Language-Based Swarm Perception: Decentralized Person Re-Identification via Natural Language Descriptions

Miquel Kegeleirs[1], Lorenzo Garattoni[2], Gianpiero Francesca[2], and Mauro Birattari[1]

***Abstract*— We introduce a decentralized person re-identification method for robot swarms that uses natural language as its primary representational modality. Unlike traditional approaches based on opaque visual embeddings, the proposed method represents observations with human-readable descriptions. Each robot detects and describes individuals using a vision-language model (VLM), and these descriptions are compared and clustered across the swarm without centralized coordination. Each cluster is distilled into a representative description by a language model, yielding an interpretable summary of the swarm's perception. This enables natural-language querying and supports explainable swarm behavior. Preliminary experiments show competitive identity consistency and interpretability compared to embedding-based methods, despite limitations in text similarity and computational load.**

I. INTRODUCTION

Swarm perception refers to the ability of a robot swarm to leverage the sensory inputs of individual robots to achieve a collective understanding of the environment [1]. Due to their distributed nature, robot swarms can gather, share, and update information in a scalable and fault-tolerant manner [2]. This is particularly useful for people (re-)identification and tracking in environments where static methods are not applicable. While previous work demonstrated decentralized visual people re-identification using visual embeddings [3], we explore a more interpretable alternative: natural language descriptions. Each robot generates textual descriptions of observed individuals—e.g., "a person wearing a black T-shirt and blue jeans"—using a vision-language model (VLM). These descriptions are compared to identify and track individuals across space and time without exchanging raw visual data.

This approach provides two main benefits:

1) It enables users to inspect the swarm's knowledge in human language.
2) It allows intuitive querying: rather than submitting an image, a user can ask, e.g., "have you seen a person in a red hoodie?" and retrieve relevant images.

To manage redundancy, each new description is compared with previous ones. Similar descriptions are merged and summarized using a large language model (LLM). This work investigates whether natural language can serve as an effective medium for decentralized person re-identification. We focus on the swarm's ability to form shared semantic representations under minimal coordination. The method is evaluated in the same simulated environment as previous work. While results are not yet optimal, they highlight the potential of natural language for decentralized swarm perception and communication. More advanced navigation strategies are left for future study.

[1]MK and MB are with IRIDIA, Université libre de Bruxelles, Belgium. [2]LG and GF are with Toyota Motor Europe. Correspondence to: mauro.birattari@ulb.be

The research received funding from Belgium's Wallonia-Brussels Federation through the ARC Advanced Project *GbO*. MB acknowledges support from the Belgian *Fonds de la Recherche Scientifique*–FNRS.

The idea was conceived by MK, LG, GF, and MB. The experiments were designed by MK, LG, GF, and MB, and conducted by MK. The original software was developed by MK. The paper was drafted by MK and edited by MB. All authors reviewed the manuscript. The research was directed by MB.

II. RELATED WORK

In swarm robotics, extensive research has focused on understanding collective behaviors [4], [5] and collective decision-making [6], [7], with perception consistently identified as a key enabler. More broadly, collective perception has gained traction in domains such as (semi-)autonomous vehicles [8], [9] and distributed monitoring systems [10], [11]. In these contexts, person re-identification [12] plays a critical role, enabling agents to consistently recognize and track individuals across different views and locations to support data fusion. Traditional re-identification pipelines rely heavily on deep learning [13], particularly on feature embeddings trained using triplet loss and its variants [14]. Although effective, these embeddings are inherently opaque, limiting transparency and making it difficult for users to interpret or audit the system's internal representations. Recent advances in vision-language models (VLMs) introduce the possibility of encoding visual information in natural language, offering interpretable and semantically rich alternatives to embedding vectors [15], [16]. Language-based interfaces have already enhanced human-robot interaction by allowing robots to follow high-level instructions and express their internal states in a more human-understandable form [17], [18]. While most re-ID research has historically focused on static CCTV surveillance [19], [20], mobile robots offer advantages such as dynamic viewpoints, close-range sensing, and interactive capabilities [21]. These have been explored in single-robot systems for tasks such as face and voice-based re-identification [22], [23], as well as for person-following and user assistance [24]. Robust methods like CARPE-ID have demonstrated reliable person tracking despite occlusions and appearance changes [25]. More recently, multi-robot and swarm-based approaches to person re-identification have begun to emerge [1], [26], [27], demonstrating promising scalability and resilience.

