LEARNING TO HOLD: FLEET-AWARE AND TEMPORALLY AUGMENTED MULTI-AGENT REINFORCEMENT LEARNING FOR BUS CONTROL

Anonymous authors
Paper under double-blind review

ABSTRACT

Effective bus holding control (BHC) is a real-time multi-agent decision-making problem where buses must coordinate under directional, time-varying, and partially observable fleet interactions. Existing BHC methods often rely on local observations, limited temporal context, or centralized training schemes that scale poorly with fleet size. We propose **SMART-HOLD**, a scalable multi-agent reinforcement learning approach for dynamic BHC. SMART-HOLD introduces a *Fleet-wide State Representation* that allows each bus agent to reason over fleet-level spatial structure, and a critic-only *Temporally-Augmented State Representation* that captures short-term system evolution for more accurate value estimation without leaking future information to the policy. A cross-attention actor-critic architecture selectively models subject, active, and passive agents to capture directional and role-dependent interactions. For scalable coordinated training, a *Regionalized Training with Distilled Coordination* strategy is developed to train multiple regional teacher policies in parallel and distills them into a shared actor for decentralized execution. Experiments on six real-world bus routes, including unseen routes with different service frequencies, fleet sizes, and temporal variability, show that SMART-HOLD substantially reduces bunching and improves service reliability over state-of-the-art BHC and multi-agent learning baselines.

1 INTRODUCTION

Bus transportation systems are central to sustainable urban mobility, but their reliability is often undermined by stochastic passenger demand, traffic congestion, and travel-time variability (Ibarra-Rojas et al., 2015). These disruptions can cause **bus bunching**, where multiple buses arrive and dwell at the same stop, breaking planned headways, increasing passenger waiting time, and degrading service reliability (Jenelius & Cats, 2015). **Bus holding control** (**BHC**) mitigates this problem by regulating dwell times to restore regular headways (Delgado et al., 2012).

Traditional BHC methods, including stochastic optimization and heuristic-based approaches (Delgado et al., 2012; Daganzo, 2009; Xuan et al., 2011; Sánchez-Martínez et al., 2016), often rely on historical data, simplified dynamics, or static assumptions, limiting their adaptability to real-time disruptions (Bartholdi & Eisenstein, 2012). Reinforcement learning (RL) and multi-agent RL (MARL) offer a more flexible paradigm by learning adaptive policies through interaction with dynamic environments (Chen et al., 2016; Liu et al., 2023). However, BHC is not a conventional MARL problem. Bus operations are decentralized, event-driven, and affected by delayed, partial, or outdated information (Liu et al., 2023; Wang & Sun, 2021; Fu & Yang, 2002). Each bus must make local holding decisions while responding to the concurrent actions of other buses, making fully centralized control brittle and impractical and motivating decentralized MARL with coordinated fleet-level learning (Wang & Sun, 2021; 2020).

Specifically, inter-bus control depends jointly on *where* other buses are and *when* they will act. At each decision event, the bus making the holding decision is the *subject agent*. Other buses may act before the subject agent reaches its next stop, directly changing the headways and passenger accumulation it will face; we call them *active agents*. Other buses do not act within this interval but still shape future fleet dynamics; we call them *passive agents*. Upstream and downstream buses

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

also affect the subject agent in different ways. Thus, BHC involves *directional*, *asymmetric*, and *time-varying* interactions that cannot be captured properly by agents' local observations or uniform inter-agent information aggregation mechanisms.

Therefore, the BHC problem raises three technical challenges tied directly to real-time deployment. **(C1) Fleet-wide and temporal awareness.** Agents must use available fleet-level information while remaining robust to partial and delayed communication, yet existing methods mainly rely on information in local contexts. **(C2) Directional and role-dependent interaction modeling.** Active, passive, upstream, and downstream agents influence the subject agent in different ways, making it difficult for standard feed-forward networks, locality-constrained graph models, or generic attention mechanisms to distinguish their distinct effects. **(C3) Scalable and coordinated learning.** Fully centralized training becomes costly as fleet size grows, whereas purely independent training can produce inconsistent local policies.

These challenges call for a new MARL approach that excels at fleet-wide representation learning, temporal value estimation, structured attention, and scalable coordination. Recent advances in attention-based architectures (Vaswani et al., 2017; Veličković et al., 2018; Iqbal & Sha, 2019), centralized-training-decentralized-execution (CTDE) paradigms (Lowe et al., 2017), and policy distillation techniques (Hinton et al., 2015; Rusu et al., 2016; Pei et al., 2025) provide useful technical foundations. However, they fall short of the combined demands of real-time BHC, where agents act asynchronously, active and passive buses play different roles, fleet interactions are directional, and deployment must scale to large systems.

We propose **SMART-HOLD**, a scalable MARL approach for dynamic BHC. SMART-HOLD addresses C1 through a *fleet-wide state representation* (FSR) for decentralized actor decisions and a *temporally augmented state representation* (TSR) for critic learning, capturing both fleet structure and short-term evolution while keeping future information hidden from the actor. It addresses C2 with a cross-attention actor-critic architecture that explicitly models directional and role-dependent interactions among subject, active, and passive agents. It addresses C3 through *Regionalized Training with Distilled Coordination* (RTDC), which trains a small number of regional teacher policies in parallel and distills them into a shared student actor to improve scalability, sample efficiency, and coordination. Experiments on five real-world bus routes, including unseen routes with longer distances, heavier loads, and higher temporal variability, show that SMART-HOLD substantially reduces bus bunching and improves service reliability over state-of-the-art BHC and MARL baselines.

Although motivated by bus holding control, SMART-HOLD captures key design principles for cooperative MARL under decentralized execution, directional interactions, and partial observability, enabling scalable coordination in complex multi-agent systems for future studies.

2 Related Work

Bus holding control. Traditional BHC methods use analytical models, stochastic optimization, linear programming, or heuristics to regulate dwell times and stabilize headways (Daganzo, 2009; Xuan et al., 2011; Sánchez-Martínez et al., 2016; Chen et al., 2013). These methods are often computationally efficient. However, they often rely on simplified assumptions or historical demand patterns, making them less adaptive to real-time disturbances such as traffic congestion and stochastic passenger arrivals (Bartholdi & Eisenstein, 2012). This limitation has motivated data-driven approaches that learn control policies directly from dynamic operating environments.

MARL for BHC. Recent work has explored RL and MARL for dynamic BHC (Yang et al., 2024; Ceder, 2015; Wang & Fang, 2026). Early DQN-based methods improve headway regularity using local observations (Chen et al., 2016; Liu et al., 2023; Alesiani & Gkiotsalitis, 2018), and later methods incorporate neighboring bus states to improve coordination (Rodriguez et al., 2023). However, these approaches rely on discrete actions, which are poorly suited to BHC where holding decisions should vary continuously over time. For the same reason, value-decomposition methods such as QMIX and QTRAN are not directly applicable to continuous BHC (Son et al., 2019; Tian et al., 2024).

Actor-critic methods support continuous actions and have become the dominant choice for learning BHC policies. (Wang & Sun, 2020) proposed a multi-agent actor-critic method, later extended by

(Wang & Sun, 2021) with an event-driven formulation and delayed reward handling. Despite these advances, existing actor-critic methods usually rely on local observations and simple feed-forward or graph-based architectures (Veličković et al., 2018; Rodriguez et al., 2023). As a result, they have limited ability to model fleet-wide, directional, and temporally evolving interactions. Moreover, passive agents are usually ignored, although they often constitute a large portion of the fleet and can strongly shape future system dynamics.

Positioning of this work. SMART-HOLD closes these gaps by unifying fleet-wide representation learning, critic-only temporal augmentation, attention-driven interaction modeling, and scalable regionalized training. Unlike prior BHC approaches, it explicitly models subject, active, and passive agents, captures directional upstream and downstream dependencies, and supports decentralized execution through RTDC. Together, these advances establish SMART-HOLD as a principled continuous-control approach for real-time BHC under partial observability, asymmetric interactions, and large-scale fleet coordination.

3 Problem Description

We consider a looped bus route with m stops and l scheduled dispatches (Figure 1). Each bus departs from Stop S_1 at its scheduled time and serves all stops $\{S_2, \ldots, S_m\}$. Upon returning to S_1, all remaining passengers alight, resulting in a full passenger turnover (Liu et al., 2023).

At decision step t, let n_t denote the number of buses currently in operation, where $n_t < l$ and may vary over time.

The objective of bus holding control (BHC) is to mitigate bus bunching by improving *headway regularity*, particularly in looped systems where instability is common. Following prior work (Liu et al., 2023; Wang & Sun, 2021; 2020), all stops are treated as holding points and buses operate under a strict *no-overtaking* policy. At each stop, the control problem is to determine the holding time for each bus.

Action. At decision step t, the action $a_{i,t} \in [0, 1]$ taken by bus i represents the fraction of the maximum allowable holding time $T_{\max}$. The resulting holding time is given by $\tau_{i,t} = a_{i,t} \cdot T_{\max}$, enabling continuous and fine-grained control.

Reward. We define the reward as a weighted combination of headway regularity and holding cost:

$$r_{i,t} = -w_1 \cdot \mathrm{CV}^2(h_t) - w_2 \cdot a_{i,t}, \tag{1}$$

where $\mathrm{CV}^2(h_t) = \frac{\mathrm{Var}(h_t)}{\bar{h}_t^2}$ denotes the squared coefficient of variation of headways across all operating buses at time t. This scale-invariant metric captures relative headway dispersion, with lower values indicating more regular service. The coefficients $w_1, w_2 \in (0, 1)$ control the trade-off between service regularity and holding cost. Following (Wang & Sun, 2020), we set $w_1 = 0.8$ and $w_2 = 0.2$. This reward formulation promotes uniform headways while discouraging excessive holding, balancing reliability and operational efficiency.

Objective. The goal is to learn a policy π^* that maximizes the expected cumulative reward over all buses:

$$\pi^* = \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{T} \sum_{i=1}^{n_t} r_{i,t}\right], \tag{2}$$

where T denotes the time horizon. This objective jointly minimizes headway variability and holding time across the fleet.

4 SMART-HOLD: A Scalable Multi-Agent Reinforcement Learning Approach

Our SMART-HOLD system introduces innovative *fleet-wide state representation* (FSR) for the actor and *temporally-augmented state representation* (TSR) for the critic, with their specific details outlined in Section 4.1.

