More Views, More Problems? A Critical Analysis of Multi-View Aggregation for Agent Perception

Extended Abstract

Federico Tavella
University of Manchester
Manchester, United Kingdom
federico.tavella@manchester.ac.uk

Amber Drinkwater
Amentum Clean Energy Ltd
Warrington, United Kingdom
amber.drinkwater@global.amentum.com

Angelo Cangelosi
University of Manchester
Manchester, United Kingdom
angelo.cangelosi@manchester.ac.uk

ABSTRACT

Multi-view perception is expected to enhance autonomous agent understanding, yet effective information integration across views remains unresolved. This work examines an LLM-based synthesis strategy where VLM-generated captions from multiple viewpoints are aggregated into a unified description. We evaluate this approach against single-view, naive average, and oracle baselines using real-world and domain-shifted 3D-printed datasets. Results demonstrate that synthesis successfully combines complementary information when per-view captions are reliable; however, under domain shift, the strategy degrades substantially, often underperforming the static baseline. These findings highlight the brittleness of current aggregation methods and the critical need for robust information-fusion mechanisms in robotic perception.

KEYWORDS

Autonomous Robots; Active Perception; VLM; Multi-view Synthesis

ACM Reference Format:
Federico Tavella, Amber Drinkwater, and Angelo Cangelosi. 2026. More Views, More Problems? A Critical Analysis of Multi-View Aggregation for Agent Perception: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/CERC5181

1 INTRODUCTION

Understanding and describing visual scenes is a fundamental capability for autonomous agents, particularly in Human-Robot Collaboration scenarios where shared understanding is essential [2, 7, 11, 32]. The emergence of Vision-Language Models (VLMs) offers a promising pathway for equipping robots with this ability [3, 25], allowing them to ground high-level instructions and translate visual data into human-understandable descriptions and physical actions [1, 12, 21, 34].

However, a critical gap exists between the static benchmarks on which these models are trained and the dynamic reality of an embodied agent [13, 18, 28]. In the physical world, single viewpoints are often insufficient due to occlusion, ambiguous angles, or poor lighting [8, 14, 17]. While recent robotic approaches have successfully integrated language models for manipulation [1, 30], they typically assume a single fixed camera view. Conversely, multi-view approaches (common in 3D reconstruction and classification [16, 20, 27]) have rarely been coupled with language output [10, 19], leaving open questions about how viewpoint aggregation affects semantic grounding.

In this paper, we reframe object description as an active, agent-driven process. We systematically evaluate distinct perceptual strategies representing a gradient of agent complexity: relying on a fixed canonical view; actively seeking an optimal 'consensus' viewpoint; and intelligently synthesizing information from all views into a holistic description using a Large Language Model (LLM). We stress-test these strategies on real-world tools and domain-shifted 3D-printed objects to answer three research questions: (1) To what extent can active perception improve accuracy over a fixed perspective? (2) Can intelligent synthesis outperform the single most informative view? (3) How do these strategies affect robustness against domain shift?

2 METHODOLOGY

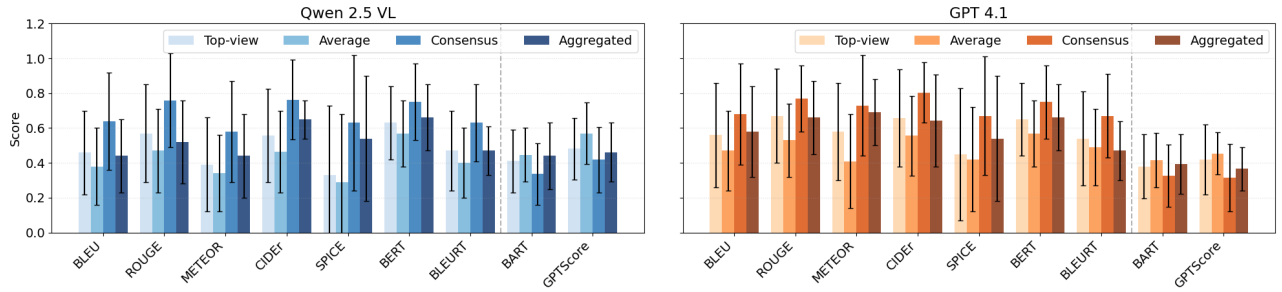
Our approach systematically evaluates how different agent perceptual strategies affect object description accuracy. We utilize a Franka Emika Research 3 robot with an eye-in-hand Intel RealSense D435i camera to collect data. The dataset O ($|O| = N$) is partitioned into two subsets: O_{real} (real-world tools) and O_{3D} (3D-printed replicas), allowing us to test robustness against domain shift. For each object o, we capture two sets of visual data: a single static top-down view v_{top}, and a set of M views V_o acquired from a hemispherical trajectory [10], segmented using SAM2 [24].

