Research Article | Open Access

Intelligence & Robotics

Yang *et al*. *Intell. Robot.* **2026**, *6*(3), 444-78 **DOI:** 10.20517/ir.2026.22

Robust control barrier function-based shared control for cognitive-physical human-robot collaboration

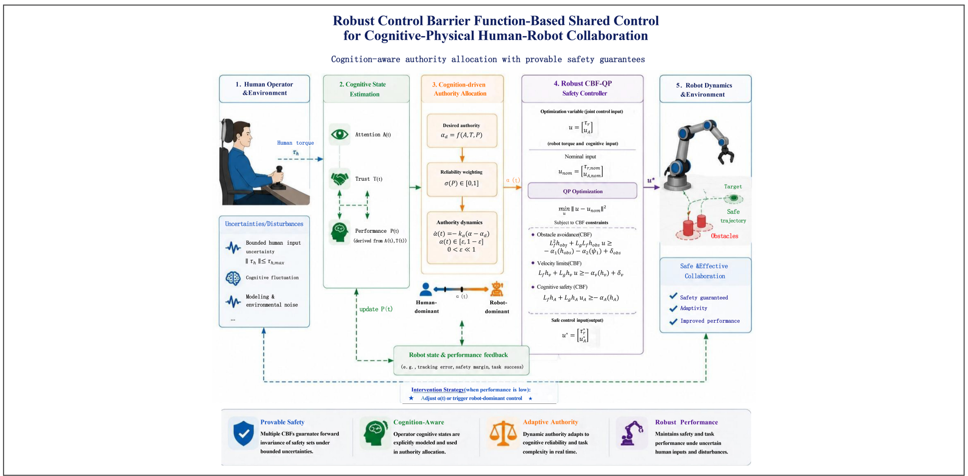
Yana Yang[1,2], Zhijia Li[1,2], Huixin Jiang[1,2], Ying Zhang[1,2]

Keywords:
Human-robot collaboration, shared control, control barrier functions, authority allocation, cognitive-state modeling, safety-critical control

Citation: Yang, Y.; Li, Z.; Jiang, H.; Zhang, Y. Robust control barrier function-based shared control for cognitive-physical human-robot collaboration. *Intell. Robot.* **2026**, *6*(3), 444-78. https://dx.doi.org/10. 20517/ir.2026.22

Received: 14 May 2026
First Decision: 12 Jun 2026
Revised: 28 Jun 2026
Accepted: 15 Jul 2026
Published: 31 Jul 2026

Academic Editors:
Hong Cheng, Jianjun Ni
Copy Editor:
Pei-Yun Wang
Production Editor:
Pei-Yun Wang



Abstract

Ensuring safe shared control in human - robot collaboration remains challenging due to uncertain human inputs and time-varying operator cognitive states. Existing methods primarily address either physical-interaction safety or authority allocation, but rarely provide a unified framework that simultaneously enables cognition-aware authority adaptation and formal safety guarantees. To address this issue, this paper proposes a robust coupled cognitive - physical shared-control framework for human - robot collaboration. First, an augmented state-space model is established by integrating robot dynamics with operator cognitive states, where the human control input is explicitly treated as a bounded disturbance. Based on this model, multiple robust control barrier functions are constructed to enforce obstacle avoidance, velocity limits, and lower bounds of cognitive safety levels via an online quadratic-programming-based controller. Furthermore, a cognition-driven dynamic authority allocation mechanism and a hierarchical intervention strategy are introduced to enable adaptive transitions between human-dominant and robot-dominant modes. The proposed framework guarantees forward invariance of the safe set and bounded closed-loop signals.

[1]Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, Hebei, China.
[2]Engineering Research Center of Intelligent Control System and Intelligent Equipment, Ministry of Education, Yanshan University, Qinhuangdao 066004, Hebei, China.

Correspondence to: Prof. Yana Yang, Institute of Electrical Engineering, Yanshan University, Qinhuangdao 066004, China. E-mail: yyn@ysu.edu.cn

Simulation results under uncertain human input and cognitive degradation scenarios demon-strate improved safety, adaptability, and collaboration compared with conventional methods.

1. INTRODUCTION

As human–robot collaboration (HRC) systems are increasingly deployed in industrial manufacturing, rehabilitation assistance, intelligent transportation, and service robotics, ensuring safe and efficient interaction between human operators and robotic systems has become a critical research issue[1–5]. In many emerging applications, robots are no longer isolated automated devices but collaborative agents that physically or cognitively interact with humans in shared workspaces. Typical examples include robot-assisted rehabilitation, semi-autonomous manipulation, intelligent vehicles, and cooperative industrial assembly, where both productivity and operator safety must be guaranteed simultaneously. Shared control provides an effective paradigm in which human decision-making capacity and robotic autonomy are combined to perform collaborative tasks with improved flexibility, robustness, and task performance[6,7]. By preserving human strategic judgment while exploiting machine precision and computational capabilities, shared control is considered a promising framework for next-generation HRC systems. However, unlike conventional automated systems, HRC environments involve uncertain human inputs, time-varying intentions, nonlinear robot dynamics, actuator nonlinearities commonly encountered in practical systems, and physical constraints that are critical to safety. Human actions may deviate from expected commands due to fatigue, delayed reaction, distraction, or imperfect situational awareness, significantly increasing the complexity of controller design. Existing methods mainly focus on the safety of physical-layer interaction, while the influence of operator cognitive states on the collaborative control process is still not adequately considered.

Existing shared-control approaches commonly rely on fixed blend weights, policy fusion, haptic authority transition, heuristic allocation mechanisms, or optimization-based teleoperation coordination frameworks[8–12]. These methods typically combine human commands and autonomous actions through predefined weighting rules or mode-switching strategies to improve operability and reduce workload. More recent studies have introduced adaptive authority regulation based on driving intention, environmental context, or task feasibility[13–18]. Such strategies demonstrate that dynamic authority adjustment can improve collaboration quality compared with static allocation. Compared with the impedance-learning-based shared control method for human-guided robots in contact-rich environments, which mainly focuses on adaptive impedance regulation and force interaction modeling[19], the proposed method emphasizes cognition-aware shared control with explicit modeling of operator attention and trust under a robust control barrier function (CBF) framework. Although these methods improve cooperation performance in structured scenarios, they typically regulate authority according to task variables or predefined rules rather than the operator's current cognitive state. In practical human-in-the-loop systems, operator capability may vary substantially over time due to stress, workload accumulation, vigilance decline, or trust fluctuation. Human factor studies have shown that attention degradation, trust miscalibration, and reduced situational awareness can significantly affect takeover quality, response speed, and collaboration reliability[20–27]. If such cognitive variations are ignored, the shared-control system may intervene too late when human performance degrades, or over-intervene during normal operation. This can lead to reduced usability and lower user acceptance. Consequently, existing shared-control methods still lack a principled mechanism for adapting cognition-aware authority to dynamically varying operator conditions.

In addition to authority-allocation strategies, rapid-convergence nonlinear control methods, such as fixed-time and prescribed-time stabilization, have demonstrated strong robustness and disturbance-rejection capability for uncertain nonlinear systems[28,29]. These properties are desirable for HRC, where unsafe deviations must be corrected promptly. However, such methods mainly focus on stability and tracking performance, while explicit safety constraint satisfaction remains insufficiently addressed. To enforce safety constraints in real time, CBFs have become an important tool for safety-critical control[30,31]. Compared with conventional constraint-handling methods, CBFs provide a systematic way to transform state safety requirements into inequality constraints on control inputs, enabling online safety filtering while preserving nominal control objectives. Due to this advantage, CBF-based methods have been successfully applied to obstacle avoidance, constrained motion regulation, multi-agent coordination, and safe human–robot interaction[32–37]. Recent work has also incorporated uncertainty prediction into barrier-function design to improve interaction safety under stochastic human motion or environmental uncertainty[38], and cooperative control with haptic shared autonomy has also been explored[39]. These developments demonstrate the strong potential of CBFs for real-time safety assurance in collaborative systems. However, most existing CBF frameworks assume accurate models or simplified. Uncertain human control inputs are rarely explicitly modeled as bounded disturbances. In addition, current constraints mainly address physical risks such as collision-avoidance, actuator limits, or velocity limits, with limited consideration of cognition-related safety degradation, such as delayed reaction or loss of attention. Moreover, the coupling between human authority variation and safety constraint activation is rarely explicitly studied. Therefore, existing CBF methods are difficult to apply directly to cognition-involved shared-control systems.

Based on the above observations, a fundamental challenge remains unresolved: how to establish a unified shared-control framework that simultaneously captures the evolution of operator cognitive states, handles uncertain human inputs, guarantees safety subject to physical constraints, and adaptively regulates human–robot authority.

To address this issue, this paper proposes a robust cognitive–physical coupled shared-control framework for HRC. The human input torque is modeled as a bounded disturbance in an augmented state-space model that integrates robot dynamics and operator cognitive states. Based on the estimated cognitive condition, a dynamic authority allocation and hierarchical intervention mechanism are developed to enable the adaptive transition between human-dominant and robot-dominant modes. In addition, multiple robust CBFs are constructed to simultaneously ensure obstacle avoidance, velocity constraints, and minimum cognitive safety requirements during the shared-control process.

The main contributions of this paper are summarized as follows:

(1) A unified cognitive–physical system model is developed by integrating robot dynamics with the operator's cognitive-state evolution. An augmented state-space representation is established to jointly characterize the robot motion states and the operator's internal states, including attention and trust. Within this framework, the human control input is explicitly modeled as a bounded external disturbance, enabling a systematic description of uncertain human interaction and providing a rigorous basis for subsequent robust controller design.

(2) A cognition-driven dynamic authority allocation and hierarchical intervention mechanism is designed to balance safety and collaboration performance through adaptive human–robot role transitions. A comprehensive performance index, constructed from cognitive states and task-tracking performance, is introduced to assess the real-time collaboration status. Based on the proposed index, control authority is continuously adjusted according to operator capability, while a hierarchical intervention policy progressively reduces human authority or triggers robot-dominant safety intervention under severely degraded conditions.

(3) A robust CBF-based safety controller is proposed to simultaneously enforce spatial, kinematic, and cognitive safety constraints under uncertain human interactions. For the cognitive–physical coupled system subject to bounded disturbances, robust CBF conditions are established to guarantee constraint satisfaction despite

uncertainty. Multiple safety requirements, including obstacle avoidance, velocity limitation, and minimum cognitive safety requirements, are integrated into an online Quadratic Programming (QP) framework to generate real-time safe control actions.

The remainder of this paper is organized as follows. Section 2 constructs a cognitive-physical coupled model for the human-robot collaborative system, comprising robot dynamics, a cognitive state model describing the evolution of the operator's attention and trust, and a unified state-space model that couples them and account for bounded human input disturbances. Section 3 details the proposed method: first, based on robust CBF theory, multiple constraints are designed to ensure spatial obstacle avoidance, velocity limits and cognitive-level safety for the perturbed system, and safety control is implemented by solving an online QP problem; second, a cognitive state-driven dynamic control authority allocation mechanism is proposed; finally, a hierarchical dynamic intervention strategy is constructed based on cognitive performance assessment. Section 4 validates the effectiveness and superiority of the proposed framework through a series of simulations across various scenarios, including the presence of disturbances in human input and fluctuations in the cognitive state of the operator. Finally, Section 5 summarizes the work and outlines the directions for future research.

2. MODELING AND FORMULATION

In human-robot collaborative systems, the dynamic response capability of the robot and the cognitive state of the human operator are critical factors that influence safety and performance. To achieve effective human-robot shared control, this paper integrates robot dynamics with a model of the human cognitive state, constructing a unified HRC model. This integration aims to optimize the human-robot interaction process while ensuring operational safety and efficiency.

2.1. System model

The robot dynamic model considered in this paper describes the manipulator's motion in the task space. The state variables include the joint angles $\boldsymbol{q}$, the joint velocities $\dot{\boldsymbol{q}}$, and the applied external forces. The dynamic equation is given by

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q},\dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{G}(\boldsymbol{q}) = \alpha(t)\,\boldsymbol{\tau}_h + \left[1-\alpha(t)\right]\boldsymbol{\tau}_r, \tag{1}$$

where $\boldsymbol{M}(\boldsymbol{q})$, $\boldsymbol{C}(\boldsymbol{q},\dot{\boldsymbol{q}})$, and $\boldsymbol{G}(\boldsymbol{q})$ are the inertia, Coriolis–centrifugal, and gravitational matrices, respectively; $\boldsymbol{\tau}_r$ and $\boldsymbol{\tau}_h$ denote the autonomous and human control inputs; and $\alpha(t) \in [0,1]$ is the dynamic control authority, with $\alpha(t)$ representing human authority and $1-\alpha(t)$ representing robot authority.

Remark 1 *In a human-robot collaborative system, the overall control inputs comprise three distinct components. First, the human input $\boldsymbol{\tau}_h$ is treated as an unknown but bounded disturbance satisfying $\|\boldsymbol{\tau}_h\|_2 \le \tau_{h,\max}$, where $\tau_{h,\max}$ denotes the upper bound of the Euclidean norm of the human input disturbance vector, enabling robust safety constraints via CBFs. Second, the autonomous robot input $\boldsymbol{\tau}_r$ is computed by a QP-based CBF controller to ensure safety and tracking performance. Third, the cognitive regulation input u_A modulates the operator's cognitive state through interventions such as prompts or alarms. In practice, the disturbance bound $\tau_{h,\max}$ can be estimated from operator calibration trials, biomechanical torque limits, or historical interaction data. To accommodate occasional transient violations caused by startle responses or unexpected physical effort, the controller additionally employs robust CBF margins, residual autonomous authority, and QP slack variables, which together provide a practical safety buffer beyond the nominal disturbance model.*

The cognitive state of the operator is characterized by two variables that vary over time: the level of attention $A(t)$ and the level of trust $T(t)$. Attention $A(t)$ can be considered a state variable whose dynamic evolution is primarily influenced by four factors: (1) natural decay; (2) the modulating effect of task complexity; (3) external regulatory input; and (4) the gain from the trust level. The dynamic equation describing the evolution of attention $A(t)$ is constructed as follows:

$$\dot{A}(t) = -\left(\lambda_0 + \mu C(t)\right) A(t) + u_A(t) + \gamma T(t)(1 - A(t)) \tag{2}$$

where $\lambda_0 > 0$ denotes the base attention decay rate, whose value is determined by reference[40]; $\mu > 0$ is a proportional coefficient that modulates the influence of task complexity, also determined by reference[40]; and $C(t) \in [0, 1]$ denotes the normalized task complexity at time t. A higher value of $C(t)$ corresponds to a more complex task, increasing the effective decay rate $\lambda = \lambda_0 + \mu C(t)$, thereby accelerating the natural decay of attention. The input $u_{\mathrm{A}}(t)$ denotes the regulatory input to modulate the attention level. In this formulation, we consider two distinct operational conditions: with and without external disturbances. The detailed formulation of $u_{\mathrm{A}}(t)$ under these conditions is provided in Section 3.1. The parameter γ represents the influence coefficient of trust on attention, quantifying the strength of the coupling between the trust level $T(t)$ and the dynamics of attention.

