EVA-Gen: When Perception Learns from Value via Generative Models in Decentralized Multi-Agent Systems

YuDuo Zheng[1,2], **XueFeng Du**[1,2], **Yanqi Cheng**[1,2], **Li Yin**[1,2] and **Fengqi Li**[2,3,*]
[1]School of Mechanical Engineering, University of Dalian Jiaotong, Dalian, China
[2]School of Rail Transit and Intelligent Engineering, University of Dalian Jiaotong, Dalian, China
[3]Blockchain and Intelligent Information Systems Laboratory
13234153818@163.com, xuefeng.du@outlook.com, fengqi-li@outlook.com

Abstract

Decentralized Multi Agent Reinforcement Learning (MARL) faces a fundamental dilemma in real world deployments: agents must operate under epistemic fragmentation, where local observations are severely occluded, while navigating heterogeneous value landscapes, where sparse, critical events carry disproportionately high stakes. Existing paradigms typically decouple state estimation from policy optimization, guiding perception modules merely to minimize uniform reconstruction error. This leads to a Perception Value Misalignment, where agents squander computational resources reconstructing task irrelevant background noise while failing to resolve uncertainties in high value regions. To bridge this gap, we propose EVA-Gen (Epistemic Value Alignment via Generative Models) that establishes a cybernetic loop between generative perception and value based decision making. We formulate the Value Conditioned Reconstruction Paradigm, establishing that optimal perception under resource constraints is functionally weighted by the gradient of the value function. EVA-Gen couples Backward Flow to steer diffusion toward high stakes manifolds, Collaborative Information Bottleneck to filter communication for value relevant consensus, and Risk Sensitive Rectification to prevent sparse signal dilution, synergistically closing the perception control loop. Empirically, we demonstrate that EVA-Gen achieves superior performance in three value-heterogeneous multi agent environments.

1 Introduction

The field of Multi-Agent Reinforcement Learning (MARL) [Zhang *et al.*, 2021; Buşoniu *et al.*, 2008] has rapidly become a cornerstone for distributed control in complex systems where centralized sensing and decision-making are impractical. These domains heavily rely on agents that must make decisions based on limited local observations and constrained communication channels. In these decentralized execution environments, each agent operates with a locally consistent yet globally incomplete view of the system, a challenge often referred to as *epistemic fragmentation* [Oliehoek and Amato, 2016; Dibangoye and Buffet, 2018]. Further compounding this issue is the presence of *heterogeneous value landscapes*, where agents must navigate through regions that carry disproportionately high stakes due to sparse, critical events. These challenges make real-world MARL deployment notoriously brittle.

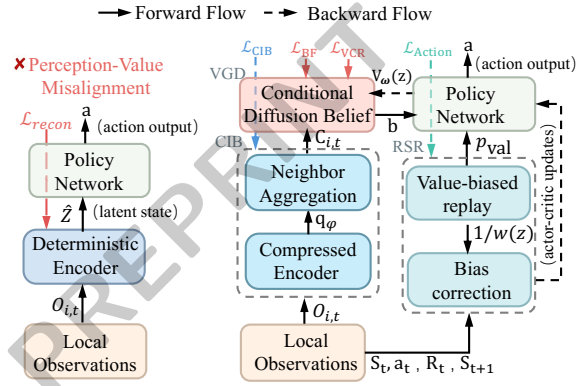

Figure 1: Overview: (a) The standard decoupled pipeline, where perception minimizes uniform reconstruction loss independent of decision relevance. (b) EVA-Gen, which closes the perception-control loop by steering generative belief inference toward value-sen·sitive regions via Backward Flow.

Traditional model-based architectures often decouple state representation from policy optimization. In these frameworks, perception modules are trained to minimize a uniform reconstruction error across the observation space, treating background noise and critical signals equally. While notable works such as DreamerV3 [Hafner *et al.*, 2025] and MuZero [Schrittwieser *et al.*, 2020] have made strides toward value-aware learning, they generally target settings with either single-agent or fully observable environments where a unified world model suffices. In decentralized systems, this separation creates a fundamental pathology: a Perception-Value Misalignment, where agents waste scarce computa-

*Corresponding author.

tional resources processing irrelevant background textures while failing to resolve uncertainties in regions where the value function gradient is most sensitive. This misalignment is especially detrimental when operating under tight communication or compute budgets.

To resolve this misalignment, we propose a paradigm shift that places the value function at the center of perception learning (Fig. 1). Our insight is rooted in the observation that under resource constraints, perceptual errors do not equally impact decision-making. Specifically, if an agent's objective is to maximize the expected return $V^\pi(s)$, the impact of a perceptual error $\varepsilon = \hat{s} - s$ can be approximated by a first-order Taylor expansion: $V^\pi(\hat{s}) \approx V^\pi(s) + \nabla_s V^\pi(s)^\top \varepsilon$. This relationship reveals that the error's impact on utility is proportional to the gradient magnitude $\|\nabla_s V^\pi(s)\|$. This observation motivates the concept of Value-Conditioned Reconstruction: optimal perception under fixed resource constraints is not uniform, but functionally weighted by the value gradient, prioritizing fidelity in high-stakes regions.

In response, we introduce **EVA-Gen** (Epistemic-Value Alignment via Generative Models), a unified framework that bridges the gap between generative perception and value-based decision-making. EVA-Gen incorporates this principle through three synergistic mechanisms: (1) **Backward Flow via Value-Guided Diffusion:** We utilize diffusion models [Ho *et al.*, 2020; Janner *et al.*, 2022] to model the multi-modal uncertainty in fragmented observations. By introducing a *value-gradient-weighted sampling* process, we bias the reverse diffusion steps to prioritize the generation of high-value latent manifolds, effectively guiding perception toward task-relevant areas. (2) **Collaborative Information Bottleneck:** To mitigate decentralized bandwidth limitations, we introduce a collaborative information bottleneck [Wang *et al.*, 2020] that filters communication. By maximizing mutual information specifically over value-relevant features, this mechanism fosters consensus on critical perceptual elements while discarding shared noise. (3) **Risk-Sensitive Gradient Rectification:** In environments with sparse high-value events, we counter signal dilution by adjusting policy gradients using uncertainty estimates derived from the generative model. This ensures the policy remains sensitive to sparse but critical signals despite the prevalence of low-value data.

In summary, the contributions of this work are as follows:

1. We formulate the Value-Conditioned Reconstruction Paradigm, establishing that optimal perception under resource constraints should be weighted by the value function gradient.
2. We introduce EVA-Gen, a novel architecture that integrates value-guided diffusion, collaborative bottlenecks, and risk-sensitive rectification to achieve tight epistemic-value alignment.
3. We provide extensive empirical validation on SMAC variants with sparse rewards, Value-oriented collaborative exploration tasks, and an edge computing testbed.

2 Related Work

Our work positions EVA-Gen at the intersection of generative belief inference, value-aware modeling, and communication-efficient coordination under resource constraints.

2.1 Generative beliefs under epistemic uncertainty

Decentralized control relies on inferring global states from partial observations. While recurrent encoders paired with CTDE methods (e.g., QMIX [Rashid *et al.*, 2020]) provide scalable coordination, they typically yield *deterministic pseudo-beliefs* that do not explicitly represent the *multimodal* posteriors induced by severe occlusion [Wang *et al.*, 2023]. This motivates generative belief modeling in POMDP RL [Igl *et al.*, 2018], where diffusion models improve mode coverage [Ho *et al.*, 2020; Janner *et al.*, 2022]. However, existing generative perception approaches optimize uniform denoising or reconstruction objectives, inducing a frequency bias: modeling capacity is allocated by visitation density rather than decision sensitivity. EVA-Gen bridges this gap by conditioning diffusion emphasis on value gradients, ensuring limited compute is spent on resolving high-stakes uncertainty.

2.2 Value equivalence vs. value conditioned reconstruction

Model-based RL (e.g., Dreamer, MuZero) learns latent dynamics for planning and control [Schrittwieser *et al.*, 2020]. A key principle is *Value Equivalence* [Grimm *et al.*, 2020]: representations need only preserve information relevant to Bellman updates. While conceptually aligned with task relevance, these methods are typically realized in centralized settings with objectives that remain effectively frequency-weighted; in heterogeneous value landscapes, this systematically underserves rare, high-impact regions. EVA-Gen complements value-equivalent modeling by shifting the focus from *state abstraction* to *resource allocation*, prioritizing perceptual fidelity according to value sensitivity under explicit compute/bandwidth budgets.

