Vision-Language Guided Identity Retrieval with Robust Appearance Disentanglement for Long-Term Person Tracking

Yuanhong Du
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
yuanhongmail@ku.edu

Walid Stanley
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
walid.stanley@unh.edu

Abstract

The sustained tracking of individuals across distributed camera networks over extended temporal windows remains an enduring systems challenge, as visual appearance can vary markedly due to clothing changes, illumination shifts, and occlusions. This paper offers a comprehensive system-level investigation into the convergence of vision-language guidance and robust appearance disentanglement for long-term person re-identification. We posit that integrating pre-trained vision-language models into the retrieval pipeline injects semantic reasoning about personal attributes that are resilient to superficial appearance transformations. A hybrid architecture is proposed in which multimodal fusion modules align visual embeddings with natural language descriptions of identity-relevant traits, while adversarial feature separation and semantic consistency regularization enforce identity-salient representations decoupled from transient clothing styles. The discussion moves beyond algorithmic design to encompass the governance, ethical deployment, and sustainability of such tracking infrastructures. We analyze structural trade-offs between model expressiveness and inference latency, energy-aware compression strategies, and privacy-preserving data flows. Drawing on smart city and public safety case illustrations, the paper examines how fairness audits and policy-aligned architectures can mitigate risks of demographic bias and erosion of civil liberties. The resulting framework provides a roadmap for designing person tracking systems that balance accuracy, robustness, accountability, and environmental responsibility in long-term operational contexts.

Keywords

person re-identification, vision-language models, appearance disentanglement, long-term tracking, system architecture, fairness, sustainable AI.

1. Introduction

Advances in vision-language pre-training, exemplified by models such as CLIP and BLIP [1][2], have demonstrated that joint embedding spaces learned from massive image-text corpora capture semantic correspondences that extend far beyond low-level pixel similarity. This capability holds profound implications for person re-identification (ReID) in long-term tracking scenarios, where persistent identity retrieval must contend with dramatic appearance variations caused by clothing changes, carried objects, and environmental factors. Traditional

ReID systems, largely developed and benchmarked on datasets like Market-1501 and CUHK03 [3][4], rely on deep feature extractors trained solely on visual appearance. While effective over short temporal windows, these models falter when a previously observed individual dons a different outfit or when lighting conditions diverge between cameras. In such cases, identity features that primarily encode color and texture become unreliable, necessitating a paradigm shift toward representations that capture more invariant attributes.

The core proposition of this paper is that vision-language guided identity retrieval, synergistically combined with robust appearance disentanglement, can anchor identity representations in semantic attributes—such as body shape, gait, facial structure, and other personal characteristics—that persist irrespective of clothing. We articulate a system architecture in which a language-aligned vision encoder and a text encoder interact through cross-modal attention layers to produce a disentangled identity embedding. This embedding is learned under an adversarial objective that discourages the encoding of clothing style, together with a consistency loss that aligns visual and linguistic descriptors of the same identity. The resulting retrieval pipeline can function across long time spans, continuously matching probe images against a gallery that may include drastically altered appearances, while also accepting natural language queries as an auxiliary input modality.

Establishing such a system in real-world deployments, however, extends far beyond algorithmic performance. The paper therefore devotes substantial attention to structural trade-offs, governance, and sustainability. We examine how the choice between cloud-centric and edge-centric inference architectures influences latency, energy consumption, and compliance with data protection regulations. We interrogate the fairness implications of training data imbalances and the potential for disproportionate misidentification across demographic groups. In addition, we address the infrastructural challenge of maintaining retrieval accuracy under concept drift as populations, clothing trends, and camera networks evolve. By weaving together these technical, operational, and ethical threads, the paper aims to offer a holistic perspective on the design, deployment, and oversight of vision-language driven long-term person tracking systems.

2. Foundational Concepts and Related System Paradigms

Person re-identification has matured from handcrafted representations to deep metric learning architectures that minimize intra-class distances while maximizing inter-class separability. The triplet loss formulation, which enforces a margin between positive and negative pairs, has become a standard supervisory signal for learning discriminative embeddings [5]. When applied across multiple camera domains, however, purely visual metrics often overfit to scene-specific appearance distributions. Self-training paradigms that propagate pseudo-labels across unlabelled target domains have been proposed to bridge this distribution gap [6], yet they remain vulnerable to confirmation bias and still depend on short-term appearance similarity.