The trust level $T(t)$ can be treated as a state variable. Its dynamics are governed by three primary factors: (1) regression to a baseline level; (2) a positive gain from the operator's attention; and (3) suppression caused by system failures. Consequently, the dynamic equation for trust can be formulated as follows:

$$\dot{T}(t) = -\beta_1(T - T_{\mathrm{base}}) + \beta_2 A(1 - T) - \beta_3 F_{\mathrm{int}} \tag{3}$$

where $\beta_1 > 0$ is the trust decay rate, $T_{\mathrm{base}} \in [0, 1]$ denotes a baseline (or equilibrium) trust level, $\beta_2 > 0$ is a positive gain coefficient that captures the influence of attention on trust, and $\beta_3 > 0$ is a scaling factor that quantifies the suppressive effect of external disturbances including system-level faults and sudden environmental or cognitive disruptions, F_{int} denotes the composite disturbance. The parameters β_1, β_2, and β_3 are introduced as tuning gains to regulate the relative influence of trust decay, attention-driven reinforcement, and the suppressive effect induced by faults and sudden disturbances. These parameters are selected empirically within bounded ranges to ensure stable and physically meaningful evolution of the cognitive state variables.

Remark 2 *The attention dynamics in Equation (2) are partially inspired by the cognitive dynamics model reported in Ref.[40], Equation (3), particularly the attention decay term and the task-complexity-dependent modulation term. In this work, this baseline structure is further extended by introducing the nonlinear coupling term $\gamma T(1 - A)$, which describes the reinforcing influence of trust on attention recovery. Moreover, Equation (3) is constructed to characterize the trust evolution in HRC by incorporating baseline recovery, attention-dependent reinforcement, and failure-induced suppression effects. Therefore, Equations (2) and (3) are not intended to propose a standalone psychological model, but to provide a control-oriented cognitive-state representation that can be embedded into the augmented control-affine system and integrated with the proposed robust CBF-based shared-control framework.*

Both variables are defined as positive and bounded functions:

$$A(t), T(t) \in [0, 1],$$

where $A(t) = 1$ denotes a state of complete focus, $A(t) = 0$ indicates complete distraction, $T(t) = 1$ indicates full trust in the autonomous system, and $T(t) = 0$ represents a complete lack of trust.

Remark 3 *Human cognitive states, such as attention and trust, are finite and decay over time. Their evolution is modeled by ordinary differential equations that capture the resource-limited nature of cognition. Attention can be recovered or enhanced by external inputs (e.g., task challenges or alerts), which are incorporated as a control input*

$u_A(t)$. According to attention allocation theory, attention decays approximately exponentially. A time-varying task complexity $C(t)$ is introduced to reflect changing operational demands. The parameters μ, γ, and β_3 are assumed to be obtained through offline calibration before task execution. Such parameters can be estimated from experimentally observed attention and trust trajectories using standard system-identification techniques or human-factor assessment methods. Since the focus of this paper is safety-critical shared-control design rather than cognitive-model learning, online parameter adaptation is not considered and is left for future investigation.

In human-robot collaborative control systems, the system's dynamic evolution is determined not solely by the mechanical plant, but is also significantly influenced by the operator's cognitive state and their control actions. Consequently, it is essential to construct a model of a unified dynamic system that captures the intrinsic coupling between cognitive and physical dynamics.

The integrated state of the collaborative human-robot system is defined by augmenting the robot's physical states with the operator's cognitive states. The complete, unified system state is obtained by concatenating the physical and cognitive states.

$$\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_p \\ \boldsymbol{x}_c \end{bmatrix} = \begin{bmatrix} \boldsymbol{q} \\ \dot{\boldsymbol{q}} \\ A \\ T \end{bmatrix} \in \mathbb{R}^{2n+2}. \tag{4}$$

where $\boldsymbol{x}_p = [\boldsymbol{q}, \dot{\boldsymbol{q}}]^\top \in \mathbb{R}^{2n}$ and $\boldsymbol{x}_c = [A, T]^\top \in [0,1]^2$.

The unified control input vector is defined as:

$$\boldsymbol{u} = \begin{bmatrix} \boldsymbol{\tau}_r \\ u_A \end{bmatrix}. \tag{5}$$

Remark 4 *Human input $\boldsymbol{\tau}_h$ is generated by the operator and is not directly controllable by the autonomous controller. Therefore, in the control design framework, it is treated as an external disturbance. However, it is explicitly taken into account in the shared control law through the dynamic authority variable $\alpha(t)$. Unlike conventional approaches where the control authority is directly defined as a static function of cognitive states, this paper models the control authority as a dynamic variable to ensure smooth and physically realizable transitions. Moreover, the inertia matrix $M(q)$ is assumed to be nonsingular in the operational workspace, which is a standard condition in robotic manipulator modeling.*

The integrated system is expressed in the standard control-affine form as follows:

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,\boldsymbol{u} + \boldsymbol{d}(\boldsymbol{x}, t), \tag{6}$$

where the state vector is $\boldsymbol{x} = [\boldsymbol{q}, \dot{\boldsymbol{q}}, A, T]^\top \in \mathbb{R}^{2n+2}$, the control input is $\boldsymbol{u} = [\boldsymbol{\tau}_r, u_A]^\top \in \mathbb{R}^{n+1}$.

Assumption 1 *The robot operates within a bounded workspace that avoids kinematic singularities. Consequently, the inertia matrix $M(q)$ remains symmetric positive definite over the admissible configuration set. As a result, $M^{-1}(q)$ exists and is uniformly bounded in the operational region, ensuring the well-posedness of the robot dynamics and the validity of the subsequent control design.*

Due to the presence of the human input, the system is subject to an external disturbance term $d(x, t)$ and the functions $f(x)$ and $g(x)$ are defined as follows.

$$f(x) = \begin{bmatrix} \dot{q} \\ M^{-1}(q)[-C(q, \dot{q})\dot{q} - G(q)] \\ -(\lambda_0 + \mu C)A + \gamma T(1 - A) \\ -\beta_1(T - T_{\text{base}}) + \beta_2 A(1 - T) - \beta_3 F_{\text{int}} \end{bmatrix}. \tag{7}$$

$$g(x) = \begin{bmatrix} \mathbf{0}_{n\times n} & \mathbf{0}_{n\times 1} \\ M^{-1}(q)(1 - \alpha) & \mathbf{0}_{n\times 1} \\ \mathbf{0}_{1\times n} & 1 \\ \mathbf{0}_{1\times n} & 0 \end{bmatrix}. \tag{8}$$

The human input torque τ_h is treated as a bounded disturbances. Specifically, we assume:

$$\|\tau_h\|_2 \le \tau_{h,\max}$$
$$d(x, t) = \begin{bmatrix} \mathbf{0} \\ M^{-1}(q)\alpha\tau_h \\ 0 \\ 0 \end{bmatrix}, \tag{9}$$

where $d(x, t)$ represents the perturbation induced by the human input, incorporated as an additive term in the system dynamics. The only non-zero entry corresponds to the influence of the human control input τ_h on the robot's acceleration, scaled by the control authority α and the inverse inertia matrix $M^{-1}(q)$. Thus, the disturbance term is bounded such that

$$\|d(x, t)\| \le d_{\max},$$

where $d_{\max}$ denotes a known uniform upper bound of the disturbance over the admissible safe set.

2.2. Performance index based on CBFs

In the human-robot collaborative system constructed, operational safety must be ensured in the presence of external disturbances induced by human input. This can be achieved by constraining the state of the system to remain within a predefined safe set. To this end, CBFs are introduced as safety performance indices.

Consider a continuously differentiable function $h : \mathbb{R}^n \to \mathbb{R}$. The associated safe set is defined as:

$$\mathcal{S} = \{x \in \mathbb{R}^n \mid h(x) \ge 0\}, \tag{10}$$

Here, $h(x) = 0$ defines the boundary of the safe set, while $h(x) < 0$ corresponds to the unsafe region.

The derivative of the safety function $h(x)$ along the dynamics of the system is given by:

$$\dot{h}(x, u) = \nabla h(x)\, f(x) + \nabla h(x)\, g(x)\, u + \nabla h(x)\, d(x, t). \tag{11}$$

Due to the presence of disturbances, an additional disturbance term appears in the derivative, where $\boldsymbol{d}(\boldsymbol{x},t)$ denotes an unknown but bounded disturbances.

To account for the bounded disturbances $\boldsymbol{d}(\boldsymbol{x},t)$ induced by human input $\boldsymbol{\tau}_h$, the standard CBF condition is extended to a robust CBF condition. The safety condition for the system is given by the inequality

$$\dot{h}(\boldsymbol{x},\boldsymbol{u}) \geq -\delta_h, \qquad \forall \boldsymbol{x} \in \partial S,$$

where $h(\boldsymbol{x})$ is a candidate for the CBF, $\dot{h}(\boldsymbol{x},\boldsymbol{u})$ denotes its time derivative along closed-loop dynamics, $\delta_h > 0$ is a robustness margin, and ∂S is the boundary of the safe set $S = \{\boldsymbol{x} \mid h(\boldsymbol{x}) \geq 0\}$.

Remark 5 *Robust safety is achieved by introducing a robust margin δ_h to counteract the worst-case disturbance, thereby guaranteeing forward invariance of the safe set. To guarantee safety under bounded disturbances, a robust margin $\delta_h = \|\nabla h(\boldsymbol{x})\| d_{\max}$ is introduced, where $d_{\max}$ is the upper bound of the disturbance satisfying $\|\boldsymbol{d}(\boldsymbol{x},t)\| \leq d_{\max}$, and $d_{\max} = \|\boldsymbol{M}^{-1}(\boldsymbol{q})\alpha\| \tau_{h,\max}$. According to Nagumo's theorem, a necessary condition for invariance of the safe set $\mathcal{S}$ is that, on the boundary $\partial\mathcal{S}$, the derivative is non-negative. Under disturbance, the invariance condition must account for the worst-case effect of d(x,t).*

To obtain a condition that can be enforced throughout the entire safe set, the notion of a CBF is introduced. A continuously differentiable function $h(\boldsymbol{x})$ is called a CBF for the system if there exists an extended class $\mathcal{K}_\infty$ function $\alpha(\cdot)$ such that for all $\boldsymbol{x} \in \mathcal{S}$:

$$\sup_u \left[\nabla h\, f + \nabla h\, g\, u\right] \geq -\alpha(h(x)) + \delta_h$$

Equivalently, this condition requires the existence of a control input $\boldsymbol{u}$ that satisfies the following:

$$\nabla h(\boldsymbol{x})\boldsymbol{f}(\boldsymbol{x}) + \nabla h(\boldsymbol{x})\boldsymbol{g}(\boldsymbol{x})\boldsymbol{u} \geq -\alpha\big(h(\boldsymbol{x})\big) + \delta_h,$$

Under the assumption of bounded disturbances and the robust CBF condition, the forward invariance of the safe set $\mathcal{S}$ can be guaranteed, provided that a control input satisfying the constraint exists for all $\boldsymbol{x} \in \mathcal{S}$. Three categories of safety constraints are considered in this paper.

First, for spatial safety, the end effector of the manipulator must maintain a safe distance from obstacles even in the presence of disturbances induced by human input. The corresponding safe set and associated safety function are defined as follows:

$$\begin{aligned} \mathcal{S}_{\text{obs}} &= \left\{\boldsymbol{x} \mid \|\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\text{obs}}\| \geq d_{\text{safe}}\right\}. \\ h_{\text{obs}}(\boldsymbol{x}) &= \|\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\text{obs}}\|^2 - d_{\text{safe}}^2 \geq 0. \end{aligned} \tag{12}$$

Second, to ensure safe motion, the joint velocities must not exceed a prescribed maximum magnitude under possible disturbances; the corresponding safe set and associated safety function are defined as follows:

$$\begin{aligned} \mathcal{S}_v &= \{\boldsymbol{x} \mid \|\dot{\boldsymbol{q}}\| \leq v_{\max}\}. \\ h_v(\boldsymbol{x}) &= v_{\max}^2 - \|\dot{\boldsymbol{q}}\|^2 \geq 0. \end{aligned} \tag{13}$$

Third, cognitive security requires that the attention level A remain above a predefined minimum threshold to ensure safe human participation in the control loop, the corresponding safe set and associated safety function are defined as follows:

$$\begin{aligned}\mathcal{S}_A &= \{\boldsymbol{x} \mid A \geq A_{\min}\}.\\ h_A(\boldsymbol{x}) &= A - A_{\min} \geq 0.\end{aligned} \tag{14}$$

Remark 6 *The threshold $A_{\min}$ is treated as a user-defined design parameter that specifies the minimum acceptable level of attention for safe human participation in the collaborative task. In practical applications, its value may vary according to task characteristics, operational risk, and human-factor considerations. For example, safety-critical tasks generally require a higher attention threshold than low-risk collaborative scenarios. The proposed control framework and theoretical analysis remain applicable for any prescribed threshold satisfying $0 < A_{\min} < 1$, since changing $A_{\min}$ only modifies the definition of the cognitive safe set without altering the controller design methodology. The above safety constraints collectively define the admissible safe sets of the system states. Due to human-induced disturbances, these constraints cannot be guaranteed directly and are subsequently enforced through the proposed robust CBF framework.*

2.3. Problem formulation

To address the safety and performance requirements in HRC under human-induced disturbances, the control problem is formulated as an optimal control framework. This framework aims to minimize a composite cost function that includes trajectory tracking errors, control efforts, and deviations of the cognitive state from its desired level, subject to the system dynamics, input constraints, and the robust safety conditions derived from spatial, velocity, and cognitive barriers. The complete problem formulation is given as follows.

$$\begin{gathered}\min_{u} \quad J = \int_0^T \left[\|\boldsymbol{q} - \boldsymbol{q}_d\|^2 + \|\dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_d\|^2 + \|\boldsymbol{\tau}_r\|^2 + u_{\mathrm{A}}^2 + (A - A_d)^2\right] \mathrm{d}t \\ \text{s.t.} \quad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\,\boldsymbol{u} + \boldsymbol{d}(\boldsymbol{x}, t), \\ \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{u} \in \mathcal{U}, \quad \boldsymbol{\tau}_r \in \mathcal{T} \\ h_{\mathrm{obs}}(\boldsymbol{x}) = \|\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\mathrm{obs}}\|^2 - d_{\mathrm{safe}}^2 \geq 0,\ L_f^2 h_{\mathrm{obs}} + L_g L_f h_{\mathrm{obs}}\, u \geq -\alpha_1(\dot{h}_{\mathrm{obs}}(\boldsymbol{x})) - \alpha_2(\psi_1(\boldsymbol{x})) + \delta_{\mathrm{obs}} \\ h_v(\boldsymbol{x}) = v_{\max}^2 - \|\dot{\boldsymbol{q}}\|^2 \geq 0,\ L_f h_v + L_g h_v\, u \geq -\alpha_v(h_v) + \delta_v \\ h_A(\boldsymbol{x}) = A - A_{\min} \geq 0,\ L_f h_A + L_g h_A\, u_A \geq -\alpha_A(h_A) \\ \|\boldsymbol{\tau}_h\|_2 \leq \tau_{h,\max},\end{gathered} \tag{15}$$

where $\boldsymbol{q}_d$ denotes the desired joint position, $\dot{\boldsymbol{q}}_d$ is the desired joint velocity, and A_d represents the desired attention level. The sets $\mathcal{X}$, $\mathcal{U}$, and $\mathcal{T}$ denote the admissible state set, the input set, and the robot torque set, respectively. The disturbance $\boldsymbol{d}(\boldsymbol{x}, t)$ is assumed to be bounded due to the physical limitations of human input.

