

Vision-Language Models for Cross-Modal Retrieval in Intelligent Surveillance Systems

Gurumurthy M^{1*}, Sundararajan R² and Kalyanasundaram D³

¹Department of Intelligent Systems, Kabini Institute of Autonomous Robotics, Philippines Email:

madhavi.gurumurthy@kabini.edu

ORCID: <https://orcid.org/6324-7753-6759-9790>

²Department of Intelligent Systems, Periyar Institute of Autonomous Systems Research, Nepal Email:

rajesh.sundararajan@periyar.edu

ORCID: <https://orcid.org/2363-3101-2032-9319>

³Department of Robotics and Automation, Chaliyar Institute of Quantum Computing, Nepal Email:

deepa.kalyanasundaram@chaliyar.edu

ORCID: <https://orcid.org/2644-9882-6612-366X>

ABSTRACT

Vision-language models (VLMs) that learn a joint embedding space for images and free-form text from web-scale, weakly aligned corpora have reshaped cross-modal retrieval, and intelligent video surveillance is one of the domains where that shift matters most in practice: an operator searching archived footage for a natural-language description, a reference photograph, or a short example clip needs exactly the similarity-ranking capability these models provide. This article synthesizes the literature connecting large-scale vision-language pretraining, from contrastive dual encoders such as CLIP and ALIGN through bootstrapped and grounding-oriented architectures such as BLIP and GLIP, to the surveillance-specific retrieval problems these backbones now underpin: text-based person search across disjoint camera views, open-vocabulary object and vehicle detection, text-to-video retrieval for reviewing archived footage, and language-conditioned framings of anomaly recognition. We trace the transformer-based fusion architectures that preceded and paralleled contrastive pretraining, examine how the fine-grained attribute matching that person search demands has been addressed through attribute disentanglement and implicit relation reasoning, and report the specific, source-verified numeric results the field has published. We close by identifying open challenges -- the domain gap between web imagery and surveillance footage, fine-grained discrimination at gallery scale, computational cost, cross-camera robustness, and privacy -- that remain substantially unresolved despite the rapid progress in the underlying vision-language foundation models.

Keywords: vision-language models; cross-modal retrieval; video surveillance; text-based person search; open-vocabulary detection

Citation: Gurumurthy M, Sundararajan R and Kalyanasundaram D [2024]. Vision-Language Models for Cross-Modal Retrieval in Intelligent Surveillance Systems. DOI: <https://doi.org/10.5281/zenodo.22843899>

Copyright: © 2024 by the authors. Open access under [CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/).

Article Information: Received: 26 April 2024 Accepted: 29 July 2024 Published: 4 November 2024

Article Class: Research Article

1. Introduction

A modern surveillance network produces far more video than any human operator can watch, and the bottleneck this creates is not sensing but search: given a text description, a sample photograph, or a short clip, find the matching person, object, or event somewhere inside days or weeks of multi-camera footage. For most of the history of video analytics that search problem was solved, if at all, by attribute tagging and closed-set classification — a fixed vocabulary of colors, clothing categories, or vehicle types against which an operator's query had to be manually translated. Vision-language models (VLMs) change the shape of that problem. By learning a joint embedding space for images and free-form text from very large, weakly aligned corpora, models such as CLIP and ALIGN made it possible, for the first time at scale, to rank a gallery of images against an arbitrary natural-language query without any task-specific labels at all, simply by measuring similarity in a shared representation space learned from hundreds of millions to billions of image-text pairs [1,2]. Cross-modal retrieval — text-to-image, image-to-text, and increasingly text-to-video — is the most direct beneficiary of this shift, and intelligent surveillance is one of the domains where the practical stakes of getting it right are highest: a forensic search for "a man in a red jacket carrying a black backpack near the north entrance" is exactly the kind of query these models were built to answer.

This article surveys the literature that connects large-scale vision-language pretraining to cross-modal retrieval specifically as it applies to surveillance-style tasks: person search by natural-language description, open-vocabulary object and vehicle detection in camera feeds, text-to-video retrieval for reviewing archived footage, and the anomaly and activity recognition problems that increasingly get framed as language-conditioned retrieval rather than fixed-class classification. The story has two intertwined threads. The first is the rapid maturation of general-purpose vision-language foundation models — contrastive models like CLIP and ALIGN, generative-contrastive hybrids like BLIP and CoCa, and grounding-oriented models like GLIP and Grounding DINO — each trained once on web-scale data and then reused, largely without retraining, across a wide range of downstream retrieval tasks [3,4,5,6]. The second is a more specialized, decade-long line of research on text-based person

search that predates CLIP by several years and has, more recently, been substantially reshaped by it: models purpose-built to match a pedestrian's textual description against a gallery of cropped person images captured across disjoint, non-overlapping cameras, the canonical retrieval task inside a surveillance network [7,8,9].

