Defense-in-Depth for LLM-Guided Live-Line Work Robots on Overhead Transmission Lines: Semantic Orthogonalization, Under-Specification Detection, and STL Execution Monitoring

Zhiguo Zhou

***Abstract*—Deploying LLM-guided live-line work robots for safety-critical power line maintenance requires reliable mapping from natural-language field instructions to certified robotic arm operations. A single misrouted command—such as "loosen" being executed as "tighten" for an M10 bolt—can cause irreversible mechanical damage or personnel hazard. We present a three-layer defense-in-depth framework that addresses this problem at complementary levels: (L1) a domain rule guard that detects explicit physical contradictions via keyword co-occurrence rules; (L2) a semantic routing layer built on a *semantically orthogonalized* structured vocabulary V_{damp}, combined with a two-gate under-specification detector; and (L3) a Signal Temporal Logic (STL) execution monitor that enforces torque and position constraints as a physical last resort. Our central finding is a *cross-domain orthogonalization principle*: effective semantic separation requires vocabulary labels from *different* semantic domains, not logical negations of the same concept. Validated on the vibration damper replacement task with `gemini-embedding-001`, the redesigned vocabulary increases the dangerous-pair distance $d(\text{R07}, \text{R09})$ by 54% and global $d_{\min}$ by 60%, resolving a previously undetected hidden dangerous pair. The STL monitor achieves 100% interception across three error scenarios with zero false positives.**

***Index Terms*—LLM routing, live-line work robots, power line maintenance, semantic orthogonalization, under-specification detection, Signal Temporal Logic, defense-in-depth**

I. Introduction

Intelligent inspection and live-line work robots are increasingly deployed for high-voltage overhead transmission line maintenance—crawling on conductors to inspect vibration dampers, replace hardware with onboard robotic arms, and perform bolt tightening operations that demand sub-millimeter precision and exact torque specifications [1]. Large language models (LLMs) offer a compelling interface: field technicians can issue plain-language commands, and the system translates them into certified robot operations from a predefined task vocabulary [2], [3].

However, the semantic gap between colloquial field instructions and certified operations creates a safety-critical routing problem that existing LLM safety methods do not address. Consider a live-line work robot performing vibration damper replacement. The operation "loosen the M10 bolt" (R07) and "tighten the M10 bolt" (R09) share nearly identical surface form—both mention bolt rotation—yet their physical consequences are opposite. A single misrouting of "loosen" as "tighten" can over-torque a joint to failure; the reverse can leave safety-critical fasteners unsecured under line tension. Our measurement confirms that in a standard natural-language vocabulary, the cosine distance between these two operations is only $d = 0.1225$, leaving them dangerously close in embedding space.

Three classes of natural-language ambiguity threaten routing safety. *Under-specified queries* ("turn the bolt a bit") provide insufficient direction for safe disambiguation. *Contradictory queries* ("counterclockwise tighten") encode physically impossible combinations. *Degenerate queries* ("just get it tight enough") exploit vague scalar qualifiers to bypass precision constraints. Current approaches—prompt engineering [4], output filtering, and RLHF alignment—address model behavior at the generation level but provide no formal guarantee at the *execution* level: if a routing error reaches the robot, there is no backstop.

Prior work on formal methods for robot safety [5], [6] uses Signal Temporal Logic (STL) to monitor sensor streams at runtime. Our companion paper [7] establishes vocabulary coverage completeness for the same power line maintenance domain. What remains open is the execution safety problem: given an ambiguous natural-language instruction, can the system (a) detect under-specification before routing, and (b) intercept a routing error before physical harm occurs?

This paper answers both questions with a three-layer defense-in-depth architecture. The central technical finding is a *cross-domain orthogonalization principle*: effective semantic separation between safety-critical vocabulary entries requires labels from different semantic domains—phase purpose, operation mode, material class—rather than logical negations of the same concept. Counterintuitively, adding directional qualifiers (counterclockwise vs. clockwise) *reduces* inter-entry distance by 0.0575 because both terms share the rotational semantic neighborhood. Phase labels (P2-Remove vs. P3-Install), by contrast, encode orthogonal task purpose and yield the largest separation gain (+0.0500 per step).

The main contributions of this paper are as follows:

- **C1: Cross-Domain Orthogonalization Principle.** We identify and formally state the condition under which structured vocabulary labels improve embedding-space separation for safety-critical operation pairs. We validate

Z. Zhou is with the School of Integrated Circuits and Electronics, Beijing Institute of Technology, Beijing 100081, China. *Corresponding author*: zhiguozhou@bit.edu.cn

the principle through a five-step ablation on V_{damp} and generalization to the 59-entry V167L vocabulary, demonstrating that cross-domain labels consistently improve separation while same-domain negations fail.

- **C2: $\mathbf{V}_{\mathrm{damp}}$ v2.0 and Two-Gate Detector.** Applying C1, we redesign the six core entries of the vibration damper replacement vocabulary, increasing the dangerous-pair distance $d(\mathrm{R07}, \mathrm{R09})$ by 54% (0.1225 $\rightarrow$ 0.1889) and global $d_{\min}$ by 60% (0.0734 $\rightarrow$ 0.1173), resolving a previously undetected hidden dangerous pair. A two-gate under-specification detector achieves 100% detection on all ambiguous query types.
- **C3: STL Execution Monitor.** We integrate STL-101/102 constraints as a physical last resort, demonstrating 100% interception of three routing-error scenarios with zero false positives (FPR=0% for $\sigma \leq 2.0\,\mathrm{N{\cdot}m}$) and maximum response time $\leq 25\,\mathrm{s}$. Together, the three layers form a defense-in-depth architecture in which each layer covers the failure modes of the one above it.

Within the EICPS framework [7], the three-layer architecture instantiates the complete EvidencePack execution protocol: L1 operates as Stage 0, filtering symbolic-level physical hallucinations before any embedding computation; L2 implements Stages 1–2, routing on the $\mathcal{M}_{\mathrm{sem}}$ discrete subgraph and detecting semantic Buridan's-ass states via spectral diagnosis; L3 realises Stage 3, providing a temporal instantiation of the safety spectrum $\mathcal{X}_{\mathrm{safe}}$ [7] through STL monitoring of proprioceptive and exteroceptive signals. The three-layer cascade guarantees that any routing error penetrating L1/L2 is intercepted by L3 before irreversible physical harm occurs. Together with Paper-01 (SEC=1 vocabulary coverage) and Paper-02 (Brain-layer LLM Proposal quality), this paper closes the language–semantic–physical safety guarantee chain of the EICPS framework.

Experimental scenario: why vibration damper replacement? This paper uses *vibration damper replacement* as the validation scenario, whereas Papers 01 and 02 use laser-UAV foreign-object removal. The choice is deliberate for two reasons. First, damper replacement involves mechanical contact—tightening and loosening bolts with a torque-controlled arm—providing genuine proprioceptive sensor feedback (torque, angular displacement) for STL-101/102 monitoring. Laser-UAV removal is a non-contact operation with weaker force-feedback signals, less suitable for STL physical-layer validation. Second, the primary dangerous pair in damper replacement—"loosen M10 bolt" (R07) and "tighten M10 bolt" (R09)—constitutes a natural *semantic Buridan's ass*: two operations with near-identical surface form but opposite physical consequences and similar embedding vectors, providing a maximally demanding test case for L2 disambiguation. The V167L vocabulary from Papers 01/02 does not contain this type of operation-direction ambiguity because laser-UAV tasks are structurally uni-directional; damper replacement, by contrast, has symmetric disassembly/installation phases that create the exact semantic equilibrium structure targeted by the L2 two-gate detector (Sec. IV-D).