3. METHOD DESIGN

3.1. Safety controller design based on robust CBF

The control input u_A is designed to regulate attention, counteract accumulated fatigue and ambient disturbances in the absence of external perturbations, and provide rapid compensation for attention deviations when a disturbance occurs[40]. The specific form of this cognitive state regulator is given by:

$$u_A(t) = \begin{cases} k_1 R_{\mathrm{recov}} - k_2 F_{\mathrm{fatig}}(t) - k_3 D_{\mathrm{distb}}(t), & \text{No disturbance} \\ -k_4 \Delta A(t), & \text{Disturbance occurs} \end{cases} \tag{16}$$

Here, R_{recov} denotes the recovery intervention, F_{fatig} represents the level of fatigue of the operator, and D_{distb} represents persistent external disturbances of small amplitude. The term $\Delta A(t)$ is defined as the deviation of attention induced by disturbances, i.e., $\Delta A(t) = (A(t) - A_d)$.

Remark 7 *From a control-theoretic perspective, attention variation exhibits two distinct time scales: a slow process under normal conditions and a fast-varying (or abrupt) response under significant external disturbance. A unified control law is insufficient to handle both regimes, necessitating a piecewise control strategy. In the absence of disturbance, the control input $u_A(t)$ balances recovery effects, accumulated fatigue, and ambient disturbances to regulate the operator's attention. In contrast, when a substantial disturbance occurs, the controller provides fast compensation proportional to the immediate attention deficit $\Delta A(t)$, acting as an emergency corrective input.*

From an implementation perspective, the control law in Equation (16) defines a nominal cognitive regulation signal that serves as the reference for the optimization-based controller. The actual implemented input is obtained through the QP formulation in Equation (24), which ensures smooth and continuous control action. Although Equation (16) is piecewise in design, the closed-loop control input remains continuous due to the QP-based optimization framework, which guarantees smooth variation of the control solution with respect to system states.

In the absence of disturbance, a nominal control torque τ_{nom} is constructed to achieve trajectory tracking. A standard computed-torque control law is adopted:

$$\boldsymbol{\tau}_{\text{nom}} = \boldsymbol{M}(\boldsymbol{q})(\ddot{\boldsymbol{q}}_d - \mathbf{K}_p\boldsymbol{e} - \mathbf{K}_d\dot{\boldsymbol{e}}) + \boldsymbol{C}(\boldsymbol{q},\dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{G}(\boldsymbol{q}) \tag{17}$$

where $\boldsymbol{e} = \boldsymbol{q} - \boldsymbol{q}_d$ is the position tracking error.

Since human input $\boldsymbol{\tau}_h$ is treated as an unknown disturbance, it cannot be used in control design. Therefore, the nominal robot command is designed only based on the robot states and the available authority variable α. Considering that the autonomous input enters the shared dynamics through the weighted term $(1-\alpha)\tau_r$, the authority-compensated nominal robot command is defined as

$$\tau_{r,\text{nom}} = \frac{\tau_{\text{nom}}}{1-\alpha}.$$

The compensation factor $1/(1-\alpha)$ compensates for the authority weighting of the autonomous input channel. Since α is available from the authority allocation mechanism, this modification does not require knowledge of the unknown human input τ_h.

The attention regulation input is defined in a piecewise manner based on the presence of external disturbances:

$$u_{A,\text{nom}} = \begin{cases} k_1 R_{\text{recov}} - k_2 F_{\text{fatig}}(t) - k_3 D_{\text{distb}}(t), & \text{No disturbance} \\ -k_4 \Delta A(t), & \text{Disturbance occurs} \end{cases}$$

Consequently, the nominal control input vector is defined as:

$$\boldsymbol{u}_{\text{nom}} = \begin{bmatrix} \boldsymbol{\tau}_{r,\text{nom}} \\ u_{A,\text{nom}} \end{bmatrix}.$$

Remark 8 *When the CBF constraints are inactive and the human input disturbance is absent, the QP solution satisfies $\tau_r^* = \tau_{r,nom}$, and the closed-loop tracking error dynamics reduce to $M\ddot{e} + C\dot{e} + K_d\dot{e} + K_p e = 0$, which guarantees convergence of the tracking error.*

The gradient of the obstacle avoidance safety function h_{obs} with respect to the state vector $\boldsymbol{x} = [\boldsymbol{q}, \dot{\boldsymbol{q}}, A, T]^\top$ is given by:

$$\nabla h_{\text{obs}} = \left[\frac{\partial h_{\text{obs}}}{\partial \boldsymbol{q}}, \frac{\partial h_{\text{obs}}}{\partial \dot{\boldsymbol{q}}}, \frac{\partial h_{\text{obs}}}{\partial A}, \frac{\partial h_{\text{obs}}}{\partial T}\right]^\top.$$

Since h_{obs} depends only on the joint position $\boldsymbol{q}$, we have the following.

$$\frac{\partial h_{\text{obs}}}{\partial \boldsymbol{q}} = 2\left(\frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}}\right)^\top (\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\text{obs}}) = 2\boldsymbol{J}(\boldsymbol{q})^\top (\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\text{obs}}),$$

where $\boldsymbol{J}(\boldsymbol{q}) = \frac{\partial \boldsymbol{p}}{\partial \boldsymbol{q}}$ is the Jacobian matrix of the end-effector position, and

$$\frac{\partial h_{\text{obs}}}{\partial \dot{\boldsymbol{q}}} = \mathbf{0}, \qquad \frac{\partial h_{\text{obs}}}{\partial A} = 0, \qquad \frac{\partial h_{\text{obs}}}{\partial T} = 0.$$

The Lie derivative of h_{obs} along the drift vector field $\boldsymbol{f}$ is:

$$L_{\boldsymbol{f}} h_{\text{obs}} = \nabla h_{\text{obs}} \cdot \boldsymbol{f}(\boldsymbol{x}) = 2(\boldsymbol{p}(\boldsymbol{q}) - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}(\boldsymbol{q})\, \dot{\boldsymbol{q}}.$$

The Lie derivative along the control input matrix $\boldsymbol{g}$ is:

$$L_{\boldsymbol{g}} h_{\text{obs}} = \nabla h_{\text{obs}} \cdot \boldsymbol{g}(\boldsymbol{x}) = \frac{\partial h_{\text{obs}}}{\partial \boldsymbol{q}} \cdot \mathbf{0} + \frac{\partial h_{\text{obs}}}{\partial \dot{\boldsymbol{q}}} \cdot \left[\boldsymbol{M}^{-1}(1-\alpha)\right] = 0.$$

Since the obstacle safety function $h_{obs}(x)$ depends only on the configuration variable q, its first Lie derivative along the input vector field satisfies $L_g h_{obs}(x) = 0$ which indicates that the control input does not explicitly appear in the first derivative of the barrier function. Taking the derivative again yields

$$\ddot{h}_{obs}(x) = L_f^2 h_{obs}(x) + L_g L_f h_{obs}(x)\tau_r + \Delta_{obs},$$

where

$$L_g L_f h_{obs}(x) = 2(p - p_{obs})^T J(q) M^{-1}(q)(1-\alpha).$$

Under the authority constraint $\alpha(t) \in [\varepsilon, 1-\varepsilon]$, the autonomous control authority satisfies $1 - \alpha(t) \geq \varepsilon > 0$. Together with the assumptions that the robot operates away from kinematic singularities and $p(q) \neq p_{obs}$ on the safety boundary, we have $L_g L_f h_{obs}(x) \neq 0$.

Therefore, the obstacle avoidance constraint has relative degree two with respect to the robot control input, and a second-order (high-order) CBF is adopted. The high-order control barrier function (HOCBF) is defined as follows.

$$\begin{aligned}\psi_1(\boldsymbol{x}) &= \dot{h}_{\text{obs}}(\boldsymbol{x}) + \alpha_1(h_{\text{obs}}(\boldsymbol{x})),\\ \psi_2(\boldsymbol{x}) &= \dot{\psi}_1(\boldsymbol{x}) + \alpha_2(\psi_1(\boldsymbol{x})) \geq 0.\end{aligned}$$

where $\alpha_1(\cdot)$ and $\alpha_2(\cdot)$ are class $\mathcal{K}$ functions.

The second-order derivative is given by:

$$\ddot{h}_{\text{obs}} = 2\dot{\boldsymbol{q}}^\top \boldsymbol{J}_p^\top \boldsymbol{J}_p \dot{\boldsymbol{q}} + 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \dot{\boldsymbol{J}}_p \dot{\boldsymbol{q}} + 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}_p \ddot{\boldsymbol{q}}.$$

Hence, the components of the second derivative are defined as follows:

$$L_f^2 h_{\text{obs}} = 2\dot{\boldsymbol{q}}^\top \boldsymbol{J}^\top \boldsymbol{J}\dot{\boldsymbol{q}} + 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \dot{\boldsymbol{J}}\dot{\boldsymbol{q}} + 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}\boldsymbol{M}^{-1}(-\boldsymbol{C}\dot{\boldsymbol{q}} - \boldsymbol{G}).$$
$$L_{\boldsymbol{g}} L_f h_{\text{obs}} = 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}\boldsymbol{M}^{-1}(1-\alpha).$$
$$\Delta_{\text{obs}} = 2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}\boldsymbol{M}^{-1}\alpha\boldsymbol{\tau}_h.$$

The disturbance term Δ_{obs} in the second derivative of the safety function can be bounded from below. Specifically, we have $\Delta_{\text{obs}} \geq -\delta_{\text{obs}}$, where the disturbance bound is given by $\delta_{\text{obs}} = \|2(\boldsymbol{p} - \boldsymbol{p}_{\text{obs}})^\top \boldsymbol{J}\boldsymbol{M}^{-1}\alpha\|\, \tau_{h,\max}$. The explicit bound $\delta_{obs}(x)$ represents a state-dependent realization of the general robust margin. This bound quantifies the worst-case influence of the bounded human input on the rate of change of the safety constraint.

The condition $\psi_2(\boldsymbol{x}) \geq 0$ can be expressed as a linear constraint with respect to $\boldsymbol{\tau}_r$. After substituting the above expressions and simplifying, we obtain the following.

$$L_{\boldsymbol{g}} L_f h_{\text{obs}}(\boldsymbol{x})\, \boldsymbol{\tau}_r \geq -L_f^2 h_{\text{obs}}(\boldsymbol{x}) - \alpha_1(\dot{h}_{\text{obs}}) - \alpha_2(\psi_1) + \delta_{\text{obs}}, \tag{18}$$

Remark 9 *To maintain nonzero authority for both the human operator and the autonomous controller during HRC, the control authority variable is constrained as $\alpha(t) \in [\varepsilon, 1-\varepsilon], 0 < \varepsilon < 0.5$. Here, $\alpha(t)$ denotes the human control authority, while $1 - \alpha(t)$ represents the autonomous robot authority. Therefore, the lower bound ε guarantees that the human operator retains a minimum level of participation and that the autonomous controller preserves a nonzero intervention capability. Consequently, the control channel remains effective under the robust CBF constraints, through the term $L_g L_f h(x)$, while avoiding complete dominance of either side. It should be noted that the constraint $\alpha(t) \in [\varepsilon, 1-\varepsilon]$ only guarantees the existence of nonvanishing authority channels for both agents. It does not by itself guarantee feasibility of the CBF-QP problem under arbitrary state and disturbance realizations. Feasibility additionally depends on the existence of admissible control inputs satisfying the safety constraints and actuator limitations.*

The Equation (18) can be written in the standard QP form as:

$$\boldsymbol{A}_{\text{obs}}(\boldsymbol{x})\, \boldsymbol{\tau}_r \leq \boldsymbol{b}_{\text{obs}}(\boldsymbol{x}), \tag{19}$$

with

$$A_{\text{obs}}(x) = -L_g L_f h_{\text{obs}}(x),$$
$$b_{\text{obs}}(x) = L_f^2 h_{\text{obs}}(x) + \alpha_1(\dot{h}_{\text{obs}}(\boldsymbol{x})) + \alpha_2(\psi_1(\boldsymbol{x})) - \delta_{\text{obs}}.$$

This linear constraint ensures that the second-order CBF condition is satisfied, thus guaranteeing forward invariance of the safe set defined by $h_{\text{obs}}(\boldsymbol{x}) \geq 0$.

Since h_v depends only on $\dot{\boldsymbol{q}}$, its gradient with respect to the full state $\boldsymbol{x} = [\boldsymbol{q}, \dot{\boldsymbol{q}}, A, T]^\top$ is:

$$\nabla h_v = \left[\frac{\partial h_v}{\partial \boldsymbol{q}},\ \frac{\partial h_v}{\partial \dot{\boldsymbol{q}}},\ \frac{\partial h_v}{\partial A},\ \frac{\partial h_v}{\partial T}\right] = \left[\mathbf{0},\ -2\dot{\boldsymbol{q}}^\top,\ 0,\ 0\right].$$

The Lie derivatives along the drift field $\boldsymbol{f}(\boldsymbol{x})$ and the control input matrix $\boldsymbol{g}(\boldsymbol{x})$ are:

$$L_f h_v = \nabla h_v \cdot \boldsymbol{f}(\boldsymbol{x}) = -2\dot{\boldsymbol{q}}^\top \boldsymbol{M}^{-1}(-\boldsymbol{C}\dot{\boldsymbol{q}} - \boldsymbol{G})],$$
$$L_{\boldsymbol{g}} h_v = \nabla h_v \cdot \boldsymbol{g}(\boldsymbol{x}) = -2\dot{\boldsymbol{q}}^\top \boldsymbol{M}^{-1}(\boldsymbol{q})(1-\alpha).$$
$$\Delta_v = -2\dot{\boldsymbol{q}}^\top \boldsymbol{M}^{-1}\alpha\boldsymbol{\tau}_h$$
$$\Delta_v \geq -\delta_v$$

where

$$\delta_v = \left\|2\dot{\boldsymbol{q}}^\top \boldsymbol{M}^{-1}\alpha\right\| \tau_{h,\max}.$$

The velocity safety constraint is enforced through the following robust CBF condition:

$$L_f h_v + L_g h_v \boldsymbol{\tau}_r \geq -\alpha_v(h_v) + \delta_v, \tag{20}$$

where $\alpha_v(\cdot)$ is a class $\mathcal{K}$ function.

