Semantic Person Search via Language-Driven Descriptions generated from CCTV cameras

Diana Souza[1], Khansa Rekik[2] and Rainer Müller[2], Grimaldo Silva[1]

***Abstract*— Person search in real-world surveillance and robotic perception often fails when facial cues are unreliable due to low resolution, motion blur, or occlusion. We propose a clothing-centric person search framework that represents people using appearance attributes (e.g., garment type, dominant colors, accessories) instead of biometrics. Given person tracklets, the system samples frames, extracts clothing cues, and generates schema-constrained structured descriptions using a large language model, enabling consistent semantic indexing. Retrieval is performed with attribute-aware semantic search over these descriptions and compared against appearance-based baselines in controlled multi-camera experiments. A full-dataset analysis of the generated descriptions highlights common failure modes under real-world conditions, including color ambiguity under lighting changes (e.g., black vs dark), posture hallucinations under occlusion (e.g., standing behind a table described as sitting), and occasional prompt deviations that introduce spurious attributes. Runtime results show that detection is lightweight (about 30 ms per frame), while semantic extraction dominates; with parallel workers, the full pipeline processes roughly 0.25-0.34 s of computation per second of video, supporting practical deployment.**

I. INTRODUCTION

Person search across multi-camera surveillance networks is a core task in public safety and urban monitoring. Several approaches rely on facial recognition, which degrades significantly under low resolution, motion blur, and occlusions, typical of real-world CCTV deployments [1]. In such conditions, solutions must fall back on non-biometric appearance cues to locate individuals across camera views. Clothing and upper-body attributes such as garment type, dominant color, and visually salient accessories, remain stable over short observation windows, and do not raise the same privacy concerns as biometric identifiers [2].

Recent work has shown that clothing-centric representations and multimodal large language models (LLM [3]) can be effectively combined for person search and re-identification [4], [5], [6], [7], [8].Yet the integration of multi-camera person detection and cross-camera person matching driven by automatically generated structured semantic descriptions into a single unified pipeline has not been addressed as a whole.

[1]Grimaldo Silva: UNEB; SENAI CIMATEC University. Diana Silva: SENAI CIMATEC University., Salvador, Brazil jose.jgrimaldo@gmail.com

[2]Khansa Rekik and Prof. Rainer Müller are with the Department of Systems Engineering, ZeMA gGmbH, Saarbrücken, Germany khansa.rekik@zema.de rainer.mueller@zema.de

The work is a part of the RICAIP project funded the European Union's Horizon 2020 research and innovation programme under grant agreement No 857306

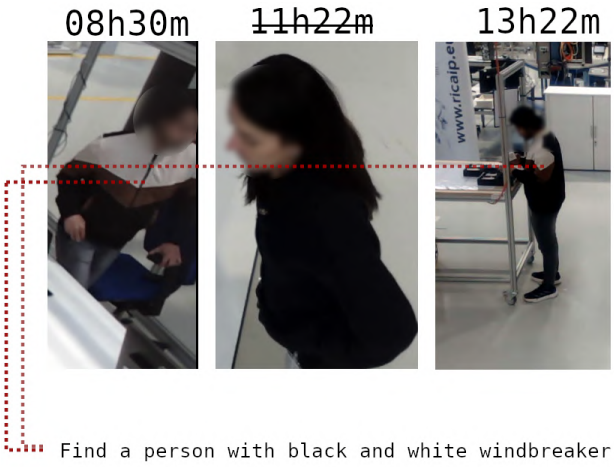


Fig. 1: Our approach enables contextual search with a wide temporal range.

Deploying such systems in realistic multi-camera environments introduces additional challenges beyond low image quality. Individuals are often observed from significantly different viewpoints across cameras , leading to strong variations in scale, pose, illumination, and visible clothing regions [9], [10]. Partial and full occlusions caused by machinery, structural elements, or other people frequently interrupt trajectories, producing fragmented tracklets and incomplete attribute evidence. Crowd density and dynamic backgrounds further increase detection ambiguity, while temporal gaps between camera fields of view make consistent cross-camera association difficult.

