ControlJEPA: Principled Trajectory Regularization via Lyapunov Tube Loss

Thai-Khanh Nguyen*
Faculty of Information Technology 1
Posts & Telecommunications Institute of Technology
khanhnt1@ptit.edu.vn

Nhi Y. Pham*
Center for AI Research
VinUniversity
nhi.py2@vinuni.edu.vn

Uyen N.B. Vo*
Faculty of Artificial Intelligence
Posts & Telecommunications Institute of Technology
uyenvnb@ptit.edu.vn

Thieu N. Vo†
Department of Computer Science
University of Bath
ntv22@bath.ac.uk

Cuong Pham†
Faculty of Information Technology 1
Posts & Telecommunications Institute of Technology
cuongpv@ptit.edu.vn

Tan M. Nguyen†
Department of Mathematics
National University of Singapore
tanmn@nus.edu.sg

Abstract

Next-token prediction leaves hidden-state trajectories largely unconstrained, while existing geometric regularizers primarily control the direction rather than the magnitude of trajectory deviations. We introduce *ControlJEPA*, a trajectory regularization framework based on a discrete-time Lyapunov contraction condition from practical Input-to-State Stability theory. Its Lyapunov Tube Loss penalizes violations of this condition on a normalized, scale-invariant measure of transversal deviation. When the contraction condition holds at every step, the hidden-state trajectory remains within an explicit tube around the reference secant. In the partial-ISS regime, we bound mean transversal energy, temporal variance, and average tube width using the Lyapunov loss and contraction parameters, with horizon-independent variance bounds under uniformly bounded violations. ControlJEPA requires no architectural modifications. Across six benchmarks and six model families, it consistently improves accuracy and data efficiency over standard fine-tuning and prior trajectory regularizers. The resulting smoother transversal profile also persists empirically during inference, when the regularizer is inactive. The code is publicly available at: https://github.com/t-khanusa/ControlJEPA.

1 Introduction

A fundamental limitation of next-token prediction (NTP) is that it uses a sum of tokenwise prediction losses without an explicit trajectory-geometry penalty, without imposing global structure on the resulting hidden-state trajectories [37, 19]. Recent trajectory regularization methods seek to address this limitation by penalizing angular deviations between trajectory segments defined by randomly sampled triplets of hidden states, thereby improving data efficiency without modifying the model architecture [19, 42]. However, these angle-based approaches constrain only the direction of displacement, rather than the magnitude; as a result, trajectories may still drift arbitrarily far from the semantic reference path without incurring any additional penalty.

*Co-first authors. †Co-last authors. Correspondence to: khanhnt1@ptit.edu.vn & tanmn@nus.edu.sg

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

This limitation reflects a deeper issue: the lack of a principled stability framework for hidden-state dynamics. Existing regularizers provide neither monotonic control over transversal deviation nor explicit bounds on trajectory divergence as a function of hyperparameters. As a result, notions such as tube width and regularization strength remain ill-defined.

Contribution. We introduce the *Controlled Joint Embedding Predictive Architecture (ControlJEPA)*, a principled trajectory regularization framework grounded in discrete-time Input-to-State Stability (ISS) theory [20]. Our central contribution is the *Lyapunov Tube Loss* $\mathcal{L}_{\text{Lyap}}$, which penalizes violations of an ISS contraction condition on a dimensionless measure of transversal deviation and is zero whenever the condition is already satisfied. ControlJEPA augments standard NTP training with $\mathcal{L}_{\text{Lyap}}$:

$$\mathcal{L}_{\text{ControlJEPA}}(\theta) = \mathcal{L}_{\text{NTP}}(\theta) + \lambda\, \mathcal{L}_{\text{Lyap}}(\theta), \tag{1}$$

requiring no architectural modifications. When the contraction condition holds, hidden states remain within a tube around a reference direction with radius $r^* = \sqrt{\tau/(1-\gamma)}||v_{\text{geo}}||_2$ (Definition 3.2). Enforcing it at every step, however, can excessively compress the transversal component and limit representational capacity (Section 4.1). We therefore operate in a *partial-ISS* regime, where the constraint is enforced softly and the achieved geometry reflects a balance between NTP and Lyapunov regularization. The effective tube width is mainly controlled by λ and γ, while τ sets the theoretical ceiling r^*. Normalization by the reference chord further makes the loss invariant to global rescaling. Our contributions are three-fold:

1. We propose a stability-based framework for hidden-state trajectory regularization using discrete-time ISS theory, yielding explicit and interpretable bounds on trajectory deviation.
2. We introduce the Lyapunov Tube Loss, a practical and architecture-agnostic regularization objective that penalizes violations of a trajectory contraction inequality while preserving representational flexibility.
3. We bound mean transversal energy, temporal variance, and average tube width in the partial-ISS regime, with horizon-independent variance control under uniformly bounded maximal violations. A spectral capacity bound links vanishing transversal deviation to limited logit expressivity under a bounded unembedding norm, motivating soft regularization.