Our work builds on these developments by introducing a language-based approach to person re-identification in

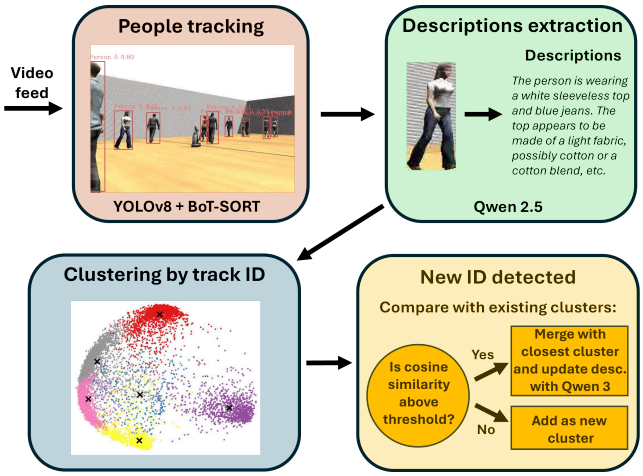

Fig. 1: Data acquisition and clustering process.

robot swarms. Instead of opaque feature vectors, we employ natural language descriptions, enabling interpretable cluster formation and intuitive querying. This approach aligns with growing interest in explainable AI and supports more transparent, accessible interactions between humans and robotic systems [28], [29].

III. METHOD

We adopt a decentralized architecture based on [3] that replaces opaque feature vectors with natural language descriptions. Each robot detects, tracks, and describes individuals with its onboard camera. Descriptions are clustered locally and refined through decentralized communication with nearby peers.

A. Local Data Acquisition

Each robot processes its video stream using a three-stage pipeline (Figure 1):

1) **Detection:** people are detected using `YOLOv8` [30], trained on COCO [31].
2) **Tracking:** `BoT-SORT` [32] assigns IDs and maintains tracks.
3) **Description generation:** each detection is passed to `Qwen-2.5` [33], which outputs a natural language description.

B. Description Similarity and Clustering

Each robot maintains a local database of clusters, where each cluster represents a hypothesized individual and contains semantically similar descriptions. Descriptions within a cluster are passed to `Qwen-3` [34] to generate a representative summary, updated when new descriptions are added. Initially, descriptions are clustered according to tracking IDs provided by `BoT-SORT` (Figure 1). When a new ID is detected, the description is compared to each existing clusters' representative summary using cosine similarity computed on sentence-level embeddings. If similarity exceeds a threshold, it is merged; otherwise, a new cluster is created.

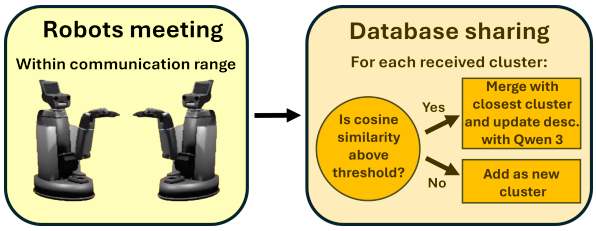

Fig. 2: Data sharing process.

C. Inter-Robot Communication and Merging

Robots periodically exchange their cluster databases upon encountering each other within communication range (Figure 2). Received cluster summaries are compared to local cluster summaries using cosine similarity. Clusters are merged if above the threshold; otherwise, added as new ones. Following a merge, `Qwen-3` regenerates the representative summary. Robots may optionally share images as additional data. These are not used for clustering and only support the final demonstration.

D. Exploration behavior

We adopt a random walk exploration strategy based on ballistic motion [3], providing a simple baseline that isolates the impact of communication and perception.

IV. EXPERIMENTAL SETUP

We adopt the same simulation framework as previous work [3]: a swarm of 4 Toyota HSR robots navigates within a 625 m^2 environment, where 6 or 50 people move freely for 10 minutes. Obstacles take the form of 5 m × 3 m × 0.2 m panels that obstruct both movement and visibility. The environment is simulated in Unity, while robot movements are simulated in ARGoS3 [35] and mirrored via ROS. Each robot follows a random walk while performing detection, tracking, and description generation.

We evaluate the proposed method (`TD-SP`) against an embedding-based baseline (`VE-SP`). Performance is measured using CMC, mAP, and cluster purity [36], [37]. Cluster purity is the proportion of the dominant ground-truth ID; duplicate clusters retain only the largest, and undetected IDs receive 0%. We assess the impact of communication (6-person case) and occlusion (50-person case). We also demonstrate natural-language querying, where each robot retrieves for a given sample description the best-matching image from their database.

