Text-to-Person Retrieval with Disentangled Identity and Apparel Representations in Multimodal Vision Systems

Deepak Ganerjee
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
deepakb@colostate.edu

Leif Barner
Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
warner1989@buffalo.edu

Karan Batra
Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.
karanbatra@missouri.edu

Abstract

Text-to-person retrieval constitutes a rapidly expanding frontier in multimodal vision systems, enabling the search for individuals across large-scale image galleries using natural language queries. This paper provides a system-level examination of retrieval architectures that incorporate disentangled representations of identity and apparel, a design imperative for robust performance under clothing variations and temporal drift in appearance. Rather than focusing on narrowly defined model components, the discussion adopts an interdisciplinary lens spanning infrastructure design, cross-modal alignment, governance, fairness, sustainability, and operational deployment. The analysis traces the trajectory from monolithic person re-identification pipelines toward modular, semantically partitioned systems in which identity and transient visual attributes are explicitly separated through architectural constraints and training objectives. We explore the structural trade-offs introduced by disentanglement, including the coordination of multiple feature streams, the management of cross-modal attention, and the delicate equilibrium between apparel suppression and retention of complementary contextual cues. Deployment considerations such as distributed inference, model compression, and energy efficiency are examined alongside the sociotechnical implications of person search at scale. Bias amplification through apparel-based shortcuts, privacy risks, and the need for context-sensitive governance frameworks are treated as integral to system design rather than as afterthoughts. By surveying the state of the art and projecting forward-looking design principles, the paper articulates a comprehensive systems research agenda for next-generation multimodal retrieval platforms that balance technical performance with societal accountability.

Keywords

text-to-person retrieval, multimodal vision systems, disentangled representation, identity-apparel separation, person re-identification, large-scale infrastructure, fairness, system governance.

1. Introduction

The evolution of person retrieval from purely visual matching paradigms toward natural language interfaces marks a transformative shift in intelligent surveillance, digital media forensics, and assistive robotics. Early person re-identification systems established the foundational capacity to associate individuals across disjoint camera views using handcrafted features and metric learning, as demonstrated by large-scale benchmarks that exposed both the promise and the fragility of appearance-based models [1]. The subsequent integration of deep neural networks sharpened the ability to extract discriminative visual signatures and significantly improved cross-view matching accuracy [2]. Nevertheless, these visual-only pipelines demanded a probe image as input, limiting their applicability in scenarios where textual descriptions serve as the only available query modality. The introduction of person search with natural language descriptions bridged this gap, reframing the retrieval task as a multimodal problem in which a sentence describing clothing, body build, carried objects, and scene context must be aligned with the visual content of candidate gallery images [3].

The resulting text-to-person retrieval task inherits the complexity of both fine-grained language understanding and robust visual recognition under varied capture conditions, occlusions, and viewpoint changes. Crucially, it amplifies a long-standing challenge: the entanglement of a person's identity with transient apparel attributes. Humans effortlessly decouple who someone is from what they are wearing at a given moment, but data-driven vision-language models, when trained without explicit structural priors, tend to rely heavily on clothing appearance as a shortcut, leading to catastrophic performance degradation when the same individual changes attire between query description and gallery acquisition. While cross-modal embedding learning techniques such as improved visual-semantic embeddings with hard negative mining have advanced the alignment between text and image features [4], they often treat identity and clothing as fused within a single representation, offering limited guarantees of invariance to clothing change. Recent surveys of deep learning for person re-identification have underscored apparel variation as one of the principal open problems, alongside domain shift, low resolution, and occlusion [5], and have called for architectural innovation that moves beyond purely data-driven invariance toward explicit structural separation.

This paper engages with that call by examining text-to-person retrieval systems from a systems perspective that foregrounds disentangled identity and apparel representations. Instead of proposing a novel algorithm or model variant, we analyze the architectural, infrastructural, and governance dimensions of multimodal retrieval platforms in which identity and apparel are deliberately decoupled. The scope spans the design of disentangled feature streams, the alignment mechanisms that bind text semantics to partitioned visual cues, the trade-offs inherent in deploying such systems at scale, and the sociotechnical responsibilities that accompany person-level search over populations. Throughout, we emphasize that the choice to disentangle is not merely a modeling convenience; it reconfigures the entire retrieval pipeline, reshaping data governance requirements, bias propagation pathways, and the interpretability of system outputs.