Our approach provides person feature matching with a wide temporal scope, as shown in Fig. 1, enabling search over extended time intervals while maintaining a reasonable number of entries. The proposed architecture supports querying across multiple videos and returns precise timestamps. Central to the method is the use of short-term tracklets, defined as short, fragmented trajectories of detected objects spanning only a few frames, which are combined to generate succinct descriptions of individuals even under partial occlusion. These descriptions are then structured in a form that can be efficiently indexed and retrieved.

As validation, we conducted an experimental validation using controlled real world experiments. The evaluation covers computational performance, robustness across heterogeneous video sources, and operational feasibility in an industrial smart factory setting.

II. State of the Art

Recent advances in person search have increasingly leveraged vision-language models to overcome the limitations of traditional appearance-based methods, particularly in challenging surveillance scenarios with low resolution, occlusion, and motion blur.

In contrast to classic approaches, such as Li et al. [11], which focuses on learning image-text affinity for person search using natural-language queries and benchmark evaluation, our approach targets a more operational CCTV setting. Instead of ranking isolated person images from a dataset, our approach generates structured semantic descriptors from tracked video segments, stores them with temporal metadata.

Wang et al.[8] proposed the use of Large Vision-Language Models (LVLMs) for person re-identification by employing instructions to guide the model in generating semantic tokens that encapsulate pedestrian appearance attributes including age, gender, and clothing, achieving 92.2% rank-1 accuracy on DukeMTMC-reID. Unlike this token-based ReID setting, we target operational multi-video search by storing structured, human-interpretable descriptors tied to tracklets and timestamps.

Building on vision-language integration, Sellam et al.[4] introduced VLM-PAR, a modular framework using SigLIP2 encoders for pedestrian attribute recognition across large-scale datasets (PA-100K with 26 attributes, PETA with 65 attributes), demonstrating effectiveness through per-attribute cosine similarity between image and prompt embeddings for applications in surveillance, retail analytics, and autonomous driving. Rather than optimizing attribute classification over fixed label sets, we generate schema-constrained natural-language attributes from tracklet frames and emphasize indexability and retrieval in long video archives.

A key paradigm shift toward flexible yet structured representations is evident in recent work [6], which introduced open "attributes" that merge the advantages of rigid attribute-based search (faster inference, consistency) with the flexibility of natural language descriptions, addressing the limitation that predefined attributes cannot capture details outside their schema while enabling dynamic introduction of new features. Our work adopts a similarly structured philosophy, but anchors it in a practical CCTV pipeline that persists descriptors in a database and supports temporal localization across cameras.

This approach aligns with the chat-driven person retrieval system by Xie et al. [5] that generates detailed clothing descriptions through multi-turn text interaction, producing outputs such as "wearing a white down jacket and black trousers" for surveillance applications.

On the tracking front, Ortac Kosun et al. [7] demonstrated clothing-based person tracking through fusion of multiple feature types (HOG, Gabor, Color, and VGG16) to create discriminative signatures that maintain identity continuity despite occlusions in complex surveillance environments.

For clothing-changing scenarios, recent work [12] addresses the challenge by leveraging Vision-Language Pre-training (VLP) models to capture clothing semantics through causality-based purification, recognizing that clothing remains an important visual attribute for person description in real-world surveillance where facial recognition may be unavailable due to low-quality images, occlusion, or privacy concerns [13], [2].

III. Semantic Feature Representation

A structured representation of semantic attributes describing a person's behavior and clothing should support efficient indexing and, consequently, search. Rather than storing a single free-text description for the entire body, we partition the textual description into four fields: three dedicated to clothing (upper clothing, lower clothing, and hair or headwear) and one dedicated to behavior. This separation improves retrieval by enabling targeted queries and straightforward grouping of related elements.

As an interchange format, we adopt JSON due to its ubiquity and the flexibility afforded by its semi-structured nature.

A simplified schema is shown below:

Listing 1: Schema

```
{
  "_id": "object_id",
  "people": [
    {
      "id": "string",
      "upper_clothing": "string",
      "lower_clothing": "string",
      "hair_or_hat": "string",
      "behavior": "string"
    }
  ],
  "video_name": "string",
  "video_creation_dt": "datetime",
  "frame": "number",
  "timestamp_seconds": "number",
  "yolo_id": "string" % Used for tracklets
}
```

A difficult design choice concerned how to represent color. Several options were considered, including RGB hex codes, HSV values, and more detailed numerical descriptions. RGB and HSV would allow the use of distance metrics but both representations become unreliable at low saturation levels.

