Reference-Guided Robotic Photography Through Natural Language Interaction

Oliver Limoyo*
University of Toronto

Jimmy Li*
Samsung AI Centre Montreal

Dmitriy Rivkhin
Samsung AI Centre Montreal

Jonathan Kelly
University of Toronto

Gregory Dudek
McGill University, Samsung AI Centre Montreal

ABSTRACT

We introduce PhotoBot, a framework for automated photo acquisition based on an interplay between high-level human language guidance and a robot photographer. We propose to communicate photography suggestions to the user via a reference picture that is retrieved from a curated gallery. We leverage a visual language model (VLM) and an object detector to characterize reference pictures via textual descriptions and use a large language model (LLM) to retrieve relevant reference pictures based on a user's language query through text-based reasoning. To correspond the reference picture and the observed scene, we exploit pre-trained features from a vision transformer capable of capturing semantic similarity across widely varied images. Using these features, we compute pose adjustments for an RGB-D camera by solving a Perspective-n-Point (PnP) problem, enabling the fully automatic capture of well-composed and compelling photographs.

CCS CONCEPTS

• **Computer systems organization** → **Robotic autonomy**.

KEYWORDS

Robot Photography, Visual Language Models, Computer Vision, Physical Human-robot Interaction

ACM Reference Format:
Oliver Limoyo, Jimmy Li, Dmitriy Rivkhin, Jonathan Kelly, and Gregory Dudek. 2024. Reference-Guided Robotic Photography Through Natural Language Interaction. In *Proceedings of Human - Large Language Model Interaction Workshop (HRI'24).* ACM, New York, NY, USA, 3 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Photographing a human subject requires nuanced interaction and clear communication between the photographer and the model.

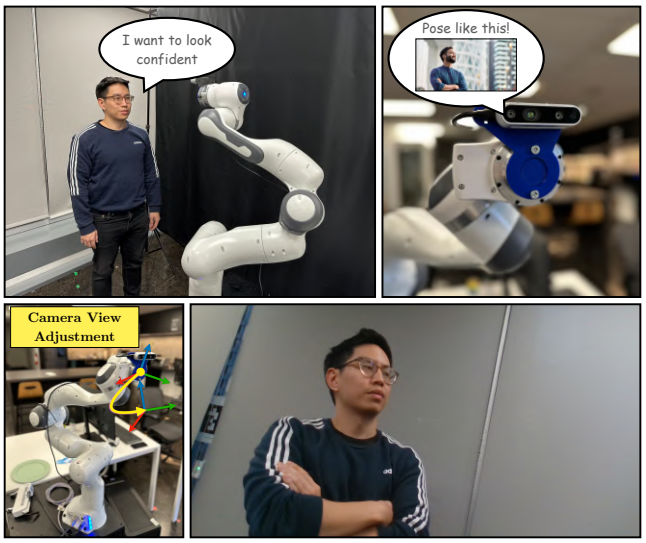

Figure 1: PhotoBot provides a reference photograph suggestion based on an observation of the scene and a user's input language query. The user strikes a pose matching that of the person in the reference photo and PhotoBot adjusts its camera accordingly to faithfully capture the layout and composition of the reference image. The lower-right panel shows an unretouched photograph produced by PhotoBot.

Beyond just capturing well-composed pictures, a professional photographer needs to understand what the client wants and to provide suggestions. Prior research in the area of robotic photography [2–4, 9, 11] has focused on the technical aspects, that is, how to navigate, plan, and control a robot to frame a picture, but not on the interaction between the photographer and model.

In this work, we introduce PhotoBot, a framework for automated photo acquisition based on an interplay between high-level human guidance and a robot photographer. PhotoBot is capable of interacting with the user by leveraging the reasoning capabilities of large language models (LLMs) [10] and the grounding capabilities of visual language models (VLMs) together. Figure 1 shows an example of a problem instance and an image produced fully automatically by PhotoBot.

2 METHOD

We first convert a curated gallery of images into text-based descriptions (e.g., including information such as a general description of the image, the mood, and the number of people in the image) using

*Both authors contributed equally to this research.

HRI'24, March 11–15, 2024, Boulder, Colorado

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/XXXXXXX.XXXXXXX

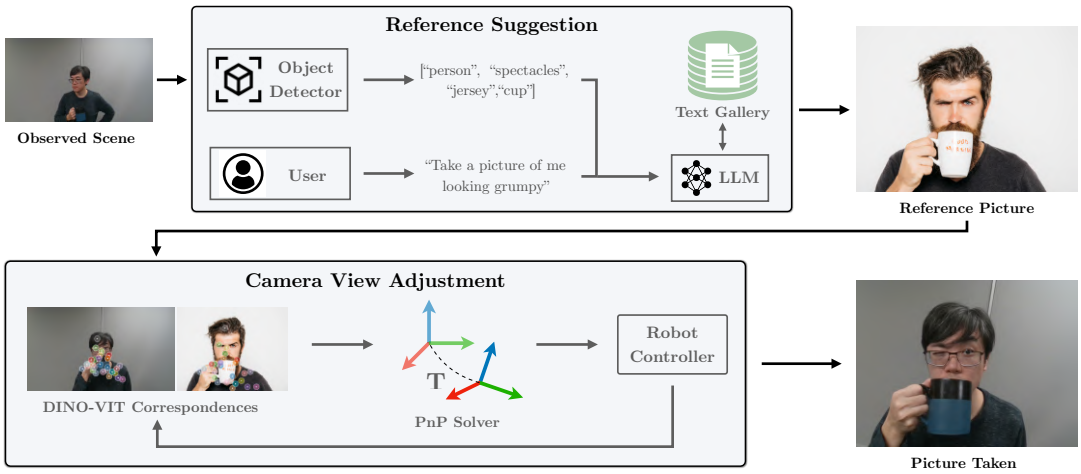

Figure 2: PhotoBot system diagram. The two main modules are shown: Reference Suggestion and Camera View Adjustment. Given the observed scene and a user query, PhotoBot suggests a reference picture to the user and adjusts the camera to take a picture with a similar layout and composition to the reference image.

