Keeping it Local, Tiny and Real: Automated Report Generation on Edge Computing Devices for Mechatronic-Based Cognitive Systems

1st Nicolas Schuler*† 0009-0007-4098-1244
2nd Lea Dewald* 0009-0004-8825-0545
3rd Jürgen Graf* 0000-0002-1354-0888

* *Department of Computer Science*
Trier University of Applied Sciences
Trier, Germany
{schulern,lxdw0338,j.graf}@hochschule-trier.de

† *Department of Engineering*
University of Luxembourg
Kirchberg, Luxembourg
nicolas.schuler.001@student.uni.lu

***Abstract*—Recent advancements in Deep Learning enable hardware-based cognitive systems, that is, mechatronic systems in general and robotics in particular with integrated Artificial Intelligence, to interact with dynamic and unstructured environments. While the results are impressive, the application of such systems to critical tasks like autonomous driving as well as service and care robotics necessitate the evaluation of large amount of heterogeneous data. Automated report generation for Mobile Robotics can play a crucial role in facilitating the evaluation and acceptance of such systems in various domains. In this paper, we propose a pipeline for generating automated reports in natural language utilizing various multi-modal sensors that solely relies on local models capable of being deployed on edge computing devices, thus preserving the privacy of all actors involved and eliminating the need for external services. In particular, we evaluate our implementation on a diverse dataset spanning multiple domains including indoor, outdoor and urban environments, providing quantitative as well as qualitative evaluation results. Various generated example reports and other supplementary materials are available via a public repository.**

***Index Terms*—Mobile Robotics, Deep Learning, Vision-Language Models, Scene Interpretation, Semantic Similarity**

I. Introduction

Recent developments in Deep Learning, especially Large Language Models (LLMs) and Vision-Language Models (VLMs), allow for mobile cognitive systems to operate in dynamic and unstructured environments, including task planning [1], object detection [2], understanding human behaviour [3] and finally reasoning over the relationships between these [4]. The incorporation of various sensors systems, including camera, LiDAR and INS/GNSS, and manipulators allow these systems to perceive their environment and to interact with it [5]. Such cognitive systems are of particular importance in the field of autonomous driving [6], service robotics and robotic caretakers [7], where the understanding of urban and indoor environments is a complex and critical tasks for human-robot interaction, cooperation and collaboration.

The evaluation, supervision and monitoring of such platforms is challenging, considering the vast range of domains and potential environments they are employed in [8]. The mobile agent might succeed in highly complex unstructured environments of a certain type but completely fail in another [9], with the difference in domain and difficulties associated to it not being obvious for a human observer. Additionally, collective and swarm robotics [10] place much more pressure on supervision, by increasing the number of agents, interactions and monitoring tasks considerably. Finally, the often sensitive nature of the domains, i.e. recording of data subject to privacy laws, and amount of received sensory data necessitate the filtering of data that can be put to human revision. Besides the challenges associated with the amount and type of data, the data interpretation itself might be challenging, requiring specialists to evaluate or make the data available for further computations to begin with.

In this paper, we introduce a pipeline for generating automated reports for mobile agents operating in arbitrary domains that may be utilized for multiple tasks including evaluation and map generation. We do this locally, on an edge computing device and with models capable of real time inference. To that end, we incorporate various Deep Learning architectures that are not fine-tuned to our particular domains to generate textual descriptions of the perceived scenes and actions therein, detect and segment objects and agents of interest to these descriptions, localize these detections within world coordinates, group these actions by similarity and anonymize the output to preserve the privacy of all agents actively and passively involved into the process. This results in automatically generated documents that can be understood by a layman who might not be knowledgeable about the details of the particular Deep Learning architectures utilized. Finally, we achieve this utilizing edge computing devices, ensuring that the data recorded is only processed locally either directly on the edge device or, depending on the requirements on accuracy and computational budgets, on local devices, completely removing the usage of cloud-based solutions and larger, external models that would not be under our direct control and data governance.

Supplementary materials are available via a public repository,[1] including code, high resolution image, video examples, evaluation data and fully generated reports. While not limited to, this work focuses on mechatronic-based cognitive systems acting in real-world scenarios.