2.3 Communication-efficient coordination and risk-sensitive learning

To handle bandwidth constraints, explicit communication methods learn what to share [Sukhbaatar *et al.*, 2016], while Information Bottleneck objectives motivate compression [Alemi *et al.*, 2017]. Yet multi-agent bottlenecks define relevance via generic predictability [Su *et al.*, 2024], which may preserve shared but low-utility features. EVA-Gen introduces a *Collaborative Information Bottleneck* shaped by value-relevant supervision, so transmitted bits reduce uncertainty along decision-critical dimensions. While prioritization (e.g., PER [Bellemare *et al.*, 2017]) can amplify rare signals, EVA-Gen couples such emphasis with closed-form importance correction and uncertainty-aware rectification to preserve gradient fidelity under value-biased perception.

3 Preliminary

We begin by defining the *resource-constrained* decentralized control problem under epistemic fragmentation, then show why *uniform* perception is theoretically misaligned with control via a first-order value-degradation analysis.

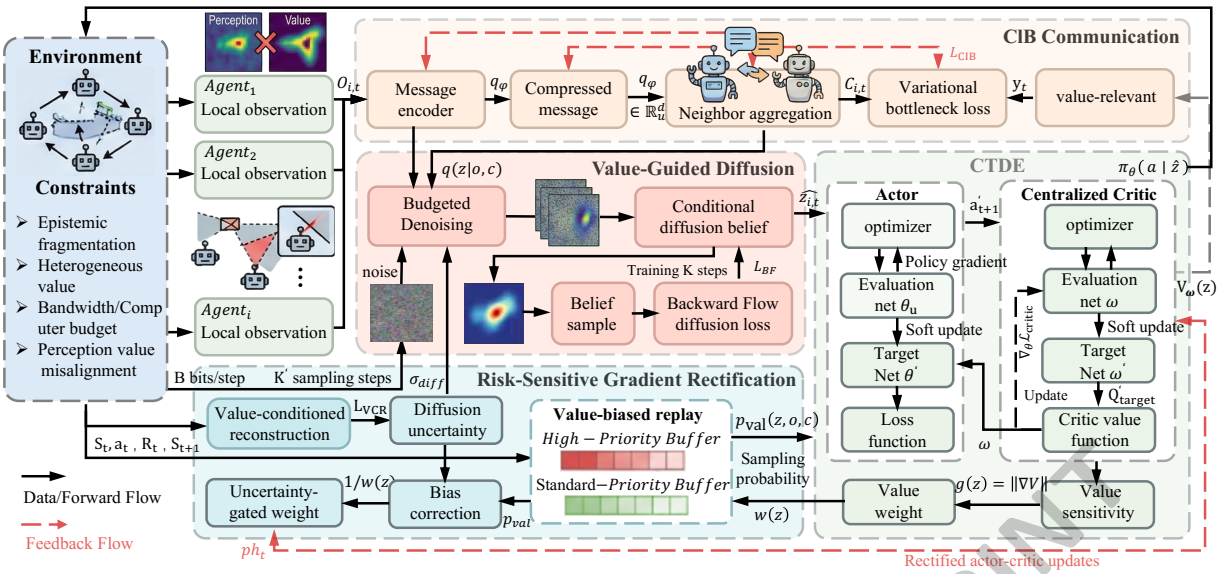

Figure 2: **The EVA-Gen Framework Architecture** establishes a cybernetic loop coupling generative perception with value-based decision-making. Starting from partial observations, agents align beliefs via a Collaborative Information Bottleneck that maximizes value-relevant mutual information under bandwidth constraints. Crucially, the perception engine operationalizes Value-Conditioned Reconstruction through Backward Flow, utilizing value gradients to bias generative sampling toward high-stakes manifolds. To ensure robust CTDE optimization, risk-sensitive weighting corrects this induced sampling bias and gates uncertainty, preventing gradient dilution from sparse critical events.

3.1 Problem Formulation

We consider a Resource-Constrained Decentralized Partially Observable Markov Decision Process (RC-Dec-POMDP),

$$\langle \mathcal{S}, \mathcal{N}, \{\mathcal{O}_i\}_{i\in\mathcal{N}}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, \Omega, \mathcal{P}, \mathcal{R}, \gamma, \mathcal{C} \rangle , \tag{1}$$

where $\mathcal{N} = \{1, \ldots, N\}$ is the agent set. At time t, the environment is in global state $s_t \in \mathcal{S}$. Each agent i receives a local observation

$$o_{i,t} \sim \Omega(\cdot \mid s_t, i), \tag{2}$$

which is severely occluded/noisy, inducing epistemic fragmentation. Agents take actions $a_{i,t} \in \mathcal{A}_i$, forming a joint action $\mathbf{a}_t = (a_{1,t}, \ldots, a_{N,t})$. The system transitions as $s_{t+1} \sim \mathcal{P}(\cdot \mid s_t, \mathbf{a}_t)$ and yields a shared reward $r_t = \mathcal{R}(s_t, \mathbf{a}_t)$.

Because s_t is unavailable at execution, EVA-Gen operates on a *control-sufficient* latent state $z_t \in \mathbb{R}^d$, defined by an encoder φ (learned under CTDE supervision):

$$z_t \triangleq \varphi(s_t). \tag{3}$$

During execution, agent i infers a belief $\hat{z}_{i,t}$ from its local observation and received messages (defined below), and executes a decentralized policy

$$\begin{aligned} a_{i,t} &\sim \pi_{\theta_i}(\cdot \mid \hat{z}_{i,t}), \\ \boldsymbol{\pi}_\theta(\mathbf{a}_t \mid \hat{\mathbf{z}}_t) &= \prod_{i=1}^{N} \pi_{\theta_i}(a_{i,t} \mid \hat{z}_{i,t}). \end{aligned} \tag{4}$$

We use CTDE to train a centralized critic on z_t, while enforcing decentralized execution through Eq. (4).

Agents can exchange compressed messages. Let $u_{i,t}$ be agent i's transmitted message and $\tilde{u}_{i,t}$ the received (possibly corrupted/dropped) message. We impose the constraint set

$$\mathcal{C} = \{\mathcal{B},\, C_{\text{comp}}\}, \tag{5}$$

where Bandwidth constraint: $\sum_{i=1}^{N} \text{bits}(u_{i,t}) \le \mathcal{B}$ for each timestep t. Compute constraint: the perception module q_ϕ has a bounded inference budget, e.g., a maximum number of denoising steps or FLOPs C_{comp}.

Let $c_{i,t}$ denote the aggregated communication context available to agent i (defined in Sec. 4.2). The generative perception module produces a posterior over latent belief:

$$\hat{z}_{i,t} \sim q_\phi(\cdot \mid o_{i,t}, c_{i,t}), \qquad \text{Cost}_{\text{infer}}(q_\phi) \le C_{\text{comp}}. \tag{6}$$

The goal is to maximize the expected discounted return:

$$J(\boldsymbol{\pi}_\theta) \triangleq \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right] \quad \text{s.t.} \quad \mathcal{C} = \{\mathcal{B}, C_{\text{comp}}\}. \tag{7}$$

Under these constraints, perception cannot be uniformly accurate everywhere; the core question becomes *how to allocate limited perceptual capacity to maximize control utility*.

3.2 Theoretical Motivation

Standard model-based MARL pipelines decouple perception from control, training perception to minimize a uniform reconstruction loss (e.g., $\mathbb{E}\|z - \hat{z}\|^2$).

Let $V^\pi(z)$ be the value function of policy π in the latent space, and define the perception error $\varepsilon \triangleq \hat{z} - z$. We denote the (discounted) visitation distribution under π by $\rho^\pi(z)$.