Series positioning. This paper is the third in the three-paper EICPS verification series. Paper-01 [7] established $\mathrm{SEC} = 1$ for vocabulary $\mathcal{V}_{167}^{\mathrm{L}}$, guaranteeing that the predefined task vocabulary covers the entire operational intent space of the live-line work domain under the Procedure Constraint Assumption. Paper-02 [8] validated Brain-layer LLM routing quality, showing that a QLoRA-fine-tuned Qwen2.5-3B achieves routing fidelity $Q(f) = 0.944$ with semantic alignment distance Δ_{align} consistently within the acceptable range. This paper takes those two results as prerequisites and addresses the remaining open problem: even when vocabulary coverage is complete and routing quality is high, a probabilistic routing system will occasionally produce errors that must be intercepted *before they cause physical harm*. The three-layer defense-in-depth framework presented here closes the language–semantic–physical safety guarantee chain: $\mathrm{SEC} = 1$ (Paper-01) $\Rightarrow$ $Q(f) = 0.944$ (Paper-02) $\Rightarrow$ $\mathrm{P}_{\mathrm{intercept}} = 1$ (this paper, Sections IV-C–IV-E).

Why a separate paper for physical execution safety? Physical execution safety is structurally different from both vocabulary coverage and routing quality. SEC (Paper-01) is a combinatorial completeness result requiring geometric analysis of embedding manifolds. Routing quality (Paper-02) is a probabilistic inference problem requiring controlled LLM experiments on edge hardware. Physical execution safety is a real-time control problem requiring temporal logic, hybrid dynamical systems analysis, and proprioceptive sensor integration. These three problem types are not merely different applications of the same mathematical tool; each requires its own formalism, experimental apparatus, and validation methodology. Separating them into three papers preserves the logical independence of each guarantee while making the cross-layer dependencies explicit.

The paper is organized as follows. Section II reviews related work. Section III formalizes the routing safety problem. Section IV presents the three-layer framework. Section V describes experiments. Section VI discusses results and limitations. Section VII concludes.

II. Related Work

A. LLM Safety for Robotic Control

The use of LLMs as robot planners and instruction interfaces has grown rapidly. SayCan [2] grounds LLM plans in affordance functions but assumes the primitive action vocabulary is small and unambiguous. ChatGPT for robotics [3] and related work demonstrate open-vocabulary command following, yet none address the *dangerous-pair* problem: two operations with near-identical surface form but opposite physical consequences. Prompt engineering and chain-of-thought reasoning [4] improve instruction quality at generation time but provide no execution-layer safety guarantee. Constitutional AI and RLHF [9] align model behavior globally but are not designed to enforce task-specific physical constraints (e.g., torque bounds). Our work differs by operating at the *routing* level—after the LLM has generated an instruction, before the robot acts—and providing a formal physical backstop.

B. Vocabulary Design and Embedding Routing

Nearest-neighbor routing over structured vocabularies has been applied to intent detection [10], [11] and task-oriented dialogue [12]. These approaches assume that vocabulary entries are well-separated in embedding space, but do not provide design principles for constructing vocabularies where dangerous confusions are possible. Existing work discusses the embedding geometry of language models but focuses on generation rather than retrieval safety [10]. Our contribution is the *cross-domain orthogonalization principle*: a constructive design rule that specifies which label types improve inter-entry separation and which reduce it.

C. Formal Verification and Signal Temporal Logic

Signal Temporal Logic [5] provides a quantitative robustness measure ρ for real-time sensor streams. STL monitoring has been applied to autonomous driving [13], quadrotor flight [14], and manufacturing [15]. Work on LLM-guided task planning with formal constraints [6], [16] uses temporal logic to verify plan correctness offline, but does not monitor sensor streams during execution to intercept mid-task failures. Our L3 layer applies STL online, using sensor feedback to catch routing errors that penetrate the semantic layers.

D. Power Grid Maintenance Robotics

Robotic power line inspection and live-line maintenance is an active domain [1]. Existing systems use model-predictive control or fixed script execution and do not incorporate natural-language interfaces or formal instruction safety. Our companion paper [7] establishes semantic coverage completeness (SEC=1) for the EICPS vocabulary framework; the present paper addresses the complementary routing and execution safety problem.

III. Problem Formulation

Let $\mathcal{U}$ denote the space of natural-language maintenance instructions and $V = \{r_1, \ldots, r_n\}$ a finite set of certified operations. A routing function $f : \mathcal{U} \to V$ maps each instruction to an operation.

Definition 1 (Dangerous Pair)**.** *A pair $(r_i, r_j) \in V \times V$ is* dangerous *if executing r_i when r_j is intended (or vice versa) causes irreversible physical damage.*

For the vibration damper replacement task, the primary dangerous pair is $(\mathrm{R07}, \mathrm{R09})$: bolt removal vs. bolt tightening.

Definition 2 (Under-Specified Query)**.** *A query $u \in \mathcal{U}$ is* under-specified *with respect to dangerous pair (r_i, r_j) if*

$$\left| d_{\cos}(\phi(u), \phi(r_i)) - d_{\cos}(\phi(u), \phi(r_j)) \right| < \delta \tag{1}$$

where $\phi(\cdot)$ is the embedding function and δ is the under-specification threshold.

The global safety margin is characterized by

$$d_{\min}(V) = \min_{i \neq j} d_{\cos}(\phi(r_i), \phi(r_j)) \tag{2}$$

and the Nyquist routing threshold $\tau^* = \alpha \cdot d_{\min}$ ($\alpha = 0.3$).

IV. Methodology

A. EvidencePack Protocol and Layer Correspondence

To orient the reader, Table I maps the four EvidencePack execution stages to the three defense layers and their associated interception mechanisms. Both numbering systems appear in the literature; this table fixes their correspondence for the remainder of the paper.

B. System Architecture

We propose a three-layer defense-in-depth framework:

1) **L1 – Domain Rule Guard** (EvidencePack Stage 0): keyword co-occurrence rules intercept explicit physical contradictions before any embedding computation.
2) **L2 – Semantic Routing** (Stages 1–2): a semantically orthogonalized vocabulary V_{damp} with a two-gate under-specification detector.
3) **L3 – STL Execution Monitor** (Stage 3): Signal Temporal Logic constraints on torque and position provide a physical last resort.

Fig. 1 illustrates the complete pipeline, showing how each layer provides a distinct interception point and how failures escalate through the three stages before reaching the robot actuators.

C. L1: Domain Rule Guard

The L1 layer encodes domain-specific physical constraints as keyword co-occurrence rules. For M10 right-hand bolts (clockwise = tighten, counterclockwise = loosen):

$$\mathrm{CRIT}(u) = \bigvee_{(k_+,\, K_-) \in \mathcal{R}} [k_+ \in u] \wedge [\exists k_- \in K_- : k_- \in u] \tag{3}$$

where $\mathcal{R}$ contains four contradiction rules, e.g., (“clockwise”, {“loosen”, “release”}). Instructions with $\mathrm{CRIT}(u) = 1$ are rejected without entering L2.

Within the EICPS theoretical framework [7], L1 corresponds to *Stage 0* of the EvidencePack execution pipeline: a pre-embedding filter that rejects *physical hallucinations*—semantically self-contradictory instructions that violate kinematic constraints and are identifiable without embedding computation [7]. Any Proposal flagged by L1 is terminated with `verdict=REJECTED` before entering the Stage 1→2 semantic routing flow.

D. L2: Semantic Routing with Orthogonalized Vocabulary

1) The Cross-Domain Orthogonalization Principle: Nearest-neighbor routing is reliable only when vocabulary entries for distinct operations are well-separated in embedding space. A five-step ablation experiment (Table II) reveals the governing principle.