We organize the discussion around four themes. First, we trace the foundations of large-scale vision-language pretraining — contrastive dual-encoder architectures and the transformer-based fusion architectures that preceded and paralleled them — because nearly every retrieval system discussed later either builds directly on one of these backbones or is evaluated against it as a baseline. Second, we examine the architectural lineage of cross-modal fusion, from early region-word alignment models through to bootstrapped and momentum-distilled pretraining, since the fusion strategy a model uses has direct consequences for whether it is fast enough for gallery-scale retrieval or accurate enough for fine-grained person search. Third, we turn to the surveillance-specific literature: text-based person retrieval, open-vocabulary detection and grounding applied to scene understanding, video-text retrieval for archived footage, and vision-language framings of anomaly detection. Fourth, we lay out the open challenges — domain gap between web-scale pretraining data and surveillance imagery, fine-grained attribute discrimination at person scale, computational cost at gallery scale, and privacy — that the literature has identified but by no means solved. We close with a conclusion that draws these threads together.

As with any literature synthesis of a fast-moving area, we are reporting what has been published and verified, not conducting new experiments. Where we cite a specific numeric result — a recall figure, an average-precision score, a relative improvement — it is drawn directly from the cited source's own reported results, and we flag explicitly where the surveillance-specific literature has, in our reading, not yet caught up to the pace of the general-purpose vision-language literature it increasingly depends on.

It is worth being explicit about what "surveillance" means for the purposes of this synthesis, because the term covers a wide range of settings with very different data characteristics. We are primarily concerned with fixed or slowly re-tasked camera networks — transit

systems, retail and campus security, municipal traffic and public-safety monitoring — where the retrieval problem is searching archived or live footage after the fact, rather than with body-worn or mobile robotic vision, where the dominant problems (real-time inference on embedded hardware, egomotion, continuously shifting scene composition) are different enough to warrant separate treatment. Within that scope, the retrieval query itself can take several forms — a free-text description typed by an operator, a reference photograph of a person or vehicle of interest, or a short example clip of a behavior — and one of the more interesting developments the literature below documents is the degree to which a single pretrained vision-language backbone can, with comparatively modest task-specific adaptation, serve all three query modalities rather than requiring a separate purpose-built system for each.

2. Foundations of Vision-Language Pretraining

The architectural idea underlying most modern cross-modal retrieval systems is deceptively simple: train two encoders, one for images and one for text, so that the embeddings they produce for a matching image-text pair are close together in a shared space and embeddings for non-matching pairs are far apart. CLIP, introduced by Radford et al., operationalized this idea at a scale previous work had not attempted, training a dual-encoder model contrastively on 400 million image-text pairs scraped from the internet, with no manual annotation beyond the pairing itself [1]. The resulting model could be used zero-shot — with no task-specific fine-tuning at all — for image classification, by comparing an image embedding against text embeddings of candidate class-name prompts, and the paper reports that CLIP's zero-shot performance matches that of the original supervised ResNet-50 on ImageNet, a model trained on 1.28 million labeled examples that CLIP never saw. For retrieval specifically, this same mechanism is exactly what is needed: encode a gallery of images once, encode a query (text or image) once, and rank the gallery by cosine similarity, an operation that scales to very large galleries because the expensive encoding step is amortized and the comparison itself is cheap.

ALIGN, from Jia et al., pursued a structurally similar contrastive objective but leaned even harder into scale as a substitute for data curation, training on over one

billion noisy image alt-text pairs collected without the expensive filtering pipeline CLIP's dataset construction used, and reported that despite this apparent noise, ALIGN set new state-of-the-art results on the Flickr30K and MSCOCO image-text retrieval benchmarks at the time of publication [2]. The comparison between CLIP and ALIGN is instructive for anyone thinking about deploying these models on surveillance-style data, because it suggests that contrastive pretraining is comparatively forgiving of noisy supervision at sufficient scale — a property that matters because surveillance deployments will rarely have access to a curated, CLIP-scale corpus of camera-domain image-text pairs, and will more plausibly have access to noisier, smaller, in-domain data that needs to be combined with general-purpose pretraining rather than replace it.

Table 1: Representative vision-language and cross-modal retrieval architectures relevant to surveillance-style search, the scale of their pretraining or design focus, and the specific numeric result each cited source reports. [1,2,3,5,20,9].

