

FotoBot: an embodied AI photography robot system its design, prototyping and application



Dawei Wang¹, Chang Chen¹, Yipeng Pan¹, Xinzheng Tang², Hua Chen³ and Jia Pan^{1,*}

¹ School of Computing and Data Science, The University of Hong Kong, Hong Kong, China

² Department of Civil Engineering, The University of Hong Kong, Hong Kong, China

³ UIUC Institute, Zhejiang University, Hangzhou, China

* Correspondence author; E-mail: jpan@cs.hku.hk.

Highlights:

- **Novel Integration:** We present the first vision-driven, user-friendly, all-terrain robotic photography system designed for real-world deployment.
- **BTS Viewpoint Planning:** We propose a compact 3-DoF Bipedal Toric Space for vision-guided and cinematographically meaningful camera viewpoint control.
- **Real-World Performance:** The proposed system outperforms baseline methods in both user satisfaction and execution efficiency through extensive real-world deployments and user studies.

Abstract: This paper introduces FotoBot, a vision-driven autonomous robot photographer designed to enhance human–robot interaction (HRI) and optimize camera parameter control through real-time visual perception. FotoBot integrates Generative Pre-trained Transformers (GPT) for seamless natural language communication, and Bipedal Toric Space (BTS) for vision-guided camera viewpoint control. Utilizing GPT, FotoBot effectively interprets and responds to user instructions, enabling intelligent behavior adjustments. BTS is introduced in this paper for camera position planning, which compresses the camera position representation into three parameters related to photo composition. The BTS representation is analytically converted into Cartesian navigation goals for robot execution. The adoption of BTS ensures the robot’s feasibility around targets and adherence to cinematographic standards. Deployed on a biped robot platform, FotoBot demonstrates comprehensive navigation capabilities, effective human-robot interaction, and outstanding auto-photography performance. User trials conducted at the Hong Kong Science Park have validated FotoBot’s proficiency in navigating complex terrains and capturing high-quality photographs while intelligently responding to user instructions. Videos and code are available on the project website: <https://sites.google.com/view/fotobot/fotobot>.

Keywords: autonomous robotic photography; human-robot interaction; embodied AI



Copyright©2026 by the authors. Published by ELSP. This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium provided the original work is properly cited.

1. Introduction

The rise of digital photography has transformed the way we capture and share moments, with the global photography market valued at 47.8 billion in 2022 (source: <https://market.us/report/digital-photography-market/>) and projected to grow as social media continues to dominate communication. While smartphones and digital cameras remain the primary tools for photography, there is increasing interest in automating the photography process. An autonomous robot photographer has vast commercial potential, particularly for capturing high-quality images at events like weddings, conferences, or social gatherings. Many works have explored the use of drones for photography due to their agility and unique aerial perspectives [1]. However, concerns about safety, noise, and limited payload capacity make drones less suitable for most indoor scenarios [2]. Recent studies on robotic systems have demonstrated that robust perception, navigation, and decision-making frameworks can be achieved across complex environments through tightly integrated sensing and control modules [3]. Meanwhile, ground-based robots present a more suitable alternative for robotic photography in various conditions, including both indoor and outdoor settings. These robots can navigate safely, carry heavier and more sophisticated camera equipment, and interact more effectively with people and obstacles.

Over the years, numerous researchers have proposed robotic systems for photography, as summarized in Table 1. Prior work in this field faces notable limitations related to mobility, human-robot interaction, and compositional intelligence. Many systems—such as Li *et al.* [4], Rivkin *et al.* [5], and Limoyo *et al.* [6]—are constrained to fixed robotic arms, which restricts their ability to explore diverse or optimal compositions. While others, like the systems proposed by Newbury *et al.* [7], Zabarauskas *et al.* [8], and Alzayer [9], are limited to flat terrain, reducing their effectiveness in complex or uneven environments. Although some studies [1,10] employ drones to enhance mobility, concerns regarding safety, noise, and battery life restrict their use in most indoor and outdoor scenarios. While a few approaches have begun to incorporate user input for composition—such as instructor cues in Li *et al.* [4] and language or reference-image guidance in Limoyo *et al.* [6]—most systems offer minimal or no human-robot interaction (HRI). However, user instruction is essential in photography, forming the foundation of effective visual storytelling. Without robust mechanisms for human-robot interaction and compositional control, the potential of robotic photography is significantly diminished; if the user—as the creative force—cannot guide the robot, there is little value in employing such technology. We believe that in the era of robotics, robots and AI should serve as valuable assistants to humans, rather than replacing them. In terms of aesthetic quality, earlier works primarily rely on simple heuristics or basic face detection and cropping, which do not guarantee satisfactory results. Furthermore, open-sourcing remains rare, with only a handful of systems (e.g., Zabarauskas *et al.* [8] and Alzayer *et al.* [9]) releasing their code, thereby limiting reproducibility and extensibility. Moreover, solutions are often tightly coupled to specific hardware platforms, making replication difficult even when software is available. In summary, existing methods are frequently constrained by limited traversability, insufficient user interaction, and basic aesthetic reasoning, leaving substantial room for advancement.

This work introduces FotoBot, a user-friendly, vision-driven, all-terrain photography robot system that integrates real-time visual perception, a large language model [11] with a novel Bipedal Toric

Space (BTS) to enable autonomous capture of aesthetically pleasing photos. The key challenges addressed are: enabling effective user communication, teaching the robot to capture “good” photos, and ensuring traversability across diverse terrains. To address the first challenge, we employ Generative Pre-trained Transformers (GPT)-based models for natural, conversational interactions, allowing FotoBot to adapt to and respond to user preferences seamlessly. For the second challenge, we present the BTS for camera position planning, inspired by [12]. BTS compresses camera position representation into three parameters, each closely related to real-world photo composition. By adjusting these parameters, the robot determines the optimal camera position within the BTS and translates it into Cartesian coordinates for execution. Our experiments demonstrate that BTS is both efficient and effective for robotic photography with user interaction. To improve traversability, we utilize a bipedal robot platform, leveraging its ability to navigate a variety of terrains. As a result, FotoBot can operate seamlessly in both indoor and outdoor environments.

Table 1. Existing works on robot photography. FotoBot stands out from the previous works in terms of two aspects: Mobility FotoBot can navigate all-terrain scenes, including staircases, indoor and outdoor uneven terrains. User Interaction FotoBot use natural-language-driven interaction combine with conventional UI interface.

