SPOT: Spatial Perception-Oriented Long-Horizon Humanoid Teleoperation

Anonymous Author(s)
Affiliation
Address
email

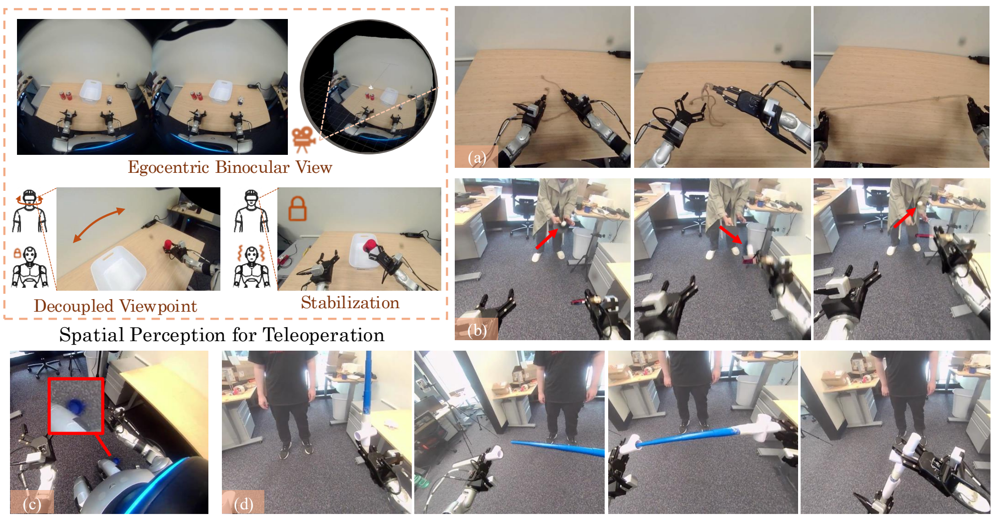


Figure 1: SPOT enhances spatial perception through: binocular egocentric visualization, viewpoint decoupling, and stabilization. It enables tasks including: (a) bimanual knot tying; (b) highly dynamic object interception; (c) dropped-object relocation; and (d) large-workspace bimanual manipulation with flexibility and ease of use.

Abstract: High-quality demonstration data is becoming a central bottleneck for training general-purpose humanoid robots. While recent humanoid teleoperation systems have made substantial progress in retargeting human motion to robot motion, long-horizon loco-manipulation requires another capability: operators must maintain task-relevant spatial awareness over time, e.g., object locations, surrounding environments, the robot's pose. We call the extent of this awareness the operator's *perceptual horizon*. However, existing methods often shorten this: narrow views miss peripheral events, robot-mounted cameras become unstable during locomotion, and coupled head-view control makes looking around interfere with robot motion. We present **SPOT**, a Spatial Perception-Oriented VR Teleoperation system for collecting long-horizon humanoid demonstration data by providing extended perceptual horizon. SPOT combines a robot-mounted binocular fisheye camera, a wide-field stereoscopic display, viewpoint-decoupled free-looking, and visual stabilization to provide a robot-centric view that is wide, stable, and actively inspectable. Unlike conventional egocentric interfaces, SPOT decouples visual exploration from robot actuation: the egocentric stereo observation is rendered on a virtual hemisphere around the operator, so natural head rotations change where the operator looks within the wide-field view rather than commanding the robot head, camera, or torso. We evaluate SPOT on perception-critical humanoid data-collection tasks spanning drop recovery, peripheral retrieval, large-workspace bimanual manipulation, fine alignment, and dynamic interaction. SPOT improves efficiency, accuracy, recovery speed, and operator workload, demonstrating its effectiveness for user-friendly and scalable long-horizon humanoid data collection.

Keywords: Humanoid, Teleoperation, Robot Learning, Data Collection

Submitted to the 10th Conference on Robot Learning (CoRL 2026). Do not distribute.

1 Introduction

High-quality demonstration data has become critical for training humanoid robots. As humanoids move from short manipulation episodes toward long-horizon loco-manipulation, teleoperation is increasingly used to collect data involving mobility, bimanual interaction, recovery from unexpected events, and precise spatial alignment [1, 2, 3, 4, 5]. Existing humanoid teleoperation systems have made substantial progress in transferring human motion to robot motion through motion tracking [3, 6, 7, 8, 9], retargeting [10, 11, 12, 13, 14, 15], and whole-body control [16, 17, 18, 19, 20].

However, reliable long-horizon teleoperation requires more than accurate motion transfer or retargeting. A central requirement in long-horizon loco-manipulation is that operators need to maintain task-relevant spatial awareness over time, including the locations of objects, the surrounding environment, robot's limb configurations, robot's pose relative to the scene, and so on. We refer to the spatial and temporal extent over which such awareness can be reliably maintained as the operator's *perceptual horizon*. When such perception horizon is limited, the operator may still be able to issue motion commands, but task execution during the teleoperation becomes less efficient and less convenient: objects must be repeatedly searched for, recovery from unexpected events becomes slower, precise alignment becomes more difficult, etc.