V. RESULTS AND DISCUSSION

The 6-person experiments show that `TD-SP` achieves lower or comparable CMC, but higher mAP and significantly higher purity than `VE-SP`. Communication improves performance across all metrics. In the 50-person experiments, `TD-SP` shows a stronger decline in CMC and mAP, while `VE-SP` remains more stable. Purity decreases for both methods, but less for `TD-SP`. This indicates that `TD-SP` is a promising approach for swarm Re-ID, albeit with trade-offs in retrieval ranking performance. With obstacles, `VE-SP` shows reduced purity, while `TD-SP` remains stable and even

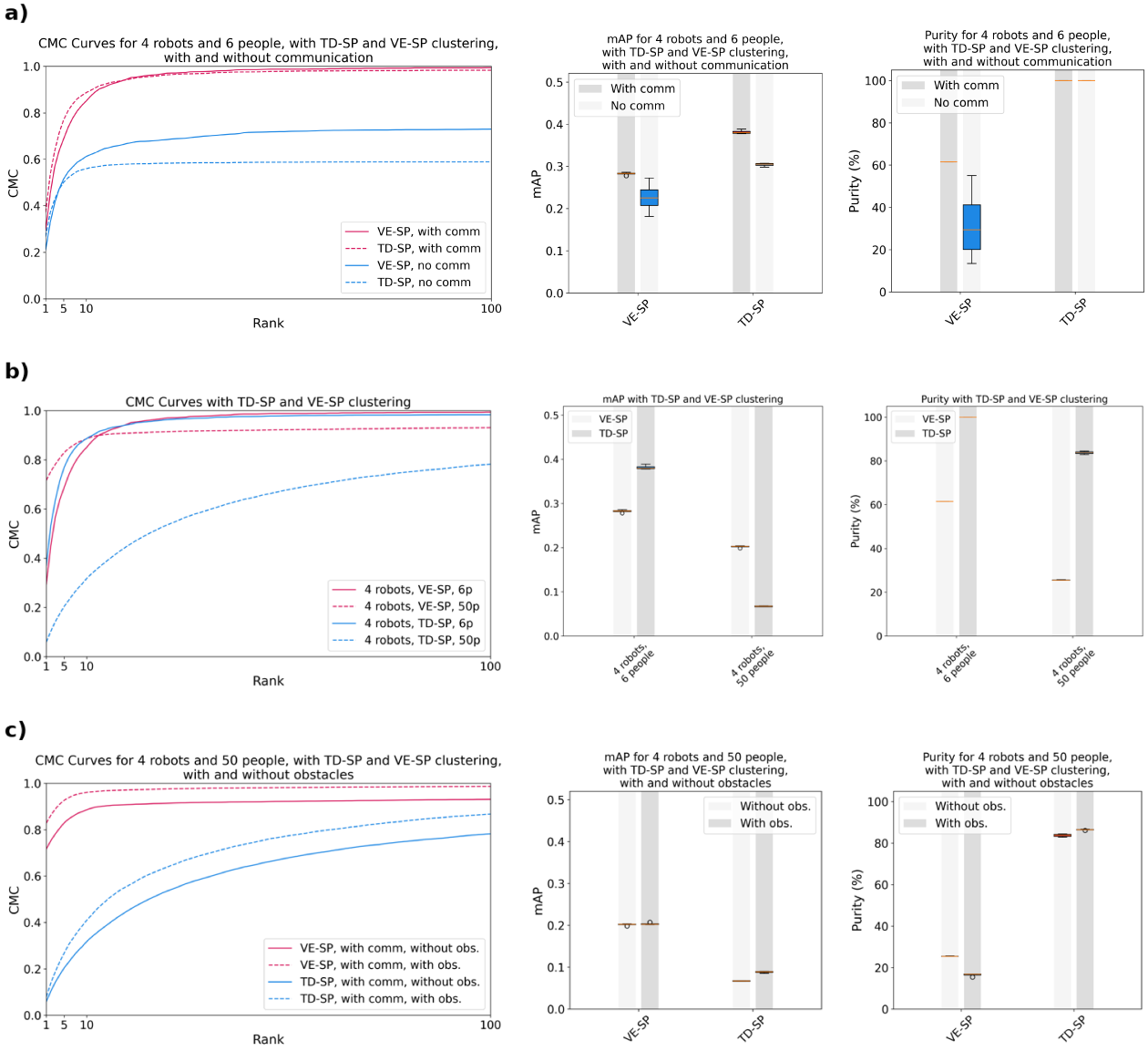

Fig. 3: (a) CMC, mAP, and purity with and without communication for 4 robots and 6 people. (b) CMC, mAP, and purity for 4 robots with 6 or 50 people. (c) CMC, mAP, and purity with and without obstacles for 4 robots and 50 people.