Theorem 1 (Value-Conditioned Reconstruction Principle)**.** *Assume V^π is differentiable and L-smooth on the support of ρ^π, i.e., $\|\nabla V^\pi(z) - \nabla V^\pi(z')\| \le L\|z - z'\|$. Then for any estimator $\hat{z}$,*

$$\begin{aligned} \mathbb{E}_{z\sim\rho^\pi}\Big[\big|V^\pi(\hat{z}) - V^\pi(z)\big|\Big] \le \mathbb{E}\Big[\|\nabla_z V^\pi(z)\| \cdot \|\varepsilon\|\Big] \\ + \frac{L}{2}\,\mathbb{E}\Big[\|\varepsilon\|^2\Big]. \end{aligned} \tag{8}$$

Method	Setup	Parameters			MAPPO		MACE		IPPO		EVA-Gen (Ours)	
		σ	D	C	SR (%)	ER	SR (%)	ER	SR (%)	ER	SR (%)	ER
PassRoom$_{\text{CRB}}$	✓	0.0	-	-	**100.0**$_{\pm 0.0}$	48.9$_{\pm 2.0}$	**100.0**$_{\pm 0.0}$	49.9$_{\pm 2.2}$	96.5$_{\pm 3.5}$	50.8$_{\pm 2.8}$	99.5$_{\pm 1.5}$	**52.4**$_{\pm 2.8}$
PassRoom$_{\text{PFS}}$	✓	0.3	-	-	90.6$_{\pm 2.8}$	44.3$_{\pm 2.8}$	90.0$_{\pm 2.5}$	46.5$_{\pm 2.1}$	85.0$_{\pm 8.5}$	45.2$_{\pm 4.8}$	**96.2**$_{\pm 1.2}$	**49.9**$_{\pm 2.2}$
PassRoom$_{\text{CAT}}$	○	0.0	0.5	-	38.5$_{\pm 4.5}$	16.2$_{\pm 3.5}$	52.3$_{\pm 3.5}$	21.5$_{\pm 2.9}$	42.0$_{\pm 15.5}$	22.8$_{\pm 8.5}$	**88.9**$_{\pm 3.2}$	**46.5**$_{\pm 3.1}$
PassRoom$_{\text{VMD}}$	✓	0.0	-	50	12.8$_{\pm 3.2}$	19.5$_{\pm 3.8}$	25.4$_{\pm 4.4}$	30.1$_{\pm 4.5}$	**78.5**$_{\pm 12.8}$	**72.5**$_{\pm 8.2}$	75.3$_{\pm 3.5}$	40.5$_{\pm 3.4}$
PassRoom$_{\text{GUD}}$	×	0.3	-	50	3.5$_{\pm 1.8}$	11.2$_{\pm 2.4}$	12.6$_{\pm 2.8}$	19.5$_{\pm 3.2}$	8.2$_{\pm 6.5}$	18.5$_{\pm 5.8}$	**43.1**$_{\pm 4.5}$	**27.5**$_{\pm 3.8}$
PassRoom$_{\text{TCB}}$	×	0.3	0.5	50	1.5$_{\pm 1.0}$	5.8$_{\pm 1.6}$	6.4$_{\pm 2.1}$	15.2$_{\pm 2.6}$	2.5$_{\pm 3.2}$	15.2$_{\pm 4.5}$	**21.3**$_{\pm 4.2}$	**18.5**$_{\pm 3.2}$
PassRoom$_{\text{Avg.}}$					41.2$_{\pm 2.2}$	31.0$_{\pm 2.7}$	47.8$_{\pm 2.5}$	30.4$_{\pm 2.9}$	52.1$_{\pm 8.3}$	37.5$_{\pm 5.8}$	**70.7**$_{\pm 2.5}$	**39.2**$_{\pm 3.1}$
SecretRoom$_{\text{CRB}}$	✓	0.0	-	-	**100.0**$_{\pm 0.0}$	47.5$_{\pm 2.2}$	**100.0**$_{\pm 0.0}$	48.3$_{\pm 1.6}$	88.5$_{\pm 7.2}$	48.2$_{\pm 3.8}$	98.9$_{\pm 1.0}$	**50.1**$_{\pm 2.2}$
SecretRoom$_{\text{PFS}}$	✓	0.3	-	-	80.2$_{\pm 3.4}$	41.9$_{\pm 2.2}$	86.5$_{\pm 2.3}$	44.5$_{\pm 2.3}$	68.0$_{\pm 12.5}$	36.5$_{\pm 6.8}$	**92.8**$_{\pm 2.4}$	**47.1**$_{\pm 2.0}$
SecretRoom$_{\text{CAT}}$	○	0.0	0.5	-	33.5$_{\pm 3.9}$	14.0$_{\pm 2.9}$	45.1$_{\pm 4.1}$	19.1$_{\pm 3.3}$	22.5$_{\pm 18.5}$	12.8$_{\pm 10.2}$	**81.9**$_{\pm 2.6}$	**41.9**$_{\pm 2.5}$
SecretRoom$_{\text{VMD}}$	✓	0.0	-	50	10.2$_{\pm 2.4}$	16.9$_{\pm 3.2}$	20.8$_{\pm 3.8}$	26.5$_{\pm 3.9}$	**72.0**$_{\pm 15.5}$	**78.5**$_{\pm 9.8}$	67.5$_{\pm 2.9}$	35.9$_{\pm 2.8}$
SecretRoom$_{\text{GUD}}$	×	0.3	-	50	2.5$_{\pm 1.2}$	9.0$_{\pm 1.8}$	10.2$_{\pm 2.2}$	16.9$_{\pm 2.4}$	3.8$_{\pm 4.8}$	22.5$_{\pm 7.2}$	**36.1**$_{\pm 3.7}$	**23.1**$_{\pm 3.2}$
SecretRoom$_{\text{TCB}}$	×	0.3	0.5	50	0.9$_{\pm 0.6}$	4.6$_{\pm 1.2}$	5.2$_{\pm 1.7}$	11.8$_{\pm 2.2}$	0.8$_{\pm 1.5}$	**18.8**$_{\pm 5.5}$	**15.9**$_{\pm 3.4}$	14.5$_{\pm 2.6}$
SecretRoom$_{\text{Avg.}}$					37.9$_{\pm 1.9}$	22.3$_{\pm 2.2}$	44.6$_{\pm 2.3}$	27.9$_{\pm 2.6}$	42.6$_{\pm 10.0}$	36.2$_{\pm 7.2}$	**65.5**$_{\pm 2.3}$	**35.4**$_{\pm 2.6}$
MultiRoom$_{\text{CRB}}$	✓	0.0	-	-	**100.0**$_{\pm 0.0}$	48.2$_{\pm 2.1}$	**100.0**$_{\pm 0.0}$	49.1$_{\pm 1.8}$	52.5$_{\pm 15.5}$	28.5$_{\pm 8.5}$	99.2$_{\pm 1.1}$	**51.1**$_{\pm 2.5}$
MultiRoom$_{\text{PFS}}$	✓	0.3	-	-	85.4$_{\pm 3.1}$	43.1$_{\pm 2.5}$	88.1$_{\pm 2.4}$	45.5$_{\pm 2.2}$	32.0$_{\pm 18.2}$	17.5$_{\pm 9.8}$	**94.5**$_{\pm 1.8}$	**48.5**$_{\pm 2.1}$
MultiRoom$_{\text{CAT}}$	○	0.0	0.5	-	36.0$_{\pm 4.2}$	15.1$_{\pm 3.2}$	48.7$_{\pm 3.8}$	20.3$_{\pm 3.1}$	8.5$_{\pm 9.5}$	4.8$_{\pm 5.2}$	**85.4**$_{\pm 2.9}$	**44.2**$_{\pm 2.8}$
MultiRoom$_{\text{VMD}}$	✓	0.0	-	50	11.5$_{\pm 2.8}$	18.2$_{\pm 3.5}$	23.1$_{\pm 4.1}$	28.3$_{\pm 4.2}$	45.0$_{\pm 20.5}$	**52.5**$_{\pm 12.5}$	**71.4**$_{\pm 3.2}$	38.2$_{\pm 3.1}$
MultiRoom$_{\text{GUD}}$	×	0.3	-	50	3.0$_{\pm 1.5}$	10.1$_{\pm 2.1}$	11.4$_{\pm 2.5}$	18.2$_{\pm 2.8}$	1.2$_{\pm 2.5}$	8.5$_{\pm 4.8}$	**39.6**$_{\pm 4.1}$	**25.3**$_{\pm 3.5}$
MultiRoom$_{\text{TCB}}$	×	0.3	0.5	50	1.2$_{\pm 0.8}$	5.2$_{\pm 1.4}$	5.8$_{\pm 1.9}$	13.5$_{\pm 2.4}$	0.0$_{\pm 0.0}$	5.2$_{\pm 3.2}$	**18.6**$_{\pm 3.8}$	**16.5**$_{\pm 2.9}$
MultiRoom$_{\text{Avg.}}$					39.5$_{\pm 2.1}$	23.3$_{\pm 2.5}$	46.2$_{\pm 2.4}$	29.1$_{\pm 2.8}$	23.2$_{\pm 11.0}$	19.5$_{\pm 7.3}$	**68.1**$_{\pm 2.5}$	**37.3**$_{\pm 2.8}$

Table 1: GridWorld Performance Comparison: Success Rate (SR) and Episodic Return (ER) across three scenarios.