Addressing clothing variation explicitly has motivated a stream of work on disentangled representation learning. By factorizing an image into separate codes for identity, pose, and clothing, generative models can synthesize novel appearances for the same individual [7]. These methods often employ adversarial decoupling to force the identity code to be invariant to changes in the other factors. In the specific context of cloth-changing person re-identification, architectures that separate body shape and gait from apparel cues have shown improved performance on dedicated benchmarks [8]. Nevertheless, purely visual disentanglement struggles to capture high-level semantic attributes that could stabilize

identity over longer durations, such as the notion of a person's typical build or the way they carry themselves. It is precisely this gap that vision-language fusion begins to close.

Recent work on decoupling feature-driven and multimodal fusion attention for clothing-changing person re-identification [9] demonstrates that incorporating textual descriptions of a person's appearance into the re-identification process significantly enhances discrimination under apparel variations. By learning to attend jointly to visual regions and corresponding linguistic phrases, such systems can contextualize transient visual details as belonging to a changeable layer of representation while preserving the identity core. This insight aligns with broader trends in video-based person re-identification, where temporal sequences and auxiliary semantic attributes are used to refine identity hypotheses [10]. Generative approaches further augment the visual data space by synthesizing cross-view and cross-pose images through generative adversarial networks, enabling models to learn view-invariant features [11]. Underpinning these multimodal architectures are the Transformer attention mechanism [12] and bi-directional language encoders such as BERT [13], which facilitate rich interactions between modalities and substitute recurrent processing with globally conditioned self-attention.

Current practice in large-scale tracking systems, however, rarely integrates these components into a unified operational pipeline that is robust over weeks or months. The system-level analysis presented in subsequent sections addresses this gap by examining how attention-based vision-language fusion, adversarial appearance disentanglement, and semantic consistency constraints can be jointly deployed, while also considering the structural and societal dimensions essential for responsible real-world application.

3. System Architecture and Multimodal Integration

The envisioned tracking infrastructure rests on a modular architecture composed of four principal processing stages: visual encoding, language encoding, cross-modal fusion and disentanglement, and identity retrieval with optional gallery maintenance. A pre-trained vision-language backbone, such as a shared image-text Transformer initialized from CLIP [1], provides aligned visual and textual embeddings that serve as the starting point for identity extraction. The visual stream processes cropped person images from multiple camera feeds, while the textual stream can ingest both structured metadata and free-form natural language descriptions generated either by human operators or by an automated scene captioning module. This dual-input design enables the system to function in hybrid query modes, where a probe may consist solely of an image, solely of a textual description, or of a paired set of both modalities.

The cross-modal fusion module employs transformer layers that compute joint attention over visual region features and tokenized language phrases, allowing the model to dynamically weight the contribution of different body parts based on linguistic context. For example, if a description specifies "tall, with a prominent nose and a slight limp," the fusion mechanism can learn to focus attention on biometric regions and gait-related spatial cues while suppressing the color histogram of the shirt. The output of the fusion module is further processed by a disentanglement head that projects the fused representation into three subspaces: an identity subspace, a clothing-style subspace, and a residual pose-illumination subspace. The identity subspace is regularized to be invariant to changes in the other two subspaces through adversarial gradient reversal and by minimizing the mutual information between the identity code and the style code. Semantic consistency is enforced by requiring

that the identity code extracted from a visual input yield a high cosine similarity with the embedding of any language description that refers to the same individual.

This architecture introduces several system-level trade-offs. First, the overhead of transformer-based cross-attention raises latency and memory consumption, which constrains edge deployment on resource-limited cameras. Offloading fusion to a cloud-based inference server yields higher throughput but incurs communication delays and raises privacy risks associated with transmitting raw person images over public networks. A viable compromise is a split-computation model, where the visual and language encoders are compressed and deployed on edge devices to extract preliminary embeddings, while the heavier cross-modal fusion and gallery search are performed on a local secure server. Model quantization and distillation techniques, applied to both the vision-language backbone and the disentanglement head, can shrink the computational footprint substantially without catastrophic accuracy loss. The granularity of the gallery also demands careful engineering: to support long-term tracking across months, an efficient index structure that supports incremental updates and approximate nearest-neighbor search in the joint embedding space is necessary, yet the dimensionality of the fused embedding must be tuned to balance recall and storage costs.