Method	Open-Sourced	Mobility	HRI Input Modality	Aesthetics & Composition
Li <i>et al.</i> [4]	No	Fixed	Subtle cues	Heuristics
Newbury <i>et al.</i> [7]	No	Flat ground	None	Face detection
Rivkin <i>et al.</i> [5]	No	Fixed	None	N/A
Zabarauskas <i>et al.</i> [8]	Yes	Flat ground	None	Heuristics
Huang <i>et al.</i> [10]	No	Aerial (limited indoor)	None	imitation learning (IL)
Bucker <i>et al.</i> [1]	No	Aerial (limited indoor)	None	Dynamic tracking
Alzayer <i>et al.</i> [9]	Yes	Flat ground	None	reinforcement learning (RL)
Limoyo <i>et al.</i> [6]	No	Fixed	Language + image	Vision-Language model (VLM)
FotoBot (Ours)	Yes	Bipedal, All-terrain	Language + UI	Biped Toric Space

The pipeline of FotoBot is illustrated in Figure 1. First, FotoBot listens to the user’s instructions and extracts their intention using a GPT engine. Upon understanding the user’s intention, the robot constructs a Bipedal Toric Space using its onboard sensors and plans a feasible trajectory for execution. Finally, once the user’s task is completed, FotoBot waits for the next instruction.

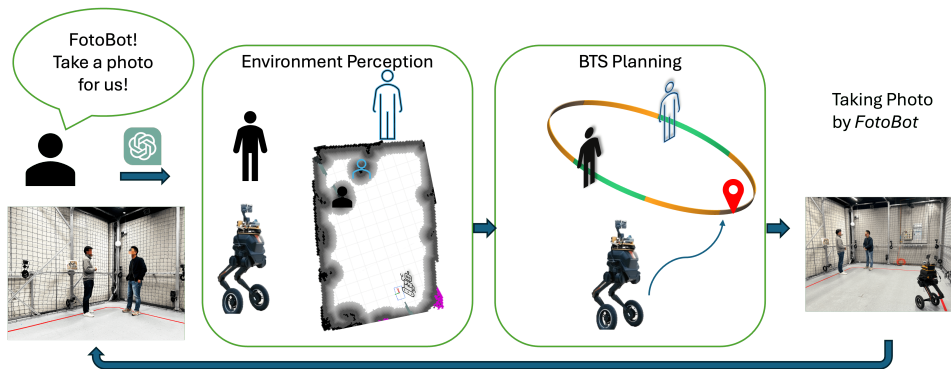


Figure 1. We demonstrate FotoBot, a user-friendly all-terrain photography robot system.

The software structure of the proposed FotoBot system is shown in Figure 2. At its core, FotoBot integrates a GPT-driven HRI module and a BTS-based high level planner for selecting the viewpoint. Additionally, the robot is equipped with comprehensive perception, planning, and control modules to support full navigation capabilities. The hardware components of FotoBot are introduced in Figure 3. Our system is deployed on a biped robot platform, which provides excellent mobility across various terrains.

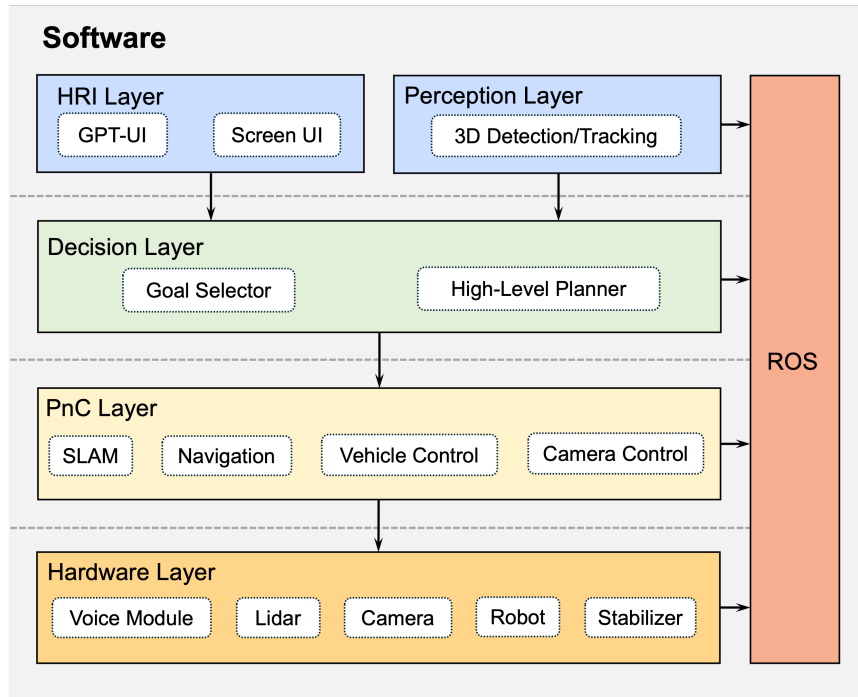


Figure 2. The software architecture of FotoBot is built on Robot Operating System (ROS) [13] and is structured into layers.

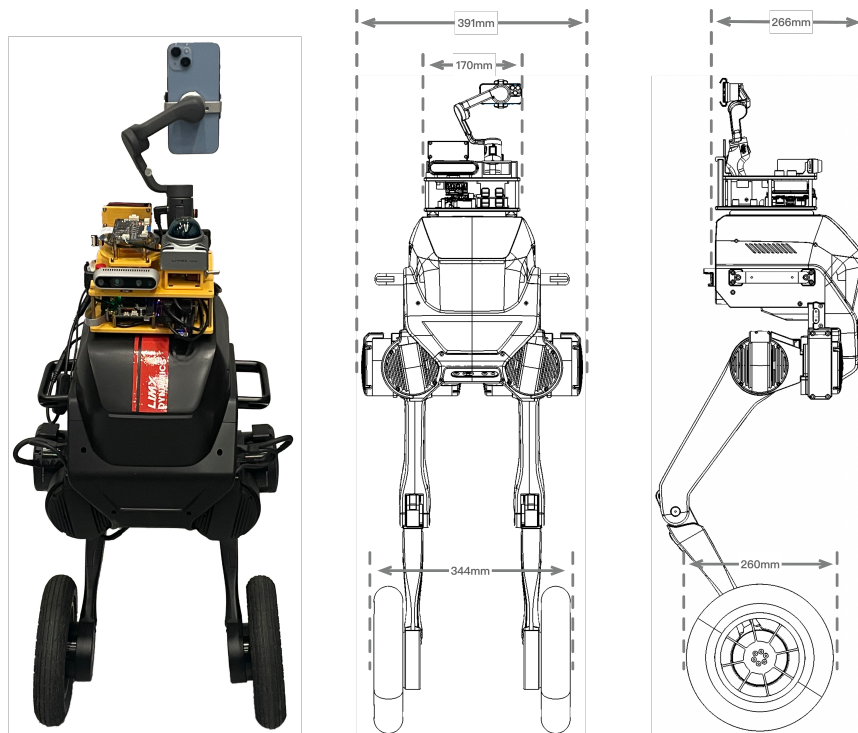


Figure 3. Hardware configuration of Fotobot.

