FORTE: Force Optimization via Riemannian Trajectory Estimation for Zero-Shot Contact-Rich Manipulation

Ege Doganay
Dept. of Electrical and Electronics Engineering
Bilkent University
Ankara, Turkey
ege.doganay@ug.bilkent.edu.tr

***Abstract*—Zero-shot vision-language manipulation systems decompose tasks well at the semantic level but degrade on phases whose success criterion is a force rather than a pose. The cost functions of their model-predictive planners treat contact as rigid and position-controlled, so the contact force is whatever the underlying tracking error happens to produce. We propose FORTE (Force Optimization via Riemannian Trajectory Estimation), a framework that grounds the planning cost in physically realisable contact dynamics while keeping the high-level phase plan zero-shot at the semantic level. A vision-language model emits a multi-phase plan with per-phase cost weights, a force band, a contact strategy, and an action prior. A typed per-task validator filters this plan against closed-form scene parameters available at execution time, such as the declared joint stiffness in simulation. A semantic monitor rotates phases on live triggers and re-queries the model on failure. The planner is an Energy-Aware MPPI whose cost augments geometric tracking with an interaction-energy term $\boldsymbol{\delta}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{\delta}$, a force upper bound that prevents jamming, and a contact-maintenance lower bound that prevents the planner from minimising energy by avoiding contact. The stiffness matrix $\hat{\boldsymbol{\Sigma}} \in \mathrm{SPD}(3)$ is identified online by a Riemannian estimator using the affine-invariant exponential map of Pennec, which keeps the estimate positive-definite for any learning rate and produces scale-invariant updates across orders of magnitude in environmental stiffness. We validate the estimator on a spring-button task with closed-form ground-truth stiffness $k \in \{50, 200, 800\}$ N/m. Every run reaches the 20-consecutive-in-band success threshold with a single learning rate, and $\hat{\Sigma}_{\max}(t)$ tracks k in the correct direction at every press across 1.2 orders of magnitude (a $16\times$ range). On two box-and-wall tasks the limiting failure traces to the contact-target representation in the language-model output, not to the estimator or the cost.**

I. Introduction

Vision-language planners read a scene, decompose a manipulation task into phases, and hand the controller a cost. They are accurate at *which face, which gripper, in what order* and brittle at *how hard to press*: their cost is force-blind, so any phase whose success criterion is a sustained interaction force degrades to whatever the underlying tracking error produces. The CoRAL architecture [1] exemplifies this trade-off, succeeding on geometric tasks while degrading on force-regulated phases. Recent VLM-guided variable-impedance methods [2], [3] address the gap reactively by tuning scalar or diagonal stiffness gains; this representation cannot encode the coupled, anisotropic dynamics that arise in real contact (e.g. angular misalignment during insertion produces cross-axis forces a diagonal stiffness cannot represent). End-to-end force-aware policies [5] sidestep the representation problem at the cost of large force-labelled demonstration datasets and the loss of zero-shot scope. Offline Riemannian methods for impedance learning [7], [8] respect the full SPD structure but have not been deployed inside a real-time control loop.

We propose FORTE, a zero-shot framework that closes the gap end-to-end inside the planner. Three research questions structure the work.

- **RQ1**. Can an online stiffness estimator on $\mathrm{SPD}(3)$ identify environmental stiffness at control frequency while keeping the estimate positive-definite by construction?
- **RQ2**. Does augmenting the MPPI cost with an interaction-energy term and explicit force-band barriers regulate contact force within a language-model-specified band, without sacrificing the zero-shot phase plan?
- **RQ3**. When the pipeline fails, is the failure inside the estimator and the cost, or upstream in the contact strategy that the language model produces?

We validate the estimator on a closed-form spring-button task (**RQ1**, **RQ2**) and trace the failures on the harder box-and-wall tasks to the contact-target representation in the language-model output (**RQ3**).

II. Method

FORTE is a closed-loop architecture with four modules: a vision-language perception and planning frontend, a Riemannian stiffness estimator, an Energy-Aware MPPI controller, and a semantic monitor that drives LLM-in-the-loop revision.

