Structured Critic Adaptation

Ahmad Mouri Zadeh Khaki [*], and Jiyun Kim [†]

Complied on August 14, 2026

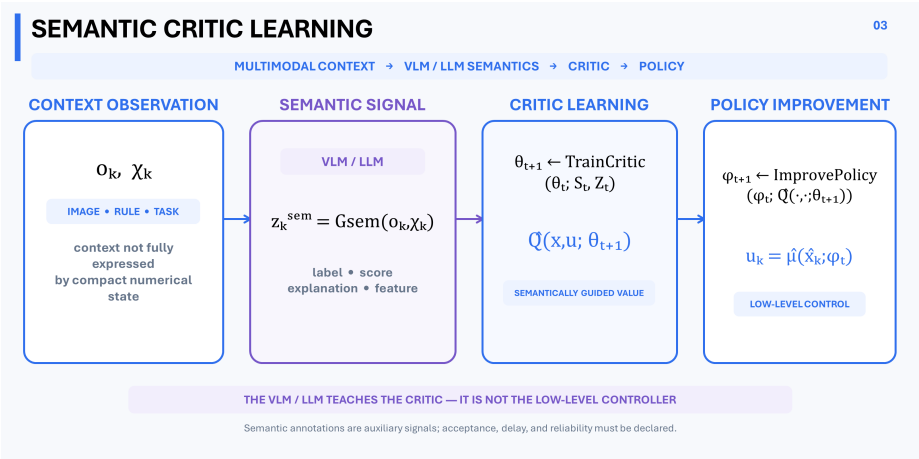

Figure 1: Multimodal observations produce semantic annotations that provide auxiliary critic supervision alongside Bellman data, and the updated critic guides policy improvement.

Physical interaction is indexed by k, whereas critic and policy updates are indexed by t. Semantic annotations may be generated offline, asynchronously, or during operation; their timing and delay must be declared. The figure shows the defining training-only route. Semantics required at action time must instead enter the deployed information state.

1 Problem definition

Online Learning-Based Optimal Control defines a policy through a value or Q-function. Semantic Critic Learning addresses cases in which multimodal or textual observations can provide evaluation-relevant supervision not readily expressed by the standard numerical transition and stage-cost record.

[*]Ahmad Mouri Zadeh Khaki is Postdoctoral Researcher at KAIST, ahmadmouri@kaist.ac.kr

[†]Jiyun Kim is Researcher at KAIST, JiyunKim@kaist.ac.kr

Let the controller receive a state estimate $\widehat{x}_k$, while o_k contains observations such as images, scene relations, task descriptions, or operator-provided rules. A VLM is suitable when visual and language context must be interpreted together; an LLM can be used when the relevant context is represented textually. Denote the selected high-level semantic interpreter by G_{sem}. It produces

$$z_k^{\text{sem}} = G_{\text{sem}}(o_k, \chi_k), \tag{1}$$

where χ_k is optional task, prompt, or rule context. z_k^{sem} is one semantic annotation, such as a label, score, preference, explanation, or feature; it is not itself a control input. Before entering a critic loss, a raw annotation must be mapped to a declared target, comparison, constraint, or feature. In the notation below, G_{sem} includes any required mapping and z_k^{sem} denotes the resulting critic-usable annotation.

The defining case uses z_k^{sem} only as training supervision. If semantic context is needed to distinguish deployed values or actions, it must enter the controller-facing information state:

$$\xi_k = (\widehat{x}_k, z_k^{\text{sem}}), \qquad \widehat{Q}_\theta(\xi_k, u), \qquad \widehat{\mu}_\phi(\xi_k). \tag{2}$$

Here z_k^{sem} is a declared control-usable encoding available at decision time, not unrestricted text. Training-only semantic targets must be representable from the deployed information state; otherwise identical deployed inputs can receive conflicting targets. Semantics that changes the desired return requires an explicitly redefined stage cost and Bellman object. It also does not by itself make a partially observed process Markov: the controller still needs a sufficient observation history, belief state, or learned information state.

Define a physical transition and the data selected at learning update t as

$$d_k = (\widehat{x}_k, u_k, g_k, \widehat{x}_{k+1}), \qquad g_k := g(\widehat{x}_k, u_k), \qquad \mathcal{S}_t = \{d_k : k \in \mathcal{I}_t\}, \tag{3}$$

where $\mathcal{I}_t$ is the set of physical transition indices selected for update t. For an available annotation, let a predeclared rule decide whether it is admitted:

$$a_k^{\text{acc}} := A_{\text{sem}}(d_k, z_k^{\text{sem}}; \mathcal{R}_{\text{acc}}) \in \{0, 1\}. \tag{4}$$

Set $a_k^{\text{acc}} = 0$ when no annotation is available, and define

$$\mathcal{I}_t^{\text{sem}} = \{k \in \mathcal{I}_t : a_k^{\text{acc}} = 1\}. \tag{5}$$

Here $\mathcal{R}_{\text{acc}}$ is the declared acceptance rule. It may check schema and provenance, temporal assignment, consistency with measurements and hard rules, and a calibrated confidence or agreement test. Raw VLM or LLM self-confidence is not assumed to be calibrated. The gate a_k^{acc} is determined before critic training; it is not optimized with θ.