An additional complication is that color detection is performed by an integrated LLM, which naturally introduces some variability because of its probabilistic nature.

Since no definitive or robust numerical color representation proved suitable for reliable retrieval at this stage, the simplest and most interpretable approach was adopted: a textual color label, optionally combined with descriptive modifiers. For example, expressions such as "blue-red striped t-shirt" can be directly indexed and queried using a keyword-based search system. This strategy favors interpretability and compatibility with language-based descriptions generated by multimodal models, while avoiding ambiguities in our current method for extracting numerical color encodings.

Another challenging aspect concerns partially occluded garments. In such cases, the corresponding clothing attribute

fields are allowed to remain empty when reliable visual evidence is not available. This prevents the system from introducing potentially misleading or hallucinated descriptors, while preserving consistency in the structured representation.

An illustrative example incorporating these design decisions is presented below with a simulated example of two people being detected while walking, one of which is partially occluded.

Listing 2: Example JSON entry generated for an image


```
{
  "_id": {
    "$oid": "6985239d06a5b859204d63fc"
  },
  "people": [
    {
      "id": "213",
      "upper_clothing": "Yellow-red jacket.",
      "lower_clothing": "",
      "hair_or_hat": "Short, dark hair.",
      "behavior": "Walking, from the side."
    },
    {
      "id": "214",
      "upper_clothing": "Dark coat or blazer.",
      "lower_clothing": "Dark pants.",
      "hair_or_hat": "Short, dark hair",
      "behavior": "Walking, front profile."
    }
  ],
  "video_name": "<path>/240126_180214_C1.mp4",
  "video_creation_dt": 2026-01-24 18:02:14,
  "frame": 238,
  "timestamp_seconds": 9.93,
  "yolo_id": "58-70-77-87-111-143"
}
```


As a result, the chosen storage method can be inspected using keyword-based queries such as “blue shirt” or “navy blue shirt” when more fine-grained filtering is required.

IV. Semantic People Analysis and Retrieval

The proposed system aims to enable semantic search of individuals in CCTV footage based on visual appearance attributes, that could allow for retrospective queries over several hours of recorded video. The architecture, as shown in Fig. 2 integrates computer vision, multimodal language models, and scalable data storage to detect, track, describe, and index people captured by surveillance cameras.

A. Building Semantic Analysis

The software architecture is composed of several connected components.

YOLOv11 [14] is used to detect individuals in each video frame. Detected persons are associated over time using a tracking mechanism, forming tracklets that represent continuous observations of the same individual. This approach avoids creating redundant database entries for every frame and significantly reduces storage and processing overhead.

When a newly detected individual cannot be associated with any existing tracklet, a representative image is selected and processed by a multimodal LLM (Gemini). The model generates a textual description of visually salient attributes, a json of Fig. 3, such as clothing type, color, footwear, and visible accessories.

The prompt itself required only two or three revisions to reach its final current wording:

Semantic Descriptor Generator (Template).

```
You are a helpful assistant.
Task: For each person visible in the
cropped video frame, write a brief
description of their appearance.
Use exactly four fields:
upper_clothing, lower_clothing,
hair_or_hat, behavior.

Constraints:
- Output must be in a structured,
  machine-parseable format (JSON).
- Do not mention the absence of items
  (e.g., do not write "no hat").
- Keep descriptions short & concrete.
- Leave fields empty if occluded
Input: {IMAGE}
```

Even if additional visual information becomes available during tracking (e.g., due to reduced occlusion or improved viewpoint), the previously available tracklet semantic information are not updated.

Generated descriptions are normalized to account for synonyms, plurals, and other variations. Afterwards, descriptions are structured in JSON format to ensure consistent representation and facilitate efficient querying.

Due to the large volume of frames generated even within a single day of operation, a database solution capable of efficiently handling millions of records was required. To address this requirement, the system adopts a NoSQL architecture [15], specifically MongoDB [16]. This choice is motivated by its horizontal scalability, flexible schema design, and efficient management of large collections of semi-structured documents, which align well with the dynamic and heterogeneous nature of the extracted semantic descriptors.

