ESCAPING LOCAL VIEWS: DISCOVERING LATENT CONCEPTS FOR INTERPRETABLE MULTI-AGENT REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient cooperation is challenging due to the usual partial observability of each agent in multi-agent reinforcement learning. Recurrent networks encode local interaction histories, but their hidden representations provide limited insight into the information underlying individual decisions. To address these challenges, we propose a novel interpretable framework, called escaping local views (ELV), which introduces semantically structured latent concepts to render policy decisions transparent. Specifically, each agent extracts low-dimensional semantic concepts from its local observation and action-observation trajectory. These concepts are jointly encoded into a contextual latent variable via a variational autoencoder (VAE), which builds a bridge between local views and global semantics. To explicitly model the decision of each agent, we employ a dual-path attention mechanism in which one module estimates the salience of individual concepts relative to the global context, while the other captures higher-order cooperative patterns with pairwise concept interactions. Furthermore, we incorporate a concept prediction module that derives an intrinsic reward from next-concept prediction errors, which incentivizes agents to explore regions of semantic novelty. Experiments in multiple environments verify that ELV not only achieves competitive performance but also explicitly provides how agents reason about their decisions.

1 INTRODUCTION

Cooperative multi-agent reinforcement learning (MARL) has demonstrated strong empirical performance on a range of complex decision-making problems, from multi-robot coordination and large-scale fleet management to autonomous driving (Cao et al., 2012; Liu et al., 2024) and real-time strategy games (Vinyals et al., 2019; Geng et al., 2025). However, the policies learned by modern deep MARL algorithms are typically parameterized by high-capacity neural networks, whose inherently opaque representations pose significant barriers to interpretability and hinder adoption in high-stakes applications (Boggess et al., 2023). In addition, many cooperative MARL problems are naturally modeled as partially observable Markov decision processes (POMDPs), where each agent has access only to its own local observations while the environment's global state remains hidden. To address the resulting non-stationarity and information asymmetry, the centralized training with decentralized execution (CTDE) paradigm has become the de facto standard: agents leverage centralized information during training but act solely based on local observations at execution time. While CTDE mitigates some optimization challenges, the split between centralized training and decentralized execution can fragment local representations and coordination, degrading generalization and destabilizing learned policies (Du et al., 2025; Barde et al., 2024).

A natural response is to equip MARL with interpretability, which largely falls into two categories. The first class provides *post-hoc* explanations for already-trained policies, typically by constructing local approximations of the decision function, e.g., via Shapley-style attributions (Wang et al., 2020), trajectory clustering (Zahavy et al., 2016), or saliency-based visualizations (Ghorbani et al., 2019). While useful, such explanations are inherently approximate and could be computationally expensive (Slack et al., 2021), and they may lead to incorrect explanations due to compounding error (Ghorbani et al., 2019), raising concerns about faithfulness and reliability. A second line aims for *intrinsically interpretable* policies by replacing a black-box agent with a simpler, human-readable

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.

surrogate (e.g., rule lists or decision trees), yielding a global summary of behavior. They generally explain decisions by inspecting inputs and outputs by visualizing intermediate weights (Liu et al., 2025) or credit assignment (Liu et al., 2023; Yang et al., 2020; Wang et al., 2022) through value decomposition. However, their explanations based on observation-level activations may not provide the most intuitive form of explanations for humans.

The quest for intrinsic explanations has led to a renewed interest in concept learning (Koh et al., 2020; Stammer et al., 2022; Havasi et al., 2022). In particular, concept bottleneck models (CBMs) (Koh et al., 2020) factorize prediction through an explicit concept layer: the model first infers a set of human-specified concepts and then maps these concepts to the final output, which yields a global, semantically grounded account of what drives decisions and enables test-time interventions by directly editing incorrect concepts (Ramaswamy et al., 2023). Subsequent studies sharpen this view by (i) analyzing and mitigating concept leakage, where the network bypasses the bottleneck (Kim et al., 2023), and (ii) formalizing interactive intervention mechanisms that improve controllability and reliability of concept edits (Vandenhirtz et al., 2024). These advances motivate a natural question in MARL: can agents discover and share concept-level abstractions that support coordination under partial observability, while retaining the intervention and inspection benefits of CBMs?

We answer this question with **ELV**, a concept-bottleneck framework for cooperative MARL that addresses both interpretability and the local view limitation of CTDE. Each agent extracts a small set of latent concepts from its action–observation history, and uses these concepts as the sole input to value estimation. To align semantics across agents, ELV aggregates the distributed concept set and learns a *variational semantic bridge* that reconstructs the global state during centralized training, thereby shaping the concept space to retain globally coherent information. The resulting latent context serves as a global query that reweights concepts through (i) *first-order* contextual salience and (ii) *second-order* interaction salience, so value estimation reflects not only what matters globally, but also what matters for coordination. Finally, ELV augments exploration with a concept-prediction intrinsic reward, which measures novelty in the learned concept dynamics and is less sensitive to task-irrelevant variation in raw observations.

Our contributions are: 1) We propose **ELV**, a plug-and-play concept-bottleneck module for CTDE MARL that computes each agent's individual value in a structured concept space, with explicit concept presence and a variational semantic bridge to align local concepts across agents. 2) We improve interpretability by **grounding concepts in local semantics** via an auxiliary **observation reconstruction** objective, enabling concept–feature attribution and semantic analysis. 3) We evaluate ELV on the Level-based foraging (LBF) (Papoudakis et al., 2021), StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019), and SMACv2 (Ellis et al., 2023) benchmarks, where it achieves strong performance while learning coherent concepts that support interpretable coordination analysis.

2 PRELIMINARIES

Dec-POMDP. A fully cooperative multi-agent task is typically formulated as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP), consisting of a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{U}, \mathcal{P}, r, O, \Omega, \gamma \rangle$, where $\mathcal{N}$ is a finite set of n agents, $s \in \mathcal{S}$ is the global state of the environment. At each timestep t, every agent $i \in \mathcal{N}$ chooses an action $u_i \in \mathcal{U}$ to formulate a joint action $\boldsymbol{u} = [u_i]_{i=1}^{n} \in \mathcal{U}^n$ by its own observation $o_i \in \Omega$ received from the environment. The environment returns a shared reward with the reward function $r(s, \boldsymbol{u}) : \mathcal{S} \times \mathcal{U}^n \to \mathbb{R}$, and transitions to the next state s' with the transition function $\mathcal{P}(s' \mid s, \boldsymbol{u}) : \mathcal{S} \times \mathcal{U}^n \to \mathcal{S}$. The action-observation history of agent i is denoted as $\tau_i \in \mathcal{T} \equiv (\Omega \times \mathcal{U})^*$ and the joint action-observation history is $\boldsymbol{\tau} \in \mathcal{T}^n$. Agent i learns its own policy $\pi_i(u_i \mid \tau_i) : \mathcal{T} \times \mathcal{U} \to [0, 1]$ conditions on its local action-observation history. The objective in a Dec-POMDP is to find a joint policy $\boldsymbol{\pi} = \langle \pi_1, \ldots, \pi_n \rangle$, which maximizes the expected cumulative discounted reward. The joint action-value function under a joint policy $\boldsymbol{\pi}$ is $Q^{\pi}(s, \boldsymbol{u}) = r(s, \boldsymbol{u}) + \gamma \mathbb{E}_{s'}[R_t \mid s, \boldsymbol{u}]$, where $\gamma \in [0, 1)$ is the discount factor and $R_t = \sum_{t=0}^{\infty} \gamma^t r_t$ is the cumulative return.

Concept Bottleneck Models (CBMs). Concept bottleneck models (CBMs) introduce an explicit intermediate representation to separate concept extraction from prediction. A concept extractor g maps an input representation (e.g., a recurrent history state h_i) to a concept vector $c_i = g(h_i)$, and a predictor f maps concepts to task outputs. In our setting, this yields the action values of agent i

$$Q(\tau_i, u_i) = f(g(h_i)). \tag{1}$$

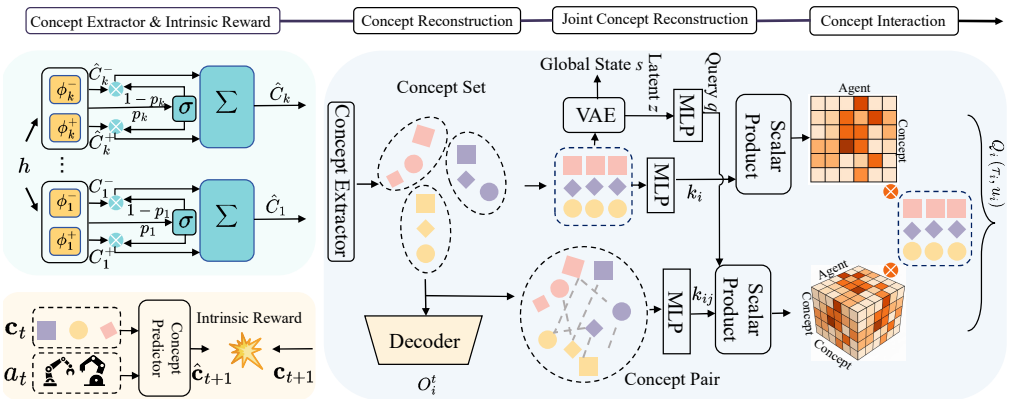


Figure 1: At each timestep t, agent i encodes its local history τ_i^t into a hidden state h_i^t, from which ELV extracts concept embeddings $\{c_{i,k}^t\}_{k=1}^K$ and keys $\{\boldsymbol{k}_{i,k}\}$ via shared projections. The concepts of all agents are aggregated and passed through a VAE to infer a global context $\boldsymbol{z}$, which yields a query vector $\boldsymbol{q}$. First-order weights $\alpha_{i,k}$ are obtained by softmax-normalizing dot-product scores $\boldsymbol{q}^\top \boldsymbol{k}_{i,k}$. To improve coordination, a second-order module scores concept pairs $(c_{i,k}, c_{j,m})$ $(j \neq i)$ against $\boldsymbol{q}$, aggregates over other agents to form $\beta_{i,k}$, and normalizes across k. Finally, ELV reweights concepts with $\alpha_{i,k} + \beta_{i,k}$ and takes a linear head to sum the individual value $Q_i(\tau_i^t, u_i^t)$.

This factorization mediates decisions through structured variables rather than opaque latent features.

Variational Autoencoders (VAEs) (Kingma & Welling, 2014). VAEs define a latent-variable model $p_\theta(s) = \int p_\theta(s \mid z)p(z)\,dz$, where $s \in \mathcal{S}$ denotes observed data and $z \in \mathbb{R}^d$ is a latent variable with prior $p(z) = \mathcal{N}(0, I)$. Since the marginal likelihood $\log p_\theta(s)$ is intractable in general, VAEs introduce an approximate posterior $q_\phi(z \mid s)$ and maximize the evidence lower bound (ELBO):

$$\log p_\theta(s) \geq \mathbb{E}_{q_\phi(z|s)}\left[\log p_\theta(s \mid z)\right] - \mathrm{KL}\left(q_\phi(z \mid s) \| p(z)\right). \tag{2}$$

It balances accurate reconstruction of s with a regularizer that keeps $q_\phi(z \mid s)$ close to the prior. Equivalently, the standard VAE objective is obtained by minimizing the negative ELBO:

$$\mathcal{L}_{\mathrm{VAE}}(s) = -\mathbb{E}_{q_\phi(z|s)}[\log p_\theta(s \mid z)] + \mathrm{KL}(q_\phi(z \mid s) \| p(z)). \tag{3}$$

The introduction of related works is presented in Appendix E.

3 Method

In this section, we detail the formulation of the *Escaping Local Views* (ELV) framework within POMDP, which is designed to induce interpretable decision behaviors from opaque observation streams. As illustrated in *Figure* 1, to handle the partial observability, ELV seeks to reconcile the inherent conflict between local partial observability and the necessity for global coordination. The framework is built upon three pillars: (1) establishing an interpretable semantic bottleneck that extracts disentangled concepts from local histories; (2) inducing a global latent bridge via a VAE to align disparate local views; and (3) incorporating a forward semantic dynamics model that utilizes prediction errors as intrinsic motivation, ensuring exploration is driven by epistemic uncertainty rather than stochastic noise.