To demonstrate the capabilities and performance of our proposed robotic photography system, we conducted user trials at the Hong Kong Science Park. Participants interacted with our robot and took pictures using it. The results, along with our demonstration video, highlight FotoBot’s ability to intelligently interpret human instructions, follow users, and navigate complex terrains with ease. Additionally, we conducted a user study to compare FotoBot with two baseline systems. The results demonstrate that FotoBot outperforms the baselines, confirming our hypothesis and showing that FotoBot offers a superior photo-taking experience. Most participants reported that FotoBot was easy to use and could achieve the desired framing directly, without requiring iterative adjustments. We will make the source code and design details for both the hardware and software publicly available on GitHub upon acceptance of this manuscript.

2. Robotic system overview

In this section, we introduce the hardware configuration, followed by an overview of the software architecture of the proposed FotoBot system.

2.1. Hardware configuration

The system is built on a bipedal robot, which serves as the platform for dynamic locomotion. On top of this robotic platform, we developed a custom sensing and computation suite to support the entire FotoBot system. It is powered by an NVIDIA Jetson Orin 16 GB module with an AI performance of 157 trillions of operations per second (TOPS), providing the necessary computational power for real-time sensor processing and deep learning applications. The system integrates a Livox light detection and ranging (LiDAR) module for high-resolution 3D scanning, enabling precise environmental mapping and obstacle detection. Additionally, an Intel Realsense depth camera provides depth perception for improved object avoidance and navigation. A DJI stabilizer ensures smooth motion by minimizing vibrations. While an iFlytek far-field microphone array module allows for accurate sound capture, enhancing speech recognition and auditory localization. The robot and Jetson module are connected to the internet via a Wi-Fi connection during operation. The sensing suite includes all necessary functionalities and devices, and can be used as a plug-and-play component for the robot.

2.2. Software architecture

The overview of the FotoBot software system is illustrated in Figure 2. The system is structured into functional layers: the HRI and perception layer, which enable interaction with and understanding of the real world; the decision layer, responsible for high-level decision-making, akin to the brain of a human; the Planning and Control (PnC) layer (containing sophisticated simultaneous localization and mapping (SLAM), navigation, and control algorithms), which executes these high-level decisions; and the hardware layer, which includes the drivers that facilitate communication with the hardware. The entire system operates on ROS Noetic [13], which provides module-to-module communication and essential services, ensuring a robust software environment for the integration of all components. The FotoBot software is modular, with well-defined interfaces between nodes, allowing for easy extension and modification with new algorithms or modules.

The core functionalities of FotoBot include following users and taking photos. Initially, the user provides instructions via natural language, which are recorded by voice module and sent to the HRI layer. The GPT-user interface (UI) module processes the input to extract the user’s intent, which is categorized into three types: chatting, following, and photo-taking. We will discuss the GPT-UI module in more detail in Section 3.1. In addition to voice interaction, we also offer a web-based user interface, as illustrated in Section 3.2. The perception layer is responsible for detecting and tracking the user, outputting 3D poses of the targets. Upon receiving the user’s intent, the goal selector switches between two modes—following or photo-taking—based on the identified action from user’s intent. In following mode, the goal selector uses 3D tracking data to generate a goal position near the user, enabling the robot to follow the user. If the user wishes to take a photo, the task is handed over to the high-level planner, which generates the optimal shooting position. This process is detailed in Section 4.

Finally, the goal is sent to the PnC layer, which includes a sophisticated robot navigation pipeline (SLAM [14], path planning [15], and robot control module) for execution.

3. Intelligent user interface

3.1. GPT interface

The goal of the GPT-driven intelligent UI is to accurately understand and interpret user’s intent with ambiguity, enabling intuitive interaction with FotoBot. By leveraging the power of Large Language Models (LLMs) [11], our system can process natural language inputs and extract meaningful instructions that guide the robot’s behavior. This approach significantly enhances the user experience by allowing for more flexible and conversational communication with the robot.

To ensure that FotoBot accurately understands and responds to user commands, we employed the technique known as “prompt engineering”. Giray *et al.* [16] This involves designing specific input prompts that guide GPT to correctly infer user intent. By tailoring the prompts, we are able to direct GPT to focus on particular aspects of the conversation—interpreting commands to follow the user, take a photo, or engage in a casual chat. For example, if the user says, “Take a picture of me,” the prompt is designed to steer GPT toward identifying the “film” intent. Similarly, commands like “Follow me” or “Stay with me” are recognized as “follow” instructions. After receiving the user’s natural language input L , we use the following prompt (Listing 1) to interact with GPT engine:

```
"role": "system", "content": "You are a human robot interaction system,
you need to identify whether the user is chatting with the robot
or ordering instuction to the robot. Now, we support two instructions,
one is following user, another is helping the user to take picture.
If the user want the robot to follow him/her, output "follow",
if the user want the robot to flim,
output "film" if the user just want to chat with the robot,
output the decent response
to chat with the user. ",
"role": "user", "content": L
```

Listing 1. GPT prompt.

To achieve state-of-the-art performance, we adopt the latest version of the OpenAI GPT engine, GPT-4o (source: <https://platform.openai.com/docs/models/gpt-4o>). The output of the GPT module can fall into one of three categories: ‘follow’ ‘film’ and ‘chat’. The first two intents are forwarded to the decision layer. If the user simply wishes to engage in a casual conversation with the robot, the GPT-generated response is directly output through the speaker module.

3.2. Web UI

To enable users to control the robot more smoothly, we also provide a conventional web-based UI for FotoBot. The web UI of Fotobot, as shown in Figure 4 is designed to provide users with an intuitive and comprehensive control experience. On the left side of the interface, a live video stream displays the current view from Fotobot’s front camera, allowing users to easily assess the scene and determine optimal photography parameters. To the right, a status panel provides real-time information about Fotobot’s operation, including the current mode, such as following or photo-taking, and immediate obstacle detection warnings to ensure safe operation. Below this panel, three function buttons—Following, Photo-taking, and Remote Control—allow users to seamlessly switch between Fotobot’s core functionalities, with the interface updating accordingly upon selection. At the bottom of the page, a real-time map offers spatial awareness: users can view a grid map generated by the SLAM module for indoor navigation or, when outdoors, see Fotobot’s location on a Google Map using GPS data from a connected mobile device. The UI is deployed on the robot’s edge computing device. Users can simply connect their mobile devices to the robot’s Wi-Fi network and access the interface through a web browser.