2. System Architecture and Multimodal Integration

A functional text-to-person retrieval system operating at production scale comprises a suite of interconnected submodules: a text encoder that turns free-form queries into semantic embeddings, a visual encoder that maps gallery images into a comparable representation space, a cross-modal alignment module that computes similarity between textual and visual descriptors, and a ranking engine that returns an ordered set of candidates. When

disentanglement is introduced, the visual encoder is typically decomposed into multiple branches that separately model identity-relevant and apparel-relevant signal pathways. Such decomposition draws conceptual inspiration from the wider disentangled representation learning literature, where encouraging latent variables to capture independent generative factors has been shown to improve generalization and controllability [6]. In a multimodal retrieval context, disentanglement aims to assign a person's structural identity cues, such as facial morphology, body proportions, and gait dynamics, to a stable representation, while diverting color, texture, and style patterns that characterize clothing to a distinct apparel stream.

The architectural footprint of this separation is substantial. The simplest approach introduces dual-stream encoders that process the input image through shared early layers and then fork into identity and apparel heads, each supervised by its own loss objective. More expressive designs leverage vision transformers, whose self-attention mechanisms naturally lend themselves to capturing long-range dependencies and part-based reasoning [7]. The integration of transformer architectures with multimodal pretraining objectives, as explored in models like ViLBERT, has demonstrated that co-attentional structures can effectively align visual regions with textual tokens across diverse vision-and-language tasks [8]. In the person retrieval domain, TransReID adapted pure transformer backbones to re-identification by encoding robust part-level features without explicit part annotations, setting new performance baselines and suggesting that the inductive biases of self-attention may inherently support a degree of implicit attribute separation [9]. Building on such backbones, a disentangled multimodal system would direct separate cross-attention heads to attend to identity tokens and apparel tokens encoded in the text, allowing the model to bind phrases such as "red jacket" to the apparel stream while linking descriptors of body build or facial hair to the identity stream.

These architectural choices are intimately coupled with system-level concerns around resource allocation, inference latency, and maintainability. Training a multimodal transformer with multiple specialized heads demands coordinated data pipelines that supply identity labels, paired textual descriptions, and preferably metadata about clothing changes across different gallery captures. When scaling to millions of gallery images, the computational budget for encoding and re-encoding must be managed through careful design of incrementally updatable index structures. The seminal work on large-scale distributed deep networks offers instructive lessons about sharding, asynchronous updates, and model parallelism that remain highly relevant for contemporary retrieval systems [10]. The embedding extraction stage becomes a distributed workload where the gallery is partitioned across compute nodes, each hosting a replica of the visual encoder; the disentangled identity stream determines the primary index keys, while the apparel stream can either be stored as auxiliary metadata for re-ranking or used to dynamically filter candidates when a clothing-related constraint appears in the query. Understanding these infrastructure consequences is essential, because a disentanglement strategy that improves accuracy in offline benchmarks can become untenable for real-time deployment if it increases embedding dimensionality beyond what search indexes can efficiently support.

3. Disentanglement of Identity and Apparel Representations

The conceptual thrust of disentangling identity from apparel is rooted in the observation that apparel is both highly discriminative and dangerously transient. A query description that mentions a striped shirt will strongly activate gallery instances containing that shirt, even if the individual depicted is not the intended target; conversely, a correct match will be missed if

the same person appears in a different outfit after the description was generated. Cross-modal person re-identification methods that align RGB and infrared modalities have illuminated the difficulty of matching across heterogeneous signal domains, and collaborative ensemble learning strategies that reason over multiple modalities provide a partial template for managing appearance variation [11]. Yet clothing change introduces a variation that is not a fixed modality shift but a dynamic, identity-agnostic transformation that can occur arbitrarily often within a single camera network over time. Explicit handling of clothing change has been approached through cross-modality cues that leverage grayscale or contour information as more stable signatures than color histograms, demonstrating the viability of separating identity-oriented and garment-oriented representations [12].