B. Fast Semantic Retrieval

During the retrieval phase, the user specifies a set of appearance-based attributes describing the individual of interest. Given a free-text search string `busca`, the method first tokenizes it into words and applies a lexical normalization and dictionary replacement step to each token, yielding a set of normalized patterns

$$T = \{t_1, \ldots, t_n\}. \tag{1}$$

The search is performed over the array field `people`, unique tracklets present in a frame, where each element is a person descriptor with the searchable fields (as represented in Sec. III)

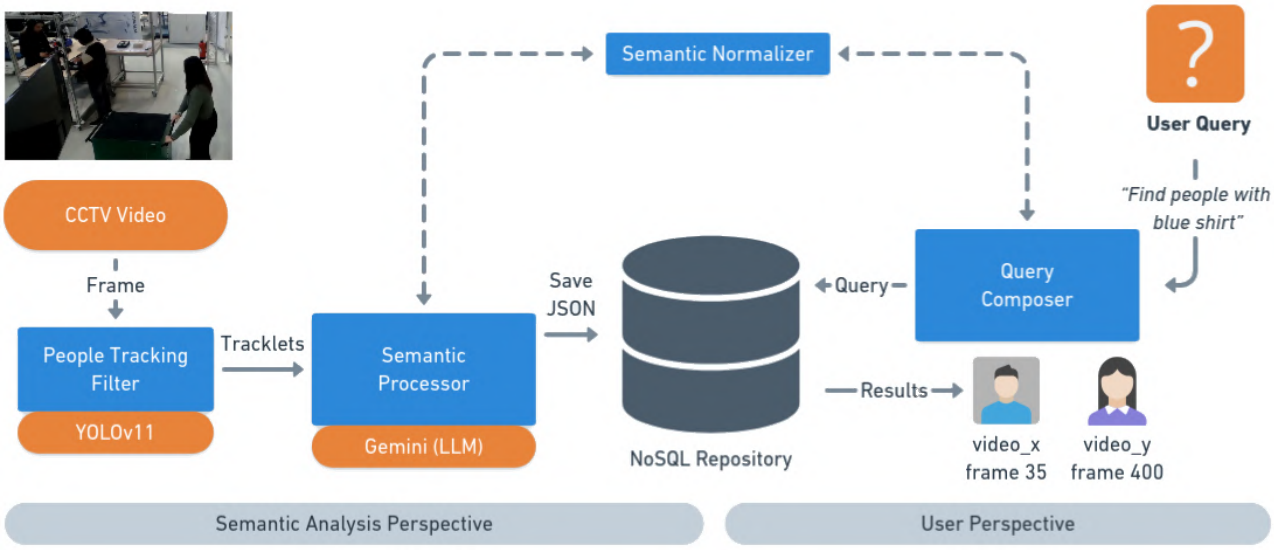


Fig. 2: Architecture of our solution, orange regions indicate existing solutions while blue regions indicate developed ones.

(a) Input frame

Upper Clothing: navy blue short-sleeve t-shirt
Lower Clothing: dark long pants (likely jeans or cargo) with dark shoes
Hair or hat: short hair, shaved on the sides, light brown or dark blond
Behavior: walking casually from right to left, slightly leaning forward, arms relaxed, observing surroundings

(b) Generated description

Fig. 3: Example semantic description generated for a cropped frame. Note that the shoe color description was hallucinated due to occlusion, actual shoe color was white.

$$F = \begin{Bmatrix} \texttt{upper\_clothing}, \\ \texttt{lower\_clothing}, \\ \texttt{hair\_or\_hat}, \\ \texttt{behavior} \end{Bmatrix}.$$

These attributes are translated into database queries that match stored descriptions. An item is considered a match if all query tokens $\{t_1, \ldots, t_n\}$ are present within the corresponding category field.

The system returns candidate results along with their temporal and spatial references, enabling users to locate individuals within historical video data.

For ease of use, the corresponding frames from each video can be presented for ease of validation. Although a more detailed interface would improve usability it is outside the scope of this work.

C. Limitations

Multimodal language models may generate semantically similar but lexically different descriptions, which can affect search precision. Moreover, visual attributes such as colors or clothing types may be subject to some ambiguity, especially under varying illumination or low image quality.

Another limitation is that the semantic normalizer, as of now, is a simplistic replacement of common terms used in search, however, a potential alternative would rely on embeddings that would more naturally encode such semantic patterns.

