Forensic Science International: Digital Investigation 56 (2026) 302057

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS EU 2026 - Selected Papers from the 13th Annual Digital Forensics Research Conference Europe

Needle in a case: Scalable search over large-scale image corpora in forensic applications

Kamil Faber [a,*], Dominik Żurek [a], Kacper Bujak [a], Monika Selegrat [b], Kamil Piętak [a]

[a] Faculty of Computer Science, AGH University of Krakow, al. Adama Mickiewicza 30, Krakow, 30-059, Lesser Poland, Poland
[b] The Internal Security Agency, Rakowiecka 2A Str., Warsaw, 00-993, Poland

ARTICLE INFO

Keywords:
Digital forensics
Image search
Artificial intelligence
Large-scale data analysis
Evidence analysis

ABSTRACT

The rapid growth of digital content has made images a key form of communication, which also extends into criminal contexts, where visual material often constitutes critical evidence. Forensic analysts face the challenge of locating relevant images within large, heterogeneous datasets, such as those extracted from mobile devices. Manual inspection of such data is time-consuming and inefficient. Although traditional automated classification methods offer partial support, they remain constrained by predefined class sets, limiting their applicability in the dynamic and unpredictable nature of real forensic investigations. Recent advances in artificial intelligence (AI) have introduced models capable of retrieving images using natural-language queries, enabling more universal and adaptive search capabilities. In this work, we conduct a comprehensive evaluation of two modern AI paradigms for large-scale forensic image retrieval: Vision–Language Models (VLMs), which generate searchable textual captions of images, and Contrastive Language–Image Pre-training (CLIP), which performs embedding-based text–image similarity search.

To the best of our knowledge, this is the first systematic comparison of these approaches in a forensic context. We evaluate 33 representative queries across three forensic use cases and nine heterogeneous datasets comprising over 80 000 images. Our results offer new insights into the trade-offs between caption-based and embedding-based retrieval methods and their applicability in practical digital forensic workflows.

1. Introduction

The widespread use of images in interpersonal communication via digital devices, including photographs, screenshots, memes, and graphics, highlights their role as carriers of information in everyday interactions. This trend naturally extends into the realm of criminal activity, where visual content can serve both as a medium of communication and as critical evidence. Consequently, the ability to quickly and accurately categorize and search for relevant images has become a key challenge in digital forensics Vasilaras et al. (2024). In practice, user devices often contain thousands of images, making manual inspection and analysis prohibitively time-consuming and resource-intensive.

Over the years, numerous methods have been developed to automate image classification Del Mar-Raave et al. (2021) and support forensic analysis. However, their effectiveness is inherently limited by the requirement to define a fixed set of classes. This constraint proves impractical in investigative contexts, where each case may involve unique patterns of interest. At the same time, developing dedicated classifiers for individual investigations is usually infeasible.

Recent advances in artificial intelligence (AI) have introduced a paradigm shift in this domain, unveiling novel avenues for the automated analysis of images. Particularly important is the ability to retrieve images based on arbitrary user queries, which makes it possible to move beyond rigid class definitions and ensures the universality of the tool.

Despite dynamic progress in AI-based image analysis, comprehensive studies on its application in digital forensics remain scarce, particularly in tasks that require universality and flexibility. Existing approaches have focused primarily on classification within limited sets of classes Del Mar-Raave et al. (2021), while forensic practice demands coverage of a much wider range of image types and unpredictable user queries. Moreover, current solutions rarely account for the critical aspect of computational performance, which, in the context of large-scale data analysis, is fundamental to their practical utility.

In this work, we address this gap by conducting a comprehensive, large-scale evaluation of two fundamentally different, modern AI paradigms for image retrieval. The first approach, based on Vision–Language

* Corresponding author.
E-mail address: kfaber@agh.edu.pl (K. Faber).

https://doi.org/10.1016/j.fsidi.2026.302057

Models (VLMs) Beyer et al. (2024), combines visual and textual understanding to generate descriptive captions that can subsequently be searched using conventional text retrieval mechanisms. The second, Contrastive Language–Image Pre-training (CLIP) Radford et al. (2021), performs retrieval directly through cross-modal embedding similarity, enabling flexible, open-vocabulary search without predefined classes.

Our study is, to the best of our knowledge, the first systematic comparisons of these two paradigms in a forensic context. We investigate 33 representative queries grouped into three diverse use cases reflecting typical forensic search scenarios. Our evaluation includes nine heterogeneous datasets comprising over 80000 images. This extensive and realistic experimental setup allows us not only to quantify retrieval accuracy but also to analyze model behavior and computational efficiency.

The remainder of this paper is organized as follows. Section 2 reviews related work on image analysis in forensic applications and provides an overview of the two investigated approaches: VLM and CLIP. Section 3 details the proposed research methodology, while Section 4 outlines the experimental setup. The obtained results are analyzed and discussed in Section 5. Finally, Section 6 concludes the paper and summarizes the key findings of our study.

2. Related work

2.1. Image analysis in commercial forensic software tools

The beginnings of machine learning in computer forensics focused on the automatic identification of illegal or evidence-relevant visual content. The first DNN-based image classifiers were designed to detect predefined categories such as weapons, drugs, pornography, and child sexual abuse material (CSAM). Magnet Axiom was one of the first to introduce its Magnet.AI module – in 2017 for categorizing chats, and in 2018 for categorizing images. A recent preliminary survey Sanna et al. (2024) tries to evaluate also the robustness of the employed AI algorithms against adversarial attacks (manipulating the content so that the AI system does not recognize the offensive/prohibited content).

At the same time, hashing-based techniques such as pHash, dHash, Facebook's PDQ, Apple's Neuralhash McKeown and Buchanan (2023) or PhotoDNA Steinebach (2023) played a fundamental role, allowing the identification of previously known images. The next evolutionary step was the transition of automatic classifiers to searching for similar images, allowing the search of evidence for photos visually similar to a specified reference image (from the evidence collection or from external sources indicated by the analyst).

The above solutions (categorization and similar image search) are currently offered by most leading forensic software vendors – Magnet Axiom, Cellebrite, Oxygen, Nuix (integrated with T3K), etc. However, software vendors do not disclose technical details about the models used. Marketing materials and technical documentation are dominated by general terms such as “proprietary models,” “deep learning,” and “proprietary AI engines,” which do not provide precise information about model architecture, training data, or validation methodology.