Existing egocentric teleoperation interfaces [1, 5, 6, 7] suffer from limited perceptual horizon in several ways. Narrow camera views may miss peripheral events that are important for task execution, such as objects falling outside the central view or robot limbs entering the workspace from the side. Robot-mounted cameras can also become unstable during locomotion, balancing, and recovery, making it difficult for the operator to maintain a consistent spatial reference. In addition, when visual exploration is coupled to robot actuation, head motion used for looking around may unintentionally command the robot's head, camera, or torso. These limitations are more pronounced in long-horizon teleoperation, where operators must continuously gather and update spatial information while simultaneously controlling locomotion and manipulation; See Fig. 2.

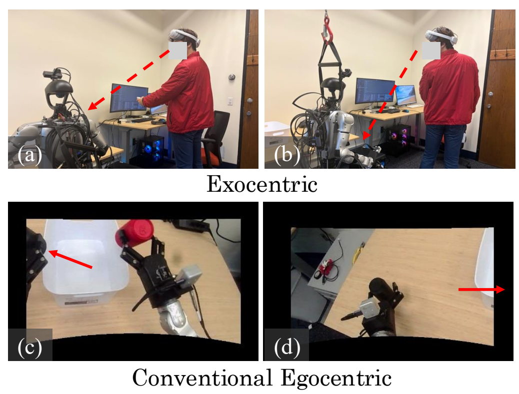


Figure 2: **Limitations of existing teleoperation interfaces.** In exocentric interfaces [3, 4, 11, 21], the separation between the operator and the robot (a) weakens depth perception and introduces additional difficulties when the operator turns (b). Conventional egocentric interfaces [1, 5, 7] may hide manipulators (c) and lose track of objects outside the camera FOV (d).

In this work, we present **SPOT** (**S**patial **P**erception-**O**riented Long-Horizon Humanoid **T**eleoperation), a perception-centered VR teleoperation system for long-horizon humanoid loco-manipulation. Unlike prior systems that primarily optimize motion retargeting [11, 13] and treat egocentric video as passive feedback [6, 7], SPOT treats spatial perception as an explicit interface design objective. The system combines a robot-mounted binocular fisheye camera, a wide-field stereoscopic display, viewpoint-decoupled free-looking, and visual stabilization to provide a robot-centric view that is wide, stable, and actively inspectable. This design aims to extend the operator's perceptual horizon by preserving access to task-relevant spatial information over longer time horizons and larger operating ranges. A key design choice in SPOT is to decouple visual exploration from robot actuation. This differs from active-camera egocentric settings, such as TWIST2-style interfaces [1], where the operator's head motion is used to change the robot-mounted camera viewpoint and visual exploration is therefore coupled to robot-side motion. In contrast, SPOT renders the egocentric stereo observation on a virtual hemisphere surrounding the operator, rather than rigidly attaching the view to the headset or directly mapping it to robot head, camera, or torso motion. As a result, natural head rotations only change where the operator looks within the wide-field observation, without commanding the robot head, camera, or torso. In parallel, visual stabilization compensates for camera rotations induced by robot motion, allowing the operator to perceive a more consistent spatial frame during locomotion and recovery.

We evaluate SPOT on a suite of perception-critical humanoid loco-manipulation tasks, including unexpected drop recovery, peripheral object retrieval, wide-range bimanual manipulation, fine spatial alignment, and dynamic interaction. These tasks are designed to stress different aspects of spatial perception, including object reacquisition, peripheral awareness, large-workspace coordination, and stable perception under robot motion. Compared with teleoperation paradigms inspired by recent humanoid data-collection systems, SPOT improves task efficiency and accuracy, reduces search and recovery time, and lowers operator workload. These results suggest that perception transfer is an important complement to motion transfer for scalable long-horizon humanoid teleoperation.

In summary, our contributions are threefold:

- We formulate the operator's perceptual horizon as a practical bottleneck in long-horizon humanoid teleoperation, complementary to motion transfer.
- We introduce SPOT, a perception-centered VR teleoperation system with a wide-field, stereoscopic, stabilized, and viewpoint-decoupled robot-centric visual interface.
- We evaluate SPOT on perception-critical humanoid loco-manipulation tasks and show that improving spatial perception improves task efficiency, accuracy, recovery, and operator workload.

2 Related Works

Bimanual teleoperation Teleoperation studies emerged in a bimanual setting, where two arms are attached steadily to a static or moving platform. Bimanual teleoperation systems like [22, 23, 24] build an isomorphic bi-arm structure under direct human operation and map the kinematics to the actuator arms. Other works use low-cost exoskeletons [25, 26] to collect human arm trajectories. Earlier VR/AR headsets also allowed relative hand position tracking, making room for VR-based bimanual teleoperation [27, 28, 29].

