ARGOS: Who, Where, and When in Agentic Multi-Camera Person Search

Anonymous CVPR submission

Paper ID 17

Abstract

*We introduce ARGOS, the first benchmark and framework that reformulates multi-camera person search as an interactive reasoning problem requiring an agent to plan, question, and eliminate candidates under information asymmetry. An ARGOS agent receives a vague witness statement and must decide what to ask, when to invoke spatial or temporal tools, and how to interpret ambiguous responses, all within a limited turn budget. Reasoning is grounded in a Spatio-Temporal Topology Graph (STTG) encoding camera connectivity and empirically validated transition times. The benchmark comprises 2,691 tasks across 14 real-world scenarios in three progressive tracks: semantic perception (*Who*), spatial reasoning (*Where*), and temporal reasoning (*When*). Experiments with four LLM backbones show the benchmark is far from solved (best TWS: 0.383 on Track 2, 0.590 on Track 3), and ablations confirm that removing domain-specific tools drops accuracy by up to 49.6 percentage points.*

1. Introduction

Identifying a target person across a multi-camera network is a fundamental requirement in surveillance, yet existing approaches remain limited. Traditional person re-identification [20, 25, 26] relies on clear visual queries, while text-based [18, 23] and interactive methods [4, 8, 10, 11, 13, 15] use appearance descriptions alone. None of these formulations equip the system with the ability to *actively plan* questions or leverage *spatial and temporal* cues— information that witnesses routinely provide in real-world scenarios (e.g., "I saw them in the warehouse, then near the lobby a few minutes later"). Recent spatial reasoning benchmarks [5, 7, 22] and agent evaluation frameworks [12, 24] remain limited to single-image or general-purpose settings, leaving interactive spatial-temporal reasoning over camera networks unaddressed.

We introduce **ARGOS** (Agentic Retrieval with Grounded Observational Search), a benchmark and agent framework that formulates person search as an active spatio-temporal reasoning problem. The agent conducts multi-turn dialogue with a witness, invokes tools grounded in a physically validated Spatio-Temporal Topology Graph (STTG), and eliminates candidates whose movements are infeasible—combining multi-modal interaction, spatial grounding, and temporal reasoning in a single evaluation protocol.

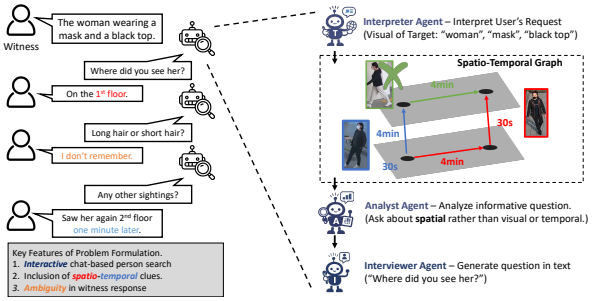

Figure 1. **Overview.** *Left:* An ARGOS agent interacts with an ambiguous witness through multi-turn dialogue, combining appearance, spatial, and temporal queries. *Right:* The four-module agent architecture (Analyst→Planner→Interviewer→Interpreter) forms an observe-think-act loop over the evaluation environment.

Our contributions are:

- We define **interactive multi-camera person search**, a task at the intersection of multi-turn witness interaction and spatio-temporal reasoning over camera networks.
- We construct the **ARGOS benchmark** (2,691 tasks, 14 scenarios, 3 progressive tracks) with Turn-Weighted Success (TWS) as the primary metric, jointly measuring correctness and turn efficiency.
- We introduce the **STTG**, a directed weighted graph encoding camera connectivity and transition-time statistics that serves as both the structural backbone for task construction and the agent's grounding tool for spatial-temporal reasoning.
- We provide a **strong baseline agent** with a four-module pipeline and eight tools, demonstrating that tool-augmented agentic interaction substantially outperforms direct LLM inference.