The remainder of this paper is structured as follows: Sect. II gives an overview of various models important to our pipeline and related work in automated, Deep Learning supported report generation with a focus on robotics. Sect. III introduces the proposed pipeline and the usage within our laboratory, with the pipeline being evaluated and discussed in Sect. IV. Finally, Sect. V gives a short summary of the findings and future work.

II. Foundations and Related Work

The following section introduces various Deep Learning architectures and models that can be applied to the task of natural language report generation with a focus on so-called foundation models, large pre-trained machine learning models capable of being applied to a large variety of tasks in the field or robotics [11]. First, Sect. II-A gives a short introduction to LLMs and VLMs, which can be utilized for various reasoning and language generation tasks. Following, Sect. II-B focuses on zero-shot object detection and segmentation models, that is models which can be used to detect arbitrary objects without the need for fine-tuning to a specific domain. Finally, Sect. II-C discusses the related work of automated report generation using such foundational models, focusing on their application in autonomous driving and robotics.

A. Vision Language Models

The recent introduction of the transformer-based architectures [12] of LLMs like ChatGPT [13] and VLMs like Gemini Robotics [14] allows mobile platforms to gain a deeper knowledge about their environment than what is possible with pure object detection, leading to more sophisticated interactions with that environment. Beyond the pure text-based interactions with LMMs, VLMs utilize multi-modal data, that is image data in combination with text, and thus can be used to generate descriptions of individual images or even entire sequences and reason over them [15], including implicit knowledge not immediately available within the provided images.

In particular, small Vision-Language Models (sVLMs) enable the usage of such models even on edge computing devices, e.g. Moondream2 [16], FastVLM [17] and SmolVLM2 [18]. While such models display strong results, their restricted size leads to inherent biases: for a recent overview see [19].

B. Zero-Shot Detection and Segmentation Models

The shift towards transformer-based architectures is also present in the classical computer vision domains of object detection and segmentation [20]. While Convolutional Neural Networks (CNNs) dominated these task for the past decade, recently attention-based [12] Vision Transformers (ViTs) [21] or hybrid architectures came to dominate vision-based object detection and segmentation, especially when utilizing multi-modal data [22], to capture long-range temporal and spatial dependencies [23], in such use cases where the general longer inference times and higher computational budget of transformers are of no immediate concern [24] and up to reasoning about the physics observed in the scene [4], [14].

The zero-shot capabilities provided by transformer-based models allow to detect and segment arbitrary object classes without any domain-specific fine-tuning. Open-vocabulary tasks [25] enable the prompting of such models with natural language texts, e.g. for object detection and segmentation. Popular models include SAM [26] and DINOv3 [27] for segmentation tasks, Grounded DINO [2] and OWLv2 [28] for object detection.

C. Related Work

Automated report generation is especially common in medical applications [29], [30]. Closer to our application, [31] apply LLMs and SAM [26] for automated construction inspection reports. Focusing on applications within the field of robotics, [32] combine LLMs and VLMs to serve as a mutual human-computer interface, while [33] utilize VLMs to generate simulated hazardous testing scenarios. For autonomous driving, [34] apply 3D-tokenized LLMs for object detection, map construction and task planning. Concerning road incidents, [35] utilize a combination of local object detection on UAVs, a local VLM (Moondream2 [16]) and the GPT-4 [13] Turbo API for automated incident localization, documentation and response.

To the best of our knowledge, the approach presented in this paper is unique in that we use only small, local Deep Learning models to generate human-readable reports utilizing Semantic Similarity measures to aggregate scene descriptions while preserving full data governance and being completely open to the type of mechatronic-based cognitive system and domain.

III. Automated Report Generation Pipeline

This section introduces the proposed pipeline to automatically generate reports for mobile cognitive systems. Sect. III-A introduces the cognitive systems and sensors we use in our laboratory and Sect. III-B gives an overview of the proposed architecture.