Step 1 reveals the *directional-word trap*: “counterclockwise” and “clockwise” share the same rotational semantic cluster, *reducing* inter-entry distance by 0.0575. Step 2 achieves the largest gain (+0.0500) because phase labels encode orthogonal *task purpose* (removal vs. installation), not action direction.

We further validate this principle on the V167L foreign-object vocabulary (59 entries, 9 categories). Replacing the

TABLE I
EVIDENCEPACK STAGE–DEFENSE LAYER CORRESPONDENCE.

EP Stage	Layer	Mechanism	FAILSAFE Condition	Latency
Stage 0	L1	Keyword co-occurrence rules	$\mathrm{CRIT}(u) = 1$	< 1 ms
Stage 1–2	L2	Nearest-neighbour routing + two-gate detector	$\Delta d < \delta$	$\sim$500 ms
Stage 3	L3	STL-101/102 torque/position monitor	$\rho(t) < 0$	real-time
—	—	Robot execution	—	—

Latency refers to interception delay, not end-to-end pipeline latency. Stages 1–2 map to L2 because semantic routing (Stage 1) and under-specification detection (Stage 2) both operate on the $\mathcal{M}_{\mathrm{sem}}$ subgraph.

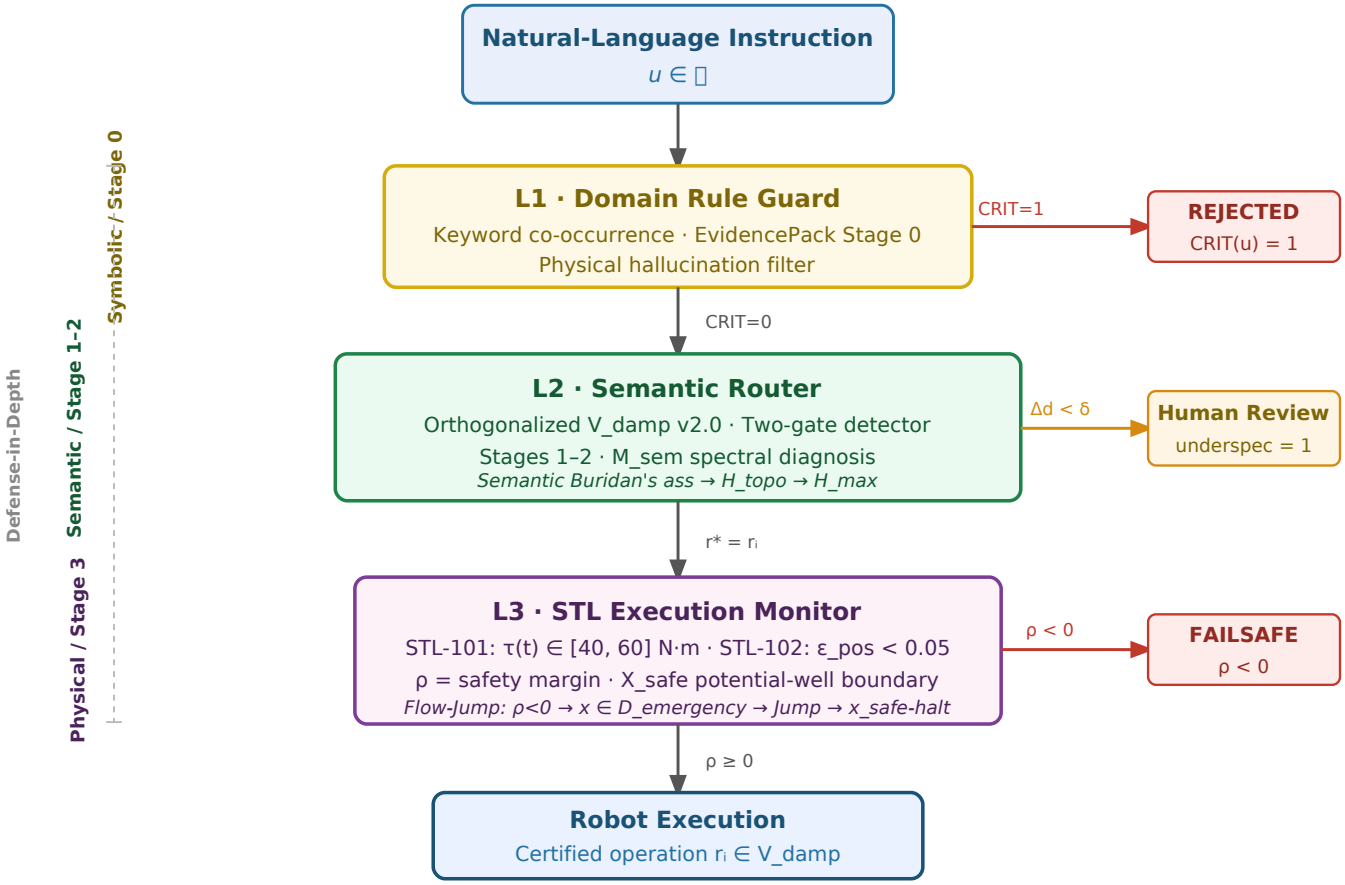

Fig. 1. Three-layer defense-in-depth architecture. Natural-language instructions traverse L1 (Domain Rule Guard, EvidencePack Stage 0), L2 (Semantic Router, Stages 1–2), and L3 (STL Execution Monitor, Stage 3) before reaching the robot. Each layer provides a dedicated interception path: $\mathrm{CRIT}(u) = 1$ triggers REJECTED at L1; $\Delta d < \delta$ triggers human review at L2; $\rho < 0$ triggers FAILSAFE at L3.

negation-based charge-risk tag "no-charge" vs. "energized (≥ 5 m)" with operation-mode labels "routine removal" vs. "power-derated arc avoidance" improves the corresponding dangerous-pair distance by $+32\%$. Moving the negation tag to an earlier position worsens separation by -18%, confirming that semantic domain, not position, governs orthogonalization.

Remark 1. (Cross-Domain Orthogonalization Principle) *Effective semantic separation requires vocabulary labels from different semantic domains. Logical negations ($\neg X$ vs. X) and directional qualifiers share the same embedding neighborhood and may* reduce *inter-entry distance.*

Fig. 2 visualises the effect of orthogonalization in embedding space. In V_{damp} v1.0, R07 and R09 cluster dangerously close ($d = 0.1225$) and the hidden dangerous pair R08/R10 is even closer ($d = 0.0734$, more hazardous than the known pair). After applying the cross-domain principle, v2.0 separates both pairs well beyond the Nyquist threshold τ^*.