3.1 Agent Concepts Network Design

At each timestep t, each agent i takes its local observation o_i^t and the hidden state h_i^{t-1}, and generates a compact set of semantic concepts bottleneck $\mathcal{C}$ that outputs the local action values Q_i. We maintain a recurrent history state $h_i^t = \mathrm{MLP}(h_i^{t-1}, o_i^t, u_i^{t-1})$, where u_i^{t-1} is the last action. This recurrent mechanism allows the agent to retain information from experience, which is helpful for agents in non-stationary and partially observable environments.

The generated history state h_i^t is then projected into K concept slots, where each slot generates separate embeddings for the presence and absence of a semantic concept denoted by $\hat{c}_i^+(\boldsymbol{h})$ and $\hat{c}_i^-(\boldsymbol{h})$

using lightweight MLP heads ϕ_i^+ and ϕ_i^-. To link the learned semantics with the target concept signal, we introduce a differentiable and learnable scoring function $\psi : \mathbb{C}^{2d} \to [0, 1]$ that maps the joint dual-state embedding to a scalar existence probability. Specifically, the probability p_i is computed as:

$$p_i \triangleq \psi(\hat{c}_i^+(\boldsymbol{h}),\, \hat{c}_i^-(\boldsymbol{h}))^T = \sigma(\boldsymbol{w}_h(\hat{c}_i^+(\boldsymbol{h}),\, \hat{c}_i^-(\boldsymbol{h}))^T + \boldsymbol{b}_h), \tag{4}$$

where d denotes the concept dimension, σ is the Sigmoid function, and $\boldsymbol{w}_h, \boldsymbol{b}_h$ are shared across all K concepts to promote consistent scoring behavior and parameter efficiency. Conditioned on p_i, the final concept embedding c_i is obtained as a continuous mixture of the active and inactive embeddings:

$$c_i = p_i\, c_i^+(\boldsymbol{h}) + (1 - p_i)\, c_i^-(\boldsymbol{h}). \tag{5}$$

This formulation implements a soft gating mechanism in which p_i functions as a differentiable switch that interpolates between active and inactive embeddings, yielding semantically distinct representations while preserving end-to-end differentiability for joint optimization.

To compensate for limited local views caused by partial observability, ELV constructs a global semantic bridge that distills the aggregated concept set $\boldsymbol{c} = [c_1; \ldots; c_n]$ across all agents into a compact latent context vector z. We approximate the posterior over this context with a variational distribution parameterized by neural networks $q_\phi(z \mid \boldsymbol{c}) = \mathcal{N}\big(\mu(\boldsymbol{c}),\, \mathrm{diag}(\sigma^2(\boldsymbol{c}))\big)$, and reconstruct the global state s from $\boldsymbol{z}$ via $p_\theta(s \mid \boldsymbol{z})$ during centralized training by minimizing:

$$\mathcal{L}_{vae} = -\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{c})}[\log p_\theta(s \mid \boldsymbol{z})] + \beta\, D_{\mathrm{KL}}(q_\phi(\boldsymbol{z} \mid \boldsymbol{c}) \,\|\, p(\boldsymbol{z})), \tag{6}$$

with a standard normal prior $p(\boldsymbol{z}) = \mathcal{N}(0, I)$. At execution, $\boldsymbol{z}$ is inferred solely from $\boldsymbol{c}$, providing the agent with global semantic context. Appendix C provides an information-theoretic analysis.

First- and Second-Order Concept Importance. Given the inferred global context z, we compute first-order concept importance via dot-product attention between z and each concept embedding:

$$\alpha_k = \frac{\exp((\boldsymbol{w}_z^1 \boldsymbol{z} + \boldsymbol{b}_z)^\top (\boldsymbol{w}_c^1 c_k + \boldsymbol{b}_1))}{\sum_j \exp((\boldsymbol{w}_z^1 \boldsymbol{z} + \boldsymbol{b}_z)^\top (\boldsymbol{w}_c^1 c_j + \boldsymbol{b}_1))}, \tag{7}$$

where $\boldsymbol{w}_z^1$, $\boldsymbol{b}_z$, $\boldsymbol{w}_c^1$, $\boldsymbol{b}_1$ are learnable parameters. The scalar $\alpha_k \in (0, 1)$ quantifies how strongly the latent context activates concept k. However, first-order attention alone may miss cross-concept dependencies critical for cooperation. To complement first-order salience, we introduce an interaction-aware aggregation branch that evaluates each concept jointly with external concepts $\tilde{\boldsymbol{c}}$ under the global context:

$$\beta_k = \frac{\exp\big(\sum_{c_m \in \tilde{\boldsymbol{c}}} (\boldsymbol{w}_z^2 z + \boldsymbol{b}_z)^\top (\boldsymbol{w}_c^2 [c_k \| c_m] + \boldsymbol{b}_2)\big)}{\sum_j \exp\big(\sum_{c_m \in \tilde{\boldsymbol{c}}} (\boldsymbol{w}_z^2 z + \boldsymbol{b}_z)^\top (\boldsymbol{w}_c^2 [c_j \| c_m] + \boldsymbol{b}_2)\big)}, \tag{8}$$

where $[\,\cdot \| \cdot\,]$ denotes concatenation and $\boldsymbol{w}_c^2$, $\boldsymbol{b}_2$ parameterize the learned pairwise compatibility scorer. The resulting weight β_k summarizes the context-conditioned compatibility scores aggregated for concept k. We combine contextual salience with this interaction-aware signal to reweight concept embeddings and compute the action value as:

$$Q_i(\tau_i, u_i) = \boldsymbol{w}_Q\big(\sum\nolimits_k (\alpha_i^k + \beta_i^k)\, c_i^k\big) + \boldsymbol{b}_Q, \tag{9}$$

where $\boldsymbol{w}_Q$ and $\boldsymbol{b}$ are learnable parameters that map the aggregated concepts to Q-values. In this formulation, the combined importance scores $\alpha_{i,k} + \beta_{i,k}$ integrate direct contextual salience with compatibility scores aggregated over external concepts for value estimation. The theoretical motivation for contextual salience and pairwise compatibility aggregation is provided in Appendix A.

3.2 Reconstruction Loss

While the concept bottleneck supports compact, interpretable representations, unconstrained concept embeddings may drift into arbitrary latent codes which lack meaningful observational grounding. To counteract this, we introduce an auxiliary reconstruction objective that explicitly anchors each agent's concept embedding to its local observation space. A lightweight decoder $\mathcal{D}$ predicts the original observation o_i^t from the corresponding concept embedding c_i^t, effectively constraining the learned concepts to retain task-relevant perceptual information and preventing representational collapse:

$$\mathcal{L}_{rec} = \mathbb{E}_t[|\mathcal{D}(c_i^t) - o_i^t|_2^2]. \tag{10}$$

3.3 INTRINSIC REWARD BY PREDICTING CONCEPTS

Efficient exploration remains a central challenge in reinforcement learning, particularly in complex multi-agent environments where rewards are sparse or delayed. To address this, we propose a concept-driven exploration mechanism that generates intrinsic rewards based on semantic dynamics rather than raw perceptual signals. Unlike intrinsic rewards defined directly on observations, which are prone to irrelevant perceptual noise, our method evaluates novelty in a structured concept space to drive exploration toward semantically meaningful transitions.

We instantiate a forward dynamics model f_θ over concept embeddings. At each timestep t, given agent i's concept embedding c_i^t and executed action a_i^t, we predict the next concept embedding:

$$\hat{c}_i^{t+1} = f_\theta(c_i^t, a_i^t), \tag{11}$$

and is trained to minimize the forward prediction loss

$$\mathcal{L}_{pred} = \mathbb{E}_t[|\hat{c}_i^{t+1} - c_i^{t+1}|_2^2], \tag{12}$$

which helps the predictor capture the causal structure of concept evolution. The overall grounding objective $\mathcal{L}_{G_w}$ jointly minimizes the variational semantic loss $\mathcal{L}_{vae}$ and the reconstruction loss $\mathcal{L}_{rec}$:

$$\mathcal{L}_{G_w} = \lambda_1 \mathcal{L}_{vae} + \lambda_2 \mathcal{L}_{rec} + \lambda_3 \mathcal{L}_{pred}, \tag{13}$$

where λ_1 encourages the latent context z to preserve global semantics via state reconstruction, λ_2 ensures that local concepts remain predictive of perceptual inputs, and λ_3 governs the learning of the forward dynamics model, ensuring the concept representation captures the temporal evolution and causal structure of the environment. In information-theoretic terms, minimizing $\mathcal{L}_{\text{rec}}$ serves to maximize a lower bound on the mutual information between the concept embedding c_i^t and the local observation o_i^t. Mutual information quantifies the shared information between two variables, i.e., how much knowing one reduces uncertainty about the other, and is widely used in representation learning to encourage representations to retain informative, task-relevant structure rather than noise. Maximizing this dependency drives concept embeddings to more faithfully encode semantically meaningful features of the observation. A formal derivation showing that reducing the reconstruction loss increases this mutual information lower bound is provided in Appendix B.

Semantic Novelty as Intrinsic Reward. We interpret the prediction error as a measure of semantic novelty: large prediction residuals indicate transitions that are unfamiliar or poorly understood. Formally, we define the intrinsic reward as the next-concept prediction error across all agents:

$$r_t^{\text{int}} = |\hat{\boldsymbol{c}}^{t+1} - \boldsymbol{c}^{t+1}|_2^2. \tag{14}$$

This formulation aligns with prediction-error-based intrinsic motivation, where the prediction error of a learned forward predictor functions as a novelty signal that encourages exploration of under-modeled dynamics. Such curiosity-driven rewards have been widely shown to improve exploration efficiency by pushing agents to sample transitions with high prediction error, especially under sparse extrinsic feedback. By deriving intrinsic rewards in the learned concept space rather than the raw observation space, the exploration signal becomes robust to irrelevant perceptual variability and remains focused on semantically meaningful transitions that are aligned with task structure.

To balance exploration and exploitation, we anneal the intrinsic component over the course of training $r_t = r_t^{\text{env}} + \beta_t \cdot r_t^{\text{int}}$, with $\beta_t = \beta_0 \cdot \exp(-\lambda t)$, where β_0 and λ control the initial scale and decay rate of the intrinsic signal, respectively. This schedule biases early learning toward structured semantic exploration when the model's uncertainty is high and progressively shifts emphasis to maximizing extrinsic task rewards as the dynamics become better learned.

3.4 OVERALL LEARNING OBJECTIVE

Under CTDE, we derive each agent's individual value from its concept representations and aggregate them using a value factorization network (e.g., QMIX (Rashid et al., 2020) or VDN (Sunehag et al., 2018)) to form the joint action-value function Q_{tot}. We optimize the value estimator by minimizing squared temporal-difference (TD) error over samples from a replay buffer $\mathcal{B}$:

$$\mathcal{L}_{TD}(\theta) = \mathbb{E}_{(\tau,u,r,\tau')\sim\mathcal{B}}[(y - Q_{\text{tot}}(\boldsymbol{\tau}, \boldsymbol{u}; \theta))^2], \tag{15}$$

where $y = r + \gamma \max_{\boldsymbol{u}'} Q_{\text{tot}}(\boldsymbol{\tau}', \boldsymbol{u}'; \theta^-)$ is the one-step Bellman target computed with the target network parameters θ^- copied from θ periodically. The overall loss jointly optimizes value estimation and semantic representation learning as

$$\mathcal{L}(\theta, w) = \mathcal{L}_{TD} + \mathcal{L}_{G_w}. \tag{16}$$

This joint optimization aligns accurate value estimation with semantically informed concept learning, ensuring that the learned representations meaningfully support coordinated decision making. We summarize the pseudo-code of our training procedure in Appendix D.

4 Experiments

We evaluate ELV on three challenging benchmarks: SMAC, SMACv2, and Level-Based Foraging (LBF). Baselines include (i) classic value decomposition methods (VDN, QMIX, QPLEX (Wang et al., 2021), QTRAN (Son et al., 2019), CDS (Li et al., 2021)) and (ii) recent CTDE approaches that strengthen per-agent value estimation and credit assignment via architectural or objective-level enhancements (NA^{2}Q (Liu et al., 2023), ReBorn (Qin et al., 2024), MA^{2}E (Kang et al., 2025), SHAQ (Wang et al., 2022)). Implementation details and benchmark configurations are provided in Appendix F. All performance curves report the mean $\pm$ standard deviation over five random seeds.