2.1 Evaluation Strategies

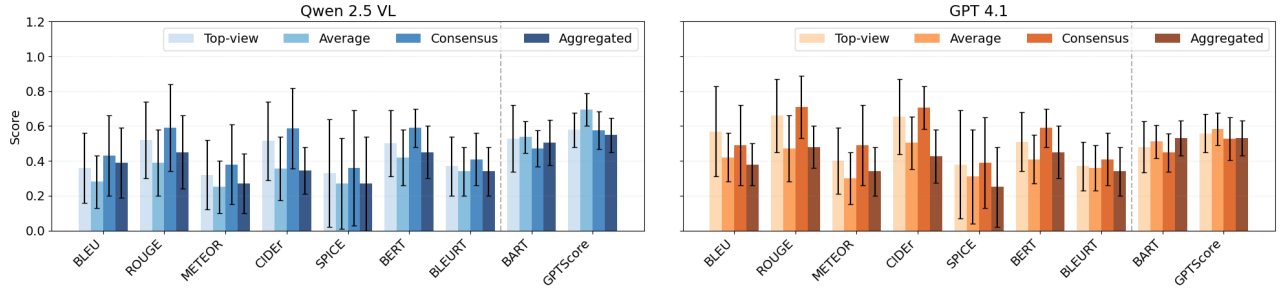
We define $\mathcal{M}_{VLM}$ as the Vision-Language Model acting as the agent's perception module, and $\mathcal{M}_{LLM}$ as the Large Language Model used for aggregation. The quality of a generated caption c is measured against a ground truth c_o^* using a scoring metric $S(c, c_o^*)$. We evaluate four strategies:

(1) **Static Canonical View (S_{top}):** Models a static agent using only the fixed top-down view $S_{\text{top}}(o) = S(\mathcal{M}_{VLM}(v_{top}), c_o^*)$.
(2) **Best Consensus View (S_{best}):** Models an active agent seeking the optimal viewpoint. To avoid metric-specific bias, we identify a consensus caption $c_{\text{consensus}}$ by ranking all $M-1$ active views across all evaluation metrics and selecting the one with the best aggregate rank $S_{\text{best}}(o) = S(c_{\text{consensus}}(o), c_o^*)$.
(3) **Average Single-View (S_{avg}):** Quantifies the naive use of multi-view data by averaging performance across all captured views $v_i \in V_o$. $S_{\text{avg}}(o) = \frac{1}{M} \sum_{i=0}^{M-1} S(\mathcal{M}_{VLM}(v_i), c_o^*)$.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/CERC5181



(a) Real objects



(b) 3D objects

Figure 1: Comparison of Top-view, Average, Consensus, and Aggregated strategies. For visualization purposes, CIDEr has been rescaled by 10, and BARTScore/GPTScore have been rescaled by -10 (where lower is better).

(4) **Multi-View Synthesis (S_{agg}):** Our proposed strategy where an LLM synthesizes a holistic description from the set of all per-view captions $C_o = \{\mathcal{M}_{VLM}(v_i)\}$

$$S_{\text{agg}}(o) = S(\mathcal{M}_{LLM}(C_o), c_o^*).$$

To assess overall effectiveness, we compute dataset-level metrics by averaging per-object scores: $S_{\text{y}}^{\text{x}} = \frac{1}{|O_{\text{x}}|} \sum_{o \in O_{\text{x}}} S_{\text{y}}(o)$, where $x \in \{\text{real}, \text{3D}\}$ and y is the strategy.