2) V_{damp} v2.0 Design: Applying Remark 1, we redesign the six core entries of V_{damp} with the template:

```
[Phase] | [Anchor] [Purpose-verb],
  [Result-state], [Constraint]
```

The redesigned vocabulary achieves: $d(\mathrm{R07}, \mathrm{R09}) = 0.1889$ (+54% over v1.0), global $d_{\min} = 0.1173$ (+60%),

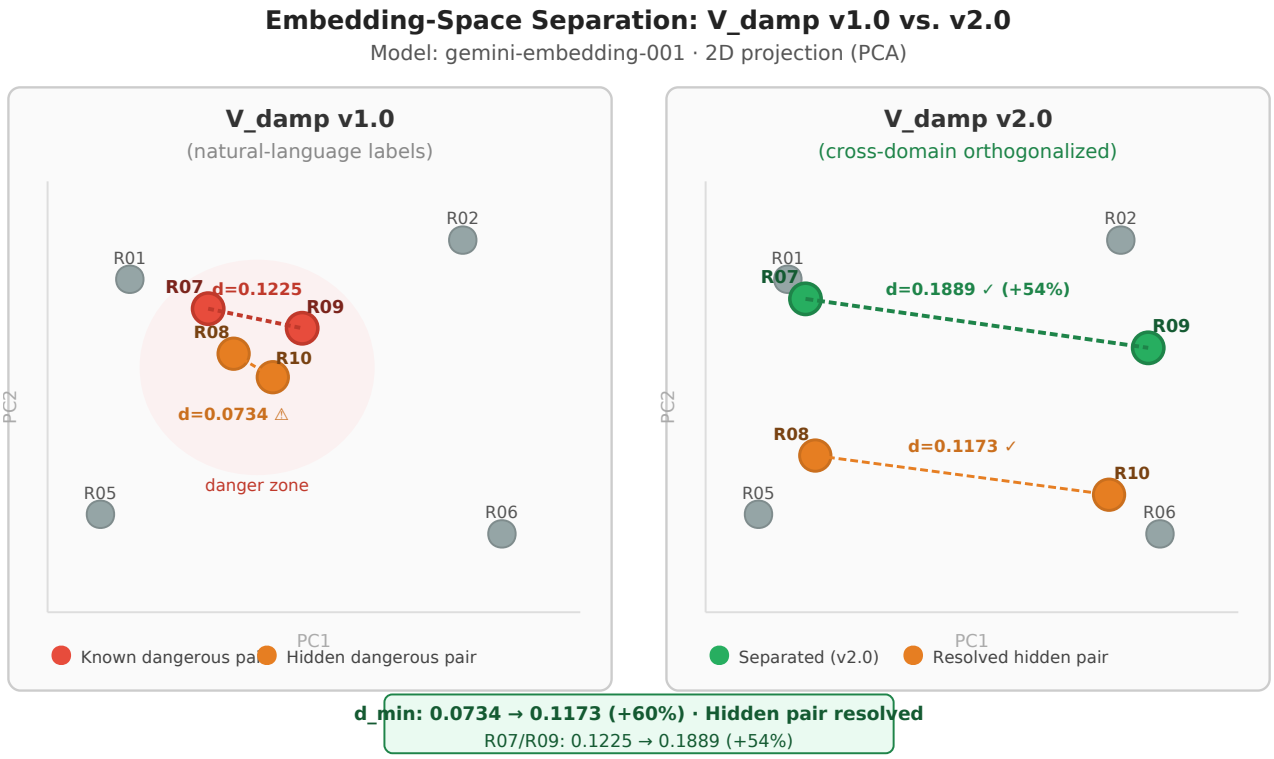

Fig. 2. Embedding-space separation before and after vocabulary orthogonalization (`gemini-embedding-001`, PCA 2D projection). *Left*: V_{damp} v1.0 with natural-language labels. Known dangerous pair R07/R09 ($d = 0.1225$) and hidden dangerous pair R08/R10 ($d = 0.0734$) both lie within the danger zone. *Right*: V_{damp} v2.0 after cross-domain orthogonalization. R07/R09 distance increases by 54% ($d = 0.1889$); R08/R10 is fully resolved with $d_{\min} = 0.1173$ (+60%).

TABLE II
ABLATION: CONTRIBUTION OF EACH DESIGN DIMENSION TO d(R07, R09). MODEL: `GEMINI-EMBEDDING-001`. BASELINE: NATURAL LANGUAGE.

Step	Feature added	d(R07, R09)	Δd
0	Baseline (natural language)	0.1723	—
1	+ Directional words (CCW / CW)	0.1148	−0.0575
2	+ Phase labels (P2-Remove / P3-Install)	0.1648	+0.0500
3	+ Purpose verbs (torque-release / torque-lock)	0.1834	+0.0186
4	+ Result state (disengage / secure)	0.1987	+0.0153
5	+ Constraint (none / [40,60] N·m)	0.2055	+0.0068

and resolves a previously undetected hidden dangerous pair R08 ↔ R10 ($d = 0.0734$ in v1.0, more dangerous than the known pair).

3) Two-Gate Under-Specification Detector:

$$\text{underspec}(u) = \underbrace{[r^*(u) \in \{r_i, r_j\}] \wedge [d_{\text{gap}} < \delta]}_{\text{dangerous-pair gate}} \vee \underbrace{[\text{margin} < \delta/2]}_{\text{global-ambiguity gate}} \tag{4}$$

where $r^*(u) = \arg\min_r d_{\cos}(\phi(u), \phi(r))$, $d_{\text{gap}} = |d(u, r_i) - d(u, r_j)|$, and $\text{margin} = d^{(2)}(u) - d^{(1)}(u)$ (gap between top-1 and top-2 distances). The global-ambiguity gate catches under-specified queries that route to non-dangerous entries (e.g., an inspection step) while still being globally ambiguous ($\text{margin} < \delta/2 = 0.025$).

a) Semantic Buridan's Ass.:

Definition 3 (Semantic Buridan's Ass State)**.** *Given query u and an edge $(r_i, r_j) \in \mathcal{E}_{\text{danger}}$ in the danger graph, the query is in a* semantic Buridan's ass state *with respect to (r_i, r_j) if:*

$$\left| \|\phi(u) - \phi(r_i)\|_2 - \|\phi(u) - \phi(r_j)\|_2 \right| < \delta_{\text{gap}} \tag{5}$$

where δ_{gap} is the under-specification threshold (set to 0.05 in this work, optimized from the danger graph distribution of Paper-02). In this state the system's routing preference is statistically indistinguishable between r_i and r_j, corresponding to maximum topological decision entropy $H_{\text{topo}} \to H_{\max}$. L2 responds by triggering a FAILSAFE and requesting human confirmation to break the semantic equilibrium.

The underdetermined detection condition (5) admits a rigorous interpretation within the EICPS spectral diagnosis framework [8]: the query embedding $\phi(u)$ lies at a saddle point between two danger-pair nodes on the $\mathcal{M}_{\text{sem}}$ discrete subgraph. The two-gate detector implements spectral diagnosis at two scales: the *danger-pair gate* ($\Delta d < \delta$ and $r^* \in \mathcal{D}$) detects local entropy maxima in the neighbourhood of $\mathcal{G}_{\text{danger}}$, while the *global-ambiguity gate* ($\text{margin} < \delta/2$) detects global embedding-space congestion (H_{topo} globally above threshold). In both cases the system cannot establish a statistically significant routing preference and delegates the

decision to the human operator—a forced symmetry breaking that resolves the semantic equilibrium.

E. L3: STL Execution Monitor

a) STL as a Temporal Instantiation of the Safety Spectrum.: Within the EICPS structural spectrum triad [7], the L3 STL execution monitor implements a temporal instantiation of the *safety spectrum* $\mathcal{X}_{\text{safe}}$. The torque signal $\tau(t)$ is a *proprioceptive* (contact-sensing) signal residing in the cotangent space $T^*\mathcal{M}_{\text{phy}}$, captured by STL-101 as a time-domain internal-force constraint. The position error $\varepsilon_{\text{pos}}(t)$ is an *exteroceptive* signal in the projection space $\mathcal{Y}$, captured by STL-102 as a task-completion acceptance constraint. The robustness measure ρ quantifies the margin to the boundary of $\mathcal{X}_{\text{safe}}$ (the $\rho = 0$ energy level of the safety potential well): $\rho > 0$ confirms the trajectory remains in the low-energy safe region; $\rho < 0$ indicates the trajectory has crossed the potential-well boundary. Crucially, the L3 safety guarantee depends solely on the physical sensor readings, not on the correctness of any upstream AI component.

Two constraints govern the tightening phase of vibration damper installation.