Humanoid teleoperation Humanoid robots' teleoperation imposes significantly greater challenges on accuracy and efficiency. Collecting [30] and retargeting whole-body joints [31, 32] is harder than bi-arm joints, as it's infeasible to have an isomorphic humanoid under the operator's direct control. [11, 21, 33, 34] relied on motion capturing (MoCap) systems to achieve the high-fidelity tracking necessary for whole-body control. [6] uses exoskeleton interfaces and [35] adopts inertial motion capture equipment for whole-body teleoperation. More recently, VR has become the dominant interface for capturing operator motion [1, 2, 3, 4, 5, 7, 12] due to its accessibility.

3 Methods

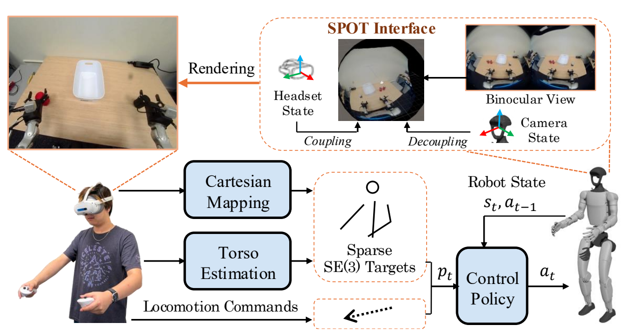


Figure 3: **Overview of the system architecture.**

We present SPOT, a perception-centered teleoperation framework designed to enhance operator spatial perception during long-horizon humanoid loco-manipulation. The key design principle of SPOT is to improve spatial perception transfer from the robot to the human operator, enabling efficient environmental understanding, object localization, and interaction while preserving intuitive teleoperation control. To achieve this goal, SPOT combines a robot-centric immersive perception interface with a VR-based teleoperation pipeline. The perception interface provides binocular visual feedback through a stabilized wide-field stereoscopic display, allowing operators to freely explore the environment while maintaining a visually consistent representation of the world. In parallel, sparse observations from a VR headset and handheld controllers are converted into robot control targets and executed by a target-conditioned control policy. Figure 3 illustrates the overall system architecture.

3.1 Immersive Perception Interface

Egocentric Binocular View We provide the operator with an immersive egocentric observation of the robot environment through a robot-mounted binocular fisheye camera. The camera stream is received in Unity [36] as a side-by-side (SBS) stereo video texture, and then rendered inside the VR headset through a hemispherical display surface. For each rendered viewing direction $\mathbf{d} \in \mathbb{S}^2$ in the front hemisphere, we project the direction into the corresponding fisheye image using an equisolid projection $r(\theta) = R \cdot \sin(\theta/2)/\sin(\theta_{\max}/2)$, where $\theta = \arccos(d_z)$, $\theta_{\max}$ is the maximum visible polar angle, and R is the normalized radius of the valid fisheye region. The left and right halves of the SBS image are sampled independently according to the rendered eye, preserving binocular disparity in the headset. Directions outside the forward-facing fisheye support are masked out. Compared with displaying a conventional planar camera image, this representation preserves wide peripheral coverage and binocular depth cues, which are particularly important in our tasks where relevant objects or robot limbs may enter the operator's peripheral field of view.

Viewpoint-Decoupled Observation The egocentric observation is rendered on a virtual hemisphere surrounding the operator rather than being rigidly attached to the operator's headset view. Consequently, the operator can naturally rotate their head to inspect different portions of the robot's wide-field observation without directly commanding a corresponding robot camera or head motion. This decouples visual exploration from robot actuation: the operator's head motion controls where they look within the currently available binocular observation, while robot commands are obtained separately from the tracked headset and handheld-controller signals. This design is especially suitable for our setting, where the robot does not require an actuated neck to provide an immersive observation interface.

Visual Stabilization Because the stereo camera is mounted on the moving robot, disturbances and recovery motions can induce abrupt camera rotations that make raw egocentric video difficult to interpret. To maintain a stable visual reference, we compensate for camera motion directly in the virtual display. Let $\mathbf{R}_t^c$ denote the camera orientation at time t and $\mathbf{R}_0^c$ the orientation at initialization. The rendered observation is rotated according to the relative camera motion $\mathbf{R}_t^{\mathrm{rel}} = (\mathbf{R}_0^c)^{-1}\,\mathbf{R}_t^c$, such that the rotational motion of the physical camera is counteracted in the displayed observation. Consequently, the operator perceives a visually stabilized scene despite camera movement induced by robot locomotion, balancing, or external interference.

The camera orientation estimate can be obtained from any source that provides the camera pose with respect to a global reference frame, such as robot state or an independent sensing device on the camera rig. In our implementation, we employ a lightweight IMU [37] attached to the stereo camera and use its orientation estimate to drive the stabilization transform.