2.2 Implementation

We evaluated a range of state-of-the-art VLMs, including Qwen 2.5 VL [5] and GPT-4.1 [22], prompted to provide concise physical descriptions. For the synthesis strategy (S_{agg}), GPT-4.1 acts as the aggregator, instructed to resolve conflicts and select the most precise details from the noisy candidate set C_o. Performance is measured using a comprehensive suite of metrics: lexical (BLEU [23], ROUGE [15], METEOR [6], CIDEr [29]), scene-graph based (SPICE [4]), and embedding-based (BERTScore [33], BLEURT [26], BARTScore [31], GPTScore [9]) to capture both surface-level and semantic accuracy.

3 EXPERIMENTS & RESULTS

Our first experiment quantified the robustness of VLMs to physical domain shift by comparing performance on real-world tools (O_{real}) versus 3D-printed replicas (O_{3D}). We defined the performance gap as $\Delta S_{\text{top}} = S_{\text{top}}^{\text{real}} - S_{\text{top}}^{\text{3D}}$. Our evaluation revealed that performance degrades consistently across all models when moving to the 3D domain, primarily due to the loss of texture and material cues. Lexical metrics (BLEU, ROUGE) often failed to capture semantic accuracy, punishing valid paraphrases, whereas semantic metrics (SPICE, BERTScore) proved more robust. Based on these results, Qwen 2.5 VL [5] and GPT-4.1 [22] were identified as the top-performing open and closed-source models respectively and were selected for the subsequent multi-view experiments.

We evaluated whether active perception improves description quality by comparing the Consensus strategy (S_{best}), Naive Averaging (S_{avg}), and LLM Synthesis (S_{agg}) against the Top-view baseline. Analysis of the consensus strategy showed that actively seeking the optimal viewpoint (S_{best}) yields consistent improvements. For Qwen 2.5 VL on real objects, scores strictly improved for up to 70% of the dataset compared to the static top view. For instance, the consensus view allowed the model to correctly identify "metal tweezers" (SPICE=1.0) where the top view saw only generic "metal tongs" (SPICE=0.0).

In contrast, Naive Averaging (S_{avg}) proved detrimental. As illustrated in Figure 1, averaging indiscriminately fuses noise from suboptimal angles, often resulting in scores lower than the single top-view baseline. The LLM-based synthesis strategy (S_{agg}) generally outperformed naive averaging but failed to match the oracle-like Consensus baseline. While aggregation can refine details (such as correcting "red and yellow pliers" to "insulated combination pliers") it is prone to errors, especially under domain shift. In the 3D dataset, GPT-4.1 aggregated noisy captions into hallucinations like "Black three-blade boat propeller" for a wing nut, causing significant metric drops. This indicates that while synthesis stabilizes output, it currently lacks the grounding necessary to reject consistent but incorrect visual evidence in ambiguous scenarios.

ACKNOWLEDGMENTS

This work was supported by the Centre for Robotic Autonomy in Demanding and Long-lasting Environments (CRADLE), UKRI Grant EP/X02489X/1.

REFERENCES

[1] Michael Ahn et al. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. *arXiv* (2022). arXiv:2204.01691 [cs.RO] https://arxiv.org/abs/2204.01691

[2] Arash Ajoudani, Andrea Maria Zanchettin, Serena Ivaldi, Alin Albu-Schäffer, Kazuhiro Kosuge, and Oussama Khatib. 2018. Progress and prospects of the human–robot collaboration. *Autonomous robots* 42, 5 (2018), 957–975.

[3] Jean-Baptiste Alayrac et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems* (2022).

[4] Peter Anderson et al. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*.

[5] Shuai Bai et al. 2025. Qwen2.5-VL Technical Report. *arXiv* (2025). arXiv:2502.13923 [cs.CV] https://arxiv.org/abs/2502.13923

[6] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*.

[7] Angelo Cangelosi, Manith Adikari, Rahul Singh Maharjan, Luca Raggioli, Francesco Semeraro, Radu Stoican, Federico Tavella, Hongbo Zhu, and Marta Romeo. 2025. Human-Centred Robotics and AI for Trustworthy Human-Robot Interaction. In *Technology as Cultural Mediator: Theories and Experiences from Different Contexts*. Springer, 191–222.

[8] Shivam Chandhok and Pranav Tandon. 2024. Do Vision-Language Foundational models show Robust Visual Perception? *arXiv* (2024). arXiv:2408.06781 [cs.CV] https://arxiv.org/abs/2408.06781

[9] Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as You Desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

[10] Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. 2024. Dream2Real: Zero-Shot 3D Object Rearrangement with Vision-Language Models. In *IEEE International Conference on Robotics and Automation (ICRA)*.