The above condition can be expressed in the following affine form:

$$A_v(x)\tau_r \leq b_v(x), \tag{21}$$

where

$$A_v(x) = -L_g h_v,$$
$$b_v(x) = L_f h_v + \alpha_v(h_v) - \delta_v.$$

The attention safety function is defined as $h_A(\boldsymbol{x}) = A - A_{\min}$, which depends only on the level of attention A. Its gradient with respect to the state vector $\boldsymbol{x} = [\boldsymbol{q}, \dot{\boldsymbol{q}}, A, T]^\top$ is:

$$\nabla h_A = \left[\frac{\partial h_A}{\partial \boldsymbol{q}},\ \frac{\partial h_A}{\partial \dot{\boldsymbol{q}}},\ \frac{\partial h_A}{\partial A},\ \frac{\partial h_A}{\partial T}\right] = \left[\mathbf{0},\ \mathbf{0},\ 1,\ 0\right].$$

The Lie derivatives along the drift vector field $\boldsymbol{f}(\boldsymbol{x})$ and the control input matrix $\boldsymbol{g}(\boldsymbol{x})$ are therefore:

$$L_f h_A = \nabla h_A \cdot \boldsymbol{f}(\boldsymbol{x}) = 1 \cdot f_A(\boldsymbol{x}) = -(\lambda_0 + \mu C)A + \gamma T(1-A),$$
$$L_{\boldsymbol{g}} h_A = \nabla h_A \cdot \boldsymbol{g}(\boldsymbol{x}) = 1 \cdot g_A(\boldsymbol{x}) = 1,$$

where $f_A(\boldsymbol{x})$ is the component of $\boldsymbol{f}(\boldsymbol{x})$ that corresponds to the dynamics of A, and $g_A(\boldsymbol{x})$ is the component of $\boldsymbol{g}(\boldsymbol{x})$ that multiplies the input to the attention regulation u_A.

The CBF condition for attention safety, $\dot{h}_A \geq -\alpha(h_A)$, leads to the linear inequality:

$$L_f h_A + L_g h_A u_A \geq -\alpha(h_A). \tag{22}$$

This can be written in the standard linear inequality form for the quadratic-programming problem as follows:

$$\boldsymbol{A}_A(\boldsymbol{x})\,\boldsymbol{u}_{\mathrm{A}} \le b_A(\boldsymbol{x}), \tag{23}$$

where

$$A_A(x) = -L_g h_A = -1,$$
$$b_A(x) = L_f h_A + \alpha_A(h_A).$$

To ensure the feasibility of the control optimization problem in complex environments, we adopt a controller design framework based on QP[30]. The general QP problem that unifies all safety constraints is formulated as follows:

$$\begin{aligned} \min_{\boldsymbol{u}} \quad & \|\boldsymbol{u} - \boldsymbol{u}_{\mathrm{nom}}\|^2 + \rho_{obs} s_{obs}^2 + \rho_v s_v^2 + \rho_A s_A^2 \\ \text{s.t.} \quad & \boldsymbol{A}_{\mathrm{obs}}(\boldsymbol{x})\,\boldsymbol{\tau}_r \le b_{\mathrm{obs}}(\boldsymbol{x}) + s_{obs}, \\ & \boldsymbol{A}_v(\boldsymbol{x})\,\boldsymbol{\tau}_r \le b_v(\boldsymbol{x}) + s_v, \\ & \boldsymbol{A}_A(\boldsymbol{x})\,\boldsymbol{u}_{\mathrm{A}} \le b_A(\boldsymbol{x}) + s_A, \\ & s_{obs}, s_v, s_A \ge 0. \end{aligned} \tag{24}$$

where the optimization variable is the augmented vector $[\mathbf{u}^\top, \mathbf{s}^\top]^\top$, where $\mathbf{u} = [\tau_r, u_A]^\top$ and $\mathbf{s} = [s_{\mathrm{obs}}, s_v, s_A]^\top$. The variables s_{obs}, s_v, and s_A denote independent slack variables. The positive constants ρ_{obs}, ρ_v, and ρ_A denote the penalty weights associated with the obstacle avoidance, velocity regulation, and cognitive safety constraints, respectively.

The QP is solved online at each sampling step, and its solution yields a continuous control input trajectory due to the continuous dependence of the optimization problem on system states.

Remark 10 *The proposed quadratic program employs independent slack variables for heterogeneous constraints, which enables selective relaxation when exact feasibility cannot be guaranteed. Compared with a shared slack-variable formulation, this design preserves the physical meaning of each constraint and avoids undesired coupling between unrelated safety requirements. The penalty parameters are selected according to a hierarchical safety-critical design principle, where obstacle avoidance constraints receive the highest penalty weight due to the risk of irreversible physical collisions, velocity constraints are assigned intermediate priority to ensure dynamic feasibility, and cognitive constraints are treated as soft performance-related constraints. This leads to the relation* $\rho_{obs} \gg \rho_v \gg \rho_A > 0$, *enforcing an implicit priority structure among heterogeneous constraints. From an optimization perspective, the resulting formulation can be interpreted as a relaxed hierarchical optimization problem, where the penalty coefficients act as implicit Lagrange multipliers that encode the priority ordering among constraints.*

This weighting structure does not affect the feasibility of the quadratic program, as the slack variables guarantee constraint relaxation whenever necessary. Consequently, the proposed formulation ensures real-time solvability of the optimization problem while preserving strict prioritization of safety-critical constraints in human–robot shared control systems.

3.2. Dynamic authority allocation via cognitive mapping

The cognitive state of the operator provides the basis for determining the desired human–robot authority allocation. Specifically, attention and trust states are first mapped into an unconstrained cognitive-based authority command:

$$\bar{\alpha}_{cog}(A,T) = \frac{1}{1+e^{-k_A(A-A_0)}} \cdot \frac{1}{1+e^{-k_T(T-T_0)}}, \tag{25}$$

where $k_A, k_T > 0$ are the sensitivity gains and $A_0, T_0 \in [0,1]$ are the threshold values for attention and trust, respectively.

The sigmoid functions provide a smooth nonlinear mapping from cognitive states to a preliminary human authority command. However, this preliminary mapping does not explicitly consider performance degradation and authority constraints, which are incorporated in the subsequent intervention mechanism.

3.3. Dynamic intervention mechanism

In human-robot collaborative control, the operator's attention is critical for maintaining operational safety. This section presents a dynamic intervention mechanism based on a real-time attention performance metric, $P(t)$. The mechanism integrates $P(t)$ with the collaborative model to actively monitor the operator's state. When attention degrades, it triggers interventions ranging from warning generation to autonomous safe takeover, thereby ensuring operational safety. This performance-driven authority adjustment is consistent with adaptive automation principles, with the objective of optimizing efficacy and safety of the collaboration.

To quantify the overall cognitive state performance, a scalar performance function $P(t)$ is constructed as follows:

$$P(t) = 1 - \frac{1}{2}\left[|\tanh(k_a e_A(t))| + |\tanh(k_t e_T(t))|\right] \tag{26}$$

where $k_a, k_t > 0$ are the sensitivity coefficients, and the tracking errors of attention and trust are defined as:

$$e_A(t) = A(t) - A_d, \tag{27}$$

$$e_T(t) = T(t) - T_d, \tag{28}$$

where A_d and T_d denote the desired attention and trust levels.

Remark 11 *The performance function $P(t)$ is bounded within the interval $(0,1]$, with $P(t) = 1$ indicating perfect alignment of the cognitive state with its desired values. Although the absolute-value operations introduce isolated non-differentiable points at $e_A = 0$ and $e_T = 0$, the performance function is used exclusively for authority scheduling and intervention assessment. It does not appear in the optimization variables or constraints of the QP problem. Consequently, the real-time QP solver operates on smooth affine constraints and is unaffected by the nonsmoothness of $P(t)$. In practice, the first-order authority dynamics further smooth the resulting authority evolution, thereby preventing abrupt switching or chattering near the target cognitive states.*

To implement a graded response to changes in the operator's cognitive state, two performance thresholds are defined: $0 < P_2 < P_1 < 1$. The performance index $P(t)$ is directly incorporated into the authority allocation mechanism through the modulation function $\sigma(P)$, allowing the target human authority to be adjusted according to real-time cognitive performance. These thresholds divide the range of the cognitive performance function $P(t)$ into three distinct operational regions. In the normal region ($P \geq P_1$), where the cognitive state of the operator is assessed as satisfactory, the dynamic control authority $\alpha(t)$ can track the desired authority level $\alpha_d(A,T,P)$, maintaining a consistently high yet bounded value to ensure that the human operator retains primary control. In the transition region ($P_2 < P < P_1$), which indicates mild cognitive degradation, $\alpha(t)$ continues to follow $\alpha_d(A,T,P)$ but remains in a high bounded range, shifting the system to a balanced shared human-robot control mode. Finally, in the intervention region ($P \leq P_2$), where cognitive performance is significantly degraded and poses a potential safety risk, the human authority $\alpha(t)$ is driven toward its minimum admissible value ε, resulting in robot-dominant control while preserving residual human participation. This three-layer structure enables a smooth and continuous transition from human-dominant to robot-dominant control based on real-time assessment of operator cognitive performance.

Remark 12 *For practical deployment, the intervention thresholds can be adjusted online according to task complexity to improve engineering adaptability in dynamic environments. Specifically, the thresholds are defined as:* $P_1(t) = P_{1,0} + k_{p1}C(t), P_2(t) = P_{2,0} + k_{p2}C(t)$, *where* $C(t)$ *denotes the normalized task complexity. As task complexity increases, the system becomes more conservative, leading to earlier intervention to ensure safety under demanding operational conditions.*

Since $C(t)$ *is bounded in practical implementations, i.e.,* $C(t) \in [0, 1]$, *the thresholds* $P_1(t)$ *and* $P_2(t)$ *remain bounded, which guarantees that the intervention switching logic remains well-defined and does not affect the stability of the overall closed-loop system.*

An intervention modulation function couples the cognitive performance assessment with the authority allocation mechanism by dynamically adjusting the cognitive-based authority command according to the real-time performance index $P(t)$. Specifically, the preliminary target human authority is obtained by scaling the cognitive-based authority command $\bar{\alpha}_{\mathrm{cog}}(A, T)$ with a performance modulation factor $\sigma(P)$, yielding

$$\bar{\alpha}_d(A, T, P) = \bar{\alpha}_{cog}(A, T)\sigma(P), \tag{29}$$

where the modulation factor $\sigma(P)$ is a piecewise-linear function of the performance index:

$$\sigma(P) = \begin{cases} 1, & P \geq P_1, \\ \dfrac{P - P_2}{P_1 - P_2}, & P_2 < P < P_1, \\ 0, & P \leq P_2. \end{cases} \tag{30}$$

The thresholds P_1 and P_2 define three cognitive operating regions corresponding to normal, degraded, and critical human cognitive states. The separation between P_1 and P_2 introduces a smooth transition region, which avoids abrupt changes in authority allocation and enables gradual switching between human-dominant and robot-dominant modes.

Although the performance-modulated authority command $\bar{\alpha}_d(A, T, P)$ provides a smooth adaptation mechanism according to the operator's cognitive condition, its value is not guaranteed to remain within the admissible authority range required by the safety controller. In particular, severe cognitive degradation may cause $\sigma(P)$ to approach zero, resulting in an excessively small desired human authority and violating the prescribed authority allocation constraint. To guarantee that both the human operator and the autonomous controller retain nonzero control authority, the target authority is further constrained within the admissible interval $[\varepsilon, 1-\varepsilon]$ by introducing a saturation operation:

$$\alpha_d(A, T, P) = \mathrm{sat}_{[\varepsilon, 1-\varepsilon]}(\bar{\alpha}_d(A, T, P)),$$
$$\mathrm{sat}_{[\varepsilon, 1-\varepsilon]}(x) = \min(1 - \varepsilon,\ \max(\varepsilon, x))$$

where $\varepsilon > 0$ denotes the minimum admissible authority reserved for each participant. This saturation operation ensures that $\alpha_d(A, T, P) \in [\varepsilon, 1 - \varepsilon]$, thereby preventing complete removal of human participation and preserving a nonvanishing autonomous intervention capability for safety enforcement.

After enforcing the admissible authority constraint, the resulting saturated target authority $\alpha_d(A,T,P)$ is tracked by the actual human control authority through a first-order dynamic system, ensuring smooth and continuous authority transitions:

$$\dot{\alpha} = -k_\alpha(\alpha - \alpha_d(A,T,P)). \tag{31}$$

Remark 13 *The proposed authority mechanism establishes a closed-loop coupling among the operator's cognitive state, performance-based intervention, and safety-constrained control execution. Specifically, the saturated target authority $\alpha_d(A,T,P)$ integrates the cognitive mapping and performance modulation through $\sigma(P)$ while satisfying $\alpha_d(A,T,P) \in [\varepsilon, 1-\varepsilon]$. The actual human control authority $\alpha(t)$ evolves according to the first-order dynamics in Equation (31). Therefore, if the initial authority satisfies $\alpha(0) \in [\varepsilon, 1-\varepsilon]$, the admissible authority interval remains forward invariant during the evolution of $\alpha(t)$, ensuring that both the human operator and the autonomous controller retain nonzero control authority.*

Moreover, the first-order authority dynamics prevent abrupt switching caused by instantaneous cognitive variations and provide a smooth transition between different HRC modes. When the cognitive performance index falls below the intervention threshold P_2, the target authority is driven toward its lower admissible bound ε, resulting in robot-dominant intervention while preserving residual human participation. At the implementation level, the final autonomous torque input τ_r^* is obtained by solving the safety-constrained QP problem, where the dynamically regulated authority allocation determines the human–robot input weighting while the CBF constraints enforce safety under degraded cognitive conditions.

3.4. Performance evaluation metrics

To comprehensively evaluate the performance of the proposed human–robot shared control framework, a unified set of performance metrics is introduced. These metrics aim to quantify both physical control performance and human–robot interaction quality, enabling a systematic evaluation of the inherent trade-off among safety, tracking accuracy, and cognitive adaptability.

The primary performance index is defined as the composite safety-performance index (CSPI), which integrates tracking accuracy, safety compliance, and control effort:

$$\mathrm{CSPI} = w_1\tilde{J}_e + w_2\tilde{J}_s + w_3\tilde{J}_u$$

where J_e represents the tracking error, J_s denotes safety constraint violation, and J_u measures the control effort. The weights w_1, w_2, and w_3 satisfy $w_1 + w_2 + w_3 = 1$. And $J_e = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\|\mathbf{q}(k) - \mathbf{q}_d(k)\|^2}$, $J_s = \frac{1}{N}\sum_{k=1}^{N}\max(0, -h_{\mathrm{obs}}(k))$, $J_u = \int_0^T \|\boldsymbol{\tau}(t)\|^2\,\mathrm{d}t$, $\tilde{J}_i = \frac{J_i}{\max_{m\in\mathcal{M}} J_i^{\mathrm{m}}}$, $i \in \{e,s,u\}$, let $\mathcal{M}$ denote the set of compared methods.

A lower CSPI indicates better overall system performance in terms of tracking accuracy, safety preservation, and energy efficiency.

To further evaluate the quality of human–robot interaction, the human–robot adaptation index (HRAI) is introduced to reflect the coordination efficiency between cognitive states and authority allocation.

$$HRAI = w_a\tilde{\mathrm{AAE}} + w_c\tilde{\mathrm{CCI}}$$

where the authority adaptation efficiency (AAE) is defined to quantify the temporal variation of control authority: $\mathrm{AAE} = \frac{1}{T}\int_0^T |\dot{\alpha}(t)|\, dt$, which reflects the responsiveness of authority adjustment over time. The cognition-consistency index (CCI) is defined as: $\mathrm{CCI} = \frac{1}{T}\sum|\alpha(t) - \alpha^*(A,T)|$, where $\alpha^*(A,T)$ denotes the cognition-driven optimal authority allocation. A lower CCI indicates stronger alignment between cognitive state and control authority. And $\tilde{\mathrm{AAE}}_i = \frac{\mathrm{AAE_i}}{\max_{m\in M}\mathrm{AAE}_i^{\mathrm{m}}}$, $\tilde{\mathrm{CCI}}_i = 1 - \frac{\mathrm{CCI_i}}{\max_{m\in M}\mathrm{CCI}_i^{\mathrm{m}}}$. To quantify the trade-off between physical performance and human–robot interaction quality, a unified metric is defined as:

$$\mathrm{CSPI}_{\mathrm{final}} = \mathrm{CSPI} + (1 - \mathrm{HRAI})$$

where $CSPI$ represents the normalized physical cost, and $(1 - HRAI)$ reflects the deficiency in human–robot adaptation. A lower $CSPI_final$ indicates a better trade-off between physical performance and human–robot coordination efficiency.

