

Autonomous robotic photography as a creative urban service

Jia Pan

jpan@cs.hku.hk

The University of Hong Kong <https://orcid.org/0000-0001-9003-2054>

Chang Chen

The University of Hong Kong <https://orcid.org/0009-0006-4198-3033>

Dawei Wang

The University of Hong Kong <https://orcid.org/0000-0003-2440-220X>

Zhilan Liao

Zhejiang University

Hang Xu

The University of Hong Kong

Yipeng Pan

The University of Hong Kong

Ignatius Japar

The University of Hong Kong

Mathieu Chambe

University of Rennes

Xiaofeng Liu

Hohai University

Hua Chen

LimX Dynamics

Marc Christie

University of Rennes

Article

Keywords: urban robotics, creative urban services, human–robot interaction, robotic photography, embodied artificial intelligence

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11118236/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.
[Read Full License](#)

Additional Declarations: There is **NO** Competing Interest.

Autonomous robotic photography as a creative urban service

Chang Chen^{1,6}, Dawei Wang¹, Zhilan Liao², Hang Xu¹,
Yipeng Pan¹, Ignatius Japar¹, Mathieu Chambe³, Xiaofeng Liu⁵,
Hua Chen^{4*}, Marc Christie^{3*}, Jia Pan^{1,4,6*}

¹The University of Hong Kong, Hong Kong SAR.

²Zhejiang University, China.

³University of Rennes, France.

⁴LimX Dynamics, China.

⁵Hohai University, China.

⁶Shenzhen Loop Area Institute, China.

*Corresponding author(s). E-mail(s): huachen.work@gmail.com;

marc.christie@irisa.fr; jpan@cs.hku.hk;

Contributing authors: changchen@connect.hku.hk;

dawei@connect.hku.hk; zhilan.liao26fall@gmail.com;

xuhang_official@outlook.com; yppancs@connect.hku.hk; iloe@hku.hk;

mathieu.a.chambe@irisa.fr; xfliu@hhu.edu.cn;

Abstract

Urban service robots are increasingly used for logistics, inspection, and construction, yet their role in cultural and recreational activities remains largely unexplored. We investigate autonomous robotic photography as a creative urban service, addressing a key limitation of existing systems: the difficulty of combining aesthetic judgment with safe, adaptive operation in dynamic public spaces. We present LEAP (Learning Embodied Aesthetics for Photography), a supervised robotic-photography system that integrates learned photography rules with fast state estimation, path planning, and obstacle avoidance. In three supervised urban deployments and 30 controlled photography tasks, LEAP completed all trials without collisions or incomplete-framing failures. It outperformed two robotic-photography baselines, achieving 9.7-19.8% higher computational aesthetic scores and 8.1-24.5% higher blinded human aesthetic ratings. LEAP also generalized to previously untrained indoor and outdoor settings and adapted to changing participant poses. These results demonstrate the potential of

autonomous robotic photography as an accessible, interactive, and culturally meaningful urban service.

Keywords: urban robotics, creative urban services, human–robot interaction, robotic photography, embodied artificial intelligence

Rapid urbanization is expanding both the scale and diversity of services required in public space. Autonomous service robots now operate in human-centred urban spaces, from streets and parks to campuses and hotels. Prominent applications include last-mile delivery [1], infrastructure inspection [2] and construction [3]. Robots in these applications typically perform utilitarian tasks that would otherwise require human labour. They must navigate these environments, avoid collisions with pedestrians, comply with social and traffic norms and complete predefined tasks. Urban life, however, also revolves around recreation, travel and social encounters, through which artistic practices such as drawing and photography emerge. Geotagged photographs, for example, reveal where people spend recreational time and how they encounter urban environments [4]. Such works do more than preserve memories of places and experiences. As cultural objects, they express how people see and relate to urban landscapes and to themselves. Urban robots could therefore support services that help people record, create and share urban life.

Autonomous robotic photography is a demanding test case for such services because composing an image is both creative and physically situated. Although cameras and drones are widely used in everyday life, they remain passive instruments that demand users’ attention. During travel, sport and social activities, a participant-facing mobile photography service could select viewpoints, compose photographs, respond to changes in the participant’s pose and print photographs on site (Fig. 1a). This would allow people to remain engaged in the experience rather than repeatedly operating a camera.

Realizing such a creative urban service raises two linked challenges. First, photography differs from established services such as last-mile delivery, for which success can usually be assessed by whether an item reaches a predefined destination. A photograph is instead co-produced with the person being photographed, and its success depends on image quality, subjective aesthetic preferences and the experience of the interaction. A participant-facing service must therefore navigate shared spaces safely, accommodate different bodies and modes of participation, make image capture legible and manage consent and identifiable data (Table 1). These requirements apply to the service as a whole, not only to the performance of the camera-control algorithm.

Second, existing robotic-photography systems have limited capacity to select, execute and revise image compositions autonomously under changing real-world conditions. A compelling photograph requires the continuous integration of aesthetic judgement and precise manoeuvring. Photography is therefore an active-sensing task [5]. The robot must move its camera to change perspective, reveal or occlude background elements and alter how the photographed person occupies the frame. It

must adapt these choices as the participant moves or changes pose. A usable service therefore requires autonomous, closed-loop camera control that updates camera motion through repeated visual feedback while avoiding collisions.

Table 1 Comparison of last-mile delivery and autonomous robotic photography as urban services. Both require safe and reliable operation in public space; robotic photography adds demands related to participant interaction, viewpoint selection and identifiable imagery.

Dimension	Last-mile delivery	Autonomous robotic photography
Service outcome	Delivery of goods to a specified destination	Photographs co-produced through participant pose and robot viewpoint
Success criteria	Timely arrival, item integrity and successful handoff	Session completion, photographic quality and participant preference
Human role	Passersby share the route; recipients interact mainly at handoff	Participants engage throughout; bystanders may enter the robot’s path or camera view
Interaction	Concentrated around requesting and receiving delivery	Initiation, posing, capture, withdrawal and assistance
Spatial operation	Route following with collision avoidance	Viewpoint adjustment around participants with safe, predictable movement
Data and context	Location, logistics and delivery records; route and sidewalk conditions	Identifiable imagery; consent and data handling; crowding, layout and lighting

Existing robotic-photography research addresses individual components of this service, but not the combined demands of participant interaction, adaptive viewpoint control, and the handling of identifiable imagery in public space. Geometric methods generate camera paths from predefined shot patterns or target poses [6–9]. Rule-based systems position photographed subjects according to fixed compositional conventions [10–16], while learned aesthetic models select or evaluate frames and viewpoints [17–22]. These advances improve camera assistance and image assessment, but individually they do not provide a complete participant-facing workflow. Such a workflow must support entry and withdrawal, respond to participants throughout a session, make image capture legible, and manage bystanders and identifiable images. Robotic photography as an urban service therefore requires both technical performance and service-level interaction and governance.

Even within viewpoint selection, aesthetic evaluation and physical camera movement generally remain separated. Geometric and rule-based approaches first specify a desired camera pose or compositional pattern and then move the camera towards it. Learned aesthetic models improve frame evaluation, but movement trajectories are generally predefined or optimized separately from the visual judgments used to assess

a Photography as an urban creative service

(i) Initiate the session



Participant consents and initiates pose

(ii) Compose and capture



Robot adjusts its viewpoint through translation (x,y) and rotation (yaw, pitch)

(iii) Select and print



Selected image is printed on site

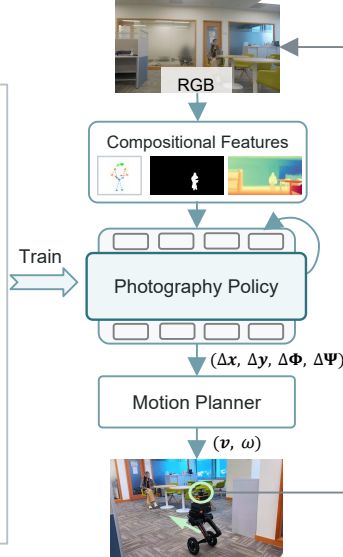
(iv) Receive and respond



Participant receives photo and gives feedback

b Method overview

Real-world photography demonstrations



Closed-loop execution

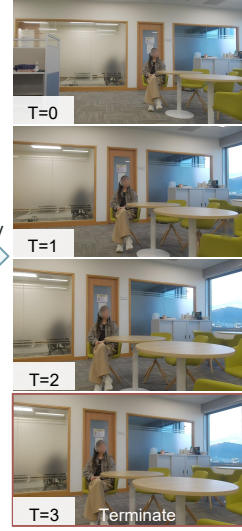


Fig. 1 Photography service workflow and system overview. (a) The participant-facing process links session initiation and consent to autonomous viewpoint adjustment, image capture and delivery of an on-site print. Human supervision remains available throughout the service. (b) (Left) An expert-operated data-collection apparatus records continuous camera trajectories and synchronized egocentric video. (Middle) The imitation-learning pipeline converts RGB images into skeletal keypoints, relative depth and a foreground mask to train a conditional flow-matching (CFM) model. (Right) A closed-loop hierarchical architecture connects learned viewpoint selection to collision-aware physical navigation.

their outcomes. This separation can support planned or static compositions, but limits repeated recomposition when participant posture, visibility, background structure, or the executed route changes. A participant-facing urban service instead requires aesthetic judgment, camera movement, collision avoidance, and visual reassessment to operate in a closed loop.

Data-driven closed-loop control offers a route beyond this separation, but existing training sources and objectives remain poorly matched to participant-facing urban photography. Some studies learn trajectories from Internet drone videos or synthetic trajectories [23–27], while others optimize viewpoints through reinforcement learning (RL) in simulation [28–30]. Internet film videos are readily accessible, but provide few examples of recovery from poor viewpoints. Moreover, aesthetic quality is difficult to formulate as a dense reward for RL training, while simulation cannot capture the full diversity of interactions with changing participants and environments. More recently, generative imitation learning has shown strong capability in representing multiple demonstrated actions for similar observations [31, 32]. These methods learn control policies directly from real-world expert demonstrations, represented as paired vision–action sequences. This property is particularly well suited to photography, where multiple camera movements may lead to distinct yet equally acceptable compositions. The remaining need is therefore a photography demonstration dataset of this kind, together with a tailored approach that learns such alternatives from human demonstrations, refines them through repeated visual feedback, and embeds them within a supervised urban-service workflow.

We developed LEAP (Learning Embodied Aesthetics for Photography) to jointly address these service and control challenges. We introduce the first photography-oriented real-world trajectory dataset, comprising 864 high-accuracy video–action sequences across 50 distinct subject poses. Learned photography rules guide camera motion, while a faster geometric planner estimates the robot state, generates safe paths, and avoids dynamic obstacles. This hierarchical architecture repeatedly observes the scene, moves, and revises the proposed composition until the predicted camera motion converges and the shutter is triggered. We embedded the system within a participant-facing workflow encompassing initiation, supervision, image capture, transfer, on-site printing, and delivery. Three supervised urban deployments evaluated this integrated service during a campus orientation day, a senior-centre recreational event, and at Shenzhen Talent Park. Together, these deployments spanned semi-public events, older-adult recreation, and public-park tourism contexts. Across 30 controlled photography tasks, LEAP significantly outperformed two robotic-photography baselines and approached the compositional performance of a professional-photographer reference. Together, these findings demonstrate the feasibility of autonomous robotic photography as a creative urban service.

1 Results

We evaluated the service workflow and the autonomous composition capabilities needed to support it. Three supervised urban deployments examined operation from

session initiation to print delivery, including completion, intervention and safety-related events. A 30-task controlled evaluation then assessed autonomous viewpoint selection and photographic quality against Bipedal Toric Space (BTS) [8], a semi-autonomous method; Visual Servoing with Rule of Thirds (VS-RoT) [12, 16], a two-stage autonomous method; and a constrained professional-photographer reference (Human). These robotic-photography baselines represent state-of-the-art implementations of two common portrait-composition strategies, face-centred framing and the Rule of Thirds. Unlike methods evaluated only in simulation, both baselines can be deployed in real-world settings.

Further experiments tested transfer to physical locations and scenes outside LEAP’s training environment, recomposition after visual change, varied framing and the visual cues and demonstrations supporting these behaviours. Movie S1 summarizes the system pipeline, comparative performance and participant experience across indoor and outdoor scenarios in Hong Kong. Movie S2 documents a one-day supervised deployment in Shenzhen Talent Park.