2. The ARGOS Benchmark

Problem formulation. An agent π receives an initial witness statement about a target g^* in a gallery $\mathcal{G}$ and en-

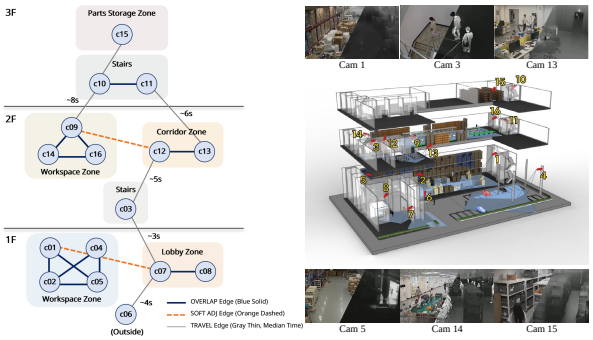

Figure 2. **Left:** STTG for a 16-camera factory environment. Nodes are grouped into zones by `OVERLAP` connectivity. Edge types: `OVERLAP` (blue), `SOFT_ADJ` (orange), `TRAVEL` (gray); labels show median transition time. **Right:** 3D camera layout with sample imagery.

gages in multi-turn dialogue with a witness simulator $\mathcal{W}$. At each turn the agent selects an action— 1) asking about visual attributes, 2) querying spatial location, or 3) invoking temporal reasoning—and receives a natural-language response. The agent accesses an STTG $\mathcal{T}$ encoding camera connectivity and transition-time statistics. A task is solved when the agent identifies the unique target within a limited turn budget.

Spatio-Temporal Topology Graph. The STTG is a directed weighted graph $\mathcal{T} = (V, E)$ where each node represents a camera with a zone label and sub-area description. Each edge e_{ij} carries a type $\lambda_{ij} \in \{\texttt{OVERLAP}, \texttt{SOFT\_ADJ}, \texttt{TRAVEL}\}$ and transition-time statistics $(t_{\min}, t_{\text{med}}, t_{\max}, n)$ computed from observed movements. `OVERLAP` edges connect cameras with shared fields of view; their connected components define *zones*. `TRAVEL` edges cover longer transitions with empirically measured times. The STTG serves two roles: the benchmark derives ground-truth tasks from it (Tracks 2 and 3), and agents receive it as an environment representation for spatio-temporal reasoning (Fig. 2).

Three tracks. ARGOS consists of 2,691 tasks across 14 scenarios, organized into three progressive tracks sharing the same gallery of 1,273 persons observed by 16 synchronized cameras in two environments (factory and campus). Each person is annotated with 24 visual attributes extracted via vision-language models [16].

Track 1: *Who* (989 tasks) tests semantic perception. The agent receives a completed multi-turn dialogue and must extract attributes to filter the gallery.

Track 2: *Where* (550 tasks) tests spatial reasoning. The witness reports a sighting in a multi-camera zone; the agent resolves the specific sub-area through spatial questions and attribute queries (oracle avg. 2.02 turns).

Track 3: *When* (1,152 tasks) tests temporal reasoning. The witness reports two sightings at different times and locations; the agent eliminates candidates whose transitions violate STTG constraints (oracle avg. 1.89 turns).

Witness simulator. ARGOS evaluation uses a deterministic witness simulator $\mathcal{W}$. For three observable attributes (visual gender, upper clothing color, and lower clothing color), selected for high class variation and perceptual salience [14], $\mathcal{W}$ returns the ground-truth value in varied natural-language templates. For other attributes it responds with uncertainty; for spatial and temporal queries it returns pre-computed responses from the ground-truth path. This design ensures reproducibility (identical witness behavior across runs) and diagnostic clarity (agent failures trace to parsing, reasoning, or strategy, rather than witness noise).

Dataset construction. All three tracks share a two-stage pipeline. Stage 1 generates deterministic ground truth through algorithmic computation: information-theoretic clue selection for Track 1, zone-based spatial disambiguation for Track 2, and STTG-based temporal feasibility classification for Track 3. No language model is involved; task validity and optimal turn sequences are computed from the attribute database, camera topology, and trajectory records. Stage 2 wraps the structured ground truth in natural-language dialogues via an LLM, preventing agents from bypassing comprehension by directly matching structured labels.