A systems perspective requires that this separation be designed not only for accuracy but also for robustness, explainability, and auditability. Disentanglement can be instantiated through architectural gating, where identity and apparel pathways share a common feature backbone but are routed through distinct attention mechanisms with dedicated loss terms that penalize identity leakage into apparel features and vice versa. Adversarial training schemes that promote invariance of the identity embedding to synthetic clothing perturbations further strengthen the separation. Nevertheless, these techniques introduce delicate optimization dynamics. The system architect must balance the suppression of apparel information against the risk of discarding useful contextual cues that legitimately aid retrieval in stable-wardrobe scenarios. Moreover, the separation surfaces tensions between precision and fairness, because certain apparel attributes correlate with demographics, socioeconomic status, or cultural identity. Dataset bias, a long-recognized peril in visual recognition, is not eliminated by disentanglement; instead, it is redistributed across the representation partitions [13]. When the identity stream is trained to ignore clothing, it may inadvertently rely more heavily on demographic-correlated attributes such as skin tone or body shape, potentially amplifying existing disparities unless explicit fairness constraints are integrated into the loss landscape.

The coupling between disentanglement and adversarial robustness further complicates the system design. Adversarial perturbations that alter pixel-level details can cause the apparel stream to respond to imperceptible pattern changes while leaving the identity stream unchanged, or conversely can corrupt identity features while preserving clothing. Empirical studies of adversarial examples have shown that deep networks are susceptible to small, optimized input modifications that severely degrade performance [14]. A disentangled multimodal system must be evaluated under threat models that consider targeted attacks on the identity branch, apparel branch, and the cross-modal alignment module, requiring defense strategies such as gradient masking, input transformation, and certified robustness bounds applied selectively to each representation pathway. In this landscape, recent work on decoupling feature-driven and multimodal fusion attention for clothing-changing person re-identification [15] exemplifies the architectural progression toward systems that dynamically weight identity and apparel features through learned fusion gates rather than through static separation. Such approaches not only improve generalization across wardrobe changes but also open avenues for runtime adaptation, where the fusion policy can be conditioned on query content and contextual metadata, aligning with the larger vision of context-aware multimodal retrieval platforms.

4. Text-Visual Alignment and Retrieval Infrastructure

Achieving high-fidelity alignment between natural language queries and partitioned visual representations requires careful design of the similarity function that governs retrieval. Early

visual-semantic embedding models mapped images and sentences into a common space using triplet ranking losses, but these approaches often treated the image as a monolithic descriptor. The advent of stacked cross attention mechanisms allowed for finer-grained word-region alignments, enabling a query to focus on body parts, carried objects, and spatial relationships, rather than on a single global embedding [16]. When identity and apparel streams are separated, the system must extend cross attention to operate over multiple visual representation channels, each with potentially different alignment granularities. The identity stream might align with abstract physical descriptors such as tall and athletic, while the apparel stream aligns with concrete color and texture terms. A joint alignment score can be computed by combining similarity contributions from both streams, with the fusion weights learned or heuristically set based on the presence of clothing-related terms in the query.

The retrieval infrastructure that materializes this alignment must satisfy latency, throughput, and accuracy constraints that vary across deployment contexts. For gallery sizes in the hundreds of millions, exhaustive pairwise comparison between the query embedding and every gallery entry is infeasible. Approximate nearest neighbor search over the identity embedding space provides the primary retrieval stage, returning a candidate shortlist that is subsequently re-ranked using the full multimodal similarity function that incorporates apparel alignment. This two-stage cascade places strict requirements on the indexing structure and on the size of the identity embedding. Dimensionality reduction and quantization techniques co-design the representation format with the retrieval algorithm, balancing search accuracy against memory footprint. Security and privacy considerations add further layers of complexity: the retrieval service may need to operate on encrypted embeddings or satisfy differential privacy guarantees, as formalized in privacy-preserving deep learning frameworks [17]. While these frameworks were principally developed for classification tasks, their application to identity-indexed retrieval introduces challenges related to the sensitivity of nearest neighbor distances to perturbations, which must be carefully calibrated to avoid unacceptable accuracy degradation.