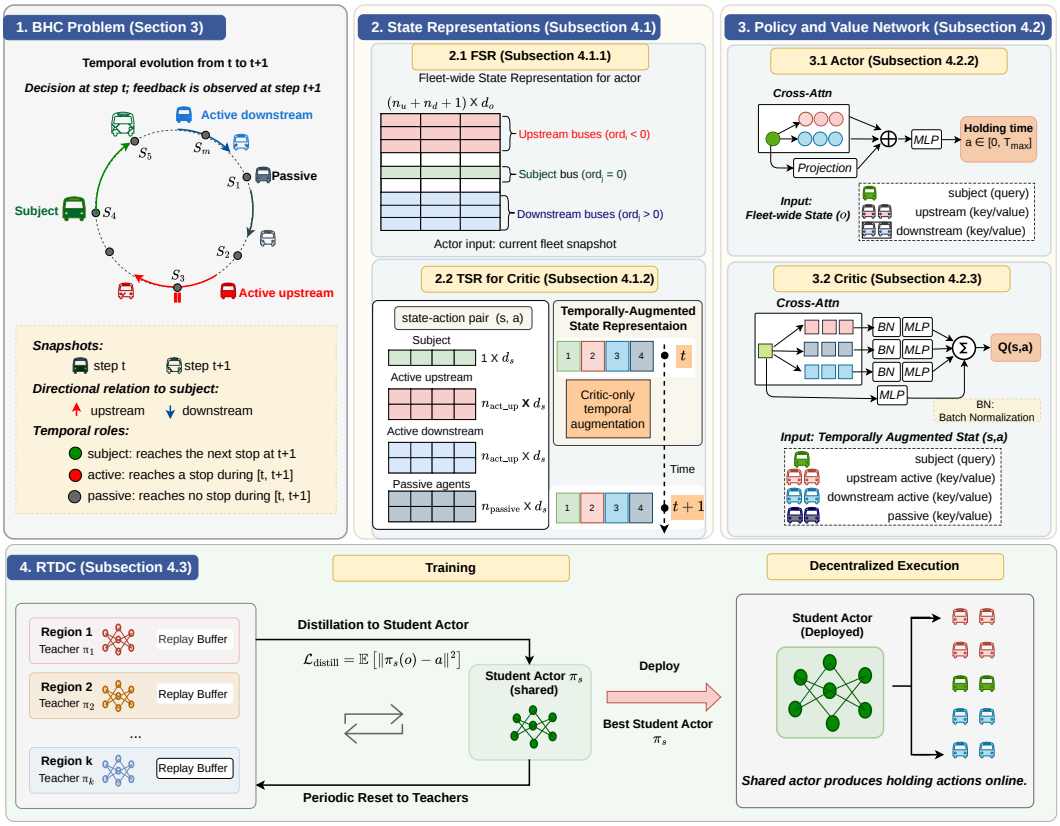


Figure 1: Overview of SMART-HOLD. The diagram illustrates the full pipeline from BHC problem modeling to state construction, actor-critic architectures, RTDC training, and online execution. FSR provides a fleet-wide snapshot for actor decisions, while TSR augments the critic input with short-term temporal evolution from t to $t+1$. RTDC trains regional teacher policies and distills them into a shared student actor for decentralized execution.

SMART-HOLD employs a cross-attention architecture to model interactions among subject, active, and passive agents. As illustrated in Figure 1, we further propose Regionalized Training with Distilled Coordination (RTDC), where a central student actor periodically distills knowledge from region-specific teacher actors via behavioral cloning. Built on DDPG, this design enables scalable and coordinated policy learning across the fleet (Section 4.3).

4.1 State Representation

Each bus is controlled by an agent A_i with its own state, action, and reward. Specifically, agent A_i relies on its FSR to process fleet-wide information, such as the locations and headways of all operating buses along a looped line, to foster coordinated BHC. Meanwhile, TSR assists the critic in accurately predicting the agent's performance by explicitly modeling and prioritizing critical temporal interactions among all agents.

4.1.1 Fleet-Wide State (Actor)

The FSR for agent A_i at time t is denoted by $o \in \mathbb{R}^{n_t \times d_o}$, representing a fleet-level snapshot from the perspective of the subject agent:

$$\mathbf{o} = [o_1, \ldots, o_i, \ldots, o_{n_t}]^T. \tag{3}$$

Each row $o_j \in \mathbb{R}^{d_o}$ encodes the state of bus j:

$$o_j = [q_j, s_j, x_j, d_j, hb_j, hf_j, m_j], \tag{4}$$

where $q_t \in \{0, 1\}$ indicates whether bus j is the subject bus currently making a holding decision; s_t is the index of its most recently visited stop; and $x_t \in [0, 1]$ is its normalized position index along the route. The variable d_t denotes the remaining dwell or holding time in minutes and is zero while the bus is travelling. The variables hb_j and hf_j are the estimated time headways from bus j to its nearest downstream and upstream buses, respectively.

We partition o into three components based on relative position:

$$\mathbf{o} = [o_{\text{up}},\ o_{\text{i}},\ o_{\text{down}}], \tag{5}$$

where $o_{\text{i}} \in \mathbb{R}^{1\times d_o}$ represents the subject agent, while $o_{\text{up}} \in \mathbb{R}^{n_u\times d_o}$ and $o_{\text{down}} \in \mathbb{R}^{n_d\times d_o}$ denote upstream and downstream buses, respectively.

Incorporating both upstream and downstream agents allows the subject agent to capture asymmetric interactions across the fleet. Upstream buses reflect emerging disruptions (e.g., delays or bunching), while downstream buses determine how local decisions propagate forward. This bidirectional context enables coordinated decision making under decentralized and partially observable conditions.

4.1.2 Temporally-Augmented State (Critic)

To improve value estimation, the critic augments the actor state o with additional information o^+ collected over the interval $(t, t + 1)$, forming the temporally augmented state:

$$s = o\|o^+. \tag{6}$$

The corresponding state-action pair is:

$$(s, a) = o\|a\|o^+\|a^+, \tag{7}$$

where o_j^+ and a_j^+ denote the updated state features and actions of agent j during $(t, t + 1)$.

We further decompose (s, a) according to agent roles:

$$(s, a) = \big[(s, a)_{\text{i}},\ (s, a)_{\text{act_up}},\ (s, a)_{\text{act_down}},\ (s, a)_{\text{passive}}\big],$$

where $(s, a)_{\text{i}}$ corresponds to the subject agent, $(s, a)_{\text{act_up}}$ and $(s, a)_{\text{act_down}}$ denote active upstream and downstream agents, and $(s, a)_{\text{passive}}$ represents passive agents.

This design enables the critic to evaluate actions using both the decision-time state and short-term system evolution. In particular, active agents may take actions within $(t, t + 1)$ that directly affect the environment faced by the subject agent, while passive agents do not act in this interval but still influence future dynamics. By incorporating both, TSR captures temporal dependencies and asymmetric interactions across the fleet, leading to more accurate value estimation.

TSR is used only by the critic during training; the actor observes only decision-time states, ensuring causal correctness.

4.2 Network Architecture

We adopt the attention mechanism (Vaswani et al., 2017) in both the actor and critic neural networks, enabling the subject agent to effectively manage its relationship with other agents.

4.2.1 Cross-Attention Mechanism

We model inter-agent interactions using a cross-attention mechanism, where the subject agent serves as the query and other agents provide keys and values. This design allows the subject agent to selectively attend to relevant agents based on their roles and states.

For example, when modeling upstream interactions in the actor, the subject state o_i acts as the query q, while upstream agents form the key-value pairs:

$$q = o_i, \quad K = o_{\text{up}}, \quad V = o_{\text{up}}.$$

Instead of using standard scaled dot-product attention (Vaswani et al., 2017), we compute attention weights via a multi-layer perceptron (MLP) over concatenated query-key embeddings:

$$\text{Cross-Attn}(\mathbf{q}, \mathbf{K}, \mathbf{V}) = \sum_i \alpha_i \mathbf{v}'_i, \quad \alpha_i = \text{softmax}\big(\text{MLP}([\mathbf{q}'\|\mathbf{k}'_i])\big), \tag{8}$$

where $\mathbf{q}' = \mathbf{q}\mathbf{W}_Q$, $\mathbf{k}'_i = \mathbf{k}_i\mathbf{W}_K$, and $\mathbf{v}'_i = \text{MLP}_v(\mathbf{v}_i)$ are learned projections.

This formulation enables flexible, non-linear modeling of inter-agent dependencies beyond similarity-based interactions. In particular, it captures *directional and role-dependent effects*, allowing the subject agent to weigh upstream, downstream, and passive agents differently based on their potential impact. Empirically, our ablation study (Appendix H) shows that this MLP-based attention consistently outperforms standard dot-product attention, aligning with recent findings on learned attention mechanisms (Morsali et al., 2023).

4.2.2 Actor Network

Figure 1 illustrates the actor network. The actor takes the FSR $\mathbf{o}$ (Equation equation 5) as input and outputs the action for the subject agent:

$$a = \sigma\big(\text{MLP}(z_{\text{up}} + z_{\text{down}} + z_{\text{i}})\big), \tag{9}$$

where $\sigma(\cdot)$ denotes the sigmoid function, ensuring $a \in (0, 1)$.

The interaction embeddings are computed via cross-attention:

$$z_{\text{up}} = \text{Cross-Attn}(o_{\text{i}}, o_{\text{up}}), \quad z_{\text{down}} = \text{Cross-Attn}(o_{\text{i}}, o_{\text{down}}).$$

The subject agent's own state is projected into the same latent space:

$$z_{\text{i}} = o_{\text{i}} W_{\text{proj}},$$

where $W_{\text{proj}} \in \mathbb{R}^{d_o \times d_{\text{proj}}}$ is a learnable projection matrix.

The final action is obtained by aggregating role-specific embeddings and passing them through an MLP, enabling the policy to jointly reason over upstream, downstream, and self information.

4.2.3 Critic Network

Figure 1 illustrates the critic architecture. The critic estimates the state-action value $Q(s, a)$ via a decomposition over different agent roles:

$$Q(s, a) = Q_i + \sum_{g \in \mathcal{G}} Q_g, \quad \mathcal{G} = \{\text{act_up}, \text{act_down}, \text{passive}\}. \tag{10}$$

Each component in Equation equation 10 is defined as:

$$Q_i = \text{MLP}((s, a)_i), \quad Q_g = \text{MLP}(\text{BN}(\text{CrossAttn}((s, a)_i, (s, a)_g))),$$

with $\text{BN}(\cdot)$ denoting batch normalization.

This decomposition explicitly separates the contributions of the subject, active upstream, active downstream, and passive agents. Although the same cross-attention operator is used for active upstream and downstream agents, they are processed as separate components in the critic, allowing their contributions to be estimated independently. Passive agents are modeled by a dedicated component because they do not act within the subject agent's decision interval but still shape future fleet dynamics. Each component is independently normalized and processed, enabling the critic to preserve heterogeneous interaction effects.

This design captures role-dependent and asymmetric interactions that are critical in BHC. By modeling these interaction types separately and aggregating them in a structured manner, the critic achieves more accurate and stable value estimation. Unlike prior approaches that aggregate neighboring agents indiscriminately or ignore passive agents (Wang & Sun, 2021), our critic treats each interaction type as a distinct source of influence. This explicit separation with structured aggregation improves the modeling of inter-agent dynamics and facilitates learning of coordinated policies in complex environments.

4.3 Regionalized Training with Distilled Coordination

We propose *Regionalized Training with Distilled Coordination (RTDC)*, a scalable training strategy that combines Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016) with periodic policy distillation. RTDC trains multiple regional teacher actors in parallel and distills them into a shared student actor via behavioral cloning.

DISTILLED COORDINATION VIA BEHAVIORAL CLONING

We partition the route into k regions and train k teacher actors for n buses, with $k \ll n$. At each decision step, an agent follows the teacher policy corresponding to its current region. Each teacher is trained using DDPG on region-specific experience.