Note: Setup indicates accessibility (✓: door open, ○: partially open, ×: closed). **CRB** (Clean); **PFS/CAT/VMD** (Single stressor); **GUD** (Gated+σ+C); **TCB** (All factors coupled).

Moreover, for any $\delta_0 > 0$,

$$\|\nabla_z V^\pi(z)\| \cdot \|\varepsilon\| \le \frac{1}{2\delta_0}\|\nabla_z V^\pi(z)\| \cdot \|\varepsilon\|^2 + \frac{\delta_0}{2}\|\nabla_z V^\pi(z)\|, \tag{9}$$

which implies that, under a fixed distortion budget $\mathbb{E}\|\varepsilon\|^2 \le \delta$, minimizing the value-conditioned *quadratic objective*

$$\mathcal{L}_{VCR} \triangleq \mathbb{E}_{z\sim\rho^\pi}\left[\underbrace{\|\nabla_z V^\pi(z)\|}_{\text{value sensitivity}} \cdot \|z - \hat{z}\|^2\right], \tag{10}$$

minimizes a tight upper bound on the expected value degradation in Eq. (8) *up to constants.*

Proof. By L-smoothness, Taylor's theorem gives $V^\pi(\hat{z}) = V^\pi(z) + \nabla V^\pi(z)^\top \varepsilon + R$ with $|R| \le \frac{L}{2}\|\varepsilon\|^2$, which yields Eq. (8) by Cauchy–Schwarz. Eq. (9) follows from the inequality $\|x\| \le \frac{1}{2\delta_0}\|x\|^2 + \frac{\delta_0}{2}$ applied to $x = \varepsilon$ and scaling by $\|\nabla V^\pi(z)\|$. Substituting into Eq. (8) shows that $\mathcal{L}_{\text{VCR}}$ controls the leading term governing value deviation under a quadratic distortion budget. □

Remark 1 Theorem 1 establishes a precise *surrogate necessity*: under finite resources, perception should prioritize regions where the value function is sensitive (large $\|\nabla_z V^\pi(z)\|$). Uniform reconstruction is recovered only in the degenerate case where $\|\nabla_z V^\pi(z)\|$ is constant (i.e., no value heterogeneity).

4 Methodology

Guided by Theorem 1, EVA-Gen Framework closes the perception–value loop via three mechanisms (Fig. 2). We utilize a CTDE-trained centralized critic $V_\omega(z)$ to compute value sensitivities in the latent space.

4.1 Backward Flow via Value-Guided Diffusion

To minimize Eq. (10) under compute constraints, we allocate capacity to value-sensitive latents and implement *Backward Flow* via value-weighted diffusion.

Value-weighted replay distribution: define the value sensitivity and sampling weight

$$g(z) \triangleq \|\nabla_z V_\omega(z)\|, \tag{11}$$

$$w(z) \triangleq \text{clip}(g(z), w_{\min}, w_{\max})^\alpha, \tag{12}$$

where $\alpha \ge 0$ controls aggressiveness and stabilizes training.

Value-weighted replay distribution: given a replay buffer distribution $p_{\mathcal{D}}(z, o, c)$, construct a value-biased distribution

$$p_{\text{val}}(z, o, c) \triangleq \frac{p_{\mathcal{D}}(z, o, c)\, w(z)}{Z}, \tag{13}$$

$$Z \triangleq \mathbb{E}_{(z,o,c)\sim p_{\mathcal{D}}}[w(z)]. \tag{14}$$

Sampling from p_{val} implements the value-conditioning in Eq. (10) while retaining the same model class.

Diffusion training in latent space: we train a conditional diffusion model ϵ_ϕ to model $q_\phi(z \mid o, c)$ by predicting injected noise in the standard denoising objective [Ho *et al.*, 2020; Song *et al.*, 2021b]. Let z_k denote the noised latent at diffusion step k. The Backward Flow objective is

$$\mathcal{L}_{\text{BF}}(\phi) = \mathbb{E}_{\substack{(z,o,c)\sim p_{\text{val}} \\ k\sim U[1,K] \\ \epsilon\sim\mathcal{N}(0,I)}}\left[\left\|\epsilon - \epsilon_\phi(z_k, k, o, c)\right\|_2^2\right]. \tag{15}$$