A. Riemannian Stiffness Estimator

Let $\boldsymbol{F}_{\text{meas}} \in \mathbb{R}^3$ denote the contact force recovered from joint-torque sensing and $\boldsymbol{\delta} = \boldsymbol{x}_{curr} - \boldsymbol{x}_{eq}$ the penetration in the task frame, with $\boldsymbol{x}_{eq}$ taken as the translation of the object

pose returned by FoundationPose [4]. We minimise the one-step force-prediction error

$$L(\boldsymbol{\Sigma}) = \tfrac{1}{2}\|\boldsymbol{F}_{\text{meas}} - \boldsymbol{\Sigma}\boldsymbol{\delta}\|^2. \tag{1}$$

The Euclidean gradient $\nabla_E L = -\boldsymbol{e}\boldsymbol{\delta}^\top$ with $\boldsymbol{e} = \boldsymbol{F}_{\text{meas}} - \boldsymbol{\Sigma}\boldsymbol{\delta}$ is not symmetric, so a naive Euclidean step leaves the SPD manifold and forfeits the structure needed by the energy term in the planning cost. We symmetrise to $\boldsymbol{G} = \frac{1}{2}(\nabla_E L + \nabla_E L^\top)$ and update via the affine-invariant exponential map on $\mathrm{SPD}(3)$ [6]:

$$\boldsymbol{\Sigma}_{t+1} = \boldsymbol{\Sigma}_t^{1/2}\,\exp\!\big(-\eta\,\boldsymbol{\Sigma}_t^{1/2}\,\boldsymbol{G}\,\boldsymbol{\Sigma}_t^{1/2}\big)\,\boldsymbol{\Sigma}_t^{1/2}. \tag{2}$$

Two properties of (2) are central. First, the matrix exponential guarantees $\boldsymbol{\Sigma}_{t+1} \in \mathrm{SPD}(3)$ for any learning rate $\eta > 0$. Eigenvalue clipping is therefore not required, which avoids the discontinuities that produce force chatter. Second, the pre- and post-multiplications by $\boldsymbol{\Sigma}_t^{1/2}$ warp the gradient by the local curvature of the manifold, so the resulting step is scale-invariant. A $10\times$ change in baseline stiffness yields the same effective step size. We exploit this in the experiments: a single $\eta = 5 \times 10^{-3}$ across all stiffness regimes.

B. Energy-Aware MPPI Cost

Let $\hat{\boldsymbol{\Sigma}}_t$ denote the current estimate and $\boldsymbol{n}$ the active contact normal in the world frame. The normal is supplied by the current phase's contact strategy, which specifies a face axis and sign on the manipulated body and propagates them through the body pose. Let $F_{\text{pred}} = |\boldsymbol{n}^\top \hat{\boldsymbol{\Sigma}}_t \boldsymbol{\delta}|$ be the predicted wall-normal force and $[F_{\min}, F_{\max}]$ the active phase's force band. The MPPI cost augments geometric tracking with three physics terms:

$$\begin{aligned} J(\boldsymbol{x}_k) = {} & \underbrace{\|\boldsymbol{x}_k - \boldsymbol{x}_{\text{goal}}\|^2_{\boldsymbol{Q}}}_{\text{task tracking}} + \underbrace{\lambda_E\,\boldsymbol{\delta}^\top \hat{\boldsymbol{\Sigma}}_t\,\boldsymbol{\delta}}_{\text{interaction energy}} \\ & + \underbrace{\rho\,\mathrm{ReLU}(F_{\text{pred}} - F_{\max})^2}_{\text{force upper bound}} \\ & + \underbrace{\gamma\,\mathrm{ReLU}(F_{\min} - F_{\text{pred}})^2}_{\text{contact maintenance}}. \end{aligned} \tag{3}$$

The interaction-energy term $\boldsymbol{\delta}^\top \hat{\boldsymbol{\Sigma}}_t \boldsymbol{\delta}$ is twice the stored elastic energy of a linear spring; the conventional $\frac{1}{2}$ factor is absorbed into λ_E. The term encodes stiffness-awareness and penalises trajectories that compress a stiff direction more than a compliant one. The $F_{\max}$ term is a soft jamming bound. The $F_{\min}$ term is essential: without it the planner minimises the energy term by avoiding contact entirely, which prevents task completion and starves the estimator of force excitation. With $F_{\min} > 0$ the planner is forced to maintain contact, which provides the persistent excitation the estimator needs.