3.2 Teleoperation Pipeline

Human Motion Acquisition We acquire sparse human motion commands through an OpenXR-compatible VR headset and two handheld controllers. The interface is implemented using OpenXR [38] and is therefore hardware-agnostic across compliant VR platforms. We have validated the system on both Meta Quest [39] and PICO [40] headsets without modifying the teleoperation pipeline.

In each Unity update cycle, the interface queries the tracked poses of the head-mounted display and the left and right controllers, yielding three rigid-body transforms: $\mathcal{X}_t^{\mathrm{human}} = \left\{\mathbf{T}_t^H, \mathbf{T}_t^L, \mathbf{T}_t^R\right\}$, where H, L, and R denote the headset, left controller, and right controller, respectively. In addition to these poses, we read the two-dimensional joystick inputs and controller-button states. The acquired signals are serialized into a lightweight UDP message and streamed to the downstream teleoperation process at a target frequency of $50 \sim 90$ Hz. This interface requires only a headset and handheld controllers: the tracked head and hand poses provide the sparse upper-body motion specification, while joystick inputs are used to issue locomotion commands. In addition, the left and right controller triggers are directly mapped to the corresponding robot gripper commands.

Upper-body Estimation To decouple robot torso control from the operator's free head motion, we do not directly map the headset orientation to the robot torso. Instead, we estimate the operator's torso orientation from a short history of the three tracked VR poses using a lightweight streaming network. The network is trained on VR motion sequences with torso orientations recorded by a body-mounted IMU, while deployment requires only the headset and two handheld controllers. Given the predicted torso orientation, we recover the torso position from the tracked headset position using fixed headset-to-neck and neck-to-torso offsets. Together with the directly observed controller poses, this yields sparse upper-body targets consisting of the estimated torso and two hands. This estimation allows the operator to freely explore the environment through natural head motion while independently commanding robot body orientation.

Cartesian Retargeting Following the Cartesian-space formulation of ExtremControl [33], we directly map the estimated human upper-body poses to robot-side SE(3) targets without online inverse kinematics or joint-space retargeting. In contrast to the full-body setting, our interface maps only the torso and two wrists, while locomotion is commanded separately through the controller joystick. A one-shot neutral-pose calibration determines the initial heading alignment, controller-to-wrist rotational offsets, and human-side arm proportions. At runtime, each hand target is expressed in the torso frame and rescaled according to the corresponding human-to-robot arm-length ratio:

$$\mathbf{p}_{t,k}^{r,\mathrm{rel}} = \mathbf{s}_k^r + \frac{l_k^r}{l_k^h} \left(\mathbf{p}_{t,k}^{h,\mathrm{rel}} - \mathbf{s}_k^h \right), \qquad k \in \{L, R\}, \tag{1}$$

where $\mathbf{s}_k$ denotes the shoulder anchor and l_k denotes arm length. The relative hand orientations are transferred after applying the calibrated offsets, and the resulting torso and wrist poses are provided directly as Cartesian targets to the downstream control policy.

Target-conditioned Control Policy Following ExtremControl [33], we train a target-conditioned policy in simulation that tracks Cartesian hand targets and locomotion commands. The policy receives robot proprioception, target commands, and recent action history, and is trained with standard motion-tracking objectives and domain randomization for deployment robustness.

4 Experimental Results

4.1 Experimental Setup

We evaluate our teleoperation interface on a set of long-horizon tasks that require continuous spatial awareness, object monitoring, and coordinated locomotion and manipulation.

We compare our method against these representative teleoperation baselines:

- **Exocentric Teleoperation** which provides a third-person external observation similar to prior exocentric humanoid teleoperation systems such as OmniH2O [3], TWIST [11], and SONIC [4]. In this setting, the operator controls the robot while observing the robot and workspace from an external viewpoint.
- **Conventional Egocentric Teleoperation** which uses a standard onboard stereo first-person camera view (e.g., ZED Mini [41]) through the VR headset, including Open-Television [2],

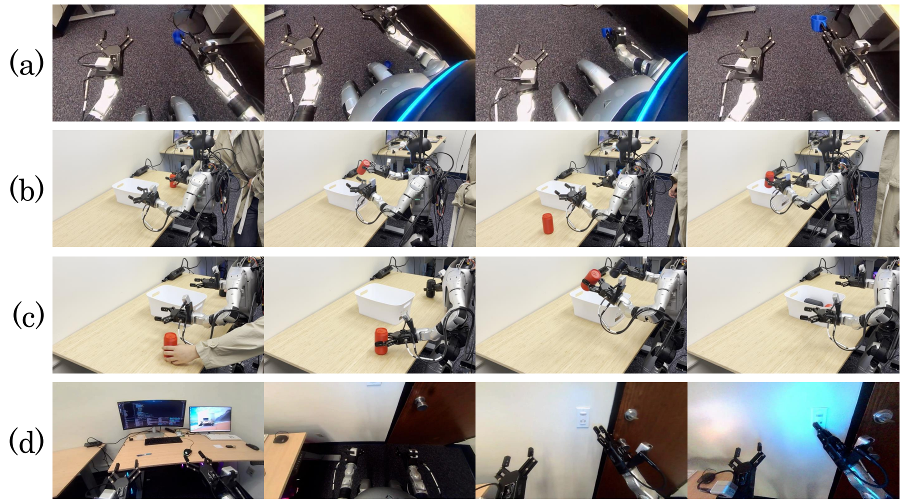


Figure 4: **Experimental task definitions.** (a) unexpected drop and recovery, (b) peripheral object retrieval, (c) multi-object bimanual retrieval, and (d) light switch.