Scenario	Setup	Parameters			IPPO		MAPPO		MACE		EVA-Gen (Ours)	
		σ	D	C	WR (%)	ER	WR (%)	ER	WR (%)	ER	WR (%)	ER
2m_vs_1z$_{\text{CBC}}$	✓	-	-	-	52.5$_{\pm 8.5}$	542.2$_{\pm 68.1}$	68.5$_{\pm 4.2}$	658.1$_{\pm 42.2}$	72.8$_{\pm 3.8}$	681.3$_{\pm 38.1}$	**76.2**$_{\pm 3.5}$	**702.3**$_{\pm 35.4}$
2m_vs_1z$_{\text{OCC-L}}$	✓	0.1	-	-	48.5$_{\pm 9.2}$	512.1$_{\pm 72.2}$	61.5$_{\pm 5.8}$	612.1$_{\pm 52.6}$	68.2$_{\pm 4.5}$	645.7$_{\pm 45.1}$	**72.8**$_{\pm 4.2}$	**675.2**$_{\pm 42.8}$
2m_vs_1z$_{\text{NOI-L}}$	✓	-	0.1	-	50.2$_{\pm 8.8}$	528.5$_{\pm 70.1}$	63.2$_{\pm 5.5}$	625.7$_{\pm 48.5}$	69.5$_{\pm 4.2}$	655.4$_{\pm 42.1}$	**74.5**$_{\pm 3.8}$	**690.9**$_{\pm 38.1}$
2m_vs_1z$_{\text{DEC-L}}$	✓	-	-	10	22.8$_{\pm 12.5}$	425.2$_{\pm 85.1}$	32.5$_{\pm 8.5}$	485.1$_{\pm 68.7}$	51.6$_{\pm 7.2}$	540.1$_{\pm 58.3}$	**68.5**$_{\pm 5.8}$	**634.4**$_{\pm 52.5}$
2m_vs_1z$_{\text{DPD}}$	○	0.1	0.1	-	42.5$_{\pm 10.5}$	485.6$_{\pm 78.7}$	55.8$_{\pm 6.8}$	578.0$_{\pm 58.1}$	62.5$_{\pm 5.5}$	615.5$_{\pm 52.4}$	**70.2**$_{\pm 4.8}$	**658.1**$_{\pm 48.2}$
2m_vs_1z$_{\text{FULL}}$	×	0.3	0.1	10	38.5$_{\pm 11.2}$	468.4$_{\pm 82.2}$	51.6$_{\pm 7.5}$	558.1$_{\pm 62.6}$	58.8$_{\pm 6.2}$	595.0$_{\pm 55.0}$	**68.5**$_{\pm 5.2}$	**634.1**$_{\pm 52.2}$
2m_vs_1z$_{\text{Avg.}}$					42.5$_{\pm 10.1}$	493.3$_{\pm 75.8}$	55.5$_{\pm 6.4}$	586.1$_{\pm 55.0}$	63.9$_{\pm 5.2}$	621.8$_{\pm 48.3}$	**71.8**$_{\pm 4.5}$	**665.5**$_{\pm 44.5}$
3m$_{\text{CBC}}$	✓	-	-	-	18.5$_{\pm 12.5}$	684.0$_{\pm 95.1}$	36.7$_{\pm 8.5}$	950.0$_{\pm 72.2}$	72.5$_{\pm 4.5}$	1188.9$_{\pm 58.6}$	**74.5**$_{\pm 4.2}$	**1220.2**$_{\pm 52.1}$
3m$_{\text{OCC-L}}$	✓	0.1	-	-	12.5$_{\pm 14.5}$	551.5$_{\pm 102.7}$	28.5$_{\pm 9.5}$	807.2$_{\pm 78.4}$	64.8$_{\pm 5.2}$	1093.1$_{\pm 62.1}$	**70.2**$_{\pm 4.8}$	**1155.9**$_{\pm 58.1}$
3m$_{\text{NOI-L}}$	✓	-	0.1	-	14.2$_{\pm 13.8}$	589.7$_{\pm 98.9}$	30.5$_{\pm 9.2}$	836.0$_{\pm 75.0}$	66.5$_{\pm 4.8}$	1121.2$_{\pm 60.1}$	**72.8**$_{\pm 4.5}$	**1185.5**$_{\pm 55.6}$
3m$_{\text{DEC-L}}$	✓	-	-	10	5.8$_{\pm 16.5}$	399.7$_{\pm 112.8}$	18.5$_{\pm 11.5}$	684.0$_{\pm 88.9}$	48.5$_{\pm 6.8}$	874.0$_{\pm 68.5}$	**62.5**$_{\pm 5.8}$	**1028.0**$_{\pm 65.0}$
3m$_{\text{DPD}}$	○	0.1	0.1	-	4.5$_{\pm 17.2}$	332.0$_{\pm 118.3}$	15.8$_{\pm 12.5}$	589.0$_{\pm 95.6}$	42.5$_{\pm 7.5}$	808.1$_{\pm 72.8}$	**54.2**$_{\pm 6.8}$	**925.8**$_{\pm 72.1}$
3m$_{\text{FULL}}$	×	0.3	0.1	10	2.5$_{\pm 18.5}$	237.0$_{\pm 125.7}$	8.5$_{\pm 14.5}$	427.0$_{\pm 105.1}$	34.3$_{\pm 8.8}$	741.1$_{\pm 78.5}$	**48.5**$_{\pm 7.8}$	**865.0**$_{\pm 82.8}$
3m$_{\text{Avg.}}$					9.7$_{\pm 15.5}$	465.3$_{\pm 108.3}$	23.1$_{\pm 11.0}$	715.5$_{\pm 85.5}$	54.8$_{\pm 6.3}$	970.8$_{\pm 66.3}$	**63.8**$_{\pm 5.6}$	**1063.0**$_{\pm 64.0}$
8m$_{\text{CBC}}$	✓	-	-	-	2.5$_{\pm 14.2}$	185.0$_{\pm 95.0}$	10.5$_{\pm 11.5}$	268.0$_{\pm 85.0}$	15.2$_{\pm 10.8}$	298.0$_{\pm 78.3}$	**17.2**$_{\pm 11.2}$	**318.0**$_{\pm 82.7}$
8m$_{\text{OCC-L}}$	✓	0.1	-	-	1.5$_{\pm 15.5}$	162.3$_{\pm 98.4}$	8.5$_{\pm 12.8}$	245.4$_{\pm 88.7}$	12.8$_{\pm 12.2}$	275.0$_{\pm 85.0}$	**14.5**$_{\pm 13.2}$	**295.0**$_{\pm 92.2}$
8m$_{\text{NOI-L}}$	✓	-	0.1	-	1.8$_{\pm 15.2}$	168.1$_{\pm 95.3}$	9.2$_{\pm 12.5}$	255.4$_{\pm 88.7}$	13.5$_{\pm 11.8}$	282.1$_{\pm 82.2}$	**15.8**$_{\pm 12.8}$	**305.0**$_{\pm 88.4}$
8m$_{\text{DEC-L}}$	✓	-	-	10	0.5$_{\pm 16.5}$	135.8$_{\pm 102.6}$	6.5$_{\pm 13.8}$	218.3$_{\pm 92.3}$	9.8$_{\pm 13.2}$	248.0$_{\pm 88.0}$	**12.5**$_{\pm 14.5}$	**275.5**$_{\pm 95.5}$
8m$_{\text{DPD}}$	○	0.1	0.1	-	0.2$_{\pm 17.2}$	118.1$_{\pm 105.0}$	4.5$_{\pm 14.8}$	195.0$_{\pm 98.1}$	7.2$_{\pm 14.2}$	225.9$_{\pm 95.6}$	**9.5**$_{\pm 15.5}$	**248.4**$_{\pm 102.2}$
8m$_{\text{FULL}}$	×	0.3	0.1	10	0.0$_{\pm 0.0}$	95.5$_{\pm 35.6}$	2.8$_{\pm 15.8}$	168.5$_{\pm 105.1}$	5.5$_{\pm 15.2}$	205.1$_{\pm 102.3}$	**7.8**$_{\pm 16.2}$	**228.5**$_{\pm 108.2}$
8m$_{\text{Avg.}}$					1.1$_{\pm 13.1}$	143.8$_{\pm 88.3}$	7.0$_{\pm 13.5}$	224.8$_{\pm 92.7}$	10.7$_{\pm 12.9}$	255.5$_{\pm 88.3}$	**12.9**$_{\pm 13.9}$	**278.2**$_{\pm 94.5}$

Table 2: SMAC Performance Comparison: Win Rate (WR) and Episode Return (ER) across three maps and six settings.

Note: Setup indicates degradation tier (✓: single mild factor, ○: two-factor coupling, ×: full stress). **CBC** (Clean); **OCC/NOI/DEC** (Single stressor); **DPD** (Dual perception degradation); **FULL** (Max stress).

To satisfy C_{comp}, we use accelerated samplers (e.g., DDIM [Song *et al.*, 2021a]) with $K' \ll K$ denoising steps at execution. Backward Flow ensures these limited steps focus on resolving uncertainty in high-stakes manifolds.

4.2 Collaborative Information Bottleneck (CIB)

Under bandwidth budget $\mathcal{B}$, agents cannot share raw observations. CIB learns compressed messages that retain value-relevant information and align agents on critical factors.

CIB encoder and aggregation (interface to diffusion): each agent encodes a message

$$u_{i,t} \sim q_\psi(\cdot \mid o_{i,t}), \qquad \sum_{i=1}^{N} \text{bits}(u_{i,t}) \le \mathcal{B}. \tag{16}$$

aggregates received neighbor messages into a context vector

$$c_{i,t} \triangleq \text{Agg}\big(\{\tilde{u}_{j,t}\}_{j\in\mathcal{N}(i)\cup\{i\}}\big), \tag{17}$$

which is provided to diffusion as the conditioning input in Eq. (6), Eq. (15). This couples communication to perception.

Information bottleneck objective: let y_t be a value-relevant training signal, e.g., $y_t = V_\omega(z_t)$, an advantage estimate $\hat{A}_t$, or a binary criticality label $\mathbb{I}\{g(z_t) \ge \tau\}$. We seek messages that preserve information about y_t while compressing $o_{i,t}$:

$$\mathcal{J}_{\text{CIB}}(\psi) = I(u_{i,t}; y_t) - \beta I(u_{i,t}; o_{i,t}), \tag{18}$$

where β trades off informativeness and rate. In practice, we optimize a standard variational information bottleneck surrogate [Alemi *et al.*, 2017]:

$$\begin{aligned}\mathcal{L}_{\text{CIB}}(\psi) = \mathbb{E}\Big[-\log p_\eta(y_t \mid u_{i,t})\Big] \\ + \beta\,\text{KL}\Big(q_\psi(u_{i,t} \mid o_{i,t}) \,\Big\|\, p(u)\Big).\end{aligned} \tag{19}$$

where $p(u)$ is a simple prior (e.g., $\mathcal{N}(0, I)$) and the KL term acts as a differentiable proxy for expected bitrate.

4.3 Risk-Sensitive Gradient Rectification

Value-Guided Diffusion biases samples toward high-sensitivity states, which is necessary for perception but can distort policy learning if uncorrected. Meanwhile, sparse critical events risk being overwhelmed by routine transitions.

Proposition 1 (Gradient dilution under value sparsity)**.** *Let $\mathcal{Z}_{crit}$ and $\mathcal{Z}_{rtn}$ be critical and routine sets with $p_{crit} = \mathbb{P}(z \in \mathcal{Z}_{crit})$ and $p_{rtn} = 1 - p_{crit}$. Decompose the expected policy gradient as $G = p_{crit}G_{crit} + p_{rtn}G_{rtn}$. If $p_{rtn}/p_{crit} = \kappa \gg 1$ and G_{crit}, G_{rtn} are comparable in magnitude, then under uniform sampling the critical-state contribution is $O(1/\kappa)$.*

Because $p_{\text{val}}(z, o, c) \propto p_{\mathcal{D}}(z, o, c)\, w(z)$ (Eq. (13)), density ratio required to correct bias is available in closed form:

$$\frac{p_{\mathcal{D}}(z, o, c)}{p_{\text{val}}(z, o, c)} = \frac{Z}{w(z)} \;\propto\; \frac{1}{w(z)}. \tag{20}$$

Thus, we *do not* perform density-ratio estimation; we directly use $1/w(z)$ (with clipping/normalization) as the bias-correction factor.