Currently, so-called copilots are gaining popularity. Since 2024, there has been a visible trend towards integration with AI Assistant software, which initially analyzed test artifacts – chats, messages, browsing history (Magnet Axiom Copilot, initially online, now available in an offline version since 2025). A novelty and a milestone is the BelkaSoft solution – BelkaGPT – advertised as the first offline copilot in computer forensics. BelkaGPT is a multimodal assistant capable of generating image descriptions and searching for them in natural language.

However, as with automatic classifiers, there is no documentation available on the effectiveness and architecture of this model. In the context of the development of computer forensics tools, it is worth mentioning Cellbrite's Guardian solution – an online platform for data management, processing, and analysis. Guardian, announced in February 2025, enables data processing using GenAI. It operates in the AWS cloud architecture (also available in the AWSgov version upon customer request). The platform offers initial integration with copilot modules, which extend the capabilities of searching data with queries formulated in natural language. Again, there is no available documentation on the effectiveness and architecture of the models used.

There is a trend towards the integration of increasingly complex and advanced GenAI into commercial solutions. However, their development brings with it a new problem of limited transparency of operation. Unlike classic methods based on explicit rules, modern AI modules function largely as “black boxes” whose decision-making logic is not fully known to the end user. The lack of insight into the architecture, training data (possible bias), and model validation methods means that trust in the generated results is based mainly on the manufacturer's declaration.

In this context, this work aims to fill the gap in the research of open AI solutions in the field of text-based image search for computer forensics. Not only the precision of the search is crucial here, but also the processing time of large data sets and the potential bias of the models. Based on the initial research, the two concepts, described in the following sections, are chosen.

2.2. CLIP and contrastive Vision–Language alignment

CLIP represents one of the earliest and most influential vision–language architectures based on contrastive alignment between image and text representations Radford et al. (2021). In such a contrastive learning approach, the model independently encodes visual and textual inputs, and then optimizes their embeddings to be close in a shared latent space, thereby enabling zero-shot generalization across multiple downstream tasks.

Radford et al. Radford et al. (2021) introduced the CLIP model (Fig. 1), which broke with traditional computer vision paradigms that relied on training networks with predefined object categories Sharma and Guleria (2022). Instead, CLIP was trained using contrastive learning on 400 million image–text pairs collected from the internet, with the objective of aligning images and their captions. This large-scale pre-training enabled zero-shot transfer across a wide range of downstream tasks without additional task-specific training. The model was evaluated on more than 30 benchmarks, including OCR, geolocation, and fine-grained classification, achieving competitive results compared to fully supervised approaches. Remarkably, CLIP matched the accuracy of ResNet-50 He et al. (2016) on ImageNet Deng et al. (2009) without using the 1.28 million labeled training examples.

Building upon this foundation, subsequent research explored the universal detection of AI-generated or manipulated content using pre-trained CLIP embeddings. Cozzolino et al. Cozzolino et al., 2023 proposed a lightweight decision strategy based on CLIP features, which requires only a few sample images from a generative model rather than a large, domain-specific training set. The CLIP-based detector demonstrated strong generalization and robustness, achieving competitive

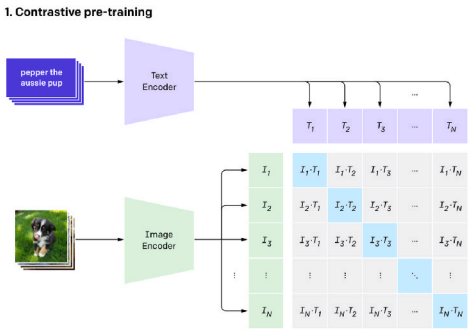


Fig. 1. Source: https://openai.com/index/clip/.

performance with state-of-the-art methods on in-distribution benchmarks while substantially improving detection on out-of-distribution samples and resilience against laundering or impaired data.

Further extending the adaptability of CLIP, Cui et al. Cui et al. (2025a) introduced the Forensics Adapter, a lightweight adapter network designed to transform CLIP into an effective and universal face forgery detector. Unlike previous approaches that treated CLIP solely as a frozen feature extractor, the adapter learns forgery-specific artifacts, such as blending boundaries in manipulated faces. With only 5.7 million trainable parameters, it improves detection accuracy across five benchmark datasets. The extended variant, Forensics Adapter++ Cui et al. (2025b), further incorporates textual information through a forgery-aware prompt learning strategy. CLIP is also tested in particular areas of applications such as medical imaging Zhao et al. (2025).

2.3. Generative Vision–Language Models and PaliGemma

In contrast to contrastive architectures such as CLIP, modern Vision–Language Models (VLMs) employ generative or autoregressive mechanisms that jointly process text and image tokens within a unified transformer backbone. This design enables fine-grained multimodal reasoning, contextual grounding, and task transfer across vision and language domains.

Google Research introduced PaliGemma Beyer et al. (2024), an open vision–language model combining the SigLIP-So400m visual encoder with the Gemma-2B language model. Designed as a versatile and knowledge-rich foundation model, PaliGemma was evaluated on nearly 40 benchmarks, including standard VLM tests as well as specialized tasks such as segmentation and remote sensing.

More recently, Steiner et al. (2024) presented PaliGemma 2 (Fig. 2), an upgraded family of models extending the original design with the full range of Gemma 2 LLMs, from 2B up to 27B parameters. The models were trained for multiple image resolutions ($224px^2$, $448px^2$, and $896px^2$) using a multi-stage training setup to facilitate efficient knowledge transfer via fine-tuning. Compared to the original PaliGemma, PaliGemma 2 expanded the range of transfer tasks to include OCR-related benchmarks, long fine-grained captioning, and radiology report generation, achieving state-of-the-art performance across these domains.

3. Methodology

In this section, we describe the methodology used to evaluate and compare the two analyzed approaches: the Vision–Language Model (VLM)-based and the Contrastive Language–Image Pre-training (CLIP)-based image retrieval methods. Given the fundamental architectural differences between the two, we developed dedicated retrieval pipelines to ensure fair and consistent evaluation.

We begin by presenting the research toolkit built for this comparison, outlining how each approach processes, indexes, and retrieves visual data. Then, we introduce the three use cases designed to reflect realistic forensic image retrieval scenarios.

