pp. 2465–2476, 2023, doi: 10.1109/TCSVT.2022.3220297.
- [8] H. Zhu, C. Zhang, Y. Wei, S. Huang, and Y. Zhao, “ESA: External Space Attention Aggregation for Image-Text Retrieval,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 10, pp. 6131–6143, 2023, doi: 10.1109/TCSVT.2023.3253548.
- [9] M. Cao, Y. Bai, Z. Cao, L. Nie, and M. Zhang, “Efficient Image-Text Retrieval via Keyword-Guided Pre-Screening,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 6, pp. 5132–5145, 2024, doi: 10.1109/TCSVT.2023.3339489.
- [10] L. Meng, F. Zhang, X. Zhang, and C. Xu, “Prototype Local-Global Alignment Network for Image-Text Retrieval,” *Int. J. Multimed. Inf. Retr.*, vol. 11, no. 4, pp. 525–538, 2022, doi: 10.1007/s13735-022-00258-1.
- [11] P. Gao et al., “CLIP-Adapter: Better Vision-Language Models With Feature Adapters,” *Int. J. Comput. Vis.*, vol. 132, no. 2, pp. 581–595, 2024, doi: 10.1007/s11263-023-01891-x.
- [12] Y. Zhu et al., “Enhancing Few-Shot CLIP With Semantic-Aware Fine-Tuning,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 7, pp. 12349–12362, 2025, doi: 10.1109/TNNLS.2024.3443394.
- [13] X. Wang, J. Jin, C. Li, J. Tang, C. Zhang, and W. Wang, “Pedestrian Attribute Recognition via CLIP-Based Prompt Vision-Language Fusion,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 35, no. 1, pp. 148–161, 2025, doi: 10.1109/TCSVT.2024.3454366.
- [14] R. Chen et al., “CLIP-Based Semantic Refinement Method for Image-Text Retrieval,” *Comput. J.*, vol. 68, no. 10, pp. 1377–1385, 2025, doi: 10.1093/comjnl/bxaf044.
- [15] J. Wu and G. Feng, “CLIP-UG: CLIP-Driven Vision-Language Model for UAV-View Geo-Localization,” *IEEE Transactions on Consumer Electronics*, vol. 71, no. 4, pp. 10096–10107, 2025, doi: 10.1109/TCE.2025.3600486.
- [16] T. Yao, S. Peng, L. Wang, Y. Li, and Y. Sun, “Cross-Modality Interaction Reasoning for Enhancing Vision-Language Pre-Training in Image-Text Retrieval,” *Applied Intelligence*, vol. 54, no. 23, pp. 12230–12245, 2024, doi: 10.1007/s10489-024-05823-1.
- [17] T. Abhiram, A. K. EB, M. KM, V. AP, and A. A., “CampusTrace: Application for Lost and Found Items on Campus,” *Journal of Trends in Computer Science and Smart Technology*, vol. 6, no. 2, pp. 168–179, 2024, doi: 10.36548/jtcsst.2024.2.006.
- [18] M. Zhou, I. Fung, L. Yang, N. Wan, K. Di, and T. Wang, “LostNet: A Smart Way for Lost and Find,” *PLoS One*, vol. 19, no. 10, p. e0310998, Oct. 2024, doi: 10.1371/journal.pone.0310998.
- [19] T. Alpay, S. Magg, P. Broze, and D. Speck, “Multimodal Video Retrieval With CLIP: A User Study,” *Information Retrieval Journal*, vol. 26, no. 1–2, p. 6, 2023, doi: 10.1007/s10791-023-09425-2.
- [20] Z. Cheng, Y. Wu, Y. Li, L. Cai, and B. Ihnaini, “A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision,” *Sensors*, vol. 25, no. 13, p. 4166, 2025, doi: 10.3390/s25134166.
- [21] A. Radford et al., “Learning Transferable Visual Models from Natural Language Supervision,” in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021, pp. 8748–8763. Accessed: Aug. 21, 2026. [Online]. Available: <https://proceedings.mlr.press/v139/radford21a>
- [22] A. Dosovitskiy et al., “An Image Is Worth 16 x 16 Words: Transformers for Image Recognition at Scale,” in *Proceedings of the 9th International Conference on Learning Representations (ICLR)*, 2021. Accessed: Aug. 21, 2026. [Online]. Available: <https://arxiv.org/pdf/2010.11929>
- [23] S. Abnar and W. Zuidema, “Quantifying Attention Flow in Transformers,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 4190–4197. doi: 10.18653/v1/2020.acl-main.385.
- [24] V. Petsiuk, A. Das, and K. Saenko, “RISE: Randomized Input Sampling for Explanation of Black-Box Models,” in *Proceedings of the British Machine Vision Conference (BMVC)*, 2018. doi: <https://doi.org/10.48550/arXiv.1806.07421>.
- [25] F. Carlsson, P. Eisen, F. Rekathati, and M. Sahlgren, “Cross-Lingual and Multilingual CLIP,” in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 2022, pp. 6848–6854. [Online]. Available: <https://aclanthology.org/2022.lrec-1.739/>
- [26] J. Johnson, M. Douze, and H. Jégou, “Billion-Scale Similarity Search with GPUs,” *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021, doi: 10.1109/TBDATA.2019.2921572.