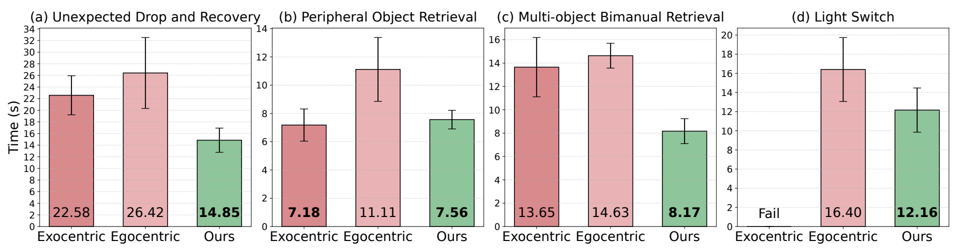


Figure 5: **Task completion time.** Our interface consistently improves efficiency on recovery, peripheral awareness, bimanual coordination, and long-horizon loco-manipulation tasks.

AMO [5], and TWIST2 [1]. In this setting, we allow the operator to freely rotate the viewing direction through head motion, and simulate the corresponding viewpoint changes similar to their robot-mounted active camera setup. However, the observable region remains constrained by the limited onboard camera field of view.

To ensure fair comparison, all methods share the same robot platform, teleoperation pipeline, task environments, and protocols; only the perception interface differs. We report task completion time across methods. Two operators (one experienced and one novice) perform 20 trials per task.

We evaluate four quantitative tasks covering different spatial perception challenges:

- **Unexpected Drop and Recovery**, as shown in Fig. 4(a). The robot transports an object that is unexpectedly dropped during locomotion. The operator must rapidly localize, recover, and continue the task.
- **Peripheral Object Retrieval**, as shown in Fig. 4(b). A target object randomly appears outside the central viewing region. The operator must detect and retrieve it as quickly as possible.
- **Multi-object Bimanual Retrieval**, as shown in Fig. 4(c). Two target objects appear simultaneously on opposite sides of the robot. Operators must retrieve both while continuously monitoring both manipulators and task-relevant objects.
- **Light Switch**, as shown in Fig. 4(d). The robot starts facing away from a wall-mounted light switch and must locomote toward and precisely toggle it.

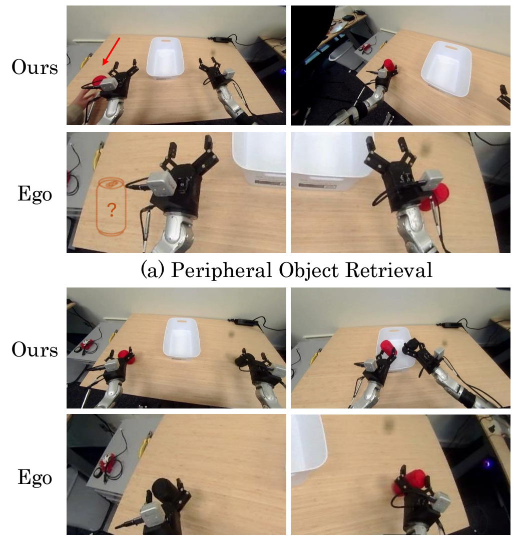
4.2 Awareness of Unexpected Events and Out-of-View Objects

We evaluate this capability using the Unexpected Drop and Recovery task and the Peripheral Object Retrieval task. Fig. 5(a) and Fig. 5(b) report the completion times of these tasks.

The exocentric baseline preserves global scene visibility but suffers from reduced spatial precision during manipulation, especially when the manipulating space is not as ideal as a tabletop setup. In contrast, the conventional egocentric baseline provides immersive first-person feedback but loses awareness of objects once they leave the narrow onboard camera field of view. It suffers from locating out-of-view objects when they come into the manipulation workspace, and requires extra search time, as shown in Fig. 6(a).

Our method consistently achieves higher success rates and lower completion times across both tasks. The immersive wide-field interface enables operators to rapidly re-localize dropped or peripheral objects while maintaining robot-centered spatial consistency throughout the task.

We also show in Fig. 1(b) that the operator can dynamically locate a ping-pong ball during teleoperation and return it to the opponent in real time. This demonstrates that our interface is capable of tracking dynamic objects across a large field of view.