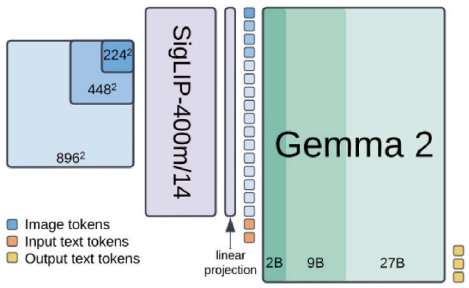


Fig. 2. Source: https://arxiv.org/pdf/2412.03555.

3.1. A retrieval methodology for CLIP and VLM

CLIP Radford et al. (2021) combines visual and language representations in a semantic vector space. It consists of an image encoder and a text encoder. Both encoders transform the input data into embeddings of the same length. CLIP learns to match these representations so that the vectors of corresponding image–text pairs are as close as possible to each other in the vector space, while mismatched ones are as far apart as possible. As a result, CLIP can later encode new texts and images, allowing for a similarity comparison between them.

Our retrieval methodology for CLIP approach starts with processing images into a set of vectors later stored in memory (to simplify subsequent tests, the set was serialized into a file). During the retrieval process, all text queries are encoded in embeddings, which are then used to find matching images using the k-nearest neighbors (kNN) algorithm Taunk et al. (2019) with cosine similarity. The returned results are sorted according to the similarity of the vectors representing the text query and the individual images. For this reason, the list of results for each query is practically limited only by the total number of images in the dataset. However, in our metrics calculation, we limit the number of considered images by K parameter.

In turn, PaliGemma 2, a VLM model that we used, is a vision language model that uses the features of language models of the Gemma 2 family, as well as the SigLIP visual model. VLM can generate appropriate descriptions based on input images and text strings (prompts).

Our retrieval methodology for VLM model starts with processing each image using a "caption" prompt, whose task is to generate a description of the image content. The generated descriptions are then saved in a dedicated Elasticsearch index configured for English language. As an extension, the index can be expanded to include definitions of synonyms for individual terms appearing in queries. After indexing the entire dataset, all queries are directed to the Elasticsearch index, without the involvement of the AI model. The image description are stored in the model as a *TEXT* field, enabling full-text search. The retrieval results are sorted based on the relevance of the query and the indexed descriptions[1]. In our paper, we use term **VLM** for retrieval using VLM model and Elasticsearch index. Moreover, we also present the results for the VLM-based approach with included synonyms (denoted in the rest of the paper as **VLMs**).

3.2. Use cases

To comprehensively evaluate the performance of the analyzed approaches, we designed three representative use cases reflecting typical forensic image retrieval tasks. Each use case targets a distinct level of semantic and contextual understanding, allowing us to assess both general retrieval capabilities and domain-specific challenges. The

Table 1
Summary of the datasets: number of samples and total size per dataset.

Dataset	Samples	Size(GB)
Flickr30k	31 783	4.11
Russia Ukraine War	28 470	9.01
Meme Images	6992	0.7
Weapon Detection	5859	1.73
Emails Screenshots	5800	0.18
Vehicle Classification	1600	0.83
Mobile Gallery	1266	0.44
Selfies & ID	435	0.94
Guns Object Detection	333	0.003
Total	82 538	17.94

[1] https://www.elastic.co/what-is/search-relevance.

queries used in all use cases are summarized in Table 2.

Use Case 1 (UC1) focuses on identifying images depicting specific **relationships between objects or persons** as well assessing the impact of **conceptual refinement** on search quality. Examples include queries such as "person holding a gun" vs "man holding a pistol", as well as "a gun in a holster." Such queries are especially relevant in forensic analyses where recognizing the relationship between objects (e.g., weapon possession, item handling) and refining search with specific concepts are critical for quick and effective identification of evidence.

Use Case 2 (UC2) examines how both models handle queries that describe **the detailed characteristics or attributes of objects**, such as "a red sedan car" or "a man in a gray hoodie and jeans". These queries evaluate the sensitivity of the models to descriptive modifiers (color, clothing, or accessories) and their ability to match visual features to specific textual attributes. In real-world forensic scenarios, such capabilities are crucial when witnesses or investigators provide partial or descriptive information about visual evidence.

Use Case 3 (UC3) investigates the models' ability to identify the **type of visual content** rather than specific objects or scenes. We focus on detecting screenshots of digital communications, such as "screenshot of WhatsApp" or "screenshot of an email." This use case reflects frequent forensic needs, such as categorizing or filtering large image collections containing mixed media.

In general, these three use cases collectively cover a wide range of forensic image retrieval challenges, from relational reasoning and attribute recognition to semantic content classification, providing a robust basis for evaluating the precision, flexibility, and practical applicability of the examined models.

4. Experimental setup

4.1. Datasets

To comprehensively evaluate both CLIP and VLM-based methods in forensic context, we used a diverse collection of publicly available datasets covering a broad range of visual domains and semantic contexts within forensic domain. Moreover, we also included a more general dataset that provides "noise" – images that are not necessarily related and useful in forensic investigation. This choice helps us create a more complex environment, ensure more realistic evaluation and avoid significant biases related to the diversity of images in the data set.

The selected datasets include everyday photographs (mainly coming from flickr30k collection, which has become the standard benchmark for verifying mechanisms for sentence-based image descriptions), social media content, and specialized imagery, such as military scenes, weapons, and document screenshots. This collection was supplemented with images relevant to the analytical cases developed, taken from other data sets. This diversity enables the assessment of model robustness, generalization, and semantic grounding across both generic and sensitive visual categories. A summary of all datasets used in the experiments, along with their descriptions and primary content domains, is presented below and in Table 1.

1. Flickr30k[2]: a collection of photos of everyday scenes.
2. Russia Ukraine War[3]: armed conflict scenes, including soldiers, military vehicles, and weapons.
3. Meme Images[4]: various types of memes.
4. Weapon Detection[5]: a collection containing various types of weapons, banknotes of different denominations, photographs of smartphones, and wallets.
5. Emails Screenshots[6]: ordinary emails and those containing spam.
6. Vehicle Classification[7]: various types of vehicles.
7. Mobile Gallery[8]: photos taken with a smartphone that contain vehicles, memes, landscapes, selfies, and screenshots of the WhatsApp application.
8. Selfies & ID[9]: selfies of different people against various backgrounds and photos of ID cards.
9. Guns Object Detection[10]: different types of weapons in everyday photographs and movie frames.