4.1 Performance on LBF

We first evaluate ELV on two constructed LBF settings. We instantiate ELV on top of the QMIX value-factorization backbone—keeping the QMIX mixing network unchanged and replacing only the individual value module with ELV. As shown in *Figure* 2, ELV achieves better test returns with competitive sample efficiency and stable learning dynamics. VDN exhibits delayed progress, which may be consistent with its limited capacity to represent complex joint action-values under coupled credit assignment. CDS plateaus at a substantially lower return, suggesting that encouraging diversity at the individual level may come at the expense of coherent team strategies in tightly coupled coordination. ReBorn also converges late, indicating inefficient exploration in the expanded joint strategy space. Compared with QMIX, ELV yields a clear performance gain while keeping the same mixing network fixed. This indicates that the improvement primarily comes from the individual-value architecture: replacing the GRU-based utility with a concept-based representation of ELV better mitigates partial observability and leads to more consistent coordination.

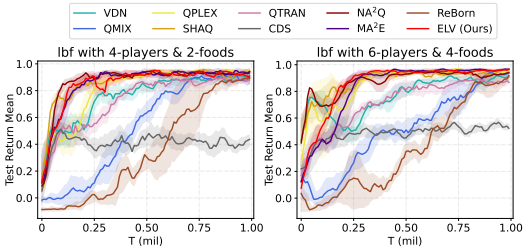


Figure 2: Performance comparison on two tasks of LBF.

4.2 Performance on SMAC

For the SMAC benchmark, we evaluate ELV on a diverse set of maps spanning easy, hard, and super-hard scenarios. As shown in *Figure* 3, ELV achieves consistently stronger win rates, with the clearest margins on super-hard tasks where partial observability and long-horizon coordination are most limiting. For example, on *3s5z_vs_3s6z* and *6h_vs_8z*, ELV continues improving after QMIX or VDN largely plateau and stabilizes at a high win rate, matching the visual trend that ELV both converges faster and sustains better asymptotic performance in tightly coupled fights. These patterns are consistent with known limitations of value factorization baselines: VDN's additive form and QMIX's monotonic mixing can express the expressivity of Q_{tot} in highly coupled settings, often leading to early saturation on harder maps. QTRAN relaxes these constraints but depends on stricter optimization objectives that can be brittle at scale, which aligns with its weaker performance in the challenging regime. CDS promotes behavioral diversity, but on maps requiring precise team-level coordination, this does not necessarily translate into coherent joint strategies and can plateau below top methods. While MA^{2}E outperforms most baselines, it still trails ELV, with the gap most pronounced on the super-hard tasks. It suggests that each agent has the ability to infer the information of other agents from partial observations. Compared to VDN and QMIX, ELV yields larger gains, which we

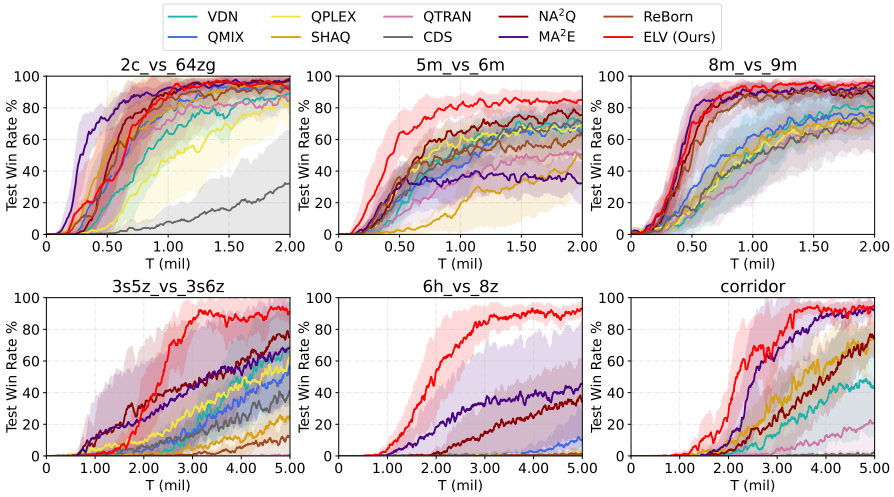


Figure 3: Performance comparison on hard and super-hard scenarios.

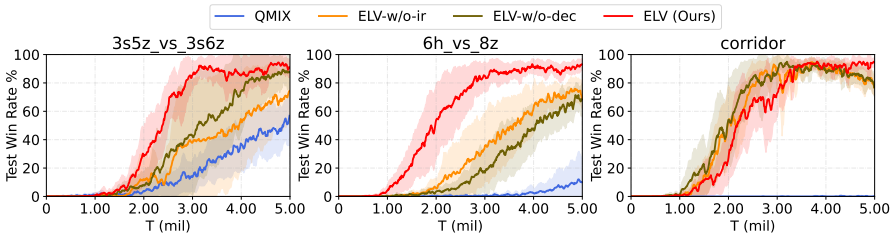


Figure 4: Ablation of intrinsic reward and concept decoder for ELV.

attribute to its semantically structured individual-value estimator: the concept bottleneck together with the variational semantic bridge mitigates information fragmentation under partial observability, enabling more consistent coordination than a GRU-based individual value function while keeping the same mixing network. More results of plugging ELV into VDN and QMIX backbones are reported in Appendix G.1 and G.2. Additional experiments on SMACv2 are provided in Appendix G.3.

4.3 Ablation Study

We perform ablations on the three core components of ELV—concept-dynamics intrinsic reward (ir), reconstruction-based grounding (dec), and concept-level interactions (inter)—by evaluating the corresponding variants without each component (ELV-w/o-ir, ELV-w/o-dec, and ELV-w/o-inter).

As shown in *Figure* 4, when removing the concept-dynamics intrinsic reward causes a sharp drop on *3s5z_vs_3s6z*, ELV-w/o-ir gets stuck in poor local solutions, whereas the full model exceeds 90% win rate. A similar effect appears on *Corridor*, where dropping ir slows learning and lowers final

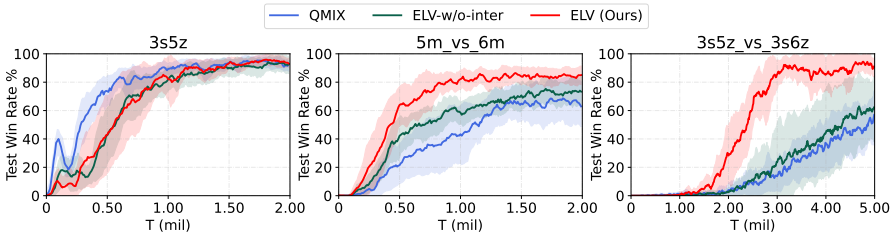


Figure 5: Ablation of the concept interaction module for ELV.

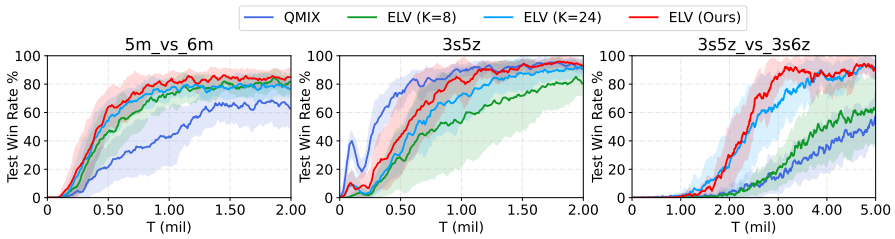


Figure 6: Ablation study with different numbers of concepts for ELV.

performance, supporting the role of concept-space prediction error as an effective exploration signal under sparse or misleading rewards. In contrast, the reconstruction grounding mainly matters under heterogeneity on *6h_vs_8z*. ELV without dec degrades the most, which suggests that grounding stabilizes concept semantics across diverse unit types, while on the navigation-heavy *Corridor*, its impact is minor. To isolate the effect of interaction from task difficulty, we compare the interaction-removed variant with the full model as shown in *Figure* 5. The benefit is that the interaction module yields only minor gains on the simpler *3s5z* map, but the gap grows markedly on the super-hard *3s5z_vs_3s6z*. This pattern suggests that explicit concept-level coordination becomes more important as the joint strategy space expands and tighter cooperation is required.

We finally study how concept capacity interacts with scalability by varying the number of concepts $k \in 8, 16, 24$ across scenarios. As shown in *Figure* 6, we can find that higher concept capacity is required to maintain both final performance and sample efficiency as the number of agents grows. The compact setting ($k = 8$) forms a clear bottleneck on complex maps such as *3s5z_vs_3s6z*, where performance drops sharply, whereas the larger setting ($k = 24$) shows better scalability: it brings little benefit on small maps but yields markedly more robust returns in high-complexity environments. These results indicate that a sufficient number of concept slots is crucial for representing the rich joint strategies needed in large-scale multi-agent coordination.

4.4 Interpretability

Concept Space Visualization. To examine the structure of the learned concepts, we visualize their embeddings with t-SNE (Maaten & Hinton, 2008) in *Figure* 7. The embeddings form well-separated, compact clusters (colors), indicating that ELV disentangles the state space into distinct semantic factors. Moreover, points from different agents (shapes) are well-mixed within each cluster, suggesting that the concepts are agent-invariant and can be shared across the team.

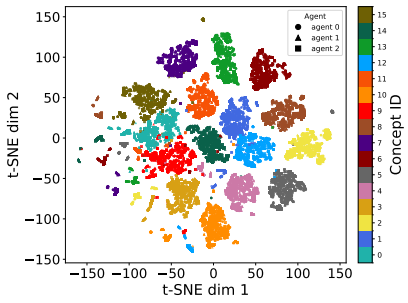


Figure 7: The t-SNE visualization of learned concept embeddings.

Grounding Concepts via Gradient-based Attribution. To further clarify concept semantics, we adopt a gradient-based attribution scheme at evaluation time to relate each concept to structured observation modalities (e.g., movement directions, ally/enemy states, self attributes). Concretely, we compute the attribution of concept c_i to observation modality f_j, which yields a concept–feature attribution matrix $G \in \mathbb{R}^{C\times F}$ with entries $G_{ij} := \|\frac{\partial \,\mathrm{Recon}_{f_j}}{\partial c_i}\|_1$. It measures how sensitive the reconstruction of feature f_j is to perturbations in concept c_i and thus serves as a proxy for their semantic influence.

To obtain a human-readable assignment, we match each observation modality to one concept by maximizing $\sum_{j=1}^{F} G_{\pi(j),j}$ over injective mappings $\pi : \{1, \ldots, F\} \to \{1, \ldots, C\}$. *Figure* 8 (right) illustrates the resulting heatmap. Clear high-intensity blocks suggest that concepts align with distinct semantic groups (e.g., specific allies or enemies) rather than entangled representations; Appendix H provides explicit assignments, test-time interventions, and temporal tracking.

Qualitative Case Study. To quantify how ELV models cooperative coupling between agents, we define an *agent dependency strength* $\mathcal{D}_{ij}$ based on the learned concept–interaction scores. For

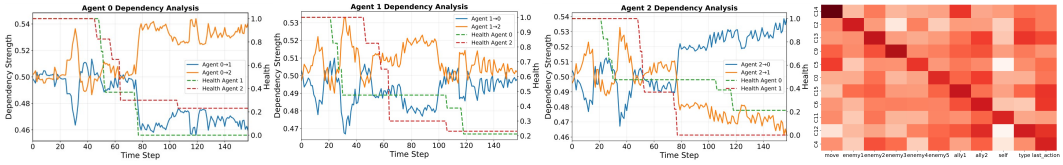
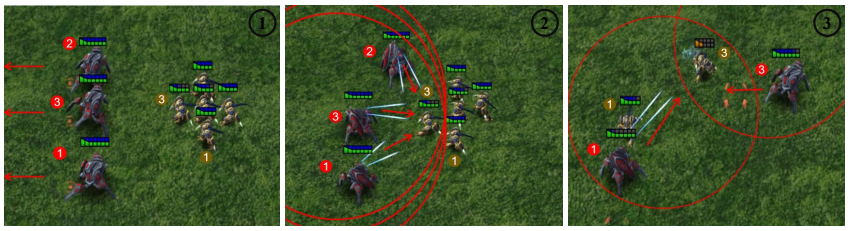


Figure 8: Agent dependency strength alongside health trajectories (left), and the concept–feature attribution heatmap (right).