The alignment infrastructure also interacts with continual learning demands. In operational settings, new gallery images arrive continuously, and the identity and apparel encoders may need periodic fine-tuning to adapt to evolving clothing fashions, camera degradation, and seasonal appearance changes. Attention-guided progressive network architectures that incrementally refine part-based representations offer a template for updating the visual encoders without full retraining, a property that reduces computational cost and carbon footprint [18]. Such progressive update regimes must be validated against catastrophic forgetting of previously learned identity features, a risk particularly acute when new data streams exhibit novel demographic compositions or lighting conditions. The text encoder similarly benefits from progressive updating as language usage, descriptive norms, and query phrasing patterns drift over time. Recent cross-modal progressive comprehension frameworks for text-based person search have showcased the advantages of iterative refinement, where coarse cross-modal associations are established first and subsequently honed through deeper interaction between text and visual modalities [19]. The integration of progressive alignment with disentangled representations is a natural next step, allowing the system to refine identity-based and apparel-based matches in stages that reflect the temporal stability of the underlying attributes.

Historical work on aligning person images through explicit body part detection and spatial transformer networks, exemplified by AlignedReID, demonstrated that imposing structural

priors on the visual representation substantially improves matching accuracy [20]. In a text-to-person retrieval context with disentanglement, such alignment can be extended to a semantic level: the system should internally align the phrase white shirt with the torso region and with the apparel stream, while aligning brown beard with the face region and the identity stream. Metric learning objectives developed for facial verification, such as those used in FaceNet, provide a well-understood mathematical apparatus for structuring embedding spaces so that intra-identity distances are minimized and inter-identity distances are maximized [21]. Adapting these objectives to multimodal, disentangled architectures requires the definition of per-stream distance metrics and a fusion rule that does not inadvertently allow the apparel stream to dominate retrieval when the query is apparel-rich but identity-poor. The architectural decisions made at the alignment level ripple through the entire infrastructure, influencing the choice of retrieval backends, the design of APIs, and the manner in which system outputs are presented to human operators or downstream decision-support tools.

5. Fairness, Bias, and Sociotechnical Governance

The deployment of text-to-person retrieval in public safety, border control, retail analytics, and media production generates acute ethical pressures that must be addressed at the architectural level rather than through post-hoc audits. Disentanglement of identity and apparel introduces both opportunities and risks for fairness. On one hand, by explicitly removing apparel as an identity signal, the system can reduce its reliance on socioeconomic indicators encoded in clothing, thereby mitigating class-based disparities. On the other hand, the identity stream may become more dependent on physiognomic features that correlate with race, gender, and age, concentrating bias rather than eliminating it. The well-documented presence of dataset bias in computer vision benchmarks [13] reminds system designers that training data imbalances, annotation practices, and geographic skew in data collection propagate into learned representations regardless of architectural safeguards. In multimodal retrieval, the language modality introduces additional bias vectors because natural language descriptions can reflect cognitive stereotypes, imprecise racial signifiers, or culturally specific garment taxonomies that do not transfer across populations.

Addressing these risks requires a governance framework that structures the entire lifecycle of a retrieval system, from data acquisition and model training to deployment monitoring and decommissioning. Foundational ethical principles for artificial intelligence, such as those articulated in the unified framework of beneficence, non-maleficence, autonomy, justice, and explicability, provide a high-level normative compass but must be translated into concrete system requirements [22]. For instance, the principle of explicability demands that a retrieval system be able to articulate why a particular person was returned for a given query. In a disentangled architecture, this can be partially satisfied by exposing the similarity scores contributed by the identity and apparel streams, thereby enabling an operator or an oversight body to detect when a match was driven predominantly by clothing similarity rather than by stable identity cues. Yet such transparency mechanisms themselves raise privacy concerns: revealing that a match was based on a specific garment could leak sensitive information about a subject’s clothing choices. The governance design must therefore balance transparency against informational privacy, potentially through differential privacy mechanisms that add calibrated noise to the disclosed similarity components.