4.2. Evaluation metrics

In order to assess the performance of the proposed models, we apply two metrics: Recall@K and Precision@K. These metrics are widely used in information retrieval to quantify both retrieval accuracy and ranking quality Japkowicz and Boukouvalas (2024).

- **Recall@K** measures the proportion of relevant items successfully retrieved within the top K results in relation to the total number of relevant items in the dataset:

$$Recall@K = \frac{\text{Number of relevant items in top } K}{\text{Total number of relevant items}} \tag{1}$$

- **Precision@K** quantifies the proportion of retrieved items in the top K that are relevant:

$$Precision@K = \frac{\text{Number of relevant items in top } K}{K} \tag{2}$$

For example, when searching a specific query, the system returns 10 images. Seven of them are correct, while there are twenty correct images in total. Therefore, Precision@10 = 7/10 = 0.7 and Recall@10 = 7/20 = 0.35.

4.3. Details about used AI models

For CLIP-based experiments, we employe the `clip-retrieval`[11] framework, which was adapted to our specific requirements and integrated with the ViT-B/32[12] model pre-trained on 224 × 224 input images. The resulting embeddings were L2-normalized, stored, and subsequently indexed using `autofaiss`, which automatically selects the optimal index type based on the available memory and the characteristics of the dataset, using the inner product as a similarity metric.

For the VLM-based experiments, we used the `Paligemma2-3b-mix-224`[13] model, which comprises approximately 3 billion parameters and has been fine-tuned on a mixture of academic multimodal tasks with 224 × 224 image inputs.

[2] https://www.kaggle.com/datasets/hsankesara/flickr-image-dataset.
[3] https://www.kaggle.com/datasets/piterfm/2022-ukraine-russia-war-equipment-losses-oryx.
[4] https://www.kaggle.com/datasets/hammadjavaid/6992-labeled-meme-images-dataset.
[5] https://www.kaggle.com/datasets/mehmetcubukcu/weapon-detection/data.
[6] https://www.kaggle.com/datasets/alihossary/emails-screenshots-enron-1-2.
[7] https://www.kaggle.com/datasets/marquis03/vehicle-classification.
[8] https://www.kaggle.com/datasets/n0obcoder/mobile-gallery-image-classification-data.
[9] https://www.kaggle.com/datasets/tapakah68/selfies-id-images-dataset.
[10] https://www.kaggle.com/datasets/issaisasank/guns-object-detection.
[11] https://github.com/rom1504/clip-retrieval.
[12] https://github.com/openai/CLIP.
[13] https://huggingface.co/google/paligemma2-3b-mix-224.

Table 2
Quantitative results for all evaluated queries across three approaches (VLM, VLMs – *VLM with synonyms*, and CLIP). The table reports the number of relevant images matching the query (Rel.), Precision@K (Prec.), Recall@K (Rec.), number of retrieved images (Ret.), and true positives (TP) for each query. K value for both Precision@K and Recall@K is equal to the number of relevant images (Rel.). Moreover, we showcase number of retrieved images only for VLM and VLMs, as CLIP can provide unbound number of results with decreasing similarity.

Query	Rel.	VLM				VLMs				CLIP		
		Prec.	Rec.	Ret.	TP	Prec.	Rec.	Ret.	TP	Prec.	Rec.	TP
UC-1												
Man holding a pistol	440	0.71	0.01	7	5	0.41	0.12	135	55	0.41	0.41	182
Man holding a rifle	14	0.00	0.00	50	0	0.00	0.00	50	0	0.00	0.00	0
Person holding a rifle	43	0.00	0.00	11	0	0.05	0.05	100	2	0.00	0.00	0
Person holding a shotgun	9	0.00	0.00	0	0	0.00	0.00	1	0	0.00	0.00	0
Person using a sniper rifle	7	0.00	0.00	0	0	0.00	0.00	0	0	0.43	0.43	3
Woman holding a pistol	51	0.00	0.00	0	0	0.04	0.04	69	2	0.43	0.43	22
Woman with a shotgun	2	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00	0
Half-naked woman with a rifle	1	0.00	0.00	0	0	0.00	0.00	0	0	0.00	0.00	0
Hand holding a knife	152	0.50	0.50	304	76	0.50	0.50	304	76	0.66	0.66	101
Black person with a gun	8	0.00	0.00	20	0	0.00	0.00	20	0	0.25	0.25	2
Gun in a holster	13	0.15	0.15	31	2	0.00	0.00	37	0	0.54	0.54	7
Gun on a table	21	0.10	0.10	50	2	0.05	0.05	50	1	0.57	0.57	12
a person carrying a suspicious bag	4	0.00	0.00	0	0	0.00	0.00	0	0	0.25	0.25	1
a person pointing a weapon at another person	3	0.00	0.00	0	0	0.00	0.00	20	0	1.00	1.00	3
a dog in a car	7	0.14	0.14	19	1	0.14	0.14	19	1	0.71	0.71	5
person next to a white building	5	0.00	0.00	1	0	0.00	0.00	12	0	0.40	0.40	2
person playing violin outdoors	9	0.00	0.00	0	0	0.00	0.00	0	0	0.67	0.67	6
A police officer standing next to a vehicle	9	0.00	0.00	0	0	1.00	0.11	1	1	0.67	0.67	6
A gun in hand	52	0.08	0.08	104	4	0.19	0.19	104	10	0.27	0.27	14
Average	45	0.09	0.05	31	5	0.13	0.06	49	8	0.38	0.38	19
UC-2												
A car with licence plate	459	0.00	0.00	0	0	0.51	0.09	82	42	0.45	0.45	208
A red sedan car	23	0.00	0.00	0	0	0.00	0.00	0	0	0.61	0.61	14
A white BMW	8	0.33	0.12	3	1	0.33	0.12	3	1	0.50	0.50	4
An orange house with a brown roof	1	0.00	0.00	0	0	0.00	0.00	0	0	1.00	1.00	1
a bus with people inside	9	0.00	0.00	0	0	0.00	0.00	0	0	0.78	0.78	7
a child holding a baloon	11	0.00	0.00	0	0	0.00	0.00	0	0	0.73	0.73	8
a dog wearing clothes	5	1.00	0.20	1	1	1.00	0.20	1	1	0.60	0.60	3
man in gray hoodie and jeans	4	0.00	0.00	0	0	0.00	0.00	1	0	0.00	0.00	0
woman in a red shirt and dark pants with a hat on her head	2	0.00	0.00	0	0	0.00	0.00	0	0	0.50	0.50	1
Family on a beach	100	0.00	0.00	0	0	0.13	0.13	146	13	0.69	0.69	69
Average	62	0.13	0.03	0	0	0.20	0.05	23	6	0.59	0.59	32
UC-3												
Screenshot of Whatsapp	164	0.00	0.00	0	0	0.47	0.47	264	77	0.98	0.98	160
Screenshot of Whatsapp with attachment	15	0.00	0.00	0	0	0.00	0.00	0	0	0.40	0.40	6
Screenshot of Whatsapp with audio attachment	6	0.00	0.00	0	0	0.00	0.00	0	0	0.67	0.67	4
Screenshot of an email	5800	1.00	0.00	16	16	0.78	0.01	64	50	0.96	0.96	5551
Average	1496	0.25	0.00	4	4	0.31	0.12	82	32	0.75	0.75	1430