1.1 Operational feasibility in three urban settings

To examine LEAP’s operational feasibility as an urban service, we deployed the system in three public and semi-public settings in 2026. These comprised Shenzhen Talent Park, used by local residents and visitors; a senior-centre recreational event for older adults; and a campus orientation day for students and parents. Deployment dates and operating hours are listed in Table 2. The settings differed in layout, lighting, pedestrian flow and participant configurations.

The service workflow is illustrated in Fig. 1a. Staff initially selected an open area with little pedestrian traffic to limit background clutter, while the robot’s collision-avoidance system remained active. Research staff invited potential participants and obtained consent before each session. Participants then chose their location and pose, while the robot was placed approximately 2 metres in front of them. LEAP alternated between observing the participant and moving its wheeled base and pan-tilt camera, capturing a photograph after each execution step. The capture sequence ended when the predicted movement converged or after five execution steps. Participants could change poses during the session or start a new one. Staff remained responsible for supervision and could take over, stop or cancel a session. After each session, staff helped participants select their preferred photographs, printed them on site using a portable Canon SELPHY CP1500 printer and delivered the prints. An eye-closure detector [33] automatically filtered images in which participants had closed their eyes. Thus, viewpoint selection and capture were autonomous, whereas enrolment, safety oversight and delivery remained human-supported.

Across a total of 14 hours of deployment, LEAP served 45 participants in 71 sessions (Table 2). All sessions reached end-to-end completion at Shenzhen Talent Park (18/18), the senior-centre event (33/33) and the campus event (20/20). No collisions were observed at any of the three sites.

Each participant could use the service multiple times, with repeat use included in the total number of sessions (n). We found that older adults and park visitors showed the greatest interest in the photography service, with session-to-participant ratios of

a Shenzhen Talent Park



b Senior-centre event



c Campus orientation day



Fig. 2 Supervised robotic-photography deployments in three urban settings. The three settings were a campus orientation day, a senior-centre recreational event and Shenzhen Talent Park. The panels document the participant-facing workflow, individual and group sessions, varied back-grounds and operation near other users of the space.

2.06 and 1.64, respectively. Despite its shorter duration (only 3 hours), the senior-centre deployment recorded the most sessions. We attribute this interest to the value placed on photographs as records of retirement experiences, combined with limited photographic expertise. The campus event, by contrast, included the most group sessions. Although visitors naturally sought a photography service, we had relatively few opportunities to meet and invite them because of the dispersed layout of the park.

To examine shooting distance, we retrospectively estimated the minimum person depth from the images recorded during each session using Depth Anything 3 [34], as the pan-tilt camera did not measure depth. We averaged these minimum depth estimates across images separately for each site. The mean image-level minimum depth was 3.397 m at the park, compared with 1.980 m at the senior-centre event and 1.962 m

Table 2 Operational outcomes of three supervised urban deployments.

Deployment metric	Shenzhen Talent Park	Senior-centre event	Campus orientation day
Deployment date	31 March	18 August	2 September
Operating hours	11:00–17:00	15:00–18:00	10:30–15:30
Deployment duration (h)	6	3	5
Unique participants (n)	11	16	18
Total sessions (n)	18	33	20
Group sessions (n)	2	2	4
End-to-end completion	18/18 (100%)	33/33 (100%)	20/20 (100%)
Mean image-level minimum depth (m)	3.397	1.980	1.962
Lowest session-level depth (m)	1.201	0.943	1.241
Mean RoT score	0.855	0.829	0.806
Mean VEN score	0.928	0.885	0.909
Mean CLIP score	0.673	0.832	0.707

Sites were Shenzhen Talent Park, a senior-centre recreational event and a campus orientation day. All dates are in 2026; times are local. Group sessions are included in session totals. Completion is reported as completed/initiated sessions (%). Mean image-level minimum depth is averaged across analysed images at each site. Aesthetic scores are means for delivered photographs: Rule of Thirds (RoT), View Evaluation Network (VEN) and CLIP-based assessment.

at the campus event. This result showed that the learned photography rules generally guided the robot to operate at a longer distance outdoors, providing a wider field of view than in the indoor scenes. As a supporting measure of closest approach, we took the minimum image-level depth within each session and then the lowest session-level value at each site. These values were 1.201 m, 0.943 m and 1.241 m, respectively, across 11, 15 and 18 analysed sessions, demonstrating that the system provided an effective safety guarantee.

We also assessed the delivered photographs using three computational aesthetic measures. RoT [35] measures how well the person’s face is positioned according to the Rule of Thirds, a common photographic convention that places key subjects near lines or intersections dividing the image into thirds. VEN [19] is a neural network that rates whole-image composition quality and returns a normalized score between 0 and 1. The CLIP-based aesthetic score [36] uses a vision–language model to measure how closely a photograph resembles a “good” rather than a “bad” portrait. Across sites, mean scores ranged from 0.806 to 0.855 for RoT, 0.885 to 0.928 for VEN and 0.673 to 0.832 for CLIP. The variation in scores revealed different preferences across sites with different participants, poses and backgrounds. Park photographs scored highest on RoT and VEN, whereas senior-centre photographs scored highest on CLIP. These results show that the service operated across diverse urban settings with generally high image-composition quality.

Alongside these measurements, informal conversations provided feedback on how participants used the service. Participants described its usefulness when their hands

were occupied, for example while holding an umbrella, or when being photographed with companions. They also suggested better printer integration, a better camera lens and greater control over background objects, such as a tower. Another comment emphasized the need for voice prompts to guide posing and indicate when to change pose. This feedback was collected informally and serves as input to service design rather than a measure of satisfaction or acceptance.

Together, these supervised deployments documented the complete robotic-photography workflow across three urban settings, from participant initiation and autonomous viewpoint adjustment to image capture and on-site print delivery.

1.2 Autonomous viewpoint selection and photographic quality

We next tested whether autonomous viewpoint selection could produce usable photographic outcomes across controlled variations in pose and setting. The evaluation comprised 12 tasks in the motion-capture laboratory, eight in untrained indoor environments and ten in untrained outdoor environments, spanning 14 new locations. Tasks included standing, sitting, squatting and lying poses. Each approach began approximately 2 m in front of the photographed subject and was repeated three times per task. Quantitative values are reported as mean $\pm$ 95% confidence interval across independent tasks.

LEAP was compared with two robotic-photography approaches and a professional photographer. BTS used manually specified geometric relationships between camera and subject [8], whereas VS-RoT combined autonomous frontal-view navigation with rule-based composition adjustment [12, 16]. The photographer used the same camera height, sensor and 4-DoF movement constraints as the robot, providing a constrained human reference rather than unrestricted professional photography. No LEAP trial met the predefined failure criteria under the tested conditions.

Prior robotic-photography studies have mainly relied on subjective evaluation. We therefore assessed complementary aspects of the final photographs using three computational measures and a blinded online evaluation. The computational measures were the Rule of Thirds (RoT) [35], the View Evaluation Network (VEN) [19] and a CLIP-based aesthetic score [36] (Fig. 3a). VS-RoT achieved the highest RoT compliance, as expected from its explicit optimization target. LEAP achieved the second-highest RoT score, with final subject positions varying across poses and backgrounds in a pattern closer to the constrained photographer.

Across the 30 tasks, LEAP’s task-weighted mean VEN score was 11.9% higher than BTS and 9.7% higher than VS-RoT. Its CLIP score was 17.0% and 19.8% higher, respectively. VS-RoT retained the highest task-weighted mean RoT score (0.905 versus 0.843 for LEAP), showing that the learned approach did not dominate every measure. Representative BTS outputs cropped parts of the body in non-upright poses, while VS-RoT sometimes reached a frontal view without producing a suitable final composition (Fig. 4). LEAP also produced composition patterns that were not explicitly encoded as rules, including *gaze room* [37], which leaves negative space in the direction of the subject’s gaze to improve balance and storytelling.

The blinded online evaluation involved 24 participants, including 17 photography enthusiasts. They viewed anonymized photographs without method labels or

capture-order cues, and image options were randomly ordered within each question for every participant. Ratings covered four dimensions: composition quality (Q1, “How well is the image composition?”), subject emphasis (Q2, “How well is the main subject emphasized?”), subject–background relationship (Q3, “How well is the relation between character and background in the image?”) and overall aesthetic impression (Q4). Q1–Q3 used three ordered categories (good, medium and bad), whereas Q4 used a nine-point Likert-type scale.

After weighting scene-level means by the 30 tasks, LEAP’s mean overall-aesthetic score was 8.1% higher than BTS and 24.5% higher than VS-RoT, while remaining 12.4% below the constrained photographer. Rating distributions for composition quality and subject emphasis were also closer to the photographer than to either robotic approach (Fig. 3b). Inter-rater agreement, assessed using Kendall’s coefficient of concordance [38], was $W = 0.725$, 0.725 and 0.678 for Q1, Q2 and Q3, respectively. These observations support a relative photographic-quality advantage under the study conditions, without equating the online image assessment with acceptance of the physical service.

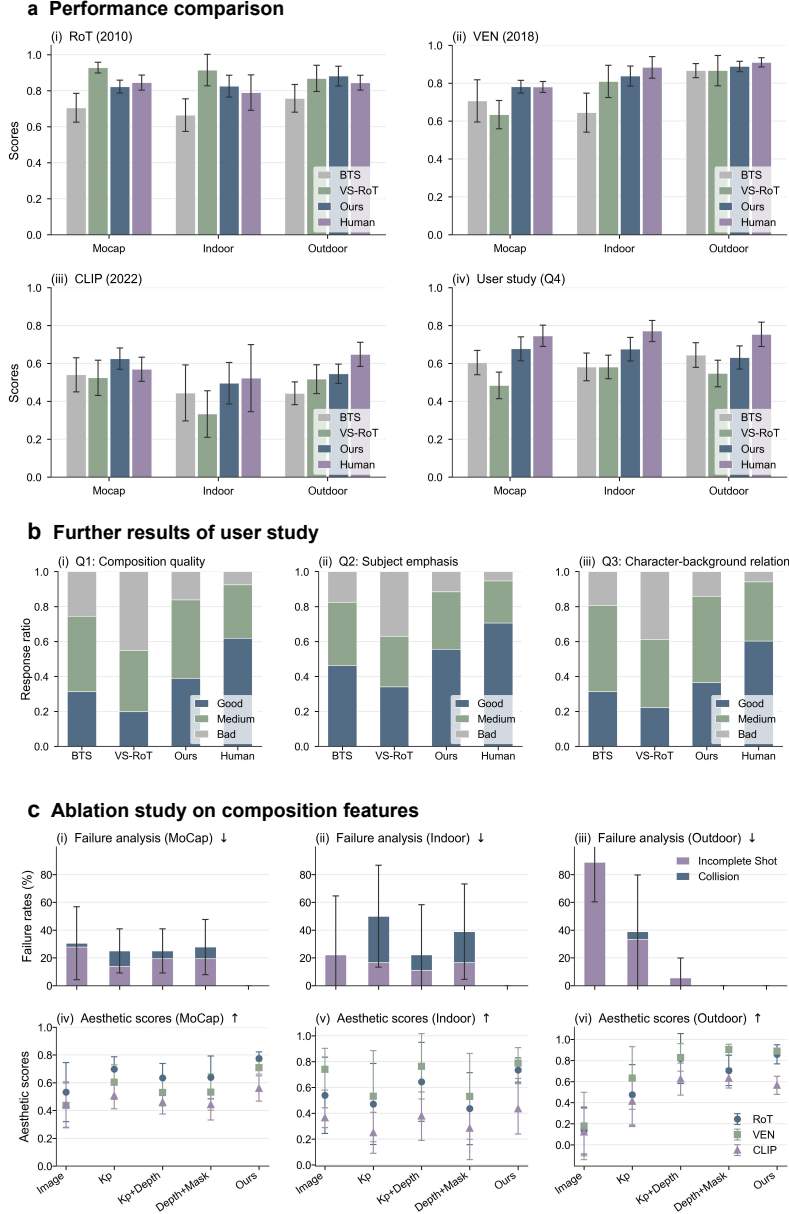


Fig. 3 Photographic outcomes and compositional cues. (a) Photographic outcomes across the motion-capture laboratory and untrained indoor and outdoor environments. LEAP was compared with BTS, VS-RoT and a constrained professional-photographer reference using RoT, VEN, CLIP and the Q4 overall aesthetic rating. (b) Blinded participant-rating distributions for composition quality (Q1), subject emphasis (Q2) and subject-background relationship (Q3). (c) Comparison of the combined compositional representation with four partial representations, reporting physical failures and computational aesthetic scores.