V. Experimental Validation

This section experimentally validates the proposed approach in a smart factory setting, with a variable number of participants performing diverse interactions over extended periods. The dataset comprises six scenarios, each recorded simultaneously from three viewpoints, enabling evaluation over longer sequences and assessment of result repeatability across cameras.

Overall, the experiments include 18 videos (2-4 min each) with 1-7 participants per scenario. To increase variability and improve generalization, recordings use two resolutions: 1920×1080 and 800×600. The total recorded duration is 59 min 58 s. Videos depict a simulated workflow of workers moving and executing tasks in a factory-like environment (Fig. 4). To preserve privacy, the database stores no physically identifying information (e.g., facial features).

We first report runtime metrics to assess practical throughput, including processing time per frame as a function of input resolution and the computational cost of detection, tracking, description generation, and database operations. Afterwards, results related to retrieval metrics are presented.

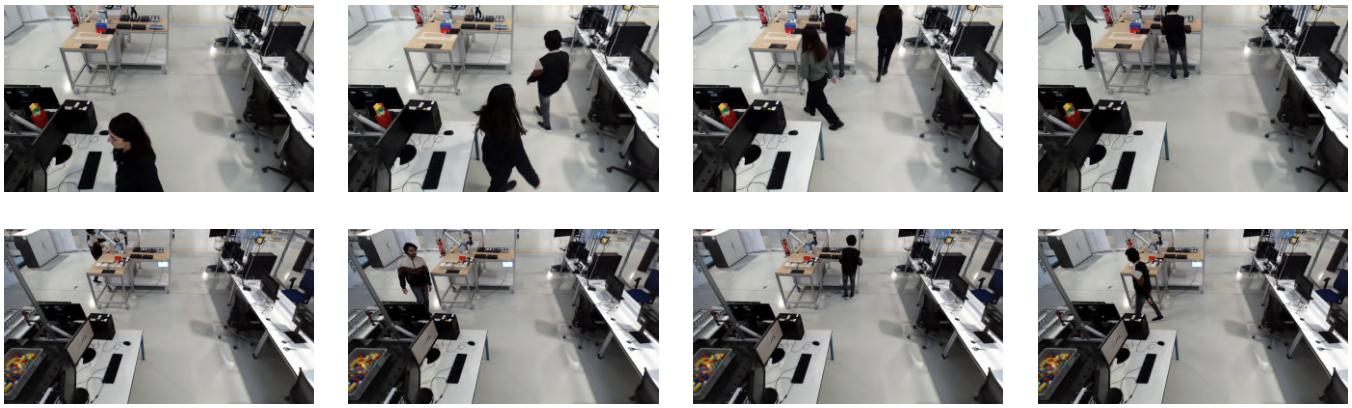


Fig. 4: Workers inside a smart factory during routine procedures. First row shows three workers rotating in and out of the screen and second row shows one worker doing activities.

TABLE I: Runtime and throughput summary (mean±std) grouped by resolution and frame rate.

Res. (px)	FPS (Hz)	N	Video dur. (s)	YOLO (ms)	Gemini (s)	Time Proc. Video (s)
1920×1080	30	6	190.9±29.9	34.2	6.51	53.7±17.3
1920×1080	60	6	196.3±30.7	31.3	6.74	66.5±19.9
800×600	30	6	212.7±20.1	30.3	6.55	53.3±19.9

A. Runtime metrics

In terms of runtime performance, our evaluation focuses on the user-perceived runtime, i.e., time spent on semantic analysis and time spent searching semantic features.

As shown in Table I, we analyze the effect of video frame resolution, frames per second (FPS), and video duration, as well as the time spent detecting people with YOLO (per frame), extracting semantics with Gemini (per frame), and the average time to process a full video. The results show that each second of a 30 FPS video requires 281.2 ms of processing time, while for 60 FPS it requires 338.9 ms. Image resolution did not significantly impact YOLO or Gemini processing time; additionally, the overall end-to-end processing time for 800×600 at 30 FPS (53.3±19.9 s) is similar to 1920×1080 at 30 FPS (53.7±17.3 s) in this setup.

Although single-frame semantic extraction (Gemini) takes about 5-8 s on average, parallel execution with 10 workers yields an effective throughput of 1.2-1.6 frames/s (about 0.6-0.9 s per frame), making the overall runtime primarily driven by Gemini plus the video-stage I/O/preprocessing overhead.