We empirically demonstrate that ControlJEPA consistently outperforms standard fine-tuning and prior trajectory regularizers across six benchmarks, including NL-RX-SYNTH, NL-RX-TURK, GSM8K, Spider, NQ-Open, and HellaSwag, and six model families. On GSM8K and NQ-Open, it improves accuracy over standard fine-tuning by approximately 10 percentage points on each benchmark, and over Semantic Tube Prediction by approximately 6 and 4 percentage points, respectively. ControlJEPA also improves data efficiency, exhibits a stable hyperparameter region, and empirically retains smoother transversal profiles during inference, when the regularizer is inactive.

Organization. We structure this paper as follows: In Section 2, we formalize hidden-state trajectories and introduce the discrete-time ISS framework. In Section 3, we derive the Lyapunov Tube Loss and the ControlJEPA objective. In Section 4, we establish theoretical guarantees on capacity preservation and transversal profile stability. Sections 5 and 6 present empirical validation and deeper analysis across six benchmarks and six model families. We discuss related work in Section 7 and conclude with future directions. Proofs and implementation details are provided in the Appendix.

2 Preliminaries

2.1 Hidden-State Trajectories in Language Models

Let $h_t \in \mathbb{R}^D$ denote the hidden-state representation at token position t, defined as the final transformer layer output prior to the output projection (unembedding) into vocabulary logits. A transformer trained with next-token prediction (NTP) induces a sequence of hidden states $h_t = f_\theta(x_{\leq t})$ that evolves according to a discrete-time dynamical system:

$$h_{t+1} = F_\theta(h_t, x_{t+1}), \qquad t \in \{0, \ldots, T-1\}, \tag{2}$$

where θ denotes the model parameters and x_{t+1} is the input at step $t+1$ (the ground-truth token under teacher forcing). The initial state h_0 is determined by the input embedding. In the terminology of dynamical systems, this defines a discrete-time *map* [43].

Training is performed via the NTP objective:

$$\mathcal{L}_{\text{NTP}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}), \tag{3}$$

which optimizes predictions independently at each position. As a consequence, it imposes no explicit constraint on the geometry of the resulting hidden-state trajectory $(h_0, \ldots, h_T)$, which is the primary object of interest in this work.

2.2 Existing Trajectory Regularization

The *geodesic hypothesis* suggests that well-structured hidden-state trajectories should follow approximately straight paths in representation space, as straighter trajectories enable more predictable dynamics and reduce first-order prediction error [15]. This behavior is observed in both biological and artificial systems, and recent evidence indicates that large language models implicitly learn to straighten representation trajectories, suggesting that linearization is a fundamental inductive bias in sequence modeling [16]. Recent methods operationalize this idea using angle-based penalties.

Semantic Tube Prediction (STP). Huang et al. [19] penalize angular deviation between displacement vectors constructed from randomly sampled triplets:

$$\mathcal{L}_{\text{STP}} = \mathbb{E}_{s<r<t}\big[1 - \cos(h_t - h_r,\ h_r - h_s)\big]. \tag{4}$$

Temporal Straightening (TS). Wang et al. [42] enforce local smoothness by penalizing curvature between consecutive latent velocities:

$$\mathcal{L}_{\text{TS}} = \mathbb{E}_{t}\big[1 - \cos(z_{t+1} - z_t,\ z_{t+2} - z_{t+1})\big]. \tag{5}$$

where z_t denotes the latent state of the world model in [42]; in our setting, the analogous quantity is h_t.

While effective in practice, these approaches share key limitations. First, they constrain only the *direction* of motion and do not control the *magnitude* of deviation, allowing trajectories to drift arbitrarily far from a reference path. Second, they operate locally and do not enforce global bounds on trajectory deviation. Finally, they provide no explicit relationship between hyperparameters and geometric quantities such as tube width, limiting interpretability.

2.3 Discrete-Time Input-to-State Stability

To overcome the aforementioned limitations, we adopt a discrete-time Lyapunov contraction condition inspired by Input-to-State Stability (ISS) theory [20]. Unlike angle-based penalties, this condition controls the magnitude of deviation and yields explicit bounds on it. The discrete-time formulation is natural for transformers and JEPA-style models, which evolve strictly in discrete steps.