5. Results and discussion

In this section, we analyze and discuss the performance of the evaluated approaches across the defined use cases and queries. Our goal is not only to report quantitative metrics, but also to interpret them in the context of real-world applicability and retrieval behavior. We focus on understanding where the models succeed, where they struggle, and what factors influence their effectiveness, such as query formulation, dataset characteristics, or the presence of semantically complex concepts. In addition to quantitative metrics, we examine selected qualitative examples that highlight the nuances behind the numerical results, offering a more comprehensive view of the strengths and limitations of each model.

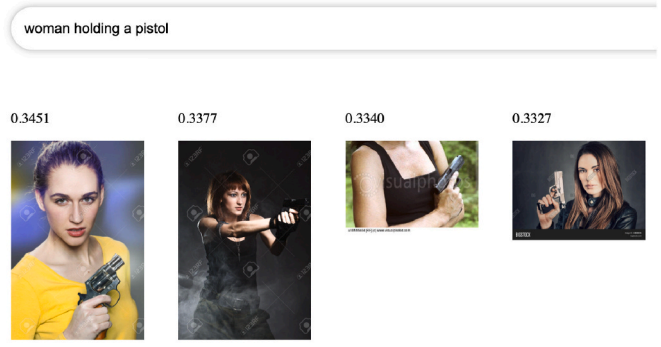
As we mentioned earlier, we evaluated 33 queries grouped into 3 Use Cases by searching through 80000 images from 9 datasets. Table 2 presents quantitative results for all evaluated queries across three approaches.

5.1. Use case 1

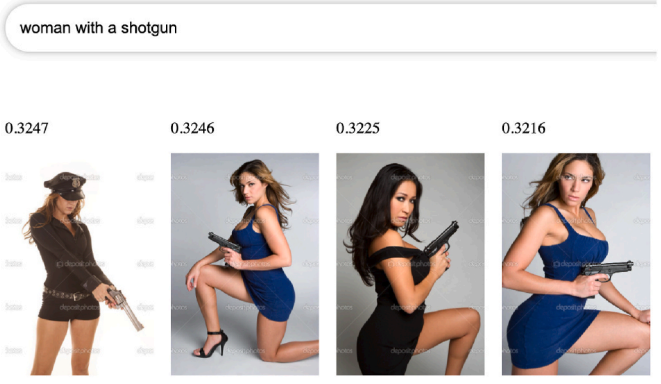
Use Case 1 focuses on searching for images in which at least two objects appear in a specific relationship, e.g. “man holding a pistol”, “a gun in a holster”. To this end, we evaluate queries that describe various relationships between objects. In addition, we assess the impact of conceptual refinement on search quality.

As we can observe in Table 2, CLIP not only searches well for images containing specified objects or people, but also copes well with determining the relationship between them. For example, the query “man holding a pistol” has a precision of 57 %, and incorrect matches often refer to images in which the gender of the person is difficult to determine. A similar query, “woman holding a pistol,” yields slightly worse results, but it is worth noting that in this case, CLIP is able to recognize images that only show a woman's hand (e.g., with painted nails or rings), so gender recognition is possible based on such attributes. It should also be emphasized that CLIP identifies the relationship between objects with high accuracy.

Interestingly, while we would expect that specifying more detailed attributes should lead to more accurate results, this is not always the case. For example, for queries such as “a man holding a rifle” or “woman with a shotgun” (which are more specific versions of the concept of “weapon”), CLIP does not return the desired images in the top K results, leading to precision and recall equal to 0. This difference is illustrated in Fig. 3. The most probable cause of such problems is a lack or a sufficient amount of images and texts related to “shotgun” and “rifle” in the



(a) Search for a query "woman holding a pistol"



(b) Search for a query: "woman with a shotgun"

Fig. 3. Sample queries to CLIP model from Use Case 1.

training data.

For most queries in this Use Case, VLM produces significantly worse results than CLIP. For many queries (e.g. "woman holding a pistol"), VLM does not find any results, and in others (e.g. "a man holding a pistol"), it finds only 5 of the 440 expected images. For a better understanding of these results, it should be noted that searching with VLM means searching through textual captions of images. Therefore, the query must match specific expected words rather than semantic similarities as in the case of CLIP. Qualitative analysis of the generated captions for expected images, VLM usually describes a man as a "person", and present verbs in the present tense (e.g. "A woman holds a black pistol in her hands"). One possible improvement is leveraging synonyms during search performance. As we can see in Table 2, VLM with synonyms (VLMs) achieves better results. For example, for "man holding a pistol", VLMs finds 55 images instead of 5, improving recall from 0.01 to 0.12. However, in most cases VLMs is unable to match the results achieved with CLIP.

Another notable example of query in this use case is "a gun in a holster". CLIP not only correctly identifies both objects, but also recognizes the relationship between them, finding 7 out of 13 images in the top 13 results, achieving the recall and precision of 54 %. On the other hand, searching through VLM-generated captions does not allow capturing such relations. One reason is the fact that the word "in" is treated as a stop word and omitted during the search process leveraging Elasticsearch. It leads to finding only two relevant images, with very low precision and recall (15 %). This case also allows us to observe very interesting behavior related to leveraging synonyms in the search. As we can see, VLMs does not find any image in the top 13 results, leading to precision and recall equal to 0 %. This is because the images found before appear further down in the search results, outside of the top 13 results. This change in order of results is directly related to the use of synonyms.