To enable coordination across regions, we periodically distill the teacher policies into a unified student actor. Specifically, we aggregate experience from all regions and train the student policy via behavioral cloning to minimize:

$$\mathcal{L}_{\text{distill}} = \mathbb{E}_{(o,a)\sim\mathcal{B}} \left[\|\pi_s(o) - a\|_2^2 \right], \tag{11}$$

where $\mathcal{B} = \bigcup_j \mathcal{B}_j$ is the union of replay buffers across regions. The distilled student policy is then used to reinitialize all teacher actors, enabling knowledge sharing and coordinated refinement. The full procedure is summarized in Algorithm 1.

Discussion. RTDC improves training scalability by learning $k \ll n$ regional teacher policies. Periodic distillation aggregates regional experience into a shared student actor, reducing redundant exploration and accelerating convergence. By resetting regional teachers from the distilled student, RTDC promotes coordination across regions, mitigating conflicting local policies during training.

Algorithm 1: Regionalized Training with Distilled Coordination

```
Initialize regional teachers {π_j, B_j}_{j=1}^k and student π_s
for each training iteration do
    Collect transitions with a_e = clip(π_j(o_e) + ε_e, 0, T_max),
    ε_e ~ N(0, σ^2)
    Update each teacher π_j with DDPG using B_j
    if distillation step then
        B ← ⋃_{j=1}^k B_j
        π_s ← arg min_π E_{(o,a)~B}[‖π(o) − a‖_2^2]
        Reset π_j ← π_s, ∀j
    end if
end for
```

5 EXPERIMENTS

5.1 EXPERIMENT SETTING

We use two benchmark datasets, denoted as R1 and R3 in (Wang & Sun, 2021),[1] together with four additional bus routes (M1, M2, M3 and M7) newly introduced in this study from the public New Zealand Metlink dataset.[2]

To evaluate generalization across unseen routes and operating conditions, SMART-HOLD is trained exclusively on R1 and evaluated on R1 together with five unseen real-world bus routes spanning diverse service patterns and fleet sizes. As summarized in Table 1, R1 is a medium-length route (17.4 km, 46 stops, 58 daily trips) with a dispatch headway of 14.8 ± 4.8 min and 9.5 ± 0.5 running buses at peak time. The remaining routes are used solely for testing and represent a diverse range of operating conditions. R3 is a long route (23.2 km, 61 stops) with large headway variability. The Metlink routes range from relatively high-frequency services such as M2, with a mean dispatch headway of 10.4 ± 4.2 min and up to 12.2 ± 0.4 running buses at peak time, to shorter routes such as M3 and M7 with smaller fleets and different service frequencies. In particular, M7 is only 6.0 km long but operates 90 daily trips, whereas M1

Table 1: Summary statistics of the six real-world bus routes used in the experiments, including route characteristics, service frequency, dispatch headway, and the average and peak numbers of running buses (mean $\pm$ standard deviation) over 10 random seeds without bus holding control.

Route	Stops	km	Trips	Headway (min)	Mean buses	Peak buses
R1	46	17.4	58	14.8 ± 4.8	5.87 ± 0.07	9.5 ± 0.5
R3	61	23.2	57	15.5 ± 5.9	7.67 ± 0.08	11.4 ± 0.5
M1	49	19.4	44	19.7 ± 7.5	5.40 ± 0.05	8.1 ± 0.3
M2	50	14.1	83	10.4 ± 4.2	8.97 ± 0.06	12.2 ± 0.4
M3	33	8.5	69	12.1 ± 5.0	5.73 ± 0.04	11.2 ± 0.4
M7	25	6.0	90	9.2 ± 3.6	5.10 ± 0.03	9.2 ± 0.4

[1] Among the four datasets studied in (Wang & Sun, 2021) (R1–R4), R1 and R3 are the only ones we found publicly available in the TransitGym repository: `https://github.com/TransitGym/TransitGym.github.io/tree/master`.

[2] Data are available at the Metlink Open Data Portal: `https://opendata.metlink.org.nz/`.

Table 2: Test performance on routes R1, R3, M1, M2, M3, and M7 in terms of bunching count, headway coefficient of variation (CV), and excess waiting time (EWT). Results are reported as mean $\pm$ standard deviation over the evaluation runs. The best mean result for each route is highlighted in bold. Lower values indicate better performance for all metrics.

Route	Metric	NC (No Control)	NHW (Daganzo, 2009)	CAAC (Wang & Sun, 2021)	IAC (Lillicrap et al., 2016)	MADDPG (Lowe et al., 2017)	MAPPO (Yu et al., 2022)	SMART-HOLD (Ours)
R1	Bunchings	45.40 ± 18.51	50.40 ± 16.40	7.80 ± 14.71	5.80 ± 15.74	16.51 ± 24.92	1.78 ± 1.69	**1.45 ± 0.91**
	Headway CV	0.49 ± 0.04	0.46 ± 0.02	0.27 ± 0.09	0.24 ± 0.08	0.30 ± 0.13	0.25 ± 0.03	**0.21 ± 0.02**
	EWT (min)	1.83 ± 0.29	1.57 ± 0.18	0.62 ± 0.48	0.51 ± 0.43	0.80 ± 0.69	0.50 ± 0.11	**0.35 ± 0.07**
R3	Bunchings	90.40 ± 26.56	75.40 ± 14.60	16.14 ± 28.70	12.67 ± 28.54	31.08 ± 43.09	6.68 ± 4.39	**5.97 ± 3.64**
	Headway CV	0.55 ± 0.05	0.51 ± 0.03	0.30 ± 0.10	0.29 ± 0.10	0.35 ± 0.14	0.30 ± 0.03	**0.24 ± 0.03**
	EWT (min)	2.48 ± 0.45	2.04 ± 0.23	0.83 ± 0.64	0.75 ± 0.63	1.12 ± 0.93	0.72 ± 0.15	**0.49 ± 0.10**
M1	Bunchings	13.50 ± 9.07	6.00 ± 0.60	2.03 ± 6.01	1.88 ± 6.23	4.19 ± 8.07	0.01 ± 0.10	**0.00 ± 0.00**
	Headway CV	0.49 ± 0.04	0.54 ± 0.02	0.31 ± 0.09	0.28 ± 0.08	0.32 ± 0.11	0.26 ± 0.03	**0.23 ± 0.03**
	EWT (min)	2.41 ± 0.43	2.95 ± 0.22	1.03 ± 0.58	0.88 ± 0.58	1.17 ± 0.76	0.72 ± 0.17	**0.55 ± 0.15**
M2	Bunchings	145.30 ± 24.88	153.10 ± 28.20	22.42 ± 43.29	14.14 ± 43.56	41.53 ± 65.21	6.02 ± 7.69	**2.25 ± 5.77**
	Headway CV	0.58 ± 0.03	0.65 ± 0.02	0.33 ± 0.11	0.30 ± 0.10	0.37 ± 0.14	0.31 ± 0.03	**0.27 ± 0.03**
	EWT (min)	1.81 ± 0.21	2.24 ± 0.16	0.64 ± 0.46	0.56 ± 0.47	0.84 ± 0.66	0.52 ± 0.09	**0.41 ± 0.08**
M3	Bunchings	65.00 ± 22.16	65.60 ± 15.40	13.14 ± 21.97	7.09 ± 21.77	19.44 ± 30.18	5.11 ± 5.17	**0.27 ± 0.72**
	Headway CV	0.57 ± 0.05	0.58 ± 0.03	0.31 ± 0.10	0.28 ± 0.10	0.33 ± 0.15	0.30 ± 0.03	**0.26 ± 0.02**
	EWT (min)	2.03 ± 0.40	2.10 ± 0.22	0.67 ± 0.51	0.56 ± 0.49	0.83 ± 0.71	0.58 ± 0.09	**0.42 ± 0.07**
M7	Bunchings	64.80 ± 17.50	90.70 ± 20.10	41.41 ± 41.95	8.34 ± 24.29	26.21 ± 38.98	10.95 ± 8.16	**1.59 ± 3.11**
	Headway CV	0.53 ± 0.03	0.58 ± 0.03	0.34 ± 0.09	0.30 ± 0.08	0.35 ± 0.13	0.31 ± 0.02	**0.25 ± 0.02**
	EWT (min)	1.33 ± 0.13	1.58 ± 0.15	0.58 ± 0.33	0.45 ± 0.31	0.66 ± 0.47	0.46 ± 0.06	**0.30 ± 0.03**

has the lowest service frequency, with 44 daily trips and a mean dispatch headway of 19.7 ± 7.5 min. Overall, the six routes span substantial variation in route length (6.0–23.2 km), service frequency (44–90 daily trips), dispatch headway (9.2–19.7 min), and peak fleet size (8.1–12.2 buses), providing a diverse test bed for assessing SMART-HOLD's ability to generalize from R1 to previously unseen route configurations and operating conditions.

Competing Approaches. To evaluate SMART-HOLD, we compare it with CAAC (Wang & Sun, 2021), Independent Actor–Critic (IAC) (Lillicrap et al., 2016), Multi-Agent DDPG (MADDPG) (Lowe et al., 2017), Multi-Agent PPO (MAPPO) (Yu et al., 2022), No Control (NC), and a naive headway-based controller (NHW) (Daganzo, 2009). IAC uses DDPG for fully decentralized training and execution. Both the actor and critic depend only on the subject bus's local observation. In contrast, in MADDPG and MAPPO, actors rely on local observations, whereas critics aggregate information from all operating buses. Specifically, MADDPG conditions the critic on both the global state and joint actions, whereas MAPPO conditions the critic only on the global state.

Hyperparameter Settings. The actor and critic use learning rates of 1×10^{-4} and 3×10^{-4}, respectively, with a batch size of 256. After every five episodes, the student undergoes 25 MSE updates using a balanced batch of 256 observations and a learning rate of 1×10^{-4}. Full settings are provided in Appendix C.1.

Evaluation Metrics. Following (Wang & Sun, 2021), we evaluate bus service reliability using the number of *Bunching* events. We additionally report *Headway Coefficient of Variation (Headway CV)*, a standard measure of headway regularity, and propose *Excess Waiting Time (EWT)*, which quantifies the additional passenger waiting caused by irregular bus spacing. Lower EWT indicates more reliable service. The definitions of all metrics are provided in Appendix G.

Experiment Platform. All methods are implemented in Python 3.10.20 and PyTorch 2.5.1 and run on a GTX 1070 Mobile GPU with 16 GB memory. SMART-HOLD averages 3.49 ms per decision, supporting real-time BHC. Training and evaluation times are reported in Appendix C.2.

5.2 Experiment Results

Main Results. Table 2 summarizes the test performance across six real-world routes. Each method is evaluated using 10 independently trained policies with identical random seeds and tested over 10 full-day operations, with results averaged across all runs. SMART-HOLD consistently achieves the

best performance across all routes and evaluation metrics. In particular, it attains the fewest bunching events and the lowest headway coefficient of variation (CV) and excess waiting time (EWT) on every route, while also exhibiting relatively small standard deviations, indicating stable performance across different training seeds.