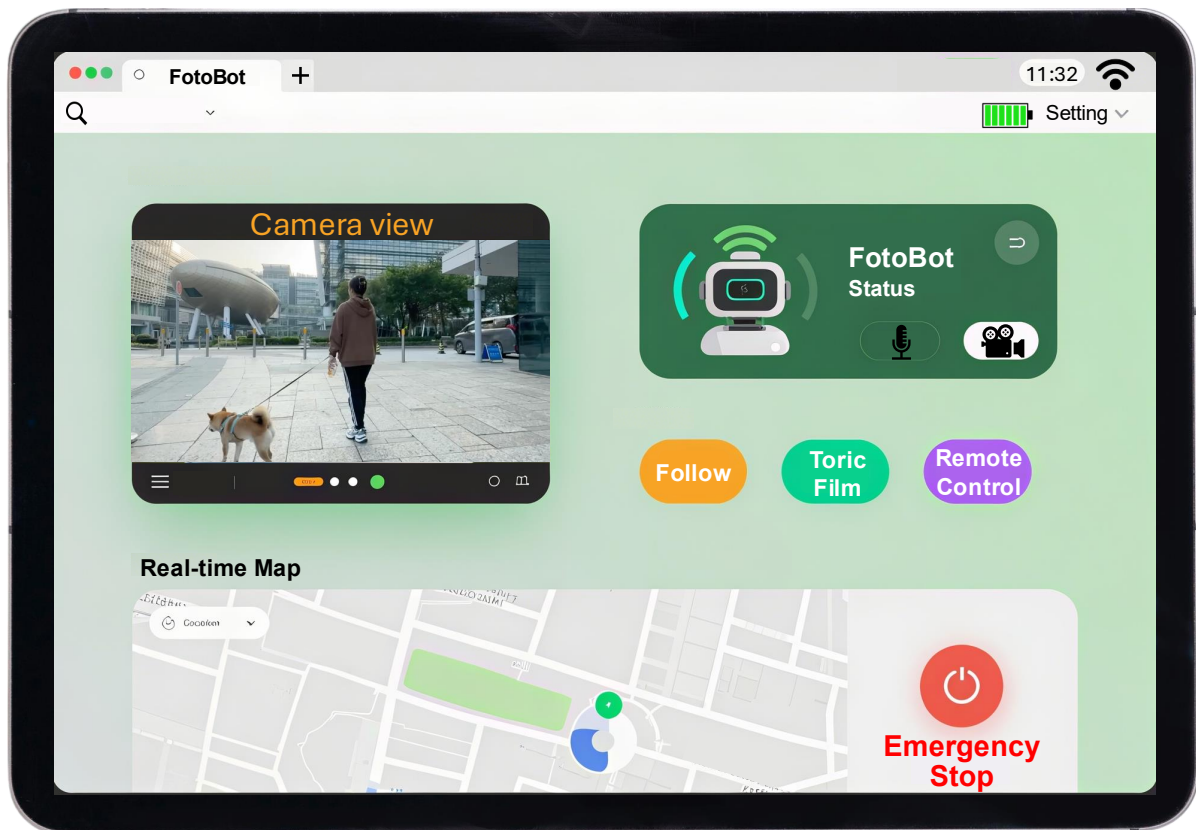


Figure 4. Web UI of FotoBot.

4. Taking good picture with FotoBot

We first enable the robot to search for the target person in the scene using a pose detector. Once detecting the person(s), we conduct Bipedal Toric Space to plan a “good” viewpoint for the robot to execute.

4.1. Perception

As the core vision-driven perception module of FotoBot, we deploy the lightweight YOLO11n-pose model [17] on the edge device to detect people in real-time and extract 17 keypoints of the face and body. These keypoints are then used to calculate an orientation vector of the person in the world frame, which we use to generate a viewpoint for heading toward the person. We project the 2D keypoint positions P_{pixel} at the pixel frame to the 3D camera frame using depth D and camera intrinsic K . These positions are then converted to the world frame P_{world} using camera extrinsics $[R|t]$,

$$P_{world} = [R|t]^{-1} K^{-1} P_{pixel} * D. \quad (1)$$

As such, we calculate an orientation vector that is perpendicular to the plane of the person in the world frame. Ultimately, this allows us to determine the 3D position and orientation of the target person(s).

4.2. Bipedal toric space

Inspired by [12], we propose BTS specifically tailored to our biped robot platform. The BTS is designed to streamline the manipulation and planning of cinematographic viewpoints for FotoBot, reducing complexity while maintaining expressiveness. The BTS simplifies the representation of camera position from seven parameters to three parameters: α , θ , and ϕ . These parameters are derived to capture the essential aspects of FotoBot viewpoint. The parameters (α, θ) defines a 2-parametric surface representing the set of possible camera positions, where α denotes the angle between the two targets and the camera, and θ represents the horizontal angle around targets. As shown in Figure 5, we depict the conceptualization of the camera viewpoint surface. For scenarios involving two targets, the set of camera viewpoints forms a spindle torus, with each point maintaining a constant angle α (computed from the screen positions) between the two targets and the camera. In the case of a single target, α represents the radius of the camera position circle, thereby unifying the notation.

We parameterize θ within the $[-1, 1]$ to translate the angle in a more meaningful manner, where $\theta = 0$ and $\theta = \pm 1$ correspond to viewpoints from behind target B and target A respectively.

The last parameter ϕ represents the vertical angle, allowing us to control the elevation of the viewpoint around the target on the screen.

This representation can be utilized for tasks such as automated viewpoint computation, significantly reducing the complexity of the planning space from 7 dimensions (for a standard camera) to 3 dimensions. This reduction applies to all classes of problems involving the viewing of one or two targets on the screen, taking into account their perceived 3D poses.

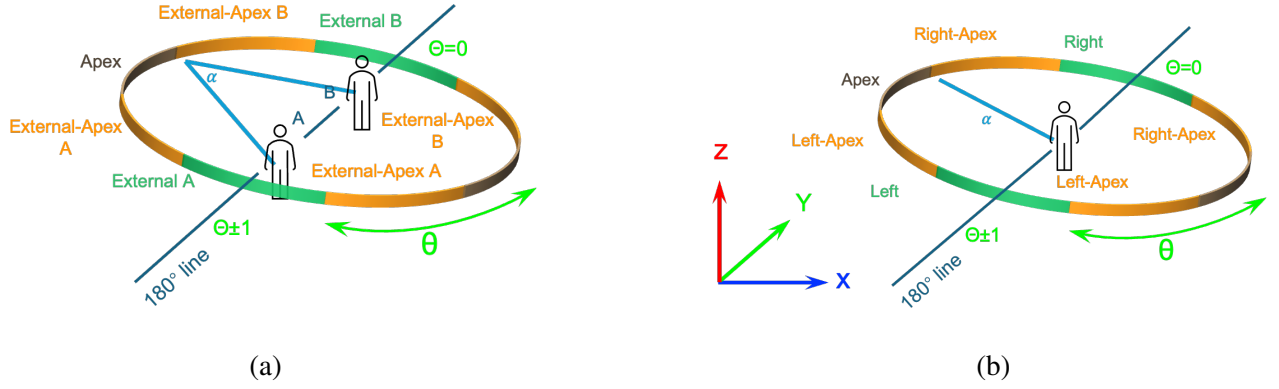


Figure 5. Camera viewpoints represented on the 2-parametric surface of the BTS.
(a) Scenario for two targets; (b) Scenario for one target.