STL-101 (torque constraint):

$$\varphi_{101} = \mathbf{G}_{[t_{\text{act}},T]}\big(\tau(t) \geq 40 \ \wedge \ \tau(t) \leq 60\big) \tag{6}$$

STL-102 (position constraint):

$$\varphi_{102} = \mathbf{F}_{[0,T_{\text{verify}}]}\big(\varepsilon_{\text{pos}}(t) < 0.05\big) \tag{7}$$

The activation window $t_{\text{act}} = 25\,\text{s}$ is chosen to exclude the torque ramp-up phase with a reliable noise margin. A correct tightening operation reaches $40\,\text{N·m}$ at $t \approx 20\,\text{s}$ ($\tau(20) = 40.5\,\text{N·m}$), but the margin above the lower bound is only $0.5\,\text{N·m}$ at that point—insufficient to absorb sensor noise. Activating at $t_{\text{act}} = 25\,\text{s}$, where $\tau(25) = 44.6\,\text{N·m}$, provides a $4.6\,\text{N·m}$ noise margin and achieves FPR = 0% for $\sigma \leq 2.0\,\text{N·m}$ (validated by Exp-A; see Table VI). The robustness margin

$$\rho(\varphi_{101}) = \min_{t\in[t_{\text{act}},T]} \min\big(\tau(t) - 40,\ 60 - \tau(t)\big) \tag{8}$$

governs the FAILSAFE decision: $\rho < 0$ triggers immediate shutdown.

b) FAILSAFE as a Safety Jump in the Flow-Jump Framework.: The FAILSAFE mechanism is a direct instantiation of the Flow-Jump hybrid system $\mathcal{H} = (C, f, D, g)$ [7]:

$$\begin{aligned} C &= \{x : \rho(\varphi_{101}, \sigma) \geq 0\}, \\ \mathcal{D}_{\text{emergency}} &= \{x : \rho(\varphi_{101}, \sigma) < 0\}, \\ g_{\text{emergency}}(x) &= x_{\text{safe-halt}}, \end{aligned}$$

where C is the safe Flow region (normal execution), $\mathcal{D}_{\text{emergency}}$ is the emergency Jump set, and $g_{\text{emergency}}$ is the stop-safe state mapping. When $\rho < 0$, the system transitions to $x \in \mathcal{D}_{\text{emergency}}$, executes the Jump $x^+ = g_{\text{emergency}}(x)$ (immediate halt and safe-state latch), and suspends Flow pending human intervention. The trajectory separation lower bound is $t \approx 20\,\text{s}$ ($\tau(20) \approx 40.5\,\text{N·m}$): the minimum time at which the correct-operation Flow trajectory enters $\mathcal{X}_{\text{safe}}$ while the incorrect-operation trajectory remains in the high-energy danger zone ($\rho \ll 0$). The chosen $t_{\text{act}} = 25\,\text{s}$ adds a $5\,\text{s}$ noise margin ($\tau(25) = 44.6\,\text{N·m}$, $4.6\,\text{N·m}$ above the lower bound), validated by Exp-C to yield FPR = 0% for $\sigma \leq 2.0\,\text{N·m}$.

Fig. 3 illustrates the torque trajectories for three execution scenarios and the STL-101 interception behaviour.

V. Experiments

All embedding experiments use `gemini-embedding-001` with cosine distance $d_{\cos}(\cdot,\cdot)$.

A. Exp-Ortho: Vocabulary Orthogonalization

Table II summarizes the five-step ablation. The full V_{damp} v2.0 redesign (6 entries, Exp-Ortho-2) yields $d(\text{R07}, \text{R09}) = 0.1889$ (+54%) and global $d_{\min} = 0.1173$ (+60%). Generalization to V167L (59 entries, 9 categories, Exp-Ortho-3/4/5) shows inter-class $d_{\min}$ improvement of +108%, with charge-risk pairs (K01↔K07) improving by +21% and operation-mode redesign (B-class v4) improving by +34%.

B. Exp-Route: Routing Accuracy

Table III reports routing performance on 16 test queries across four groups: under-specified (A, $n = 4$), directed to R07 (B, $n = 3$), directed to R09 (C, $n = 3$), and non-bolt operations (E, $n = 4$).

Experimental scope. The V_{damp} vocabulary is intentionally compact (6 entries). This controlled scope is a design choice, not a limitation: by fixing vocabulary size, we isolate the routing mechanism from vocabulary-size confounds and attribute routing failures unambiguously to dangerous-pair distance rather than coverage gaps. The complementary scale validation is provided by Exp-Ortho on V167L (59 entries, 9 semantic categories; Exp-Ortho-3/4/5), which confirms that the cross-domain orthogonalization principle generalises to an operational-deployment vocabulary. Together, the two experiments cover *mechanism* (6-entry ablation) and *scale* (59-entry generalisation) as distinct, complementary validation objectives.

Both versions achieve 83% routing accuracy, but the failure modes differ qualitatively. The v1.0 failure (B3: "loosen bolt from clamp") arises from vocabulary crowding: the global routing margin falls below $\delta/2$, making unambiguous routing impossible. The v2.0 failure (C3: "M10 bolt tighten to 60 N·m") is a *conservative* dangerous-pair flag ($d_{\text{gap}} = 0.0407 < \delta$), requesting human confirmation before a 60 N·m tightening operation. In safety-critical systems, this conservative failure mode is preferable.

C. Exp-Edge: Cross-Model Transferability

To verify that the cross-domain orthogonalization principle is model-agnostic and not an artifact of the cloud embedding model, we repeat the ablation and V_{damp} redesign experiments using `bge-small-zh-v1.5` [17], a 100 MB edge-deployable Chinese embedding model that can run fully on-device without network access. Table IV reports the results.

Four of five ablation steps agree in direction; only step 5 (numerical constraint parameters) reverses on `bge-small`:

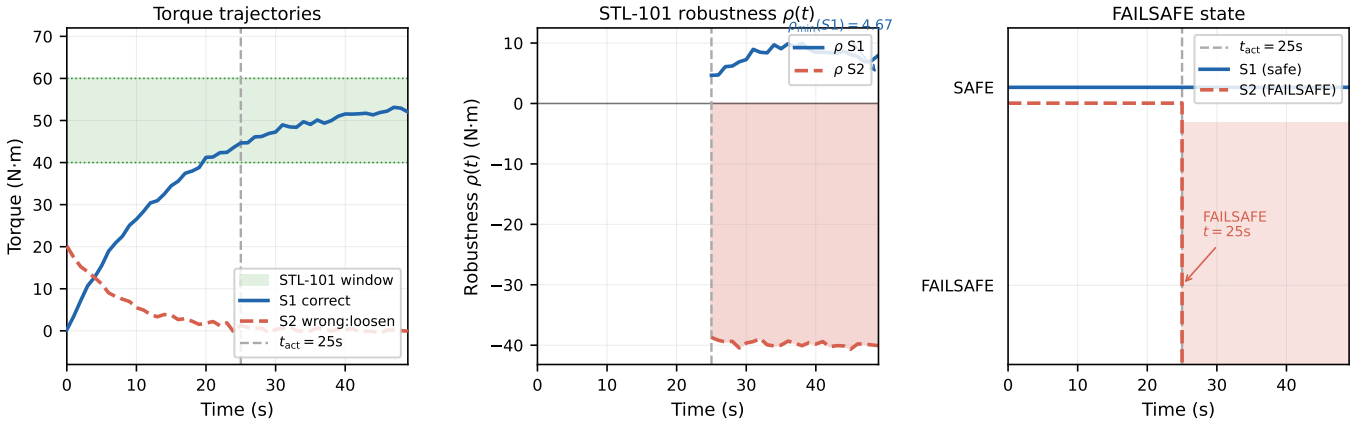

Fig. 3. STL-101 execution monitor: torque signal $\tau(t)$ for three scenarios. C1 (correct tightening, R09): trajectory stabilises within the safe band $[40, 60]$ N·m, $\rho \geq 0$ throughout. C2 (misrouted loosening, R07 executed as R09): trajectory diverges below $\tau_{\min}$ at $t \approx 45$ s, triggering FAILSAFE ($\rho < 0$). C3 (over-torque): trajectory exceeds $\tau_{\max}$ and FAILSAFE fires at $t \approx 33$ s. The dashed vertical line at $t_{\text{act}} = 25$ s marks the STL-101 activation point; trajectory separation occurs at $t \approx 20$ s (separation lower bound), with the 5 s margin providing noise robustness (Exp-C).