For example, on route R3, SMART-HOLD records only 5.97 ± 3.64 bunching events, a headway CV of 0.24 ± 0.03, and an EWT of 0.49 ± 0.10 min. Compared with MAPPO, the strongest baseline for bunching and EWT on this route, SMART-HOLD reduces the mean bunching count from 6.68 to 5.97 and EWT from 0.72 to 0.49 min, while also substantially reducing headway variability. Similar improvements are observed on routes with markedly different baseline levels of instability. For instance, on M3 the bunching count decreases to 0.27 ± 0.72, compared with 5.11 ± 5.17 for MAPPO, whereas on M7 it decreases to 1.59 ± 3.11, compared with 8.34 ± 24.29 for IAC. Notably, SMART-HOLD almost completely eliminates bunching on M1, achieving 0.00 ± 0.00 bunching events while retaining the lowest headway CV and EWT.

Overall, SMART-HOLD consistently improves bunching, headway regularity, and passenger waiting time across diverse routes, indicating effective coordination and strong generalization to different operating conditions.

Attention Analysis. Figure 2 shows that SMART-HOLD generally assigns more attention to nearby buses across R1, R3, and M1. Upstream attention varies more by route, while downstream attention remains relatively stable. This suggests that SMART-HOLD learns a transferable, locality-aware interaction pattern that adapts to different operating conditions.

Ablation Study. The ablation study in Appendix H confirms the importance of FSR, TSR, and RTDC. TSR is particularly critical: replacing it with a non-augmented critic state degrades value estimation and overall performance. These results show that fleet-wide information and short-term temporal context are essential for stable learning and reliable BHC.

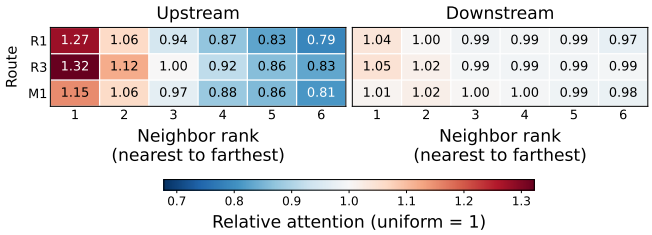


Figure 2: Learned attention for buses ordered from nearest to farthest on R1, R3, and M1.

As further shown in Appendix H, SMART-HOLD exhibits significantly higher learning speed than training a single shared policy. Powered by the RTDC strategy, SMART-HOLD converges quickly during training. Compared to centralized training, it also requires substantially fewer samples to learn coordinated behaviors, offering a clear advantage in complex and dynamic transit systems.

Appendix I examines sensitivity to the number of teacher actors k. Performance remains robust across tested values, with $k = 3$ giving the best balance of bunching reduction, travel efficiency, and policy stability. Larger k increases training cost without clear performance gains.

6 Conclusion

In this paper, we introduced SMART-HOLD, a scalable MARL approach for dynamic bus holding control. SMART-HOLD combines fleet-wide actor states, critic-only temporal augmentation, cross-attention-based interaction modeling, and regionalized training with distilled coordination to address decentralized decision-making, directional inter-agent dependencies, and scalable fleet-level learning. Across six real-world routes, SMART-HOLD achieves superior overall performance, reducing bus bunching on most routes while consistently improving headway regularity and service reliability compared with strong BHC and MARL baselines. These results demonstrate its effectiveness for real-time BHC under diverse operating conditions. However, the current evaluation focuses on individual bus routes under standard BHC assumptions and does not consider shared-stop multi-route networks, multimodal interactions, or external factors such as weather and roadworks. Extending SMART-HOLD to multimodal and network-wide transit control is an important direction for future work.

ETHICS STATEMENT

This work focuses on multi-agent reinforcement learning for dynamic bus holding control in public transit systems. The study does not involve human subjects, personally identifiable information, or individual passenger records. Experiments are conducted in simulation using real-world bus route data and operational settings. The proposed method is intended to improve service regularity, reduce bus bunching, and decrease passenger waiting time, potentially contributing to more reliable and efficient public transportation. Any real-world deployment would require appropriate operational validation and safety monitoring. We do not identify significant ethical concerns arising from the current study.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide: (1) a detailed description of the bus holding control problem formulation and the SMART-HOLD method, including the Fleet-wide State Representation (FSR), Temporally-Augmented State Representation (TSR), cross-attention actor-critic architecture, and Regionalized Training with Distilled Coordination (RTDC); (2) detailed experimental settings, including the seven evaluated bus routes, competing baselines, evaluation metrics, network configurations, and hyperparameters; (3) pseudo-code and implementation details for the training procedure; and (4) additional ablation, sensitivity, robustness, and scalability analyses in the appendix. The implementation environment and computational resources used for the experiments are also reported to facilitate reproducible evaluation.

LLM USAGE STATEMENT

During the preparation of this manuscript, we used large language models solely to assist with language refinement, including polishing, grammar correction, and rephrasing. All LLM-assisted text was carefully reviewed and revised by the authors to ensure its accuracy and originality. The LLM was not used to generate experimental results, design algorithms or neural network architectures, or develop any of the core technical contributions of this work.

REFERENCES

Francesco Alesiani and Konstantinos Gkiotsalitis. Reinforcement learning-based bus holding for high-frequency services. In *2018 21st International Conference on Intelligent Transportation Systems*, pp. 3162–3168, 2018. doi: 10.1109/ITSC.2018.8569473.

Christopher Amato. An introduction to centralized training for decentralized execution in cooperative multi-agent reinforcement learning, 2024.

John J. Bartholdi and Donald D. Eisenstein. A self-coördinating bus route to resist bus bunching. *Transportation Research Part B: Methodological*, 46(4):481–491, 2012. ISSN 0191-2615. doi: 10.1016/j.trb.2011.11.001.

Avishai Ceder. *Public Transit Planning and Operation: Modeling, Practice and Behavior*. CRC Press, Boca Raton, FL, 2 edition, 2015. ISBN 9781466563919. doi: 10.1201/b18689.

Qin Chen, Elodie Adida, and Jane Lin. Implementation of an iterative headway-based bus holding strategy with real-time information. *Public Transport*, 4(3):165–186, 2013. doi: 10.1007/s12469-012-0057-1.

Weiya Chen, Kunlin Zhou, and Chunxiao Chen. Real-time bus holding control on a transit corridor based on multi-agent reinforcement learning. In *2016 IEEE 19th International Conference on Intelligent Transportation Systems*, pp. 100–106, 2016. doi: 10.1109/ITSC.2016.7795538.

Carlos F. Daganzo. A headway-based approach to eliminate bus bunching: Systematic analysis and comparisons. *Transportation Research Part B: Methodological*, 43(10):913–921, 2009. ISSN 0191-2615. doi: 10.1016/j.trb.2009.04.002.

Felipe Delgado, Juan Carlos Muñoz, and Ricardo Giesen. How much can holding and/or limiting boarding improve transit performance? *Transportation Research Part B: Methodological*, 46(9): 1202–1217, 2012. doi: 10.1016/j.trb.2012.04.005.

Liping Fu and Xuhui Yang. Design and implementation of bus-holding control strategies with real-time information. *Transportation Research Record*, 1791(1):6–12, 2002. doi: 10.3141/1791-02.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

O. J. Ibarra-Rojas, F. Delgado, R. Giesen, and J. C. Muñoz. Planning, operation, and control of bus transport systems: A literature review. *Transportation Research Part B: Methodological*, 77: 38–75, 2015. ISSN 0191-2615. doi: 10.1016/j.trb.2015.03.002.

Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, pp. 2961–2970. PMLR, 2019. URL https://proceedings.mlr.press/v97/iqbal19a.html.

Erik Jenelius and Oded Cats. The value of new public transport links for network robustness and redundancy. *Transportmetrica A: Transport Science*, 11(9):819–835, 2015. doi: 10.1080/23249935.2015.1087232.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1509.02971.

Dong Liu, Feng Xiao, Jian Luo, and Fan Yang. Deep reinforcement learning-based holding control for bus bunching under stochastic travel time and demand. *Sustainability*, 15(14):10947, 2023. doi: 10.3390/su151410947.

VJM Low, HL Khoo, and WC Khoo. Robust dynamic real-time control strategies for high-frequency bus service: a multi-agent reinforcement learning framework. *Journal of Intelligent Transportation Systems*, 2024. doi: 10.1080/15472450.2024.2425293.

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.02275.

Alireza Morsali, Moein Heidari, Samin Heydarian, and Tohid Abedini. MLP-attention: Improving transformer architecture with MLP attention weights, 2023. URL https://openreview.net/forum?id=99XvUeDFYTD.

Yuhang Pei, Tao Ren, Yuxiang Zhang, Zhipeng Sun, and Matys Champeyrol. Policy distillation for efficient decentralized execution in multi-agent reinforcement learning. *Neurocomputing*, 626: 129617, 2025. doi: 10.1016/j.neucom.2025.129617.

Joseph Rodriguez, Haris N. Koutsopoulos, Shenhao Wang, and Jinhua Zhao. Cooperative bus holding and stop-skipping: A deep reinforcement learning framework. *Transportation Research Part C: Emerging Technologies*, 155:104308, 2023. doi: 10.1016/j.trc.2023.104308.

Andrei A. Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. Policy distillation. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06295.

Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. QTRAN: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5887–5896. PMLR, 2019. URL https://proceedings.mlr.press/v97/son19a.html.

G. E. Sánchez-Martínez, H. N. Koutsopoulos, and N. H. M. Wilson. Real-time holding control for high-frequency transit with dynamics. *Transportation Research Part B: Methodological*, 83:1–19, 2016. ISSN 0191-2615. doi: 10.1016/j.trb.2015.11.013.

Shucong Tian, Meng Yang, Rongling Xiong, Xingxing He, and Sutharshan Rajasegarar. A sequential multi-agent reinforcement learning framework for different action spaces. *Expert Systems with Applications*, 258:125138, 2024. ISSN 0957-4174. doi: 10.1016/j.eswa.2024.125138.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJXMpikCZ.

Jiawei Wang and Lijun Sun. Dynamic holding control to avoid bus bunching: A multi-agent deep reinforcement learning framework. *Transportation Research Part C: Emerging Technologies*, 116:102661, 2020. doi: 10.1016/j.trc.2020.102661.

Jiawei Wang and Lijun Sun. Reducing bus bunching with asynchronous multi-agent reinforcement learning. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pp. 426–433. International Joint Conferences on Artificial Intelligence Organization, August 2021. doi: 10.24963/ijcai.2021/60.

Zhe Wang and Zhixiang Fang. A deep reinforcement learning approach for real-time bus operation control using departure timetabling, stop-skipping, and re-routing. *Travel Behaviour and Society*, 43:101165, 2026. ISSN 2214-367X. doi: 10.1016/j.tbs.2025.101165.

Weitiao Wu, Yanchen Zhu, and Ronghui Liu. Dynamic scheduling of flexible bus services with hybrid requests and fairness: Heuristics-guided multi-agent reinforcement learning with imitation learning. *Transportation Research Part B: Methodological*, 190:103069, 2024. ISSN 0191-2615. doi: 10.1016/j.trb.2024.103069.

Yiguang Xuan, Juan Argote, and Carlos F. Daganzo. Dynamic bus holding strategies for schedule reliability: Optimal linear control and performance analysis. *Transportation Research Part B: Methodological*, 45(10):1831–1845, December 2011. doi: 10.1016/j.trb.2011.07.009.