A. Our Mobile Platforms and Sensory Systems

Within our laboratory, we deploy various different mobile platforms including wheelchairs, humanoid, quadruped robots, wheeled vehicles as well as drones, with a selection displayed in Fig. 1. These different platforms are used in domains including indoor and outdoor campus as well as urban environments. For sensory systems, we utilize a flexible combination of multiple video cameras, LiDAR, INS and GNSS that can be adapted to our use cases. For our proposed pipeline, we assume a minimum of one video camera and INS/GNSS to be available.

[1] https://datahub.rz.rptu.de/hstr-csrl-public/publications/automated-report-generation-robotics/

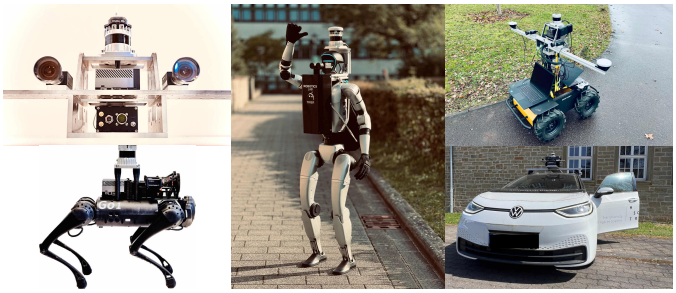

Fig. 1. The different mobile cognitive systems used within our laboratory, utilizing the same sensory system (top left) used for data acquisition in the present paper. Description and images modified from [36].

We want to highlight the versatility of our proposed pipeline within the context of our cognitive systems and domains, that is we make no assumptions about the system the pipeline is used with and the domains it is applied to, giving us full flexibility without any need to fine-tune the used models to a specific task or domain. As the edge device, we utilize either one or multiple NVIDIA Jetson AGX Orin, which host a custom C++ software stack developed within the laboratory to control these systems to handle the recording of all data and inference of Deep Learning applications as well as the pipeline implementation. Additionally, the pipeline itself can be used in a post-processing context on other local machines. For implementation details of the Deep Learning models specifically relevant to the pipeline refer to Sect. IV-A.

B. Automated Report Generation for Mobile Robotics

For the generation of report documents for mobile agents, we focus on several distinct tasks important to the ability to interact with unstructured and unpredictable environments. This includes the detection and segmentation of objects, point cloud data of agents and objects of importance, the semantic description of sequences beyond the obvious and finally the relative and absolute localization of such agents, objects and descriptions within the world.

To manage the large amount of data and make them accessible for human collaboration, our goal is to condense these information into digestible portions. We achieve this by clustering the generated descriptions by semantic similarity [37] and by limiting the segmentation process to entities that are relevant to the described actions. Fig. 2 provides an overview of the proposed architecture.

First, the system generates a natural language description of a sequence of images in real time, using a local sVLM model. Note that we do not use all images within a sequence but sample them in an interval, for example one image per second, to enable real-time inference while retaining accuracy. These descriptions are not only used for report generation, but also during runtime. The generated descriptions and images are then saved for the final report generation. These textual descriptions are then embedded in a common embedding space using transformer models for textual embeddings [39]. The resulting embedding vectors are clustered, forming groups of