The proposed metrics are used solely for performance evaluation and do not influence the control design.

Remark 14 *The weighting coefficients used in the proposed performance evaluation metrics are selected as* $w_1 = 0.3$, $w_2 = 0.5$, $w_3 = 0.2$ *for CSPI, and* $w_a = w_c = 0.5$ *for HRAI. These values are chosen according to a safety-oriented design principle in human–robot shared control systems. Specifically, obstacle avoidance is assigned the highest weight due to its critical importance in preventing irreversible physical damage, while tracking performance is assigned a moderate weight reflecting its role in task execution. Control effort is assigned a lower weight since it primarily reflects energy consumption rather than safety-critical behavior. For the HRAI metric, AAE and CCI are assigned equal weights as they capture complementary aspects of interaction quality, namely responsiveness and cognitive alignment.*

These weights are not tuned to improve specific experimental results but are determined based on general principles of safety-critical control and human–robot interaction design. Furthermore, moderate variations in these coefficients do not affect the relative ranking among different methods, indicating that the proposed evaluation framework is robust to weight selection.

3.5. Stability analysis

This section presents a rigorous theoretical analysis of the proposed human–robot collaborative control framework. The stability guarantees are developed hierarchically across four aspects. First, the dynamic authority allocation mechanism is shown to be input-to-state stable (ISS) while remaining within its admissible range. Second, the associated safe sets are proven to be forward invariant, ensuring persistent collision avoidance and safe-state invariance. Third, all closed-loop signals, including physical states, cognitive states, and shared-control variables, are shown to remain uniformly bounded. Finally, the tracking error dynamics are established to be uniformly ultimately bounded under bounded human input disturbances.

We begin the theoretical analysis with the authority allocation dynamics, since the boundedness of the shared-control ratio plays a fundamental role in the subsequent safety and closed-loop stability results.

Proposition 1 (ISS and Forward Invariance of the Authority Allocation Dynamics) .

Consider the authority allocation dynamics

$$\dot{\alpha}(t) = -k_\alpha(\alpha(t) - \alpha_d(t)), \qquad k_\alpha > 0, \tag{32}$$

where the desired authority signal $\alpha_d(t) = \alpha_d(A(t), T(t), P(t))$ *is piecewise-continuously differentiable and satisfies* $\alpha_d(t) \in [\varepsilon, 1-\varepsilon], \forall t \geq 0,$ *for some constant* $0 < \varepsilon < 1/2.$

Then, for any initial condition $\alpha(0) \in [\varepsilon, 1-\varepsilon],$ *the following properties hold:*

1. *The tracking error* $e_\alpha(t) = \alpha(t) - \alpha_d(t)$ *is ISS with respect to the input* $\dot{\alpha}_d(t)$.
2. *If* $\dot{\alpha}_d(t) = 0,$ *then the equilibrium* $e_\alpha = 0$ *is exponentially stable.*
3. *The interval* $[\varepsilon, 1-\varepsilon]$ *is forward invariant; namely,* $\alpha(t) \in [\varepsilon, 1-\varepsilon], \forall t \geq 0.$

Proof. Define the tracking error $e_\alpha = \alpha - \alpha_d(t)$. From the authority dynamics Equation (32), the error dynamics are

$$\dot{e}_\alpha = \dot{\alpha} - \dot{\alpha}_d = -k_\alpha e_\alpha - \dot{\alpha}_d. \tag{33}$$

Consider the Lyapunov function

$$V_\alpha = \frac{1}{2}e_\alpha^2.$$

Its derivative along Equation (33) is

$$\dot{V}_\alpha = -k_\alpha e_\alpha^2 - e_\alpha\dot{\alpha}_d.$$

Using Young's inequality, for any $\mu > 0,$

$$|e_\alpha\dot{\alpha}_d| \leq \frac{\mu}{2}e_\alpha^2 + \frac{1}{2\mu}\dot{\alpha}_d^2.$$

Choosing $\mu = k_\alpha$, we obtain

$$\dot{V}_\alpha \leq -\frac{k_\alpha}{2}e_\alpha^2 + \frac{1}{2k_\alpha}\dot{\alpha}_d^2.$$

Equivalently,

$$\dot{V}_\alpha \leq -k_\alpha V_\alpha + \frac{1}{2k_\alpha}\dot{\alpha}_d^2.$$

Hence, the error dynamics are ISS with respect to the input $\dot{\alpha}_d$. In particular, if $\dot{\alpha}_d = 0$, then $e_\alpha = 0$ is exponentially stable.

Next, we prove forward invariance of $\Omega = [\varepsilon, 1-\varepsilon]$. At the lower boundary $\alpha = \varepsilon$, we have $\dot{\alpha} = -k_\alpha(\varepsilon - \alpha_d) \geq 0$, since $\alpha_d \geq \varepsilon$. At the upper boundary $\alpha = 1-\varepsilon$, we similarly have

$\dot{\alpha} = -k_\alpha((1-\varepsilon) - \alpha_d) \leq 0$, since $\alpha_d \leq 1-\varepsilon$.

Therefore, the vector field points inward on both boundaries. By Nagumo's theorem, the interval Ω is forward invariant. Thus, for any initial condition $\alpha(0) \in \Omega$, we have

$$\alpha(t) \in \Omega, \qquad \forall t \geq 0.$$

This completes the proof.

Remark 15 *Proposition 1 provides the theoretical basis for the proposed shared-control allocation law. The ISS property ensures that the implemented authority ratio $\alpha(t)$ smoothly follows the desired target $\alpha_d(t)$, thereby avoiding abrupt switching and improving interaction robustness. The forward-invariance property guarantees that $\alpha(t) \in [\varepsilon, 1-\varepsilon]$ for all time, so that both the human and robot always retain a nonzero level of authority. In particular, the lower bound ε guarantees that the autonomous controller always retains a nonvanishing control channel. Consequently, the control effectiveness term appearing in the robust CBF constraints remains nonzero, thereby preserving the robot's ability to influence the safety-critical dynamics. It should be emphasized that the forward invariance of the authority interval does not, by itself, imply the feasibility of the associated safety-constrained QP for arbitrary values of ε or for all possible disturbance realizations. The feasibility of the optimization problem additionally requires the existence of admissible control inputs satisfying the imposed safety constraints and actuator limitations, which is a standard assumption in CBF-based control frameworks. The admissibility of $\alpha_d(t)$ is ensured by construction through bounded cognitive states and a saturated target-mapping function.*

Next, we establish the forward invariance of the hard-constrained safe set under the proposed QP-based controller.

The closed-loop system, given by Equation (6), is expressed as

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) + \boldsymbol{g}(\boldsymbol{x})\boldsymbol{u}^*(\boldsymbol{x}) + \boldsymbol{d}(\boldsymbol{x}, t), \tag{34}$$

where $\boldsymbol{u}^*(\boldsymbol{x})$ is the optimal control input obtained from solving a QP problem, and $\boldsymbol{d}(\boldsymbol{x}, t)$ is the bounded disturbances due to human input, satisfying $\|\boldsymbol{d}(\boldsymbol{x}, t)\| \leq d_{\max}$. The QP-based controller is defined as:

$$\boldsymbol{u}^*(\boldsymbol{x}) = \arg\min_{\boldsymbol{u}} \|\boldsymbol{u} - \boldsymbol{u}_{\text{nom}}\|^2 \quad \text{subject to the CBF constraints}, \tag{35}$$

ensuring that the applied control minimally deviates from the nominal input while satisfying all safety constraints derived from the CBFs.

Theorem 1 (Forward Invariance of the Physical Safe Set) . *Consider the closed-loop system. Let the physical safe set be defined as:*

$$\mathcal{S}_p = \{\mathbf{x} \mid h_{obs}(\mathbf{x}) \geq 0,\ h_v(\mathbf{x}) \geq 0\}.$$

If the safety-constrained QP remains feasible for all $\mathbf{x} \in \mathcal{S}_p$, and the relaxation variables satisfy $s_{obs} = 0, s_v = 0$, then $\mathcal{S}_p$ is forward invariant. That is, $\mathbf{x}(0) \in \mathcal{S}_p \quad \Rightarrow \quad \mathbf{x}(t) \in \mathcal{S}_p,\ \forall t \geq 0$.

Proof. For each physical safety function $h_i(\mathbf{x}) \in \{h_{obs}, h_v\}$, the time derivative along the disturbed closed-loop dynamics is

$$\dot{h}_i(\boldsymbol{x}) = L_{\boldsymbol{f}} h_i(\boldsymbol{x}) + L_{\boldsymbol{g}} h_i(\boldsymbol{x})\boldsymbol{u}^*(\boldsymbol{x}) + L_{\boldsymbol{d}} h_i(\boldsymbol{x}),$$

where $L_{\boldsymbol{d}} h_i(\boldsymbol{x}) = \nabla h_i(\boldsymbol{x}) \boldsymbol{d}(\boldsymbol{x}, t)$. Using the Cauchy–Schwarz inequality and the disturbance bound, we have

$$|L_{\boldsymbol{d}} h_i(\boldsymbol{x})| \leq \|\nabla h_i(\boldsymbol{x})\| \|\boldsymbol{d}(\boldsymbol{x}, t)\| \leq \|\nabla h_i(\boldsymbol{x})\| d_{\max} = \delta_i(\boldsymbol{x}).$$

Consequently,

$$L_{\boldsymbol{d}} h_i(\boldsymbol{x}) \geq -\delta_i(\boldsymbol{x}).$$

Since $s_i = 0$, the QP solution satisfies the exact robust barrier condition

$$L_f h_i + L_g h_i \mathbf{u}^* \geq -\alpha_i(h_i) + \delta_i(\mathbf{x}).$$

Hence,

$$\dot{h}_i \geq -\alpha_i(h_i).$$

On the boundary $h_i = 0$, one has $\alpha_i(0) = 0$, thus

$$\dot{h}_i \geq 0.$$

By Nagumo's theorem, trajectories cannot leave $\mathcal{S}_p$ through the boundary. Since this holds for both physical constraints, the intersection set $\mathcal{S}_p$ remains forward invariant.

Remark 16 *The robust CBF condition explicitly compensates for worst-case bounded disturbances through the margin term $\delta_i(x)$. By assigning zero slack variables to hard physical constraints, strict forward invariance of the physical safe set is preserved. Soft constraints, such as cognitive-performance limits, may be temporarily relaxed through their associated slack variables to maintain QP feasibility. This hard-soft decomposition guarantees safety while improving the robustness and feasibility of the optimization-based controller.*

Building upon the previous invariance results, we now analyze the boundedness of all closed-loop signals.

Theorem 2 (Uniform Boundedness of Closed-Loop Signals) . *For any admissible initial condition satisfying the hard safety constraints, all closed-loop signals, including robot states, cognitive states, and authority allocation variables, remain uniformly bounded for all $t \geq 0$.*

Proof. The proof proceeds in four steps: construction of a composite Lyapunov function, analysis of its time derivative, bounding of the cross terms, and concluding global boundedness.

Construction of a composite Lyapunov function, define the tracking errors for the physical subsystem:

$$\boldsymbol{e} = \boldsymbol{q} - \boldsymbol{q}_d, \qquad \dot{\boldsymbol{e}} = \dot{\boldsymbol{q}} - \dot{\boldsymbol{q}}_d. \tag{36}$$

Consider the following positive definite and radially unbounded Lyapunov function candidate:

$$V(\boldsymbol{x}) = V_p(\boldsymbol{e}, \dot{\boldsymbol{e}}) + V_c(A, T), \tag{37}$$

where

$$V_p(\boldsymbol{e}, \dot{\boldsymbol{e}}) = \frac{1}{2}\dot{\boldsymbol{e}}^\top \boldsymbol{M}(\boldsymbol{q})\dot{\boldsymbol{e}} + \frac{1}{2}\boldsymbol{e}^\top \mathbf{K}_p \boldsymbol{e}, \tag{38}$$

$$V_c(A, T) = \frac{1}{2}(A - A_d)^2 + \frac{1}{2}(T - T_d)^2. \tag{39}$$

The inertia matrix $\boldsymbol{M}(\boldsymbol{q})$ is known to be uniformly positive and definite and bounded for all $\boldsymbol{q}$; that is, there exist constants $m_1, m_2 > 0$ such that

$$m_1 \boldsymbol{I} \le \boldsymbol{M}(\boldsymbol{q}) \le m_2 \boldsymbol{I}.$$

Meanwhile, the gain matrix $\mathbf{K}_p$ is designed to be positive definite. Denote by $\lambda_{\min}(\mathbf{K}_p)$ and $\lambda_{\max}(\mathbf{K}_p)$ the minimum and maximum eigenvalues of $\mathbf{K}_p$, respectively. For the state vector $z = [\boldsymbol{e},\ \dot{\boldsymbol{e}}]^\top$, the following inequality can be derived:

$$\frac{1}{2}\min(m_1, \lambda_{\min}(\mathbf{K}_p))\, \|z\|^2 \ \le\ V_p(\boldsymbol{e}, \dot{\boldsymbol{e}}) \ \le\ \frac{1}{2}\max(m_2, \lambda_{\max}(\mathbf{K}_p))\, \|z\|^2.$$

Consequently, there exist positive constants

$$c_1 = \frac{1}{2}\min(m_1, \lambda_{\min}(\mathbf{K}_p)), \qquad c_2 = \frac{1}{2}\max(m_2, \lambda_{\max}(\mathbf{K}_p)),$$

such that

$$c_1\|z\|^2 \ \le\ V_p \ \le\ c_2\|z\|^2, \qquad \forall z.$$

This shows that V_p is positive definite and radially unbounded, and its upper and lower bounds can be expressed explicitly in terms of the physical parameters of the system. Moreover, since attention and trust levels are bounded by construction ($A, T \in [0, 1]$), the cognitive part satisfies $V_c \le 1$.

Using robot dynamics $\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{q}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{q}} + \boldsymbol{G}(\boldsymbol{q}) = (1 - \alpha)\boldsymbol{\tau}_r^* + \alpha\boldsymbol{\tau}_h$, the error dynamics can be written as

$$\boldsymbol{M}(\boldsymbol{q})\ddot{\boldsymbol{e}} + \boldsymbol{C}(\boldsymbol{q}, \dot{\boldsymbol{q}})\dot{\boldsymbol{e}} + \mathbf{K}_d\dot{\boldsymbol{e}} + \mathbf{K}_p\boldsymbol{e} = \boldsymbol{\Delta}, \tag{40}$$

The disturbance term $\boldsymbol{\Delta}$ aggregates the correction of the autonomous control input and the effect of the human operator's input:

$$\boldsymbol{\Delta} = (1 - \alpha)(\boldsymbol{\tau}_r^* - \boldsymbol{\tau}_{\mathrm{r,nom}}) + \alpha\boldsymbol{\tau}_h. \tag{41}$$

Here, $(\boldsymbol{\tau}_r^* - \boldsymbol{\tau}_{\mathrm{r,nom}})$ denotes the safety correction generated by the QP relative to the authority-compensated nominal robot command.