	Agent 1	Agent 2	Agent 3
Agent 1	\	0.502	0.498
Agent 2	0.499	\	0.501
Agent 3	0.497	0.503	\

	Agent 1	Agent 2	Agent 3
Agent 1	\	0.505	0.495
Agent 2	0.494	\	0.505
Agent 3	0.493	0.507	\

	Agent 1	Agent 2	Agent 3
Agent 1	\	0.459	0.541
Agent 2	0.511	\	0.489
Agent 3	0.538	0.462	\

Figure 9: Visualization of agent collaboration dynamics.

each pair of agents $i \neq j$, we first aggregate interaction weights over all pairs of latent concepts $\mathcal{M}_{ij} = \sum\sum m(c_i^{k_1}, c_j^{k_2})$, $m(c_i^{k_1}, c_j^{k_2}) = (\boldsymbol{w}_z \boldsymbol{z} + \boldsymbol{b}_z)^\top (\boldsymbol{w}_2 [c_i^{k_1} \,\|\, c_j^{k_2}] + \boldsymbol{b}_2)$, and then normalize via a row-wise softmax to get a relative dependency distribution $\mathcal{D}_{ij} = \frac{\exp(\mathcal{M}_{ij})}{\sum_j \exp(\mathcal{M}_{ij})}$. Here, $\mathcal{D}_{ij}$ measures how agent i depends on agent j as it aggregates concept-level interactions.

Figure 9 illustrates three representative phases from a SMAC rollout. At the beginning of the episode, concept activations are relatively uniform, and no specific tactical target stands out. Accordingly, dependency scores between agents remain close to a balanced split (e.g., Agent 1 allocates its attention to teammates roughly $0.502/0.498$), suggesting that only weak mutual coupling is required for simple coordinated retreat. In the middle phase, an abrupt increase in the activation of a particular concept (Concept 9, associated in this run with a specific enemy unit) coincides with the onset of a coordinated counter-attack. Although the agents move into a tightly clustered formation, the dependency matrix stays relatively stable, indicating that once a common tactical focus emerges, ELV maintains coherent group behavior without needing large shifts in pairwise attention. In the final phase, agents split spatially: Agent 3 chases 'Enemy 3' while Agent 1 provides support from a distance. Here, we observe a clear dependency asymmetry: the dependency of Agent 1 on Agent 3 rises to 0.541, while its dependency on nearby agents decreases. This pattern shows that ELV prioritizes concept-level interaction (who engages the key target) over mere spatial proximity, enabling consistent joint tactics under physical dispersion. Tracking $\mathcal{D}_{ij}$ over time further reveals that dependencies on a given teammate tend to decrease as that agent's health drops, consistent with *Figure* 8, suggesting that ELV gradually down-weights coordination with vulnerable agents to stabilize the overall policy.

5 Conclusion

We presented an interpretable cooperative MARL framework with a shared concept bottleneck. ELV decomposes local observations and histories into semantically structured latent concepts and aligns them via a variational semantic bridge, thereby mitigating partial observability and improving global coordination. Dual-path attention captures context-dependent salience and higher-order concept interactions, while reconstruction-based grounding and concept-dynamics intrinsic rewards regularize representations and promote exploration. Across a range of MARL benchmarks, ELV consistently attains competitive or superior performance relative to strong baselines and yields concept-level explanations of coordination strategies. By coupling value estimation with explicit semantic structure, ELV advances transparent, data-efficient multi-agent systems in complex domains.

AI USE STATEMENT

In preparing this manuscript, we used generative AI tools to improve the readability and concision of the text and to assist with LaTeX formatting checks. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

REFERENCES

Paul Barde, Jakob Foerster, Derek Nowrouzezahrai, and Amy Zhang. A model-based solution to the offline multi-agent reinforcement learning coordination problem. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, pp. 141–150, 2024.

Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. In *Advances in Neural Information Processing Systems*, pp. 2494–2504, 2018.

Kayla Boggess, Sarit Kraus, and Lu Feng. Explainable multi-agent reinforcement learning for temporal queries. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 55–63, 2023.

Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. In *Proceedings of the International Conference on Learning Representations*, 2019.

Yongcan Cao, Wenwu Yu, Wei Ren, and Guanrong Chen. An overview of recent progress in the study of distributed multi-agent coordination. *IEEE Transactions on Industrial Informatics*, 9(1): 427–438, 2012.

Eric Chen, Zhang-Wei Hong, Joni Pajarinen, and Pulkit Agrawal. Redeeming intrinsic rewards via constrained optimization. In *Advances in neural information processing systems*, pp. 4996–5008, 2022.

Quentin Delfosse, Sebastian Sztwiertnia, Mark Rothermel, Wolfgang Stammer, and Kristian Kersting. Interpretable concept bottlenecks to align reinforcement learning agents. In *Advances in Neural Information Processing Systems*, pp. 66826–66855, 2024.

Wei Du, Shifei Ding, Wei Guo, Yuqing Sun, Guoxian Yu, and Lizhen Cui. Multi-agent communication with information preserving graph contrastive learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 64–71, 2025.

Benjamin Ellis, Jonathan Cook, Skander Moalla, Mikayel Samvelyan, Mingfei Sun, Anuj Mahajan, Jakob Foerster, and Shimon Whiteson. Smacv2: An improved benchmark for cooperative multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 37567–37593, 2023.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems*, pp. 21400–21413, 2022.

Minghong Geng, Shubham Pateria, Budhitama Subagdja, Lin Li, Xin Zhao, and Ah-Hwee Tan. L2m2: A hierarchical framework integrating large language model and multi-agent reinforcement learning. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 99–107, 2025.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 3681–3688, 2019.

Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *Proceedings of the International Conference on Machine learning*, pp. 1792–1801, 2018.

Marton Havasi, Sonali Parbhoo, and Finale Doshi-Velez. Addressing leakage in concept bottleneck models. In *Advances in Neural Information Processing Systems*, pp. 23386–23397, 2022.

Sehyeok Kang, Yongsik Lee, Gahee Kim, Song Chong, and Se-Young Yun. MA^{2}E: Addressing partial observability in multi-agent reinforcement learning with masked auto-encoder. In *Proceedings of the International Conference on Learning Representations*, 2025.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, pp. 16521–16540, 2023.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the International Conference on Machine Learning*, pp. 5338–5348, 2020.

Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 3991–4002, 2021.

Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. Reasoning multi-agent behavioral topology for interactive autonomous driving. In *Advances in Neural Information Processing Systems*, pp. 92605–92637, 2024.

Zichuan Liu, Yuanyang Zhu, and Chunlin Chen. NA^{2}Q: Neural attention additive model for interpretable multi-agent Q-learning. In *Proceedings of the International Conference on Machine Learning*, pp. 22539–22558, 2023.

Zichuan Liu, Yuanyang Zhu, Zhi Wang, Yang Gao, and Chunlin Chen. Mixrts: Toward interpretable multi-agent reinforcement learning via mixing recurrent soft decision trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):4090–4107, 2025.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

Augustine Mavor-Parker, Kimberly Young, Caswell Barry, and Lewis Griffin. How to stay curious while avoiding noisy tvs using aleatoric uncertainty estimation. In *International conference on machine learning*, pp. 15220–15240, 2022.

Alexander McCaffrey, Eduardo Alonso, and Esther Mondragon. Predictive improvement through latent space optimisation. In *Proceedings of International Conference on Autonomous Agents and Multiagent Systems*, pp. 2669–2671, 2025.

Tuomas Oikarinen, Subhro Das, Lam Nguyen, and Lily Weng. Label-free concept bottleneck models. In *Proceedings of the International Conference on Learning Representations*, 2023.

Yiyuan Pan, Zhe Liu, and Hesheng Wang. Wonder wins ways: Curiosity-driven exploration through multi-agent contextual calibration. In *Advances in Neural Information Processing Systems*, pp. 116109–116137, 2025.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V. Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the International Conference on Machine Learning*, pp. 2778–2787, 2017.

Haoyuan Qin, Chennan Ma, Deng Deng, Zhengzhu Liu, Songzhu Mei, Xinwang Liu, Cheng Wang, and Siqi Shen. The dormant neuron phenomenon in multi-agent reinforcement learning value factorization. In *Advances in Neural Information Processing Systems*, pp. 35727–35759, 2024.

Vikram V Ramaswamy, Sunnie SY Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10932–10941, 2023.

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.

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 International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2186–2188, 2019.

Dylan Slack, Anna Hilgard, Sameer Singh, and Himabindu Lakkaraju. Reliable post hoc explanations: Modeling uncertainty in explainability. In *Advances in Neural Information Processing Systems*, pp. 9391–9404, 2021.

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 *Proceedings of the International Conference on Machine Learning*, pp. 5887–5896, 2019.

Wolfgang Stammer, Marius Memmel, Patrick Schramowski, and Kristian Kersting. Interactive disentanglement: Learning concepts by interacting with their prototype representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10317–10328, 2022.

Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems*, pp. 2085–2087, 2018.

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia Vogt. Stochastic concept bottleneck models. In *Advances in Neural Information Processing Systems*, pp. 51787–51810, 2024.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *nature*, 575(7782):350–354, 2019.

Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. In *International Conference on Learning Representations*, 2021.

Jianhong Wang, Yuan Zhang, Tae-Kyun Kim, and Yunjie Gu. Shapley q-value: A local reward approach to solve global reward games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7285–7292, 2020.

Jianhong Wang, Yuan Zhang, Yunjie Gu, and Tae-Kyun Kim. Shaq: Incorporating shapley value theory into multi-agent q-learning. In *Advances in Neural Information Processing Systems*, pp. 5941–5954, 2022.

Xuefei Wu, Yuanyang Zhu, Caihua Chen, and Chunlin Chen. Gcm: Interpretable multiagent reinforcement learning via graph cooperation modeling. *IEEE Transactions on Neural Networks and Learning Systems*, 2025. doi: 10.1109/TNNLS.2025.3626000.

Yaodong Yang, Jianye Hao, Ben Liao, Kun Shao, Guangyong Chen, Wulong Liu, and Hongyao Tang. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*, 2020.

Changmin Yu, Neil Burgess, Maneesh Sahani, and Samuel J Gershman. Successor-predecessor intrinsic exploration. In *Advances in neural information processing systems*, pp. 73021–73038, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *Proceedings of the International Conference on Learning Representations*, 2023.

Tom Zahavy, Nir Ben-Zrihem, and Shie Mannor. Graying the black box: Understanding dqns. In *Proceedings of the International Conference on Machine Learning*, pp. 1899–1908, 2016.

A THEORETICAL JUSTIFICATION FOR HIERARCHICAL CONCEPT MODELING

In this section, we provide theoretical grounding for explicitly modeling first-order concept importance (individual semantics) and second-order concept interactions (pairwise synergies) within the ELV framework. Our justification is based on the functional ANOVA (Hoeffding-style) decomposition together with a second-order truncation as an inductive bias. This perspective clarifies that our architecture is not merely heuristic: it corresponds to a structured approximation of a general value (or policy) function defined over a learned concept space.

A.1 FUNCTIONAL ANOVA–STYLE DECOMPOSITION

Consider the learned action-value function as a multivariate function defined on the joint concept variables. Let $\mathbf{C} = (C_1, \ldots, C_M)$ denote the set of all concepts extracted across agents (under the data distribution induced by the environment and the current policy), and let $f(\mathbf{C}, a)$ represent the target function (e.g., a per-agent Q-function or a utility component) conditioned on action a.

A classical result in functional ANOVA (often attributed to Hoeffding) states that, for $f(\cdot, a) \in L_2(\mathbb{P}_{\mathbf{C}})$, one can express f as a sum of components of increasing order,

$$f(\mathbf{C}, a) = f_0(a) + \sum_{i=1}^{M} f_i(C_i, a) + \sum_{1 \le i < j \le M} f_{ij}(C_i, C_j, a) + \cdots, \tag{17}$$

where $f_0(a)$ is a global baseline and f_i, f_{ij}, etc., capture effects attributable to individual concepts and their interactions, respectively. To make these components *identifiable*, a common choice is to impose centering (purification) constraints, e.g.,

$$\mathbb{E}[f_i(C_i, a)] = 0, \qquad \mathbb{E}[f_{ij}(C_i, C_j, a) \mid C_i] = 0, \qquad \mathbb{E}[f_{ij}(C_i, C_j, a) \mid C_j] = 0, \tag{18}$$

with expectations taken under $\mathbb{P}_{\mathbf{C}}$. Under additional conditions (e.g., a product measure over inputs), these components are orthogonal in L_2; in general, we adopt this decomposition as a modeling lens and structural prior that separates main semantic effects from pairwise semantic synergies.