3. ARGOS Agent

The ARGOS Agent is a tool-augmented LLM agent that processes each turn through four modules (Fig. 1, right): (1) **Analyst** queries the gallery and computes attribute elimination power over the current candidate set; (2) **Planner** decides the next action (attribute question, spatial query, or temporal check); (3) **Interviewer** executes the action via the appropriate tool; (4) **Interpreter** parses the witness's response and applies filters.

The agent leverages recent multimodal LLMs [1, 6, 9] as backbone reasoning engines. The agent accesses eight tools: gallery queries (T1–T2), zone structure retrieval (T3), witness interaction (T4), temporal feasibility checking via the STTG (T5), and filtering/prediction actions (T6–T8). The same tool set serves all three tracks, but the Planner must make qualitatively different strategic decisions in each. A critical design property is the information boundary: the agent does not know which attributes the witness can answer (only 3 of 21 are observable), forcing strategic decisions under uncertainty.

4. Experiments

Metrics. For Track 1 the primary metric is Top-1 Accuracy. For Tracks 2 and 3 the primary metric is Turn-Weighted Success (TWS):

$$\text{TWS} = \frac{1}{N}\sum_{i=1}^{N} s_i \cdot \frac{\tau_i^*}{\max(\tau_i,\, \tau_i^*)} \qquad (1)$$

Table 1. **Main results on Tracks 2 and 3.** Best TWS per track is **bolded**. Blue highlights the best value per column.

	Track 2 (Spatial)				Track 3 (Temporal)			
Backbone	TWS	Top-1	$SR_t@5$	$AvgT_s$	TWS	Top-1	$SR_t@5$	$AvgT_s$
Oracle	1.000	100.0	100.0	2.05	1.000	100.0	100.0	1.88
GPT-5.2	0.338	73.1	33.6	7.04	**0.590**	88.2	65.8	3.91
GPT-4o	0.323	74.5	31.6	7.49	0.567	80.6	60.8	3.91
GPT-5-mini	0.319	74.9	32.2	7.48	0.556	88.0	60.5	4.40
Cl. Sonnet 4	**0.383**	76.0	38.4	6.71	0.548	83.6	59.5	4.25

$SR_t@T$: success within budget T; $AvgT_s$: avg. turns (success only).

Table 2. **Track 1 results** (GPT-4o). LLM Direct outputs a single ID (SR@1 = SR@5).

Agent	SR@1	SR@5	Parse Acc.
Oracle	100.0	100.0	100.0
LLM ToolCall	81.1	82.3	90.8
LLM Direct	73.3	73.3	93.0
Rule-based	32.2	48.0	66.0

All values in %.

where $s_i \in \{0, 1\}$ indicates correctness, τ_i is the agent's turn count, and τ_i^* is the oracle-optimal count. TWS follows the design principle of SPL [2] in embodied navigation, weighting success by efficiency but replacing path length with dialogue turns.

Setup. We evaluate across all three tracks using four LLM backbones (GPT-5.2 [17], GPT-4o [1], GPT-5-mini, Claude Sonnet 4 [3]) at temperature 0.0 with a 20-turn budget. Additional metric definitions (AUC-CRR, premature prediction) and agent behavioral statistics are provided in Sec. D.6 and Sec. D.7.

Main results. On Track 1, structured tool-calling outperforms end-to-end LLM inference (Table 2): LLM ToolCall achieves 81.1% SR@1, +7.8 pp over LLM Direct (73.3%). Although LLM Direct attains slightly higher parsing accuracy (93.0% vs. 90.8%), it cannot recover from filtering errors.

For Tracks 2 and 3, the benchmark remains far from solved (Table 1): the best TWS reaches 0.383 (Track 2, Claude Sonnet 4) and 0.590 (Track 3, GPT-5.2), well below the Oracle. No single backbone dominates both tracks—Claude Sonnet 4 leads Track 2 but ranks last on Track 3—indicating that spatial and temporal reasoning stress different capabilities.