(a) Peripheral Object Retrieval

(b) Multi-object Bimanual Retrieval

Figure 6: **Comparison between our interface (Ours) and conventional egocentric teleoperation (Ego).** The limited FOV in conventional egocentric interfaces (a) requires additional search and re-localization time, and (b) requires sequential instead of simultaneous manipulation.

4.3 Bimanual Spatial Awareness

We evaluate the ability to simultaneously monitor both manipulators and multiple task-relevant objects during coordinated bimanual interaction by using the Multi-object Bimanual Retrieval task. Fig. 5(c) summarizes the quantitative results.

The conventional egocentric baseline loses visibility of one manipulator when the operator focuses on the other hand, particularly during large-arm motions or object transfer. The operator needs to switch focus and complete the task on each manipulator one at a time, as shown in Fig.6. Meanwhile, although the exocentric baseline preserves visibility of both arms, the detached external viewpoint weakens depth perception and precise spatial judgment during manipulation. In our experiments, the operator also experiences significantly reduced manipulation accuracy on the arm farther from the viewing position.

Our method improves the task efficiency by enabling continuous awareness of both manipulators and nearby objects within a coherent immersive spatial representation.

We further present several challenging qualitative demonstrations that highlight additional bimanual capabilities. Fig. 1(a) shows the robot performing rope knotting while continuously monitoring both manipulators and rope configuration, highlighting long-horizon bimanual spatial awareness. Fig. 1(d) shows the robot inserting a toy sword into a sheath during large-range dual-arm motions, requiring simultaneous global spatial awareness and precise bimanual alignment.

4.4 Spatial Consistency Across Locomotion and Manipulation

We further evaluate spatial perception during transitions between navigation and precise manipulation using the Light Switch task. This task requires operators to maintain environmental awareness during locomotion while also performing accurate final-stage interaction. Fig. 5(d) shows the quantitative comparison.

The exocentric baseline frequently fails because the operator loses awareness of the robot after reorientation. Conventional egocentric teleoperation preserves immersion but requires additional viewpoint realignment after robot turning, since the visual frame rotates together with the robot during navigation.

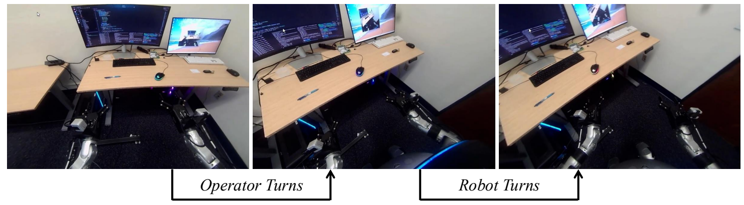


Figure 7: **Stabilized egocentric perception during turning motions.** When the operator turns in VR, their view turns accordingly. When the robot turns while the operator stops, the coordinate frame in VR remains stabilized, preserving spatial continuity and reducing abrupt viewpoint changes.

Our method maintains stable robot-centered spatial perception across both locomotion and manipulation stages. As shown in Fig. 7, the viewpoint in VR decouples with robot motion and follows the head motion solely, avoiding abrupt viewpoint changes during teleoperation, which results in less viewpoint alignment time and improves task completion efficiency and reliability.

5 Conclusion

We presented SPOT, a spatial perception-oriented VR teleoperation system for long-horizon humanoid data collection. Motivated by the observation that effective humanoid teleoperation requires not only accurate motion transfer but also sustained task-relevant spatial awareness, we introduced the notion of the operator's *perceptual horizon*. SPOT extends this perceptual horizon through a wide-field stereoscopic robot-centric interface that combines binocular fisheye sensing, viewpoint-decoupled free-looking, and visual stabilization. By decoupling visual exploration from robot actuation, SPOT allows operators to actively inspect the environment without interfering with robot motion, enabling more flexible and user-friendly control during long-horizon loco-manipulation. Across perception-critical humanoid tasks involving drop recovery, peripheral retrieval, large-workspace bimanual manipulation, fine alignment, and dynamic interaction, SPOT improves task efficiency, accuracy, recovery speed, and operator workload compared with conventional egocentric teleoperation interfaces. These results indicate that perception transfer is a key complement to motion transfer for scalable humanoid data collection. More broadly, SPOT highlights spatial perception as an explicit interface design objective for teleoperation: improving how operators perceive, search, and maintain task-relevant spatial context can improve the quality, consistency, and scalability of long-horizon humanoid demonstrations.

6 Limitations and Future Work

Our current system primarily focuses on visual perception through immersive stereo observation. However, effective teleoperation also depends on other sensing modalities, including tactile feedback, force interaction, and audio. Incorporating additional sensory channels into immersive humanoid teleoperation remains an important future direction. Although the proposed interface improves operator spatial awareness, end-to-end camera streaming latency remains a major bottleneck for responsive teleoperation. Delays introduced by image transmission can reduce responsiveness during rapid manipulation or recovery behaviors. While our system remains usable under current latency conditions, further reducing visual feedback latency will likely be critical for highly responsive whole-body teleoperation. While SPOT improves spatial awareness and teleoperation performance across loco-manipulation tasks, we have not yet evaluated its impact on large-scale downstream policy learning. Understanding how improved operator perception affects the scalability and quality of learned policies remains an important direction.