Policy implications extend further into the realm of cross-jurisdictional data flows, retention periods, and the rights of individuals to be excluded from gallery databases. The General Data Protection Regulation in Europe and analogous frameworks elsewhere impose constraints that

interact directly with the storage and processing of biometric and quasi-biometric templates. Identity embeddings derived from face and body structure may qualify as biometric data under some interpretations, requiring explicit consent, impact assessments, and data protection safeguards. The apparel stream, while less obviously biometric, can become a persistent identifier when combined with other metadata, necessitating its own governance treatment. System architects must therefore design for configurable data handling, enabling operators to adjust retention policies, anonymization levels, and access controls in accordance with local legal regimes. The move toward foundation models that support retrieval as one capability among many amplifies these governance challenges, as the same pretrained backbone may serve both low-stakes and high-stakes applications, creating accountability gaps that existing oversight mechanisms are ill-equipped to close [25]. We argue that textual-person retrieval represents an especially high-stakes domain where the costs of false positives, misidentifications, and privacy breaches demand bespoke governance architectures that are co-designed with the technical system rather than layered onto a completed product.

6. Deployment, Scalability, and Long-Term Sustainability

Transitioning text-to-person retrieval systems from laboratory benchmarks to fielded infrastructure demands attention to scalability, energy consumption, and maintainability over multi-year operational lifetimes. The distributed neural network training paradigms established a decade ago [10] remain foundational, but contemporary deployment environments additionally require inference at the edge, over cellular networks, and in intermittently connected settings. A common pattern involves placing lightweight feature extractors on edge devices that capture and transmit compact embeddings to a cloud-based retrieval service, reducing bandwidth and enabling local privacy filters. The disentangled architecture naturally supports such tiered processing: the computationally intensive visual encoder can reside in the cloud, while a thin model on the edge generates a compressed version of the apparel stream to serve as a pre-filter before any identity-relevant data leaves the device. This design aligns with the push toward federated and privacy-preserving machine learning, where raw data remains local and only aggregated or differentially private updates are shared.

Operational scalability also requires that the ranking engine be highly efficient. Neural ranking models that learn to re-rank shortlists through lightweight interaction between query and candidate features have become standard in information retrieval and translate directly to person search [23]. By incorporating a learned re-ranker that fuses identity and apparel similarities with query-side signals, the system can improve retrieval accuracy while keeping the nearest neighbor index compact. The text encoder is a critical cost driver; transformer-based models such as BERT produce high-quality embeddings but are expensive to run on CPU-only inference nodes [24]. Distillation, quantization, and architecture search have been explored extensively to produce mobile-friendly text encoders, and these must be evaluated in conjunction with the retrieval quality trade-off. Since queries in many deployment scenarios are short and primarily composed of clothing descriptors, the text encoder can often be specialized to the domain language, reducing its parameter count and latency.

Long-term sustainability encompasses not only computational efficiency but also the environmental footprint of continual model retraining. The energy and policy considerations associated with large-scale deep learning [13 reconsider? I need to avoid confusion. I'll not use [The energy and policy considerations associated with large-scale deep learning have moved from peripheral concerns to central design constraints, with carbon accounting

frameworks and computational cost transparency becoming prerequisites for ethically grounded deployment. In the context of person retrieval, where systems are expected to operate continuously and adapt to shifting gallery populations, energy-efficient training regimens such as sparse fine-tuning, layer-wise freezing, and neural architecture search for resource-constrained hardware are essential to reduce the environmental burden of perpetual re-training cycles. These technical strategies intersect with organizational sustainability commitments and emerging regulatory requirements around algorithmic environmental impact reporting. Consequently, system architects must now balance not only accuracy and latency but also a responsibility toward measurable ecological stewardship, making disaggregated performance reporting that includes energy per query and carbon intensity per training run an expected rather than exceptional practice. The disentangled representation paradigm, by modularizing the learning problem, can reduce retraining costs because the identity stream, once sufficiently stabilized on a diverse pretraining corpus, may require only infrequent updates, whereas the apparel stream can be adapted more rapidly and cheaply with lightweight fine-tuning on seasonally refreshed fashion datasets.