Model	Source	Pretraining data / design focus	Reported result
CLIP (contrastive dual encoder)	Radford et al. [1]	400M web image-text pairs	Matches supervised ResNet-50 zero-shot top-1 accuracy on ImageNet without using its 1.28M training labels
ALIGN (contrastive, noisy alt-text)	Jia et al. [2]	Over 1B noisy image alt-text pairs	New state-of-the-art results on Flickr30K and MSCOCO image-text retrieval at publication
BLIP (bootstrapped captions, CapFilt)	Li et al. [3]	Web image-text pairs filtered/re-captioned by the model itself	+2.7% average recall@1 retrieval, +2.8% CIDEr captioning, +1.6% VQA score over prior best
GLIP (grounded language-image pre-training)	Li et al. [5]	27M pairs: 3M human-annotated + 24M web-crawled grounding data	49.8 AP zero-shot on COCO; 60.8 AP on COCO val after fine-tuning
SCAN (stacked cross attention)	Lee et al. [20]	Region-word cross-attention matching, Flickr30K/MSOCO	+22.1% / +18.2% relative Recall@1 text/image retrieval on Flickr30K

Model	Source	Pretraining data / design focus	Reported result
IRRA (implicit relation reasoning)	Jiang & Ye [9]	Text-to-image person retrieval, three benchmarks	About 3-9 percentage point Rank-1 accuracy gain over prior state of the art

A parallel line of work treats vision-language pretraining as a joint contrastive-and-generative problem rather than contrastive-only. BLIP introduced a bootstrapping mechanism, CapFilt, that uses a captioner to generate synthetic captions for web images and a filter to discard noisy ones, iteratively cleaning its own training data, and reported gains of +2.7% in average recall@1 on image-text retrieval, +2.8% in CIDEr on image captioning, and +1.6% in VQA score over the prior state of the art at publication [3]. BLIP-2 extended this line by freezing large pretrained image encoders and large language models and training only a lightweight "Querying Transformer" bridge between them, substantially reducing the compute needed to reach strong vision-language performance by reusing off-the-shelf frozen components rather than training a full model end to end [10]. CoCa unified the contrastive and generative objectives into a single encoder-decoder architecture trained jointly, reporting strong zero-shot and fine-tuned performance across recognition, retrieval, and captioning tasks from one set of weights [4], while Flamingo demonstrated that a frozen vision encoder can be interleaved with a frozen large language model via cross-attention layers to support few-shot, in-context adaptation to new visual-language tasks without any gradient updates to the underlying pretrained components at all [11]. Table 1 places several of these models, together with two of the surveillance-adjacent retrieval systems discussed later, side by side on the scale of their pretraining data and the specific numeric result each source reports.

Underlying essentially all of these architectures is the Transformer, introduced by Vaswani et al. as a sequence model built entirely from attention rather than recurrence or convolution [12], and the Vision Transformer (ViT), which showed that a nearly unmodified Transformer applied directly to sequences of image patches, rather than a convolutional backbone, could match or exceed convolutional networks on image recognition when pretrained at sufficient scale [13]. The image encoders in CLIP, ALIGN, BLIP, and most subsequent vision-language models are either ViT

variants or convolutional networks retrofitted with the same large-scale contrastive training recipe, and the recent survey by Zhang et al. of vision-language models for vision tasks gives a systematic taxonomy of this design space — pretraining objectives, architectures, and the transfer-learning and knowledge-distillation techniques used to adapt a general-purpose VLM to a specific downstream recognition or retrieval task — that is a useful map for anyone trying to situate a surveillance-specific retrieval system within the broader field [14].

3. Vision-and-Language Transformer Architectures

Before contrastive dual-encoder pretraining became dominant, a substantial body of work explored a different architectural strategy: fuse visual region features and textual tokens early, inside a single cross-modal transformer, rather than encoding each modality separately and comparing embeddings afterward. ViLBERT extended BERT-style pretraining to a two-stream architecture, processing image regions and text tokens in separate transformer streams that interact through co-attentional layers, and pretrained on the Conceptual Captions dataset before transferring to visual question answering, visual commonsense reasoning, and caption-based retrieval with strong results across all four evaluated tasks [15]. LXMERT pursued a closely related single-stream-of-cross-attention design, pretraining a cross-modality encoder with several masked and matching objectives simultaneously [16]. UNITER generalized this family further with four pretraining objectives spanning masked language modeling, masked region modeling, image-text matching, and word-region alignment via optimal transport, reporting state-of-the-art results across six vision-and-language task families and nine datasets from a single pretrained backbone [17]. Oscar added a further refinement: rather than relying on the model to implicitly discover which image regions correspond to which words, Oscar uses object tags detected in the image (via an off-the-shelf detector) as explicit anchor points to make that alignment easier to learn, reporting improved downstream performance on retrieval and captioning as a direct consequence of this simplification [18].

The region-based image encoders that these fusion architectures depend on were themselves built on the

bottom-up attention mechanism introduced by Anderson et al., which uses a Faster R-CNN detector trained on Visual Genome to extract a variable-length set of salient image-region features that downstream captioning and VQA models then attend over top-down; this bottom-up feature set won the 2017 VQA Challenge and set a new state of the art in image captioning on MSCOCO at the time [19]. (We return to this architectural lineage below in the context of person search, where the same region-versus-global tension between fine-grained local features and holistic embeddings recurs directly.)