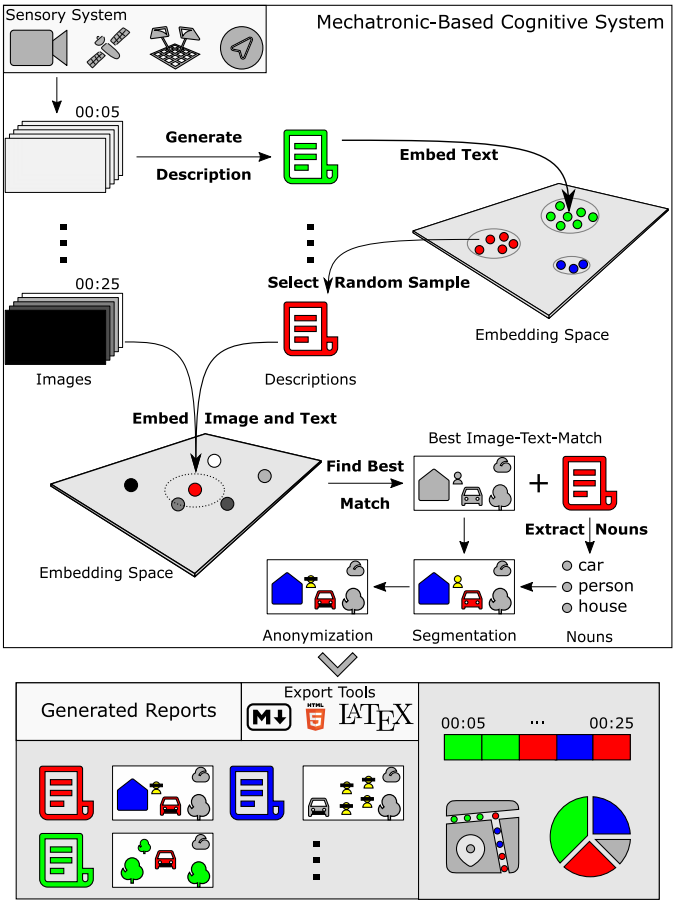

Fig. 2. The proposed pipeline for automated report generation for Mobile Robotics. We utilize a combination of VLM, text and image embedders like CLIP [38], NLP and zero-shot detection and segmentation models to generate reports of multi-modal data recorded by a variety of mobile cognitive systems, agnostic to the application domain.

semantically similar descriptions. Next, a random description from each identified group is selected and embedded with its associated images using CLIP [38] to a shared embedding space to calculate the best match between description and image candidates. We do this to exclude those images from the sequence that are not relevant to the generated description, e.g. in a batch of five images only the last three actual might contain the relevant agent mentioned in the description. Once the best image-text pair is found, we extract all nouns from the generated description using classical Natural Language Processing (NLP) via spaCy [40]. These extracted nouns are used as text prompts for open-vocabulary detection and segmentation. The resulting images are finally anonymized using local Deep Learning models before being used in the report to preserve the privacy of all actively and passively involved agents. This is done for every identified cluster of descriptions, with the individual image-text pairs representing the cluster in the final document.

Finally, we generate a report with the image-text pairs, the generated clusters, all generated descriptions and the localization of these grouped descriptions within the world.

We assume no fixed output data representation format within the pipeline and thus are flexible to generate multiple reports in different formats, include Markdown, HTML and LaTeX, all serving different purposes. Markdown documents might be used within documentations and repositories, also allowing for quick conversion into HTML documents, which contain interactive maps. These then can be easily displayed in web browsers or other applications. LaTeX finally generates highly customizable and structured professional PDF documents.

IV. Experimental Evaluation

During preliminary evaluation, we identified two critical points of failure for the proposed pipeline. First, the correctness and quality of the generated descriptions by the local VLM and second, the results of the clustering algorithm that groups these descriptions by Semantic Similarity. We already investigated the first aspect, that is the correctness of the generated textual descriptions by VLMs running on local, mobile platforms in a previous work [36], and thus will focus on the second aspect for this evaluation. Note that we use the same dataset for both evaluations.

A. Experimental Setup

The proposed pipeline is implemented in Python 3.13 using PyTorch [41], Transformers [42] and Sentence Transformers [43]. We chose Grounded DINO [2] as the zero-shot object detection model, SAM [26] as the zero-shot segmentation model and SmolVLM2 [18] as the sVLM. The communication between the C++ software stack and the Python modules is handled using LitServe [44]. For the purpose of this work, we generate the reports via LaTeX as a PDF file as well as interactive HTML-based maps. For evaluation, we use a dataset consisting of 234 minutes of data, recorded in the German city of Trier by our laboratory. Examples from the dataset are given in Fig. 4. Within the dataset itself, we differentiate between three domains: *Campus Indoor*, *Campus Outdoor* and *City*, with a split of 107 minutes, 74 minutes and 53 minutes respectively. We further split the dataset 20/80, with the first 20 % being used for hyperparameter tuning and 80 % for the final evaluation, keeping the ratio between domains. The individual sequences are split into five second clips that are fed into the VLM, generating textual descriptions of the scene, with the VLM sampling one frame per second to enable real-time inference on the edge device. The generated descriptions are then clustered by human annotators by similarity. We provide the instructions for the annotator in the supplementary materials. We do not evaluate the correctness of the generated descriptions itself and solely focus on the clustering of these descriptions, since the correctness of the descriptions in relation to the real world is of no concern to the quality of the semantic clustering itself. For a detailed study on the correctness of the generated description within our domains see our previous work [36].