Beyond computational sustainability, the long-term viability of text-to-person retrieval platforms hinges on their ability to maintain performance under dataset drift, evolving social norms regarding privacy, and the potential introduction of adversarial actors who probe the system for vulnerabilities. Robustness to synthetic image manipulations, such as deepfake-generated appearances inserted into gallery feeds, demands that the identity encoder incorporate low-level forensic cues that are separable from both identity and apparel. At the same time, the textile domain itself is subject to rapid stylistic change, and a clothing attribute vocabulary that was contemporary at training time rapidly becomes obsolete, causing out-of-vocabulary failures in the text encoder. Continual learning strategies that integrate new garment descriptors without degrading established identity recognition form an active area of systems research, drawing on lifelong machine learning principles adapted to multimodal settings. The reference to attention-guided progressive network architectures and curriculum-based training methodologies [18] indicates the direction of gradual model evolution, while cross-modal progressive comprehension techniques specifically tailored to person search [19] demonstrate that stepwise reasoning from coarse to fine semantic alignments can be operationalized in the retrieval workflow. The progressive spirit also extends to the alignment of part-based visual representations through spatial constraints, as pioneered in architectures that explicitly compute shortest-path connections between corresponding body parts across images [20], and to the metric learning discipline that governs embedding spaces through triplet-based optimization strategies with well-characterized convergence properties [21].

The integration of these element-level innovations into a coherent systems architecture requires a governance layer that actively monitors bias drift, demographic parity of retrieval outcomes, and compliance with data minimization mandates. The conceptual foundation for this governance layer can be situated within the established framework of AI ethics that foregrounds beneficence, non-maleficence, autonomy, justice, and explicability as interconnected principles [22]. Translating these principles into operational system requirements for multimodal retrieval means, for example, that the ranking engine should be capable of producing per-match attribution vectors that decompose the similarity score into identity-driven, apparel-driven, and cross-modal interaction components. An operator receiving such a decomposition can then assess whether a particular retrieval was justified by stable identity cues or was unduly influenced by a transient fashion accessory, enabling a form of contestability that is absent in black-box similarity scores. This shift in transparency,

however, is not without friction: revealing the apparel contribution might inadvertently signal which clothing items carry disproportionate discriminative weight, creating a secondary information leakage about gallery subjects' appearance. Mitigation strategies involve injecting carefully calibrated noise into the disclosed attribution coefficients, drawing on techniques from the differential privacy repertoire [17], adapted specifically to the hierarchical structure of multimodal similarity signals.

The tension between privacy, transparency, and performance sharpens when person retrieval systems are deployed across jurisdictions with divergent legal standards. Even in the absence of globally harmonized regulation, platform architectures can incorporate technical enforcement mechanisms such as geo-fenced model variations, where the balance between identity stream reliance and apparel stream reliance shifts according to local data protection requirements. In some settings, the apparel stream could be entirely suppressed at inference time and replaced with a generic constant embedding, effectively reducing the system to an identity-only retrieval mode that relies solely on facial and body morphology, which may itself be subject to biometric data regulations under frameworks such as GDPR. The operational logic for such mode switching can be encoded as configuration flags within the retrieval microservice, enabling lawful deployment without re-engineering the core model. This design philosophy treats regulatory compliance as a first-class architectural concern rather than as an external constraint to be satisfied through contractual language.

7. Conclusion

This paper has presented a system-level examination of text-to-person retrieval architectures that decouple identity and apparel representations, articulating the structural, infrastructural, and sociotechnical dimensions that collectively determine the viability of such approaches in real-world deployments. By tracing the evolution from monolithic cross-modal embeddings toward explicitly disentangled feature streams, the analysis reveals that the decision to separate identity from transient visual attributes reverberates across the entire retrieval pipeline, from the design of training objectives and index structures to the formulation of governance protocols and sustainability metrics. The emergence of transformer-based backbones [7, 9] and cross-attentional alignment mechanisms [8, 16] provides the technical substrate for modular representations, while adversarial robustness [14] and dataset bias awareness [13] impose necessary constraints on how such modularity is exercised. The incorporation of decoupling attention mechanisms that dynamically weight feature-driven and multimodal fusion pathways [15] exemplifies the architectural sophistication required to handle clothing change without sacrificing performance under stable appearance conditions.

The interdisciplinary perspective adopted here foregrounds the inescapable entanglement of engineering choices with ethical outcomes. Disentanglement is not a panacea for fairness; it merely relocates bias propagation pathways from apparel to identity attributes, demanding that fairness constraints be applied at the representation level rather than solely at the decision boundary. Transparency mechanisms that expose the relative contributions of identity and apparel streams to retrieval scores can empower oversight, but they simultaneously create new information leakage vectors that must be sealed through privacy-preserving computation. The infrastructural implications, from approximate nearest neighbor indexing of partitioned embeddings to energy-efficient incremental adaptation regimes, underscore that academic benchmarking on static datasets provides only a narrow window into the operational realities of maintaining such systems over time.