Diffusion defines a posterior distribution over $\hat{z}_{i,t}$. Quantify epistemic uncertainty by Monte Carlo posterior spread:

$$\begin{aligned}&\{\hat{z}_{i,t}^{(m)}\}_{m=1}^{M} \sim q_\phi(\cdot \mid o_{i,t}, c_{i,t}),\\ &\sigma_{\text{diff},i,t} \triangleq \text{tr}\left(\text{Cov}\big(\{\hat{z}_{i,t}^{(m)}\}_{m=1}^{M}\big)\right).\end{aligned} \tag{21}$$

In practice, M is small (e.g., $M \in \{2, 4\}$) and can be computed only during training or amortized at execution.

Method	Setup	Parameters				MAPPO		CMiMC		ExplabOff		EVA-Gen (Ours)	
		p	σ	λ	ρ	SR (%)	VR_{LS} (%)	SR (%)	VR_{LS} (%)	SR (%)	VR_{LS} (%)	SR (%)	VR_{LS} (%)
$\text{Urban}_{\text{base}}$	✓	5	0.2	1.5	20	$60.6_{\pm 1.5}$	$39.3_{\pm 1.5}$	$79.8_{\pm 3.0}$	$20.1_{\pm 3.0}$	$70.4_{\pm 1.6}$	$29.5_{\pm 1.6}$	$\mathbf{90.8}_{\pm 1.9}$	$\mathbf{9.2}_{\pm 1.9}$
$\text{Urban}_{\text{noise}}$	×	5	0.3	1.5	20	$45.8_{\pm 1.9}$	$54.1_{\pm 1.9}$	$67.8_{\pm 1.7}$	$32.1_{\pm 1.7}$	$59.0_{\pm 2.2}$	$40.9_{\pm 2.2}$	$\mathbf{86.0}_{\pm 2.4}$	$\mathbf{13.9}_{\pm 2.4}$
$\text{Urban}_{\text{loss}}$	✓	10	0.2	1.5	20	$46.1_{\pm 1.3}$	$53.8_{\pm 1.3}$	$68.7_{\pm 2.2}$	$31.2_{\pm 2.2}$	$60.5_{\pm 2.2}$	$39.5_{\pm 2.2}$	$\mathbf{86.6}_{\pm 1.9}$	$\mathbf{13.3}_{\pm 1.9}$
$\text{Urban}_{\text{load}}$	×	5	0.2	2.5	20	$40.3_{\pm 0.9}$	$59.6_{\pm 0.9}$	$62.2_{\pm 3.4}$	$37.7_{\pm 3.4}$	$55.5_{\pm 1.5}$	$44.4_{\pm 1.5}$	$\mathbf{82.1}_{\pm 2.5}$	$\mathbf{17.9}_{\pm 2.5}$
$\text{Urban}_{\text{qos}}$	✓	5	0.2	1.5	35	$43.7_{\pm 1.8}$	$56.2_{\pm 1.8}$	$65.7_{\pm 2.4}$	$34.2_{\pm 2.4}$	$58.8_{\pm 2.4}$	$41.1_{\pm 2.4}$	$\mathbf{84.8}_{\pm 3.0}$	$\mathbf{15.1}_{\pm 3.0}$
$\text{Urban}_{\text{Avg.}}$						$47.3_{\pm 1.5}$	$52.6_{\pm 1.5}$	$68.9_{\pm 2.5}$	$31.0_{\pm 2.5}$	$60.8_{\pm 2.0}$	$39.1_{\pm 2.0}$	$\mathbf{86.0}_{\pm 2.4}$	$\mathbf{13.9}_{\pm 2.4}$
$\text{Emergency}_{\text{base}}$	✓	10	0.2	2.5	25	$38.4_{\pm 0.9}$	$61.5_{\pm 0.9}$	$60.5_{\pm 2.6}$	$39.4_{\pm 2.6}$	$53.0_{\pm 0.7}$	$46.9_{\pm 0.7}$	$\mathbf{81.7}_{\pm 2.6}$	$\mathbf{18.2}_{\pm 2.6}$
$\text{Emergency}_{\text{noise}}$	×	10	0.3	2.5	25	$36.5_{\pm 0.9}$	$63.4_{\pm 0.9}$	$59.4_{\pm 1.4}$	$40.6_{\pm 1.4}$	$49.5_{\pm 1.3}$	$50.4_{\pm 1.3}$	$\mathbf{78.3}_{\pm 1.5}$	$\mathbf{21.6}_{\pm 1.5}$
$\text{Emergency}_{\text{loss}}$	×	15	0.2	2.5	25	$38.5_{\pm 0.8}$	$61.5_{\pm 0.8}$	$57.3_{\pm 1.4}$	$42.6_{\pm 1.4}$	$52.5_{\pm 2.3}$	$47.4_{\pm 2.3}$	$\mathbf{81.6}_{\pm 2.6}$	$\mathbf{18.3}_{\pm 2.6}$
$\text{Emergency}_{\text{burst}}$	×	10	0.2	3.5	25	$31.8_{\pm 1.5}$	$68.1_{\pm 1.5}$	$52.8_{\pm 2.2}$	$47.1_{\pm 2.2}$	$45.6_{\pm 1.1}$	$54.3_{\pm 1.1}$	$\mathbf{76.6}_{\pm 2.5}$	$\mathbf{23.3}_{\pm 2.5}$
$\text{Emergency}_{\text{critical}}$	×	20	0.3	4.0	40	$24.1_{\pm 0.8}$	$75.8_{\pm 0.8}$	$43.0_{\pm 1.6}$	$56.9_{\pm 1.6}$	$36.7_{\pm 1.7}$	$63.2_{\pm 1.7}$	$\mathbf{69.1}_{\pm 1.9}$	$\mathbf{30.9}_{\pm 1.9}$
$\text{Emergency}_{\text{Avg.}}$						$33.8_{\pm 1.0}$	$66.1_{\pm 1.0}$	$54.6_{\pm 1.8}$	$45.3_{\pm 1.8}$	$47.5_{\pm 1.4}$	$52.4_{\pm 1.4}$	$\mathbf{77.5}_{\pm 2.2}$	$\mathbf{22.4}_{\pm 2.2}$
$\text{Dense}_{\text{base}}$	✓	10	0.3	1.5	30	$42.1_{\pm 0.6}$	$57.8_{\pm 0.6}$	$63.5_{\pm 1.3}$	$36.4_{\pm 1.3}$	$57.9_{\pm 2.0}$	$42.4_{\pm 2.0}$	$\mathbf{86.8}_{\pm 3.5}$	$\mathbf{13.1}_{\pm 3.5}$
$\text{Dense}_{\text{noise}}$	×	10	0.4	1.5	30	$39.5_{\pm 1.1}$	$60.4_{\pm 1.1}$	$62.8_{\pm 2.2}$	$37.1_{\pm 2.2}$	$53.9_{\pm 1.3}$	$46.0_{\pm 1.3}$	$\mathbf{85.2}_{\pm 5.1}$	$\mathbf{14.7}_{\pm 5.1}$
$\text{Dense}_{\text{loss}}$	×	15	0.3	1.5	30	$41.9_{\pm 0.8}$	$58.0_{\pm 0.8}$	$63.7_{\pm 2.6}$	$36.2_{\pm 2.6}$	$55.8_{\pm 1.8}$	$44.1_{\pm 1.8}$	$\mathbf{82.9}_{\pm 3.4}$	$\mathbf{17.0}_{\pm 3.4}$
$\text{Dense}_{\text{load}}$	×	10	0.3	2.0	30	$38.5_{\pm 1.2}$	$61.4_{\pm 1.2}$	$58.3_{\pm 2.7}$	$41.6_{\pm 2.7}$	$53.8_{\pm 1.6}$	$46.1_{\pm 1.6}$	$\mathbf{80.7}_{\pm 3.5}$	$\mathbf{19.2}_{\pm 3.5}$
$\text{Dense}_{\text{extreme}}$	×	15	0.5	2.0	40	$32.0_{\pm 1.4}$	$67.9_{\pm 1.4}$	$54.0_{\pm 1.1}$	$45.9_{\pm 1.1}$	$46.0_{\pm 1.6}$	$53.9_{\pm 1.6}$	$\mathbf{78.7}_{\pm 1.8}$	$\mathbf{21.2}_{\pm 1.8}$
$\text{Dense}_{\text{Avg.}}$						$38.8_{\pm 1.0}$	$61.1_{\pm 1.0}$	$60.5_{\pm 2.0}$	$39.4_{\pm 2.0}$	$53.4_{\pm 1.7}$	$46.5_{\pm 1.7}$	$\mathbf{82.8}_{\pm 3.5}$	$\mathbf{17.1}_{\pm 3.5}$

Table 3: UAV-MEC Performance Comparison: Success Rate (SR), Violation Rate of latency-sensitive tasks (VR_{LS}) across three maps. **Note: Setup** indicates workload sparsity (✓: sparse/benign regime; ×: dense/stress regime). **Params** are reported as $p/\sigma/\lambda/\rho$.