[11] Ali Keshvarparast, Daria Battini, Olga Battaia, and Amir Pirayesh. 2024. Collaborative robots in manufacturing and assembly systems: literature review and future research agenda. *Journal of Intelligent Manufacturing* 35, 5 (2024), 2065–2118.

[12] Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan P Foster, Pannag R Sanketi, Quan Vuong, et al. 2025. OpenVLA: An Open-Source Vision-Language-Action Model. In *Conference on Robot Learning*. PMLR, 2679–2713.

[13] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems* 37 (2024), 100428–100534.

[14] Yifan Li et al. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. https://doi.org/10.18653/v1/2023.emnlp-main.20

[15] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*.

[16] Huei-Yung Lin, Shih-Cheng Liang, and Yu-Kai Chen. 2021. Robotic Grasping With Multi-View Image Acquisition and Model-Based Pose Estimation. *IEEE Sensors Journal* (2021).

[17] Hanqing Liu et al. 2025. When Lighting Deceives: Exposing Vision-Language Models' Illumination Vulnerability Through Illumination Transformation Attack. *arXiv* (2025). arXiv:2503.06903 [cs.CV] https://arxiv.org/abs/2503.06903

[18] Yang Liu, Weixing Chen, Yongjie Bai, Xiaodan Liang, Guanbin Li, Wen Gao, and Liang Lin. 2025. Aligning Cyber Space with Physical World: A Comprehensive Survey on Embodied AI. *IEEE/ASME Transactions on Mechatronics* (2025).

[19] Tiange Luo, Chris Rockwell, Honglak Lee, and Justin Johnson. 2023. Scalable 3D captioning with pretrained models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NIPS '23)*.

[20] Ben Mildenhall, , et al. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* (2021).

[21] NVIDIA et al. 2025. GR00T N1: An Open Foundation Model for Generalist Humanoid Robots. arXiv:2503.14734 [cs.RO] https://arxiv.org/abs/2503.14734

[22] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[23] Kishore Papineni et al. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics.

[24] Nikhila Ravi et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714* (2024).

[25] Scott Reed, Konrad Zolna, Emilio Parisotto, Sergio Gomez Colmenarejo, Alexander Novikov, Gabriel Barth-Maron, Mai Gimenez, Yury Sulsky, Jackie Kay, Jost Tobias Springenberg, et al. 2022. A generalist agent. *arXiv preprint arXiv:2205.06175* (2022).

[26] Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. BLEURT: Learning Robust Metrics for Text Generation. In *Proceedings of ACL*.

[27] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. 2015. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE international conference on computer vision*. 945–953.

[28] Andrew Szot, Bogdan Mazoure, Omar Attia, Aleksei Timofeev, Harsh Agrawal, Devon Hjelm, Zhe Gan, Zsolt Kira, and Alexander Toshev. 2025. From Multimodal LLMs to Generalist Embodied Agents: Methods and Lessons. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10644–10655.

[29] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-Based Image Description Evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[30] Jianing Yang et al. 2024. Llm-grounder: Open-vocabulary 3d visual grounding with large language model as an agent. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE.

[31] Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating Generated Text as Text Generation. In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (Eds.), Vol. 34. Curran Associates, Inc., 27263–27277. https://proceedings.neurips.cc/paper/2021/file/e4d2b6e6fdeca3e60e0f1a62fee3d9dd-Paper.pdf

[32] Muhammad Hamza Zafar, Even Falkenberg Langås, and Filippo Sanfilippo. 2024. Exploring the synergies between collaborative robotics, digital twins, augmentation, and industry 5.0 for smart manufacturing: A state-of-the-art review. *Robotics and Computer-Integrated Manufacturing* 89 (2024), 102769. https://doi.org/10.1016/j.rcim.2024.102769

[33] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SkeHuCVFDr

[34] Brianna Zitkovich et al. 2023. RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control. In *Proceedings of The 7th Conference on Robot Learning (Proceedings of Machine Learning Research, Vol. 229)*, Jie Tan, Marc Toussaint, and Kourosh Darvish (Eds.). PMLR, 2165–2183. https://proceedings.mlr.press/v229/zitkovich23a.html