For the quantitative evaluation, we compare the results of the automated semantic clustering using Sentence Transformers [43] to the manual clustered ground truth, reporting the Adjusted Rand Index (ARI) [45], Normalized Mutual Information (NMI) [46] and Fowlkes-Mallows Index (FMI) [47]. We select the best hyperparameters per domain for the semantic clustering, using the highest ARI score on the first 20 % of the data and report the discussed metrics on the final 80 % used for evaluation. The hyperparameters optimized are the text embedding model as well as the metric and distance threshold used by the clustering algorithm. We provide the full data of the hyperparameter tuning process in the supplementary materials, see Sect. I. Note that we do not fine-tune any of the Deep Learning models involved for a specific domain and only apply hyperparameter tuning to the clustering algorithm to better align them with our different domains and to evaluate if such a tuning is needed for specific domains to begin with. In addition to the quantitative evaluation, we also present qualitative results, including generated reports and findings that highlight successes and failures of the proposed pipeline.

B. Experimental Results

The quantitative results as a measure of similarity between the ground truth clustering and the result of our pipeline are given in Tab. I. A value closer to 1.0 indicates higher similarity and for ARI and NMI, a value close to 0.0 indicates a similarity close to chance level. The domain with the highest similarity between ground truth and semantic clustering is *Campus Indoor* with an ARI of 0.620, NMI of 0.793 and an FMI of 0.721. The domain *City* show the lowest values in similarity to the ground truth with 0.476, 0.654 and 0.628 respectively.

For qualitative results, we provide example figures from an authentic report generated by the proposed pipeline, see Fig. 3. In addition, we provide multiple full reports in the supplementary material, see Sect. I.

TABLE I
Results of the evaluation by domain.

Domain	ARI	NMI	FMI	Model	Metric	Dist. Thres.
City	0.476	0.654	0.628	all-distilroberta-v1	euclidean	1.5
Campus Outdoor	0.572	0.769	0.651	all-distilroberta-v1	euclidean	1.5
Campus Indoor	0.620	0.793	0.721	all-MiniLM-L6-v2	euclidean	1

C. Discussion

First, the discussion focuses on the quantitative evaluation, that is the measured similarity of the ground truth to the semantic clusters of the textual scene and action descriptions generated by a local VLM. We follow this up by discussing an authentic, generated report from our pipeline in greater detail.

All three domains show values well above 0.0 for all metrics respectively, displaying results of the clustering above the chance level at 0.0 for the ARI and NMI scores. Looking at the individual domains, *City* displays the lowest similarity to the ground truth across all three metrics used, while *Campus Indoor* displays the highest. This is contrary to the findings in one of our previous works [36], where this domain had the highest correctness concerning the generated descriptions, with *Campus Indoor* having the lowest. This fact might be

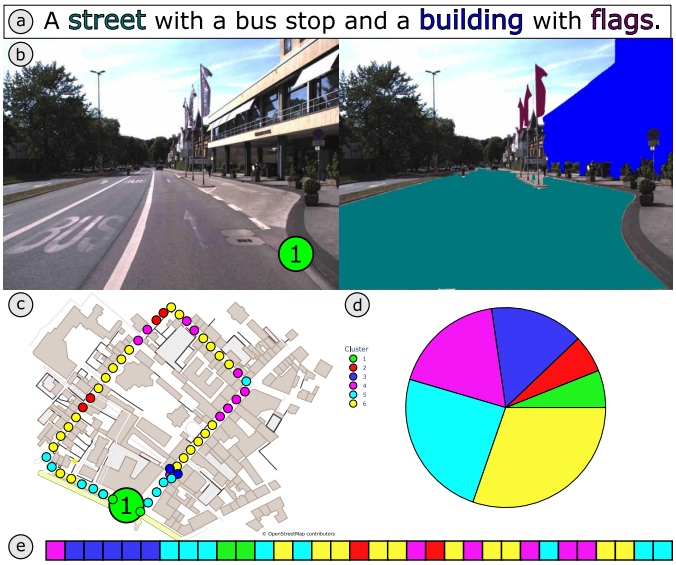

Fig. 3. Examples of an automatically generated report from a car drive through the German city of Trier. a) generated description, b) original image and segmentation results, c) location of individual descriptions colored by cluster, d) relative distribution of clusters and e) timeline of individual descriptions colored by cluster. Map modified from OpenStreetMap [48].