We define the rectification weight

$$\rho_t = \underbrace{\text{clip}\left(\frac{1}{w(z_t)}, \rho_{\min}, \rho_{\max}\right)}_{\text{bias correction via Eq. (20)}} \cdot \underbrace{(1 - \tanh(\lambda\, \bar{\sigma}_{\text{diff},t}))}_{\text{uncertainty gating}}, \quad (22)$$

where $\bar{\sigma}_{\text{diff},t}$ aggregates per-agent uncertainties (e.g., mean over i) and $\lambda > 0$ controls risk sensitivity. The rectified actor update is then

$$\nabla_\theta J \approx \mathbb{E}_{\tau \sim \mathcal{D}_{\text{val}}} \left[\rho_t \cdot \nabla_\theta \log \boldsymbol{\pi}_\theta(\mathbf{a}_t \mid \hat{\mathbf{z}}_t)\, \hat{A}_t \right], \quad (23)$$

with an analogous ρ_t-weighted critic regression. The bias-correction restores gradient fidelity under 4.1's value-biased sampling, while the uncertainty gate suppresses unstable updates under highly ambiguous beliefs.

5 Experiments

We conduct comprehensive experiments using the **EVA-Gen** framework across three scenarios characterized by epistemic fragmentation and heterogeneous value landscapes: (i) **MACE-GridWorld** [Jiang *et al.*, 2024] for fundamental mechanism verification, (ii) **SMAC** for complex combat coordination [Samvelyan *et al.*, 2019], and (iii) a custom **UAV-MEC** testbed for system-level validation.

5.1 Baselines

We compare **EVA-Gen** with five baselines: **IPPO** [De Witt *et al.*, 2020], an independent on-policy learner without explicit coordination; **MAPPO** [Yu *et al.*, 2022], a strong CTDE PPO baseline with centralized value estimation and decentralized execution; **MACE** [Jiang *et al.*, 2024], using decentralized novelty sharing for coordinated exploration; **CMiMC** [Li *et al.*, 2024], maximizing mutual information between local and collaborative views for communication-efficient MARL; and **ExplabOff** [Ren *et al.*, 2025], an MI-enhanced collaborative offloading baseline. All methods follow the same training protocol and network scale for fair comparison.

5.2 Experimental Environments

GridWorld: We adopt GridWorld benchmarks, *Pass*, *SecretRoom*, and *MultiRoom*, for sequential critical state reasoning. We introduce an Epistemic-Value Trap via perturbations: (1) Positional Drift (σ) for epistemic fragmentation via stochastic coordinate shifts; (2) Ambiguity (D) for perceptual aliasing using dummy triggers; and (3) Distractors (C) for value heterogeneity via dense low-value rewards.

SMAC: Experiments are conducted on SMAC (*2m-1z*, *3m*, *8m*) under sparse rewards [Jiang *et al.*, 2024]. We augment the standard environment with three stressors to test robustness: (i) Dynamic Occlusion (stochastic feature masking); (ii) Observation Noise (attribute perturbation); and (iii) Value Decoys (dense, low-value targets), which create *value traps* to distract agents from global coordination.

UAV-MEC System Verification: In our custom UAV-MEC simulation, tasks exhibit heterogeneous QoS requirements: a small fraction of high-priority, latency-sensitive tasks (LS, high value) and a majority of best-effort tasks (LT, low value). Agents operate under partial observability due to channel occlusion and incomplete communications.

5.3 Experimental Analysis

GridWorld Results

Table 1 shows Success Rate (SR) and Episodic Return (ER) in scenarios under six configurations of epistemic fragmentation and value heterogeneity. The near-perfect performance

Scenario	EVA-Gen (Full)			w/o BF			w/o CIB			w/o RSR		
	SR/WR (%)	VR_{LS}/ER	AUC (%)	SR/WR (%)	VR_{LS}/ER	AUC (%)	SR/WR (%)	VR_{LS}/ER	AUC (%)	SR/WR (%)	VR_{LS}/ER (%)	AUC (%)
$\text{PassRoom}_{\text{GUD}}$	**43.1**$\pm$ 4.5	**27.5**$\pm$ 3.8	**26.8%**$\pm$ 5.2	35.0$\pm$ 5.7	20.6$\pm$ 4.6	22.5%$\pm$ 5.9	30.9$\pm$ 7.4	17.9$\pm$ 6.2	20.7%$\pm$ 8.7	34.6$\pm$ 5.9	21.9$\pm$ 5.0	22.9%$\pm$ 7.4
$\text{SecretRoom}_{\text{GUD}}$	**36.1**$\pm$ 3.7	**23.1**$\pm$ 3.2	**19.5%**$\pm$ 4.8	26.4$\pm$ 5.1	17.2$\pm$ 3.8	14.6%$\pm$ 6.3	20.8$\pm$ 7.1	12.0$\pm$ 6.0	10.7%$\pm$ 7.4	26.7$\pm$ 5.3	16.6$\pm$ 5.4	12.6%$\pm$ 7.7
$\text{MultiRoom}_{\text{GUD}}$	**39.6**$\pm$ 4.1	**25.3**$\pm$ 3.5	**21.2%**$\pm$ 5.5	29.4$\pm$ 5.9	19.0$\pm$ 4.7	15.8%$\pm$ 7.7	23.4$\pm$ 8.1	11.7$\pm$ 7.0	12.3%$\pm$ 9.3	29.4$\pm$ 7.6	17.6$\pm$ 6.0	13.4%$\pm$ 8.8
2m vs $1z_{\text{DPD}}$	**70.2**$\pm$ 4.8	**658.1**$\pm$ 48.5	**58.5%**$\pm$ 6.2	64.4$\pm$ 6.2	595.5$\pm$ 57	50.7%$\pm$ 7.5	59.4$\pm$ 7.8	578.2$\pm$ 71.3	46.6%$\pm$ 9.5	64.1$\pm$ 5.8	594.6$\pm$ 58.1	51.3%$\pm$ 8.6
$3m_{\text{DPD}}$	**54.2**$\pm$ 6.8	**925.9**$\pm$ 72.3	**38.5%**$\pm$ 8.5	44.6$\pm$ 9.0	795$\pm$ 91	29.1%$\pm$ 10.7	38.0$\pm$ 10.7	707.6$\pm$ 104.6	26.0%$\pm$ 12.0	43.6$\pm$ 10.1	745.3$\pm$ 100	30.3%$\pm$ 11.4
$8m_{\text{DPD}}$	**9.5**$\pm$ 15.5	**248**$\pm$ 102	**6.2%**$\pm$ 12.8	5.3$\pm$ 17.4	145.9$\pm$ 110.8	2.8%$\pm$ 15.5	2.8$\pm$ 18.7	111$\pm$ 129.4	2.4%$\pm$ 16.5	5.3$\pm$ 18.5	142.6$\pm$ 125.3	3.3%$\pm$ 15.8
$\text{Urban}_{\text{noise}}$	**86.1**$\pm$ 2.4	**13.9**$\pm$ 2.4	**75.8%**$\pm$ 3.2	79.4$\pm$ 3.5	20.6$\pm$ 3.5	65.5%$\pm$ 4.3	75.0$\pm$ 4.5	24.9$\pm$ 4.4	57.1%$\pm$ 6.0	73.8$\pm$ 4.0	26.2$\pm$ 3.4	56.3%$\pm$ 4.7
$\text{Emergency}_{\text{noise}}$	**78.4**$\pm$ 1.5	**21.6**$\pm$ 1.5	**68.5%**$\pm$ 2.8	69.2$\pm$ 2.8	30.8$\pm$ 2.9	58.2%$\pm$ 4.1	67.0$\pm$ 4.3	33.0$\pm$ 3.8	55.2%$\pm$ 6.2	61.6$\pm$ 4.4	38.4$\pm$ 4.7	49.8%$\pm$ 5.8
$\text{Dense}_{\text{noise}}$	**85.3**$\pm$ 5.1	**14.7**$\pm$ 5.1	**74.5%**$\pm$ 6.2	76.6$\pm$ 7.1	23.4$\pm$ 6.6	54.8%$\pm$ 7.8	73.7$\pm$ 7.3	26.3$\pm$ 7.8	53.0%$\pm$ 8.8	74.5$\pm$ 6.9	25.5$\pm$ 7.4	51.7%$\pm$ 8.6