Consider a system with state $e_t \in \mathbb{R}^D$ and exogenous signal $\mu_t \in \mathbb{R}^m$:

$$e_{t+1} = g_\theta(e_t, \mu_t). \tag{6}$$

Here μ_t collects every quantity other than e_t that drives the update, such as the preceding hidden states and the next input. Equation (6) is therefore a representation of the dynamics, not a modeling assumption.

Definition 2.1 (ISS-type Lyapunov function)**.** A continuous function $V : \mathbb{R}^D \to \mathbb{R}_{\geq 0}$ is an *ISS-type Lyapunov function* along a trajectory $(e_t)_{t \geq t_s}$ of (6) if there exist $\gamma \in (0, 1)$ and $\tau \geq 0$ such that

$$V_{t+1} \leq \gamma V_t + \tau, \qquad V_t = V(e_t), \qquad t \geq t_s. \tag{7}$$

Iterating this inequality from the initial step t_s yields

$$V_t \leq \frac{\tau}{1-\gamma} + \gamma^{t-t_s} V_{t_s}, \qquad t \geq t_s. \tag{8}$$

The first term is a steady-state bound determined by (γ, τ). The second term is a transient that decays geometrically. Consequently, the sublevel set $\{V \leq \tau/(1-\gamma)\}$ is forward invariant, and $\limsup_{t\to\infty} V_t \leq \tau/(1-\gamma)$ along every trajectory satisfying (7).

This explicit bound is central to our approach. We construct a Lyapunov function that measures transversal deviation. Whenever the contraction condition holds, the bound translates into a quantitative limit on the width of the trajectory tube.

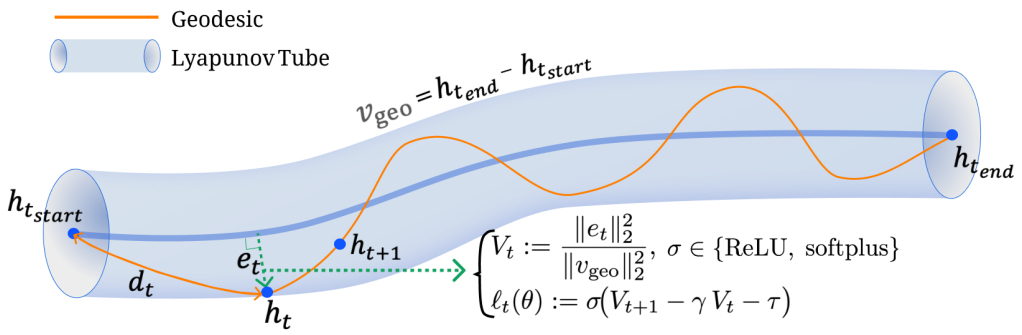


Figure 1: **Lyapunov Tube.** The displacement $d_t = h_t - h_{t_{\text{start}}}$ is decomposed into a longitudinal component along the reference secant v_{geo} and a transversal error $e_t \perp v_{\text{geo}}$. The Lyapunov Tube Loss penalizes violations of $V_{t+1} \leq \gamma V_t + \tau$, confining hidden-state trajectories within the tube. (Definition 3.2).

3 Principled Trajectory Regularization via Lyapunov Tube Loss

We develop ControlJEPA in three steps. First, we decompose hidden-state displacement into longitudinal and transversal components (§3.1). Next, we define a dimensionless Lyapunov function that induces a tube via the ISS bound (§3.2). Finally, we derive the Lyapunov Tube Loss by penalizing violations of the contraction condition and integrate it with the NTP objective (§3.3).

3.1 Transversal Decomposition

We instantiate the ISS framework of §2.3 for transformer hidden-state trajectories. For each token position t within a semantic segment, we define the *transversal error* $e_t \in \mathbb{R}^D$ as the component of the hidden-state displacement that lies perpendicular to a reference direction connecting two semantic anchor states (formally derived in Eq. (10)). This error plays the role of the state in system (6), and the Lyapunov candidate V_t is constructed so that the decrease condition (7) directly bounds the tube width. Let $h_t \in \mathbb{R}^D$ denote the hidden state at position t and let $h_{t_{\text{start}}}, h_{t_{\text{end}}} \in \mathbb{R}^D$ be semantic anchor states. The *reference secant* v_{geo} is the vector connecting these anchors; its unit vector $\hat{v}$ defines a preferred direction along which the trajectory is expected to evolve:

$$v_{\text{geo}} = h_{t_{\text{end}}} - h_{t_{\text{start}}}, \quad \hat{v} = \frac{v_{\text{geo}}}{\|v_{\text{geo}}\|_2}, \quad d_t = h_t - h_{t_{\text{start}}}, \quad \|v_{\text{geo}}\|_2 > 0 \tag{9}$$