explained by the difference in task at hand. The domain *City* contains fewer radically different actions and scenes due to the relative monotony of urban traffic compared to highly dynamic indoor domains in a campus setting containing kitchens, offices, laboratories et cetera, which differences in descriptions often depend on small, hard to observe details. This facilitates the task of generating correct textual descriptions for the domain *City*, but complicates the clustering of the generated descriptions. In some cases, the entire sequence might contain identical fragments in each description, e.g. 'A street with cars parked on the side and a few pedestrians walking on the sidewalk.', 'A street with cars parked on the side and a few people walking on the sidewalk.' and 'A street with cars parked on the side and a building in the background.' While we would consider the first two description to belong to the same cluster, one can argue that the third description belongs to a different cluster. The semantic clustering algorithm however might struggle to differentiate between this difference, since half of the sentence is identical for all three examples provided. This can be tweaked using hyperparameters, but that might lead to other descriptions not being clustered correctly together anymore.

The results of the hyperparameter optimization show identical optimal parameters for the domains *City* and *Campus Outdoor* and we want to note that for the domain *Campus Indoor*, the same hyperparameter configuration is in the top five combinations as well. This indicates to us that the semantic clustering using transformers is stable enough to use identical tuned hyperparameters for all domains while achieving satisfactory clustering. Since our approach aims to be agnostic to its domain, environment and cognitive systems used, this is a highly desirable property to achieve, especially since manual evaluation using human annotators is labor-intensive and subjective, thus prone to biases [49]. While we aim to reduce this subjectivity by guiding the annotation process (see supplementary materials), this challenge remains.

We want to discuss qualitative results of our evaluation, that is authentic results from the generated reports. As displayed in Fig. 3, the clustering and localization within the report gives us insights into our domain and environment. The magenta cluster describes actions exemplified by 'A cyclist is riding down a city street.' and is clustered in the side streets of this particular drive (to the top of the map), while the turquoise cluster, exemplified by 'A street scene with cars driving down the road.' mainly occurs on the main streets within the city (to the bottom of the map). The green cluster, exemplified by 'A street with a bus stop and a building with flags.', highlights a bus lane and close-by hotel with flags in front of it located at the city centre. Finally, the yellow cluster, exemplified by 'A street with cars and people walking on the sidewalk.', is a more generic cluster which contains descriptions that are frequent both on the side and main streets. These results can not only be used for manual reviews by laymen, but also to augment semantic maps, highlighting spatio-temporal similarities and frequency of descriptions connected to certain streets and locations. The semantically guided zero-shot segmentation process further highlights the area of interest and thus location of objects related to the description. Since the absolute pose w.r.t. various geodetic data and time of the mobile cognitive system is known, this allows us to exactly localize these objects of interest within the world of the mobile system. In addition, we are able to differentiate between active agents, e.g. humans and vehicles, and passive objects we assume to not move, e.g. buildings, enabling us to build dynamic semantic maps comprised of description and object heatmaps.

V. Conclusion and Future Work

In this work, we presented a privacy-preserving, local pipeline for automated report generation for mechatronic-based cognitive systems. We demonstrated that our system is able to generate semantic clusters similar to a human annotator, with the resulting report offering valuable insights into the environment of the mobile cognitive system.

For future work, we aim to add more sensor information provided by our platform specific to the task at hand, in particular by incorporating AI responsible for decision-making on our mobile cognitive system. Further, we want to expand the internal data representation to incorporate scene graphs and additional semantic information, especially to increase the explainability and internal consistency within our pipeline, expanding on the topic of automated semantic map generation and world models [50]. We also want to utilize the aspect of collective robotics in building shared, detailed maps, leveraging consistent internal world representations.