a VLM and an object detector. The VLM and object detector provide an automated approach to describe curated images using language. Given a language query from a user and the detected objects in the scene observed by the camera, we use an LLM to retrieve a relevant *reference picture* (i.e., an existing image from the curated gallery of high-quality photographs) to suggest to the user through text-based reasoning (much like a professional photographer would do). We use InstructBLIP [6] for the VLM, Detic [12] for the object detector and GPT-4 [8] for the LLM. The user then imitates what is shown in the image and PhotoBot solves for the respective camera motion and image crop such that the camera view matches the reference picture. We formulate camera view adjustment as a Perspective-n-Point (PnP) problem [7] with pre-trained features from a vision transformer capable of capturing semantic similarity across significantly varying images [1, 5].

3 RESULTS

We evaluated the PhotoBot framework using a real Franka Emika robot manipulator equipped with a RealSense D435 RGB-D camera. To assess the reference-based photography approach used in this work, we had a group of users ($N = 8$) interact with and have their pictures taken by PhotoBot. We prompted users to query PhotoBot for picture suggestions from three general categories of emotions: *Confident*, *Happy*, and *Surprised*. Based on each query, three reference photograph suggestions were provided, from which the user chose one. Users then posed in a manner similar to the reference image (to the best of their ability) and PhotoBot took their pictures. We used two baselines, which we named "No PhotoBot" and "Reference Suggestion Only." For "No Photobot," we asked the same set of users to come up with a gesture and expression that matched the category of emotion and to pose for a photo in front of a static camera. For "Reference Suggestion Only," where we asked users to try their best to position themselves and pose in front of the static camera in a manner that re-created the reference picture as closely as possible. Examples of pictures taken by PhotoBot and the respective baselines are shown in Figure 3.

Figure 3: Photos of users evoking various emotions. The user prompts, from top to bottom, are *surprised*, *confident*, *guilty*, *confident*, *happy*, and *confident*. Columns, from left to right, are: user's own creative posing; user mimicking the suggested reference using a static camera; photo taken by our PhotoBot system; and reference picture suggested by PhotoBot. The checkered background indicates cropping. The black background indicates padding of the reference image to facilitate the PnP solution. PhotoBot automatically crops the pictures it takes to match the picture template.

4 CONCLUSION

We briefly described PhotoBot, a novel interactive robot photography system. PhotoBot is capable of suggesting reference pictures based on a natural language query from a user and visual observation of the current scene. In addition, PhotoBot can adjust the camera to match the layout and composition of a chosen reference picture. Incorporating a LLM and VLM into our robot photographer imbues it with the ability to interact with human users — a crucial part of a professional photographer's job. For future work, we aim to develop an even more interactive robot photographer by investigating methods to provide language-based corrective posing feedback to the human user.

ACKNOWLEDGMENTS

All reference images are used under license from Shutterstock.com. Individuals in the photographs gave consent for their images to appear in this paper.

REFERENCES

[1] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. 2021. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814* 2, 3 (2021), 4.

[2] Z. Byers, M. Dixon, K. Goodier, C.M. Grimm, and W.D. Smart. 2003. An autonomous robot photographer. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003) (Cat. No.03CH37453)*, Vol. 3. 2636–2641 vol.3. https://doi.org/10.1109/IROS.2003.1249268

[3] Zachary Byers, Michael Dixon, William Smart, and Cindy Grimm. 2003. Say Cheese!: Experiences with a Robot Photographer. *Ai Magazine - AIM* 25, 65–70.

[4] J. Campbell and P. Pillai. 2005. Leveraging Limited Autonomous Mobility to Frame Attractive Group Photos. In *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*. 3396–3401. https://doi.org/10.1109/ROBOT.2005.1570635

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 9650–9660.

[6] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500 [cs.CV]

[7] Martin A. Fischler and Robert C. Bolles. 1981. Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Commun. ACM* 24, 6 (jun 1981), 381Ã¢ÂĂÀŞ395. https://doi.org/10.1145/358669.358692

[8] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[9] Dmitriy Rivkin, Gregory Dudek, Nikhil Kakodkar, David Meger, Oliver Limoyo, Michael Jenkin, Xue Liu, and Francois Hogan. 2023. ANSEL Photobot: A Robot Event Photographer with Semantic Intelligence. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. 8262–8268. https://doi.org/10.1109/ICRA48891.2023.10161403

[10] Alon Talmor, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant. 2021. CommonsenseQA 2.0: Exposing the Limits of AI through Gamification. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[11] Manfredas Zabarauskas and Stephen Cameron. 2014. Luke: An autonomous robot photographer. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*. 1809–1815. https://doi.org/10.1109/ICRA.2014.6907096

[12] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. 2022. Detecting Twenty-thousand Classes using Image-level Supervision. In *ECCV*.