For feature search, we measured time required to find a certain semantic feature in the database and also the time required to determine that a given item does not exist. Even for thousands of entries the search time was almost negligible well below 0.02 seconds.

Finally, for over an hour of content, the total price considering both input and structured output tokens was U$ 2.90.

B. Metrics Evaluation

After generating semantic descriptions for all videos and inserting them into the database, the next objective was to evaluate the quality of these descriptions and identify their most common limitations. The full dataset was used for this assessment to ensure a comprehensive analysis across different scenarios, viewpoints, and interaction types.

The results presented in Table II reveal several notable patterns. First, there is a recurring ambiguity between the use of black and dark when describing clothing colors. In some cases, garments labeled as green in one frame were described as dark or black in subsequent frames, likely due to changes in illumination and viewpoint. This indicates a sensitivity of the language model to lighting conditions, which affects color consistency across frames as shown in see Fig. 5. These cases were not considered false positives due to a lack of frame-by-frame consistency in the model.

Additionally, contextual occlusion contributed to a significant number of false positives. For example, individuals standing behind a table were frequently described as sitting when their lower body was not visible (see Fig. 6). The limited visual context led the model to infer posture incorrectly.

Another source of false positives arises from the model that occasionally deviates from the prompt instructions. Despite being instructed not to mention absent attributes, the model sometimes explicitly stated that a person was not wearing a hat, which triggered matches for headwear-related queries. These observations highlight limitations related to semantic interpretation under occlusion, lighting variability, and prompt adherence.

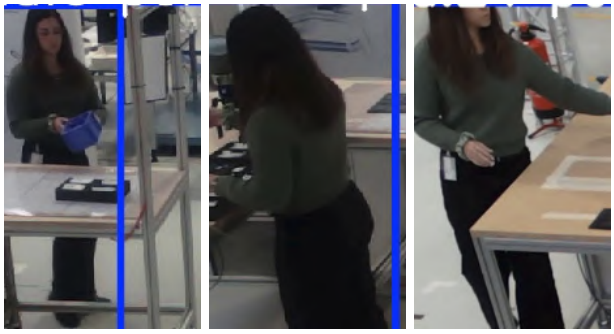
Despite these shortcomings, there were no detectable cases where a given color was visible in clear manner and the model was found to have made a clear error. The lack of frame-by-frame consistency is certainly a point to be evaluated.

VI. CONCLUSIONS

This paper presented an end-to-end system for semantic person search in videos based on non-biometric, clothing-centric descriptors. The proposed pipeline integrates person detection and tracking with multimodal description generation, semantic normalization, and database indexing, enabling keyword-based retrieval across multiple videos and over extended time intervals. In contrast to approaches that

TABLE II: Search performance per keyword query for 952 entries containing at least one person.

Search Keywords	Exists	True Positives	False Positives
hat/cap/beanie/hood	No	0	52
blouse	Yes	323	–
dark blouse	Yes	204	?
green blouse	Yes	233	-
dark green blouse	Yes	36	?
olive green blouse	Yes	66	-
black jacket	Yes	18	?
dark jacket	Yes	64	?
sitting	Yes	117	273



(a) Dark color (b) Dark green color (c) Green color

Fig. 5: Example of limitation in color description of blouse between frames showing the same person under different illumination.

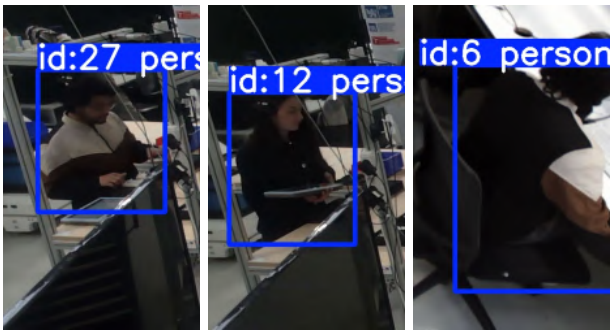
depend on facial cues, our representation focuses on apparel and observable behavior, which remain useful under typical CCTV constraints and better align with privacy-preserving requirements.

Experimental validation was conducted under three complementary settings with promising results. However, the solution still faces limitations. Descriptions produced by multimodal models may vary lexically while remaining semantically equivalent, which can reduce precision for strict keyword matching. In addition, color and apparel type remain sensitive to illumination changes and image quality, occasionally leading to false negatives.