Appendix

Fig. 4 gives selected examples of the dataset used for evaluation. The dataset was recorded in the German city of

Trier, containing three domains: *City*, *Campus Outdoor* and *Campus Indoor*. Note that we anonymize the images only for the purpose of this publication, the models and mobile cognitive systems utilize the raw data instead.

Fig. 4. Examples from the dataset used in the evaluation of this paper. The dataset contains various domains, challenging lighting conditions, crowded scenes and complex scenarios. The dataset was captured by different mobile cognitive systems within our laboratory. Images taken from [36].

REFERENCES

[1] Annie S. Chen et al., "Commonsense reasoning for legged robot adaptation with vision-language models," 2024, arXiv:2407.02666.

[2] Shilong Liu et al., "Grounding dino: Marrying dino with grounded pre-training for open-set object detection," 2023, arXiv:2303.05499.

[3] M. Karim, S. Khalid, S. Lee, S. Almutairi, A. Namoun, and M. Abohashrh, "Next generation human action recognition: A comprehensive review of state-of-the-art signal processing techniques," *IEEE Access*, vol. 13, pp. 135 609–135 633, 2025.

[4] A. Cherian, R. Corcodel, S. Jain, and D. Romeres, "Llmphy: Complex physical reasoning using large language models and world models," 2024, arXiv:2411.08027.

[5] M. Beetz, G. Kazhoyan, and D. Vernon, "Robot manipulation in everyday activities with the cram 2.0 cognitive architecture and generalized action plans," *Cognitive Systems Research*, vol. 92, p. 101375, 2025.

[6] J. Zhao, Y. Wu, R. Deng, S. Xu, J. Gao, and A. Burke, "A survey of autonomous driving from a deep learning perspective," *ACM Computing Surveys*, vol. 57, no. 10, pp. 1–60, 2025.

[7] D. Silvera-Tawil, "Robotics in healthcare: A survey," *SN Computer Science*, vol. 5, no. 1, 2024.

[8] C. Tang, B. Abbatematteo, J. Hu, R. Chandra, R. Martín-Martín, and P. Stone, "Deep reinforcement learning for robotics: A survey of real-world successes," 2024.

[9] Jiaqi Wang et al., "Large language models for robotics: Opportunities, challenges, and perspectives," *Journal of Automation and Intelligence*, vol. 4, no. 1, pp. 52–64, 2025.

[10] Z. Feng, G. Hu, Y. Sun, and J. Soon, "An overview of collaborative robotic manipulation in multi-robot systems," *Annual Reviews in Control*, vol. 49, pp. 113–127, 2020.

[11] Xuan Xiao et al., "Robot learning in the era of foundation models: a survey," *Neurocomputing*, vol. 638, p. 129963, 2025.

[12] Ashish Vaswani et al., "Attention is all you need," 2017, arXiv:1706.03762.

[13] OpenAI et al., "Gpt-4 technical report," 2023, arXiv:2303.08774.

[14] Gemini Robotics Team et al., "Gemini robotics: Bringing ai into the physical world," 2025, arXiv:2503.20020.

[15] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, "From recognition to cognition: Visual commonsense reasoning," 2018, arXiv:1811.1083.

[16] M87 Labs, "moondream2," [Online]. Available: https://moondream.ai/

[17] Pavan Kumar Anasosalu Vasu et al., "Fastvlm: Efficient vision encoding for vision language models," 2024, arXiv:2412.13303.

[18] Andrés Marafioti et al., "Smolvlm: Redefining small and efficient multimodal models," 2025, arXiv:2504.05299.

[19] Nitesh Patnaik et al., "Small vision-language models: A survey on compact architectures and techniques," 2025, arXiv:2503.10665.

[20] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[21] Alexey Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.

[22] K. Carolan, L. Fennelly, and A. F. Smeaton, "A review of multi-modal large language and vision models," 2024, arXiv:2404.01322.

[23] Z. Wang and Y. Liu, "Staa: Spatio-temporal attention attribution for real-time interpreting transformer-based video models," 2024, arXiv:2411.00630.