Fig. 4 Qualitative comparison of photographic compositions. Representative photographs captured by LEAP, BTS, VS-RoT and a professional photographer from the same initial conditions and under the same 4-DoF constraints. Among the robotic approaches shown, LEAP produced the compositional style closest to that of the constrained professional photographer across subject poses and settings. Aggregate computational and participant-rating comparisons are reported in Fig. 3. The original questionnaire and extended comparison examples are available on the [user-study website](#).

1.3 Compositional cues support transfer to untrained settings

Operation across unfamiliar settings requires a view representation that is less dependent on clothing, lighting and background appearance. We therefore compared LEAP’s skeletal keypoints, relative depth and foreground mask with four partial alternatives: raw RGB images, keypoints alone, keypoints with relative depth and relative depth with a foreground mask but without keypoints.

All partial representations produced more failures than the full representation across the 24 tasks (Fig. 3c). Failures were most frequent in spatially constrained laboratory and indoor settings and included collisions and incomplete framing. The raw-image variant failed in 16 of 18 untrained outdoor trials (88.9%), indicating sensitivity to changes in appearance and lighting.

Keypoints alone described body configuration but provided no explicit scene geometry, and sometimes led to large movements, collisions or incomplete photographs. Adding relative depth reduced this ambiguity but did not describe the full foreground boundary. The combined representation provided the lowest observed failure rates across the tested scene layouts.

Removing keypoints from the depth-and-mask variant also increased failures and reduced aesthetic scores in the laboratory and indoor settings. Nonlinear depth normalization retained nearby spatial relations while compressing variation in distant backgrounds. Together, these comparisons show that body configuration, relative scene geometry and the foreground boundary contributed complementary information for operation beyond the training environment.

1.4 Repeated observation supports recomposition and varied framing

The controlled evaluation established transfer from the motion-capture laboratory to eight untrained indoor and ten untrained outdoor tasks. We further examined how repeated observation affected responses to participant movement, partial visibility and background changes. Repeated trials from a common initial condition tested whether the service could produce more than one framing. Figure 5 presents illustrative cases rather than a population-level sample of dynamic encounters.

In the selected recovery cases, participant movement or proximity to an image boundary reduced facial visibility or partially excluded the body (Fig. 5a-i,ii). LEAP moved backwards to restore shooting distance and then recomposed the image. It also adjusted its viewpoint after posture changes, including a transition from standing to a deep squat (Fig. 5a-iii,iv). The expert demonstrations retained corrections after initially unsuitable compositions, unlike Internet video datasets that tend to emphasize successful shots without paired corrective actions [23, 28]. Repeated observation allowed comparable corrections during physical execution.

Repeated trials produced more than one final composition from the same initial condition (Fig. 5b). Across 24 tasks, LEAP produced a task-weighted average of 2.0 spatial composition types in three trials, varying among left, middle and right shooting positions. This outcome is consistent with the alternative demonstrated movements represented by conditional flow matching [32]. Within the tested conditions, LEAP therefore transferred to unfamiliar scenes, recomposed after visual change and avoided reducing every session to one fixed framing.

a Qualitative demonstrations for robustness toward dynamic changes



b Qualitative examples of the diverse shot photos

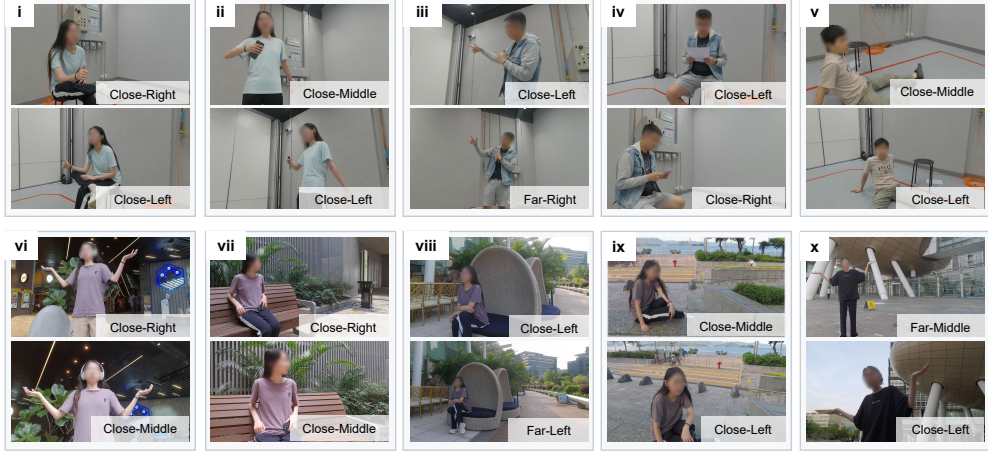


Fig. 5 Illustrative closed-loop recovery, adaptation and compositional diversity. (a) Selected physical trajectories. Rows (i) and (ii) show recovery from degraded initial viewpoints. Rows (iii) and (iv) show recomposition after changes in subject posture, with black arrows marking the change between robot timesteps $t = 2$ and $t = 3$. Muted red boxes show the final photographs captured after the predicted movement converged. (b) Repeated trials from the same initial condition produced several spatial framing types represented in the demonstrations.

1.5 Demonstration volume and photographer preference

We finally examined the amount and source of expert demonstrations needed to support the service (Supplementary Fig. A2). We used Dataset *A* and Dataset *B* (see data description in Sec. 3.3) to evaluate how the learned model improved as demonstration volume and source increased. Models trained on 20% of Dataset *A* failed in 26 of 36

laboratory trials and 13 of 18 untrained indoor trials, compared with no observed failures across 72 trials at 100%. Adding Dataset *B* increased task-weighted mean framing diversity from 1.75 to 2.00 spatial types and retained the wider compositions preferred by the second photographer. These results show that the current demonstration volume enables failure-free operational reliability, while increasing demonstration volume and diversifying its sources can further scale aesthetic performance and framing diversity.

2 Discussion

In this work, we examined the participant-facing workflow in three supervised urban deployments: Shenzhen Talent Park, a senior-centre recreational event and a campus orientation day. Across 14 hours, 45 participants took part in 71 sessions, all of which reached end-to-end completion, with generally good composition scores across sites. We then tested the autonomous composition capabilities required to support this service across 30 controlled photography tasks in the training laboratory and untrained indoor and outdoor settings. No LEAP trial met the two predefined failure criteria of collision or incomplete framing. Its task-weighted mean VEN and CLIP scores were 9.7–19.8% higher than those of two robotic-photography approaches. In a blinded online evaluation, LEAP’s overall aesthetic score was 8.1% higher than that of BTS and 24.5% higher than that of VS-RoT. Additional controlled experiments assessed transfer to untrained settings, closed-loop recomposition after participant or route changes and compositional diversity across repeated trials from common initial conditions. Further analyses identified the visual cues and demonstration data that supported these behaviours. Together, these findings establish autonomous robotic photography as a feasible, participant-facing creative service in urban public space.

LEAP linked autonomous viewpoint adjustment and image capture to the delivery of printed photographs, embedding camera control within an end-to-end service. The service therefore depended not only on camera movement, but also on participant entry, navigation around other people, successful print delivery and timely staff intervention. The supervised deployments allowed these interdependent functions to be evaluated together. Site-specific reporting captured differences in manoeuvring space, pedestrian interactions and supervision. This context-sensitive evaluation accords with the wider argument that autonomous urban mobility must adapt to local spatial and social practices [39].

Robotic photography also broadens the role commonly assigned to urban robots. Delivery, inspection and construction primarily transport objects, collect information or alter material environments. Photography instead produces an experience and a cultural object whose value is partly subjective. Machine participation in visual culture raises questions about how intelligent systems generate and transmit cultural products [40]. In urban settings, art and technology can contribute to culture, identity and innovation [41]. A mobile photography service could complement tourism and leisure by helping visitors record encounters with landmarks, waterfronts and parks. It could also support exhibitions, performances, festivals and community events through participant-controlled documentation. Camera position shapes how a person and place

are represented, making robot motion part of the creative output rather than merely a route to a destination.

A hands-free photography service can broaden participation in urban image-making. Solo visitors can adopt more varied poses, couples can document shared memories and groups can include every member. Active users, such as runners, can record their experiences without repeatedly handling a device. The service is also readily accessible to older adults, including those with mobility limitations, and requires no prior photographic expertise. Delivering these benefits at scale requires accessible interfaces and operating procedures. These should address interface height and reach, speech and touch alternatives, interaction pace, route accessibility, affordability and the availability of assistance. Accessibility therefore depends on the complete service design, not only on the mobility of the platform.

Photography intensifies interaction and governance requirements because it is directed towards people and produces identifiable images. The photographed person participates actively, while bystanders may enter the camera field without joining the service. Predictable, socially informed motion from the participant’s perspective can improve comfort more effectively than movement optimized only for efficiency in the robot’s space [42]. Legible signals should indicate when the service is active and where the camera is directed. These design features must be paired with participant initiation, affirmative consent, practical withdrawal, visible operator responsibility and clear retention and deletion procedures. The supervised protocol provides one participant-facing arrangement, while routine public operation will require site-specific governance of bystanders, data access, storage and accountability.

The controlled experiments identified capabilities needed beyond the deployment sites. Body pose, relative scene geometry and foreground extent contributed complementary information when appearance and setting changed. Across 30 photography tasks, LEAP completed the tested movements and produced higher data-driven quality scores than the two robotic approaches. Transfer to untrained indoor and outdoor settings matters because participants, backgrounds and lighting cannot be fixed in advance. Repeated observation also allowed the robot to recompose after participant movement, posture changes and altered navigation. Finally, the system retained alternative demonstrated viewpoints rather than reducing every session to one predetermined framing. These results connect the learning approach to practical requirements for a photography service: adaptation to place, recovery during interaction and variation in the creative outcome.

LEAP also has limitations that should be addressed in future work. First, several aspects of service feasibility require broader evaluation. The deployments covered three sites in Hong Kong and Shenzhen, used one robot configuration and were short relative to routine operation. Human supervision remained available throughout the participant-facing workflow, and the completeness of session-level records differed among sites. Longer deployments should recruit participants across demographic, cultural and mobility groups and use a common prospective protocol. They should quantify service uptake, intervention-free completion, social comfort, accessibility, long-term reliability and operating cost under different crowd densities, weather conditions and event types. Comparisons with practical alternatives, including tripods, selfie

sticks, companions and nearby photographers, would clarify the value of autonomy, attention saved and willingness to reuse the service.

Second, the photographic scope remains bounded. LEAP adjusts planar position, yaw and pitch, and bases composition mainly on body pose, relative depth and foreground shape. It does not explicitly reason about lighting, colour, facial-expression timing, lens choice, or camera height. Interaction is participant-initiated rather than conversational. Richer visual and language-based preferences could expand the supported compositions and help the robot follow individual intent. For public deployment, improvements in signalling, consent, accessibility and predictable motion are as important as a richer set of learned photography rules.

In conclusion, autonomous robotic photography establishes a new modality for creative and experiential urban services. LEAP realizes this vision by coupling learned photography rules with safety-aware navigation and obstacle avoidance within a supervised service workflow. The controlled experiments demonstrated autonomous completion, transfer to untrained settings, closed-loop recomposition and varied framing under the tested conditions. Combined with deployment feasibility across three distinct urban sites, these findings signal a future in which robots act as genuine creative partners in public life.