Future work will focus on improving robustness to illumination and appearance variability, for example by introducing color normalization or more stable categorical encodings for clothing attributes. We also plan to replace the current keyword-based normalization with embedding-based retrieval to better handle paraphrases and semantically similar descriptions. Finally, strengthening tracklet-level aggregation, so that descriptors remain consistent across partial occlusions and pose changes, is expected to reduce fragmentation and further improve retrieval reliability.

REFERENCES

[1] P. Li, L. Prieto, D. Mery, and P. Flynn, "On low-resolution face recognition in the wild: Comparisons and new techniques," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2000–2012, 2019.



(a) Incorrect 'sitting' description (b) Incorrect 'sitting' description (c) Correct 'sitting' description

Fig. 6: Example of limitation in behavior description due to occlusion, that can cause hallucinated descriptions. In some cases users standing next to a table were considered to be sitting.

[2] K. Iyshwarya Ratthi, B. Yogameena, M. Swathi, A. Harini, P. Shanmugapriya, and S. Saravana Perumaal, "Attire in attention: Enhancing cctv surveillance with cloth-based image retrieval," in *Computer Vision and Image Processing (CVIP 2024)*, ser. Communications in Computer and Information Science, vol. 2474. Springer, 2024.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[4] A. Sellam, S. Bekhouche, F. Dornaika, C. Distante, and A. Hadid, "Vlm-par: A vision language model for pedestrian attribute recognition," *arXiv preprint arXiv:2512.22217*, December 2024.

[5] Z. Xie, C. Wang, Y. Wang, S. Cai, S. Wang, and T. Jin, "Chat-driven text generation and interaction for person retrieval," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, November 2025, pp. 5259–5270.

[6] Y. Alayary, D. Waref, N. Fathallah, N. Ahmed, S. Abbas, M. A. Abd El Ghany, and M. A.-M. Salem, "Attribute-based people search: open attribute recognition for person re-identification in surveillance and security," *PeerJ Computer Science*, vol. 11, 2025.

[7] G. Ortac Kosun, S. Yilmaz, and R. Samli, "Querytrack: identifying and tracking a person of interest using clothing-based hybrid features," *The Visual Computer*, vol. 42, p. 152, 2026.

[8] Q. Wang, B. Li, and X. Xue, "When large vision-language models meet person re-identification," *arXiv preprint arXiv:2411.18111*, November 2024.

[9] K. Rekik, G. Silva, A. Bashir, and R. Müller, "Towards global awareness in human-robot-collaborative multi-cell assembly system," in *2024 IEEE 20th International Conference on Automation Science and Engineering (CASE)*. IEEE, 2024, pp. 5–10.

[10] G. Silva, K. Rekik, and R. Müller, "Multiperspective approach for semi-autonomous robot control using fusion of exocentric cameras," in *Proceedings of the 2026 International Conference on Robotics, Computer Vision and Intelligent Systems (ROBOVIS)*, ser. Communications in Computer and Information Science. Springer, March 2026.

[11] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *IEEE Conf. on computer vision and pattern recognition*, 2017, pp. 1970–1979.

[12] Z. Yang, H. Zhu, N. Lei, B. Fernando, and Z. Wang, "Clothing purification with causality meets vision-language pretraining models," *Int. Journal of Computer Vision*, vol. 133, no. 11, 2025.

[13] G. Lin, W. Zheng, Z. Li, Y. Lu, X. Ma, and Z. Huang, "Vision-language models for person re-identification: A survey and outlook," *Information Fusion*, vol. 130, p. 104095, 2026.

[14] R. Khanam and M. Hussain, "Yolov11: An overview of the key architectural enhancements," *arXiv preprint arXiv:2410.17725*, 2024.

[15] C. Strauch, U.-L. S. Sites, and W. Kriha, "Nosql databases," *Lecture Notes, Stuttgart Media University*, vol. 20, no. 24, p. 79, 2011.

[16] A. Boicea, F. Radulescu, and L. Agapin, "Mongodb vs oracle–database comparison," in *2012 3rd Conference on Emerging Intelligent Data and Web Technologies*. IEEE, 2012, pp. 330–335.