4.3. Adjusting shot angle

One of the key advantages of BTS is the ease of adjusting the shot angle using a few parameters that are directly related to the final picture composition. We then explain how to adjust photography composition using BTS.

First, we adjust the α via adjusting the relative on-screen position of two targets. In this case, we assume that the screen is one-dimension and define two targets' normalized on-screen position as $P_A, P_B \in [-1, 1]$. Given the camera's field of view (FOV) ϕ_{FOV} and relative on-screen positions of two targets, we can compute the α as following:

$$\begin{aligned} \alpha &= (P_B^{cam}, P_A^{cam}) \\ P_B^{cam} &= \left(\frac{P_B}{\tan(\phi_{FOV})}, 1 \right) \\ P_A^{cam} &= \left(\frac{P_A}{\tan(\phi_{FOV})}, 1 \right) \end{aligned} \quad (2)$$

The user can adjust the relative on-screen position of two targets to adjust the viewpoint of camera.

We can adjust the shot angle by modifying the parameter θ within the BTS system. As illustrated in Figure 5, the shot angle for two target persons is divided into five categories, modeled as intervals of the parameter θ , based on the triangle principle. Similarly, we categorize the shot angle for a one-person scenario using the same logic. This categorization is summarized in Table 2. Users can specify their preferred θ value, prompting FotoBot to replan and execute a new goal accordingly.

Table 2. Shot angle categories and corresponding parameter θ in the BTS surface.

Shot Angle (two targets)	Apex	External-Apex A	External-Apex B	External A	External B
Shot Angle (one target)	Apex	Left-Apex	Right-Apex	Left	Right
θ	$(-\frac{5}{8}, -\frac{3}{8}), (\frac{3}{8}, \frac{5}{8})$	$(-\frac{7}{8}, -\frac{5}{8}), (\frac{5}{8}, \frac{7}{8})$	$(-\frac{3}{8}, -\frac{1}{8}), (\frac{1}{8}, \frac{3}{8})$	$(-1, -\frac{7}{8}), (\frac{7}{8}, 1)$	$(-\frac{1}{8}, \frac{1}{8})$

The relationship between the BTS parameters and screen framing is illustrated in Figure 6b. The parameter α controls the angular separation between two targets, or equivalently the camera–target distance in the single-target case. The parameter θ adjusts the horizontal filming angle around the target(s), allowing the camera to move left or right and thereby change the relative screen prominence and visual focus between targets. The parameter ϕ independently controls the vertical pitch of the camera, enabling fine-grained adjustment of viewpoint elevation without affecting horizontal composition.



Figure 6. Overview of the BTS for photography viewpoint control. **(a)** Geometric definition of the BTS surface; **(b)** Effect of BTS parameters on screen framing, illustrating how variations in α , θ , and ϕ influence target separation.

Upon initialization, FotoBot uses default parameters listed in Table 3. Users interact with the system through devices such as an iPhone or iPad (via a web browser, as detailed in Section 3.2), enabling them to adjust BTS parameters while viewing a real-time scene preview. This interface supports iterative parameter tuning during operation.

4.4. Planning with BTS

Given a representation of camera in BTS, it is crucial to convert it into Cartesian coordinates for the robot to execute. To transform the BTS representation (α, θ, ϕ) to its Cartesian representation $\mathcal{C}(x, y)$, we first convert the normalized horizontal angle $\theta \in [-1, 1]$ to the angle δ as following:

$$\delta = \theta * (2\pi - \alpha) \quad (3)$$

As shown in Figure 6a, δ represents the angle between $\overrightarrow{OP}$ and $\overrightarrow{OB}$. In this paper, we focus exclusively on scenarios where the robot is oriented towards the targets.

Then, the corresponding Cartesian representation $C(x, y)$ can be calculated as:

$$C = A + (q_\theta * AB) * \frac{\sin(\alpha + \theta/2)}{\sin(\alpha)} \quad (4)$$

where A and B represent the location of two target A and B respectively, q_θ represents the rotation angle $\theta/2$ from vector AB to vector AP as shown in Figure 6a.

Next, we need to calculate the orientation of FotoBot to ensure it is directed towards the target(s). Since the vertical angle ϕ has already been specified by the user, only the yaw angle ψ of the robot needs to be computed. In two-target scenarios, the robot's orientation will toward the mid-point of AB .

In the single-target case, we position the robot in front of the target and orient it to face the target. Following the methodology outlined in Chen *et al.* [18], we generate a robot goal C along the orientation vector of the target. This guides the robot to face the front of the detected person at a specified distance α , ensuring a feasible initial photo pose.

Finally, we determine the initial shooting position (x, y, ψ) for FotoBot to navigate to.

5. Robot navigation

To navigate to the viewpoint (x, y) , we construct the occupancy grid maps online by FAST-LIO [14] and employ the Dijkstra algorithm to plan a path with a fixed replan rate in case of getting stuck. The dynamic window approach (DWA) [15] is adopted to guarantee collision avoidance in the local space. Subsequently, the velocity commands are converted to torque commands by the low-level controller. We utilize the manufacturer-provided default policy model without additional fine-tuning. The robot operates in wheel mode at approximately 5 m/s, which is comparable to the average human walking speed.

The controller includes two modes: WALK mode and STAIR mode, each corresponding to different operational scenarios—flat ground movement and stair climbing, respectively. When the robot approaches stairs, the controller is manually switched to STAIR mode to enhance its all-terrain navigation capabilities. For flat ground navigation, the robot remains in WALK mode to optimize its high-speed movement. To minimize human intervention, the robot can continuously operate in STAIR mode for both flat ground and stair scenarios. However, performance may be compromised since the control algorithms are specifically trained for individual scenarios using proximal policy optimization (PPO) [19] in the GPU-based Isaac Gym simulator [20].

6. Experiments

In this section, we evaluate the proposed FotoBot system through real-world experiments and user studies. The default hyperparameters used in the experiments are detailed in Table 3.

Table 3. The default hyperparameters of FotoBot.

Parameter	Value
α (one-target)	1.5 m
α (two-target)	20°
θ	$\frac{1}{2}$
FOV	71°

6.1. User study

We recruited 20 volunteers to try our FotoBot prototype in our lab. All participants were students or postdoctoral researchers (ages 21 to 35; Mean: 27.8; SD: 4.07) with no professional photography training. Participants were asked to complete several tasks:

- Take a photo with the FotoBot system.
- Take a photo using an auto-tracking stabilizer and teleoperate the robot for navigation (Baseline 1).
- Take a photo using teleoperation of the robot and stabilizer (Baseline 2).