Ablation study. We ablate using GPT-4o (Table 3).

(a) Domain-specific tools are indispensable. Without any tool, Track 3 Top-1 collapses to 11.4%, on par with random guessing. Removing only the track-specific tool is equally destructive: w/o Temporal Tool falls to 31.0% (−49.6 pp) and w/o Spatial Tool to 40.7% (−33.8 pp), with TWS near zero.

(b) Strategic reasoning improves efficiency beyond accuracy. On Track 3, w/o Strategy nearly matches the full agent in Top-1 (76.9% vs. 80.6%), yet its TWS is sharply lower (0.373 vs. 0.567) because it requires twice the turns.

Table 3. **Ablation results** (GPT-4o). Red marks the steepest drop per track.

		Track 2			Track 3		
Agent	**Removed**	TWS	Top-1	$AvgT_s$	TWS	Top-1	$AvgT_s$
Oracle	*(upper bound)*	1.000	100.0	2.05	1.000	100.0	1.88
ARGOS Agent	*(none)*	0.323	74.5	7.49	0.567	80.6	3.91
w/o Strategy	Reasoning	0.136	47.6	10.78	0.373	76.9	7.73
w/o Spatial	Spatial tool	0.063	40.7	14.20	—		
w/o Temporal	Temporal tool		—		0.054	31.0	13.44
Single-Pass	Interaction	—	64.7	—	—	11.3	—
w/o All Tools	All tools		N/A		—	11.4	—

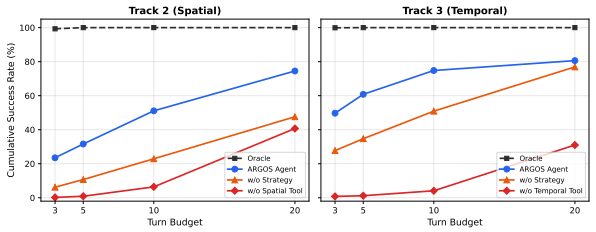

Figure 3. **Cumulative success rate ($SR_t@T$) vs. turn budget** for Track 2 (left) and Track 3 (right), GPT-4o. The ARGOS Agent (blue) resolves tasks substantially faster than w/o Strategy (orange). w/o Spatial/Temporal Tool (red) stays near zero.

(c) Agentic interaction is the only viable path. Single-pass baselines converge near 11% Top-1 on Track 3—a structural floor that stronger LLMs cannot overcome; only tool-augmented multi-turn interaction reaches 80.6%.

Failure pattern analysis. Failure patterns reveal distinct behavioral profiles across backbones. On Track 3, GPT-4o fails primarily through premature commitment: 91% of its failures are wrong predictions, with negligible timeout (1.6%). GPT-5.2 shows the opposite profile: 46% of failures are timeouts from extended evidence gathering, yielding fewer total errors and the highest TWS despite higher per-task latency. These profiles suggest that TWS rewards sustained disambiguation over rapid commitment. The two tracks also probe distinct capabilities: Track 2's bottleneck is natural-language understanding (78.6% parse success vs. 94.2% for Track 3), while Track 3's bottleneck is temporal reasoning (w/o Temporal Tool: −49.6 pp, the steepest single-component drop). This explains the backbone crossover in Table 1 and confirms that the two tracks evaluate complementary aspects of agentic person search.

Figure 3 visualizes the efficiency advantage: at turn 5 on Track 3, the full agent reaches 60.8% success vs. 34.7% for w/o Strategy, confirming that TWS captures a real behavioral difference beyond what Top-1 alone reveals.