4. Robust Appearance Disentanglement for Long-Term Identity Preservation

The central challenge in long-term person tracking is the instability of appearance-based features over time. A robust system must be able to recognize an individual who changes from a heavy winter coat to light summer attire, who alters their hairstyle, or who appears under drastically different lighting conditions across indoor and outdoor environments. Disentangling identity-relevant cues from transient visual attributes is thus not merely a representation learning objective but a foundational requirement for sustained retrieval. The proposed architecture achieves this disentanglement through a combination of adversarial training, style transfer supervision, and variational factorization, each contributing to a separation logic that aligns with the semantic priors introduced by language signals.

Adversarial disentanglement operates by training a domain classifier to predict the clothing style from the identity code, and then using a gradient reversal layer to prevent the encoder from encoding style information in that code. This encourages the identity subspace to capture only cues that are invariant across clothing changes. Complementarily, the model is exposed to synthetic training data generated by style transfer techniques, such as adaptive instance normalization [14], which replaces the clothing style of one individual with that of another while preserving body shape and pose. By training the identity encoder to be indifferent to these style-transferred images, the system internalizes a notion of appearance variability that goes beyond the original dataset distribution. A variational autoencoder framework [15] can be adopted to model the latent distribution of clothing appearances, allowing sampling of diverse style codes at training time and further strengthening the invariance of the identity representation.

Language guidance acts as a stabilizing force throughout this process. When the model possesses a textual anchor—for instance, a fixed description of an individual's height, build, and biometric traits—those semantic cues become an additional supervisory signal that constrains the identity subspace. Even if visual evidence becomes ambiguous due to a sudden appearance change, the language embedding can pull the identity code toward a semantically consistent region, effectively functioning as a memory augmentation. This synergy between vision and language reduces the risk of identity drift, a phenomenon in which gradual feature shifts over time cause the system to incorrectly split or merge identities. The design also

permits intentional obfuscation of sensitive appearance features: by architecting the disentanglement to isolate racialized or gender-correlated visual attributes into the style subspace rather than the identity subspace, the system can be tailored to minimize the re-identification salience of such attributes, aligning with fairness objectives that will be examined in the following section.

The longevity of the identity representation also depends on continuous adaptation. As the underlying appearance distribution of a monitored population shifts—due to seasonal clothing changes, fashion trends, or demographic changes—the disentanglement model must be periodically updated. This introduces a dataset shift management problem that is best addressed through lightweight fine-tuning on newly collected unlabeled data using self-training with pseudo-labels, constrained by language anchors to prevent label drift. Balancing adaptation agility against the risk of catastrophic forgetting requires architectural choices such as expanding the style-code vocabulary while freezing the identity backbone, or employing replay buffers that interleave old and new samples during incremental training.

5. Governance, Fairness, and Ethical Deployment

Deploying vision-language guided identity retrieval systems in public or semi-public spaces raises profound governance and ethical questions that must be addressed at the architectural and policy levels, not retrofitted after development. The capacity to track individuals persistently over long periods, across different domains of life, amplifies the power asymmetries between system operators and the monitored population. Consequently, the design of such systems must embed principles of data minimization, purpose limitation, and procedural fairness from the outset. One architectural manifestation of data minimization is the enforcement of disentanglement that explicitly isolates biometric and clothing attributes, making it technically feasible to store only identity codes that are not invertible to raw images, and to discard style information when it is not operationally necessary.

Fairness concerns in person re-identification are well-documented. Unequal representation in training datasets across demographic groups can lead to disparate error rates, where individuals from underrepresented groups experience higher false positive and false negative match rates. Vision-language models, despite their broad pretraining, can inherit cultural and representational biases present in web-scale image-text corpora. These biases may manifest in the language-guided retrieval pipeline when textual descriptions inadvertently encode stereotypical associations, or when the model's attention mechanisms attend disproportionately to visual features that correlate with race or gender. Rigorous auditing protocols are required, as advocated in fairness evaluation frameworks for computer vision [16], and must be embedded in the continuous integration and deployment cycles. Audits should measure differential recall and precision across self-identified demographic categories, as well as false positive rates that could lead to wrongful accusation or surveillance. Where disparities are detected, mitigation strategies may include adversarial debiasing of the identity subspace, resampling-based data augmentation, or the use of separate calibration thresholds for different groups, though the latter must be carefully weighed against the risk of legitimizing differential treatment.