The accepted paired dataset is

$$\mathcal{Z}_t = \{(d_k, z_k^{\text{sem}}) : k \in \mathcal{I}_t^{\text{sem}}\}. \tag{6}$$

Rejected or unavailable annotations do not remove d_k from $\mathcal{S}_t$; they only omit semantic supervision for that transition. For compactness, the notation

uses a common update batch. An asynchronous implementation may reselect a retained transition when its annotation arrives.

The learning path is represented conceptually as

$$\begin{aligned} \theta_{t+1} &\leftarrow \mathrm{TrainCritic}(\theta_t; \mathcal{S}_t, \mathcal{Z}_t)\,, \\ \phi_{t+1} &\leftarrow \mathrm{ImprovePolicy}\Big(\phi_t; \widehat{Q}_{\theta_{t+1}}\Big)\,, \end{aligned} \tag{7}$$

where θ and ϕ are the critic and policy parameters. These operators are conceptual rather than a commitment to one algorithm. $\mathcal{S}_t$ supplies stage-cost and successor-state evidence for ordinary Bellman or TD learning; $\mathcal{Z}_t$ supplies auxiliary semantic guidance for the accepted subset.

The equations use a Q-factor because a sampled transition provides a direct model-free critic target. A state-value critic $\widehat{J}_\theta$ can be used instead when its successor-state and policy dependence are declared, as in the overview figure.

2 Why this problem matters

A scalar stage cost provides one scalar evaluation channel, while images, relations, or rules may support additional labels and comparisons: whether another agent is yielding, whether a maneuver conflicts with a contextual rule, or whether two outcomes should be evaluated differently.

A VLM or LLM may provide such high-level context as an auxiliary learning signal without being placed directly in the fast control loop. The research question is whether Bellman-compatible semantic supervision can improve task-level evaluation by the critic and, through that critic, improve the policy. Because the semantic model is fallible and distribution-dependent, richer context is not by itself evidence of safety, optimality, or generalization.

3 Key challenges

- **Physical grounding and semantic reliability:** a semantic annotation can conflict with measurements, dynamics, action feasibility, or operating constraints. Designing and calibrating the acceptance rule, rejection behavior, and fallback path is therefore part of the method.
- **Closed-loop timing and causal attribution:** semantic annotations may be delayed, asynchronous, or stale. They must be assigned to the state–action transitions that caused the evaluated outcome without future-data leakage.
- **Control-objective and safety consistency:** semantic supervision may change which decisions the critic prefers. Its role in the objective or information state must be declared, and closed-loop evidence must separate its effect from extra data, reward tuning, or imitation while evaluating performance and constraint violations.

4 Established approaches — semantic rewards and policy guidance

Semantic reward specification or shaping. An LLM can act as a proxy reward function, while a VLM can score visual observations against a language-defined goal (Kwon et al., 2023; Rocamonde et al., 2024). The resulting semantic score modifies the cost used by an otherwise standard RL algorithm:

$$\widetilde{g}_k = g_k + \lambda_{\text{rew}} s_k^{\text{sem}}, \qquad \lambda_{\text{rew}} \geq 0, \tag{8}$$

so the critic and policy optimize the modified return. This is a reward-level route, and an arbitrary semantic term can change the control objective. Here s_k^{sem} is defined as a semantic **cost** or penalty, so a larger value is less desirable; if a semantic model produces a reward-like score, its sign must be converted before it is added to the cost.

Preference-derived reward learning. Instead of requesting a raw scalar score, human or VLM feedback can compare trajectory segments or observations. A separate reward model is then learned from those preferences and supplied to RL (Christiano et al., 2017; Wang et al., 2024). Semantics still reaches the critic through a scalar learned reward.

Direct policy guidance. A preferred semantic action or demonstration can supervise the policy without defining a critic-level learning signal:

$$\mathcal{L}_{\text{IL}}(\phi) = -\sum_{k \in \mathcal{I}_{\text{IL}}} \log \pi_\phi(u_k^{\text{sem}} \mid \widehat{x}_k)\,. \tag{9}$$

$\mathcal{I}_{\text{IL}}$ indexes semantic demonstrations, and u_k^{sem} is the preferred or demonstrated action. Language-conditioned imitation is an established example of this direct route (Lynch and Sermanet, 2021). The stochastic-policy symbol π_ϕ is used here because the displayed loss is a likelihood-based imitation objective; the default control-policy symbol elsewhere in these pages remains μ_ϕ.

5 Research direction — critic-level semantic supervision

The proposed direction keeps physical transitions and stage costs in an explicit Bellman loss while adding a separate semantic supervision term. The equations below use the training-only case defined in Section 1, with deployed critic and policy inputs $\widehat{x}$.