Ying Yang, Junchi Cheng, and Yang Liu. An overview of solutions to the bus bunching problem in urban bus systems. *Frontiers of Engineering Management*, 11(4):661–675, December 2024. ISSN 2096-0255. doi: 10.1007/s42524-024-0297-1.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24611–24624. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9c1535a02f0ce079433344e14d910597-Abstract-Datasets_and_Benchmarks.html.

Jiajie Yu, Yuhong Wang, and Wei Ma. Large language model-enhanced reinforcement learning for generic bus holding control strategies, 2024. URL https://arxiv.org/abs/2410.10212.

Xian Zhang, Changlei Gu, Hong Wang, Guibin Wang, Yinliang Xu, and Ahmed Rabee Sayed. Hydrogen energy storage system participated decentralized voltage control with multi-agent deep reinforcement learning method. *IEEE Transactions on Industry Applications*, 61(2):2578–2588, 2025. doi: 10.1109/TIA.2025.3531321.

Yuwei Zhao, Ziyan Feng, and Xiang Li. Transfer-based customized bus network design with holding control and heterogeneous fleet. *IEEE Transactions on Evolutionary Computation*, 29(6):2567–2581, 2025. doi: 10.1109/TEVC.2024.3498315.

APPENDIX CONTENTS

A Problem Setup: Looped Bus Route and Assumptions **14**

B Related Work **14**

B.1 Classical Methods 14

B.2 Reinforcement Learning: Actor-Critic-Based Approaches 15

B.3 Multi-Agent Reinforcement Learning Approaches 15

C Implementation Settings and Runtime **16**

C.1 Hyperparameter Settings 16

D State Representations **16**

E Neural Networks **16**

E.1 Cross-Attention 16

F Training Algorithm **19**

G Performance Metrics **20**

H Baseline Methods and Ablation Study **21**

H.1 Baseline Methods 21

H.2 Performance Comparison 22

H.3 Ablation Studies 22

I Parameter Sensitivity Analysis **25**

A Problem Setup: Looped Bus Route and Assumptions

We consider a looped bus route for the BHC as illustrated in Figure 3. Each bus is dispatched from Stop 1 at its designated time and sequentially serves all stops {Stop 2, . . ., Stop m}, before returning to Stop 1 to repeat the cycle.

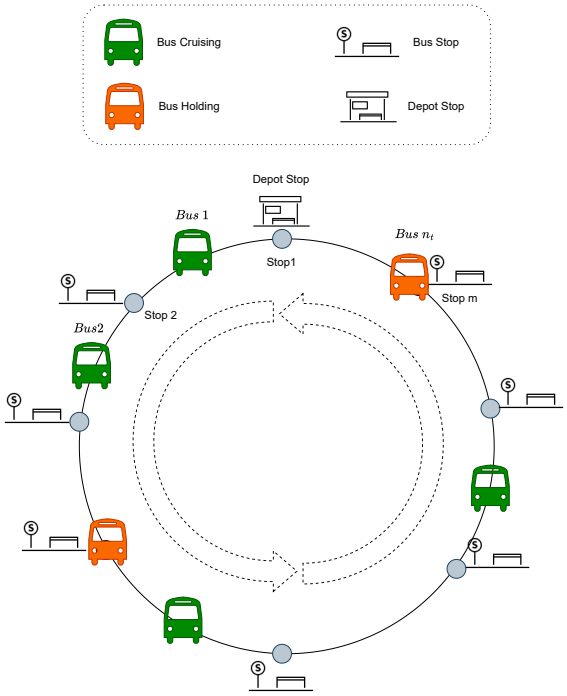


Figure 3: Looped Bus Route

Compatible with existing studies (Liu et al., 2023; Wang & Sun, 2020; 2021), the following operational assumptions are adopted for the Bus Holding Control (BHC) problem:

1. All stops function as *holding stops*, serving as control points for managing bus headways (Liu et al., 2023; Wang & Sun, 2021).
2. Buses follow the *no-overtaking* policy, preserving their sequence throughout the route (Liu et al., 2023).

These assumptions facilitate the analysis by isolating the impact of headway control strategies, without introducing complexities from capacity constraints or overtaking rules.

B Related Work

In this appendix, we present a comprehensive review of existing methods for BHC, including classical methods, DQN-based approaches, and BHC using the actor-critic learning techniques. Finally, we will discuss multi-agent reinforcement learning methods for BHC and highlight their practical advantages over other method types.

B.1 Classical Methods

Classical bus holding control strategies aim to improve transit reliability by maintaining regular headways and minimizing passenger waiting times. These approaches often rely on deterministic

or stochastic optimization models and predefined rules for operational control. Daganzo and Pilachowski (Daganzo, 2009) introduced a headway-based approach to mitigate bus bunching through systematic adjustments to bus departures. Delgado et al. (Delgado et al., 2012) further investigated the potential benefits of holding and limiting boarding strategies to improve transit performance under varying operational scenarios. Similarly, Xuan et al. (Xuan et al., 2011) developed an optimal linear control model for bus holding, analyzing its impact on schedule reliability and demonstrating its effectiveness in reducing bus bunching. Chen et al. (Chen et al., 2013) implemented an iterative headway-based holding strategy that integrates real-time information to dynamically adjust bus operations. These methods have demonstrated strong analytical foundations but often lack the adaptability needed for dynamic and stochastic urban transit environments.

REINFORCEMENT LEARNING: DQN-BASED APPROACHES

Recent advancements in deep reinforcement learning (RL) have enabled the development of adaptive bus holding strategies capable of addressing the complexities of dynamic transit environments. Among these, methods based on deep Q-networks (DQN) have shown promise in handling the stochastic nature of urban bus operations. Chen et al. (Chen et al., 2016) implemented a real-time bus holding strategy using a multi-agent DQN framework, demonstrating its potential to optimize transit performance on high-frequency corridors. Liu et al. (Liu et al., 2023) proposed a deep reinforcement learning-based holding control approach that accounts for stochastic travel times and demand variations, highlighting significant improvements in service regularity and passenger satisfaction. Wang and Sun (Wang & Sun, 2020) explored a dynamic holding strategy using a DQN-based multi-agent framework, showcasing the effectiveness of distributed cooperation among buses in minimizing headway deviations. Wu et al. (Wu et al., 2024) developed a hybrid reinforcement learning and imitation learning approach for dynamic scheduling in flexible bus services, incorporating holding control as a key element to enhance efficiency. These methods leverage the ability of DQNs to approximate optimal policies in high-dimensional state-action spaces, enabling robust performance under uncertainty.

B.2 REINFORCEMENT LEARNING: ACTOR-CRITIC-BASED APPROACHES

Actor-critic-based reinforcement learning frameworks have also emerged as a powerful tool for optimizing bus holding control in complex, multi-agent transit networks. Wang and Sun (Wang & Sun, 2021) developed a multi-agent actor-critic framework for adaptive headway control, enabling coordinated learning across buses to maintain service regularity. Rodriguez et al. (Rodriguez et al., 2023) introduced a deep reinforcement learning framework incorporating cooperative bus holding and stop-skipping strategies, leveraging centralized training and decentralized execution for enhanced operational efficiency. Yu et al. (Yu et al., 2024) explored large language model-enhanced reinforcement learning to design robust bus holding strategies, addressing asynchronous decision-making and system complexity. Zhao et al. (Zhao et al., 2025) proposed an actor-critic model with a focus on passenger transfer optimization and heterogeneous fleets, demonstrating improvements in service reliability and passenger experience. These actor-critic approaches enable the integration of continuous action spaces and multi-agent coordination, making them particularly effective in dynamic and uncertain environments.

B.3 MULTI-AGENT REINFORCEMENT LEARNING APPROACHES

Multi-agent reinforcement learning (MARL) has emerged as an effective framework for addressing complex, dynamic problems in transit systems by enabling cooperation among multiple agents. One prominent approach is centralized training with decentralized execution (CTDE) (Amato, 2024), where a centralized entity trains the agents collectively, but decisions are made locally by individual agents during operation. This paradigm enhances scalability and adaptability, making it particularly suitable for large-scale bus or train networks. Low et al. (Low et al., 2024) applied MARL to develop robust, real-time control strategies for high-frequency bus services, leveraging CTDE to optimize holding control and reduce service irregularities. Wu et al. (Wu et al., 2024) proposed a hybrid MARL framework that integrates imitation learning for flexible bus scheduling with holding and passenger assignment optimization.

Another key development in MARL for transit systems is the incorporation of cooperative learning mechanisms to improve system performance and passenger experience. Zhao et al. (Zhao et al., 2025) introduced a cooperative MARL framework for customized bus network design, integrating holding control and heterogeneous fleet optimization. Furthermore, Zhang et al. (Zhang et al., 2025) proposed a decentralized MARL framework for hydrogen energy storage systems that can be adapted for bus networks, ensuring stability and reliability in power-intensive transit systems. These MARL-based approaches highlight the potential of cooperative multi-agent systems in achieving decentralized yet coordinated decision-making for dynamic transit operations.

C Implementation Settings and Runtime

C.1 Hyperparameter Settings

Table 3 summarizes the hyperparameters used for SMART-HOLD. We train three regional teachers for 100 episodes using Adam. The actor and critic learning rates are 1×10^{-4} and 3×10^{-4}, respectively. Each regional replay buffer stores up to 20,000 transitions, and 128 gradient updates are shared across the three teachers after each episode.

Exploration noise is initialized at 0.3, multiplied by 0.995 after each episode, and bounded below by 0.15. Holding actions are restricted to $[0, 2]$ minutes. Every five episodes, the regional teachers are distilled into the student using a balanced batch of 256 observations and 25 mean-squared-error updates. Distillation is unconditional and begins at episode 5. Route R1 contains approximately $2{,}584 \pm 11$ event-driven simulator steps per episode.

D State Representations

The temporally augmented state-action pair defined in Equation 7 is:

$$(s, a) = o \| a \| o^+ \| a^+ \qquad = \begin{bmatrix} o_1 & 0 & o_1^+ & a_1^+ \\ \vdots & \vdots & \vdots & \vdots \\ o_i & a_i & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ o_n & 0 & o_n^+ & a_n^+ \end{bmatrix},$$

For agents that are active during this time period, these terms capture the corresponding states and actions of the agents at their decision times. Conversely, for passive agents, these values are set to 0. The feature vector o_j^+ employs the same types of features as o_j in Equation 4. The resulting (s, a) matrix has a shape of $n \times d_s$, where n is the number of buses, and $d_s = 2 \times d_o + 2$. This dimensionality accounts for two observation vectors as defined in Equation 4 and two action components.

Figure 4 compares state representation used in CAAC and our approach.

E Neural Networks

E.1 Cross-Attention

In the implementation of the cross-attention mechanism, as described in Equation 10, the model computes an *attended representation* from the subject agent to other agents (e.g., upstream agents), as illustrated in Figure 5 for the upstream cross-attention z_{up} in Section *Actor Network Design*. The mechanism employs learnable projection matrices $\mathbf{W}_Q \in \mathbb{R}^{d_q \times d_{qk}}$ and $\mathbf{W}_K \in \mathbb{R}^{d_k \times d_{qk}}$ to map the query vector $\mathbf{q}$ and key matrix $\mathbf{K}$ into a shared latent space of dimension d_{qk}, resulting in the transformed representations $\mathbf{q}'$ and $\mathbf{K}'$. A separate feedforward module, MLP_v, projects the value matrix $\mathbf{V}$ into another latent space of dimension d_{proj}, yielding $\mathbf{V}'$.