Accountability and transparency pose additional system design challenges. When a retrieval decision is contested—such as the misidentification of an individual in a law enforcement context—operators and affected individuals must be able to interrogate the evidence on which the decision was based. This requires that the system provides interpretable explanations, for instance by highlighting the image regions and language tokens that most influenced the

match score. The architectural inclusion of attention map visualization and natural language rationales, while computationally modest, becomes a governance mandate. Drawing from sociotechnical perspectives on fairness in machine learning systems [17], the very abstraction of "identity" within the embedding space must be continually re-evaluated against the social context of deployment. Policies governing the retention period of identity embeddings, the linkage of identities across separate camera networks operated by different entities, and the circumstances under which human operators can override automated decisions must be codified and made transparent to the public. Deployment in international settings further requires alignment with diverse regulatory regimes, such as the European Union's General Data Protection Regulation and its evolving provisions on biometric processing.

6. Sustainability and Infrastructure Considerations

The computational demands of large vision-language models, particularly when combined with transformer-based cross-modal fusion, entail a significant carbon footprint that must be accounted for in any responsible deployment strategy. Training a high-capacity joint encoder from scratch is environmentally expensive, yet the current paradigm of fine-tuning pre-trained weights substantially reduces this burden. To further shrink the operational carbon budget, model compression through knowledge distillation and quantization can be applied before deployment to edge devices, allowing the bulk of inference to be executed on energy-efficient processors. Structural sparsity techniques can be incorporated to prune attention heads that are not instrumental for identity disentanglement, yielding models that maintain accuracy while reducing floating-point operations by more than half. These practices align with the broader Green AI movement, which advocates for reporting energy and carbon costs alongside accuracy metrics [18].

Infrastructure sustainability also involves decisions about data storage and transmission. In a centralized cloud architecture, the streaming of high-resolution video from hundreds of cameras to a remote data center consumes substantial network bandwidth and energy. Federated learning and hierarchical processing present alternatives: video feeds can be processed locally to extract lightweight feature descriptors, with only those descriptors and associated language captions transmitted to a regional server for fusion and gallery search. This not only reduces communication energy but also provides a natural privacy barrier. Life-cycle analyses of large-scale AI systems have shown that the inference phase often dominates total energy consumption over a multi-year operational life [19], underscoring the importance of optimizing inference efficiency even if it entails modest reductions in absolute accuracy. The trade-off between performance and energy must be explicitly managed through a service-level agreement that defines acceptable latency and accuracy targets under a fixed energy budget.

In addition to the direct environmental impact, the resilience of the tracking infrastructure to resource constraints in under-provisioned environments must be considered. Deployments in smart city contexts in low-resource regions may have intermittent power and bandwidth, requiring models to function in an offline or partially online mode. Lightweight vision-language encoders optimized for mobile processors, combined with on-device gallery caches, can sustain identity retrieval during connectivity outages. The system must be designed with graceful degradation: when full cross-modal fusion is unavailable, a fallback to a purely visual or purely textual retrieval mode should minimize the drop in tracking continuity. Such resilience planning expands the equitable applicability of the technology beyond well-resourced metropolitan centers.

7. Evaluation Frameworks and Deployment Challenges

Evaluating the effectiveness of a long-term person tracking system demands metrics that go far beyond the rank-1 accuracy commonly reported on static ReID benchmarks. A system may exhibit high short-term matching scores yet fail catastrophically when clothing changes occur hours or days later. Evaluation protocols must therefore incorporate temporal decay curves that measure matching probability as a function of time elapsed since the last gallery enrollment, as well as consistency metrics that capture the frequency of identity switches within a tracked trajectory. Cross-dataset generalization, tested by training on one set of camera environments and evaluating on another, provides essential insight into the robustness of disentanglement and fusion designs. The inclusion of natural language queries introduces additional evaluation dimensions: retrieval accuracy conditioned on varying levels of description specificity, and the ability of the system to alert the operator when a description is too ambiguous to yield reliable results.