Here d_t is the displacement of the intermediate hidden state h_t from the start anchor. Projecting d_t onto $\hat{v}$ decomposes it into a *longitudinal* component (the longitudinal coordinate along the reference direction) and a *transversal* error e_t (the deviation perpendicular to it):

$$d_t = \underbrace{\langle d_t, \hat{v} \rangle \hat{v}}_{\text{longitudinal}} + \underbrace{e_t}_{\text{transversal}}, \qquad e_t = (I - \hat{v}\hat{v}^\top) d_t, \qquad e_t \perp \hat{v}. \tag{10}$$

The norm $\|e_t\|_2$ measures how far the trajectory deviates from the reference line at step t, and serves as the tube radius. We define the Lyapunov Tube in Definition 3.2 below. Figure 1 illustrates this decomposition.

3.2 Lyapunov Candidate

We now convert the transversal decomposition in Section 3.1 into a scalar stability certificate. The goal is to measure not only whether the trajectory is aligned with the reference direction, but also how far it deviates from the reference secant in magnitude. This scalar quantity will serve as the Lyapunov candidate whose ISS contraction yields an explicit tube bound around the semantic reference path.

A Lyapunov candidate is a nonnegative scalar function $V_t = V(e_t)$ that plays the role of an energy for the transversal deviation. We define the *reference line* as the affine line $\{h_{t_{\text{start}}} + \alpha\, v_{\text{geo}} \mid \alpha \in \mathbb{R}\}$ passing through the two semantic anchors, along which a perfectly straight trajectory would evolve. In our setting, V_t should satisfy three requirements: it should vanish exactly when the hidden state lies on the reference line, increase with the squared transversal deviation, and remain invariant under global rescaling of the hidden-state representation. To achieve this, we normalize the squared transversal error by the squared length of the reference secant.

Definition 3.1. (Lyapunov Candidate). Given the transversal error e_t and reference secant v_{geo}, we define:

$$V_t = \frac{\|e_t\|_2^2}{\|v_{\text{geo}}\|_2^2}. \tag{11}$$

This quantity is dimensionless and scale-invariant. Moreover, $V_t = 0$ if and only if h_t lies on the reference line through $h_{t_{\text{start}}}$ with direction v_{geo}. Thus, V_t measures the relative transversal deviation of the hidden-state trajectory from the semantic reference path.

The normalization by $\|v_{\text{geo}}\|_2^2$ is essential. Directly minimizing the unnormalized error $\|e_t\|_2^2$ would encourage the trajectory to collapse onto the reference line, which can reduce the rank of the hidden-state matrix and destroy representational capacity, as shown in Proposition 4.1. By contrast, the normalized candidate allows the model to reduce relative deviation either by controlling the transversal error or by increasing the semantic separation between the anchor states. This distinction is important because it permits geometric stabilization without forcing all hidden states into a low-dimensional geodesic subspace.

The ISS condition applied to this Lyapunov candidate takes the form

$$V_{t+1} \le \gamma V_t + \tau, \qquad \gamma \in (0,1), \quad \tau \ge 0, \tag{12}$$

where γ controls the contraction rate and τ specifies the allowable residual transversal deviation. Applying the iterated bound (8) with $t_s = t_{\text{start}}$ gives

$$V_t \le \frac{\tau}{1-\gamma} + \gamma^{t-t_{\text{start}}} V_{t_{\text{start}}}. \tag{13}$$

In our setting, the start anchor satisfies $d_{t_{\text{start}}} = h_{t_{\text{start}}} - h_{t_{\text{start}}} = 0$, hence $e_{t_{\text{start}}} = 0$ and $V_{t_{\text{start}}} = 0$. The transient term therefore vanishes identically, and the bound reduces to

$$V_t \le \frac{\tau}{1-\gamma}, \qquad t \ge t_{\text{start}}. \tag{14}$$

This motivates the following geometric notion.