3 Methods

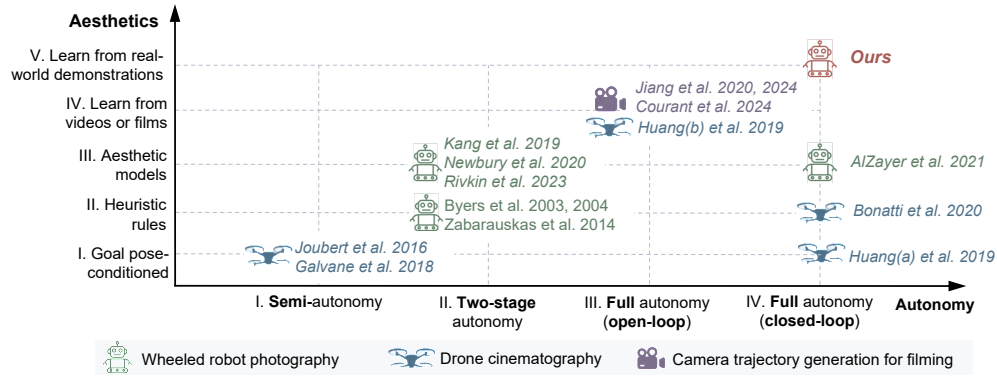


Fig. 6 Conceptual map of prior work in robotic photography. Studies are organized by degree of autonomy (horizontal axis) and the way aesthetic decisions are specified or learned (vertical axis). Muted green wheeled-robot icons denote wheeled robot photography, steel-blue drone icons denote drone cinematography, grey-purple camera icons denote camera-trajectory generation for filming, and the muted red robot denotes LEAP.

3.1 System overview

LEAP represents each view using skeletal keypoints, relative depth and a foreground segmentation mask. Skeletal keypoints encode body configuration and position within the image. Relative depth describes spatial relationships among the participant, nearby

objects and the background, while the mask delineates the full foreground boundary. This representation retains compositional structure while reducing dependence on clothing, texture and other appearance details. Conditional flow matching (CFM) represents alternative expert movements within a four-degree-of-freedom (4-DoF) camera space. After each movement, the robot observes the participant again and can recompose if the participant moves or changes pose, or if obstacle avoidance alters the executed route. We curated a motion-captured expert dataset to provide the paired visual observations and precise actions required for this closed-loop control. It contains 864 synchronized trajectories and egocentric videos collected with 16 actors, 50 poses and three starting directions (Fig. 1b).

LEAP integrates learned aesthetic viewpoint selection with physical execution through a fast-slow hierarchical architecture (Fig. 7). The high-level component decides where the camera should move, whereas the low-level component handles state estimation, path generation and collision avoidance. At robot timestep t , the system extracts an abstract compositional observation O_t from the current RGB image I_t . A high-level photography model $\pi(\cdot)$, trained on the expert dataset $\mathcal{D}$ through conditional flow matching (CFM), maps this observation and Gaussian noise ϵ to an H -step sequence of future egocentric waypoints, $\mathcal{A}_t = \{\Delta x, \Delta y, \Delta \Phi, \Delta \Psi\}_{1:H}$. These waypoints represent candidate changes in planar camera position, yaw and pitch relative to the current frame. Starting from ϵ , the model integrates a learned vector field for T denoising steps to recover a trajectory drawn from the demonstrated distribution rather than a single deterministic pose.

A receding-horizon strategy links the generated trajectory to robot movement. The l -th waypoint $[\mathcal{A}_t^1]_l$ is passed to a low-level motion planner $P(\cdot)$, which converts the target into collision-aware velocity commands (v, ω) for the wheeled-legged base and commands for the pan-tilt camera. Executing only part of the predicted sequence allows the system to reassess changes in subject pose, visibility and background before committing to the remaining waypoints. The robot then acquires a new image and predicts another trajectory. If obstacle avoidance changes the executed route, the next observation allows the photography model to revise its target from the robot’s actual position. This perceive-plan-act cycle continues until the model predicts near-zero translation and rotation, when the shutter is triggered. Algorithm 1 summarizes the complete procedure. Figure 6 situates LEAP within prior robotic-photography research.

3.2 Robotic platform and hardware architecture

We deployed LEAP on a LimX Dynamics Tron 1 bipedal wheeled-legged robot (Fig. 8a), following our earlier embodied photography system FotoBot [43]. The wheels provided stable rolling for image capture, while leg articulation supported traversal of uneven terrain. A custom top-mounted sensor suite integrated the sensing, actuation, communication and computation required for onboard operation (Fig. 8b).

An Autolabor C1 pan-tilt monocular camera provided egocentric RGB observations. The mobile base executed planar x, y translation and coarse yaw, whereas the camera controlled pitch and fine yaw, compensating for residual navigation error after the base reached a waypoint. A Livox Mid-360 LiDAR provided local 3D obstacle

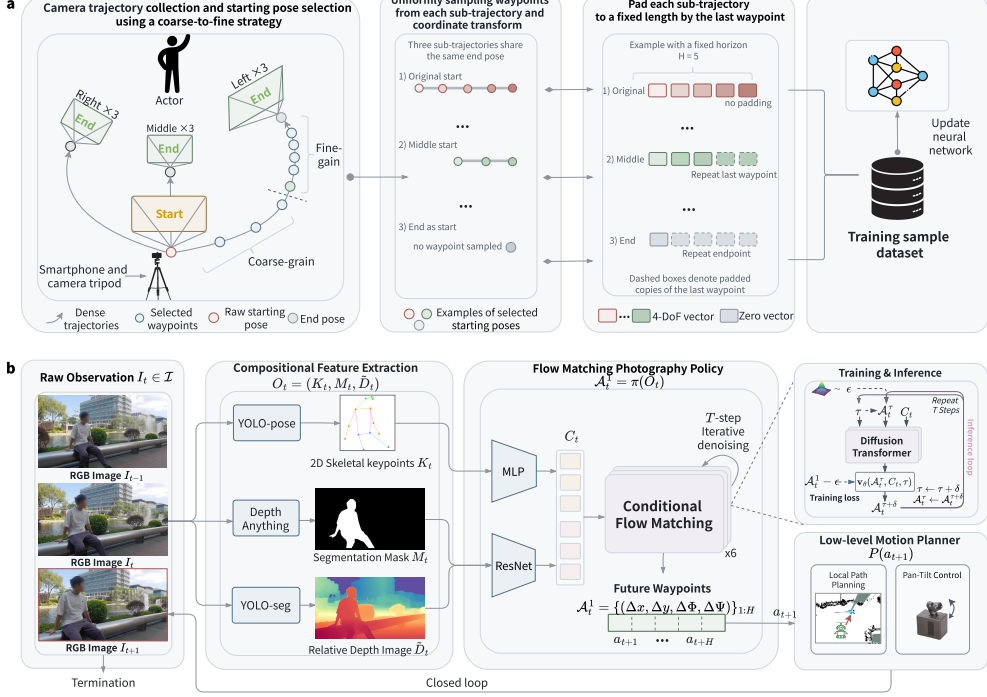


Fig. 7 Data generation and closed-loop execution of LEAP. (a) Expert demonstrations are divided into overlapping sub-trajectories through coarse-to-fine waypoint selection. Each sub-trajectory is transformed into the local frame of its initial pose and padded to a fixed horizon ($H = 5$). (b) At deployment, skeletal keypoints, relative depth and a foreground mask condition the flow-matching model. Iterative integration produces an H -step 4-DoF trajectory. A low-level planner executes the next waypoint, and observation and prediction repeat until the generated trajectory converges and the shutter is released.

observations. A RealSense D435i supplied metric depth only to the two comparison methods that required it; LEAP used monocular relative depth. Compositional feature extraction, CFM inference, state estimation and trajectory planning ran on an NVIDIA Jetson Orin NX edge computer rated at 157 tera operations per second (TOPS). A network switch connected all modules within the same local-area network. Onboard computation removed dependence on off-board inference during closed-loop control.

3.3 Expert demonstration data collection

Learning continuous photographic control requires expert motion paired with the images that motivate each action. Joystick teleoperation can introduce constraints, delay and jitter that alter natural camera movement. We therefore mounted a high-resolution smartphone on a tripod at a height of 95 cm, matching the robot camera (Fig. 8c). In a motion-capture (MoCap) laboratory, an expert photographer moved and rotated the apparatus around human subjects while using the smartphone screen

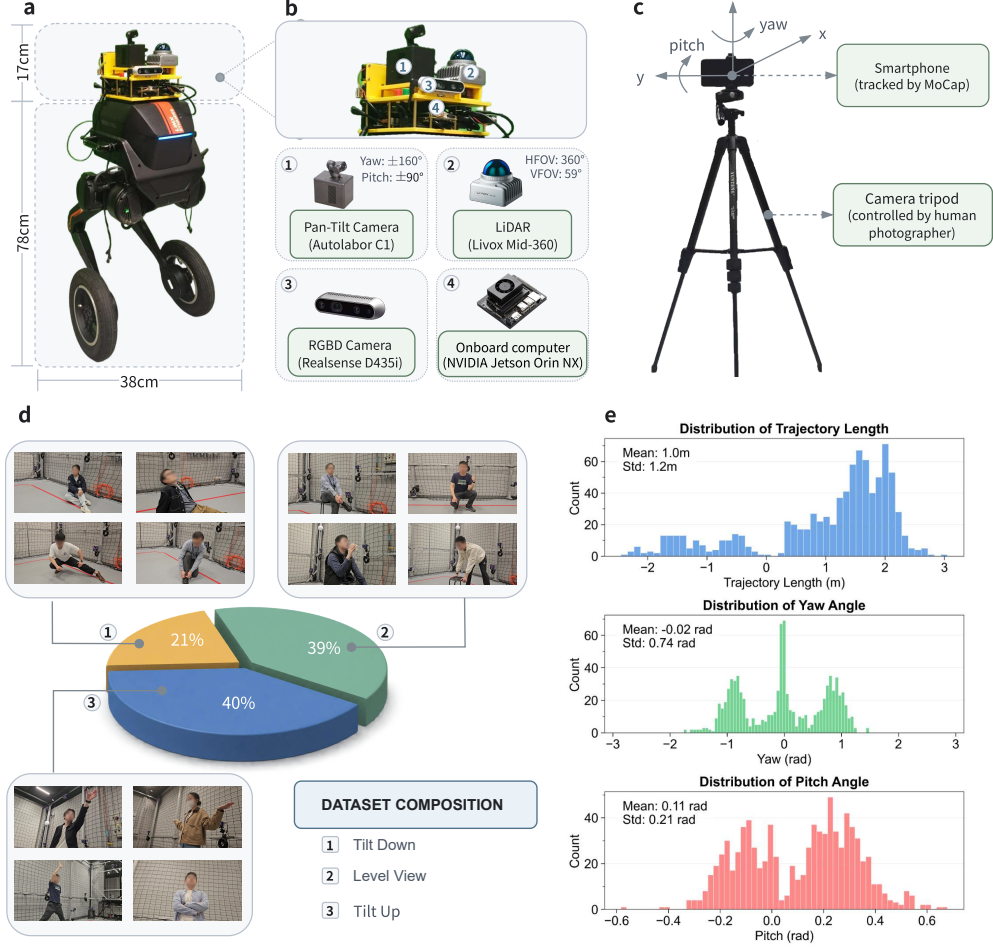


Fig. 8 Robot platform and expert demonstration dataset. (a) Integrated wheeled-legged photography robot. (b) Pan-tilt camera, LiDAR and onboard computer used for perception and execution. (c) Human-operated apparatus used to record synchronized 4-DoF camera motion and egocentric video. (d) Distribution of demonstrated camera pitch categories. (e) Distributions of trajectory length and camera rotation in the demonstration data.

as a viewfinder. Twenty-one Vicon Vero cameras recorded the apparatus position (x, y) with sub-centimetre accuracy, together with camera yaw and pitch, matching the four controllable degrees of freedom of the robot base and pan-tilt camera. Pose labels were synchronized with continuous 1080×1920 egocentric video to produce paired observation-action trajectories. During inference, the crop ratio $\gamma_f = \text{FOV}_1/\text{FOV}_2$ aligned the fields of view of the smartphone (FOV_1) and robot camera (FOV_2), preserving the demonstrated image-space composition across devices.

The primary dataset (Dataset *A*) contains 864 trajectories, or approximately 1.7 hours of recorded movement. This demonstration set is markedly compact and data-efficient relative to the 5–100 hours reported for adapting recent vision–language–action models [44], making coverage of pose and movement modes important. Sixteen adult actors aged 21–30 years and 158–183 cm tall performed 50 standing, sitting, squatting and lying poses. The demonstrations were grouped by camera-pitch intent as tilt-down, level-view and tilt-up (Fig. 8d) and approached from left, middle and right starting orientations, with several repetitions from each direction. These repetitions retained corrective movements and alternative routes to a final composition. A second photographer collected Dataset *B*, approximately 30% the size of Dataset *A*, with a preference for wider framing and longer shooting distances. Dataset *B* tested whether one generative model could represent more than one demonstrated framing preference. Trajectory-length, yaw and pitch distributions are summarized in Fig. 8e.