A key limitation of the early fusion approach, deep cross-attention between every region and every token before producing a similarity score, is computational: it is well suited to re-ranking a small candidate set but poorly suited to searching a gallery of millions of images, because the fusion has to be recomputed for every query-candidate pair rather than amortized. Stacked Cross Attention (SCAN), from Lee et al., addressed this specifically for image-text matching by discovering the full latent alignment between image regions and sentence words through attention computed with each modality's elements used as context for the other, rather than treating similarity as a single global comparison, and reported substantial gains over prior methods: 22.1% relative improvement in text retrieval and 18.2% relative improvement in image retrieval at Recall@1 on Flickr30K, and 17.8% and 16.6% relative improvements respectively on the MSCOCO 5K test split [20]. ALBEF partially reconciled the efficiency-versus-accuracy tension between dual-encoder and fusion architectures by aligning image and text representations with a contrastive loss *before* fusing them through cross-modal attention, plus a momentum distillation mechanism to make the model more robust to the noisy web image-text pairs it was trained on, reporting improved performance across retrieval and vision-language reasoning tasks relative to prior fusion-only baselines [21]. This two-stage pattern — cheap global filtering with a contrastive dual encoder, followed by expensive fine-grained re-ranking with a fusion model on only the top candidates — has become the standard practical recipe for making transformer-fusion accuracy affordable at retrieval scale, and it recurs directly in the person-search literature discussed next.

4. Cross-Modal Person and Object Retrieval for

Surveillance

Text-based person search is the surveillance domain's most direct and long-standing instance of cross-modal retrieval: given a natural-language description of a pedestrian, retrieve the matching cropped-person images from a gallery collected across multiple, non-overlapping camera views, a task that arises directly whenever an eyewitness description rather than a reference photograph is the only evidence available to search on. Li, Xiao, and colleagues introduced this task formally along with the CUHK-PEDES benchmark dataset and a gated neural attention RNN baseline, establishing both the problem formulation and the first large-scale annotated dataset that the subsequent decade of work has been built and evaluated on [7]. The task is meaningfully harder than generic image-text retrieval: the visual differences between the correct match and a hard negative are frequently confined to a small region of the image (a bag's color, a logo on a shirt, the pattern on a scarf), and standard image-text retrieval systems trained on captioning-style datasets like MSCOCO or Flickr30K, whose captions describe whole scenes rather than fine-grained garment and accessory attributes, transfer to this setting comparatively poorly without further adaptation.

ViTAA, from Wang et al., addressed this fine-grained attribute problem directly by adopting an attribute-aligning perspective: rather than treating person search as generic whole-image-to-whole-text matching, ViTAA disentangles the person's feature space into subspaces corresponding to distinct visual attributes (using an auxiliary attribute segmentation branch for supervision) and aligns each subspace with the corresponding attribute phrase in the description via contrastive learning, reporting state-of-the-art performance on both natural-language and structured attribute-phrase person-search queries [8]. This attribute-disentanglement strategy is one of two dominant approaches the field converged on for handling fine granularity; the other, exemplified by IRRa, keeps the representation unified but strengthens the *reasoning* about implicit cross-modal relations. Jiang and Ye's IRRa framework performs implicit relation reasoning via a masked-language-modeling-style module combined with a cross-modal encoder that fuses visual information directly into textual tokens, and aligns the resulting distributions using a similarity distribution matching objective based on KL divergence,

without requiring any additional attribute-level supervision beyond the paired description itself; the paper reports improvements of roughly 3 to 9 percentage points in Rank-1 accuracy over prior state-of-the-art methods across three standard text-to-image person-retrieval benchmarks [9].

The arrival of CLIP-scale pretraining reshaped this subfield directly. CFine, from Yan et al., builds a text-to-image person re-identification framework driven by a CLIP backbone, performing fine-grained information excavation that mines both intra-modal discriminative cues and inter-modal correspondences directly in the original CLIP embedding space rather than mapping into a newly learned joint space, with multi-grained global feature learning, cross-grained feature refinement, and fine-grained correspondence discovery modules layered on top [22]. That a general-purpose web-pretrained contrastive model transfers usefully, with lightweight adaptation, to a domain as visually narrow and fine-grained as pedestrian attribute matching is itself informative: it suggests the representations CLIP-style pretraining learns from open-domain web images retain enough low-level texture, color, and part-level structure to be useful even for a task whose training distribution (cropped, low-resolution, often poorly lit surveillance-camera pedestrian crops) looks nothing like CLIP's web-image training data. Mang Ye and colleagues' broader survey of deep learning for person re-identification situates text-based person search within the wider re-identification literature — image-to-image re-identification across cameras, attribute-based re-identification, and the "open-world" re-identification settings (unlabeled or partially labeled galleries, cross-resolution and cross-modality matching) that most closely resemble what an operational surveillance deployment actually faces, as opposed to the closed-world, cleanly labeled benchmark conditions most re-identification papers still evaluate under [23].