Within this view, $f_i(C_i, a)$ corresponds to the first-order (main) effect of concept C_i on the decision, aligning with our first-order attention branch. The term $f_{ij}(C_i, C_j, a)$ corresponds to the second-order interaction effect, capturing the residual dependence on (C_i, C_j) beyond what is explained by the individual main effects, aligning with our interaction branch.

A.2 SECOND-ORDER TRUNCATION AS AN INDUCTIVE BIAS

While the full expansion includes higher-order interactions up to order M, explicitly modeling all terms is statistically and computationally prohibitive: the number of possible components grows exponentially with M (on the order of 2^M), and higher-order components are typically harder to estimate reliably from finite data.

ELV therefore truncates the expansion at second order and approximates f by

$$f(\mathbf{C}, a) \approx f_0(a) + \sum_{i=1}^{M} f_i(C_i, a) + \sum_{1 \le i < j \le M} f_{ij}(C_i, C_j, a). \tag{19}$$

This choice should be interpreted as a deliberate inductive bias: it enforces a transparent hierarchy in which (i) individual concepts provide interpretable semantic evidence for an action, and (ii) pairwise interactions capture salient synergies that cannot be expressed by purely additive effects. Importantly, we do **not** claim that higher-order interactions are absent in principle; rather, the second-order truncation provides a favorable trade-off between expressivity, robustness, and interpretability, which we validate empirically. This is particularly appropriate for coordination patterns that are naturally pairwise (e.g., complementary roles or two-factor conditions that jointly trigger a behavior).

A.3 CONNECTION TO COORDINATION GRAPHS

Our decomposition is also conceptually aligned with coordination-graph-based factorization in cooperative MARL, where a joint utility function is expressed as a sum of unary and pairwise

potentials over graph nodes and edges. ELV instantiates an analogous unary + pairwise structure, but crucially applies it in the latent semantic (concept) space rather than directly over agent identities. As a result, ELV can capture semantic interactions (e.g., "low health" interacting with "enemy proximity") regardless of which agent expresses these concepts, supporting improved interpretability while retaining sufficient capacity for complex decision-making.

B Theoretical Justification for Concept-based Reconstruction

In this section, we provide the theoretical justification for the auxiliary reconstruction objective. We demonstrate that minimizing the reconstruction loss is equivalent to maximizing a variational lower bound of the Mutual Information (MI) between the local observation and the extracted concept, under the empirical distribution induced by the current policy.

B.1 Problem Formulation

Let O be the continuous random variable representing the local observation, and C be the continuous random variable for the concept representation. The concept extractor defines a (possibly stochastic) conditional distribution $p_\phi(c|o)$ (e.g., by injecting noise or outputting a parametric distribution), and we aim to maximize the Mutual Information $I(O;C)$:

$$I(O;C) = H(O) - H(O|C), \tag{20}$$

where $H(\cdot)$ denotes the differential entropy.

B.2 Addressing Non-Stationarity in RL

In Reinforcement Learning, the distribution of observations $p_\pi(o)$ is non-stationary as it depends on the evolving policy π. However, at any given training step t, we optimize the representation with respect to the empirical distribution $p_{\text{emp}}(o)$ (e.g., sampled from a replay buffer or a fixed minibatch). Conditioned on this empirical distribution, the marginal entropy term $H(O)$ can be treated as a constant with respect to the representation parameters within each optimization step. Our goal thus simplifies to minimizing the conditional differential entropy $H(O|C)$.

B.3 Derivation of the Lower Bound

The conditional differential entropy is defined as:

$$H(O|C) = -\mathbb{E}_{p_{\text{emp}}(o,c)}[\log p_{\text{emp}}(o|c)], \tag{21}$$

where $p_{\text{emp}}(o,c) = p_{\text{emp}}(o)p_\phi(c|o)$.

Here, $p_{\text{emp}}(o|c)$ denotes the conditional distribution induced by the empirical joint distribution, which is generally not available in closed form. To approximate it, we introduce a variational decoder $q_\theta(o|c)$.

Using the non-negativity of the Kullback-Leibler (KL) divergence, we have:

$$D_{\text{KL}}(p_{\text{emp}}(\cdot|c) \,\|\, q_\theta(\cdot|c)) \geq 0, \tag{22}$$

where

$$\begin{aligned} D_{\text{KL}}(p_{\text{emp}}(\cdot|c) \,\|\, q_\theta(\cdot|c)) &= \mathbb{E}_{p_{\text{emp}}(o|c)}\left[\log \frac{p_{\text{emp}}(o|c)}{q_\theta(o|c)}\right] \\ &= \mathbb{E}_{p_{\text{emp}}(o|c)}[\log p_{\text{emp}}(o|c)] - \mathbb{E}_{p_{\text{emp}}(o|c)}[\log q_\theta(o|c)]. \end{aligned} \tag{23}$$

Therefore, $D_{\text{KL}}(\cdot) \geq 0$ implies $\mathbb{E}_{p_{\text{emp}}(o|c)}[\log p_{\text{emp}}(o|c)] \geq \mathbb{E}_{p_{\text{emp}}(o|c)}[\log q_\theta(o|c)]$. Taking the expectation over concepts c extends this inequality to the joint distribution $p_{\text{emp}}(o,c)$:

$$\mathbb{E}_{p_{\text{emp}}(o,c)}[\log p_{\text{emp}}(o|c)] \geq \mathbb{E}_{p_{\text{emp}}(o,c)}[\log q_\theta(o|c)]. \tag{24}$$

Substituting this inequality into the MI equation yields:

$$\begin{aligned} I(O;C) &= H(O) + \mathbb{E}_{p_{\text{emp}}(o,c)}[\log p_{\text{emp}}(o|c)] \\ &\geq H(O) + \mathbb{E}_{p_{\text{emp}}(o,c)}[\log q_\theta(o|c)]. \end{aligned} \tag{25}$$

We define the variational lower bound objective $\mathcal{L}_{\text{MI}}$ as:

$$\mathcal{L}_{\text{MI}} \triangleq H(O) + \mathbb{E}_{p_{\text{emp}}(o,c)}[\log q_\theta(o|c)]. \tag{26}$$

Maximizing $\mathcal{L}_{\text{MI}}$ maximizes a lower bound on $I(O;C)$.

B.4 Equivalence to Reconstruction Loss

Since our decoder is a deterministic neural network $\hat{o} = g_\theta(c)$, we interpret it probabilistically by assuming $q_\theta(o|c)$ follows a fixed-variance Gaussian distribution centered at $\hat{o}$:

$$q_\theta(o|c) = \mathcal{N}(o; g_\theta(c), \sigma^2 \mathbf{I}), \tag{27}$$

where σ^2 is a fixed constant and $\mathbf{I}$ is the identity matrix. The log-likelihood term becomes:

$$\log q_\theta(o|c) = -\frac{1}{2\sigma^2}\|o - g_\theta(c)\|^2 + \text{const.} \tag{28}$$

Consequently, maximizing the MI lower bound is equivalent (up to an additive constant and a positive scaling factor) to minimizing the Mean Squared Error (MSE) reconstruction loss:

$$\mathcal{L}_{\text{rec}} = \mathbb{E}_{p_{\text{emp}}(o,c)}\left[\|o - g_\theta(c)\|^2\right]. \tag{29}$$

Conclusion The reconstruction loss serves as a proxy for maximizing a variational lower bound on the mutual information between observations and concepts under the current empirical data distribution. This promotes faithful concept extraction by encouraging the concepts to retain information sufficient for reconstructing the local observations (while additional bottlenecks or regularizers, when used, can further discourage trivial copying and encourage compact representations).

C Enforcing Concept Sufficiency via Variational Information Bottleneck

In this appendix, we justify our auxiliary global-state reconstruction objective. We consider the architecture where extracted concepts c are compressed into a latent bottleneck z, which is then used to reconstruct the global state s. We show that minimizing the reconstruction loss maximizes a variational lower bound of $I(s;z)$, and by the Data Processing Inequality (DPI) this pushes up a certified lower bound of $I(s;c)$.

C.1 Graphical Model and Assumptions

Let $s \in \mathcal{S}$ denote the global state and $c \in \mathcal{C}$ denote the aggregated concept representation (e.g., $c = \{c_i\}_{i=1}^N$). We introduce a bottleneck latent $z \in \mathcal{Z}$ sampled from an encoder $q_\phi(z \mid c)$, and reconstruct s using a decoder $p_\theta(s \mid z)$.

Assumption 1 (Markov property). Since z is generated solely from c, the induced variables satisfy the Markov chain

$$S \to C \to Z, \tag{30}$$

equivalently $q_\phi(z \mid c, s) = q_\phi(z \mid c)$ (i.e., $Z \perp S \mid C$). Denote the encoder-induced joint distribution by

$$q(s,c,z) = p_{\text{env}}(s)\, q_\psi(c \mid s)\, q_\phi(z \mid c), \quad q(s,z) = \int q(s,c,z)\, dc. \tag{31}$$

C.2 Connecting Reconstruction to Mutual Information

Our goal is to encourage the concepts c to preserve information about the global state s, i.e., to increase $I(S;C)$. We first show that reconstruction maximizes a variational lower bound of $I(S;Z)$.

Proposition 1 (Variational information maximization). Maximizing the expected log-likelihood $\mathbb{E}_{q(s,z)}[\log p_\theta(s \mid z)]$ maximizes a variational lower bound of $I_q(S;Z)$.

Proof. By definition,

$$I_q(S;Z) = H_q(S) - H_q(S \mid Z) = H_q(S) + \mathbb{E}_{q(s,z)}[\log q(s \mid z)]. \tag{32}$$

Introduce a variational decoder $p_\theta(s \mid z)$. Using the non-negativity of KL divergence,

$$\mathbb{E}_{q(s|z)}[\log q(s \mid z)] \geq \mathbb{E}_{q(s|z)}[\log p_\theta(s \mid z)], \tag{33}$$

we obtain the Barber–Agakov lower bound:

$$I_q(S;Z) \geq H_q(S) + \mathbb{E}_{q(s,z)}[\log p_\theta(s \mid z)]. \tag{34}$$

Since $H_q(S)$ is constant w.r.t. (ϕ, θ), maximizing $\mathbb{E}_{q(s,z)}[\log p_\theta(s \mid z)]$ (or minimizing the negative log-likelihood reconstruction loss) maximizes this lower bound. □

C.3 Enforcing Concept Sufficiency via DPI

Proposition 2 (Sufficiency propagation via DPI). Increasing $I_q(S;Z)$ pushes up a certified lower bound of $I_q(S;C)$.

Proof. From the Markov chain $S \to C \to Z$, the Data Processing Inequality (DPI) implies

$$I_q(S;C) \geq I_q(S;Z). \tag{35}$$

Combining Eq. (34) and Eq. (35) yields

$$I_q(S;C) \;\geq\; I_q(S;Z) \;\geq\; H_q(S) + \mathbb{E}_{q(s,z)}[\log p_\theta(s \mid z)]. \tag{36}$$

Therefore, improving reconstruction from z increases a guaranteed lower bound on $I_q(S;C)$: since Z is computed solely from C, Z cannot contain information about S that is not already present in C. Thus, demanding accurate reconstruction $Z \to S$ forces the upstream concept representation C to preserve globally predictive information. □

C.4 Role of the Bottleneck and KL Regularization (Variational IB)

We optimize a bottlenecked objective of the form

$$\mathcal{L}_{\text{aux}} = \mathbb{E}_{q(s,c)}\mathbb{E}_{q_\phi(z|c)}[\log p_\theta(s \mid z)] - \beta\, \mathbb{E}_{q(c)}\mathrm{KL}\big(q_\phi(z \mid c) \,\|\, p(z)\big), \tag{37}$$

where $p(z)$ is a simple prior. The KL term imposes an information bottleneck on the channel $C \to Z$. In particular, under $q(c)q_\phi(z \mid c)$, we have the decomposition

$$\mathbb{E}_{q(c)}\mathrm{KL}\big(q_\phi(z \mid c) \,\|\, p(z)\big) = I_q(C;Z) + \mathrm{KL}\big(q(z) \,\|\, p(z)\big), \tag{38}$$

showing that the regularizer controls the information capacity $I_q(C;Z)$ and aligns the aggregated posterior $q(z)$ with the prior. Hence, the auxiliary objective implements a Variational Information Bottleneck for predicting S from C: it encourages Z (and thus C) to retain compressible, robust information that is maximally predictive of the global state.