Table 4: Combined Ablation Study Results: Impact of Core Components across Benchmarks.
Note: GridWorld: Door=closed, $\sigma = 0$, Dummy=low, Coins=50; **SMAC**: $\sigma = D = 0.1$; **UAV-MEC**: obs.-noise.

of the clean baseline (CRB) confirms task reachability. EVA-Gen excels: in PassRoom, it lifts SR from 52.1% (IPPO) to 70.7%; in MULTIROOM, it improves from 46.2% (MACE) to 68.1%. While value-agnostic baselines show high variance in difficult regimes—indicating *policy divergence*—EVA-Gen maintains tight error margins ($\pm 2 \sim 3\%$). Higher ER indicates that EVA-Gen succeeds efficiently without time penalties. This validates that conditioning perception on downstream value gradients suppresses task-irrelevant noise, resolving perception–value misalignment.

SMAC Results

Table 2 reports WR and ER on three SMAC maps under six stress settings with increasing perceptual degradation and value heterogeneity. EVA-Gen achieves the strongest robustness across maps, improving the average WR on 2m_vs_1z from 63.9% (MACE) to 71.8% and maintaining 68.5% WR under the FULL setting. On the harder 8m map, EVA-Gen obtains the highest average WR (12.9%) and ER (278.2), outperforming MACE (10.7% WR, 255.5 ER). As stress increases, baselines degrade more sharply, whereas EVA-Gen preserves higher WR/ER, indicating more reliable coordination under severe partial observability and value decoys. These results support that value-guided perception mitigates perception–value misalignment in sparse-reward MARL.

UAV-MEC Results

Table 3 presents UAV-MEC performance across Urban, Emergency, and Dense scenarios under varying stress (noise, packet loss, QoS criticality). The results demonstrate a critical joint improvement: EVA-Gen simultaneously increases SR while reducing the LS Task Violation Rate (VR). Specifically, in Urban scenarios, it attains 86.09% SR with only 13.91% VR, substantially outperforming MAPPO and strong offloading baselines like CMiMC. Unlike aggressive strategies that trade safety for marginal success gains, EVA-Gen pushes the Pareto frontier outward, indicating superior risk-sensitive trade-offs under uncertainty. This supports our argument: by aligning reconstruction resources with value and risk, the system prevents sparse high-stakes signals from being diluted by abundant low-value channel fluctuations, ensuring reliability where it matters most.

Ablation Study Results

Table 4 confirms the synergy of EVA-Gen, which consistently outperforms its ablated variants in performance (SR/WR), reliability (VR_{LS}/ER), and efficiency (AUC). **w/o BF:** Removing Backward Flow reduces AUC and success rates, especially under severe uncertainty such as UAV-MEC, showing that value-guided belief reconstruction is crucial for stable convergence. **w/o CIB:** Without CIB, agents transmit value-irrelevant signals, leading to higher constraint violations and weaker coordination under limited communication. **w/o RSR:** Removing RSR sharply increases VR_{LS} in UAV-MEC despite moderate success rates, revealing a *value trap*: agents overfit frequent low-value rewards while neglecting sparse high-priority constraints. This validates RSR's role in preserving reliability under long-tailed value distributions.

6 Conclusion

In this work, we proposed EVA-Gen, a unified decentralized MARL framework for mitigating *Perception–Value Misalignment* under epistemic fragmentation and heterogeneous value landscapes. We established the Value-Conditioned Reconstruction Principle, showing that value degradation induces a perception objective weighted by value sensitivity, and instantiated it through a closed loop of Backward Flow, Collaborative Information Bottleneck, and Risk-Sensitive Gradient Rectification. Our results suggest that resource-constrained agents should prioritize decision-critical uncertainty over uniform world reconstruction.

Acknowledgements

This work was supported in part by the Artificial Intelligence Scientific and Technological Innovation Program of Liaoning Province (No.2023JH26/10100008)

Contribution Statement

YuDuo Zheng and XueFeng Du contributed equally to the formulation, method design, experiments, analysis, and writing of this work. Yanqi Cheng supported the implementation and experimental analysis. Li Yin contributed to validation and revision. Fengqi Li supervised the project, revised the manuscript, and serves as the corresponding author. All authors have read and approved the final manuscript.

References

[Alemi *et al.*, 2017] Alexander A Alemi, Ian Fischer, Joshua V Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations (ICLR)*, 2017.

[Bellemare *et al.*, 2017] Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International conference on machine learning*, pages 449–458. PMLR, 2017.

[Buşoniu *et al.*, 2008] Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[De Witt *et al.*, 2020] Christian Schroeder De Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

[Dibangoye and Buffet, 2018] Jilles S Dibangoye and Olivier Buffet. Learning to act in decentralized partially observable mdps. In *International Conference on Machine Learning (ICML)*, pages 1233–1242. PMLR, 2018.

[Grimm *et al.*, 2020] Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 5541–5552, 2020.

[Hafner *et al.*, 2025] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, pages 1–7, 2025.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[Igl *et al.*, 2018] Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for pomdps. In *International Conference on Machine Learning (ICML)*, pages 2117–2126. PMLR, 2018.

[Janner *et al.*, 2022] Michael Janner, Yilun Du, Joshua B Tenenbaum, and Sergey Levine. Planning with diffusion for flexible behavior synthesis. In *International Conference on Machine Learning (ICML)*, pages 9902–9915. PMLR, 2022.

[Jiang *et al.*, 2024] Haobin Jiang, Ziluo Ding, and Zongqing Lu. Settling decentralized multi-agent coordinated exploration by novelty sharing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17444–17452, 2024.

[Li *et al.*, 2024] Jiawei Li, Yiran Wang, Chen Chen, Peng Liu, and Yi Wu. Cmimc: Cascaded mutual information maximization for cooperative multi-agent reinforcement learning. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 13562–13570, 2024.

[Oliehoek and Amato, 2016] Frans A Oliehoek and Christopher Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

[Rashid *et al.*, 2020] Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.

[Ren *et al.*, 2025] Tao Ren, Zheyuan Hu, Jianwei Niu, and Yiming Yao. Explaboff: Towards explorative and collaborative task offloading via mutual information-enhanced marl. In *IEEE INFOCOM 2025-IEEE Conference on Computer Communications*, pages 1–10. IEEE, 2025.

[Samvelyan *et al.*, 2019] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The StarCraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, 2019.

[Schrittwieser *et al.*, 2020] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

[Song *et al.*, 2021a] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021.

[Song *et al.*, 2021b] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

[Su *et al.*, 2024] Ronshee Chawla Su, Qing Zhou, Yaqi Yang, and Baohe Wang. Collaborative multi-agent heterogeneous multi-armed bandits. In *International Conference on Machine Learning (ICML)*, pages 32819–32848. PMLR, 2024.

[Sukhbaatar *et al.*, 2016] Sainbayar Sukhbaatar, Rob Fergus, et al. Learning multiagent communication with backpropagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pages 2244–2252, 2016.

[Wang *et al.*, 2020] Rundong Wang, Xu He, Runsheng Yu, Wei Qiu, Bo An, and Zinovi Rabinovich. Learning efficient multi-agent communication: An information bottleneck approach. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 9908–9918. PMLR, 2020.

[Wang *et al.*, 2023] Andrew Wang, Andrew C Li, Toryn Q Klassen, Rodrigo Toro Icarte, and Sheila A McIlraith. Learning belief representations for partially observable deep RL. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, pages 35970–35988. PMLR, 2023.

[Yu *et al.*, 2022] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:24611–24624, 2022.

[Zhang *et al.*, 2021] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.