Operational deployment further exposes the system to data shift challenges that are rarely captured in academic benchmarks. Seasonal changes, fashion cycles, and demographic shifts in a monitored space cause a gradual divergence between the distribution of training data and the live stream. Without continuous adaptation, retrieval accuracy will degrade. Designing a reliable continuous learning loop that updates embeddings without requiring manual re-labeling is a non-trivial systems engineering problem. The fusion module must be capable of detecting when its confidence in identity matching falls below a safety threshold, triggering either conservative rejection or escalating to human-in-the-loop review. This ties directly to the governance requirement for contestability: a well-instrumented system log records every match decision along with the contributing sensory and linguistic evidence, enabling post-hoc audits.

A case illustration drawn from a hypothetical airport security setting highlights these interwoven concerns. The system must track passengers from check-in to boarding, accommodating clothing changes as passengers don travel attire, as well as partial occlusions from luggage and crowds. Language descriptions entered by security personnel, such as “wearing a bright red scarf” become ephemeral once the scarf is removed, but a correctly disentangled identity code should still anchor the match. At the same time, the system’s energy envelope must be consistent with the airport’s sustainability commitments, and bias audits must ensure that error rates remain equitable across an international traveler population. The governance framework must specify that identity embeddings are purged within a legally defined period after the individual leaves the secured area. This multi-layered scenario illustrates that technical, operational, and normative requirements cannot be treated in isolation but must be co-designed.

8. Conclusion

This paper has presented an interdisciplinary analysis of vision-language guided identity retrieval with robust appearance disentanglement as a foundation for long-term person tracking systems. By integrating semantic language cues into the visual re-identification pipeline and architecting a disentangled representation that separates enduring identity attributes from changeable clothing styles, the proposed paradigm addresses the fundamental fragility that limits the temporal scope of current tracking infrastructures. The analysis extended beyond algorithmic design to examine the structural trade-offs in cloud-to-edge deployment, fairness auditing, privacy-preserving data flows, and the carbon footprint of large-scale multimodal models. We argued that sustainable and ethically defensible

deployment requires embedding governance mechanisms directly into the system architecture, including interpretable matching, demographic bias monitoring, and strict data retention policies. Looking ahead, the convergence of vision-language modeling with lifelong learning and federated processing will likely drive the next generation of person tracking infrastructures. Realizing this potential responsibly will demand continued collaboration among computer vision researchers, systems engineers, ethicists, and policymakers to ensure that technical performance never outpaces societal safeguards.

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. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

2. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Advances in Neural Information Processing Systems (Vol. 35, pp. 12888–12900).

3. 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).

4. Xiao, T., Li, S., Wang, B., Lin, L., & Wang, X. (2017). Joint detection and identification feature learning for person search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3415–3424).

5. Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. arXiv preprint, arXiv:1703.07737.

6. Ge, Y., Chen, D., & Li, H. (2020). Mutual mean-teaching: Pseudo label refinery for unsupervised domain adaptation on person re-identification. In International Conference on Learning Representations.

7. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., & Van Gool, L. (2019). Disentangled person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 12089–12098).

8. Eom, C., Lee, G., Lee, J., & Ham, B. (2021). Cloth-changing person re-identification using pose and shape. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 4220–4228).

9. 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.

10. Liu, X., Song, M., Tao, D., Zhou, Z., Chen, C., & Bu, J. (2019). Video-based person re-identification with accumulative motion context. IEEE Transactions on Circuits and Systems for Video Technology, 29(9), 2776–2789.

11. Sun, P., Zhang, R., Jiang, Y., Kong, F., Xu, C., Zhan, W., Tomizuka, M., Li, Z., Yuan, Z., & Wang, C. (2019). MV-GAN: Multi-view generative adversarial network for two-stage cross-view person re-identification. In Proceedings of the IEEE International Conference on Computer Vision (pp. 5655–5664).

12. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998–6008).

13. 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 (pp. 4171–4186).

14. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1501–1510).

15. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. In International Conference on Learning Representations.

16. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Proceedings of the Conference on Fairness, Accountability and Transparency (pp. 77–91).

17. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 59–68).

18. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

19. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint, arXiv:2104.10350.