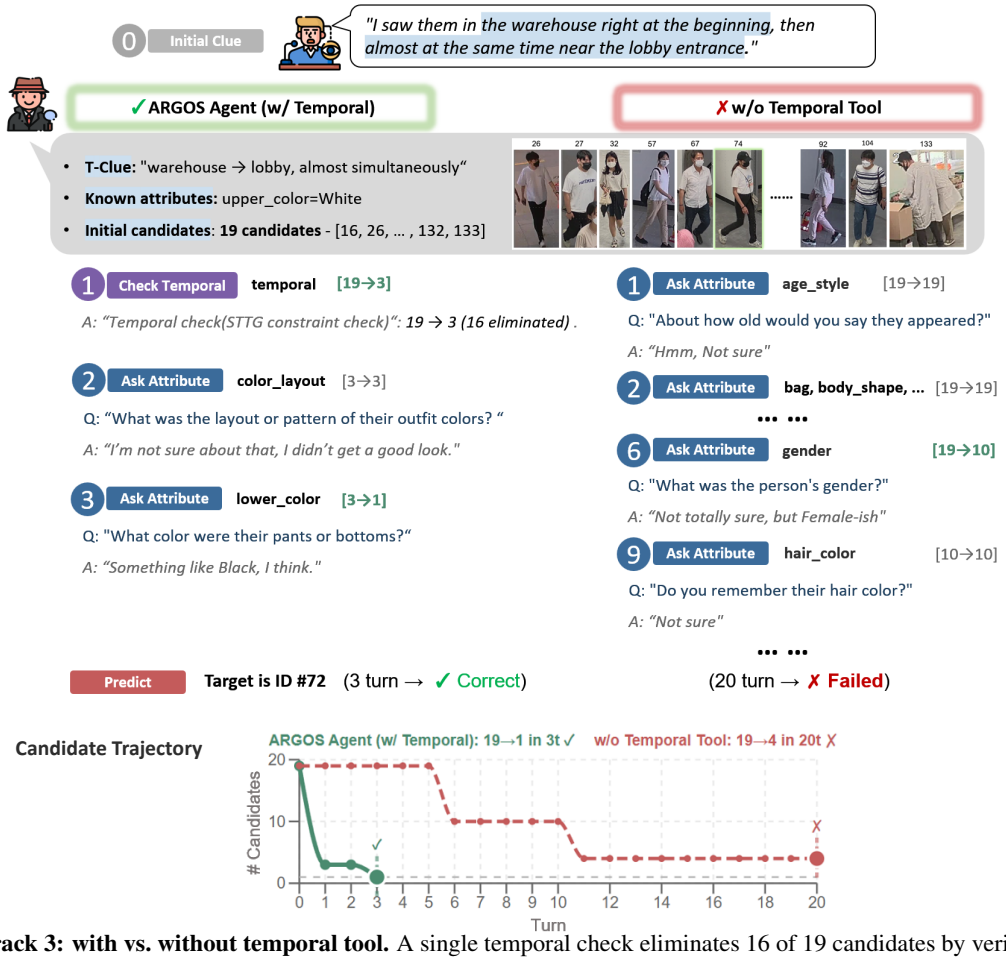

Figure 4. **Track 3: with vs. without temporal tool.** A single temporal check eliminates 16 of 19 candidates by verifying spatio-temporal transition feasibility against the STTG, demonstrating the tool's information density.

5. Qualitative Analysis

Figure 4 shows a Track 3 case where a temporal feasibility check eliminates the majority of candidates in a single turn. The STTG encodes that the transition from camera c05 to c08 takes 7.6–20.7 seconds; candidates whose trajectories violate this constraint are immediately ruled out. Without this tool, the agent is forced to rely on attribute-based narrowing alone, exhausting its turn budget without convergence. A qualitative comparison for the spatial tool (Track 2) is provided in the supplementary material.

6. Conclusion

We presented ARGOS, the first benchmark for interactive multi-camera person search that integrates spatio-temporal reasoning with agentic dialogue. Our experiments establish that direct LLM inference is fundamentally insufficient: only tool-augmented interaction grounded in the STTG achieves meaningful performance across spatial and temporal tracks. The benchmark remains far from solved, and the backbone crossover between tracks confirms that spatial and temporal reasoning demand distinct capabilities—offering a challenging testbed for future work on multimodal spatial intelligence.