slightly improves. This may reflect better generalization from partially visible features or increased viewpoint diversity introduced by constrained navigation paths. Overall, `TD-SP` improves interpretability, cluster purity and robustness to occlusion, but suffers from limitations that affect scalability and ranking-based metrics. In particular is the over-fragmentation issue: `TD-SP` produces significantly more clusters than `VE-SP`, likely due to rigid cosine similarity over sentence embeddings. This yields many fine-grained clusters with high purity but lower overall matching accuracy—hence, lower CMC and mAP. More flexible matching methods—such as cross-encoders or WMD [38]—could reduce fragmentation at the cost of increase computational cost.

Natural Language Query Evaluation

In the 50-person scenario, each robot is queried with three sample descriptions and returns the best-matching image from its database (Figure 4). All robots consistently retrieve images of the correct individual. For Query 2, robots return different images, suggesting independent cluster formation prior to communication. For Queries 1 and 3, identical images indicate that the cluster was shared across robots through communication. This indicates both independent perception and semantic information sharing within the swarm.

VI. CONCLUSIONS AND FUTURE WORK

We introduced a decentralized person re-identification method based on natural language descriptions. This improves interpretability and enables intuitive querying. While performance is limited by similarity measures and computation, results demonstrate the viability of language-based re-identification. Future work includes improved similarity metrics, lightweight models, contextual reasoning, selective communication, and multimodal integration.

Fig. 4: Natural-language query results for the swarm. Each group corresponds to one query, with four images returned by the robots.

REFERENCES

[1] M. Kegeleirs, *et al.*, "Collective perception for tracking people with a robot swarm," in *ICRA@40*. IEEE, 2024.

[2] M. Brambilla, E. Ferrante, M. Birattari, and M. Dorigo, "Swarm robotics: a review from the swarm engineering perspective," *Swarm Intell.*, vol. 7, no. 1, pp. 1–41, 2013.

[3] M. Kegeleirs, *et al.*, "Assessing the impact of feature communication in swarm perception for people re-identification," *Front. Robot. AI*, vol. 12, p. 1671952, 2025.

[4] V. Trianni and A. Campo, "Fundamental collective behaviors in swarm robotics," in *Handbook of Computational Intelligence*, ser. Springer Handbooks, J. Kacprzyk and W. Pedrycz, Eds. Heidelberg, Germany: Springer, 2015, pp. 1377–1394.

[5] L. Garattoni and M. Birattari, "Swarm robotics," in *Wiley Encyclopedia of Electrical and Electronics Engineering*, J. G. Webster, Ed. Hoboken, NJ, USA: John Wiley & Sons, 2016, pp. 1–19.

[6] G. Valentini, E. Ferrante, H. Hamann, and M. Dorigo, "Collective decision with 100 kilobots: Speed versus accuracy in binary discrimination problems," *Autonomous agents and multi-agent systems*, vol. 30, pp. 553–580, 2016.

[7] V. Strobel, E. Castello Ferrer, and M. Dorigo, "Managing byzantine robots via blockchain technology in a swarm robotics collective decision making scenario," in *AAMAS 2018*. Richland, SC, USA: International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 2018, pp. 541–549.

[8] H.-J. Günther, B. Mennenga, O. Trauer, R. Riebl, and L. Wolf, "Realizing collective perception in a vehicle," in *2016 IEEE Vehicular Networking Conference*, 2016, pp. 1–8.

[9] G. Thandavarayan, M. Sepulcre, and J. Gozalvez, "Analysis of message generation rules for collective perception in connected and automated driving," in *2019 IEEE Intelligent Vehicles Symposium*, 2019, pp. 134–139.

[10] W. Choi and S. Savarese, "A unified framework for multi-target tracking and collective activity recognition," in *Computer Vision–ECCV 2012*, ser. LNCS, A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, Eds., vol. 7575. Springer, 2012, pp. 215–230.

[11] D. Montero, *et al.*, "Multi-camera BEV video-surveillance system for efficient monitoring of social distancing," *Multimedia Tools and Applications*, vol. 82, no. 22, pp. 34 995–35 019, 2023.

[12] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *CoRR*, vol. abs/1610.02984, 2016.