3.4 Compositional feature extraction and nonlinear depth normalization

LEAP converts each RGB observation into a three-part representation that retains posture, spatial relationships and foreground–background structure while reducing sensitivity to lighting, clothing and texture. The representation comprises three complementary cues: *skeletal keypoints*, a *foreground mask* and a *relative depth map*. First, a pose detector (YOLO-pose [45]) extracts skeletal keypoints $K_t \in \mathbb{R}^{17 \times 3}$ from the image I_t , encoding the image coordinates and visibility of 17 joints. These keypoints describe body configuration and position within the frame while reducing dependence on appearance. Second, a lightweight segmentation model (YOLO-seg [45]) produces a binary foreground mask M_t of the person, supplying explicit subject extent and boundaries. The mask is particularly useful when limbs are close together or framing depends on the area occupied by the whole body. Finally, a monocular depth estimator (Depth Anything V2 Base [46]) predicts a non-metric depth map D_t . This map represents spatial relationships among the foreground subject, nearby obstacles and the background.

Transfer to outdoor environments with open sky or distant landscapes was hindered by relative-depth distributions that differed markedly from those in the bounded MoCap laboratory. Standard linear normalization retains large variations in distant background depth. We therefore apply nonlinear depth normalization to reduce this indoor-to-outdoor distribution shift. The raw depth map is transformed using

$$\tilde{D}_t = \log(\lambda D_t + 1), \quad (1)$$

where $\lambda = 3$, followed by min–max scaling. This transformation compresses gradients far behind the subject while retaining relative depth changes in the region where camera motion occurs. It consequently reduces one source of distribution shift between the MoCap environment and the outdoor evaluation scenes.

The model input is $O_t = (K_t, M_t, \tilde{D}_t)$. A three-layer multilayer perceptron (MLP) projects the keypoints into a latent embedding. The depth map and mask are concatenated as image channels and encoded by a ResNet-18 backbone [47], allowing

foreground boundaries to be interpreted together with scene structure. A final MLP combines the keypoint and visual embeddings into the conditioning vector C_t . The feature extractors remain active during both training-data processing and deployment, exposing the model to their bounded estimation noise. Training on these imperfect signals discourages reliance on exact detector outputs and makes the observation pipeline consistent between recorded demonstrations and robot execution.

3.5 Generative imitation via conditional flow matching

At each timestep, the high-level photography model predicts an H -step action trajectory $\mathcal{A}_t = \{\Delta x, \Delta y, \Delta \Phi, \Delta \Psi\}_{1:H}$. The translations and rotations are expressed in the egocentric coordinate frame of the initial waypoint, so the model learns relative camera movement rather than absolute laboratory coordinates. More than one movement can yield an acceptable composition from a similar observation. For example, a photographer may move left or right while maintaining the subject at a comparable scale. Direct regression can average such distinct demonstrated movements into an action that corresponds to neither mode [48]. We therefore formulate trajectory prediction as a conditional generative problem using CFM [32].

CFM learns a vector field that transports an initial Gaussian sample $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to an expert trajectory $\mathcal{A}_t^1$, conditioned on O_t . Under the Optimal Transport formulation [49], the interpolated action at flow time $\tau \in [0, 1)$ follows the linear path

$$\mathcal{A}_t^\tau = \tau \mathcal{A}_t^1 + (1 - \tau) \epsilon, \quad (2)$$

with target vector field $\mathbf{u}(\mathcal{A}_t^\tau | \mathcal{A}_t^1) = \mathcal{A}_t^1 - \epsilon$. A network parameterized by θ is trained offline to predict $\mathbf{v}_\theta(\mathcal{A}_t^\tau, O_t)$ by minimizing the mean squared error (MSE) between the predicted and target vector fields over expert trajectories, observations, noise samples and flow times. The learned field defines a deterministic transport for each sampled ϵ , while repeated noise samples permit different action modes for the same observation.

The vector-field network uses a Diffusion Transformer (DiT) [50] rather than a convolutional action decoder. The compositional embedding C_t is combined with a sinusoidal representation of τ to condition the transformer. The interpolated noisy action is linearly projected into query tokens and combined with a learnable positional encoding $\text{PE}(\cdot)$ [51], preserving the temporal order of the H waypoints. Adaptive layer normalization with zero initialization injects the high-dimensional condition into the transformer blocks without a separate cross-attention module. The network returns one four-dimensional velocity for every waypoint token.

During deployment, the flow-matching model solves the neural ordinary differential equation from ϵ using an explicit forward Euler solver with T steps and $\delta = 1/T$:

$$\mathcal{A}_t^{\tau+\delta} = \mathcal{A}_t^\tau + \delta \mathbf{v}_\theta(\mathcal{A}_t^\tau, O_t). \quad (3)$$

Only waypoint l of the resulting sequence is executed before the next observation. Using a waypoint inside the predicted horizon filters very short action fluctuations while retaining closed-loop correction. A session terminates when $\|(\Delta x^*, \Delta y^*)\|_2 < \eta_t$ and $(|\Delta \Phi^*| + |\Delta \Psi^*|) < \eta_r$; both thresholds were set to 0.1 in the experiments. Repeating the demonstrated terminal waypoint when padding trajectories supplies

near-zero actions at completed compositions, allowing this condition to be learned from the data rather than from a manually specified final pose.

Two data-sampling procedures support termination and multimodality. Coarse-to-fine state sampling selects waypoints sparsely when the photographer is far from the final composition and densely during the small corrective movements near the end. Observation-mode alignment pairs geometrically similar starting observations with distinct left, middle or right demonstrated trajectories, rather than treating one direction as the only correct output. Network optimization, input degradation, shifted-Beta flow-time sampling, architectural settings and full hyperparameters are reported in the Supplementary Text.

3.6 Hierarchical system deployment and closed-loop execution

The learned photography model predicts compositional targets at a lower frequency, whereas the navigation system handles state estimation and collision avoidance at higher frequency. This separation prevents the learned model from directly producing high-frequency joint commands and leaves geometric obstacle response to a conventional local planner. FAST-LIO2 supplies LiDAR-inertial odometry [52], and a two-dimensional local cost map is continuously updated from the LiDAR point cloud. At 30 Hz, a Dynamic Window Approach planner samples admissible velocities and computes a collision-aware path towards the planar translation and coarse-yaw target. Velocity commands are limited to $v_{\max} = 0.5 \text{ m s}^{-1}$ and $\omega_{\max} = 2.5 \text{ rad s}^{-1}$ and are converted to joint torques by the locomotion controller. Coarse point-goal navigation requires approximately 3 s per predicted pose, after which the pan-tilt camera refines pitch and yaw before the next observation or capture.

If the proposed route intersects an obstacle, the local planner modifies the path using the cost map rather than executing the high-level waypoint exactly. The next image then reflects the robot’s altered position, allowing the photography model to revise its target. Thus, collision avoidance and aesthetic adjustment remain coupled through observation even though they operate in separate control modules. This repeated observation and correction produces the backtracking and viewpoint-adaptation behaviours evaluated in the experiments.

3.7 Controlled photography evaluation and outcome measures

The primary controlled evaluation comprised 30 physical photography tasks: 12 in-distribution tasks in the MoCap laboratory and eight indoor and ten outdoor out-of-distribution tasks across 14 previously untrained locations. Subjects adopted standing, sitting, squatting and lying poses, requiring changes in camera distance, lateral position and pitch. The indoor and outdoor scenes differed from the training laboratory in background structure, illumination and available navigation space. An independent set of 24 tasks (12 MoCap, six indoor and six outdoor) was used for the compositional-feature and demonstration-volume analyses, so these mechanism studies did not reuse the primary comparison suite.

Each method was initialized approximately 2 m in front of and directly facing the subject and was run three times per task. This standardized initialization reduced

variation attributable to different starting positions while still requiring each method to select its final viewpoint. Reported metrics are presented as mean $\pm$ 95% confidence interval across independent tasks. A trial was classified as a failure if the robot made physical contact with a subject or obstacle, or if the final image omitted or substantially occluded the subject’s face or upper body. Collision and incomplete shooting were retained as separate failure types in the underlying records.

We compared LEAP with two robotic-photography approaches selected to represent geometric and rule-based control. Bipedal Toric Space (BTS) is a semi-autonomous method that computes collision-free camera trajectories from user-specified geometric poses [8]. We used its deterministic face-centred configuration because the face is the coordinate origin of its target pose and because this framing represents many tracking-based photography systems. Visual Servoing with Rule of Thirds (VS-RoT) is a two-stage autonomous pipeline [12, 16]. It first estimates a frontal viewpoint using a depth camera, navigates to that viewpoint and then adjusts the image according to the Rule of Thirds using a proportional-derivative controller. We augmented frontal-view estimation with the active-perception module of Chen et al. [53] to support varied subject orientations. A professional photographer using the same camera height, sensor and 4-DoF movement constraints provided a human reference; this comparison represents performance under the study’s constrained action space rather than unrestricted professional photography.

Because no single measure represents photographic preference, we used three computational quality measures, one spatial-diversity measure and a blinded human evaluation. The computational and spatial measures were:

1. **Rule of Thirds (RoT) score.** The geometric measure [35] combines the normalized distance d_P from the face centre F to third-line intersections P_j and the distance d_L to third-lines L_j . With $\rho(d) = e^{-d^2/\gamma_d}$, $\gamma_d = 0.07$ and $\omega_{\text{point}} = 1/3$, the RoT score is defined as

$$S_{\text{RoT}} = \omega_{\text{point}} \rho\left(\min_j d_P(F, P_j)\right) + (1 - \omega_{\text{point}}) \rho\left(\min_j d_L(F, L_j)\right). \quad (4)$$

2. **View Evaluation Network (VEN).** This data-driven model was pretrained on large-scale human-annotated photographic composition data using contrastive learning [19]. It evaluates the complete image and returns a normalized score $S_{\text{VEN}} \in [0, 1]$, providing a learned alternative to the geometric RoT measure.
3. **CLIP-based aesthetic score.** Following Wang et al. [36], cosine similarities s_1 and s_2 between the photograph and the prompts “a good photo with a person” and “a bad photo with a person” are converted using temperature $\gamma_c = 0.01$:

$$S_{\text{CLIP}} = \frac{\exp(s_1/\gamma_c)}{\exp(s_1/\gamma_c) + \exp(s_2/\gamma_c)}. \quad (5)$$

4. **Compositional diversity.** Final camera poses were assigned to six spatial types defined by shooting distance and lateral position: Far-Left, Far-Middle, Far-Right, Close-Left, Close-Middle and Close-Right. We clustered the final poses and report

Algorithm 1 Closed-loop photography process of *LEAP*

Require: RGB image I_t , denoising steps T , look-ahead index l , feature extractors $\{\phi_K, \phi_D, \phi_M\}$, photography model $\pi(\cdot)$, motion planner $P(\cdot)$ and termination thresholds (η_t, η_r)

```
1:  $t \leftarrow 0, \delta \leftarrow 1/T$ 
2: while True do
3:    $I_t \leftarrow \text{getObs}()$ 
4:    $K_t \leftarrow \phi_K(I_t), M_t \leftarrow \phi_M(I_t), D_t \leftarrow \phi_D(I_t)$ 
5:    $\tilde{D}_t \leftarrow \sigma(D_t), O_t \leftarrow (K_t, M_t, \tilde{D}_t)$ 
6:    $\epsilon \sim \mathcal{N}(0, \mathbf{I}), \mathcal{A}_t^0 \leftarrow \epsilon, \tau \leftarrow 0$ 
7:   for  $n = 1$  to  $T$  do
8:      $\mathcal{A}_t^{\tau+\delta} \leftarrow \mathcal{A}_t^\tau + \delta \mathbf{v}_\theta(\mathcal{A}_t^\tau, O_t)$ 
9:      $\tau \leftarrow \tau + \delta$ 
10:  end for
11:   $(\Delta x^*, \Delta y^*, \Delta \Phi^*, \Delta \Psi^*) \leftarrow [\mathcal{A}_t^1]_l$ 
12:  if  $\|(\Delta x^*, \Delta y^*)\|_2 < \eta_t$  and  $(|\Delta \Phi^*| + |\Delta \Psi^*|) < \eta_r$  then
13:    capture photograph; break
14:  else
15:    move to next waypoint using planner  $P$ 
16:  end if
17:   $t \leftarrow t + 1$ 
18: end while
```

the mean number of distinct types produced across repeated trials from the same initial condition. This measure captures variation in spatial composition rather than aesthetic quality.