C. Phase Plan with Typed Validators

A first-frame call to a vision-language model emits a structured phase plan with a stiffness-prior topology assigning one of three discrete levels per axis, a task-frame orientation, an ordered list of phases, and recovery hints. Each phase carries its own cost-weight matrix $\boldsymbol{Q}$, force band $[F_{\min}, F_{\max}]$, contact strategy (face axis, sign, standoff, vertical offset, gripper command), and action prior. A semantic monitor consumes live metrics (end-effector-to-contact distance, wall gap, wall-normal force, height) and rotates phases when the active phase's trigger predicate is satisfied. Two trigger families cover the experiments: a force-threshold predicate that fires when wall-normal force exceeds a target, and a geometric-proximity predicate that fires when the end-effector enters a neighborhood of the contact target. The monitor re-queries the language model on stall, drop, or repeated over-force events; the revision step may rewrite cost weights, force bands, contact strategy, or the entire phase plan.

Two recurring failure modes of the raw language-model output motivate a typed contract. First, the model imports triggers from one task family into another, for example a proximity trigger referencing a box face in a scene that contains only a spring button. Second, the model returns target forces that exceed the joint range or saturate the controller. We add a small per-task validator that runs after every model call: it pins the contact axis to the task-appropriate direction, replaces invalid triggers with the force-threshold predicate, and clips the force band to $[0.7\,kd^\star, 1.3\,kd^\star]$ using any closed-form scene parameter (in simulation, the declared slide-joint stiffness k). The validator is the difference between zero and three successful runs on the spring-button sweep.

TABLE I
SPRING-BUTTON SWEEP (POST-VALIDATOR). RUNS TERMINATE EARLY ON REACHING 20 CONSECUTIVE IN-BAND STEPS; "IN-BAND" COUNTS QUALIFYING STEPS BEFORE TERMINATION.

k (N/m)	success	steps	in-band	max F_n (N)	revisions
50	yes	89	51	1.63	7
200	yes	43	24	6.22	2
800	yes	56	33	21.33	10

III. EXPERIMENTS

A. Setup

All experiments use a simulated 7-DoF Franka Emika Panda with an OSC operational-space controller in MuJoCo [10] via the LIBERO [9] task framework. The MPPI samples $N{=}128$ rollouts of horizon $H{=}24$ at 20 Hz in parallel; warm-start uses the previous step's optimal trajectory for half the samples. The Riemannian estimator runs at control frequency with $\eta = 5 \times 10^{-3}$ and a smallest-eigenvalue clamp of 10^{-1} for floating-point robustness only; the SPD property is already preserved by (2). FoundationPose [4] tracks the manipulated object's 6-DoF pose from RGB-D. We measure two task-level metrics: *steps-in-band*, the number of control steps with $F_n \in [0.7\,F^\star, 1.3\,F^\star]$ during a press/hold phase, and the binary *success* flag, set to true after 20 consecutive in-band steps.

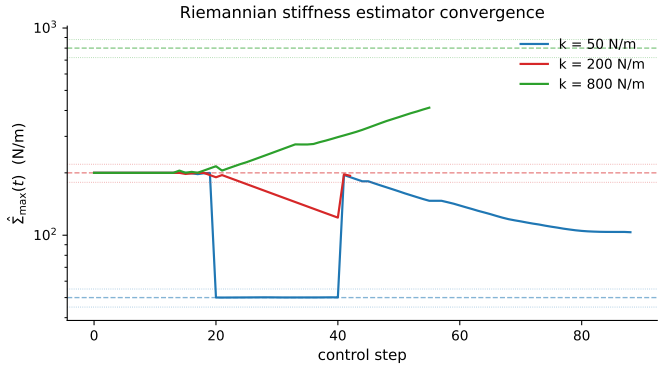

Fig. 1. Spring-button stiffness sweep, 200-step budget. Solid: $\hat{\Sigma}_{\max}(t)$ for $k \in \{50, 200, 800\}$ N/m. Dashed: ground-truth k_{true}. The same $\eta = 5 \times 10^{-3}$ produces motion of comparable shape across 1.2 orders of magnitude (a 16× range), consistent with the affine-invariant scale-invariance of (2). The simultaneous jumps near step 40 reflect a single artifact: each language-model revision re-initialises the estimator to the new prior. Updating only the prior topology and retaining the matrix removes this artifact.