D PSEUDO CODE

Algorithm 1 Escaping Local Views (ELV)

Initialize a set of agents $\mathcal{N} = \{1, \ldots, n\}$
Initialize concept networks and policy networks $Q_i(\tau_i, u_i; \theta)$, target networks parameterized by $\hat{\theta} \leftarrow \theta$
Initialize VAE components, concept predictor and concept decoder with parameters ϕ
Initialize a replay buffer $\mathcal{B}$ for storing episodes
repeat
  Initialize a history state h_i^0 and previous action u_i^0 for each agent
  Observe each agent's partial observation $[o_i^1]_{i=1}^n$
  for $t = 1 : T$ **do**
    Get current history h_i^t and extract latent concepts c_i^t
    Infer global latent context z from concepts and calculate value function $Q_i(\tau_i^t, u_i^t)$
    Select action u_i^t via ϵ-greedy exploration based on Q_i
    Execute joint action $\boldsymbol{u}^t$ to receive reward r^t and next state s^{t+1}
  end for
  Store the episode trajectory to $\mathcal{B}$
  Sample a batch of episode trajectories with batch size b from $\mathcal{B}$
  for $t = 1 : T$ **do**
    Predict next concepts $\hat{c}^{t+1}$ and calculate intrinsic reward r_t^{int} via prediction error in Eq. (14)
    Calculate augmented reward $r_t = r_t^{env} + \beta_t \cdot r_t^{int}$
    Get reconstructed observation $\tilde{o}_i^t$ from concepts and calculate grounding loss $\mathcal{L}_{rec}$ via Eq. (10)
    Construct global latent variable $z \sim q_\phi(z|c)$ and calculate VAE loss $\mathcal{L}_{vae}$ via Eq. (6)
    Calculate concept importance weights α_k, β_k using dual-path attention in Eq. (7) and (8)
    Compute individual Q-values incorporating second-order concept interactions via Eq. (9)
  end for
  Construct the total loss function defined in Eq. (16)
  Update θ and ϕ by minimizing the total loss
  Periodically update target parameters $\hat{\theta} \leftarrow \theta$
until Q_i converges

E RELATED WORK

Interpretability in MARL. The opacity of deep MARL policies has traditionally been addressed via post-hoc explanations, such as saliency-based analyses (Greydanus et al., 2018) or Shapley-style attributions for value decomposition (Wang et al., 2020). While informative, post-hoc explanations can be sensitive to perturbations and may provide primarily correlational evidence rather than a structured account of multi-agent coordination (Slack et al., 2021). This motivates intrinsically interpretable architectures that expose intermediate decision structure, e.g., differentiable trees or symbolic distillation (Bastani et al., 2018; Liu et al., 2025). However, such designs may struggle to simultaneously maintain readability and expressiveness when coordination is high-dimensional and nontrivial. Graph-based interpretability methods can reveal who interacts (Wu et al., 2025), but may not directly ground what semantic content is exchanged. In contrast, ELV emphasizes cross-agent semantic coherence by training concepts to be globally informative under a VAE-style reconstruction objective, mitigating fragmented local views.

Concept Bottleneck Models (CBMs). Concept Bottleneck Models (CBMs) introduce an interpretable intermediate layer where input features are first mapped to human-understandable concepts before making final predictions (Koh et al., 2020). While early CBMs suffered from a "performance-interpretability trade-off," Concept Embedding Models (CEMs) (Espinosa Zarlenga et al., 2022) mitigated this by representing concepts as high-dimensional vectors. To bypass the need for exhaustive manual annotations, Post-hoc CBMs (Yuksekgonul et al., 2023) and Label-free CBMs (Oikarinen et al., 2023) leveraged pre-trained vision-language models like CLIP to automatically define concept spaces. Recent advances, such as Probabilistic CBMs (Kim et al., 2023) and Soft-CBMs (Havasi et al., 2022), further enhance robustness by modeling concept uncertainty, ensuring

that semantic abstractions remain reliable even under distribution shifts or noisy inputs. Recent research has sought to bridge such semantic abstractions with reinforcement learning. For instance, SCoBots (Delfosse et al., 2024) successfully utilizes hierarchical concepts to align single-agent policies, yet extending this transparency to decentralized coordination remains non-trivial.

Semantic Exploration and Intrinsic Motivation. Efficient exploration in sparse-reward MARL remains challenging. Classic intrinsic motivation methods such as curiosity or RND (Pathak et al., 2017; Burda et al., 2019) often operate in raw observation spaces and can be distracted by task-irrelevant stochasticity. Beyond this, recent studies observe that naively mixing intrinsic and extrinsic objectives may bias optimization and cause "over-exploration"; accordingly, constrained or adaptive schemes have been proposed to automatically regulate intrinsic reward strength and reduce such interference (Chen et al., 2022). To further improve robustness, uncertainty-aware curiosity explicitly down-weights inherently unpredictable (aleatoric) transitions, mitigating the noisy-TV failure mode (Mavor-Parker et al., 2022). Meanwhile, structure-aware intrinsic rewards leverage transition regularities (e.g., successor–predecessor representations) to encourage bottleneck-seeking exploration beyond purely local novelty (Yu et al., 2023). Recent methods also incorporate social context or calibration to stabilize intrinsic signals (Pan et al., 2025). ELV is complementary: it computes prediction errors in a learned latent concept space, encouraging exploration toward semantically uncertain transitions rather than pixel-level noise (McCaffrey et al., 2025).

F Experimental Details



(a) Level Based Foraging (b) StarCraft Multi-Agent Challenge (c) StarCraft Multi-Agent Challenge v2

Figure 10: Three benchmarks used in our experiments.

F.1 Benchmark and Settings

The majority of our experiments are carried out on Level Based Foraging (LBF), StarCraft Multi-Agent Challenge (SMAC) and SMACv2, as shown in *Figure* 10. We describe the details and settings of these benchmarks in this section. Comprehensive experimental details are enumerated in Table 1 and Table 2.

Table 1: Experimental Settings of Level-Based Foraging.

Hyperparameter	Value	Description
Max Player Level	3	Maximum Agent Level Attribute
Max Episode Length	50	Maximum Timesteps Per Episode
Batch Size	32	Number of Episodes Per Update
Test Interval	10,000	Frequency of Evaluating Performance
Test Episodes	32	Number of Episodes to Test
Replay Batch Size	5000	Maximum Number of Episodes Stored in Memory
Discount Factor γ	0.99	Degree of Impact of Future Rewards
Total Timesteps	1,050,000	Number of Training Steps
Start ϵ	1.0	The Start ϵ Value to Explore
Finish ϵ	0.05	The Finish ϵ Value to Explore
Anneal Steps for ϵ	50,000	Number of Steps of Linear Annealing
Target Update Interval	200	The Target Network Update Cycle

Table 2: Experimental settings of SMAC and SMACv2.

HYPERPARAMETER	VALUE	DESCRIPTION
DIFFICULTY	7	ENEMY UNITS WITH BUILT-IN AI DIFFICULTY
BATCH SIZE (SMAC)	32	NUMBER OF EPISODES PER UPDATE
BATCH SIZE (SMACV2)	128	NUMBER OF EPISODES PER UPDATE
TEST INTERVAL	10,000	FREQUENCY OF EVALUATING PERFORMANCE
TEST EPISODES	32	NUMBER OF EPISODES TO TEST
REPLAY BATCH SIZE	5000	MAXIMUM NUMBER OF EPISODES IN MEMORY
DISCOUNT FACTOR γ	0.99	DEGREE OF IMPACT OF FUTURE REWARDS
START ϵ	1.0	THE START ϵ VALUE TO EXPLORE
FINISH ϵ	0.05	THE FINISH ϵ VALUE TO EXPLORE
ANNEAL STEPS FOR EASY & HARD	50,000	NUMBER OF STEPS OF LINEAR ANNEALING ϵ
ANNEAL STEPS FOR SUPER HARD	100,000	NUMBER OF STEPS OF LINEAR ANNEALING ϵ
TARGET UPDATE INTERVAL	200	THE TARGET NETWORK UPDATE CYCLE

Table 3: The detailed specifications of SMAC scenarios used in our evaluation.

MAP NAME	ALLY UNITS	ENEMY UNITS	TIMESTEPS	SCENARIO TYPE
2s3z	2 Stalkers & 3 Zealots	2 Stalkers & 3 Zealots	2M	Easy
3s5z	3 Stalkers & 5 Zealots	3 Stalkers & 5 Zealots	2M	Easy
1c3s5z	1 Colossus, 3 Stalkers & 5 Zealots	1 Colossus, 3 Stalkers & 5 Zealots	2M	Easy
2s_vs_1sc	2 Stalkers	1 Spine Crawler	2M	Easy
5m_vs_6m	5 Marines	6 Marines	2M	Hard
8m_vs_9m	8 Marines	9 Marines	2M	Hard
3s_vs_5z	3 Stalkers	5 Zealots	2M	Hard
2c_vs_64zg	2 Colossi	64 Zerglings	2M	Hard
MMM2	1 Medivac, 2 Marauders & 7 Marines	1 Medivac, 3 Marauders & 8 Marines	2M	Super Hard
3s5z_vs_3s6z	3 Stalkers & 5 Zealots	3 Stalkers & 6 Zealots	5M	Super Hard
6h_vs_8z	6 Hydralisks	8 Zealots	5M	Super Hard
corridor	6 Zealots	24 Zerglings	5M	Super Hard

Level-Based Foraging (LBF). LBF simulates a cooperative multi-agent task on a 10×10 grid, where agents with initialized skill levels must coordinate to collect food items under partial observability (restricted to a 5×5 local view). The core challenge lies in its enforced cooperation mechanism: a food item with difficulty L_f can only be consumed if the sum of the levels of participating agents meets the threshold ($\sum l_i \geq L_f$), necessitating dynamic subgroup formation. To encourage efficient exploration, the environment applies a step penalty of -0.002 at each timestep, and we evaluate our method on two specific configurations: 4-Players & 2-Foods and 6-Players & 4-Foods.

StarCraft Multi-Agent Challenge. The StarCraft Multi-Agent Challenge (SMAC) (Samvelyan et al., 2019) is based on StarCraft II (version SC2.4.10), which provides a suite of cooperative multi-agent reinforcement learning (MARL) tasks where a team of allied units is trained to defeat enemy units controlled by the built-in StarCraft II AI. The enemy AI operates at difficulty level 7 in all scenarios. Each allied unit is controlled by an individual decentralized agent, which observes the environment locally and learns to act cooperatively through reinforcement learning. Agents are trained to maximize cumulative rewards by dealing as much damage as possible to enemy units while avoiding unnecessary damage to themselves. These tasks demand a wide range of cooperative behaviors, including focus fire, kiting, flanking, and spatial coordination. The evaluation is performed across multiple challenging combat scenarios with varying levels of complexity, agent types, and asymmetry. Some scenarios involve homogeneous units, while others include heterogeneous unit compositions. A summary of the selected scenarios is provided in Table 3.

StarCraft Multi-Agent Challenge v2. StarCraft Multi-Agent Challenge v2 (SMACv2) (Ellis et al., 2023) is a significant evolution of the original benchmark designed to rigorously test the generalization capabilities of MARL algorithms. Unlike the static scenarios in v1, SMACv2 introduces stochasticity

Table 4: The SMACv2 scenarios used in our evaluation.

SCENARIO NAME	ALLIES	ENEMIES	TIMESTEPS	UNIT COMPOSITION (RANDOMIZED)
Terran_5_vs_5	5	5	1M	MARINES, MARAUDERS, MEDIVACS
Terran_10_vs_10	10	10	2M	
Zerg_5_vs_5	5	5	1M	ZERGLINGS, HYDRALISKS, BANELINGS
Zerg_10_vs_10	10	10	2M	
Protoss_5_vs_5	5	5	1M	STALKERS, ZEALOTS, COLOSSI
Protoss_10_vs_10	10	10	2M	

through randomized initial configurations, with unit compositions and initial positions varying across episodes. It also adjusts unit sight and attack ranges to increase diversity among agents. We use SC2.4.10 and train for 1M environment steps on 5-unit scenarios and 2M steps on 10-unit scenarios. Table 4 summarizes the detailed specifications for each scenario.