Limitations. The witness simulator is deterministic; real witnesses exhibit memory errors and inconsistencies. ARGOS currently covers two environments; broader validation across diverse layouts is needed. The underlying technology carries inherent surveillance risks; we emphasize that ARGOS is evaluated on fully consented, public datasets [21].

Future work. Promising directions include adversarial witness settings where the agent must detect and correct contradictory or misleading statements, multi-hop STTG reasoning over indirect camera paths (currently only direct edges are used), and transferring the ARGOS evaluation protocol to additional multi-camera datasets with diverse architectural layouts and camera densities.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2, 3

[2] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018. 3

[3] Anthropic. Introducing Claude 4. Online; https://www.anthropic.com/news/claude-4, 2025. 3

[4] Yang Bai, Yucheng Ji, Min Cao, Jinqiao Wang, and Mang Ye. Chat-based person retrieval via dialogue-refined cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3952–3962, 2025. 1

[5] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024. 2

[7] An-Chieh Cheng, Hongxu Yin, Yang Fu, Qiushan Guo, Ruihan Yang, Jan Kautz, Xiaolong Wang, and Sifei Liu. Spatialrgpt: Grounded spatial reasoning in vision-language models. *Advances in Neural Information Processing Systems*, 37:135062–135093, 2024. 1

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 1

[9] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 2

[10] Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. Interactive text-to-image retrieval with large language models: A plug-and-play approach. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 791–809, 2024. 1

[11] Matan Levy, Rami Ben-Ari, Nir Darshan, and Dani Lischinski. Chatting makes perfect: Chat-based image retrieval. *Advances in Neural Information Processing Systems*, 36:61437–61449, 2023. 1

[12] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023. 1

[13] Yiding Lu, Mouxing Yang, Dezhong Peng, Peng Hu, Yijie Lin, and Xi Peng. Llava-reid: Selective multi-image questioner for interactive person re-identification. *arXiv preprint arXiv:2504.10174*, 2025. 1

[14] Christian A. Meissner, Siegfried L. Sporer, and Jonathan W. Schooler. Person descriptions as eyewitness evidence. In *The Handbook of Eyewitness Psychology: Volume II: Memory for People*, pages 3–34. Psychology Press, 2007. 2

[15] Ke Niu, Haiyang Yu, Mengyang Zhao, Teng Fu, Siyang Yi, Wei Lu, Bin Li, Xuelin Qian, and Xiangyang Xue. Chatreid: Open-ended interactive person retrieval via hierarchical progressive tuning for vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24245–24254, 2025. 1

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[17] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025. 3

[18] Wentan Tan, Changxing Ding, Jiayu Jiang, Fei Wang, Yibing Zhan, and Dapeng Tao. Harnessing the power of mllms for transferable text-to-image person reid. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17127–17137, 2024. 1

[19] Rejin 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)*, pages 1–6. IEEE, 2024. 1

[20] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 79–88, 2018. 1

[21] Sanghyun Woo, Kwanyong Park, Inkyu Shin, Myungchul Kim, and In So Kweon. Mtmmc: a large-scale real-world multi-modal camera tracking benchmark. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22335–22346, 2024. 4, 1

[22] Peiran Xu, Sudong Wang, Yao Zhu, Jianing Li, and Yunjian Zhang. Spatialbench: Benchmarking multimodal large language models for spatial cognition. *arXiv preprint arXiv:2511.21471*, 2025. 1

[23] Shuyu Yang, Yinan Zhou, Zhedong Zheng, Yaxiong Wang, Li Zhu, and Yujiao Wu. Towards unified text-based person retrieval: A large-scale multi-attribute and language search benchmark. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4492–4501, 2023. 1

[24] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*, 2022. 1

[25] Mang Ye, Jianbing Shen, Gaojie Lin, Tao Xiang, Ling Shao, and Steven CH Hoi. Deep learning for person re-identification: A survey and outlook. *IEEE transactions on pattern analysis and machine intelligence*, 44(6):2872–2893, 2021. 1

[26] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015. 1