References

[1] Y. Ze, S. Zhao, W. Wang, A. Kanazawa, R. Duan, P. Abbeel, G. Shi, J. Wu, and C. K. Liu. Twist2: Scalable, portable, and holistic humanoid data collection system. *arXiv preprint arXiv:2511.02832*, 2025.

[2] X. Cheng, J. Li, S. Yang, G. Yang, and X. Wang. Open-television: Teleoperation with immersive active visual feedback. *arXiv preprint arXiv:2407.01512*, 2024.

[3] T. He, Z. Luo, X. He, W. Xiao, C. Zhang, W. Zhang, K. Kitani, C. Liu, and G. Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024.

[4] Z. Luo, Y. Yuan, T. Wang, C. Li, S. Chen, F. Castaneda, Z.-A. Cao, J. Li, D. Minor, Q. Ben, et al. Sonic: Supersizing motion tracking for natural humanoid whole-body control. *arXiv preprint arXiv:2511.07820*, 2025.

[5] J. Li, X. Cheng, T. Huang, S. Yang, R.-Z. Qiu, and X. Wang. Amo: Adaptive motion optimization for hyper-dexterous humanoid whole-body control. *arXiv preprint arXiv:2505.03738*, 2025.

[6] Q. Ben, F. Jia, J. Zeng, J. Dong, D. Lin, and J. Pang. Homie: Humanoid loco-manipulation with isomorphic exoskeleton cockpit. *arXiv preprint arXiv:2502.13013*, 2025.

[7] Y. Li, Y. Lin, J. Cui, T. Liu, W. Liang, Y. Zhu, and S. Huang. Clone: Closed-loop whole-body humanoid teleoperation for long-horizon tasks. In *9th Annual Conference on Robot Learning*, 2025.

[8] N. Myers, O. Kwon, S. Yamsani, and J. Kim. Child (controller for humanoid imitation and live demonstration): A whole-body humanoid teleoperation system. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 1–6. IEEE, 2025.

[9] A. Purushottam, J. Yan, C. Xu, and J. Ramos. Heavy lifting tasks via haptic teleoperation of a wheeled humanoid. In *2025 IEEE-RAS 24th International Conference on Humanoid Robots (Humanoids)*, pages 345–350. IEEE, 2025.

[10] T. He, Z. Luo, W. Xiao, C. Zhang, K. Kitani, C. Liu, and G. Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8944–8951. IEEE, 2024.

[11] Y. Ze, Z. Chen, J. P. Araújo, Z.-a. Cao, X. B. Peng, J. Wu, and C. K. Liu. Twist: Teleoperated whole-body imitation system. *arXiv preprint arXiv:2505.02833*, 2025.

[12] A. Allshire, H. Choi, J. Zhang, D. McAllister, A. Zhang, C. M. Kim, T. Darrell, P. Abbeel, J. Malik, and A. Kanazawa. Visual imitation enables contextual humanoid control. *arXiv preprint arXiv:2505.03729*, 2025.

[13] J. P. Araujo, Y. Ze, P. Xu, J. Wu, and C. K. Liu. Retargeting matters: General motion retargeting for humanoid motion tracking. *arXiv preprint arXiv:2510.02252*, 2025.

[14] C. Lu, X. Cheng, J. Li, S. Yang, M. Ji, C. Yuan, G. Yang, S. Yi, and X. Wang. Mobile-television: Predictive motion priors for humanoid whole-body control. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5364–5371. IEEE, 2025.

[15] Y. Sun, R. Chen, K. S. Yun, Y. Fang, S. Jung, F. Li, B. Li, W. Zhao, and C. Liu. Spark: Safe protective and assistive robot kit. IFAC Symposium on Robotics, 2025.

[16] X. Cheng, Y. Ji, J. Chen, R. Yang, G. Yang, and X. Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.

[17] S. Feng, E. Whitman, X. Xinjilefu, and C. G. Atkeson. Optimization based full body control for the atlas robot. In *2014 IEEE-RAS International Conference on Humanoid Robots*, pages 120–127. IEEE, 2014.

[18] B. Henze, M. A. Roa, and C. Ott. Passivity-based whole-body balancing for torque-controlled humanoid robots in multi-contact scenarios. *The International Journal of Robotics Research*, 35(12):1522–1543, 2016.

[19] A. Herzog, L. Righetti, F. Grimminger, P. Pastor, and S. Schaal. Balancing experiments on a torque-controlled humanoid with hierarchical inverse dynamics. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 981–988. IEEE, 2014.

[20] L. Sentis and O. Khatib. A whole-body control framework for humanoids operating in human environments. In *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006.*, pages 2641–2648. IEEE, 2006.