[13] M. Ye, *et al.*, "Deep learning for person re-identification: a survey and outlook," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 6, pp. 2872–2893, 2022.

[14] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," https://arxiv.org/abs/1703.07737, 2017.

[15] L. H. Li, *et al.*, "Grounded language-image pre-training," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2022, pp. 10 965–10 975.

[16] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *2023 IEEE/CVF Conference on Computer Vision*. IEEE, 2023, pp. 11 975–11 986.

[17] J. Atuhurra, "Leveraging large language models in human-robot interaction: A critical analysis of potential and pitfalls," https://arxiv.org/abs/2405.00693, 2024.

[18] H. Rahimi, *et al.*, "User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions," https://arxiv.org/abs/2502.10636, 2025.

[19] N. Ukita, Y. Moriguchi, and N. Hagita, "People re-identification across non-overlapping cameras using group features," *Computer Vision and Image Understanding*, vol. 144, pp. 228–236, 2016.

[20] K. Koide, E. Menegatti, M. Carraro, M. Munaro, and J. Miura, "People tracking and re-identification by face recognition for RGB-D camera networks," in *European Conference on Mobile Robots (ECMR)*, 2017.

[21] Y. Murata and M. Atsumi, "Person re-identification for mobile robot using online transfer learning," in *2018 Joint 10th International Conference on Soft Computing and Intelligent Systems (SCIS) and 19th International Symposium on Advanced Intelligent Systems (ISIS)*. IEEE, 2018, pp. 977–981.

[22] Y. Wang, J. Shen, S. Petridis, and M. Pantic, "A real-time and unsupervised face re-identification system for human-robot interaction," *Pattern Recognition Letters*, vol. 128, pp. 559–568, 2019.

[23] Z. Lu, A. Ashok, and K. Berns, "Roboreid: Audio-visual person re-identification by social robot," in *2024 10th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, 2024, pp. 1758–1763.

[24] H. Ye, *et al.*, "Robot person following under partial occlusion," in *ICRA 2023*. IEEE, 2023, pp. 7591–7597.

[25] F. Rollo, A. Zunino, N. Tsagarakis, E. M. Hoffman, and A. Ajoudani, "Continuous adaptation in person re-identification for robotic assistance," in *ICRA 2024*. IEEE, 2024, pp. 425–431.

[26] M.-A. Popovici, I. E. Gil, E. Montijano, and D. Tardioli, "Distributed dynamic assignment of multiple mobile targets based on person re-identification," in *Iberian Robotics conference*. Springer, 2022.

[27] M. Kegeleirs, *et al.*, "Leveraging swarm capabilities to assist other systems," https://arxiv.org/abs/2405.04079, 2024.

[28] F. Xu, *et al.*, "Explainable ai: A brief survey on history, research areas, approaches and challenges," in *Natural language processing and Chinese computing*, Springer, Ed., 2019, pp. 563–574.

[29] R. Dwivedi, *et al.*, "Explainable ai (XAI): Core ideas, techniques, and solutions," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–33, 2023.

[30] R. Varghese and M. Sambath, "YOLOv8: A novel object detection algorithm with enhanced performance and robustness," in *2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS)*, 2024, pp. 1–6.

[31] T.-Y. Lin, *et al.*, "Microsoft COCO: common objects in context," in *Computer vision–ECCV 2014*. Springer, 2014, pp. 740–755.

[32] N. Aharon, R. Orfaig, and B.-Z. Bobrovsky, "BoT-SORT: Robust associations multi-pedestrian tracking," https://arxiv.org/abs/2206.14651, 2022.

[33] A. Yang, *et al.*, "Qwen2.5 technical report," https://arxiv.org/abs/2412.15115, 2025.

[34] J. Bai, *et al.*, "Qwen technical report," https://arxiv.org/abs/2309.16609, 2023.

[35] C. Pinciroli, *et al.*, "ARGoS: a modular, parallel, multi-engine simulator for multi-robot systems," *Swarm Intell.*, vol. 6, no. 4, pp. 271–295, 2012.

[36] H. Schütze, C. D. Manning, and P. Raghavan, *Introduction to information retrieval*. Cambridge, MA, USA: Cambridge University Press, 2008.

[37] L. Zheng, *et al.*, "Scalable person re-identification: A benchmark," in *2015 IEEE/CVF Conference on Computer Vision*. IEEE, 2015, pp. 1116–1124.

[38] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *ICML 2015*. PMLR, 2015, pp. 957–966.