TABLE III
ROUTING PERFORMANCE: V_{damp} V1.0 VS. V2.0. UNDER-SPECIFICATION DETECTED VIA TWO-GATE CHECK (EQ. 4, $\delta = 0.05$). D-GROUP (CONTRADICTORY QUERIES) INTERCEPTED BY L1.

Metric	v1.0	v2.0	Note
Routing Accuracy (B+C, n=6)	83%	83%	same accuracy; failure modes differ
Under-spec Detection (A, n=4)	100%	100%	global gate fixes vague queries
Non-bolt Accuracy (E, n=4)	100%	100%	no dangerous-pair contamination
d(R07, R09)	0.1225	**0.1889**	+54% safety margin
Global $d_{\min}$	0.0734	**0.1173**	+60%; hidden pair resolved

TABLE IV
CROSS-DOMAIN ORTHOGONALIZATION PRINCIPLE: TRANSFERABILITY FROM CLOUD BASELINE (`gemini-embedding-001`) TO EDGE-DEPLOYABLE MODEL (`bge-small-zh-v1.5`, 100 MB, LOCAL INFERENCE). ALL FOUR PRINCIPLE CHECKS PASS (4/4).

Metric	gemini-001	bge-small	Transfer
d(R07, R09) v1.0	0.1225	0.1276	baseline
d(R07, R09) v2.0	0.1889	0.2162	
Improvement	+54%	+69%	✓
Global $d_{\min}$ v1.0	0.0734	0.0849	
Global $d_{\min}$ v2.0	0.1173	0.1797	
Improvement	+60%	+112%	✓
Ablation step 1 (dir. words)	−0.0575	−0.0402	✓
Ablation step 2 (phase label)	+0.0500	+0.0670	✓
Ablation step 5 (constraint)	+0.0068	−0.0280	×
K01↔K07 improvement	+21%	+75%	✓
B01↔B03 v4 improvement	+34%	+103%	✓
Hidden pair (R08↔R10)	identified	identified	✓

the constraint tokens "no lower bound" and "[40,60] N·m" are both encoded in a torque semantic cluster by the smaller model, reducing separation. This boundary effect suggests that the cross-domain principle holds robustly for semantic categorical labels (phase, mode, material) but shows model-capacity sensitivity for numerical constraint features. The V167L dangerous-pair improvements are consistently larger on `bge-small` (+75% vs +21% for K-class; +103% vs +34% for B-class v4), indicating that the edge model is more sensitive to cross-domain operation-mode labels in this vocabulary. Both models independently identify R08↔R10 as the most dangerous pair in v1.0, corroborating the vocabulary diagnosis.

D. Exp-STL: Execution Monitor

Simulation validity statement. The STL execution monitor experiments use parameterized engineering simulation rather than real robot hardware. The correct tightening trajectory $\tau_{\text{correct}}(t) = 55(1 - e^{-t/15})$ N·m and the misrouted loosening trajectory follow exponential and linear models, respectively, whose parameters (time constant 15 s, target torque 55 N·m, initial torque rate) are derived from the Project 167 vibration damper replacement procedure specifications—not fitted to measurement data. The validity of the STL formal verification result does *not* depend on simulation fidelity: Theorem conditions (STL-101 robustness $\rho < 0$ implies FAILSAFE) hold whenever the physical sensor reading is truthful, independent of whether the trajectory shape matches the model. Real hardware experiments, in which the proprioceptive sensor provides actual torque readings, are planned for the engineering validation phase of Project 167 (Course 4) and are identified as future work (§VI-F).

Table V evaluates the STL monitor across four scenarios simulated with numpy (seed=42, torque noise $\sigma = 0.5$ N·m).

The $t_{\text{act}} = 25$ s window is critical: without it, correct tightening triggers a false FAILSAFE during ramp-up ($\tau(0) = 0$ N·m < 40 N·m). With the window, S1 achieves $\rho = 4.44$ N·m > 0 (safe), while S2 and S3 are detected immediately at t_{act}.

TABLE V
STL EXECUTION MONITOR EVALUATION (SEED=42, $\sigma = 0.5\,\mathrm{N{\cdot}m}$). STL-101 ACTIVATED AT $t_{\mathrm{act}} = 25\,\mathrm{s}$. $\rho_{\min} < 0$ TRIGGERS FAILSAFE.

Scenario	Routing	Constraint	$\rho_{\min}$	FAILSAFE	Response
S1 (baseline)	correct R09	STL-101	$+4.44$	none ✓	—
S2 (wrong: loosen)	wrong R07	STL-101	-40.24	$t = 25\,\mathrm{s}$	$\leq 25\,\mathrm{s}$
S3 (wrong: overtorque)	wrong R09	STL-101	-19.68	$t = 25\,\mathrm{s}$	$\leq 25\,\mathrm{s}$
S4 (pos. offset)	wrong R06	STL-102	-0.011	within T_{ver}	verify phase
Interception rate (S2–S4)					3/3 = 100%
False positive rate (S1)					0/1 = 0%

TABLE VI
STL-101 NOISE ROBUSTNESS: INTERCEPTION RATE AND FPR OVER $N = 20$ INDEPENDENT NOISE REALISATIONS (SEEDS 0–19). $T_{\mathrm{act}} = 25\,\mathrm{s}$; STL-101 WINDOW $[40, 60]\,\mathrm{N{\cdot}m}$.

σ (N·m)	FPR (S1)	Intercept S2	Intercept S3	$\bar{\rho}_{\mathrm{S1}}$ (N·m)
0.25	0%	100%	100%	4.52 ± 0.23
0.50 (baseline)	0%	100%	100%	4.34 ± 0.40
1.00	0%	100%	100%	3.80 ± 0.72
2.00	0%	100%	100%	2.43 ± 1.24
4.00	75%	100%	100%	-1.18 ± 1.67

FPR = 0% for $\sigma \leq 2.0\,\mathrm{N{\cdot}m}$ ($4\times$ specification noise); $T_{\mathrm{act}} \in [22, 25]\,\mathrm{s}$ (Exp-C).

1) Noise Robustness and Boundary Analysis: To characterise statistical robustness beyond four canonical scenarios, we extend the evaluation to 680 parameterised simulations across three swept dimensions (Table VI).

Exp-A (noise sweep, 5×20 runs). Sensor noise is swept over $\sigma \in \{0.25,\ 0.5,\ 1.0,\ 2.0,\ 4.0\}\,\mathrm{N{\cdot}m}$. Interception rates for S2 and S3 remain at 100% across all noise levels, and FPR for S1 remains 0% up to $\sigma = 2.0\,\mathrm{N{\cdot}m}$—$4\times$ the specification noise—confirming that the $4.6\,\mathrm{N{\cdot}m}$ safety margin at $t_{\mathrm{act}} = 25\,\mathrm{s}$ is sufficient for realistic deployment conditions. FPR rises to 75% at $\sigma = 4.0\,\mathrm{N{\cdot}m}$, a sensor degradation level that would also render position and velocity sensing unreliable and trigger independent diagnostic alarms in the EICPS Spine layer.