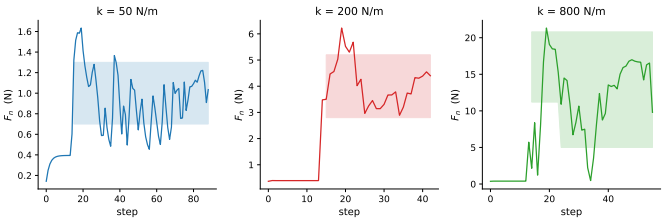

Fig. 2. Measured wall-normal force $F_n(t)$ during the spring-button sweep. Shaded band: target $[0.7\,F^\star,\ 1.3\,F^\star]$ with $F^\star = kd^\star$, $d^\star = 2$ cm. Each run enters the band shortly after contact and the trajectory remains anchored to the band thereafter, demonstrating that the ρ and γ terms in (3) regulate the contact force as designed. The band scales linearly with k across the sweep, set by the per-task validator from the ground-truth joint stiffness.

B. Stiffness Estimator Convergence (RQ1, RQ2)

The spring-button task is a closed-form validation environment for (2). A cylindrical cap on a prismatic joint with declared stiffness k and rest position q_{ref}=0 gives the estimator a ground-truth target. We run three independent rollouts at $k \in \{50, 200, 800\}$ N/m with target depth $d^\star = 2$ cm, hence target force $F^\star = kd^\star \in \{1, 4, 16\}$ N.

After the validator is applied, all three runs satisfy the in-band success criterion and terminate early (Table I). Table II isolates the validator's effect: without it, $0/3$ runs reach success because the language model imports wall-task triggers into the spring scene and proposes force bands outside the achievable joint range. The validator is the smallest sufficient contract identified in this study.

The estimator behaviour is the headline result. Figure 1 plots $\hat{\Sigma}_{\max}(t)$ alongside each k_{true} on a log axis. The estimator is initialised at the MEDIUM prior (∼200 N/m). The k=200 trace begins near ground truth. The k=50 trace descends and reaches the ground-truth level within ∼5 contact steps after the press phase starts. The k=800 trace ascends toward its target with the same step shape as the other two. The single learning rate $\eta = 5 \times 10^{-3}$ drives all three traces, confirming

TABLE II
VALIDATOR ABLATION ON THE SPRING-BUTTON SWEEP. "IN-BAND RATIO" AVERAGES PER-RUN *steps-in-band / total steps*; THE WITHOUT-VALIDATOR RUNS NEVER TRANSITIONED TO THE PRESS PHASE.

configuration	success	in-band ratio (avg)
Full FORTE	3/3	0.46
No per-task validator	0/3	0.00

the scale-invariance prediction of the affine-invariant metric. Figure 2 shows $F_n(t)$ entering and remaining inside the per-run target band $[0.7\,kd^\star, 1.3\,kd^\star]$, directly demonstrating the ρ and γ terms in (3). **RQ1** (positive-definite online identification at control frequency) and **RQ2** (force-band regulation under the language-model phase plan) are answered affirmatively in this regime.

C. Failure Analysis on Box-and-Wall Tasks (RQ3)

We probe the framework at a regime that goes beyond the estimator's clean test bed. The *compliant force-hold* task uses a foam-pad wall ($k_{\text{eff}} \approx 200$ N/m) and asks the planner to maintain $F_n \in [3, 8]$ N for 20 consecutive steps with no lift objective. The VLM emits a three-phase plan, contact latches at step 92, but only $2/200$ steps fall in band; the box tilts to $180°$ before stable wall pressure is established. The *wall-lift* task asks the planner to slide a 16 cm box up a vertical wall via friction. At the default box mass (∼570 g), $\mu N \geq mg$ at μ=0.5 requires $N \geq 11.2$ N, above the OSC saturation of ∼7.5 N per axis; the task is force-budget-infeasible. After reducing box mass to 90 g, the same contact-failure recurs in a different guise: the lift plateaus at 2.6 cm before relaxing.