[24] K. T. Chitty-Venkata, S. Mittal, M. Emani, V. Vishwanath, and A. K. Somani, "A survey of techniques for optimizing transformer inference," *Journal of Systems Architecture*, vol. 144, p. 102990, 2023.

[25] X. Yu, Y. Wang, and H. Sun, "Open-vocabulary object detection survey," *Procedia Computer Science*, vol. 266, pp. 62–71, 2025.

[26] Alexander Kirillov et al., "Segment anything," 2023, arXiv:2304.02643.

[27] Oriane Siméoni et al., "Dinov3," 2025, arXiv:2508.10104.

[28] M. Minderer, A. Gritsenko, and N. Houlsby, "Scaling open-vocabulary object detection," 2023, arXiv:2306.09683.

[29] J. Ji, Y. Hou, X. Chen, Y. Pan, and Y. Xiang, "Vision-language model for generating textual descriptions from clinical images: Model development and validation study," *JMIR Formative Research*, vol. 8, p. e32690, 2024.

[30] Vishwanatha M. Rao et al., "Multimodal generative ai for medical image interpretation," *Nature*, vol. 639, no. 8056, pp. 888–896, 2025.

[31] H. Pu, X. Yang, Z. Shi, and N. Jin, "Automated inspection report generation using multimodal large language models and set-of-mark prompting," in *Proceedings of the 41st International Symposium on Automation and Robotics in Construction*, ser. ISARC2024. International Association for Automation and Robotics in Construction (IAARC), 2024.

[32] A. N. Abbas and C. Beleznai, "Talkwithmachines: Enhancing human-robot interaction for interpretable industrial robotics through large/vision language models," pp. 253–258, 2024.

[33] J. Wu, C. Lu, A. Arrieta, S. Ali, and T. Peyrucain, "Vision language model-based testing of industrial autonomous mobile robots," 2025, arXiv:2508.02338.

[34] Yifan Bai et al., "Is a 3d-tokenized llm the key to reliable autonomous driving?" May 2024, arXiv:2405.18361.

[35] M. Farhan, H. Eesaar, A. Ahmed, K. T. Chong, and H. Tayara, "Transforming highway safety with autonomous drones and ai: A framework for incident detection and emergency response," *IEEE Open Journal of Vehicular Technology*, vol. 6, pp. 829–845, 2025.

[36] N. Schuler, L. Dewald, N. Baldig, and J. Graf, "From the laboratory to real-world application: evaluating zero-shot scene interpretation on edge devices for mobile robotics," *in press*.

[37] P. Yang, H. Wang, J. Yang, Z. Qian, Y. Zhang, and X. Lin, "Deep learning approaches for similarity computation: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 36, no. 12, pp. 7893–7912, 2024.

[38] Alec Radford et al., "Learning transferable visual models from natural language supervision," 2021, arXiv:2103.00020.

[39] H. Cao, "Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark," 2024, arXiv:2406.01607.

[40] Explosion AI, "spacy." [Online]. Available: https://spacy.io/

[41] Adam Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," 2019, arXiv:1912.01703.

[42] Thomas Wolf et al., "Huggingface's transformers: State-of-the-art natural language processing," 2019, arXiv:1910.03771.

[43] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019, arXiv:1908.10084.

[44] Lightning AI, "Litserve." [Online]. Available: https://lightning.ai/docs/litserve

[45] D. Steinley, "Properties of the hubert-arable adjusted rand index." *Psychological Methods*, vol. 9, no. 3, pp. 386–396, 2004.

[46] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[47] E. B. Fowlkes and C. L. Mallows, "A method for comparing two hierarchical clusterings," *Journal of the American Statistical Association*, vol. 78, no. 383, pp. 553–569, 1983.

[48] OSMF, "Openstreetmap." [Online]. Available: https://www.openstreetmap.org/

[49] S. Gautam and M. Srinath, “Blind spots and biases: Exploring the role of annotator cognitive biases in nlp,” 2024, arXiv:2404.19071.
[50] Y. Guan, H. Liao, Z. Li, J. Hu, R. Yuan, Y. Li, G. Zhang, and C. Xu, “World models for autonomous driving: An initial survey,” 2024, arXiv:2403.02622.