The two baseline systems were selected to reflect dominant control paradigms in current camera-controlling practice. Teleoperation (baseline 2) corresponds to the standard control method used in professional production environments, including robotic camera arms and mobile platforms deployed in studios, where camera motion and positioning are manually controlled by human operators. The auto-tracking gimbal baseline (baseline 1) represents widely adopted consumer products (e.g., Insta360 and DJI mobile gimbals), which provide automatic subject tracking and stabilization but require manual navigation and framing by users. These two baselines therefore capture both professional and consumer usage scenarios and span different levels of automation, enabling a meaningful comparison with FotoBot’s autonomous navigation and viewpoint planning. Representative user study scenarios are shown in Figure 7.



Figure 7. User study scenarios for FotoBot evaluation. Blue insets show the robot’s perspective during autonomous photography. The proposed system simplifies capturing well-composed photos for non-experts in diverse environments.

At the beginning of the study, a tutorial session was conducted to teach participants how to use the robot’s remote controller and how to adjust the viewpoint using BTS parameters. After the tutorial, each participant was asked to complete each task twice. To mitigate learning and ordering effects, the execution order of the three conditions was randomized across participants. We recorded user satisfaction ratings (based on both operation satisfaction and photography result satisfaction, on a scale from 0 to 10) and the execution time (in seconds) for each photo-taking trial. After completing the tasks, participants were asked to fill out a questionnaire comprising the following:

Participant Questionnaire

Satisfaction Rating (0–10):

Please rate your satisfaction with both the operational experience and the resulting photographs for each algorithm.

- Algorithm 1: _____
- Algorithm 2: _____
- Algorithm 3: _____

Feedback:

Please provide any comments or suggestions regarding the FotoBot system.

The results of the user study are illustrated in Figure 8 and Figure 9. This user study followed a within-subject (paired-samples) design, in which all participants evaluated all three conditions. To avoid treating repeated measurements from the same participant as independent samples, the two trials under each condition were aggregated by averaging, yielding one paired satisfaction score per condition per participant.

Prior to hypothesis testing, distributional assumptions were examined. Shapiro-Wilk tests indicated that the paired difference scores between conditions deviated from normality. Given the violation of normality, non-parametric paired tests were employed for all comparisons.

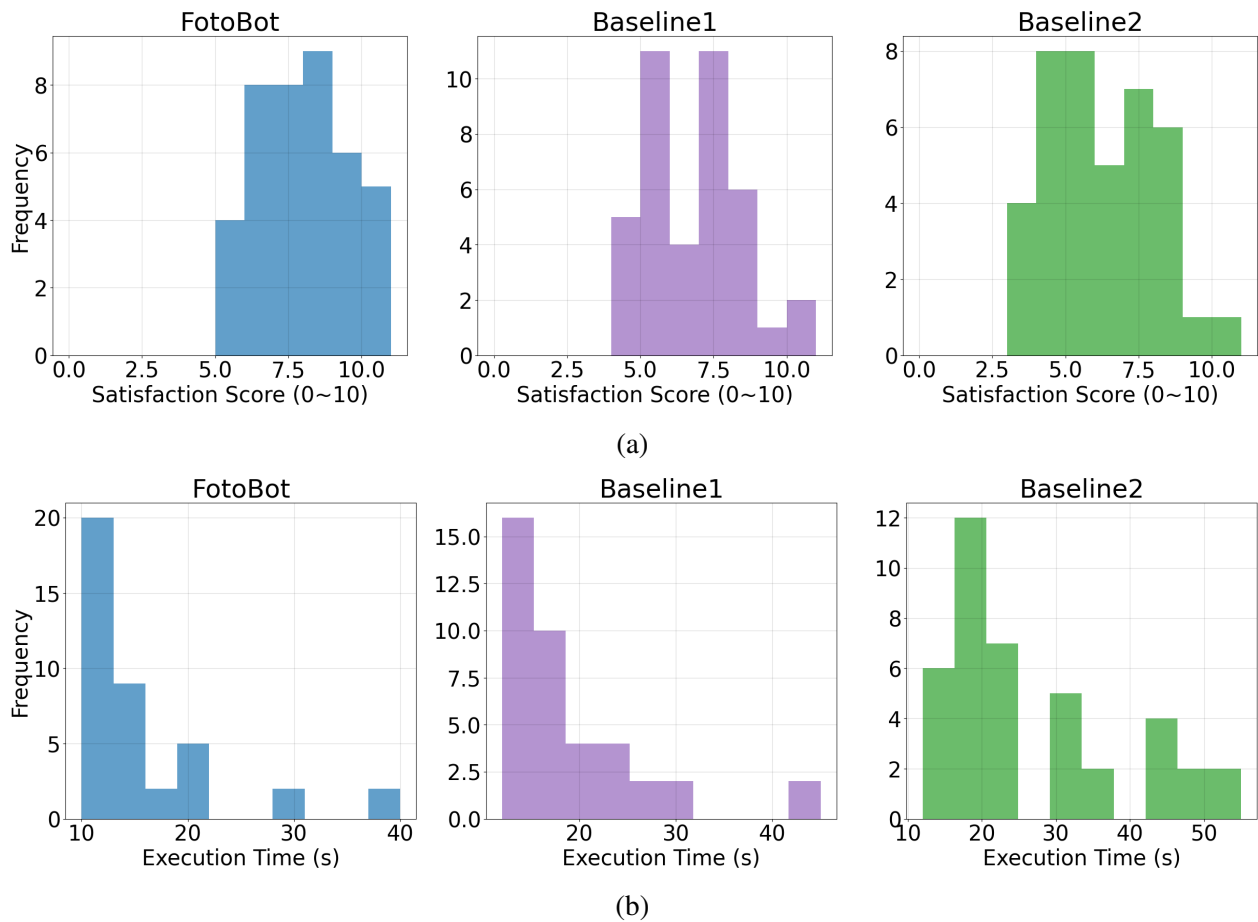


Figure 8. Quantitative user study results comparing FotoBot with two baseline systems. **(a)** Distribution of user satisfaction scores (0–10 scale) across all participants; **(b)** Distribution of total execution times for completing the photography task. FotoBot achieves both significantly higher user satisfaction and faster task completion than both baselines.