Multimodal fusion more broadly, of which text-based person search is one instance, has its own foundational taxonomy independent of any specific architecture. Baltrušaitis, Ahuja, and Morency's widely cited survey organizes the general multimodal machine learning problem into five core challenges — representation, translation, alignment, fusion, and co-learning — and this taxonomy remains a useful lens for classifying where

a given cross-modal retrieval architecture sits: a dual-encoder model like CLIP is fundamentally solving representation and alignment, while a fusion architecture like UNITER or VITAA's attribute-disentanglement module is additionally solving fusion in the sense the survey defines it [24].

5. Open-Vocabulary Detection, Video Retrieval, and Anomaly Recognition

Person search assumes the query is a natural-language description of a pedestrian; a broader class of surveillance retrieval problems requires locating and naming arbitrary objects, vehicles, or events in a scene without a fixed, pre-enumerated label set, since a real deployment cannot anticipate every object category an operator might eventually want to search for. Open-vocabulary object detection addresses exactly this gap by using vision-language pretraining to detect object categories that were never seen with bounding-box supervision during training. ViLD, from Gu et al., distills knowledge from a pretrained vision-language classification model (the teacher) into a two-stage detector (the student), aligning the student's region embeddings with the teacher's image and text embeddings so that the detector can recognize novel categories described only in natural language at inference time; the paper reports 16.1 mask AP on rare LVIS categories with a ResNet-50 backbone, exceeding a supervised baseline by 3.8 points, and 26.3 AP on rare categories when using ALIGN as the stronger teacher model [25].

GLIP unifies object detection and phrase grounding into a single pretraining objective, treating detection as a grounding problem where the "vocabulary" is simply whatever noun phrases appear in a paired caption, and pretrains on a combination of 3 million human-annotated and 24 million web-crawled image-text pairs; without any task-specific fine-tuning, GLIP reports 49.8 AP on COCO and 26.9 AP on LVIS, and after fine-tuning on COCO reaches 60.8 AP on the validation set and 61.5 AP on test-dev, surpassing the prior state of the art [5]. Grounding DINO extends this grounded-pretraining idea onto the stronger DINO detector architecture, combining a feature enhancer, language-guided query selection, and a cross-modality decoder, and reports 52.5 AP on the COCO zero-shot transfer benchmark along with new state-of-the-art results on the ODinW open-world detection benchmark

suite [6]. For a surveillance operator, the practical consequence of this line of work is direct: a camera feed can, in principle, be searched for "unattended luggage," "a person climbing a fence," or "a delivery van," categories that would previously have required collecting bounding-box training data for each new query type, simply by describing the target in language at query time. Figure 1 illustrates the scale of the accuracy gap that grounded pretraining closes between a purely zero-shot open-vocabulary detector and one given task-specific fine-tuning, using GLIP's own reported COCO results as the concrete anchor.

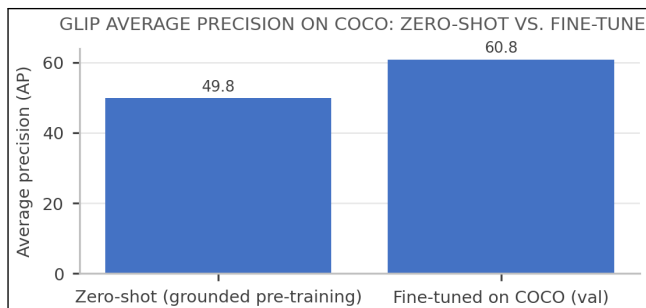


Figure 1: GLIP's reported average precision (AP) on COCO object detection, comparing zero-shot grounded pre-training against the same model after task-specific fine-tuning on COCO -- the scale of accuracy that grounded language-image pre-training still leaves on the table for a downstream detector to recover, and the gap open-vocabulary surveillance object detection has to close without full re-training on every new deployment [5].

Surveillance footage is video, not a single frame, and retrieving the right *moment* inside hours of archived video is a distinct problem from retrieving the right still image. CLIP4Clip adapted CLIP directly to video-text retrieval, empirically studying whether CLIP's still-image representations transfer to video, how large-scale video-text pretraining affects downstream performance, and which temporal-aggregation strategy (mean pooling, sequential modeling, or a tight transformer over frame embeddings) works best, reporting state-of-the-art results on MSR-VTT, MSVD, LSMDC, ActivityNet, and DiDeMo video-text retrieval benchmarks [26]. Frozen in Time trained a joint video-and-image encoder end to end, using a curriculum that begins by treating still images as "frozen" single-frame video and gradually introduces temporal context, and introduced the WebVid-2M dataset of weakly captioned web videos to support this training at scale, reporting state-of-the-art results on the same family of retrieval benchmarks despite training on a smaller video-text corpus than several contemporaneous approaches [27]. ActionCLIP