Exploiting the skew-symmetry property $\dot{\boldsymbol{M}} - 2\boldsymbol{C}$ and differentiating Equation (38) yields

$$\dot{V}_p = -\dot{\boldsymbol{e}}^\top \mathbf{K}_d \dot{\boldsymbol{e}} + \dot{\boldsymbol{e}}^\top \boldsymbol{\Delta}. \tag{42}$$

The first term on the right-hand side, $-\dot{\boldsymbol{e}}^\top \mathbf{K}_d \dot{\boldsymbol{e}} \le -\lambda_{\min}(\mathbf{K}_d)\|\dot{\boldsymbol{e}}\|^2$, is negative definite and therefore contributes to stability. The second term, $\dot{\boldsymbol{e}}^\top \boldsymbol{\Delta}$, represents the coupling between the disturbance and the error, and its boundedness must be analyzed.

From Theorem 1, the hard physical safe set $\mathcal{S}_h$ is forward invariant. Hence, the physical states $(q, \dot{q})$ evolve in a closed and bounded admissible region. Moreover, the cognitive variables satisfy $A, T \in [0, 1]$, and by Proposition 1, $\alpha(t) \in [\varepsilon, 1 - \varepsilon]$. Therefore, the closed-loop state remains in a compact operating domain $\Omega \subset \mathbb{R}^n$, on which both $u^*(x)$ and $u_{nom}(x)$ are continuous. Hence there exists $c_u > 0$ such that $\|u^*(x) - u_{nom}(x)\| \leq c_u, \forall x \in \Omega.$. Together with the bounded human input $\|\tau_h\|_2 \leq \tau_{h,\max}$, we obtain

$$\|\mathbf{\Delta}\| \leq (1 - \alpha)c_u + \alpha\tau_{h,\max} \leq c_3, \tag{43}$$

for some constant $c_3 > 0$. Applying Young's inequality to the cross term in Equation (42) gives, for any $\lambda > 0$,

$$\dot{\boldsymbol{e}}^\top \mathbf{\Delta} \leq \frac{\lambda}{2}\|\dot{\boldsymbol{e}}\|^2 + \frac{1}{2\lambda}\|\mathbf{\Delta}\|^2. \tag{44}$$

Substituting Equation (44) into Equation (42) and using the positive definiteness of $\mathbf{K}_d$ produces

$$\begin{aligned}\dot{V}_p &\leq -\lambda_{\min}(\mathbf{K}_d)\|\dot{\boldsymbol{e}}\|^2 + \frac{\lambda}{2}\|\dot{\boldsymbol{e}}\|^2 + \frac{1}{2\lambda}c_3^2 \\ &= -\left(\lambda_{\min}(\mathbf{K}_d) - \frac{\lambda}{2}\right)\|\dot{\boldsymbol{e}}\|^2 + \frac{1}{2\lambda}c_3^2.\end{aligned}$$

Choosing $\lambda = \lambda_{\min}(\mathbf{K}_d) > 0$ and defining the positive constants

$$c_4 = \frac{\lambda_{\min}(\mathbf{K}_d)}{2}, \qquad c_5 = \frac{c_3^2}{2\lambda_{\min}(\mathbf{K}_d)},$$

we obtain the compact derivative inequality

$$\dot{V}_p \leq -c_4\|\dot{\boldsymbol{e}}\|^2 + c_5.$$

From the cognitive dynamics Equations (2) and (3) and the boundedness of the input to the regulation u_A (forced by the QP constraints), the derivatives $\dot{A}$ and $\dot{T}$ are continuous and bounded on the compact set $[0, 1]^2$. Consequently,

$$|\dot{V}_c| = |(A - A_d)\dot{A} + (T - T_d)\dot{T}| \leq c_6, \tag{45}$$

for some constant $c_6 > 0$.

Composite Lyapunov inequality and conclusion of boundedness, combining the bound on $\dot{V}_p$ and the bound on $\dot{V}_c$, the total derivative satisfies

$$\dot{V} = \dot{V}_p + \dot{V}_c \leq -c_4\|\dot{\boldsymbol{e}}\|^2 + c_5 + c_6 = -c_4\|\dot{\boldsymbol{e}}\|^2 + c, \tag{46}$$

We have $c_1\|z\|^2 \leq V_p$ with $z = [\boldsymbol{e},\ \dot{\boldsymbol{e}}]^\top$. Since $\|\dot{\boldsymbol{e}}\|^2 \leq \|z\|^2$, substituting this into Equation (46) yields

$$\dot{V} \leq -\frac{c_4}{c_1}V_p + c.$$

Recalling that $V = V_p + V_c$ and $V_c \geq 0$, it follows that $V_p \leq V$. Therefore, we obtain a differential inequality for the composite Lyapunov function V:

$$\dot{V} \leq -\kappa V + c, \qquad \text{where } \kappa = \frac{c_4}{c_1} > 0. \tag{47}$$

Equation (47) is the standard form for the uniform ultimate boundedness (UUB). Applying the comparison lemma gives the explicit upper bound for $V(t)$:

$$V(t) \le V(0)e^{-\kappa t} + \frac{c}{\kappa}(1 - e^{-\kappa t}) \le V(0) + \frac{c}{\kappa}, \qquad \forall t \ge 0.$$

The hard CBF constraints guarantee that the robot position remains inside the safe workspace and collision avoidance conditions are preserved for all time. If the velocity constraint is treated as a hard constraint, then the prescribed velocity bound is also maintained. Soft cognitive constraints may experience temporary bounded relaxation through their associated slack variables, but the variables $A, T \in [0, 1]$ remain bounded by construction. Furthermore, the authority allocation dynamics guarantee $\alpha(t) \in [\varepsilon, 1 - \varepsilon]$. Therefore, all closed-loop signals remain uniformly bounded for all future time. This completes the proof of Theorem 2.

Finally, the tracking performance of the physical subsystem is characterized through UUB of the tracking errors.

Theorem 3 (UUB of Tracking Errors) . *Under bounded human input disturbances, the tracking error e of the system is ultimately uniformly bounded.*

Proof. From Theorem 2, all closed-loop signals are uniformly bounded, and the composite Lyapunov function satisfies

$$\dot{V} \le -\kappa V + c.$$

Since

$$V = V_p + V_c, \qquad V_c \ge 0,$$

it follows that

$$V_p \le V.$$

Moreover, from the quadratic bounds of V_p,

$$c_1 \|z\|^2 \le V_p \le c_2 \|z\|^2, \qquad z = [e, \dot{e}]^\top.$$

Therefore,

$$\|z(t)\|^2 \le \frac{1}{c_1} V(t) \le \frac{1}{c_1}\left(V(0)e^{-\kappa t} + \frac{c}{\kappa}\right).$$

Hence,

$$\limsup_{t\to\infty} \|z(t)\| \le \sqrt{\frac{c}{\kappa c_1}}.$$

Thus, the tracking error state is uniformly ultimately bounded.

Remark 17 *The task complexity $C(t)$ introduces a time-varying effect into the cognitive dynamics and the authority allocation mechanism through the attention evolution equation. Since $C(t)$ is bounded, the term $-\mu C(t) A$ can be regarded as a uniformly bounded perturbation in the closed-loop system. Therefore, the overall system can be interpreted as a nonlinear system with a bounded time-varying disturbance. The Lyapunov function derivative derived* *in Theorem 2 remains valid, and the bounded perturbation induced by $C(t)$ only affects the convergence rate but does not alter the stability property. Consequently, the previously established ISS and UUB results still hold in the presence of task-complexity variations.*

Theorem 3 further shows that bounded human intervention and safety-driven control corrections do not compromise closed-loop stability, but only enlarge the residual tracking error bound. In the nominal low-disturbance case, the ultimate tracking radius becomes small, yielding high-accuracy motion tracking. This robustness property is particularly important for human-robot shared control systems subject to uncertain operator actions and time-varying task complexity.

Collectively, the foregoing results establish the main theoretical properties of the proposed human–robot collaborative control framework. The authority-allocation dynamics remain stable and confined to the admissible sharing interval, the hard-constrained safe set is forward invariant under the QP-based controller, all closed-loop signals remain uniformly bounded, and the tracking errors are uniformly ultimately bounded in the presence of bounded human disturbances. Therefore, the proposed scheme guarantees safety, stability, and robustness of the overall human–robot system, providing a rigorous theoretical foundation for the simulation studies presented in the next section.

4. SIMULATION

This section presents simulations conducted on the MATLAB platform to validate the effectiveness of the proposed robust CBF safety control framework and the cognitive performance-based dynamic intervention mechanism.

4.1. Comprehensive simulation

This experiment is designed to comprehensively validate the proposed integrated cognition–physical modeling framework, the robust CBF-QP-based safety controller, and the cognitive-driven dynamic intervention mechanism. The simulation considers a complete HRC scenario that incorporates robot dynamics, safety constraints, operator cognitive evolution, external disturbances, and the proposed control architecture in a unified setting. To evaluate the effectiveness of the proposed method, comparative simulations are conducted between the full proposed framework and two baseline cases, including a fixed-weight CBF-based shared control strategy and a nominal controller without CBF constraints, enabling a systematic assessment of safety, tracking performance, and cognitive adaptability. Table 1 lists the relevant simulation parameters.

To evaluate the response of the proposed framework under time-varying disturbances, two prescribed disturbance intervals are introduced during the simulation, namely $t \in [8, 10]$ s and $t \in [18, 20]$ s. During these intervals, external disturbances are applied to the cognitive system, resulting in temporary degradation of the operator's attention and trust states. The same disturbance intervals are used consistently across all related simulations unless otherwise stated.

As shown in Figure 1, the proposed method successfully achieves collision-free trajectory tracking in the presence of environmental obstacles. Compared with the baseline method without CBF constraints, which exhibits significant safety violations, the proposed controller ensures strict adherence to the safety boundary. The fixed-weight CBF method also guarantees safety but produces overly conservative trajectories due to the lack of adaptive authority allocation. In contrast, the proposed cognition-aware robust CBF framework achieves a better balance between safety and tracking performance by dynamically adjusting control authority based on human cognitive state and task conditions.

Table 1. Simulation parameters

Parameter	Value	Unit
Robot Physical Parameters		
m_1, m_2	1.0, 0.5	kg
l_1, l_2	1.0, 0.7	m
g	9.81	m/s^2
Safety Parameters		
$\boldsymbol{p}_{\text{obs}}$	[1.2; 0.3]	m
d_{safe}	0.3	m
$v_{\max}$	4.0	rad/s
$A_{\min}$	0.4	-
$\tau_{h,\max}$	1	N m
Cognitive Model Parameters		
λ_0	0.1	-
μ	0.05	-
γ	0.15	-
β_1	0.2	-
β_2	0.3	-
β_3	0.5	-
T_{base}	0.3	-
A_d	0.85	-
T_d	0.8	-
Control Authority Parameters		
A_0	0.6	-
T_0	0.5	-
k_A	10	-
k_T	10	-
k_a	1.0	-
k_t	2.0	-
k_α	3.0	-
ε	0.15	-
Dynamic Intervention Thresholds		
P_1	0.7	-
P_2	0.5	-
Cognitive Regulator Parameters		
k_1	1.1	-
k_2	1.0	-
k_3	1.0	-
k_4	0.2	-
Simulation Time		
Δt	0.01	s
T_{total}	30	s

The deviation of the experimental trajectory from the nominal circular reference, including the semi-circular shape and the absence of full re-convergence after obstacle avoidance, is expected in the CBF-QP framework.

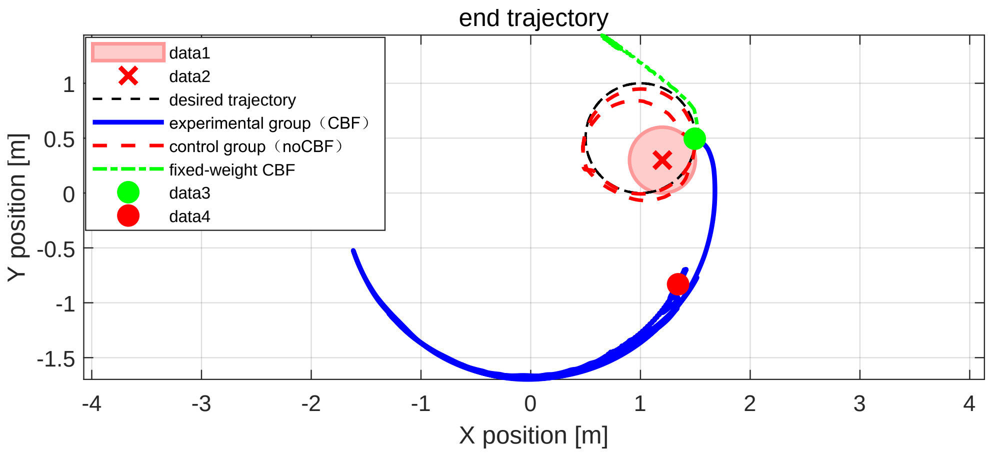


Figure 1. Comparison of end-effector trajectories in the task space. CBF: Control barrier function.

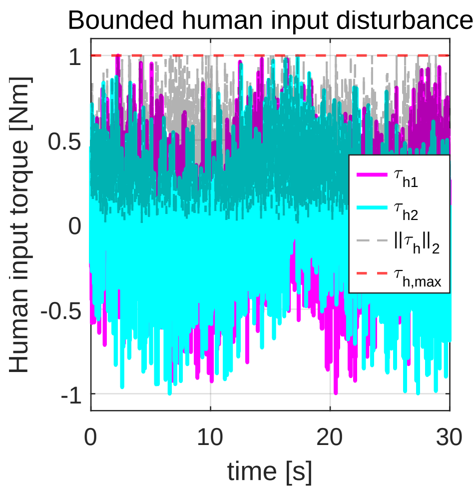
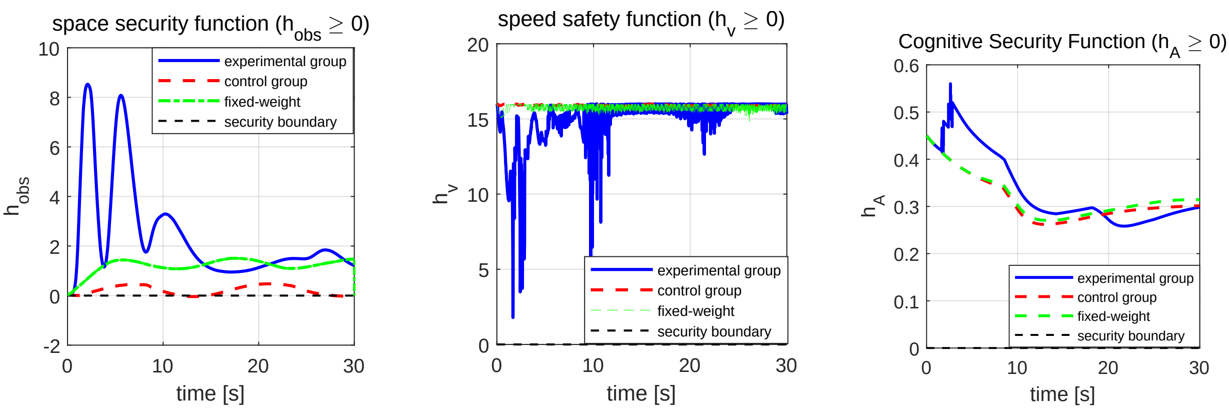


Figure 2. Bounded human input disturbance. The two joint torque components τ_{h1} and τ_{h2} are generated under the constraint $\|\tau_h\|_2 \le \tau_{h,\max} = 1\ \mathrm{N \cdot m}$.