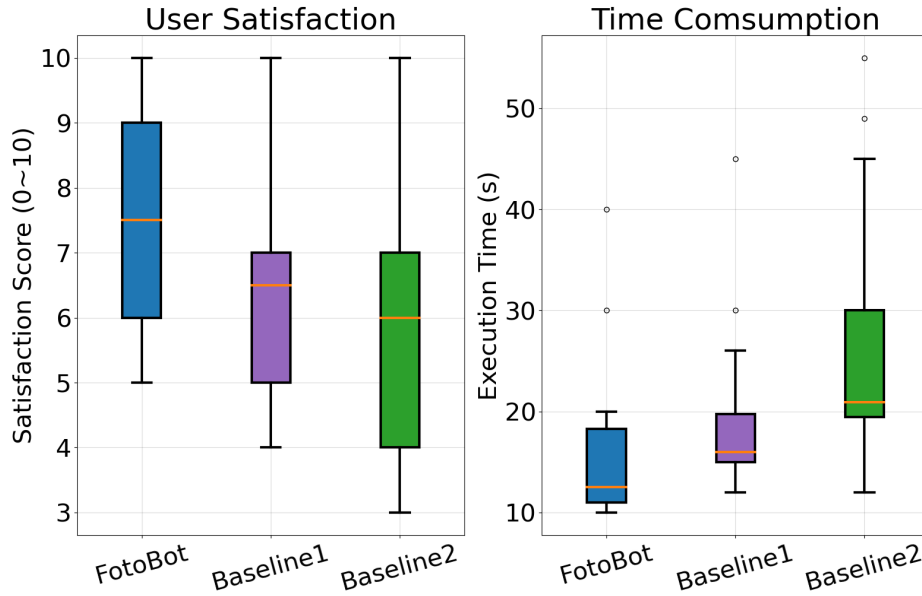


Figure 9. We compare user satisfaction (higher is better) and time consumption (lower is better) when taking a photo with FotoBot versus two baselines (see Section 6.1). The results indicate that our proposed FotoBot system is more efficient and easier to use, as evidenced by higher satisfaction scores and lower time consumption.

Pairwise comparisons between conditions were conducted using the Wilcoxon signed-rank test, with Holm–Bonferroni correction applied to account for multiple comparisons. The results indicate that FotoBot achieved significantly higher user satisfaction than Baseline 1 (mean difference = 1.175, $p < 0.0001$, effect size $r = 0.85$) and Baseline 2 (mean difference = 1.750, $p < 0.0001$, effect size $r = 0.83$).

Effect sizes were quantified using rank-biserial correlation (r), indicating large effects for FotoBot compared to both baselines. Overall, the statistical analysis provides strong evidence that FotoBot offers a superior photo-taking experience compared to the two baseline methods.

Another piece of evidence can be found in the time consumption results. The time required to achieve a good photography view with FotoBot is significantly reduced compared to the other two experimental settings, further demonstrating the high efficiency of the FotoBot system.

6.2. Quantitative evaluation of GPT-based intent classification

The GPT-based user interface in FotoBot is responsible for mapping free-form natural language input to a discrete set of robot intents. To quantitatively evaluate the reliability of this module, we conduct an explicit intent classification analysis.

6.2.1. Experiment setup

Intent understanding is formulated as single-label classification problem over three predefined categories: chat (conversational interaction without physical action), follow (user-following behavior), and film (photo-taking behavior). Only the follow and film intents are forwarded to the robot decision and planning modules, while chat produces no motion commands.

We construct an evaluation dataset of 240 natural-language utterances collected from pilot interactions and manually designed paraphrases covering common usage patterns and edge cases. Each utterance is

annotated with a single ground-truth intent label. During evaluation, the GPT prompt and model version are fixed, and each utterance is processed in a single forward pass without retries.

6.2.2. Results

We report the confusion matrix, per-class precision, recall, and F1 score, together with macro-averaged metrics to account for class balance.

The GPT-UI module achieves an overall classification accuracy of 95.42%. Table 4 shows the confusion matrix, and Table 5 summarizes the quantitative performance. The macro-averaged F1 score reaches 0.954, indicating strong and balanced intent recognition across all categories.

Table 4. Confusion matrix for GPT-UI intent classification. Rows denote ground-truth intents and columns denote predicted intents.

GT \ Pred	Follow	Film	Chat
Follow	79	0	2
Film	0	71	9
Chat	0	0	79

Table 5. Intent classification performance of GPT-UI.

Intent	Precision	Recall	F1
Follow	1.000	0.975	0.988
Film	1.000	0.888	0.940
Chat	0.878	1.000	0.935
Macro Average	0.959	0.954	0.954

A total of 11 misclassifications are observed. Most errors arise when short or underspecified commands related to camera operation (e.g., “start the camera”, “recording mode”) are interpreted as conversational input. In these cases, GPT-UI generates a clarification-style response rather than directly outputting the film intent. Misclassifications between follow and chat occur for socially phrased but weakly imperative utterances (e.g., “can the robot follow me?”).

This evaluation demonstrates that GPT-UI provides reliable intent parsing for real-world human-robot interaction, while maintaining a clear safety boundary between conversational interaction and motion-triggering commands.

6.3. Real-world experiments

To evaluate the proposed Fotobot system, we conducted experiments at the sea front area near Hong Kong Science Park, Pak Shek Kok, Hong Kong (Figure 10). The robot was required to follow the user across different terrains, including flat ground and staircases, to assess its ability to navigate and respond to dynamic changes in the environment. During the testing, the system’s language interaction capabilities were evaluated, enabling natural language communication between the robot and a human

user. Additionally, the robot was tasked with capturing pictures based on the user’s verbal instructions, demonstrating its ability to understand and execute commands intelligently. The robot was evaluated in an open-world scenario, where it navigated without any pre-existing maps, testing its adaptability and real-time decision-making capabilities in an unknown environment. This experiment provided valuable insights into the robot’s interaction, mobility, and task execution abilities, highlighting its potential for real-world applications.



Figure 10. Real-world experiment of FotoBot.

7. Discussion

Despite the promising results of our FotoBot system, several limitations need addressing to enhance its functionality and user experience.

First, the robot currently requires manual intervention to switch between operational modes (WALK/STAIR), which can hinder its efficiency in dynamic environments. This limitation arises because the control policy of the underlying robotic platform cannot handle all terrain types using a single locomotion policy. Given this constraint, the current prototype should be regarded as an autonomous photography robot with limited human intervention, as users are required to specify the photography intent and manually adjust the operation mode when necessary. As future work, we plan to integrate an advanced terrain perception system to enable automatic terrain recognition and mode switching, which is expected to significantly improve the robot’s adaptability and overall operational efficiency.

Additionally, the photography performance under the BTS planning framework still relies heavily on user-specific parameter tuning, which can be time-consuming and may lead to sub-optimal results without sufficient expertise. Moreover, the current system does not explicitly prune BTS parameters based on navigability or environmental constraints; instead, physical feasibility (e.g., collision avoidance and reachability) is handled at the downstream navigation stage, which may result in execution failure when the environment is highly constrained. Incorporating RL or IL techniques offers a promising direction to address both limitations in a unified manner. By jointly considering aesthetic objectives (e.g., composition and framing quality) and physical constraints (e.g., collision-free reachability and

terrain feasibility), a learned policy could automatically optimize BTS parameters, thereby enabling more consistent photography performance with minimal user intervention.