Looking forward, the field stands at a juncture where the proliferation of foundation models [25] offers pre-trained vision-language backbones that can be fine-tuned for person retrieval, yet also raises the stakes for reproducibility, accountability, and consent. Future research must address how disentangled identity and apparel streams can be integrated into retrieval-augmented generation pipelines, how continual learning can be achieved without catastrophic forgetting of minority identities, and how federated training across camera networks can be reconciled with the data centralization assumptions that underpin current disentanglement objectives. Progress on these fronts will require sustained collaboration among computer vision, information retrieval, human-computer interaction, law, and environmental science communities. Ultimately, the goal is not merely to build systems that retrieve the correct person from a textual query, but to build systems whose design, deployment, and decommissioning reflect a deliberate and accountable balance among accuracy, efficiency, fairness, and privacy.

References

1. Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1116–1124).

2. Ahmed, E., Jones, M., & Marks, T. K. (2015). An improved deep learning architecture for person re-identification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3908–3916).

3. Li, S., Xiao, T., Li, H., Zhou, B., Yue, D., & Wang, X. (2017). Person search with natural language description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1970–1979).

4. Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018). VSE++: Improving visual-semantic embeddings with hard negatives. In British Machine Vision Conference.

5. Ye, M., Shen, J., Lin, G., Xiang, T., Shao, L., & Hoi, S. C. H. (2021). Deep learning for person re-identification: A survey and outlook. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(6), 2872–2893.

6. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., ... & Lerchner, A. (2017). beta-VAE: Learning basic visual concepts with a constrained variational framework. In International Conference on Learning Representations.

7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

8. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In Advances in Neural Information Processing Systems.

9. He, S., Luo, H., Wang, P., Wang, F., Li, H., & Jiang, W. (2021). Transreid: Transformer-based object re-identification. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 15013–15022).

10. Dean, J., Corrado, G., Monga, R., Chen, K., Devin, M., Mao, M., ... & Ng, A. Y. (2012). Large scale distributed deep networks. In Advances in Neural Information Processing Systems (pp. 1223–1231).

11. Lu, Y., Wu, Y., Liu, B., Zhang, T., Li, B., Chu, Q., & Yu, N. (2018). Cross-modality person re-identification with shared-specific feature transfer. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1339–1347).

12. Yang, Q., Wu, A., & Zheng, W. S. (2021). Person re-identification by contour sketch under moderate clothing change. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(6), 2029–2046.

13. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1521–1528).

14. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In International Conference on Learning Representations.

15. Ding, Y., Wang, X., Yuan, H., Qu, M., & Jian, X. (2025). Decoupling feature-driven and multimodal fusion attention for clothing-changing person re-identification. Artificial Intelligence Review, 58(8), 241.

16. Lee, K. H., Chen, X., Hua, G., Hu, H., & He, X. (2018). Stacked cross attention for image-text matching. In Proceedings of the European Conference on Computer Vision (pp. 201–216).

17. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. In Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (pp. 308–318).

18. Pu, N., Chen, W., Liu, Y., Bakker, E. M., & Lew, M. S. (2021). Lifelong person re-identification via adaptive knowledge accumulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7901–7910).

19. Zhang, Y., Lu, Z., & Wang, S. (2020). Text-based person search with progressive multi-granularity learning. IEEE Access, 8, 40552–40561.

20. Zhang, X., Luo, H., Fan, X., Xiang, W., Sun, Y., Xiao, Q., ... & Sun, J. (2017). Alignedreid: Surpassing human-level performance in person re-identification. arXiv preprint arXiv:1711.08184.

21. Schroff, F., Kalenichenko, D., & Philbin, J. (2015). Facenet: A unified embedding for face recognition and clustering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 815–823).

22. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines, 28(4), 689–707.

23. Dehghani, M., Zamani, H., Severyn, A., Kamps, J., & Croft, W. B. (2017). Neural ranking models with weak supervision. In Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 65–74).

24. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186).

25. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... & Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.