To compute attention scores, a multi-layer perceptron (MLP) is applied to each concatenated pair $[\mathbf{q}' \| \mathbf{k}'_i] \in \mathbb{R}^{2d_{qk}}$, producing a scalar compatibility score for each key. These scores are passed through a softmax function to form a valid probability distribution over the keys. The resulting attention weights are used to compute a convex combination of the transformed value vectors $\mathbf{V}'$,

Table 3: Hyperparameters used to train SMART-HOD on Route R1.

Parameter	**Value**
Regional teachers (k)	3
Training episodes	100
Actor learning rate	1×10^{-4}
Critic learning rate	3×10^{-4}
Discount factor (γ)	0.99
Target update rate (τ)	0.001
Training batch size	256
Replay capacity	20,000 per region
Updates per episode	128 total across teachers
Initial action-noise std.	0.3
Minimum action-noise std.	0.15
Action-noise decay	0.995 per episode
Action range	$[0, 2]$ minutes
Distillation interval	5 episodes
First distillation	Episode 5
Distillation batch size	256, region-balanced
Distillation updates	25
Distillation learning rate	1×10^{-4}
Distillation objective	Action MSE
Validation interval	5 episodes
Validation episodes	3
Optimizer	Adam
Observation dimension	7 per bus
Reconciled critic dimension	16 per bus
Attention embedding dimension	32
Attention output dimension	400
MLP hidden dimension	400
Actor parameters	357,043
Critic parameters	808,401
Bunching threshold	0.2 km
Bunching penalty coefficient	0.2
Holding penalty coefficient	0.05

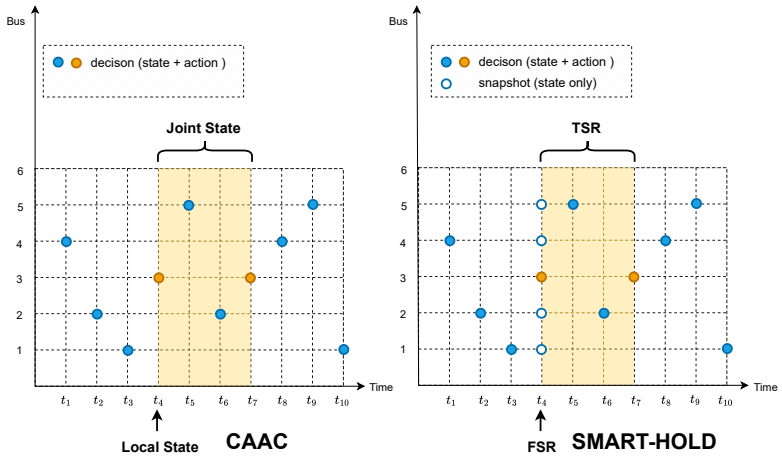


Figure 4: State Representation Comparison: CAAC uses a local snapshot for the actor and a joint state of active agents for the critic. In contrast, SMART-HOLD employs a fleet-wide snapshot (FSR) for the actor and a temporally augmented state (TSR) for the critic.

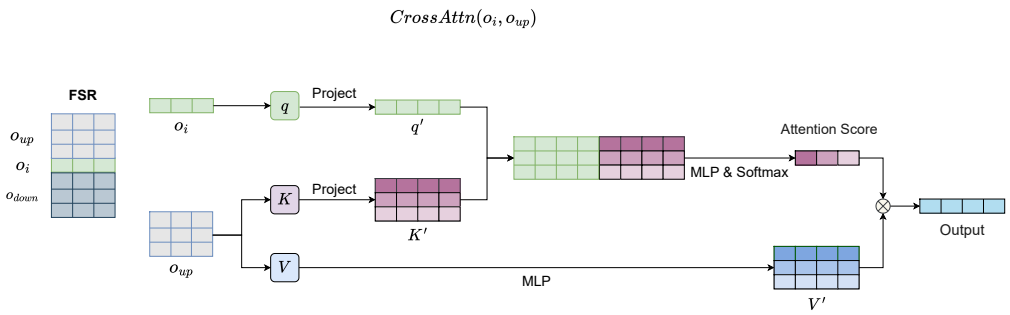


Figure 5: MLP-based Cross-Attention: The subject agent (o_i) "attends to" upstream agents (o_{up}).

yielding the final attended output. This process allows the subject agent to incorporate contextual information from other agents' states into its own representation.

Actor Network

The actor processes the subject bus observation, o_i, together with the observations of upstream and downstream buses. Separate cross-attention modules aggregate the two directional neighborhoods:

$$z = z_i + z_{\text{up}} + z_{\text{down}}, \tag{12}$$

where z_i is obtained by projecting the seven-dimensional subject observation through $\texttt{Linear}(7, 400)$. Each cross-attention module embeds its query and neighboring observations into a 32-dimensional space and produces a 400-dimensional output. Its attention scorer follows $64 \to 32 \to 8 \to 1$. If no bus is present on one side, the corresponding attention output is zero.

The aggregated representation is processed by an MLP with dimensions $400 \to 400 \to 400 \to 1$ and ReLU activations in its hidden layers. The final output is transformed using Tanh and rescaled to the allowable holding range:

$$a_{i,t} = [\tanh(f_\theta(z)) + 1] \frac{a_{\max}}{2}, \tag{13}$$

where $a_{\max} = 2$ minutes. Thus, the actor produces an additional holding time within $[0, 2]$ minutes, or $[0, 120]$ seconds.

During training, zero-mean Gaussian noise is added to the actor output before action clipping. Its standard deviation is initialized at 0.3 minutes, multiplied by 0.995 after each episode, and bounded below by 0.15 minutes. Table 4 reports the corresponding 95% noise interval before clipping.

Table 4: Decay of the Gaussian action-noise interval before action clipping.

Episode	95% Noise Interval (seconds)
0	± 35.28
25	± 31.12
50	± 27.46
75	± 24.22
100	± 21.37

Critic Network

The centralized critic estimates $Q(s, a)$ from the reconciled fleet state (TSR). An active bus that makes another decision within the subject bus's reconciliation window is represented by its current and second observation–action pairs, whereas the subject and passive buses retain only their current pair. Only buses operating in the current fleet snapshot are considered.

Algorithm 2 Regionalized Training with Distilled Coordination (RTDC)

Require: Regions k, episodes $E = 100$, updates per region $U = 256$, distillation start $E_{\text{start}} = 20$, interval $I = 5$

Initialize student actor π_θ, regional teacher actors $\{\pi_{\phi_j}\}_{j=1}^{k}$, shared critic Q_ψ, target networks, and regional replay buffers $\{\mathcal{B}_j\}_{j=1}^{k}$
for $e = 1, \ldots, E$ **do**
  Collect transitions using $a_t = \pi_\theta(o_t) + \epsilon_t$
  Store each transition in its corresponding regional buffer $\mathcal{B}_j$
  for $j = 1, \ldots, k$ **do**
    for $u = 1, \ldots, U$ **do**
      Sample a minibatch $\mathcal{D}_j \sim \mathcal{B}_j$
      Update the shared critic Q_ψ
      Update regional DDPG actor π_{ϕ_j}
      Soft-update the corresponding target networks
    end for
  end for
  if $k > 1 \wedge e > E_{\text{start}} \wedge e \bmod I = 0$ **then**
    Sample approximately balanced observation sets $\{\mathcal{S}_j\}_{j=1}^{k}$
    Update π_θ by minimizing

$$\mathcal{L}_{\text{distill}} = \frac{1}{\sum_j |\mathcal{S}_j|} \sum_{j=1}^{k} \sum_{o \in \mathcal{S}_j} \left\| \pi_\theta(o) - \pi_{\phi_j}(o) \right\|_2^2$$

    Evaluate π_θ and retain it if it is the best distilled student so far
    Synchronize all teachers and actor targets with π_θ
  end if
end for

Ensure: Best-performing distilled student actor π_{θ^*}

Using the subject's current observation–action pair as the query, separate cross-attention modules aggregate active upstream, active downstream, and passive buses. Each attention output is batch-normalized and processed by a two-layer MLP with hidden dimension 400 and ReLU activations. The subject pair is processed by a separate MLP of the same width. The four scalar components are summed to obtain

$$Q(s, a) = Q_i + Q_{\text{act_up}} + Q_{\text{act_down}} + Q_{\text{passive}}.$$

This role-separated architecture preserves the asymmetric effects of upstream, downstream, and passive buses on the subject bus's delayed return.

F Training Algorithm

As shown in Algorithm 2, Regionalized Training with Distilled Coordination (RTDC) partitions a route of m stops into k regions. Each region maintains a teacher actor and replay buffer, while all teachers share a centralized critic. A student actor collects experience and is periodically trained to imitate the regional teachers. After every distillation, the student is copied to all teachers, whereas the best-performing distilled student is retained for deployment.

DDPG Updates

For a regional batch $\mathcal{D}_j$, the actor receives the Fleet Snapshot Representation (FSR), denoted by o, while the shared critic receives the Temporally-Augmented State Representation (TSR) and actions, denoted by (s, a). For the next transition, the subject action in the next TSR is replaced by the target teacher output $\pi_{\phi_j^-}(o')$, yielding $(s', \tilde{a}')$. The critic target and loss are

$$y = r + \gamma(1 - d) Q_{\psi^-}(s', \tilde{a}'), \tag{14}$$

$$\mathcal{L}_Q = \mathbb{E}_{\mathcal{D}_j} \left[\ell_{\text{H}}(Q_\psi(s, a) - y) \right], \tag{15}$$

where ℓ_{H} is the Huber loss.

To update teacher j, the subject action in the current TSR is replaced by $\pi_{\phi_j}(o)$, yielding $(s, \tilde{a})$. The teacher minimizes

$$\mathcal{L}_{\pi_j} = -\mathbb{E}_{\mathcal{D}_j}\left[Q_\psi(s, \tilde{a})\right].$$

The teacher and shared critic target networks are updated using

$$\phi_j^- \leftarrow (1-\tau)\phi_j^- + \tau\phi_j, \qquad \psi^- \leftarrow (1-\tau)\psi^- + \tau\psi.$$

For $k = 1$, distillation is disabled, and the student actor is trained directly using these DDPG updates.

Behavioral Cloning

For $k > 1$, distillation is performed every five episodes after episode 20, with the first distillation occurring at episode 25. The student is optimized for 25 gradient steps using observations sampled approximately equally across regions, with a nominal batch size of 256. Teacher actions are generated online as supervised labels. After distillation, the current student is copied to all regional teachers and their target networks, enabling region-specific refinement from a common policy. The deterministic student return is used to retain the best distilled policy for deployment. Thus, RTDC consolidates knowledge from k regional policies into a single deployable actor while preserving scalable region-specific training.

G Performance Metrics

We evaluate service reliability using bunching events, headway coefficient of variation, and excess waiting time.