reframes a closely related problem, video action recognition, explicitly as video-text matching, using label text's semantic content rather than a fixed classifier head, through a "pre-train, prompt, fine-tune" pipeline, and reports 83.8% top-1 accuracy on Kinetics-400 with a ViT-B/16 backbone while retaining the ability to recognize action categories unseen during supervised training [28]. Taken together, this body of work suggests that a surveillance archive-search system built on video-text retrieval can plausibly reuse the same contrastively pretrained backbones already deployed for still-image person and object search, rather than requiring an entirely separate video-specific pipeline.

Anomaly and unusual-event detection in surveillance video has historically been framed as a self-supervised or weakly supervised classification problem rather than a retrieval problem, and Sultani, Chen, and Shah's multiple-instance-learning formulation remains a foundational reference for that framing: using only video-level labels (normal versus anomalous) rather than frame-level annotations, their deep ranking model learns to assign high anomaly scores to anomalous segments, and the paper released a 128-hour, 1,900-video dataset spanning 13 real-world anomaly categories including fighting, robbery, and burglary that remains a standard benchmark in this subfield [29]. The connection between this classification-oriented framing and the retrieval-oriented framing this article otherwise emphasizes is increasingly direct in more recent work: once a vision-language model can score arbitrary text against video content in a shared embedding space, "is this segment anomalous" and "does this segment match the description 'a person falling to the ground'" become the same underlying operation, and we expect — though the literature synthesized here does not yet fully demonstrate it at scale — that future anomaly detection systems will increasingly be built directly on top of the video-text retrieval backbones surveyed above rather than as a separate, purpose-built classification pipeline.

6. Open Challenges for Vision-Language Retrieval in Deployed Surveillance Systems

Several gaps recur across this literature and are worth naming plainly. First, domain gap: essentially every foundation model discussed here — CLIP, ALIGN, BLIP, GLIP, Grounding DINO — is pretrained on web-scraped, high-resolution, well-composed photographs, a distribution that differs substantially from surveillance

imagery, which is frequently low-resolution, poorly lit, captured from oblique overhead angles, and subject to compression artifacts from long-term archival storage. The person-search literature's own trajectory is instructive on this point: CFine's CLIP-driven approach needed fine-grained adaptation modules specifically because directly reusing CLIP's embedding space for cropped, low-resolution pedestrian images was not, on its own, sufficient to match dedicated architectures [22]. How much of this gap can be closed with lightweight adaptation versus how much genuinely requires in-domain pretraining on surveillance-style imagery remains, in our reading, an open empirical question rather than a settled one.

Second, the granularity mismatch between what general-purpose vision-language models were trained to align and what surveillance retrieval queries actually need. Web image-caption pairs describe whole scenes; a forensic person-search query needs to discriminate on a shirt logo, a bag color, or a gait pattern, distinctions the attribute-disentanglement and implicit-relation-reasoning architectures reviewed here were built specifically to address [8,9], but which the general-purpose foundation models were not designed around and do not obviously solve out of the box. Third, gallery-scale computational cost: the strongest cross-modal accuracy in this literature consistently comes from architectures that perform some form of cross-attention fusion between the query and each candidate [20,21], which is computationally prohibitive to run against a gallery of millions of person crops or hours of archived video without a cheaper first-stage filter; the two-stage retrieve-then-rerank pattern that ALBEF and similar architectures popularized is a practical mitigation, but the literature synthesized here reports relatively little systematic study of that tradeoff specifically at surveillance-relevant gallery sizes.

Fourth, privacy and governance. None of the papers surveyed here treat privacy as a central design constraint; person-search and re-identification research in particular optimizes retrieval accuracy on benchmark datasets collected, in some cases, with looser consent and disclosure standards than would be acceptable for an operational deployment, and Ye et al.'s survey is candid that most reported re-identification results still come from closed-world benchmark conditions that do not reflect the open-world, less-controlled data an actual

surveillance network produces [23]. Fifth, evaluation standardization: the video-text retrieval literature reports results on a fairly consistent set of benchmarks (MSR-VTT, MSVD, DiDeMo) [26,27], but no equivalently standardized, surveillance-specific benchmark spanning person search, open-vocabulary object detection, and anomaly retrieval together yet exists, which makes it difficult to compare end-to-end system designs rather than individual component accuracies. Sixth, cross-camera and cross-condition robustness deserves separate mention from the domain-gap point above, because it is a distinct failure mode: even a model that adapts well to surveillance imagery in general still has to match the same person or vehicle across cameras with different color calibration, viewing angle, illumination, and compression settings, a challenge the re-identification literature has studied directly for over a decade under the heading of cross-camera and cross-resolution matching [23], but which the newer CLIP-driven person-search architectures have, to date, evaluated mostly within single-benchmark conditions rather than across genuinely heterogeneous camera hardware. We do not think any of these gaps undermine the core claim that vision-language pretraining has made cross-modal surveillance retrieval qualitatively more capable than fixed-vocabulary alternatives; they are, instead, the honest list of what remains to be demonstrated before that capability can be called mature.