Blinded aesthetic evaluation. Twenty-four participants, including 17 photography enthusiasts, evaluated anonymized images through an [online website](#). Method labels and capture-order cues were withheld, and image options within each question were randomly ordered for every participant. Q1–Q3 used three ordered categories (good, medium and bad), whereas the primary overall-aesthetic outcome, Q4, used a nine-point Likert-type scale. Kendall’s coefficient of concordance quantified inter-rater agreement for Q1–Q3.

Compositional-feature and demonstration-volume experiments. The full representation and four variants (Image, Kp, Kp+Depth and Depth+Mask) were evaluated on the independent 24-task suite. The demonstration-volume analysis compared five fractions of Dataset A, from 20% to 100%, and a combined Dataset A+B condition. Each condition used one trained model and three physical rollouts per task.

In the controlled experiments, metrics were computed for all captured images, including failed trials, so that a method did not benefit from excluding physically unsuccessful runs. For a collision, the last frame before contact was used. VEN and CLIP were not designed specifically for human-centred photography and can assign non-zero scores to images that do not contain a usable portrait. Images in which any keypoint on the face, shoulders or arms was missing were classified as *incomplete*

framing and assigned a score of zero. Further baseline implementation details and the RoT metric formulation are provided in the Supplementary Text.

3.8 Urban deployment protocol and outcome definitions

The service was deployed at a campus orientation day, at a senior-centre recreational event and in Shenzhen Talent Park. The campus event served students and parents, the senior-centre event served older adults, and the park served local residents and visitors. Deployment dates, operating hours and site-level outcomes are reported in Table 2.

For prospective recording and retrospective coding where records permit, a session begins when a consenting participant accepts the service and the robot begins observation. It ends at photograph delivery, participant withdrawal, operator termination or a pre-specified timeout. A group request counts as one session, and retries before delivery remain within that session. Withdrawn, timed-out and operator-terminated sessions remain in the denominator.

Staff recruited participants, obtained consent, initiated sessions, supervised operation and supported image transfer, printing and delivery. Photographs were printed on site using a portable Canon SELPHY CP1500 printer. LEAP performed observation, viewpoint generation, collision-aware movement, pan-tilt adjustment, convergence-based termination and shutter release. Predefined staff roles were not counted as interventions; unplanned assistance to perception, navigation, aesthetic control or capture was counted.

Deployment exposure comprises total on-site time and autonomous operating time, excluding setup, charging, maintenance and other unavailable periods. We report initiated sessions, captured and delivered photographs, and sessions per autonomous hour with explicit denominators. End-to-end completion requires observation, viewpoint adjustment, capture, transfer or processing, and delivery. Intervention-free completion additionally requires no unplanned operator assistance.

Session records retain site and timing, participation and consent status, stage-level outcomes, interventions, failures and safety events. Unsuccessful sessions receive a primary category of perception, navigation, aesthetic control, human interaction, hardware or infrastructure, or safety termination. Safety interventions include takeover, emergency stop or cancellation because continuation was judged unsafe; collisions are reported separately. Recorded operating conditions include crowd density, lighting, weather, available space, occlusion and background complexity.

3.9 Statistical analysis

For the controlled physical experiments, each method was run three times per task, and computational metrics are reported as mean $\pm$ 95% confidence interval across independent tasks. When scene-level means are summarized across environments, they are weighted by the number of tasks in each environment: 12 motion-capture, eight indoor and ten outdoor tasks in the 30-task comparison, and 12, six and six tasks, respectively, in the 24-task compositional-feature and demonstration-volume experiments. Relative differences are calculated from these task-weighted means before

rounding. Failure outcomes are retained as counts and rates with their denominators. The blinded evaluation reports Q4 as the scalar overall-quality outcome, Q1–Q3 as ordinal rating distributions and Kendall’s W as inter-rater agreement. No rating from the urban deployments is combined with this blinded photographic-quality evaluation.

Deployment results will be summarized by site rather than pooled across incompatible protocols. Rates will include raw numerators and denominators, with Wilson or exact binomial intervals when sample sizes permit. Session duration and intervention counts will be reported using medians and interquartile ranges when their distributions are skewed. Site and condition comparisons will be made only when definitions and recording procedures are compatible.

3.10 Ethics, privacy, consent and data handling

Offline participation and public deployment were approved by the Human Research Ethics Committee of the University of Hong Kong (reference EA250531). Participants provided informed consent under the approved protocol. The online $N = 24$ aesthetic evaluation was analytically separate from the deployment study.

Incidental bystander footage was restricted to documenting system operation, and informal deployment comments were used as formative service-design observations. Recognizable participant images and supplementary footage will be reported only where research use and publication permission are documented. Other identifiable material will be cropped, sufficiently anonymized or excluded. Raw recordings containing identifiable people will not be deposited openly unless the consent and institutional data-governance conditions permit this; non-identifiable session summaries and aggregate outcomes will be provided as source data.

Supplementary information

The Supplementary Information includes Supplementary Text, Table [A1](#), Figures [A1](#), [A2](#) and [A3](#), Movies [S1](#) and [S2](#) with their captions, the ethics approval letter from the Human Research Ethics Committee (HREC) of HKU and the informed-consent form used for offline data collection and public-deployment interactions. The movies show system operation, closed-loop recovery, pose adaptation and the participant-facing deployment workflow.

Acknowledgements

We thank Chengyu Deng for assistance with expert-demonstration collection and for taking reference photographs in the performance-comparison experiment, and Lei Yang for discussions and comments. We thank Zejia Liu, Chen Wang, Wenbo Dong, Wentao Chen, Kin-Fung Chu, Hao Zheng, Yizhou Chen, Sihan Chen, Zeqing Zhang, Jimmy Ng, Hubert Fung, and Haibing He for acting in data collection and real-world experiments. We thank Haobo Zuo and Shanghang Li for assisting with the deployments in the senior-center event and campus orientation day. We also thank Shuang Wu, Yanhui Yang, Xiangwen Chen, Chengyu Deng, Yuhan Xie, Guangze Zheng, Tran Hoang Nguyen, Taisarinov Yerassyl, Eason Liu, Wentao Chen, Sihan Chen, Meng Lin,

ZeJia Liu, Can Yang, Guowei Du, Youze Zhang, Wenxin Yue, Chen Wang, Qi Chen and additional anonymous volunteers for participating in the online evaluation. Chat-GPT, DeepSeek and Gemini assisted with language revision, color style and layout of the figures; the authors verified all manuscript content. We upscaled first-person-view images by a factor of two using [PixelCut](#) and [Upscayl](#) to improve presentation clarity.

Declarations

- Funding: This work was partially supported by the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative and by the Research Grants Council (Ref: 17200924). Chang Chen was funded by the Shenzhen Loop Area Institute (SLAI).
- Conflict of interest/Competing interests: There are no competing interests to declare.
- Ethics approval and consent to participate: The Human Research Ethics Committee (HREC) approved the ethical aspects of the offline data-collection and public-deployment protocol (reference number: EA250531; approval expiration date: December 7, 2029). We obtained informed consent from all offline participants after explaining the nature and potential consequences of the study. The HREC approval letter and informed-consent form are provided in the Supplementary Information. The online blinded aesthetic evaluation was conducted separately from this offline protocol.
- Consent for publication: All authors consent to publication. Identifiable participant images or footage will be included only where publication permission has been documented; materials without the required permission will be cropped, sufficiently anonymized or excluded before submission.
- Data availability: Processed controlled-evaluation results, participant-level anonymized aesthetic ratings, source data underlying the main and supplementary figures, non-identifiable deployment summaries and the public components of the LEAP training dataset will be deposited in Zenodo at <https://doi.org/10.5281/zenodo.21888808>. Access to the training dataset is tiered because the original RGB videos contain potentially identifiable visual information. Public derived data and privacy-filtered videos, and the controlled-access procedure for the original RGB videos are described in the Supplementary Text [A.2](#). Raw urban-deployment images, video and audio containing identifiable participants or incidental bystanders will not be released openly unless the applicable consent and publication permissions allow it. The deposited record will provide aggregate session data and metadata sufficient to audit the reported deployment outcomes. Access to restricted recordings will follow the HREC approval and institutional data-governance requirements.
- Code availability: Code will be made available through the [project GitHub repository](#) upon publication.
- Author contributions: C.C. formulated the main ideas, collected the dataset, designed the training and inference algorithms, constructed the robotic software system, conducted all experiments, analyzed and interpreted the results, designed all

figures and demonstration videos, and wrote the paper. D.W. helped with formulating the main ideas, discussion, data collection, real-world experiments, and paper revision. Z.L. helped with the real-world experiments, data analysis, and figure revision. Y.P. contributed to sensor design and hardware development. I.J. acted as the photographer to collect the training data and invited participants to engage in the user study. H.X., Mathieu Chambe, X.L. and H.C. provided suggestions on the research design and experimental evaluation and helped revise the paper. M.C. provided insights into the problem formulation, data collection, algorithm development, experiment design, and evaluation, and helped with paper revision. J.P. proposed the initial research idea and scope, supervised the project, provided insights into the experiment and demonstration-video design, and helped revise the paper.

Appendix A Supplementary Text

A.1 Hyperparameters

The main text reports the training and deployment protocol at a conceptual level; this section lists the implementation settings used in our experiments. The network was trained using AdamW [54] with a learning rate of 10^{-4} , 100 linear warm-up steps and an exponential moving average (EMA) update strategy. Batch-normalization layers in the ResNet-18 visual encoders were replaced with GroupNorm. The feature encoders used LeakyReLU activations with layer normalization, and the Diffusion Transformer used GELU activations. Training was performed on a single NVIDIA GeForce RTX 4090 GPU for 500 epochs with a batch size of 32. During training, depth maps and segmentation masks were downsampled to 45×80 , while a scale jitter with a ratio of 0.03 was applied to the depth maps, and the semantic-mask boundaries were smoothed by morphological erosion with a kernel of size 5. Isotropic Gaussian noise $\mathcal{N}(0, 0.03^2 I)$ was injected into normalized skeletal keypoints and depth-map background regions with 50% probability. For flow matching, timesteps $\tau \in [0, 1]$ were sampled from a shifted beta distribution $\mathcal{B}(\tau) = \text{Beta}(\frac{s-\tau}{s}; 1.5, 1)$ with shift parameter $s = 0.999$.

A.2 Dataset access and privacy protection

The LEAP training dataset uses tiered access to protect participant privacy while supporting reproducibility. The public dataset at <https://doi.org/10.5281/zenodo.21888808> contains synchronized camera and robot trajectories, calibration data, annotations and dataset splits. It also includes pose tracks extracted from the original videos and privacy-filtered videos. Original RGB videos may contain identifiable visual information and are available only through controlled access. Requests should be submitted to changchen@connect.hku.hk and will be reviewed within ten working days. Approved users must sign a data-use agreement prohibiting re-identification, participant contact, redistribution and public display. The agreement also requires secure storage and deletion after completion of the approved project. Access does not carry an authorship requirement.