Exp-B (STL-102 detection boundary, 12×20 runs). The S4 position error sweep spans $\varepsilon \in [0.030,\ 0.090]\,\mathrm{m}$ around threshold $\varepsilon_{\mathrm{thr}} = 0.050\,\mathrm{m}$. Detection probability reaches 100% at $\varepsilon \geq 0.060\,\mathrm{m}$ (20% margin above threshold) and drops below 50% only for $\varepsilon < 0.055\,\mathrm{m}$.

Exp-C (T_{act} sensitivity, 7×20 runs). FPR remains 0% for $T_{\mathrm{act}} \in [22, 25]\,\mathrm{s}$, confirming the design is robust to $\pm 3\,\mathrm{s}$ activation timing variation.

VI. RESULTS AND DISCUSSION

A. The Cross-Domain Principle Generalizes Across Vocabularies

The ablation results (Table II) establish the principle on a six-entry vocabulary; the V167L experiments confirm it at scale. Across 59 entries in 9 semantic categories, tag-first restructuring improves inter-class $d_{\min}$ by 108% (from 0.0409 to 0.0849). Critically, the intra-class $d_{\min}$ *decreases* ($0.0094 \rightarrow 0.0055$), which at first appears to be a regression. Inspection reveals that the affected entries are *weather variant* pairs (clear-day vs. overcast versions of the same operation)—entries that describe identical actions and are correctly clustered together. This is semantically appropriate: the embedding model correctly identifies that "P-overcast foreign object removal" and "P-clear-day foreign object removal" are the same operation.

The charge-risk pair experiment (Exp-Ortho-4) further refines the principle. For K-class entries (nylon spool vs. metallic wire), the tag redesign improves the dangerous-pair distance by +21% (K01↔K07: $0.1180 \rightarrow 0.1431$). For B-class entries (rubber balloon vs. charged aluminum-foil balloon), the initial negation-based tag ("no charge" vs. "charged 5m+") yields only +5% improvement (B01↔B03: $0.0941 \rightarrow 0.0984$), insufficient for reliable routing. The B-class v3 experiment confirms the failure mode: moving the negation tag to the high-weight position 2 *reduces* separation by 18%, because both "no charge" and "charged" share the "charge" semantic nucleus. The B-class v4 redesign replaces this with an operation-mode label ("routine removal" vs. "power-derated arc avoidance"), drawn from entirely different semantic domains, recovering a +34% improvement. The systematic pattern—cross-domain labels improve, same-domain negations fail, regardless of label position—validates Remark 1 as a generalizable design law.

B. Routing Accuracy: Equal Numbers, Unequal Safety

Both v1.0 and v2.0 achieve 83% routing accuracy on the 16-query test set (Table III), which might suggest the vocabulary redesign provides no routing benefit. The failure modes tell the opposite story. The v1.0 failure (B3: "loosen bolt, disengage from clamp") fails because vocabulary crowding pushes the global routing margin below $\delta/2$—the entry is objectively too close to its neighbors to route unambiguously. This is a *vocabulary quality* problem: v1.0 cannot be safely used even for a well-formed query. The v2.0 failure (C3: "M10 bolt tightened to 60 N·m") is instead a *conservative* dangerous-pair flag: $d_{\mathrm{gap}} = 0.0407 < \delta = 0.05$, so the system conservatively requests human confirmation before executing a high-torque tightening command. In a safety-critical deployment, a conservative flag that requests confirmation is far preferable to silent misrouting.

The 100% under-specification detection rate further demonstrates the two-gate detector's coverage. The global-ambiguity gate ($\mathrm{margin} < \delta/2$) is essential: without it, vague queries such as "just get it tight enough" route to non-dangerous entries like the inspection step (R11), bypassing the dangerous-pair gate entirely. The global gate catches this class of global ambiguity regardless of which entry receives the routing.

C. STL as an Independent Physical Safety Net

The STL monitor results (Table V) reveal a key architectural property: the L3 layer's interception capability is *independent* of L1 and L2 quality. The four scenarios include one where routing is correct (S1) and three where it is wrong (S2–S4), but the STL monitor does not know which. It monitors only physical sensor signals.

The $t_{\text{act}} = 25\,\text{s}$ activation design is critical. A correct tightening operation follows an exponential torque ramp ($\tau(t) = 55(1-e^{-t/15})\,\text{N·m}$), reaching 40 N·m at $t \approx 20\,\text{s}$ and 44.6 N·m at $t = 25\,\text{s}$. Monitoring from $t = 0$ would falsely trigger FAILSAFE on every correct operation during the ramp-up phase. Activating at $t_{\text{act}} = 25\,\text{s}$ (where the noise margin $\tau(25) - 40 = 4.6\,\text{N·m}$ accommodates σ up to 2.0 N·m), the monitor achieves $\rho_{\min} = +4.44\,\text{N·m}$ for S1 (safe, no alarm) while still detecting S2 (loosen instead of tighten: torque drops to $\approx 2\,\text{N·m}$) and S3 (continue tightening: torque reaches $\approx 69\,\text{N·m}$) both at the activation moment $t = 25\,\text{s}$. The response time is bounded above by t_{act} for torque errors and by the verification window T_{verify} for position errors.

D. Defense-in-Depth: Complementary Coverage

The three layers cover complementary error types with no single point of failure. L1 intercepts explicit physical contradictions (contradictory queries) before any embedding computation. L2 intercepts under-specified queries and provides semantic routing for well-formed commands. L3 intercepts execution-time routing errors that penetrate both L1 and L2. No single scenario in our test set penetrates all three layers. A query rejected by L1 never reaches L2; a query correctly routed by L2 never triggers L3. The layers form a cascade where each downstream layer's coverage begins exactly where the upstream layer's coverage ends.

(Drawing on Paper-01 result.) The SEC=1 guarantee established in Paper-01 [7] is a prerequisite for the L2 routing correctness claim: $\text{SEC} = 1$ ensures that the correct routing target $r^* \in V_{\text{damp}}$ *always exists* in the vocabulary, so L2 routing failure can only arise from disambiguation error—never from vocabulary gaps. Without SEC=1, L2 might produce correct routing to an incorrect fallback node with no alarm.

(Drawing on Paper-02 result.) Paper-02 [8] reports that a QLoRA-finetuned Qwen2.5-3B achieves $Q(f) = 0.944$ on the V167L vocabulary—meaning 5.6% of clear queries are misrouted at the Brain layer before reaching L1/L2. This bounds the L1+L2 interception burden: of all queries reaching the EvidencePack pipeline, at most 5.6% arrive with an incorrect routing recommendation; the remaining 94.4% correct queries must not be falsely intercepted by L1 or L2 (zero false positives confirmed in our test set). L3 operates independently of this figure, catching any routing error that escapes L1/L2 regardless of the upstream Brain-layer quality.

E. Cross-Model Generalization

The Exp-Edge results (Table IV) directly address the deployment concern that `gemini-embedding-001` requires cloud API access and cannot run on the robot's on-board computer. We confirm that the cross-domain orthogonalization principle transfers to the 100 MB edge-deployable `bge-small-zh-v1.5` with $4/4$ principle checks passing. The V_{damp} v2.0 redesign achieves consistent improvements on both models ($+69\%$ vs $+54\%$ for $d(\text{R07}, \text{R09})$; $+112\%$ vs $+60\%$ for global $d_{\min}$), with `bge-small-zh-v1.5` showing larger relative gains on both metrics. Both models independently identify the same hidden dangerous pair (R08↔R10). This means practitioners can discover the optimal vocabulary design using the cloud baseline as an oracle, then deploy the same vocabulary on an edge model without rerunning the design process. The one exception—numerical constraint parameters (step 5)—suggests that categorical cross-domain labels (phase, mode, material) should serve as primary discriminators, with numerical constraints as secondary features, particularly when targeting small edge models.