We also emphasize that the interpretation of the metrics requires careful consideration, as certain queries initially appear to yield inefficient results. However, a more detailed analysis, allowing for a larger number of retrieved images, reveals a different perspective. For example, for the query "a person carrying a suspicious bag" (see Fig. 4), CLIP retrieves only one of the four relevant images within the top four most similar results. However, when extending the evaluation to the top 13 retrieved images, all four relevant images are found. Even though the resulting Precision@13 is relatively low (approximately 0.30), such performance can still be considered acceptable in practical scenarios, where reviewing a small set of around a dozen images to locate all relevant cases remains feasible and efficient.

5.2. Use case 2

Use Case 2 aims to verify how specifying the characteristics of an object affects the precision of the search. A simple query for a general object (for example, 'car' or 'person') can lead to having too many found images making it cumbersome to analyze. However, both models offer the possibility of specifying the appearance or characteristics of objects.

For example, the query "car with license plate" yields very promising results for CLIP, which finds 208 out of 459 relevant images, achieving precision and recall of 45 %. At the same time, VLMs is able to identify 42 relevant images. Analyzing results for other queries such as "a red sedan car" and "white BMW", we can observe that the VLM and VLMs do not have the capability to handle very specific descriptions. On the other hand, CLIP achieves much better results for both cases, finding 14 out of 23 and 4 out of 8 images for "a red sedan car" and "a white BMW" respectively.

Focusing on searching person, we adopt precise description of appearance, creating queries such as "man in a gray hoodie and jeans" and "woman in a red shirt and dark pants with a hat on her head". Such precise queries in the case of the CLIP model return many results that only partially match the description (Fig. 5). For example, we find a man in jeans and a sweatshirt (instead of hoodie) and a woman with a hat on her head wearing a dress (instead of pants). Focusing on VLM, both versions, with and without synonyms, do not find any relevant images for both queries. This result is consistent with our observations from the previous use case, where the specificity of full-text search and a certain selectivity of image description to text form make it more difficult for VLM to find a complete set of matching files. However, in the real-world investigation, it is possible to use an iterative approach, in which single words are added to queries gradually. - since such a search naturally returns fewer matches, it can be assumed that fewer details in the query will yield sufficiently good matches.

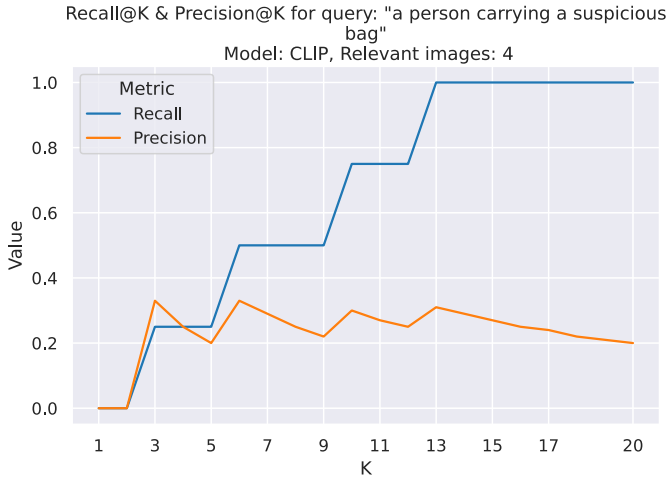
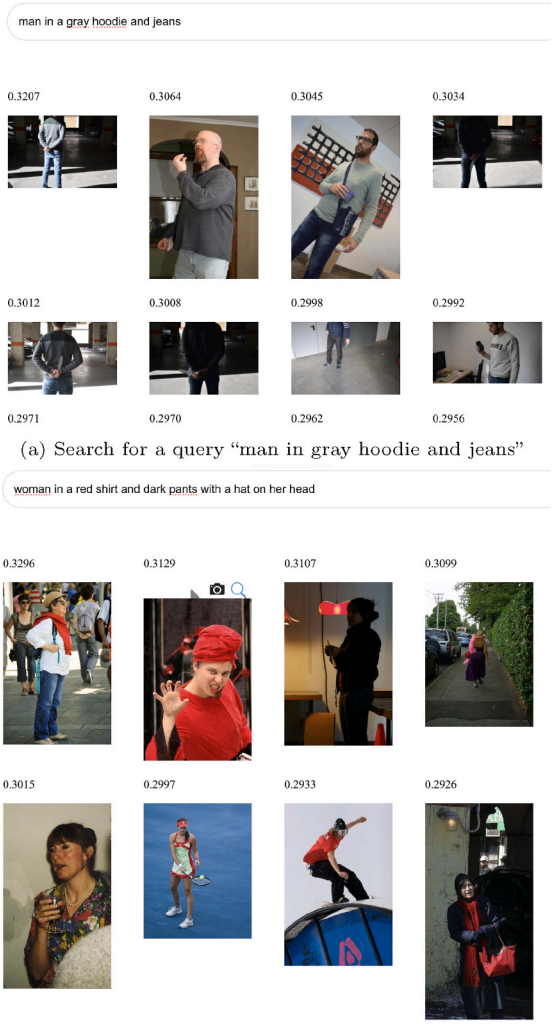


Fig. 4. Precision@K and Recall@K for multiple values of K for query "a person carrying a suspicious bag".



(a) Search for a query "man in gray hoodie and jeans"

(b) Search for a query: "woman in a red shirt and dark pants with a hat on her head"

Fig. 5. Sample queries to CLIP model from Use Case 2.

5.3. Use case 3

Use Case 3 focuses on recognizing the type of image content rather than specific objects or scenes. In this context, we evaluate the models' ability to identify screenshots of WhatsApp conversations (with various subconditions) and email messages, both of which are common and relevant elements in digital forensic analyses.