Both failures localise to a common mechanism that is logically independent of (2) and (3). The LLM-generated contact target is a fixed point on a moving rigid body; the pose-tracking term in (3) pulls the end-effector to that point while the force-tracking and height-tracking terms push the body away. The two are geometrically incompatible during the lift phase, and no setting of $(\lambda_E, \rho, \gamma)$ that preserves the spring-button validation closes this gap. The corrective fix is structural: the contact target must be a contact *set* (a face normal plus a permissible range along the face) rather than a single point. This is a representation change to the LLM contract, not a change to the estimator or the cost.

An ablation supports the diagnosis. We hardcoded a post-contact heuristic that locks the contact target to the box's lower edge and forces the action prior to a $45°$ vector (equal wall-normal and upward components). The lift increases from 2.6 cm to 5.7 cm and the maximum tilt drops from $21°$ to $16°$; the residual gap is a time-budget effect, not a control-law effect.

IV. CONCLUSION

FORTE augments a vision-language-initialised, validator-constrained MPPI control loop with an online Riemannian stiffness estimator on $\mathrm{SPD}(3)$ and an Energy-Aware MPPI cost with explicit force-band barriers. The estimator update is

positive-definite by construction, which is a sufficient condition for the energy term in the cost to remain well-posed but not a closed-loop passivity proof. The update is also scale-invariant across orders of magnitude in stiffness, and the cost adds a contact-maintenance bound that prevents the planner from minimising energy by avoiding contact. On a closed-form spring-button validation the estimator tracks ground-truth k in the correct direction at every press across $\{50, 200, 800\}$ N/m with a single learning rate. The full pipeline reaches the in-band success criterion on three of three runs once the per-task validator is applied to the language-model output. On harder box-and-wall tasks the bottleneck is upstream of FORTE: the language-model-generated contact target is a fixed point on a moving rigid body, which is geometrically incompatible with the force-tracking objectives in the cost. The follow-up is a contact-*set* representation in the validator together with revision-persistent estimator state. Both can be implemented without changing (2) or (3). Three further open directions are a 6D impedance that couples translation and rotation, a formal closed-loop passivity argument for the estimator and plant together, additional ablations against Euclidean updates with eigenvalue clipping and diagonal stiffness, and a real-robot port that probes scene parameters online rather than reading them from the simulator.

REFERENCES

[1] B. Cicek, M. K. Er, and O. S. Oguz, "CoRAL: Contact-Rich Adaptive LLM-based Control for Robotic Manipulation," *Robotics: Science and Systems (RSS)*, 2026.

[2] Y. Zhang et al., "OmniVIC: A Self-Improving Variable Impedance Controller with Vision-Language In-Context Learning," arXiv:2510.17150, 2025.

[3] Y. Zhang et al., "CompliantVLA-adaptor: VLM-Guided Variable Impedance Action for Safe Contact-Rich Manipulation," arXiv:2601.15541, 2026.

[4] B. Wen et al., "FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects," *Proc. CVPR*, 2024.

[5] X. Yu et al., "ForceVLA: Enhancing VLA Models with a Force-aware MoE for Contact-rich Manipulation," *Proc. NeurIPS*, 2025.

[6] X. Pennec, P. Fillard, and N. Ayache, "A Riemannian Framework for Tensor Computing," *Int. J. Comput. Vis.*, vol. 66, no. 1, pp. 41–66, 2006.

[7] N. Jaquier, L. Rozo, and S. Calinon, "Bayesian Optimization Meets Riemannian Manifolds in Robot Learning," *Proc. CoRL*, 2020.

[8] S. Calinon, "Gaussians on Riemannian Manifolds: Applications for Robot Learning and Adaptive Control," *IEEE Robot. Autom. Mag.*, vol. 27, no. 2, pp. 71–82, 2020.

[9] B. Liu et al., "LIBERO: Benchmarking Knowledge Transfer for Lifelong Robot Learning," *Proc. NeurIPS*, 2023.

[10] E. Todorov, T. Erez, and Y. Tassa, "MuJoCo: A physics engine for model-based control," *Proc. IROS*, 2012.