For a nonempty update set $\mathcal{S}_t$ and a declared nonnegative TD-residual loss ℓ_{TD}, write

$$\begin{aligned} \mathcal{L}_{\text{Bellman}}(\theta; \mathcal{S}_t) &= \frac{1}{|\mathcal{S}_t|} \sum_{k \in \mathcal{I}_t} \ell_{\text{TD}}\left(\widehat{Q}_\theta(\widehat{x}_k, u_k) - y_{k,t}^{\text{TD}}\right), \\ y_{k,t}^{\text{TD}} &= g_k + \alpha \widehat{Q}_{\bar{\theta}_t}\left(\widehat{x}_{k+1}, \widehat{\mu}_{\bar{\phi}_t}(\widehat{x}_{k+1})\right), \end{aligned} \tag{10}$$

where $(\bar{\theta}_t, \bar{\phi}_t)$ are fixed target parameters during update t, and $0 < \alpha < 1$ is the discount factor. This is an approximate one-step policy-evaluation target associated with the target policy $\widehat{\mu}_{\bar{\phi}_t}$, not an optimal-Q target. Using data

generated by another behavior policy requires the usual coverage or off-policy assumptions.

The accepted annotations provide a separate critic-level supervision term:

$$\theta_{t+1} \approx \arg\min_{\theta} \left\{ \mathcal{L}_{\text{Bellman}}(\theta; \mathcal{S}_t) + \lambda_{\text{sem},t} \mathcal{L}_{\text{sem}}(\theta; \mathcal{Z}_t) \right\}, \qquad \lambda_{\text{sem},t} \geq 0, \tag{11}$$

$$\mathcal{L}_{\text{sem}}(\theta; \mathcal{Z}_t) = \begin{cases} \dfrac{1}{|\mathcal{Z}_t|} \displaystyle\sum_{(d_k, z_k^{\text{sem}}) \in \mathcal{Z}_t} \ell_{\text{sem}}\left(\widehat{Q}_\theta; d_k, z_k^{\text{sem}} \right), & |\mathcal{Z}_t| > 0, \\ 0, & \text{otherwise}, \end{cases} . \tag{12}$$

The acceptance rule has already removed rejected annotations, so this overview weights the accepted pairs equally. A later implementation may use separately calibrated confidence weights, but their source and calibration must be declared; raw VLM or LLM self-confidence is not treated as a probability of correctness.

$\lambda_{\text{sem},t}$ controls the relative contribution of semantic supervision. To retain $\widehat{Q}_\theta$ as a critic for the original stage cost g, ℓ_{sem} is limited to accepted value comparisons or terminal labels that are demonstrably compatible with that return. A comparison may reference another transition or trajectory segment through z_k^{sem}. A new preference or risk specification instead requires an explicitly redefined objective or constraint; conflicting labels that reveal missing decision context require an augmented deployed information state.

The updated critic then guides policy improvement:

$$\phi_{t+1} \leftarrow \text{ImprovePolicy}\left(\phi_t; \widehat{Q}_{\theta_{t+1}} \right). \tag{13}$$

A deployable method also needs annotation provenance, delay and cache handling, rejection, and fallback rules.

Evaluation should compare a nonsemantic critic, numerical reward shaping, semantic reward or reward-model baselines, direct semantic imitation, and the critic-level formulation above. Relevant metrics include Bellman error, critic calibration, control performance, semantic acceptance and rejection, constraint violations, inference and update time, annotation sensitivity, and contextual distribution shift.

6 Application domains

- **Autonomous-driving decision and motion planning**, where visual scene relations, road conventions, and interaction context may provide auxiliary critic supervision or, when available online, augment the decision state.
- **Human–robot and language-conditioned control**, where task semantics may supervise value evaluation without unrestricted language output driving actuators.
- **Mobility and energy-management systems with contextual operating rules**, including exceptional modes difficult to encode in one handcrafted scalar cost.

7 References

- M. Kwon, S. M. Xie, K. Bullard, and D. Sadigh, "Reward Design with Language Models," *International Conference on Learning Representations (ICLR)*, 2023. Paper
- J. Rocamonde, V. Montesinos, E. Nava, E. Perez, and D. Lindner, "Vision-Language Models are Zero-Shot Reward Models for Reinforcement Learning," *ICLR*, 2024. Paper
- P. F. Christiano et al., "Deep Reinforcement Learning from Human Preferences," *Advances in Neural Information Processing Systems*, vol. 30, 2017. Paper
- Y. Wang et al., "RL-VLM-F: Reinforcement Learning from Vision Language Foundation Model Feedback," *Proceedings of the 41st International Conference on Machine Learning*, PMLR 235, pp. 51484–51501, 2024. Paper
- C. Lynch and P. Sermanet, "Language Conditioned Imitation Learning Over Unstructured Data," *Robotics: Science and Systems XVII*, 2021. Paper