F.2 HYPERPARAMETERS OF BASELINES

We implement a diverse set of baselines using the PyMARL framework (Samvelyan et al., 2019). These include five classic value-decomposition methods—such as QMIX (Rashid et al., 2020), VDN (Sunehag et al., 2018), QPLEX (Wang et al., 2021), CDS (Li et al., 2021) and QTRAN (Son et al., 2019)—as well as recent approaches focusing on exploration and communication, including $\mathrm{NA}^2\mathrm{Q}$ (Liu et al., 2023), ReBorn (Qin et al., 2024), $\mathrm{MA}^2\mathrm{E}$ (Kang et al., 2025), and SHAQ (Wang et al., 2022). For fair comparison, all baselines adhere to the default hyperparameters provided in their respective official implementations or the PyMARL codebase.

During evaluation, we employ decentralized greedy action selection, testing the policy every $10,000$ environment steps over 32 episodes. To ensure statistical robustness, all experiments are conducted across five random seeds in environments characterized by clear success/failure dynamics. The metric *test win rate* is defined as the percentage of episodes where allied agents successfully eliminate all enemies within the time limit. Optimization is performed using RMSprop with a learning rate of 5×10^{-4}, with target networks updated every 200 training episodes.

F.3 HYPERPARAMETERS OF ELV

We implement ELV based on the PyMARL framework and optimize the entire ELV architecture using the Adam optimizer with a learning rate of 5×10^{-4}. The target networks are updated every 200 training episodes to stabilize value estimation. For a detailed listing of the structural hyperparameters and loss coefficients specific to ELV, please refer to Table 5.

Table 5: Hyper-parameters of ELV.

COMPONENT	HYPER-PARAMETERS	VALUE
CONCEPT EXTRACTION	CONCEPT NUMBER (K)	16
	CONCEPT DIMENSION (d)	16
	HISTORY HIDDEN DIM	64
VARIATIONAL AUTOENCODER	LATENT CONTEXT DIM (z)	32
	VAE LOSS WEIGHT λ_1	0.1
	REC. LOSS WEIGHT λ_2	0.1
INTRINSIC REWARD	PRED. LOSS WEIGHT λ_3	0.05
	INITIAL SCALE β_0	1
	DECAY COEFFICIENT λ	9.2×10^{-8}
	DECAY SCHEDULE	Exponential
DUAL-PATH CONCEPT IMPORTANCE	QUERY DIMENSION	16
	KEY DIMENSION	16

G Additional Experimental Results

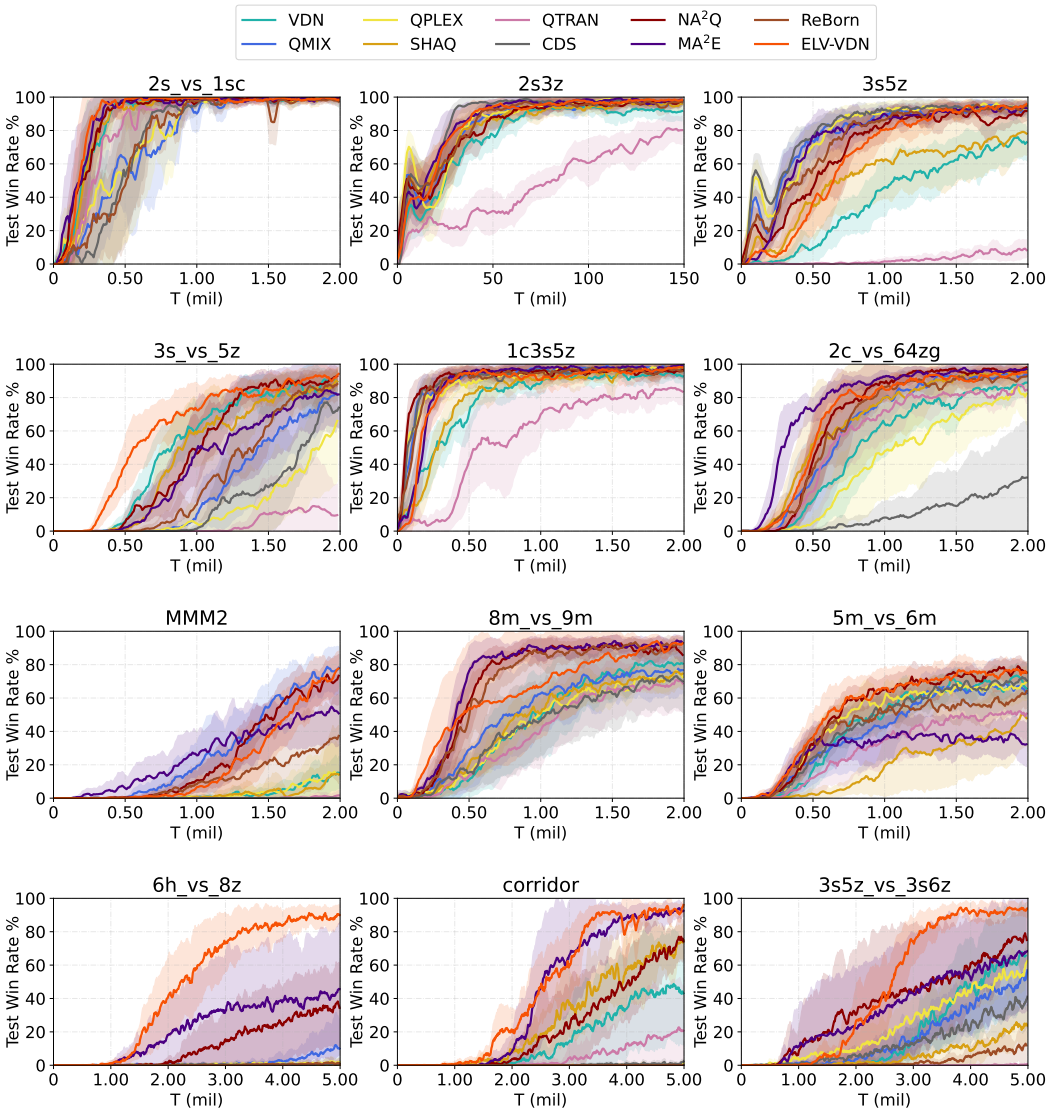


Figure 11: Test win rate (%) of ELV with the VDN mixing network.

G.1 Generality across Mixing Architectures

Since the proposed ELV framework focuses on restructuring the local agent's policy network via latent concept extraction, it is theoretically compatible with arbitrary global mixing networks. To empirically validate this architectural flexibility, we integrated ELV with Value-Decomposition Networks (VDN), which employs a simple linear summation for value aggregation, contrasting with the non-linear mixing of QMIX used in our main experiments. As illustrated in *Figure* 11, ELV-VDN demonstrates remarkable performance improvements over the vanilla VDN baseline. This advantage is particularly evident in super-hard scenarios such as 6h_vs_8z and corridor. In these tasks, the standard VDN fails to learn effective policies due to its limited representational capacity, hovering near a 0% win rate. In stark contrast, ELV-VDN achieves high win rates competitive with distinctively more complex architectures like QMIX and QPLEX. This result underscores that the semantic concepts learned by ELV provide a robust inductive bias, enabling simple mixing mechanisms to solve complex coordination problems. It confirms that ELV is a generalizable framework compatible with diverse value factorization backbones.

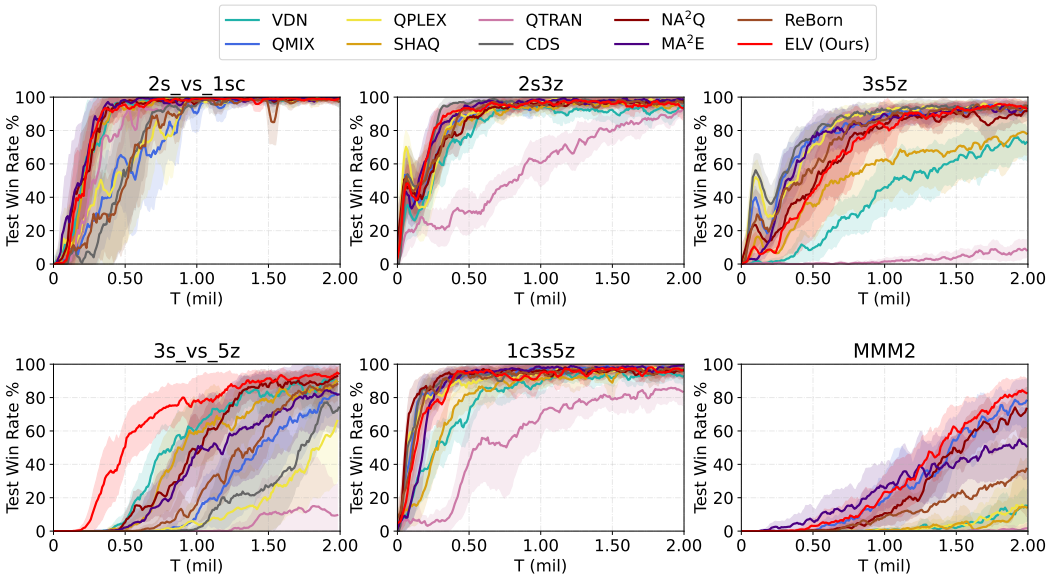


Figure 12: Test win rate (%) on six additional SMAC scenarios.

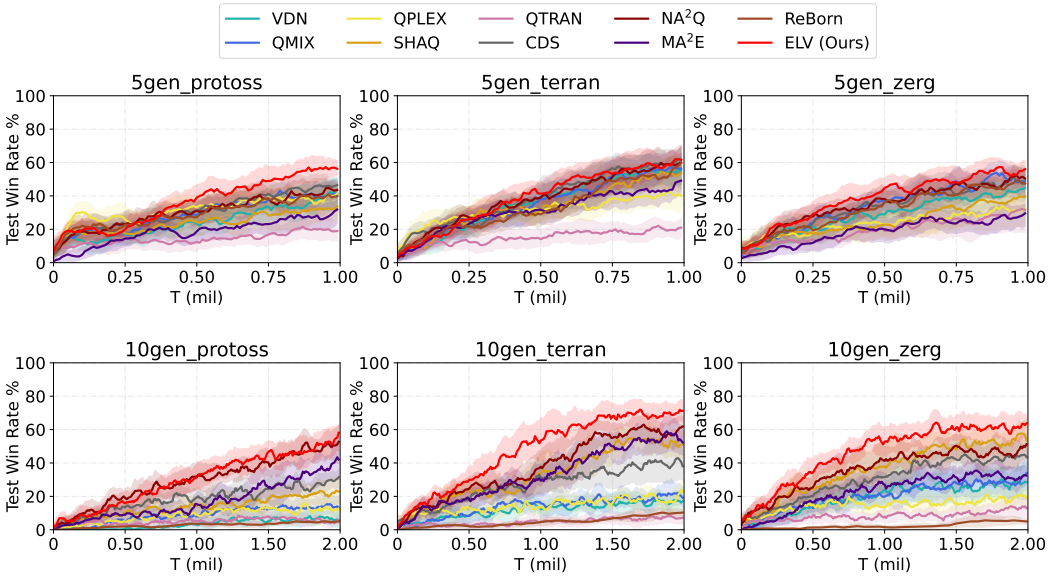


Figure 13: Test win rate (%) on SMACv2.

G.2 Additional Experimental Results on SMAC Extra Scenarios

To assess ELV across varying difficulty levels, we further evaluate its performance on six additional SMAC scenarios. As illustrated in *Figure* 12, ELV reaches near-saturated win rates on simpler scenarios such as *2s3z*, *3s5z*, and *1c3s5z*, matching strong baselines such as QMIX and QPLEX. These results suggest that introducing the concept bottleneck and auxiliary reconstruction objectives does not compromise performance on these tasks. The super-hard *MMM2* scenario exhibits a different pattern: win rates remain well below 100% within the training budget. ELV continues improving and achieves a higher final win rate than QMIX and VDN, while QPLEX remains competitive. These results indicate that ELV also remains competitive under more demanding coordination requirements, although substantial room for improvement remains on *MMM2*.