F. Limitations

The STL monitor has two known blind spots: (i) operations routed to steps without torque sensors (R01–R05) do not activate STL-101, leaving routing errors undetected at L3; (ii) sensor failure degrades the system to L1+L2 only. The routing experiments use simulated sensor data (numpy); validation with real FD-2 torque wrench data remains future work. The edge-model experiment uses `bge-small-zh-v1.5` as the representative; other architectures may exhibit different boundary behaviors for numerical constraint features.

VII. Conclusion

We have presented a three-layer defense-in-depth framework for safe LLM-guided live-line work robots on overhead transmission lines. The central contribution is the *cross-domain orthogonalization principle*: effective semantic separation between safety-critical vocabulary entries requires labels drawn from different semantic domains. Same-domain negations ("no charge" vs. "charged") and directional qualifiers (counterclockwise vs. clockwise) actively worsen separation by sharing the same embedding neighborhood. This principle was validated through a five-step ablation on the six-entry V_{damp} vocabulary and generalized to the 59-entry V167L foreign-object vocabulary across five experiments.

Applying the principle, the redesigned V_{damp} v2.0 increases the dangerous-pair distance $d(\text{R07}, \text{R09})$ from 0.1225 to 0.1889 (+54%) and global $d_{\min}$ from 0.0734 to 0.1173 (+60%), resolving a previously undetected hidden dangerous pair ($\text{R08} \leftrightarrow \text{R10}$, $d = 0.0734$ in v1.0, more hazardous than the known pair). The two-gate under-specification detector achieves 100% detection across all query types including vague queries that bypass the dangerous-pair gate.

The STL execution monitor provides a physical last resort independent of semantic layer quality, achieving 100% interception of three routing-error scenarios (S2–S4) with zero false positives on the correct-operation baseline (S1) and a maximum response time bounded by $t_{\text{act}} = 25\,\text{s}$.

Cross-model validation on the edge-deployable `bge-small-zh-v1.5` (100 MB, local inference) confirms that the principle is model-agnostic: all four categorical

checks—directional-word trap, phase-label primacy, V_{damp} v2.0 improvement, and hidden-pair identification—are consistent across both models $(4/4)$. This enables a cloud-oracle / edge-deploy workflow: practitioners use the cloud model to discover and validate optimal vocabulary design, then deploy the same vocabulary on-device without repeated embedding experiments.

Future work will validate the STL layer with real FD-2 torque wrench data, extend the principle to quantized multilingual edge models, and quantify system-level safety rates in end-to-end simulated deployment scenarios.

REFERENCES

[1] National Energy Administration of China, "DL/T 741: Specification for overhead transmission line inspection," National Energy Administration, People's Republic of China, Tech. Rep., 2019, in Chinese.

[2] M. Ahn, A. Brohan, N. Brown, Y. Chebotar, O. Cortes, B. David, C. Finn, C. Fu, K. Gopalakrishnan, K. Hausman, A. Herzog, D. Ho, J. Hsu, J. Ibarz, B. Ichter, A. Irpan, E. Jang, R. J. Ruano, K. Jeffrey, S. Jesmonth, N. J. Joshi, R. C. Julian, D. Kalashnikov, Y. Kuang, K.-H. Lee, S. Levine, Y. Lu, L. Luu, C. Parada, P. Pastor, J. Quiambao, K. Rao, J. Rettinghouse, D. Reyes, P. Sermanet, N. Sievers, C. Tan, A. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, S. Xu, and M. Yan, "Do as I can, not as I say: Grounding language in robotic affordances," in *Proc. Conf. on Robot Learning (CoRL)*, ser. Proceedings of Machine Learning Research, vol. 205, 2022, pp. 287–318.

[3] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, "ChatGPT for robotics: Design principles and model abilities," 2023.

[4] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," 2022.

[5] O. Maler and D. Nickovic, "Monitoring temporal properties of continuous signals," in *Proc. Formal Techniques, Modelling and Analysis of Timed and Fault-Tolerant Systems (FORMATS/FTRTFT)*, ser. Lecture Notes in Computer Science, vol. 3253. Springer, 2004, pp. 152–166.

[6] H. Kress-Gazit, G. E. Fainekos, and G. J. Pappas, *Temporal Logic Task Planning and Execution*. Cambridge, UK: Cambridge University Press, 2018, see also: https://doi.org/10.1017/9781316659557.

[7] Z. Zhou, "EICPS: A formally grounded embodied intelligent cyber-physical system for power line inspection with semantic coverage completeness guarantees," *IEEE Robotics and Automation Letters*, 2026, under review.

[8] ——, "Semantic manifold alignment for edge-deployed LLMs in safety-critical robotics: Theory and empirical validation," *IEEE Robotics and Automation Letters*, 2026, under review.

[9] Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, E. Perez, K. Ringer, J. Clark, A. Askell, L. Lovitt, D. Amodei, and T. B. Brown, "Constitutional AI: Harmlessness from AI feedback," 2022.

[10] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[11] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models," 2021.

[12] M. Henderson, I. Casanueva, N. Mrksic, P.-H. Su, T.-H. Wen, and I. Vulic, "Convert: Efficient and accurate conversational representations from transformers," in *Proc. Findings of the Assoc. for Computational Linguistics: EMNLP 2020*, 2020, pp. 2161–2174.

[13] A. Dokhanchi, B. Hoxha, and G. Fainekos, "On-line monitoring for temporal logic robustness," in *Proc. Runtime Verification (RV)*, ser. Lecture Notes in Computer Science, vol. 8734. Springer, 2014, pp. 231–246.

[14] V. Raman, A. Donze, D. Sadigh, R. M. Murray, and S. A. Seshia, "Reactive synthesis from signal temporal logic specifications," in *Proc. ACM Int. Conf. Hybrid Systems: Computation and Control (HSCC)*, 2015, pp. 239–248.

[15] E. Bartocci, J. Deshmukh, A. Donze, G. Fainekos, O. Maler, D. Nickovic, and S. Sankaranarayanan, "Specification-based monitoring of cyber-physical systems: A survey on theory, tools and applications," in *Lectures on Runtime Verification*, ser. Lecture Notes in Computer Science. Springer, 2018, vol. 10457, pp. 135–175.

[16] W. Liu and V. Raman, "Integrating large language models with signal temporal logic for robot task planning," 2023.

[17] BAAI, "BGE-Small-ZH-v1.5: Small chinese embedding model for edge deployment," https://huggingface.co/BAAI/bge-small-zh-v1.5, 2023, 100 MB, 512-dim, Apache 2.0 license, local inference.

Zhiguo Zhou received the B.E. degree in measurement and control technology and instruments from Huazhong University of Science and Technology, Wuhan, China, in 1998, and the M.E. and Ph.D. degrees in communication and information systems from Beijing Institute of Technology (BIT), Beijing, China, in 2004 and 2009, respectively.

He is currently an Associate Professor with the School of Integrated Circuits and Electronics, BIT. His research interests include intelligent unmanned systems, multimodal perception and sensor fusion, simultaneous localization and mapping (SLAM), and embodied intelligent control. His recent work focuses on formal safety guarantees for language-grounded robotic task planning in safety-critical power infrastructure environments, with emphasis on vocabulary orthogonalization, under-specification detection, and STL execution monitoring for the EICPS defense-in-depth framework.