Table A1 Key hyperparameters for model training, closed-loop deployment, and evaluation metrics

Variable	Description	Value
λ	coefficient for nonlinear depth normalization	3
n_{epoch}	total training epochs	500
n_{layer}	number of transformer decoder layers	6
n_{head}	number of transformer attention heads	4
n_{hidden}	hidden dimension size of the transformer decoder	256
p_{drop}	dropout rate of the transformer decoder	0.1
lr	initial learning rate	10^{-4}
H	horizon length of predicted future actions	5
T	number of flow matching timesteps during inference	5
l	look-ahead index of the waypoint executed per replanning cycle	2
η_t	autonomous termination threshold for translation	0.1
η_r	autonomous termination threshold for rotation	0.1
γ_f	crop ratio derived from FOV_1 and FOV_2	0.647
γ_d	decay coefficient in the Rule of Thirds metric	0.07
γ_c	temperature parameter in the CLIP metric softmax	0.01
v_{max}	maximum linear velocity of low-level execution module	0.5 m s^{-1}
ω_{max}	maximum angular velocity of low-level execution module	2.5 rad s^{-1}

A.3 Data processing

After dataset collection, we discarded low-quality segments and smoothed the recorded dense trajectories. To construct training data for closed-loop model learning, we further augmented each demonstration by extracting multiple sub-trajectories from intermediate waypoints along the original path. Specifically, several waypoints were selected as candidate starting poses, and for each starting pose, the remaining trajectory was uniformly downsampled into a fixed-length H -step future sequence. Although these sub-trajectories shared the same terminal target, they originated from different spatial locations and viewpoints. Each sub-trajectory was then transformed into the local coordinate frame of its starting pose and padded to length H by repeating the final state. In this way, each demonstration yielded a large set of aligned H -step training samples. In addition, we treated the endpoint of each original trajectory as a special starting pose and assigned an all-zero future trajectory as the ground-truth label. These processed sub-trajectories therefore represent the desired future motion conditioned on different intermediate observations along the same demonstration.

In practice, we found that two data-processing strategies were particularly important for enabling autonomous termination and diverse trajectory generation.

Coarse-to-fine sampling for autonomous termination: To improve the model’s ability to recognize when the task should terminate, we adopted a nonuniform sampling strategy over the starting waypoints. Waypoints far from the goal were sampled sparsely, whereas those near the endpoint were sampled more densely. This increases the proportion of near-terminal training samples and helps the model associate goal-like visual observations with near-zero future motion. During deployment,

the model can therefore identify observations resembling the terminal states seen during training and terminate autonomously. Empirically, we set the transition between coarse and fine sampling at a distance of 1 m from the endpoint.

Observation-mode alignment for generative diversity: To encourage multimodal predictions, such as approaching the subject from the left, centre or right, the model must observe nearly identical initial visual conditions paired with distinct future motion styles. Small errors in the starting camera pose can otherwise cause the model to overfit to pose discrepancies rather than learn genuine multimodality. We therefore augmented the dataset with ground-truth trajectories that originated from nearby locations but followed different directional modes. These trajectories were transformed into the local coordinate frame of the reference starting pose, aligning several stylistically distinct trajectories with the same initial observation.

A.4 Model training and optimization strategies

The generative photography model was optimized to map high-dimensional visual observations to continuous spatial action distributions. To bridge static dataset modelling and dynamic closed-loop deployment, we implemented targeted architectural and training optimizations.

Architecture and optimization protocol. The network was trained using AdamW [54] with stabilization strategies for conditional generative models. Batch-normalization layers in the ResNet-18 [47] visual encoders were replaced with Group-Norm to improve stability with small visual batches and support exponential moving average (EMA) weight updates. The feature encoders used LeakyReLU activations with layer normalization, whereas the Diffusion Transformer used GELU activations [55]. Detailed hyperparameters, training hardware and deployment thresholds are reported above.

Robustness through data augmentation. To reduce the perceptual domain gap and sensitivity to noise from downstream vision detectors, we degraded inputs during training. Depth maps and segmentation masks were downsampled to suppress high-frequency distractors, and semantic-mask boundaries were smoothed by morphological erosion. We also injected isotropic Gaussian noise into normalized skeletal keypoints and depth-map background regions, encouraging the model to learn compositional structure rather than exact pixel values. The augmentation parameters are reported above.

Shifted-beta timestep sampling for flow matching. To bias the network towards minimizing errors in highly corrupted states, represented by lower timesteps and noisier action samples, we used a shifted-beta sampler instead of uniform timestep sampling. This strategy improved denoising during the initial high-variance phase of trajectory generation; the distributional parameters are reported above.

Strategic sampling for autonomy and diversity. Training-data curation used two sampling strategies to support autonomous termination and multimodal trajectory prediction. First, *coarse-to-fine state sampling* sampled states distant from the terminal viewpoint sparsely and states near it densely. This exposed the model to near-zero-velocity actions around completed compositions and reduced reliance on an

externally specified stopping pose. Second, *observation-mode alignment* paired geometrically similar initial states with different expert directional modes, such as framing the subject from the left, centre or right. The model could therefore represent several demonstrated actions for similar observations instead of one averaged trajectory.

A.5 Conditional flow-matching formulation

At robot timestep t , the target action $\mathcal{A}_t^1$ is an H -step sequence of relative planar translations and camera rotations. Conditional flow matching learns a vector field that transports Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to the expert action distribution while conditioning on the compositional observation O_t . Using the optimal-transport path,

$$\mathcal{A}_t^\tau = \tau \mathcal{A}_t^1 + (1 - \tau) \epsilon, \quad \tau \in [0, 1],$$

the target vector field is $\mathcal{A}_t^1 - \epsilon$. A Diffusion Transformer [50] predicts $\mathbf{v}_\theta(\mathcal{A}_t^\tau, O_t)$ and is optimized to match this field. At inference, explicit Euler integration with step size $\delta = 1/T$ generates the action sequence:

$$\mathcal{A}_t^{\tau+\delta} = \mathcal{A}_t^\tau + \delta \mathbf{v}_\theta(\mathcal{A}_t^\tau, O_t).$$

Only the selected look-ahead waypoint is executed before the image is observed again. The loop terminates when the predicted translation and rotation fall below η_t and η_r , respectively. Table A1 reports the trajectory horizon, integration steps, look-ahead index and termination thresholds.

A.6 Baseline implementation

For both baselines, we used a RealSense D435i camera to provide the depth maps needed to compute shooting poses.

Bipedal Toric Space (BTS): Following our prior work, we adapted the semi-autonomous drone-based method [8] to single-person photography with the bipedal robot. The YOLO-pose model detected the face centre, which was transformed from the image plane to the camera frame P_c using depth Z_c and camera intrinsics $\mathbf{K}$. This 3D face-centre position served as the origin for manual input. Users specified shooting poses by distance, yaw and pitch relative to the origin, and the same low-level planner used by LEAP generated collision-aware trajectories.

Visual Servoing with Rule of Thirds (VS-RoT): We first calculated a frontal viewpoint using the active-perception module of Chen et al. [53]. The module detected the frontal body plane, used its normal vector as the viewing direction and placed the target viewpoint 2m from the plane. After the robot reached this viewpoint, image-based visual servoing adjusted the composition according to the Rule of Thirds. A proportional-derivative (PD) controller sent yaw and pitch angular velocities to the pan-tilt camera, using the RoT score as feedback until the system converged. The proportional gain K_p and derivative gain K_d were 0.0015 and 0.0008, respectively.

A.7 Metric calculation

Rule of Thirds (RoT): We provide the detailed formulation for the RoT metric used in our evaluation. The metric relies on the normalized Manhattan distance between two 2D image coordinates, defined as:

$$d((x_1, y_1), (x_2, y_2)) = \frac{|x_1 - x_2|}{w} + \frac{|y_1 - y_2|}{h},$$

where h and w denote image height and width, respectively. Let F denote the 2D coordinate of the detected face centre. Spatial deviation from the ideal RoT composition is evaluated pointwise and linewise. The pointwise distance $d_P(F, P_j)$ is the normalized Manhattan distance between F and the j -th standard intersection P_j of the third lines. The linewise distance $d_L(F, L_j)$ is the normalized orthogonal distance from F to the j -th third line L_j . The composite score S_{RoT} is a weighted sum of these components:

$$S_{\text{RoT}} = \omega_{\text{point}} \cdot \rho \left(\min_j d_P(F, P_j) \right) + (1 - \omega_{\text{point}}) \cdot \rho \left(\min_j d_L(F, L_j) \right).$$

In our experiments, the pointwise weight ω_{point} was set to 1/3, and the decay coefficient γ_d was set empirically to 0.07.

A.8 Network architecture

The complete network structure is presented in Fig. [A1](#).

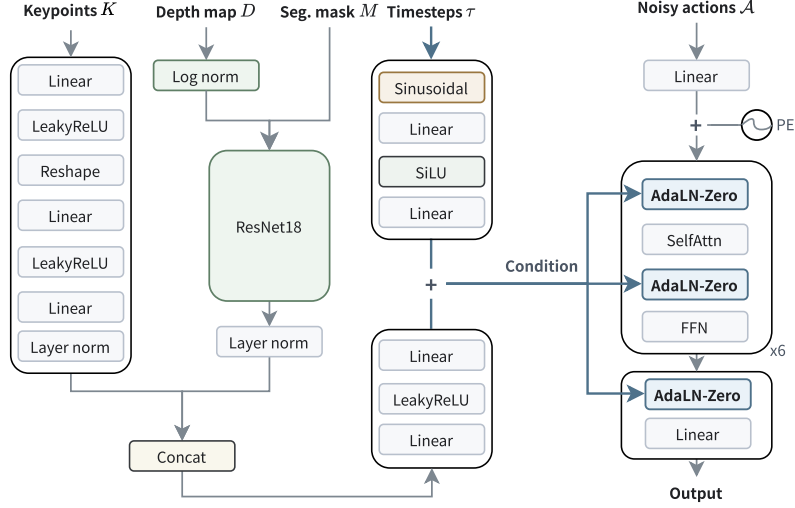
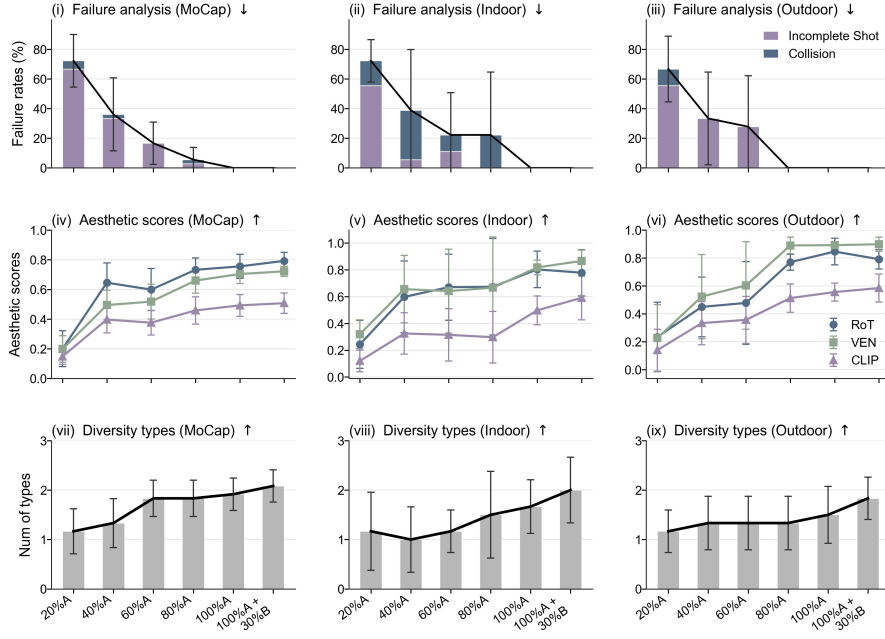


Fig. A1 Network architecture of the learned photography model. **(Left)** The perception module uses an MLP to encode skeletal keypoints and a shared ResNet-18 encoder to encode relative depth and the segmentation mask. **(Right)** The generative backbone is implemented as a Diffusion Transformer (DiT). Adaptive layer normalization with zero initialization (AdaLN-Zero) provides stable conditioning. Abbreviations and symbols: *Segmentation mask* denotes the foreground segmentation mask; *PE* represents learnable positional embeddings; *SelfAttn* denotes the self-attention mechanism; *FFN* indicates the feed-forward network within each standard transformer decoder block; and + signifies element-wise embedding addition.