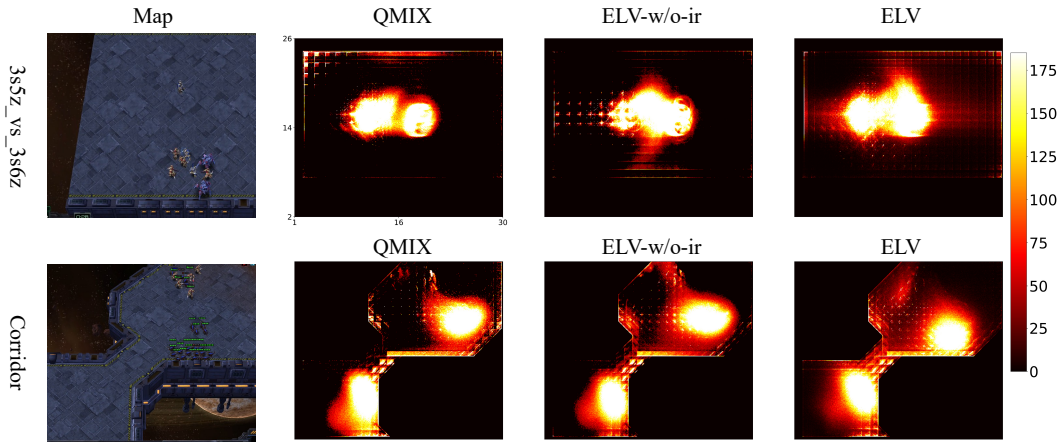


Figure 14: Visualization of exploration coverage via spatial heatmaps.

G.3 Performance on SMACv2

Performance on SMACv2. We further assess the generalization capability of ELV under the high stochasticity of SMACv2, where unit compositions and start positions are randomized. As shown in *Figure* 13, ELV consistently achieves superior win rates across all scenarios, establishing a significant performance margin in the more complex 10-unit tasks (e.g., 10gen_terran and 10gen_zerg). Notably, CDS, despite being designed for heterogeneous agents, exhibits high variance and suboptimal convergence. This failure suggests that merely encouraging behavioral diversity is insufficient to adapt to dynamic tactical shifts without grounded semantic understanding. Similarly, ReBorn struggles in large-scale settings (10gen), implying that its exploration mechanism may be inefficient in handling the exponentially growing state space of randomized scenarios. In contrast, QMIX and QPLEX require significantly more samples to reach competitive performance, struggling with the non-stationarity of the environment. ELV's success stems from its ability to extract invariant semantic concepts from variable observations, effectively filtering out stochastic noise and enabling robust coordination where baselines falter.

G.4 Qualitative Analysis of Exploration

To intuitively verify how the proposed concept-based intrinsic reward facilitates exploration, we visualize the spatial heatmaps of agent trajectories throughout the training process in *Figure* 14. In the super-hard scenario 3s5z_vs_3s6z, which demands complex kiting and positioning, QMIX exhibits a highly concentrated visitation pattern, indicating that agents tend to stagnate in local optima with limited movement. Although removing the intrinsic reward (ELV-w/o-ir) slightly expands the range, it still fails to cover the map effectively. In contrast, the full ELV model demonstrates significantly broader spatial coverage, suggesting that agents actively utilize the entire terrain to execute superior micro-tactics. In the navigation-heavy Corridor map (Bottom Row), the disparity is even more pronounced. Agents must pass through a narrow bottleneck to reach enemies. Baselines like QMIX and ELV-w/o-ir show high-density clustering at the bottom-left entry, failing to effectively penetrate the choke point. ELV, however, not only successfully traverses the narrow passage but also achieves a more even and widespread distribution within the top-right target area relative to the baselines. This indicates that the intrinsic reward effectively drives agents to break through bottlenecks and explore the task-critical region more thoroughly, rather than clustering at the entrance.

H CONCEPT SEMANTICS AND BEHAVIORAL VALIDATION

We examine learned concepts through semantic assignments, test-time interventions, and temporal tracking over an episode.

H.1 SEMANTIC ROLES AND TEST-TIME CONCEPT INTERVENTIONS

Semantic roles of learned concepts. ELV learns concept representations without ground-truth concept annotations. To interpret the resulting slots, we use the concept–feature attribution matrix G from Section 4.4, which measures the sensitivity of each observation modality's reconstruction to each concept. The maximum-attribution matching yields a semantic assignment of concept slots to structured observation features. Table 6 presents the resulting assignments for a representative model on *3s_vs_5z*. The assigned concepts cover movement, individual enemies, teammates, self-state, unit type, and action history, providing a concrete interpretation of the concept–feature structure visualized in *Figure* 8. For example, C2 is associated with tracking enemy unit 1, while C11 and C14 are associated with self-state and movement direction, respectively.

Table 6: Attribution-based semantic assignments of learned concept slots on *3s_vs_5z*. Concept IDs refer to slots in the analyzed model.

Concept ID	Dominant feature	Semantic label
C14	move	Movement direction
C2	enemy1	Enemy unit 1 tracking
C13	enemy2	Enemy unit 2 tracking
C9	enemy3	Enemy unit 3 tracking
C5	enemy4	Enemy unit 4 tracking
C0	enemy5	Enemy unit 5 tracking
C15	ally1	Teammate 1 monitoring
C6	ally2	Teammate 2 monitoring
C11	self	Self-state
C12	type	Unit type recognition
C4	last_action	Action history

Influence on policy behavior. We next perform test-time interventions on a trained ELV model on *3s_vs_5z* by setting the existence probability of one concept slot to either 0 or 1 during evaluation. Under the concept mixture defined in Section 3.1, these settings select the slot's absence or presence embedding, respectively. Each intervention setting is evaluated over 10 episodes; the unmodified policy achieves a win rate of 1.00. Table 7 reports the results for all 11 assigned slots.

The effects differ substantially across concepts. Setting the existence probability of C9 or C11 to 0 reduces the win rate to 0.00, while the same intervention on C14 or C2 reduces it to 0.10. In contrast, interventions on C15 and C6 have smaller effects, with win rates remaining between 0.80 and 0.90. Fixing a concept's probability to 1 also changes performance: for example, C2 yields a win rate of 0.20, compared with 0.90 for C4 and C6. These results show that the concept variables materially influence policy behavior and contribute unequally to coordination. The reductions under both intervention settings further highlight the importance of context-dependent concept gating rather than keeping a slot permanently present or absent.

Table 7: Test-time concept interventions on *3s_vs_5z*. Each setting uses 10 evaluation episodes. The unmodified policy's win rate is 1.00; all values are proportions.

Concept ID	Win rate ($p = 0$)	Win rate ($p = 1$)
C9	0.00	0.30
C11	0.00	0.60
C14	0.10	0.70
C2	0.10	0.20
C13	0.60	0.60
C4	0.60	0.90
C5	0.60	0.70
C0	0.70	0.80
C12	0.80	0.70
C15	0.90	0.80
C6	0.90	0.90

H.2 Temporal Evolution of a Semantically Assigned Concept

To examine how an assigned concept evolves with its corresponding semantic factor, we follow Agent 0's Concept 2 together with enemy unit 1 over a full episode. *Figure* 15 shows the agent–enemy distance and enemy visibility, concept activation, and enemy health on a shared time axis. The activation decreases as the agent–enemy distance grows during approximately $t = 15$–30, then returns to a higher level when the enemy is nearby during approximately $t = 30$–63. At the marked enemy death event ($t = 64$), activation drops sharply and remains lower for the rest of the episode. This temporal pattern connects C2's attribution-based assignment to changes in the tracked enemy's state. Together with the semantic assignment and intervention results, it illustrates how learned concept slots provide a meaningful interface for inspecting policy behavior over time.

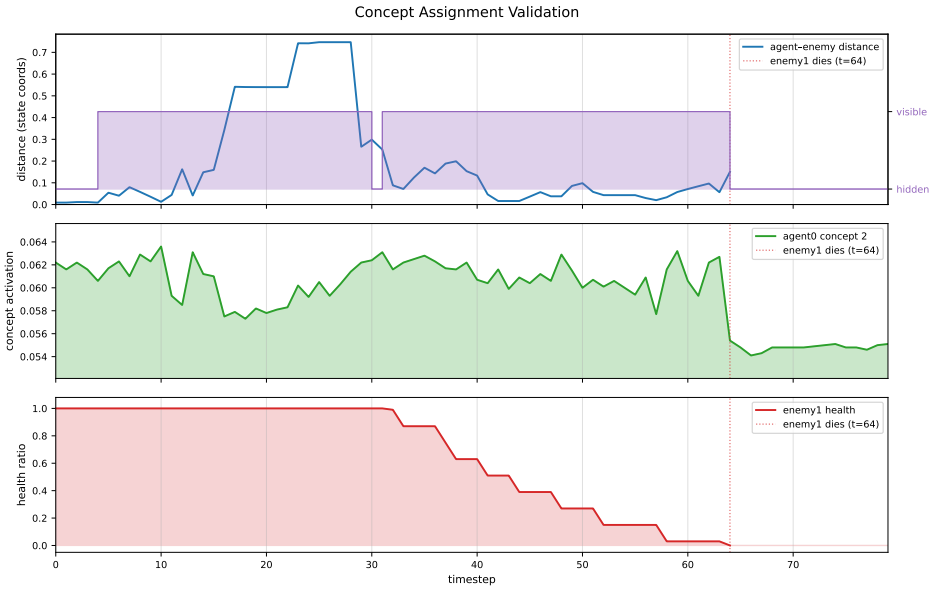


Figure 15: Temporal validation of the assignment of Agent 0's Concept 2 to enemy unit 1. The panels show agent–enemy distance with enemy visibility (top), concept activation (middle), and enemy health (bottom). The dotted vertical line marks the enemy death event at $t = 64$.

I PERFORMANCE UNDER RESTRICTED LOCAL OBSERVATIONS

We evaluate the effect of partial observability in LBF by varying the local observation radius, $\text{sight} \in \{2, 4, 8\}$, in the 6-players and 4-foods setting on a 10×10 grid. *Figure* 16 compares ELV with MA^2E over one million environment steps using the mean test return. Both methods use the same QMIX mixing network, allowing the comparison to focus on their individual-value representations under restricted observations.

ELV maintains similar final returns across all three observation radii, including the most restrictive setting, $\text{sight} = 2$. In contrast, MA^2E achieves substantially lower final returns at radii 2 and 4, while the two methods reach similar final performance at radius 8. ELV also learns faster in the two more restricted settings. The concentration of its advantage at smaller observation radii supports the role of the concept-based representation in handling limited local information, rather than attributing the improvement to a stronger global mixer.

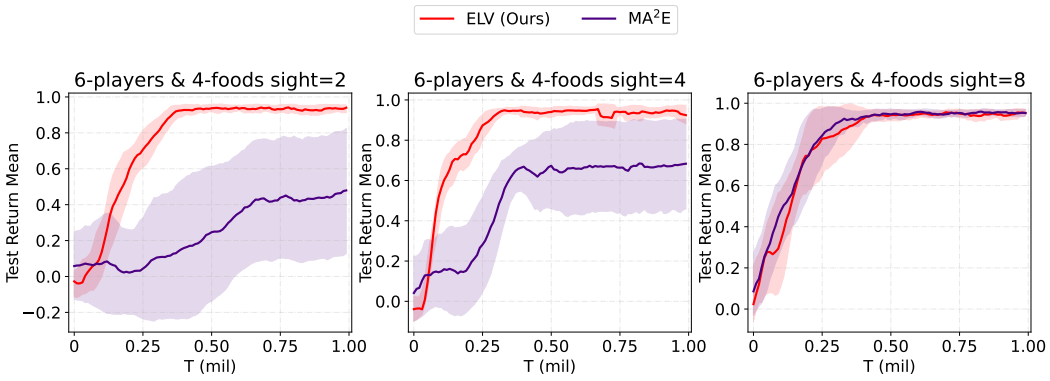


Figure 16: Mean test return on LBF with 6 players and 4 foods on a 10×10 grid, using observation radii of 2, 4, and 8. ELV and MA^2E use the same QMIX mixing network.