- **Bunching.** An adjacent bus pair is classified as bunched when its physical spacing $d_{j,t}$ is below $d_{\mathrm{b}} = 0.2\,\mathrm{km}$:

  $$b_{j,t} = \mathbb{I}(d_{j,t} < d_{\mathrm{b}})\,. \tag{16}$$

  The reported number of bunching events counts newly formed bunched pairs:

  $$B = \sum_t \sum_j \mathbb{I}\big(b_{j,t} = 1,\ b_{j,t^-} = 0\big)\,, \tag{17}$$

  where t^- denotes the preceding decision epoch. Consequently, a persistent bunched pair is counted once, but is counted again if it separates and later reforms. Only adjacent operating buses are considered, without circular terminal-to-origin adjacency.

- **Headway Coefficient of Variation.** The route contains m stops, denoted by stop_i, $i = 1, \ldots, m$. Let $t_{i,j}$ be the arrival time of the j-th bus at stop_i. The corresponding arrival headway is

  $$h_{i,j} = t_{i,j} - t_{i,j-1}.$$

  Given n_i observed headways at stop_i, their mean and variance are

  $$\bar{h}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} h_{i,j}, \qquad \sigma_i^2 = \frac{1}{n_i}\sum_{j=1}^{n_i}(h_{i,j} - \bar{h}_i)^2. \tag{18}$$

  The Headway CV at stop_i is

  $$\mathrm{CV}_i = \frac{\sigma_i}{\bar{h}_i}.$$

  The episode-level Headway CV is averaged across stops with at least two observed headways:

  $$\mathrm{CV} = \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}} \mathrm{CV}_i, \qquad \mathcal{I} = \{i : n_i \geq 2\}. \tag{19}$$

  A lower Headway CV indicates more regular bus arrivals, with zero representing perfectly equal headways.

- **Excess Waiting Time.** Assuming that passengers arrive independently of the timetable at a constant rate, a passenger arriving during a headway of length h waits $h/2$ on average. However, longer headways contain proportionally more passenger arrivals and are therefore encountered with probability proportional to their length. Consequently, the expected passenger waiting time at stop_i is

$$\mathbb{E}[W_i] = \frac{\mathbb{E}[h_i^2]}{2\mathbb{E}[h_i]}. \tag{20}$$

For perfectly regular service with the same mean headway, the expected waiting time is $\mathbb{E}[h_i]/2 = \bar{h}_i/2$. Excess Waiting Time is therefore the additional waiting caused by headway irregularity:

$$\mathrm{EWT}_i = \frac{\mathbb{E}[h_i^2]}{2\mathbb{E}[h_i]} - \frac{\mathbb{E}[h_i]}{2} \tag{21}$$

$$= \frac{\sigma_i^2}{2\bar{h}_i}, \tag{22}$$

where the second equality follows from $\mathbb{E}[h_i^2] = \sigma_i^2 + \bar{h}_i^2$.

The episode-level EWT is averaged across stops with at least two observed headways:

$$\mathrm{EWT} = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \mathrm{EWT}_i, \qquad \mathcal{I} = \{i : n_i \geq 2\}. \tag{23}$$

EWT is measured in minutes. A lower value indicates more regular service and less additional passenger waiting, while $\mathrm{EWT} = 0$ corresponds to perfectly regular headways.

Lower values indicate better performance for all three metrics.

H Baseline Methods and Ablation Study

H.1 Baseline Methods

We first consider two non-learning baselines. No Control (NC) dispatches buses according to the timetable without applying additional holding. Naive Headway Control (NHW)(Daganzo, 2009) applies a deterministic holding rule based on the difference between the scheduled headway and the estimated forward headway to the preceding bus. The resulting holding time is clipped to the allowable action range. Because NC and NHW do not require policy training, they are included in the performance evaluation but excluded from the neural-network comparison in Table 5.

We further compare SMART-HOLD with four learning-based baselines: CAAC (Wang & Sun, 2021), IAC (Lillicrap et al., 2016), MADDPG (Lowe et al., 2017), and MAPPO(Yu et al., 2022). CAAC uses a three-hidden-layer local actor and an event-attention critic that separately aggregates upstream and downstream decision events occurring before the subject bus reaches its next control point. Both components are trained using DDPG.

IAC uses only the subject bus's local information. From the raw seven-dimensional observation, it constructs six normalized, route-independent features comprising route progress, remaining dwell time, forward- and backward-headway proportions, log-scaled total headway, and motion state. Its actor uses a $6-400-400-1$ MLP. Its critic concatenates the same six features with the subject action and uses a $7-400-400-1$ MLP. Thus, IAC uses no information from other operating buses.

MADDPG and MAPPO follow centralized training with decentralized execution. Their actors receive the subject bus's raw seven-dimensional observation, from which the constant subject indicator is removed, resulting in six actor inputs. The MADDPG actor uses two 400-unit ReLU layers followed by a sigmoid output, whereas the MAPPO actor uses two 400-unit Tanh layers and parameterizes a Beta action distribution.

During training, both centralized critics use a single fleet snapshot containing only currently operating buses. A shared two-layer MLP with 400-unit Tanh layers encodes each bus independently. The representations of all non-subject buses are mean-pooled into a permutation-invariant "others" representation. This representation is concatenated with the subject representation, its absolute

Table 5: Comparison of the information inputs and neural-network architectures of SMART-HOLD and the learning-based baselines.

Property	SMART-HOLD	CAAC	IAC	MADDPG	MAPPO
Training Method	RTDC–DDPG	DDPG	DDPG	MADDPG	PPO
Actor Input	FSR	Local	Local	Local	Local
Critic Input	TSR	Subject and active state + action	Local state + action	Global state + action (single snapshot)	Global state + action (single snapshot)
Fleet Aggregation	Role-separated cross-attention	Event-based attention	None	Mean pooling over all other buses	Mean pooling over all other buses
Actor Network	Cross-attention + 400-unit MLP	Three-hidden-layer 400-unit MLP	Two-hidden-layer 400-unit MLP	Two-hidden-layer 400-unit MLP	Two-hidden-layer 400-unit MLP
Critic Network	Cross-attention + 400-unit MLPs	Event attention + 400-unit MLPs	Two-hidden-layer 400-unit MLP	Mean pooling + 400-unit MLP	Mean pooling + 400-unit MLP

route progress, and the normalized active-fleet size. MADDPG additionally includes each bus's action entry from the original snapshot and estimates $Q(s, a)$; because decisions are asynchronous, only the subject action is nonzero. MAPPO excludes actions and estimates $V(s)$. The resulting $400 + 400 + 1 + 1 = 802$-dimensional representation is processed by an $802-400-400-1$ Tanh value network. For a capacity-matched comparison, IAC, MADDPG, and MAPPO use two 400-unit hidden layers.

Table 5 illustrates the five learning based methods we used as baseline methods.

H.2 Performance Comparison

Table 6 compares SMART-HOLD with heuristic and learing–based baselines across six test routes. SMART-HOLD achieves the lowest Headway CV and EWT on every route, as well as the best Bunching and reward results on all six routes. Overall, the results demonstrate that SMART-HOLD consistently improves service regularity and passenger waiting performance across diverse operating conditions.

We assess statistical significance using two-sided paired Wilcoxon signed-rank tests, treating each independently trained policy as one experimental unit and averaging its evaluation episodes. For each route and metric, Holm's procedure is applied across the four comparisons between SMART-HOLD and the learning-based baselines at $\alpha = 0.05$. Of the 112 comparisons, 38 remain significant after correction. The evidence is strongest for Headway CV and EWT, with 13 significant comparisons for each metric. In particular, SMART-HOLD significantly outperforms MAPPO in both metrics on all seven routes. On M1 and M7, the improvements in Headway CV and EWT are significant against all four learning-based baselines. The corresponding rank-biserial correlations range from 0.782 to 1.000, indicating large and consistently directed effects. By contrast, significance is less consistent for Bunching and reward, with five and seven significant comparisons, respectively. These findings support the robustness of SMART-HOLD's improvements in service regularity and passenger waiting time, while indicating that its advantages in Bunching and reward are more dependent on the route and baseline.

H.3 Ablation Studies

To evaluate the contribution of each component in SMART-HOLD, we conduct five ablation studies targeting its state representation, attention architecture, and training strategy. Each ablation modifies only one component while retaining the remaining architecture and experimental settings. Table 7

Table 6: Test performance on routes R1, R3, M1, M2, M3, and M7 in terms of bunching count, headway coefficient of variation (CV), excess waiting time (EWT), and reward. Results are reported as mean ± standard deviation over the evaluation runs. The best mean result for each route is highlighted in bold. Lower values indicate better performance for all Bunchings, Headway CV, EWT, and higher values are better for reward. The best and second-best mean results for each route and metric are highlighted in bold and underlined, respectively.