A.9 Demonstration volume and photographer preference

a Results of scaling effect



b Learned style preferences A and B

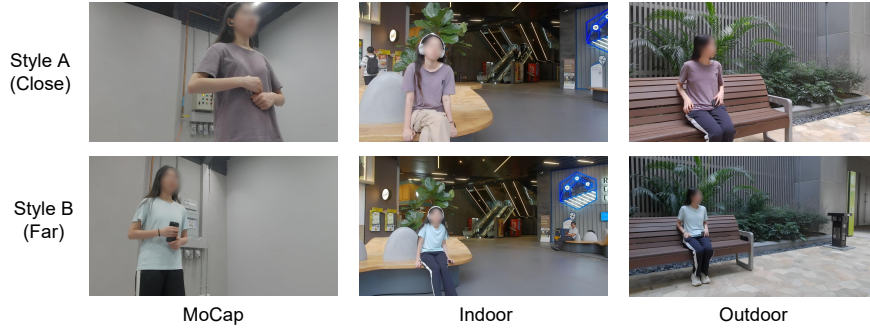


Fig. A2 Effect of demonstration volume and source on the photography model. (a) Performance of models trained with increasing fractions of Dataset A, from 20% to 100%, and a model trained with Dataset A plus Dataset B. Outcomes across motion-capture, untrained indoor and untrained outdoor tasks include physical failure rate, computational aesthetic scores and compositional diversity. (b) Representative close- and long-distance framing patterns associated with the preferences in Datasets A and B.

A.10 Image gallery for experiment demonstrations

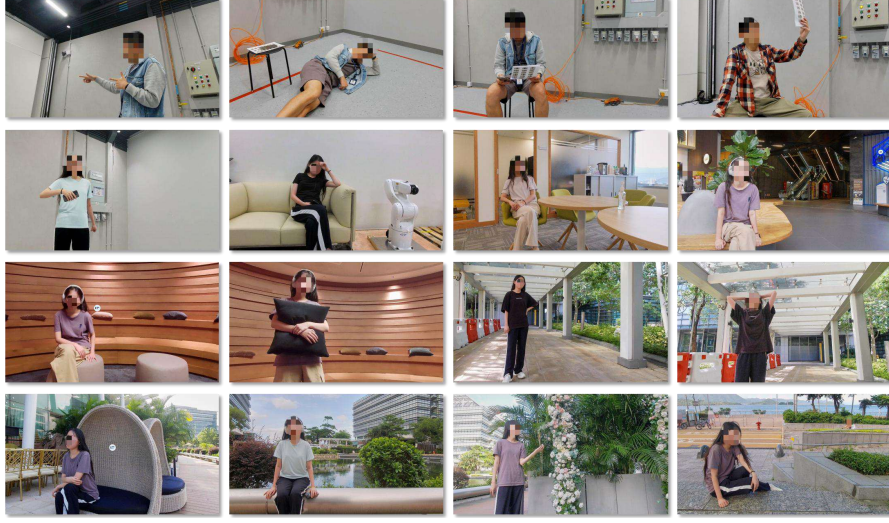


Fig. A3 Illustrative image gallery from the controlled photography evaluation. Selected photographs captured by LEAP across indoor and outdoor tasks show variation in subject pose, body scale and environmental context. These examples complement the aggregate results of the 30-task evaluation.

A.11 Captions for movies

Movie S1: Real-world evaluation and closed-loop behaviours of LEAP for autonomous human-centred photography

This video illustrates the learned photography model operating in a closed loop across indoor and outdoor environments. The movie is organized into several segments. (0) Motivation for the work. (1) *System pipeline*: Vision-to-action inference and physical camera actuation. (2) *Comparative performance*: LEAP, BTS, VS-RoT and the constrained professional-photographer reference under the study conditions. (3) *Recovery and pose adaptation*: Selected recoveries from degraded initial viewpoints and responses to subject movement. (4) *Continuous indoor execution*: Sequential photography tasks with varied subject postures and body scales. (5) *Long-horizon outdoor execution*: Tracking and photography in a waterfront setting, including obstacle avoidance and variation in framing. Together, the sequences visualize physical execution, transfer to untrained settings, closed-loop recomposition and compositional diversity.

Movie S2: One-day deployment in Shenzhen Talent Park

Beyond the controlled evaluation, the supervised system operated for one day in Shenzhen Talent Park and delivered printed photographs to consenting participants. The movie follows the participant-facing workflow from session initiation and autonomous viewpoint adjustment to image capture and print delivery.

References

- [1] Lee, J. *et al.* Learning robust autonomous navigation and locomotion for wheeled-legged robots. *Science Robotics* **9**, eadi9641 (2024).
- [2] Ri, S., Ye, J., Toyama, N. & Ogura, N. Drone-based displacement measurement of infrastructures utilizing phase information. *Nature Communications* **15**, 395 (2024).
- [3] Johns, R. L. *et al.* A framework for robotic excavation and dry stone construction using on-site materials. *Science Robotics* **8**, eabp9758 (2023).
- [4] Kaur, A. *et al.* Functional brain imaging and population-level visits to urban spaces. *Nature Cities* **1**, 880–887 (2024). URL <https://doi.org/10.1038/s44284-024-00158-x>.
- [5] Bajcsy, R. Active perception. *Proceedings of the IEEE* **76**, 966–1005 (1988).
- [6] Joubert, N. *et al.* Towards a drone cinematographer: Guiding quadrotor cameras using visual composition principles. *arXiv preprint arXiv:1610.01691* (2016).
- [7] Galvane, Q., Fleureau, J., Tariolle, F.-L. & Guillotel, P. *Automated cinematography with unmanned aerial vehicles* (2016).
- [8] Galvane, Q. *et al.* Directing cinematographic drones. *ACM Transactions on Graphics (TOG)* **37**, 1–18 (2018).
- [9] Lino, C. & Christie, M. Intuitive and efficient camera control with the toric space. *ACM Transactions on Graphics (TOG)* **34**, 1–12 (2015).
- [10] Grill, T. & Scanlon, M. *Photographic Composition* (Amphoto, New York, 1990).
- [11] Krages, B. *Photography: The Art of Composition* (Allworth, New York, 2005).
- [12] Byers, Z., Dixon, M., Goodier, K., Grimm, C. M. & Smart, W. D. *An autonomous robot photographer*, Vol. 3, 2636–2641 (2003).
- [13] Byers, Z., Dixon, M., Smart, W. D. & Grimm, C. M. Say cheese! experiences with a robot photographer. *AI magazine* **25**, 37–37 (2004).
- [14] Campbell, J. & Pillai, P. *Leveraging limited autonomous mobility to frame attractive group photos*, 3396–3401 (2005).
- [15] Ahn, H. *et al.* A robot photographer with user interactivity, 5637–5643 (2006).
- [16] Zabarauskas, M. & Cameron, S. *Luke: An autonomous robot photographer*, 1809–1815 (2014).

- [17] Murray, N., Marchesotti, L. & Perronnin, F. *Ava: A large-scale database for aesthetic visual analysis*, 2408–2415 (2012).
- [18] Kong, S., Shen, X., Lin, Z., Mech, R. & Fowlkes, C. *Photo aesthetics ranking network with attributes and content adaptation*, 662–679 (2016).
- [19] Wei, Z. *et al.* *Good view hunting: Learning photo composition from dense view pairs*, 5437–5446 (2018).
- [20] Newbury, R., Cosgun, A., Koseoglu, M. & Drummond, T. *Learning to take good pictures of people with a robot photographer*, 11268–11275 (2020).
- [21] Kang, H. *et al.* Lerop: A learning-based modular robot photography framework. *arXiv preprint arXiv:1911.12470* (2019).
- [22] Rivkin, D. *et al.* *Ansel photobot: A robot event photographer with semantic intelligence*, 8262–8268 (2023).
- [23] Huang, C. *et al.* *Learning to film from professional human motion videos*, 4244–4253 (2019).
- [24] Jiang, H., Wang, B., Wang, X., Christie, M. & Chen, B. Example-driven virtual cinematography by learning camera behaviors. *ACM Transactions on Graphics* **39**, 45 (2020).
- [25] Courant, R., Dufour, N., Wang, X., Christie, M. & Kalogeiton, V. *Et the exceptional trajectories: Text-to-camera-trajectory generation with character awareness*, 464–480 (2024).
- [26] Jiang, H., Wang, X., Christie, M., Liu, L. & Chen, B. *Cinematographic camera diffusion model*, Vol. 43, e15055 (2024).
- [27] Zhang, M. *et al.* *Gendop: Auto-regressive camera trajectory generation as a director of photography* (2025).
- [28] Huang, C. *et al.* *Learning to capture a film-look video with a camera drone*, 1871–1877 (2019).
- [29] Bonatti, R. *et al.* Autonomous aerial cinematography in unstructured environments with learned artistic decision-making. *Journal of Field Robotics* **37**, 606–641 (2020).
- [30] AlZayer, H., Lin, H. & Bala, K. *Autophoto: Aesthetic photo capture using reinforcement learning*, 944–951 (2021).
- [31] Ho, J., Jain, A. & Abbeel, P. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems* **33**, 6840–6851 (2020).

- [32] Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M. & Le, M. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747* (2022).
- [33] Soukupova, T. & Cech, J. *Eye blink detection using facial landmarks*, Vol. 2, 4 (2016).
- [34] Lin, H. *et al.* Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647* (2025).
- [35] Liu, L., Chen, R., Wolf, L. & Cohen-Or, D. Optimizing photo composition. *Computer Graphics Forum* **29**, 469–478 (2010).
- [36] Wang, J., Chan, K. C. & Loy, C. C. *Exploring clip for assessing the look and feel of images*, Vol. 37, 2555–2563 (2023).
- [37] Mercado, G. *The filmmaker’s eye: Learning (and breaking) the rules of cinematic composition* (Routledge, New York, 2013).
- [38] Kendall, M. G. & Babington Smith, B. The problem of m rankings. *The Annals of Mathematical Statistics* **10**, 275–287 (1939).
- [39] Mubiru, I. Autonomous mobility beyond formal cities. *Nature Cities* **3**, 298–299 (2026). URL <https://doi.org/10.1038/s44284-026-00422-2>.
- [40] Brinkmann, L. *et al.* Machine culture. *Nature Human Behaviour* **7**, 1855–1868 (2023). URL <https://doi.org/10.1038/s41562-023-01742-2>.
- [41] Nature Cities. Echoes of art in urban spaces. *Nature Cities* **2**, 359 (2025). URL <https://doi.org/10.1038/s44284-025-00246-6>.
- [42] Zhou, C. *et al.* Human-behaviour-based social locomotion model improves the humanization of social robots. *Nature Machine Intelligence* **4**, 1040–1052 (2022). URL <https://doi.org/10.1038/s42256-022-00542-z>.
- [43] Wang, D. *et al.* Fotobot: an embodied ai photography robot system its design, prototyping and application. *Robot Learning* **3** (2026).
- [44] Black, K. *et al.* π_0 : A vision-language-action flow model for general robot control (2025).
- [45] Jocher, G. & Qiu, J. Ultralytics yolo11 (2024). Software version 11.0.0. Available at <https://github.com/ultralytics/ultralytics>.
- [46] Yang, L. *et al.* Depth anything v2, 21875–21911 (2024).
- [47] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition, 770–778 (2016).

- [48] Chi, C. *et al.* Diffusion policy: Visuomotor policy learning via action diffusion. *The International Journal of Robotics Research* **44**, 1684–1704 (2023).
- [49] McCann, R. J. A convexity principle for interacting gases. *Advances in Mathematics* **128**, 153–179 (1997).
- [50] Peebles, W. & Xie, S. *Scalable diffusion models with transformers*, 4195–4205 (2023).
- [51] Vaswani, A. *et al.* *Attention is all you need* (2017).
- [52] Xu, W., Cai, Y., He, D., Lin, J. & Zhang, F. Fast-lío2: Fast direct lidar-inertial odometry. *IEEE Transactions on Robotics* **38**, 2053–2073 (2022).
- [53] Chen, C. *et al.* Signage-aware exploration in open world using venue maps. *IEEE Robotics and Automation Letters* **10**, 3414–3421 (2025).
- [54] Loshchilov, I. & Hutter, F. *Decoupled weight decay regularization* (2019).
- [55] Hendrycks, D. & Gimpel, K. Gaussian error linear units (gelus) (2023). URL <https://arxiv.org/abs/1606.08415>. [arXiv:1606.08415](https://arxiv.org/abs/1606.08415).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [3indoorimmersivedemo.mp4](#)
- [EA250531AwardLetter.pdf](#)
- [HREConsentForm.docx](#)
- [2comparisonXrobustness.mp4](#)
- [1pipeline.mp4](#)
- [0intro.mp4](#)
- [4outdoorimmersivedemo.mp4](#)
- [5onedaydeployment.mp4](#)