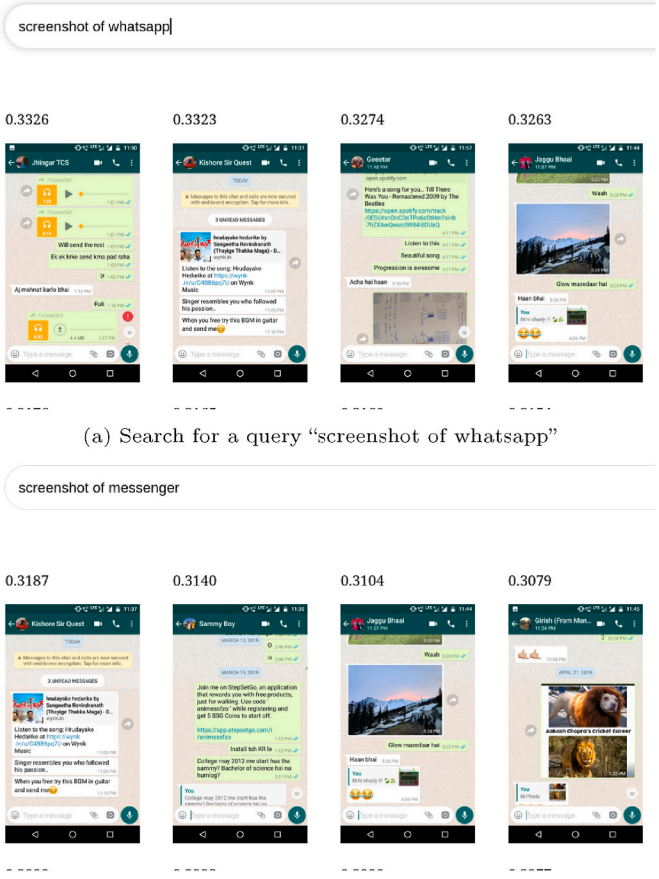
As we can observe in Table 2, VLM does not retrieve any relevant images for the given queries in UC3. However, when a synonym-based query reformulation is applied, the model successfully identifies approximately 47 % of WhatsApp screenshots. This improvement highlights the sensitivity of caption-based retrieval to linguistic variations and the dependence of such models on query phrasing. Qualitative analysis of the captions generated for WhatsApp screenshots reveals uniform outputs such as "a page displaying a conversation in a social/chat app." Although these descriptions are factually correct, their generality prevents precise differentiation and search using application names. Consequently, relevant images could only be retrieved through broader queries, such as "conversation/chat app." This case highlights a key limitation of caption-based retrieval: while it provides semantically valid and human-readable descriptions, it lacks the necessary granularity for domain-specific forensic searches. Moreover, caption-based retrieval requires putting much focus on crafting queries as well as preparing a synonyms dictionary.

In contrast, CLIP demonstrates near-perfect performance in identifying WhatsApp screenshots using the simple query "screenshot of WhatsApp." Nonetheless, its performance degrades when the query becomes more specific ("screenshot of WhatsApp with attachment" or "screenshot of WhatsApp with audio attachment.)" These results suggest that while CLIP effectively captures the general concept, it struggles with fine-grained contextual details that are less visually and semantically distinguishable.

To further investigate the behavior of the model, we conduct additional experiments by adjusting the query formulations to test semantic consistency and generalization. Interestingly, when queried with "screenshot of Messenger," CLIP retrieves WhatsApp screenshots but assigns them lower similarity scores (see Fig. 6). This behavior aligns with the principles of semantic retrieval, as the model recognizes both "WhatsApp" and "Messenger" as conceptually related entities: they are both messaging applications under the same parent company (Meta). However, from a forensic perspective, this blend is problematic, as it can lead to false positives between distinct communication platforms.

5.4. Times comparison

Table 3 presents the execution time results for all evaluated approaches, covering: i) time of processing images by model ("Model processing"); ii) time of indexing data into database (Indexing); iii) Synonyms creating, only for VLMs (Synonyms); iv) Time required to search through the database, averaged per a single query (Avg per query); v) Time required to search through the database, divided by the number of retrieved images in each query (Avg per retrieved image). As expected, substantial differences can be observed between multimodal models such as VLM/VLMs and the CLIP-based approach.



(a) Search for a query "screenshot of whatsapp"

(b) Search for a query: "screenshot of messenger"

Fig. 6. Sample queries to CLIP model from Use Case 3.

Table 3
Execution time results for all evaluated approaches.

	Model processing	Indexing time	Creating Synonyms	Avg per query	Avg per retrieved image
VLM	27.51 h	41.27 s	–	6.1 ms	0.7 ms
VLMs	27.51 h	41.27 s	3 h	49.8 ms	3.9 ms
CLIP	0.08 h	61.76 s	–	2175 ms	53.2 ms

Both VLM and VLMs require approximately 27.5 h for model processing, primarily due to the computational complexity of extracting image features and then leveraging large language model to generate captions. In contrast, CLIP demonstrates significantly faster processing, completing this stage in less than 0.1 h. The difference in processing time is related also to the size of model, as the model used in CLIP approach has around 88 million of parameters, while Paligemma2 leveraged in VLM approaches has around 3 billion parameters.

The indexing phase was relatively similar between all methods, taking around 40–60 s, indicating that the differences in total execution time are primarily due to the feature extraction step. However, the introduction of synonym generation in VLMs resulted in an additional 3 h of preprocessing, reflecting the cost of enhancing query flexibility.

When considering retrieval efficiency per query, CLIP exhibited the highest average time (2.2 s per query), mainly due to the need to use the text encoder for every query and its heavier text-image embedding comparison overhead. In contrast, VLM-based search does not require involvement of model, as the text is already indexed in Elasticsearch and the only execution time related to search is due to elasticsearch search. Due to that, the VLM-based search achieves considerably lower latencies (6–50 ms per query). Similarly, the average processing time per retrieved image remained below 5 ms for both VLM-based methods, compared to more than 50 ms for CLIP.

Overall, both approaches demonstrate satisfactory retrieval efficiency suitable for real-world forensic scenarios, where quick access to relevant visual evidence is essential. While VLM-based methods provide very fast query responses thanks to efficient Elasticsearch indexing, CLIP stands out in terms of data processing speed, completing embeddings extraction in a fraction of the time required by VLM. This contrast highlights a practical trade-off: VLM offers slightly faster retrieval during use, whereas CLIP enables faster system preparation and indexing, as well as better retrieval performance.