Route	Metric	NC (No Control)	NHW (Daganzo, 2009)	CAAC (Wang & Sun, 2021)	IAC (Lillicrap et al., 2016)	MADDPG (Lowe et al., 2017)	MAPPO (Yu et al., 2022)	SMART-HOLD (Ours)
R1	Bunchings	45.40 ± 18.51	50.40 ± 16.40	7.80 ± 14.71	5.80 ± 15.74	16.51 ± 24.92	1.78 ± 1.69	**1.45 ± 0.91**
	Headway CV	0.49 ± 0.04	0.46 ± 0.02	0.27 ± 0.09	0.24 ± 0.08	0.30 ± 0.13	0.25 ± 0.03	**0.21 ± 0.02**
	EWT (min)	1.83 ± 0.29	1.57 ± 0.18	0.62 ± 0.48	0.51 ± 0.43	0.80 ± 0.69	0.50 ± 0.11	**0.35 ± 0.07**
	Reward	-676.54 ± 110.77	-590.07 ± 69.44	-218.44 ± 160.44	-157.04 ± 157.14	-282.68 ± 253.83	-189.75 ± 37.28	**-114.58 ± 19.18**
R3	Bunchings	90.40 ± 26.56	75.40 ± 14.60	16.14 ± 28.70	12.67 ± 28.54	31.08 ± 43.09	6.68 ± 4.39	**5.97 ± 3.64**
	Headway CV	0.55 ± 0.05	0.51 ± 0.03	0.30 ± 0.10	0.29 ± 0.10	0.35 ± 0.14	0.30 ± 0.03	**0.24 ± 0.03**
	EWT (min)	2.48 ± 0.45	2.04 ± 0.23	0.83 ± 0.64	0.75 ± 0.63	1.12 ± 0.93	0.72 ± 0.15	**0.49 ± 0.10**
	Reward	-1159.96 ± 197.06	-933.00 ± 100.43	-345.36 ± 274.82	-302.32 ± 292.78	-485.35 ± 418.50	-324.30 ± 60.48	**-246.53 ± 48.24**
M1	Bunchings	13.50 ± 9.07	6.00 ± 0.60	2.03 ± 6.01	1.88 ± 6.23	4.19 ± 8.07	0.01 ± 0.10	**0.00 ± 0.00**
	Headway CV	0.49 ± 0.04	0.54 ± 0.02	0.31 ± 0.09	0.28 ± 0.08	0.32 ± 0.11	0.26 ± 0.03	**0.23 ± 0.03**
	EWT (min)	2.41 ± 0.43	2.95 ± 0.22	1.03 ± 0.58	0.88 ± 0.58	1.17 ± 0.76	0.72 ± 0.17	**0.55 ± 0.15**
	Reward	-547.33 ± 106.05	-698.56 ± 35.73	-272.69 ± 121.34	-220.14 ± 123.36	-277.06 ± 158.01	-255.06 ± 45.21	**-170.26 ± 27.98**
M2	Bunchings	145.30 ± 24.88	153.10 ± 28.20	22.42 ± 43.29	14.14 ± 43.56	41.53 ± 65.21	6.02 ± 7.69	**2.25 ± 5.77**
	Headway CV	0.58 ± 0.03	0.65 ± 0.02	0.33 ± 0.11	0.30 ± 0.10	0.37 ± 0.14	0.31 ± 0.03	**0.27 ± 0.03**
	EWT (min)	1.81 ± 0.21	2.24 ± 0.16	0.64 ± 0.46	0.56 ± 0.47	0.84 ± 0.66	0.52 ± 0.09	**0.41 ± 0.08**
	Reward	-1615.95 ± 122.86	-2224.60 ± 175.36	-408.35 ± 314.94	-321.10 ± 358.78	-544.50 ± 506.69	-346.05 ± 50.13	**-302.08 ± 103.04**
M3	Bunchings	65.00 ± 22.16	65.60 ± 15.40	13.14 ± 21.97	7.09 ± 21.77	19.44 ± 30.18	5.11 ± 5.17	**0.27 ± 0.72**
	Headway CV	0.57 ± 0.05	0.58 ± 0.03	0.31 ± 0.10	0.28 ± 0.10	0.33 ± 0.15	0.30 ± 0.03	**0.26 ± 0.02**
	EWT (min)	2.03 ± 0.40	2.10 ± 0.22	0.67 ± 0.51	0.56 ± 0.49	0.83 ± 0.71	0.58 ± 0.09	**0.42 ± 0.07**
	Reward	-722.46 ± 164.14	-722.45 ± 79.17	-203.04 ± 167.54	-151.35 ± 172.99	-266.10 ± 248.80	-173.50 ± 34.56	**-110.10 ± 21.02**
M7	Bunchings	64.80 ± 17.50	90.70 ± 20.10	41.41 ± 41.95	8.34 ± 24.29	26.21 ± 38.98	10.95 ± 8.16	**1.59 ± 3.11**
	Headway CV	0.53 ± 0.03	0.58 ± 0.03	0.34 ± 0.09	0.30 ± 0.08	0.35 ± 0.13	0.31 ± 0.02	**0.25 ± 0.02**
	EWT (min)	1.33 ± 0.13	1.58 ± 0.15	0.58 ± 0.33	0.45 ± 0.31	0.66 ± 0.47	0.46 ± 0.06	**0.30 ± 0.03**
	Reward	-670.81 ± 81.82	-768.48 ± 96.74	-296.62 ± 153.37	-184.64 ± 153.11	-307.73 ± 234.37	-197.61 ± 25.93	**-132.06 ± 23.25**

Table 7: Ablation study on routes R1 and R3. Results are reported as mean ± standard deviation. Best and second-best results are highlighted in bold and underlined, respectively.

Route	Metric	SMART-HOLD	No Distillation	Local Actor	Single-Snapshot Critic	Dot-Product Attention	Unified Attention
R1	Bunchings	**1.45 ± 0.91**	3.20 ± 3.56	11.62 ± 23.09	2.76 ± 3.71	11.32 ± 18.94	3.94 ± 6.52
	Headway CV	**0.2118 ± 0.0216**	0.2535 ± 0.0270	0.2649 ± 0.1117	0.2372 ± 0.0449	0.2818 ± 0.1102	0.2636 ± 0.0576
	EWT (min)	**0.3493 ± 0.0684**	0.5007 ± 0.1092	0.6346 ± 0.5950	0.4527 ± 0.1717	0.7028 ± 0.5971	0.5516 ± 0.2403
	Reward	**−114.58 ± 19.18**	−156.79 ± 33.84	−223.04 ± 217.90	−149.72 ± 44.78	−240.14 ± 217.19	−172.12 ± 80.63
R3	Bunchings	5.97 ± 3.64	6.76 ± 4.42	21.16 ± 37.99	**5.94 ± 3.73**	23.60 ± 38.12	8.78 ± 6.82
	Headway CV	**0.2436 ± 0.0292**	0.2837 ± 0.0261	0.3154 ± 0.1277	0.2529 ± 0.0454	0.3243 ± 0.1208	0.3145 ± 0.0669
	EWT (min)	**0.4912 ± 0.1042**	0.6542 ± 0.1208	0.9253 ± 0.8264	0.5389 ± 0.1864	0.9591 ± 0.7974	0.8184 ± 0.3512
	Reward	**−246.53 ± 48.24**	−258.93 ± 38.29	−401.51 ± 368.45	−247.58 ± 58.50	−413.88 ± 366.60	−323.73 ± 119.16

reports the resulting Bunching, Headway CV, excess waiting time (EWT), and reward on three representative test routes.

H.3.1 Knowledge Distillation

This ablation removes the knowledge-distillation stage from Regionalized Training with Distilled Coordination (RTDC). The state representation, actor–critic architecture, hyperparameters, and evaluation procedure remain identical to those of SMART-HOLD. The only difference is the training strategy: SMART-HOLD distills the knowledge learned by region-specific teacher policies into a shared student actor, whereas the ablation trains the shared policy without teacher-to-student distillation. This comparison isolates the contribution of knowledge transfer to the performance and generalization of the resulting policy. In addition to the four evaluation metrics in Table 7, we compare training reward and Bunching to examine the effect of distillation on training stability and convergence.

H.3.2 Local-Observation Actor

This ablation restricts the actor to the subject bus's local observation and removes fleet-level attention from action selection. The critic remains identical to the full SMART-HOLD critic, including its role-separated cross-attention and reconciled temporal state. This setting isolates the contribution

Table 8: Training performance with and without RTDC at different training stages. Results are reported as mean ± standard deviation over 10 seeds. The best mean result at each stage is highlighted in bold. Higher reward and lower Bunching indicate better performance.

Stage	Metric	No Distillation ($k = 1$)	SMART-$k3$	SMART-$k4$
Episode 30	Reward	-518.66 ± 126.54	-594.52 ± 141.63	$\mathbf{-464.49 \pm 221.58}$
	Bunchings	$\mathbf{28.33 \pm 22.79}$	46.51 ± 19.52	31.03 ± 26.54
Episode 60	Reward	-344.45 ± 180.77	-304.36 ± 198.61	$\mathbf{-252.87 \pm 130.86}$
	Bunchings	11.74 ± 20.70	11.19 ± 20.41	$\mathbf{9.03 \pm 14.79}$
Final	Reward	-244.89 ± 150.53	$\mathbf{-174.98 \pm 49.99}$	-178.53 ± 58.74
	Bunchings	6.18 ± 14.52	$\mathbf{1.62 \pm 0.60}$	2.37 ± 2.52

of fleet-wide information to decentralized action selection while retaining centralized information during training.

H.3.3 Single-Snapshot Critic

This ablation replaces the reconciled temporal state used by the SMART-HOLD critic with the initial fleet snapshot and corresponding action vector. The actor remains unchanged. The critic retains subject-centric cross-attention and separates non-subject buses into upstream and downstream groups, but it cannot distinguish active and passive buses because subsequent decision events are unavailable. This comparison evaluates the effect of temporal state reconciliation on credit assignment and value estimation.

H.3.4 Dot-Product Attention

This ablation replaces SMART-HOLD's customized MLP-based attention scoring with standard scaled dot-product attention in both the actor and critic. The subject-relative inputs, role-separated interaction groups, network dimensions, and training procedure remain unchanged. This comparison evaluates whether the proposed MLP-based scoring function captures asymmetric bus interactions more effectively than conventional dot-product similarity.

H.3.5 Unified Attention

This ablation replaces the separate upstream and downstream attention modules in both the actor and critic with a single cross-attention module over all non-subject operating buses. The FSR, TSR and remaining network architecture are unchanged. This setting evaluates whether explicitly separating buses by their relative roles and directions provides an advantage over learning a unified interaction representation.

Ablation Results. Table 7 reports Bunching, Headway CV, EWT, and reward on three representative test routes. SMART-HOLD achieves the lowest Headway CV and EWT on every route, demonstrating the most consistent performance in service regularity and passenger waiting time. It also achieves the best results for all four metrics on R1 and the best Headway CV, EWT, and reward on R3. The single-snapshot critic produces marginally fewer Bunching events on R3, with 5.94 compared with 5.97 for SMART-HOLD, while their rewards are also similar.

Removing fleet-level information from the actor or replacing the proposed MLP-based attention with dot-product attention consistently degrades all four metrics and generally increases variability. This result supports the use of fleet-wide information during decentralized execution and the proposed attention-scoring mechanism. Unified attention also performs worse than SMART-HOLD on R1 and R3, indicating that explicitly separating buses according to their interaction roles is beneficial under these conditions.

I PARAMETER SENSITIVITY ANALYSIS

Table 8 compares the training progress of SMART-HOLD with $k = 3$ and $k = 4$ regional teachers against training without distillation. By episode 60, both RTDC variants achieve higher rewards than the no-distillation baseline, while SMART-$k4$ also obtains the lowest Bunching count. At the end of training, both variants substantially outperform the baseline, with SMART-$k3$ achieving the best final reward and Bunching performance. These results indicate that regionalized training with knowledge distillation accelerates the development of effective policies and improves final training performance. Figure 6 further shows the trajectories of both metrics over the complete training process from episodes 1 to 100.

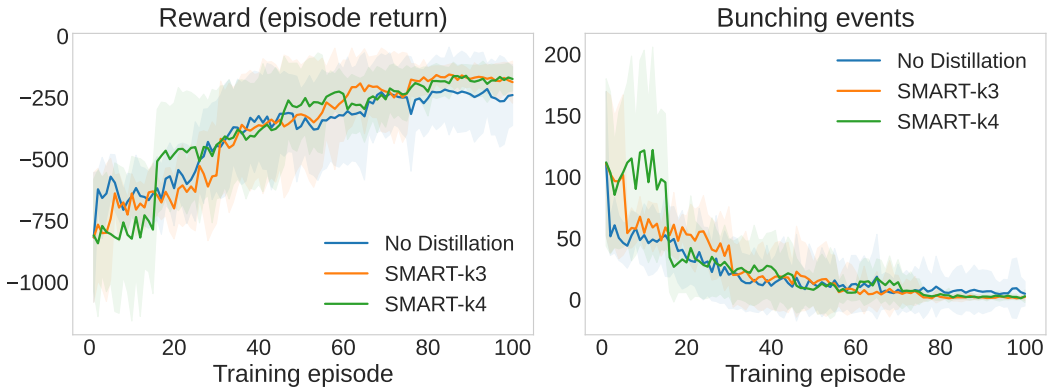


Figure 6: Training performance comparison of SMART-HOLD using different numbers of teacher agents (k = 1(no distillation), 3, 4).