This is because safety constraints are enforced as hard constraints and become active when the system approaches the obstacle boundary, temporarily overriding trajectory tracking objectives. After bypassing the obstacle, the controller operates in a locally optimal safe regime rather than performing global trajectory re-planning, leading to a safe but non-identical recovery path. In this work, task completion is defined as safe traversal along the reference workspace rather than exact reproduction of the nominal geometric path.

Figure 2 illustrates the evolution of the human input disturbance $\tau_h(t)$ over a 30-second simulation period. The simulation emulates the unexpected control inputs exerted by the human operator during a human-robot collaborative task. The two disturbance profiles exhibit bounded randomness, simulating the uncertain, time-varying, and non-ideal nature of human operation in real HRC. This provides a physical basis for the subsequent design of the robust CBF controller, which is required to guarantee system safety in the presence of bounded human disturbances.

As shown in Figure 3, the proposed cognition-aware robust CBF framework ensures simultaneous satisfaction of spatial, kinematic, and cognitive safety constraints under bounded human input disturbances. The spatial



(A) Space security function **(B)** Speed safety function **(C)** Cognitive security function

Figure 3. Security function.

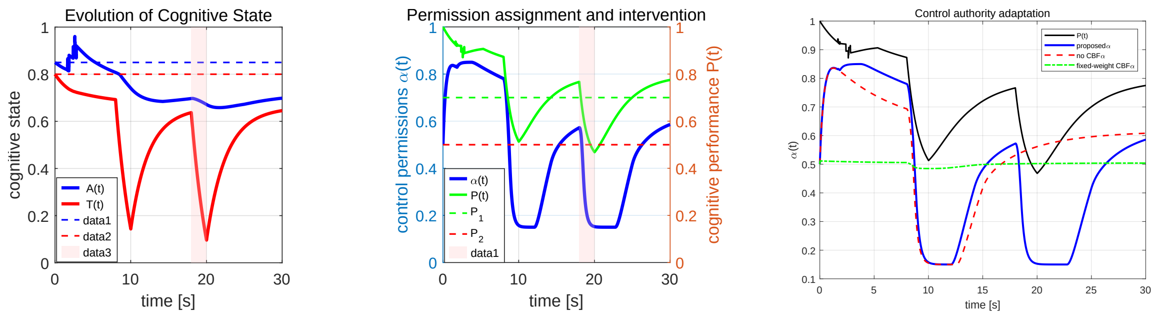
barrier function h_{obs} remains non-negative throughout the task, demonstrating strict obstacle avoidance. The velocity barrier function h_v exhibits transient degradation during obstacle avoidance but recovers once safety constraints are satisfied. The cognitive barrier function h_A represents the safety margin of the operator's attention level relative to the prescribed minimum threshold. Although trust T affects the attention dynamics through the cognitive coupling term, it is not directly represented by h_A. The non-negativity of h_A therefore confirms that the attention safety requirement is maintained throughout the task.

The large initial fluctuation observed in all safety-related functions is attributed to transient controller initialization and active-set switching in the CBF-QP optimization process before convergence to a steady feasible region. Furthermore, the sharp variations occurring in the intervals of 8-10 s and 18-20 s correspond to two external disturbance events injected into the cognitive system, which temporarily affect attention and trust dynamics, leading to momentary degradation in h_A, h_v, and h_{obs}. After each disturbance event, the proposed authority allocation mechanism and robust CBF constraints restore system stability, ensuring recovery to a safe operating regime.

The observed differences among the three methods can be attributed to their distinct control architectures. The baseline method without CBF lacks explicit safety constraints, resulting in occasional violations under disturbances. The fixed-weight CBF approach enforces safety in a conservative manner due to constant authority allocation, leading to reduced performance flexibility. In contrast, the proposed cognition-aware framework adaptively adjusts control authority based on real-time cognitive states, enabling a better balance between safety preservation and performance recovery under time-varying disturbances.

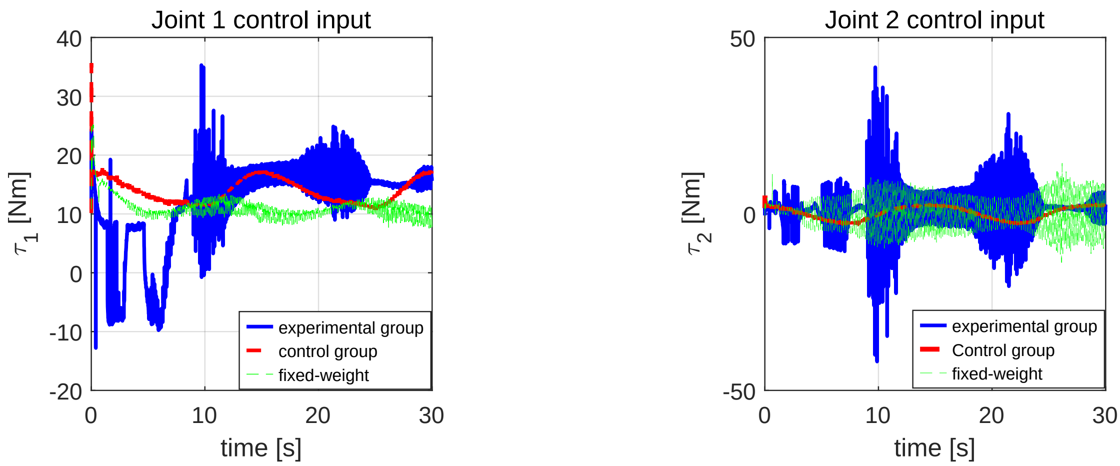
As shown in Figure 4, the proposed cognition-aware intervention mechanism dynamically adjusts control authority based on the evolution of cognitive states. The permission assignment $\alpha(t)$ exhibits clear adaptive switching behavior in response to variations in cognitive performance $P(t)$, with two distinct intervention phases observed during the disturbance intervals of 8–10 s and 18–20 s. These abrupt variations are intentionally introduced in the simulation as external disturbance events injected into the cognitive system to evaluate the robustness of the proposed framework. During these periods, reductions in $P(t)$ trigger a decrease in $\alpha(t)$, indicating increased robot intervention to maintain task safety and performance.

The evolution of cognitive states further demonstrates the effectiveness of the proposed framework. Specifically, the trust variable $T(t)$ shows higher sensitivity to system failure events, exhibiting sharp declines followed by grad-



A B C

Figure 4. Interplay between cognitive dynamics and control allocation.



(A) Joint 1 control input **(B)** Joint 2 control input

Figure 5. Joint control input.

ual recovery, while the attention state $A(t)$ evolves more smoothly under continuous workload and intervention effects. This asymmetric response highlights the distinct roles of attention and trust in cognitive regulation.

Overall, the results confirm that the proposed method achieves a closed-loop interaction between cognitive state evolution and authority allocation, enabling adaptive intervention under time-varying task conditions and deliberately introduced disturbance scenarios.

As shown in Figure 5, the control inputs of both joints exhibit distinct behaviors under different methods. The proposed method demonstrates higher control activity during the initial transient phase and disturbance intervals, which is attributed to the active enforcement of safety constraints through the CBF-QP framework. In particular, sharp variations in the control torques during the disturbance intervals of 8-10 s and 18-20 s correspond to external disturbance events, requiring rapid reallocation of control authority to maintain system safety. After these intervals, the control inputs gradually converge to smoother profiles as the system enters a safe operating region.

Compared with the baseline without CBF constraints, which produces smoother but unsafe control signals, and the fixed-weight CBF method, which exhibits conservative but less adaptive behavior, the proposed approach achieves a balance between responsiveness and safety assurance under bounded human input disturbances.

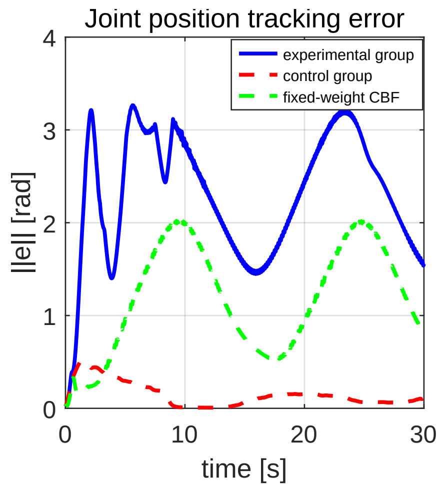


Figure 6. Joint position tracking error. CBF: Control barrier function.

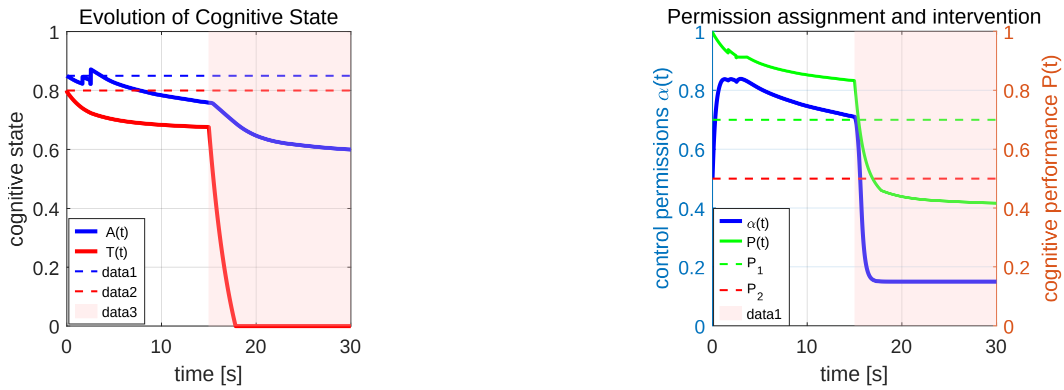
As shown in Figure 6, the joint position tracking error varies significantly among different control strategies. The no-CBF method achieves the lowest tracking error due to the absence of safety constraints, allowing the controller to strictly follow the nominal trajectory. However, this comes at the cost of safety violations in constrained environments. The fixed-weight CBF method introduces moderate tracking deviations, resulting from the constant trade-off between safety enforcement and tracking performance.

In contrast, the proposed cognition-aware robust CBF framework exhibits larger tracking error, which is primarily induced by the activation of safety constraints and adaptive authority allocation under cognitive disturbances. In particular, the pronounced error peaks observed during the disturbance intervals of 8-10 s and 18-20 s correspond to two external disturbance events injected into the system, during which the CBF-QP controller actively reconfigures its control authority to maintain safety, leading to temporary degradation in tracking performance. After each disturbance phase, the tracking error gradually decreases as the system returns to a safe and feasible operating regime.

Despite increased tracking deviation, the proposed method ensures strict satisfaction of safety constraints and bounded system behavior, highlighting the inherent trade-off between safety and tracking accuracy in constrained control systems.

The aforementioned simulation results were all obtained under scenarios involving two sudden disturbance events. To further validate the effectiveness of the multi-factorial influence and intervention mechanism of cognitive decline, the following two figures present the results obtained under a scenario where a system fault occurs around 15 s.

As shown in Figure 7, the proposed cognition-aware intervention framework effectively responds to a system fault injected at $t = 15$ s. Prior to the fault, the human control authority $\alpha(t)$ remains relatively high, indicating balanced human–robot shared control. Once the system failure occurs, the cognitive performance $P(t)$ significantly decreases, triggering a gradual reduction in $\alpha(t)$, which corresponds to an increased level of robot intervention authority. When $P(t)$ further drops below the critical threshold P_2, the intervention mechanism is fully activated, resulting in control authority being transferred to the robotic system to the maximum extent, where $\alpha(t)$ is maintained at its minimum value to ensure safety and system stability. After the fault event, the system stabilizes in a degraded but safe operating regime, with $\alpha(t)$ remaining at a lower level to maintain robustness under persistent uncertainty.



(A) Evolution of cognitive state

(B) Permission assignment and intervention

Figure 7. Evolution of cognitive performance and dynamic authority allocation under fault-induced cognitive degradation.

Table 2. Sensitivity analysis under different $\tau_{h_{max}}$

$\tau_{h_{max}}$ (Nm)	$\bar{\delta}_{\text{obs}}$	$\delta_{\text{obs}_{max}}$	RMS error	Safety violation count
0.2	0.65	2.91	1.27	0
0.5	1.41	6.69	1.95	0
1.0	2.01	10.24	2.39	0
1.5	2.83	11.95	2.43	0

RMS: Root mean square.

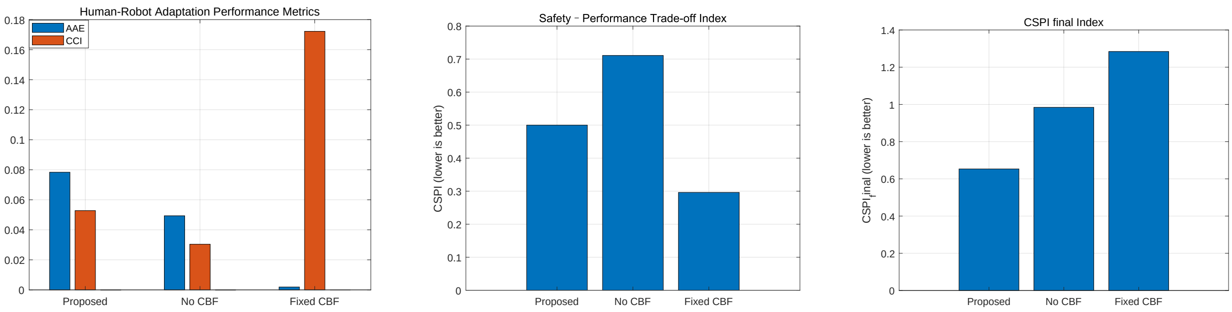
The evolution of cognitive states further confirms the effectiveness of the proposed mechanism. Specifically, the trust variable $T(t)$ exhibits a sharp collapse immediately after the fault occurrence, reflecting its high sensitivity to system failures, while the attention state $A(t)$ decreases more gradually due to its inertia-driven dynamics. Although partial recovery is observed in later stages, the cognitive system does not fully return to its initial state, indicating a persistent impact of system failure on human–robot interaction dynamics. These results demonstrate that the proposed framework can effectively capture and respond to fault-induced cognitive degradation through adaptive authority reallocation.

4.2. Sensitivity analysis of robust margin

To evaluate the influence of the assumed upper bound of human-input disturbances on the proposed robust CBF framework, a sensitivity analysis is conducted by varying the disturbance bound parameter $\tau_{h,\max}$ while keeping all other controller parameters unchanged. Four cases are considered here, namely $\tau_{h,\max} = \{0.2, 0.5, 1, 1.5\}$ Nm. As shown in Table 2, for each case, the root-mean-square tracking error, the number of safety violations, and the average and maximum values of the robustness margin δ_{obs} are recorded.