5.5. Limitations of experimental results

The research shows differences between the models in terms of application in digital forensics, including triage. However, we acknowledge the inherent limitations of experimental results. Initially, the utilization of a dataset comprising diverse everyday photographs facilitates the evaluation of general cases (a central objective of this study), yet it doesn't show the capacity of models to process specific queries within a collection of similar and domain-specific images. Next, to simplify results analysis, VLM was prompt using a simple and general query „caption" – in specific cases the prompt can be improved and adjusted to particular use cases what probably would improve results precision. Finally, a comparison of a single model from CLIP and VLM enables the conclusions previously outlined. However, a comparison of various model variants and sizes will result in the development of guidelines for selecting a model for a specific case, determining the volume of analyzed data, and assessing the available computing infrastructure. These limitations, therefore, become the focal points of our research directions.

6. Conclusion

This study presented a comprehensive evaluation of two fundamentally different artificial intelligence approaches: Contrastive Language-Image Pre-training (CLIP) and Vision-Language Models (VLM) for large-scale forensic image retrieval. By analyzing 33 representative queries across three distinct use cases and nine heterogeneous datasets containing over 80 000 images, we assessed their performance in terms of retrieval accuracy, generalization, and computational efficiency.

For universal open-vocabulary queries, CLIP proves to be the more suitable tool. It processes data significantly faster, is purpose-built for semantic image retrieval, and delivers acceptable-quality results even without task-specific tuning. Although CLIP occasionally produces inaccurate matches, increasing the number of retrieved results often ensures retrieval of all relevant images. Its primary limitation lies in handling highly specific or underrepresented concepts, particularly those absent from its training distribution, where retrieval precision can degrade. An interesting yet demanding future research direction is finetuning CLIP specifically for forensic applications, including more detailed description of potentially relevant content (for example describing specific types of guns instead of using a generic words). In this context, it is noteworthy that CLIP can be successfully applied not only in the long-lasting, detailed forensic analyses, but also during the triage process. The summary times of model processing, indexing, and query time, in conjunction with sufficient precision for common cases, are suitable for such usage.

On the other hand, VLM is a much more sophisticated tool capable of generating detailed textual descriptions and supporting more complex reasoning over image content. However, these benefits come at the cost of substantially longer processing times and less effective text-based retrieval when used with generic prompts. Still, our experiments demonstrate that introducing synonym expansion can partially mitigate these issues, improving search coverage and recall.

The use of VLM models is not an unjustified direction for further research. In this study, we leverage a general captioning prompt. However, the use of dedicated prompts can enable more domain-oriented captioning, detection of specific objects, and even the construction of conditional image processing chains depending on the content.

In future work, besides directions mentioned in Section 5.5, we plan to extend this study toward hybrid architectures that integrate both paradigms, and to explore new tasks such as searching for images containing embedded text, either from document scans or natural scene images or other typical objects for criminal analysis such as electronics, chemical substances or drugs.

Acknowledgment

The research leading to these results has received funding from the project No. DOB-SZAFIR/11/A/036/01/2020 "Advanced techniques for analysing information obtained from digital media, in particular from mobile devices.", funded by the Polish National Centre for Research and Development.

References

Beyer, L., Steiner, A., Pinto, A.S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., Unterthiner, T., Keysers, D., Koppula, S., Liu, F., Grycner, A., Gritsenko, A., Houlsby, N., Kumar, M., Rong, K., Eisenschlos, J., Kabra, R., Bauer, M., Bošnjak, M., Chen, X., Minderer, M., Voigtlaender, P., Bica, I., Balazevic, I., Puigcerver, J., Papalampidi, P., Henaff, O., Xiong, X., Soricut, R., Harmsen, J., Zhai, X., 2024. Paligemma: A versatile 3b vlm for transfer. ArXiv preprint, arXiv:2407.07726.

Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L., 2023. Raising the bar of ai-generated image detection with clip. IEEE, pp. 4356–4366.

Cui, X., Li, Y., Luo, A., Zhou, J., Dong, J., 2025a. Forensics adapter: adapting clip for generalizable face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 19207–19217.

Cui, X., Li, Y., Zhu, D., Zhou, J., Dong, J., Lyu, S., 2025b. Forensics adapter: unleashing clip for generalizable face forgery detection. https://arxiv.org/abs/2411.19715, arXiv:2411.19715.

Del Mar-Raave, J.R., Bahşi, H., Mršić, L., Hausknecht, K., 2021. A machine learning-based forensic tool for image classification - a design science approach. Forensic Sci. Int.: Digit. Invest. 38, 301265. https://doi.org/10.1016/j.fsidi.2021.301265.

Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L., 2009. Imagenet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Japkowicz, N., Boukouvalas, Z., 2024. Machine Learning Evaluation: towards Reliable and Responsible AI. Cambridge University Press, Cambridge.

McKeown, S., Buchanan, W.J., 2023. Hamming distributions of popular perceptual hashing techniques. Forensic Science International: Digital Investigation 44, 30150.

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML). PMLR, pp. 8748–8763.

Sanna, S.L., Regano, L., Maiorca, D., Giacinto, G., 2024. Exploring the robustness of ai-driven tools in digital forensics: a preliminary study. arXiv:2412.01363.

Sharma, S., Guleria, K., 2022. Deep learning models for image classification: comparison and applications. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE). IEEE, pp. 1733–1738.

Steinebach, M., 2023. An analysis of photodna. In: Proceedings of the 18th International Conference on Availability, Reliability and Security. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3600160.3605048.

Steiner, A., Susano Pinto, A., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., Qin, S., Ingle, R., Bugliarello, E., Kazemzadeh, S., Mesnard, T., Alabdulmohsin, I., Beyer, L., Zhai, X., 2024. Paligemma 2: a family of versatile vlms for transfer. arXiv preprint arXiv: 2412.03555.

Taunk, K., De, S., Verma, S., Swetapadma, A., 2019. A brief review of nearest neighbor algorithm for learning and classification. In: 2019 International Conference on Intelligent Computing and Control Systems (ICCS), pp. 1255–1260. https://doi.org/10.1109/ICCS45141.2019.9065747.

Vasilaras, A., Papadoudis, N., Rizomiliotis, P., 2024. Artificial intelligence in mobile forensics: a survey of current status, a use case analysis and ai alignment objectives. Forensic Sci. Int.: Digit. Invest. 49, 301737. https://doi.org/10.1016/j.fsidi.2024.301737. URL: https://www.sciencedirect.com/science/article/pii/S2666281724000568.

Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., Shen, D., 2025. Clip in medical imaging: a survey. Med. Image Anal. 102, 103551. https://doi.org/10.1016/j.media.2025.103551.