Thirdly, from an evaluation perspective, the user study was conducted with a relatively small sample size ($n = 20$), which is common and generally acceptable in real-world mobile robotics research due to constraints on hardware availability, safety, and deployment costs. Nevertheless, this limits the generalizability of the findings across broader user populations and varying levels of expertise. Looking forward, we plan to conduct larger-scale studies with more diverse participant groups, including expert photographers and users with prior experience operating robotic systems. Such studies will enable more rigorous evaluation of usability, learning curves, and compositional quality. In addition, extensive real-world deployments across varied environments, complemented by expert assessments, will provide more comprehensive benchmarks and deeper insights to inform future system improvements.

8. Conclusion

In this work, we present FotoBot, an innovative embodied AI robot system designed for photography. FotoBot is powered by GPT-based human interaction and a BTS-based planner for robotic photography. FotoBot can intelligently understand and respond to user instructions, take high-quality photos, and navigate various terrains. User study experiments demonstrate that the FotoBot system is user-friendly, enables users to achieve optimal shot positions with lower execution times, and results in high user satisfaction. The robot has been successfully tested in an outdoor area at the HK Science Park, achieving impressive results in real-world deployment.

Data availability statement

The data or datasets generated or analyzed in this study are available in <https://sites.google.com/view/fotobot>.

Declaration of generative AI and AI assisted technologies

During the preparation of this manuscript, the authors used generative AI tools only to improve language and readability. Specifically, the authors used ChatGPT for language polishing and text revision only. The authors take full responsibility for the entire content of the manuscript.

Acknowledgments

This work was supported by the Natural Science Foundation of China (62461160309), the NSFC-RGC Joint Research Scheme (N_HKU705/24), Hong Kong RGC (GRF 17201025, GRF17200924)

Authors' contribution

Dawei Wang: conceptualization, methodology, formal analysis, writing—original draft; Chen Chang: software, validation; Yipeng Pan: software, validation, data curation; Xinzheng Tang: validation, writing—review and editing, data curation; Hua Chen: resources, supervision; Jia Pan: conceptualization,

methodology, supervision, writing—review and editing, funding acquisition. All authors have read and agreed to the published version of the manuscript.

Conflicts of interest

Jia Pan holds the position of Associate Editor for *Robot Learning* and has not peer reviewed or made any editorial decisions for this paper.

References

- [1] Buckner A, Bonatti R, Scherer S. Do you see what I see? coordinating multiple aerial cameras for robot cinematography. In *Proceedings of 2021 IEEE International Conference on Robotics and Automation (ICRA)*, Xi'an, China, May 30–June 5, 2021, pp. 7972–7979.
- [2] Valavanis KP, Vachtsevanos GJ. *Handbook of Unmanned Aerial Vehicles*. Dordrecht: Springer, 2014.
- [3] Liu R, Hu X, Jiang Z, Wang J, Zhang W. Survey on heterogeneous aquatic robot systems: communication, perception, navigation, control, decision-making and energy management. *Robot Learn.* 2025, 2(1):0003.
- [4] Li J, Sousa M, Mahadevan K, Wang B, Aoyagui PA, *et al.* Stargazer: an interactive camera robot for capturing how-to videos based on subtle instructor cues. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, Hamburg, Germany, April 22–28, 2023, pp. 1–16.
- [5] Rivkin D, Dudek G, Kakodkar N, Meger D, Limoyo O, *et al.* ANSEL Photobot: a robot event photographer with semantic intelligence. In *Proceedings of 2023 IEEE International Conference on Robotics and Automation (ICRA)*, London, UK, May 29–June 2, 2023, pp. 8262–8268.
- [6] Limoyo O, Li J, Rivkin D, Kelly J, Dudek G. PhotoBot: reference-guided interactive photography via natural language. In *Proceedings of 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Abu Dhabi, UAE, October 14–18, 2024, pp. 2479–2486.
- [7] Newbury R, Cosgun A, Koseoglu M, Drummond T. Learning to take good pictures of people with a robot photographer. In *Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Las Vegas, USA, October 25–29, 2020, pp. 11268–11275.
- [8] Zabarauskas M, Cameron S. Luke: an autonomous robot photographer. In *Proceedings of 2014 IEEE International Conference on Robotics and Automation (ICRA)*, Hong Kong, China, May 31–June 5, 2014, pp. 1809–1815.
- [9] AlZayer H, Lin H, Bala K. Autophoto: aesthetic photo capture using reinforcement learning. In *Proceedings of 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, September 27–October 1, 2021, pp. 944–951.
- [10] Huang C, Dang Y, Chen P, Yang X, Cheng KT. One-shot imitation drone filming of human motion videos. *IEEE Trans. Pattern Anal. Mach. Intell.* 2021, 44(9):5335–5348.
- [11] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, *et al.* Language models are few-shot learners. In *Proceedings of Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, Virtual, December 6–12, 2020, pp. 1877–1901.
- [12] Galvane Q, Lino C, Christie M, Fleureau J, Servant F, *et al.* Directing cinematographic drones. *ACM Trans. Graphics.* 2018, 37(3):1–18.

- [13] Quigley M, Conley K, Gerkey B, Faust J, Foote T, *et al.* ROS: an open-source robot operating system. In *Proceedings of ICRA Workshop on Open Source Software*, Kobe, Japan, May 12–17, 2009, p. 5.
- [14] Xu W, Zhang F. FAST-LIO: a fast, robust lidar-inertial odometry package by tightly-coupled iterated kalman filter. *IEEE Rob. Autom. Lett.* 2021, 6(2):3317–3324.
- [15] Fox D, Burgard W, Thrun S. The dynamic window approach to collision avoidance. *IEEE Rob. Autom. Mag.* 1997, 4(1):23–33.
- [16] Giray L. Prompt engineering with ChatGPT: a guide for academic writers. *Ann. Biomed. Eng.* 2023, 51(12):2629–2633.
- [17] Khanam R, Hussain M. Yolov11: an overview of the key architectural enhancements. *arXiv* 2024, arXiv:2410.17725.
- [18] Chen C, Lu L, Yang L, Zhang Y, Chen Y, *et al.* Signage-aware exploration in open world using venue maps. *IEEE Rob. Autom. Lett.* 2025, 10(4):3414–3421.
- [19] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. *arXiv* 2017, arXiv:1707.06347.
- [20] Makoviychuk V, Wawrzyniak L, Guo Y, Lu M, Storey K, *et al.* Isaac gym: high performance gpu-based physics simulation for robot learning. *arXiv* 2021, arXiv:2108.10470.