7. Conclusion

The literature surveyed here supports a specific, reasonably well-evidenced claim: contrastively and generatively pretrained vision-language models, first developed for open-domain web imagery, have become the substrate on which nearly all current cross-modal retrieval research for surveillance is built, whether the task is matching a pedestrian description against a multi-camera gallery [7,8,9,22], detecting an open-vocabulary object category never seen with box supervision [25,5,6], retrieving the matching moment inside archived video [26,27,28], or scoring a segment against a natural-language description of anomalous behavior [29]. The foundational contrastive and fusion architectures — CLIP and ALIGN's dual encoders, BLIP and CoCa's bootstrapped and unified objectives, and the region-word alignment lineage running from ViLBERT and UNITER through SCAN and ALBEF — supply the representational backbone that the surveillance-specific

literature adapts, extends, and, where the domain gap is largest, substantially re-engineers [1,2,3,4,15,17,20,21].

What remains open is less a question of whether this architectural pattern works and more a question of how far it generalizes beyond the benchmark conditions under which it has been demonstrated: whether fine-grained, attribute-level retrieval accuracy holds up on genuinely low-quality surveillance imagery rather than curated re-identification benchmarks, whether gallery-scale retrieval latency and cost are compatible with real-time operational use, and whether the privacy and governance questions this technology raises are addressed with anything like the rigor its accuracy claims have received. Read together, the consistency with which independent research groups — working on person search, open-vocabulary detection, video retrieval, and anomaly recognition largely independently of one another — have converged on the same underlying vision-language pretraining substrate is, in our view, the strongest evidence that cross-modal retrieval is not a peripheral application of vision-language models but one of their central use cases, and intelligent surveillance is likely to remain one of the domains where the gap between benchmark accuracy and deployment readiness is scrutinized most closely.

8. References

1. 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. Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139: 8748-8763. DOI: [10.48550/arXiv.2103.00020](https://doi.org/10.48550/arXiv.2103.00020)
2. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q.V., Sung, Y.H., Li, Z., Duerig, T. (2021) Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), PMLR 139: 4904-4916. DOI: [10.48550/arXiv.2102.05918](https://doi.org/10.48550/arXiv.2102.05918)
3. Li, J., Li, D., Xiong, C., Hoi, S. (2022) BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Proceedings of the 39th International Conference on Machine Learning (ICML), PMLR 162: 12888-12900. DOI: [10.48550/arXiv.2201.12086](https://doi.org/10.48550/arXiv.2201.12086)
4. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y. (2022) CoCa: Contrastive Captioners are Image-Text Foundation Models. Transactions on Machine Learning Research (TMLR). DOI: [10.48550/arXiv.2205.01917](https://doi.org/10.48550/arXiv.2205.01917)
5. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J. (2022) Grounded Language-Image Pre-training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2022. DOI: [10.48550/arXiv.2112.03857](https://doi.org/10.48550/arXiv.2112.03857)
6. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., Zhu, J., Zhang, L. (2023) Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. arXiv preprint. DOI: [10.48550/arXiv.2303.05499](https://doi.org/10.48550/arXiv.2303.05499)
7. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., Wang, X. (2017) Person Search with Natural Language Description. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017. DOI: [10.48550/arXiv.1702.05729](https://doi.org/10.48550/arXiv.1702.05729)
8. Wang, Z., Fang, Z., Wang, J., Yang, Y. (2020) ViTAA: Visual-Textual Attributes Alignment in Person Search by Natural Language. Proceedings of the European Conference on Computer Vision (ECCV) 2020. DOI: [10.48550/arXiv.2005.07327](https://doi.org/10.48550/arXiv.2005.07327)
9. Jiang, D., Ye, M. (2023) Cross-Modal Implicit Relation Reasoning and Aligning for Text-to-Image Person Retrieval. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) 2023. DOI: [10.48550/arXiv.2303.12501](https://doi.org/10.48550/arXiv.2303.12501)
10. Li, J., Li, D., Savarese, S., Hoi, S. (2023) BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. Proceedings of the 40th International Conference on Machine Learning (ICML), PMLR 202: 19730-19742. DOI: [10.48550/arXiv.2301.12597](https://doi.org/10.48550/arXiv.2301.12597)
11. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K. (2022) Flamingo: a Visual Language Model for Few-Shot Learning. Advances in Neural Information Processing Systems (NeurIPS) 35. DOI: [10.48550/arXiv.2204.14198](https://doi.org/10.48550/arXiv.2204.14198)
12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I. (2017) Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS) 30. DOI: [10.48550/arXiv.1706.03762](https://doi.org/10.48550/arXiv.1706.03762)
13. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N. (2021) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. International Conference on Learning Representations (ICLR) 2021. DOI: [10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929)