[21] Y. Pan, R. Qiao, L. Chen, K. Chitta, L. Pan, H. Mai, Q. Bu, C. Zheng, H. Zhao, P. Luo, and H. Li. Agility meets stability: Versatile humanoid control with heterogeneous data. *arXiv preprint arXiv:2511.17373*, 2025.

[22] Z. Fu, T. Z. Zhao, and C. Finn. Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation. *arXiv preprint arXiv:2401.02117*, 2024.

[23] O. Kwon, S. Yamsani, N. Myers, S. Taylor, J. Hong, K. Park, A. Alspach, and J. Kim. Paprle (plug-and-play robotic limb environment): A modular ecosystem for robotic limbs. *arXiv preprint arXiv:2507.05555*, 2025.

[24] P. Wu, Y. Shentu, Z. Yi, X. Lin, and P. Abbeel. Gello: A general, low-cost, and intuitive teleoperation framework for robot manipulators. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12156–12163. IEEE, 2024.

[25] H. Fang, C. Wang, Y. Wang, J. Chen, S. Xia, J. Lv, Z. He, X. Yi, Y. Guo, X. Zhan, et al. Airexo-2: Scaling up generalizable robotic imitation learning with low-cost exoskeletons. *arXiv preprint arXiv:2503.03081*, 2025.

[26] S. Yang, M. Liu, Y. Qin, R. Ding, J. Li, X. Cheng, R. Yang, S. Yi, and X. Wang. Ace: A cross-platform visual-exoskeletons system for low-cost dexterous teleoperation. *arXiv preprint arXiv:2408.11805*, 2024.

[27] T. Zhang, Z. McCarthy, O. Jow, D. Lee, X. Chen, K. Goldberg, and P. Abbeel. Deep imitation learning for complex manipulation tasks from virtual reality teleoperation. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5628–5635. Ieee, 2018.

[28] Y. Pan, C. Chen, D. Li, Z. Zhao, and J. Hong. Augmented reality-based robot teleoperation system using rgb-d imaging and attitude teaching device. *Robotics and Computer-Integrated Manufacturing*, 71:102167, 2021.

[29] R. Ding, Y. Qin, J. Zhu, C. Jia, S. Yang, R. Yang, X. Qi, and X. Wang. Bunny-visionpro: Real-time bimanual dexterous teleoperation for imitation learning. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 12248–12255. IEEE, 2025.

[30] K. Yin, W. Zeng, K. Fan, M. Dai, Z. Wang, Q. Zhang, Z. Tian, J. Wang, J. Pang, and W. Zhang. Unitracker: Learning universal whole-body motion tracker for humanoid robots. *IEEE Robotics and Automation Letters*, 2026.

[31] K. Darvish, Y. Tirupachuri, G. Romualdi, L. Rapetti, D. Ferigo, F. J. A. Chavez, and D. Pucci. Whole-body geometric retargeting for humanoid robots. In *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*, pages 679–686. IEEE, 2019.

[32] L. Yang, X. Huang, Z. Wu, A. Kanazawa, P. Abbeel, C. Sferrazza, C. K. Liu, R. Duan, and G. Shi. Omniretarget: Interaction-preserving data generation for humanoid whole-body loco-manipulation and scene interaction. *arXiv preprint arXiv:2509.26633*, 2025.

[33] Z. Xiong, L. Fang, J. Huang, K. Yamazaki, H. Zhang, and C. Gan. Extremcontrol: Low-latency humanoid teleoperation with direct extremity control. *arXiv preprint arXiv:2602.11321*, 2026.

[34] T. Zhu, G. Cai, Y. Zhaohui, G. Ren, H. Xie, Z. Wang, J. Wu, J. Wang, X. Yang, Y. Mu, et al. Clot: Closed-loop global motion tracking for whole-body humanoid teleoperation. *arXiv preprint arXiv:2602.15060*, 2026.

[35] J. Li, B. Tang, and F. Wu. Telegate: Whole-body humanoid teleoperation via gated expert selection with motion prior. *arXiv preprint arXiv:2602.09628*, 2026.

[36] Unity Technologies. Unity, 2026. URL https://unity.com/.

[37] WitMotion HK Co.ltd. 6-axis accelerometer: Precision in motion tracking, 2026. URL https://witmotion-sensor.com/.

[38] The Khronos Group Inc. Khronos openxr registry, 2026. URL https://registry.khronos.org/OpenXR/.

[39] Meta. Meta quest vr headsets and accessories — meta store, 2026. URL https://www.meta.com/quest/.

[40] PICO Immersive Pte.ltd. Pico 4 ultra enterprise-vr-mr-headset — pico global, 2026. URL https://www.picoxr.com/global/products/pico4-ultra-enterprise.

[41] Stereolabs Inc. Introducing zed mini — stereolabs, 2026. URL https://www.stereolabs.com/blog/introducing-zed-mini.