It can be observed that both δ_{obs} and $\delta_{\text{obs}_\text{m}\text{ax}}$ increase monotonically with respect to $\tau_{h_{max}}$, which is consistent with the theoretical formulation that the robust CBF margin scales proportionally with the upper bound of the human-input disturbance. Specifically, δ_{obs} increases from 0.65 to 2.83, while $\delta_{\text{obs}_\text{m}\text{ax}}$ increases from 2.91 to 11.95, indicating that the proposed method effectively enlarges the safety margin to compensate for stronger external disturbances.

Meanwhile, the root mean square (RMS) tracking error shows a gradual increase from 1.27 to 2.43 as $\tau_{h_{max}}$ grows. This behavior is expected since a larger disturbance bound leads to more conservative safety constraints, resulting in a slight reduction in tracking accuracy due to increased safety-oriented control effort. Notably, the



A B C

Figure 8. Comparison of the safety-performance trade-off index. AAE: Authority adaptation efficiency; CCI: cognition-consistency index; CBF: control barrier function; CSPI: composite safety-performance index.

growth of RMS error tends to saturate at higher disturbance levels, suggesting that the proposed framework maintains stable tracking performance even under strong uncertainty.

Importantly, the safety violation count remains zero across all tested cases, demonstrating that the proposed robust CBF-QP controller consistently preserves forward invariance of the safe set. Overall, these results verify that the proposed method achieves a desirable trade-off between robustness, safety, and tracking performance under increasing human-input uncertainty.

4.3. Comparison of performance evaluation metrics

To comprehensively evaluate the performance of the proposed method, this subsection presents a quantitative comparison of the CSPI, AAE, and CCI metrics proposed in Section 3.4 under different control strategies. Because human–robot shared control inherently involves a trade-off between tracking performance and safety constraints, a single metric is insufficient to assess the overall system behavior. Therefore, the composite index $\mathrm{CSPI}_{\mathrm{final}}$ is introduced to provide a holistic evaluation of the safety–performance trade-off. The numerical values of the relevant metrics and the data comparison of the final $\mathrm{CSPI}_{\mathrm{final}}$ index are shown in the figure below.

Figure 8A illustrates the human–robot adaptation performance in terms of AAE and CCI under three control strategies, including the proposed method, the no-CBF baseline, and the fixed-weight CBF method. It can be observed that the proposed method achieves a balanced trade-off between AAE and cognition consistency. Specifically, compared with the fixed-weight CBF method, the proposed approach slightly increases AAE but significantly reduces CCI, indicating improved efficiency in maintaining cognition-consistent interaction while preserving acceptable authority adaptation performance. In contrast, the fixed-weight CBF baseline achieves lower AAE at the expense of substantially higher cognitive inconsistency, while the no-CBF method exhibits moderate cognitive burden but weaker overall coordination stability. These results demonstrate the superiority of the proposed adaptive strategy in balancing human–robot interaction performance.

Figure 8B presents the Safety–Performance Trade-off Index, which integrates tracking accuracy, safety violations, and control effort into a unified metric to evaluate the overall system performance across different control strategies. As shown in the figure, the fixed-weight CBF method achieves the lowest trade-off index, indicating that it provides strong safety enforcement with relatively low overall cost. However, this improvement is mainly attributed to its conservative safety regulation, while the fixed authority allocation lacks the capability to adaptively adjust the human–robot control balance according to cognitive variations. The proposed method achieves a slightly higher trade-off index than the fixed-weight CBF approach but significantly outperforms the no-CBF baseline, demonstrating its capability to achieve an effective balance between safety preservation and task performance. More importantly, the proposed cognition-aware adaptive authority allocation provides enhanced adaptability and cognitive-aware intervention capability, which enables dynamic adjustment of human–robot interaction and avoids overly conservative behaviors, making it more suitable for long-term HRC scenarios.

Figure 8C presents the final CSPI, providing a unified metric for overall system evaluation across different control strategies. Results indicate that the proposed method achieves the lowest CSPI value, demonstrating the best overall trade-off among tracking performance, safety assurance, and control effort. In contrast, the no-CBF baseline and fixed-weight CBF method yield higher CSPI values, due to safety violations and overly conservative control behavior, respectively. These results further confirm the effectiveness of the proposed approach in achieving balanced and efficient HRC.

Overall, the results in Figure 8A-C consistently demonstrate that the proposed method achieves a superior balance between human cognitive adaptation and robot safety control. By introducing a unified composite evaluation framework, the proposed approach effectively quantifies the inherent trade-off between safety enforcement and performance degradation, thereby providing greater interpretability and comparability compared with conventional methods.

5. CONCLUSIONS

This paper investigated safety-critical HRC under uncertain human inputs and fluctuating operator cognition by developing a unified cognitive-physical control framework. A robust CBF-QP controller was designed to enforce obstacle avoidance, velocity limits, and cognitive-state safety constraints, while an adaptive authority allocation mechanism enabled smooth transitions between shared control and autonomous intervention. Theoretical analysis established stable, admissible authority allocation, forward invariance of the hard-constrained safe set, uniform boundedness of all closed-loop signals, and uniformly ultimately bounded tracking errors under bounded disturbances. Simulation studies verified the effectiveness of the proposed method in mitigating cognitive risk and preserving safety and performance. Future work will consider multimodal cognitive sensing, online adaptation, dynamic environments, multi-robot coordination, and real-world experimental validation.

DECLARATIONS

Authors' contributions

Conception and design of the study: Yang, Y.
Manuscript writing: Li, Z.
Manuscript review and correction: Jiang, H.
Performed data acquisition: Zhang, Y.

Availability of data and materials

The data used in this study are private and confidential due to privacy and confidentiality concerns. Therefore, we declare that the data will not be made publicly available. However, readers who require further information may contact the corresponding author to obtain the relevant data.

AI and AI-assisted tools statement

During the preparation of this manuscript, ChatGPT (OpenAI) was used to generate the initial graphical abstract. The generated graphical abstract was subsequently reviewed, manually revised, and finalized by the authors to ensure technical accuracy and full consistency with the manuscript. The AI tool did not influence the study design, data collection, analysis, interpretation, or the scientific content of the work. All authors take full responsibility for the accuracy, integrity, and final content of the manuscript.

Financial support and sponsorship

This work was supported in part by the National Natural Science Foundation of China (Grant 62373319, 62473327); and in part by the Natural Science Foundation of Hebei Province (Gran F2024203114, F2025203121).

Conflicts of interest

All authors declared no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Ajoudani, A.; Zanchettin, A. M.; Ivaldi, S.; Albu-Schäffer, A.; Kosuge, K.; Khatib, O. Progress and prospects of the human-robot collaboration. *Auton. Robot.* **2018**, *42*, 957–75. DOI
2. Yang, C.; Zhang, X.; Zhang, L.; et al. Coordinated energy-efficient walking assistance for paraplegic patients by using the exoskeleton-walker system. *Intell. Robot.* **2024**, *4*, 107–24. DOI
3. Tong, L.; Cui, D.; Wang, C.; Peng, L. A novel zero-force control framework for post-stroke rehabilitation training based on fuzzy-PID method. *Intell. Robot.* **2024**, *4*, 125–45. DOI
4. Wang, Z.; Chen, M.; Liu, Q. A review on multimodal communications for human-robot collaboration in 5G: from visual to tactile. *Intell. Robot.* **2025**, *5*, 579–606. DOI
5. Gao, Z.; Liao, Z.; Li, C. Better interaction experience: human-machine interface for soft robotic systems. *Intell. Robot.* **2025**, *5*, 520–40. DOI
6. Abbink, D. A.; Carlson, T.; Mulder, M.; et al. A topology of shared control systems - finding common ground in diversity. *IEEE Trans. Hum. Mach. Syst.* **2018**, *48*, 509–25. DOI
7. Parasuraman, R.; Sheridan, T. B.; Wickens, C. D. A model for types and levels of human interaction with automation. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **2000**, *30*, 286–97. DOI
8. Dragan, A. D.; Srinivasa, S. S. A policy-blending formalism for shared control. *Int. J. Robot. Res.* **2013**, *32*, 790–805. DOI
9. Abbink, D. A.; Mulder, M.; Boer, E. R. Haptic shared control: smoothly shifting control authority? *Cogn. Technol. Work* **2012**, *14*, 19–28. DOI
10. Flemisch, F.; Heesen, M.; Hesse, T.; Kelsch, J.; Schieben, A.; Beller, J. Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations. *Cogn. Technol. Work* **2012**, *14*, 3–18. DOI
11. Sheridan, T. B. Adaptive automation, level of automation, allocation authority, supervisory control, and adaptive control: distinctions and modes of adaptation. *IEEE Trans. Syst. Man Cybern. A Syst. Hum.* **2011**, *41*, 662–7. DOI
12. Jiang, H.; Yang, Y.; Hua, C.; Li, J. A novel ADP-based neurooptimal control methodology for teleoperation systems under interactive shared-control framework. *IEEE Trans. Autom. Sci. Eng.* **2026**, *23*, 7036–48. DOI
13. Wang, W.; Zhang, Y.; Yang, C.; et al. A human-machine shared dual fuzzy authority allocation control strategy for automatic driving vehicle considering driver intention judgement. *Expert Syst. Appl.* **2025**, *274*, 126971. DOI
14. Su, C.; Yang, H.; Li, J.; Wu, X. Steering authority allocation strategy for human-machine shared control based on driver take-over feasibility. *Comput. Electr. Eng.* **2024**, *120*, 109753. DOI
15. Ma, Y.; Liu, H.; Yu, Z. Adaptive shared control for robot manipulator obstacle avoidance with dynamic authority allocation. In *AI Enabled Robotic Loco-Manipulation. Lecture Notes in Networks and Systems*. Springer Nature Switzerland; 2025. pp. 27–38. DOI
16. Sarabia, J.; Marcano, M.; Diaz, S.; Rastelli, J. P.; Zubizarreta, A. Evaluating shared control in real-world conditions. *IEEE Open J. Veh. Technol.* **2026**, *7*, 418–31. DOI
17. Matsumoto, S.; Riek, L. D. Shared control in human robot teaming: toward context-aware communication. *arXiv* **2022**, arXiv:2203.10218. Available online: https://doi.org/10.48550/arXiv.2203.10218. (accessed on 22 Jul 2026)
18. Xiao, J.; Wang, B.; Huang, K.; Terzi, S.; Wang, W.; Macchi, M. Intelligent disassembly scenario understanding for human behavior and intention recognition towards self-perception human-robot collaboration system. *J. Manuf. Syst.* **2025**, *79*, 937–62. DOI
19. Xing, X.; Burdet, E.; Si, W.; Yang, C.; Li, Y. Impedance learning for human-guided robots in contact with unknown environments. *IEEE Trans. Robot.* **2023**, *39*, 3705-21. DOI

20. Johnson, A. W.; Duda, K. R.; Sheridan, T. B.; Oman, C. M. A closed-loop model of operator visual attention, situation awareness, and performance across automation mode transitions. *Hum. Factors* **2017**, *59*, 229–41. DOI
21. Lee, J. D.; See, K. A. Trust in automation: designing for appropriate reliance. *Hum. Factors* **2004**, *46*, 50–80. DOI
22. Kraus, J.; Scholz, D.; Stiegemeier, D.; Baumann, M. The more you know: trust dynamics and calibration in highly automated driving and the effects of take-overs, system malfunction, and system transparency. *Hum. Factors* **2020**, *62*, 718–36. DOI
23. Elgohr, A. T.; Rashad, M.; El-Gendy, E. M.; Shaaban, W.; Saafan, M. M. Trust as a design principle in human–robot collaboration: a review of explainable and adaptive control. *Artif. Intell. Rev.* **2026**, *59*, 134. DOI
24. Hergeth, S.; Lorenz, L.; Krems, J. F. Prior familiarization with takeover requests affects drivers' takeover performance and automation trust. *Hum. Factors* **2017**, *59*, 457–70. DOI
25. Muir, B. M. Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* **1994**, *37*, 1905–22. DOI
26. Mackworth, N. H. The breakdown of vigilance during prolonged visual search. *Q. J. Exp. Psychol.* **1948**, *1*, 6–21. DOI
27. Warm, J, S.; Parasuraman, R.; Matthews, G. Vigilance requires hard mental work and is stressful. *Hum. Factors* **2008**, *50*, 433–41. DOI
28. Hua, C.; Li, Y.; Guan, X. Finite/fixed-time stabilization for nonlinear interconnected systems with dead-zone input. *IEEE Trans. Autom. Control* **2017**, *62*, 2554–60. DOI
29. Hua, C.; Ning, P.; Li, K. Adaptive prescribed-time control for a class of uncertain nonlinear systems. *IEEE Trans. Autom. Control* **2022**, *67*, 6159–66. DOI
30. Ames, A. D.; Xu, X.; Grizzle, J. W.; Tabuada, P. Control barrier function based quadratic programs for safety critical systems. *IEEE Trans. Autom. Control* **2017**, *62*, 3861–76. DOI
31. Jankovic, M. Robust control barrier functions for constrained stabilization of nonlinear systems. *Automatica* **2018**, *96*, 359–67. DOI
32. Salehi, I.; Rotithor, G.; Dani, A. Safe adaptive trajectory tracking control of robot for human-robot interaction using barrier function transformation. In *Collaborative and Humanoid Robots*. London: IntechOpen; 2021. DOI
33. Maithani, P.; Arab, A.; Khorrami, F.; Krishnamurthy, P. Proactive hierarchical control barrier function-based safety prioritization in close human-robot interaction scenarios. *arXiv* **2025**, arXiv:2505.16055. Available online: https://doi.org/10.48550/arXiv.2505.16055. (accessed on 22 Jul 2026)
34. Janwani, N. C.; Daş, E.; Touma, T.; Wei, S. X.; Molnar, T. G.; Burdick, J. W. A learning-based framework for safe human-robot collaboration with multiple backup control barrier functions. *arXiv* **2023**, arXiv:2310.05865. Available online: https://doi.org/10.48550/arXiv.2310.05865. (accessed on 22 Jul 2026)
35. Xiong, Y.; Zhai, D. H.; Xia, Y. Robust safety-critical control design for robotic systems under input disturbances with multiple time-varying constraints. *Int. J. Syst. Sci.* **2026**. DOI
36. Zhang, Y.; Xie, W.; Ma, R. Disturbance observer-based terminal sliding mode control for the training safety improvement in robot-assisted rehabilitation. *Intell. Robot.* **2025**, *5*, 333–54. DOI
37. He, L.; Peng, Z.; Zhang, X.; et al. Reinforcement learning-based fixed-time optimal impedance control for human-robot collaboration with input disturbances. *IEEE Internet Things J.* **2025**, *12*, 54638–51. DOI
38. Busellato, L.; Cunico, F.; Dall'Alba, D.; et al. Uncertainty aware-predictive control barrier functions: safer human-robot interaction through probabilistic motion forecasting. *Robot. Auton. Syst.* **2026**, *197*, 105291. DOI
39. Sato, E.; Wada, T. Human-centered cooperative control coupling autonomous and haptic shared control via control barrier function. *IFAC-Pap.* **2025**, *58*, 280–5. DOI
40. Feng, N.; Cai, X. Attention-based and individualizable adaptive cruise control using control barrier function. *J. Donghua Univ.* **2026**, *52*, 255-62. DOI