[10.48550/arXiv.2010.11929](https://doi.org/10.48550/arXiv.2010.11929)[DOI: 10.48550/arXiv.1705.09406](https://doi.org/10.48550/arXiv.1705.09406)

14. Zhang, J., Huang, J., Jin, S., Lu, S. (2024) Vision-Language Models for Vision Tasks: A Survey. arXiv preprint. DOI: [10.48550/arXiv.2304.00685](https://doi.org/10.48550/arXiv.2304.00685)
15. Lu, J., Batra, D., Parikh, D., Lee, S. (2019) ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. Advances in Neural Information Processing Systems (NeurIPS) 32. DOI: [10.48550/arXiv.1908.02265](https://doi.org/10.48550/arXiv.1908.02265)
16. Tan, H., Bansal, M. (2019) LXMERT: Learning Cross-Modality Encoder Representations from Transformers. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP). DOI: [10.48550/arXiv.1908.07490](https://doi.org/10.48550/arXiv.1908.07490)
17. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J. (2020) UNITER: UNiversal Image-TEXT Representation Learning. Proceedings of the European Conference on Computer Vision (ECCV) 2020. DOI: [10.48550/arXiv.1909.11740](https://doi.org/10.48550/arXiv.1909.11740)
18. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., Gao, J. (2020) Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. Proceedings of the European Conference on Computer Vision (ECCV) 2020. DOI: [10.48550/arXiv.2004.06165](https://doi.org/10.48550/arXiv.2004.06165)
19. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L. (2018) Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018. DOI: [10.48550/arXiv.1707.07998](https://doi.org/10.48550/arXiv.1707.07998)
20. Lee, K.H., Chen, X., Hua, G., Hu, H., He, X. (2018) Stacked Cross Attention for Image-Text Matching. Proceedings of the European Conference on Computer Vision (ECCV) 2018. DOI: [10.48550/arXiv.1803.08024](https://doi.org/10.48550/arXiv.1803.08024)
21. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H. (2021) Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Advances in Neural Information Processing Systems (NeurIPS) 34. DOI: [10.48550/arXiv.2107.07651](https://doi.org/10.48550/arXiv.2107.07651)
22. Yan, S., Dong, N., Zhang, L., Tang, J. (2023) CLIP-Driven Fine-grained Text-Image Person Re-identification. arXiv preprint. DOI: [10.48550/arXiv.2210.10276](https://doi.org/10.48550/arXiv.2210.10276)
23. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., Hoi, S.C.H. (2020) Deep Learning for Person Re-identification: A Survey and Outlook. arXiv preprint. DOI: [10.48550/arXiv.2001.04193](https://doi.org/10.48550/arXiv.2001.04193)
24. Baltrušaitis, T., Ahuja, C., Morency, L.P. (2019) Multimodal Machine Learning: A Survey and Taxonomy. arXiv preprint.
25. Gu, X., Lin, T.Y., Kuo, W., Cui, Y. (2022) Open-vocabulary Object Detection via Vision and Language Knowledge Distillation. International Conference on Learning Representations (ICLR) 2022. DOI: [10.48550/arXiv.2104.13921](https://doi.org/10.48550/arXiv.2104.13921)
26. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T. (2022) CLIP4Clip: An Empirical Study of CLIP for End to End Video Clip Retrieval. arXiv preprint. DOI: [10.48550/arXiv.2104.08860](https://doi.org/10.48550/arXiv.2104.08860)
27. Bain, M., Nagrani, A., Varol, G., Zisserman, A. (2021) Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) 2021. DOI: [10.48550/arXiv.2104.00650](https://doi.org/10.48550/arXiv.2104.00650)

28. Wang, M., Xing, J., Liu, Y. (2021) ActionCLIP: A New Paradigm for Video Action Recognition. arXiv preprint. DOI: [10.48550/arXiv.2109.08472](https://doi.org/10.48550/arXiv.2109.08472)
29. Sultani, W., Chen, C., Shah, M. (2018) Real-world Anomaly Detection in Surveillance Videos. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018. DOI: [10.48550/arXiv.1801.04264](https://doi.org/10.48550/arXiv.1801.04264)

9. Declarations

Funding

This work received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported here.

Data Availability Statement

This is a literature synthesis; no new primary data were generated or analyzed. All numeric results reported are drawn from the cited sources, which readers should consult directly for underlying data and code availability.

Ethical Approval

Not applicable. This article did not involve human participants, animal subjects, or the collection of any primary surveillance or personal data.