Definition 3.2. (Lyapunov Tube).
Given contraction parameters $\gamma \in (0,1)$ and $\tau \ge 0$, the Lyapunov tube is the set of hidden states whose Lyapunov value lies within the ISS steady-state bound:

$$\mathcal{B}_{\text{tube}} = \left\{ h_t \;\middle|\; V_t \le \frac{\tau}{1-\gamma} \right\}. \tag{15}$$

Equivalently, in terms of the transversal error, this corresponds to the tube-radius bound:

$$\|e_t\|_2 \le r^* := \sqrt{\frac{\tau}{1-\gamma}}\, \|v_{\text{geo}}\|_2. \tag{16}$$

The tube $\mathcal{B}_{\text{tube}}$ serves as the target attractor set of the regularization: when the ISS condition (12) holds at every step, all hidden states are guaranteed to remain inside $\mathcal{B}_{\text{tube}}$; the Lyapunov Tube Loss (§ 3.3) is designed precisely to penalize violations of this condition. In practice, however, the ISS condition need not hold at every step to achieve better trade-offs between the model's stability and capacity. Section 4 analyzes this partial-ISS regime and characterizes what our method achieves when the condition holds only approximately.

Therefore, enforcing the ISS contraction condition on V_t provides an explicit and interpretable geometric guarantee: the hidden-state trajectory is confined to a tube around the reference secant whose width is determined by the contraction rate γ, the slack parameter τ, and the reference scale $\|v_{\text{geo}}\|_2$. Unlike angle-based penalties, this construction controls the magnitude of transversal deviation while preserving a nonzero slack region for expressive hidden-state representations.

3.3 ControlJEPA: Stable Trajectory Regularization

Violations of $V_{t+1} - \gamma V_t - \tau \le 0$ are penalized through a nonnegative activation σ:

$$\ell_t(\theta) = \sigma\big(V_{t+1} - \gamma V_t - \tau\big), \qquad \sigma \in \{\text{ReLU},\ \text{softplus}\}, \tag{17}$$

so $\ell_t = 0$ whenever the ISS condition is already satisfied ($\sigma = \mathrm{ReLU}$) or near zero ($\sigma = \mathrm{softplus}$). Following Huang et al. [19], we restrict penalization to valid transitions $\mathcal{T} \subseteq \{t_{\text{start}}+1, \ldots, t_{\text{end}}-1\}$ within contiguous semantic segments. We define the *Lyapunov Tube Loss* as follows:

Definition 3.3. (Lyapunov Tube Loss).

$$\mathcal{L}_{\text{Lyap}}(\theta) = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \sigma\left(\frac{\|e_{t+1}\|_2^2}{\|v_{\text{geo}}\|_2^2} - \gamma \frac{\|e_t\|_2^2}{\|v_{\text{geo}}\|_2^2} - \tau \right), \tag{18}$$

with $\gamma \in (0, 1)$ (contraction rate) and $\tau \geq 0$ (orthogonal slack).

The **ControlJEPA** objective augments NTP with the Lyapunov Tube Loss:

$$\mathcal{L}_{\text{controlJEPA}}(\theta) = \mathcal{L}_{\text{NTP}}(\theta) + \lambda \, \mathcal{L}_{\text{Lyap}}(\theta), \tag{19}$$

where $\lambda > 0$ controls the regularization strength. The only additions to standard fine-tuning are the projection in Eq. (10) and the scalar comparisons in Eq. (17).

The geometric effects of $\mathcal{L}_{\text{Lyap}}$ are characterized formally in Section 4: minimizing the loss provably confines the mean transversal level (Theorem 4.2, bound (27)) and suppresses its temporal variance (bound (28)), providing bounds on their relative transverse geometry. Crucially, this operates in a *partial-ISS* regime where $\mathcal{L}_{\text{Lyap}} > 0$, preserving the representational capacity that NTP requires (Section 4.2).

4 Theoretical Analysis

Scope of the theoretical analysis. The bounds in this section hold for trajectories on the training distribution; no formal guarantee extends to unseen inputs. Whether the induced geometric structure generalizes is an empirical question addressed in Section 5. Figure 5 shows the stable transversal profile persists at inference time, suggesting training reshapes the representation manifold globally. A theoretical characterization of this generalization remains an open question. The Lyapunov Tube Loss is designed to enforce the geodesic hypothesis [19] without destroying the representational capacity that next-token prediction requires. We establish two preliminary results that jointly motivate why the method must operate in a *partial-ISS regime* (§4.1–4.2), and then state the central theorem characterizing what minimizing $\mathcal{L}_{\text{Lyap}}$ achieves in that regime (§4.3).

4.1 Why Minimizing $\|e_t\|_2$ Collapses Model Capacity

To see why the transversal error $\|e_t\|_2$ must be *controlled* rather than minimized, consider the effect on the model's output distribution. Stack the hidden states as rows of $H \in \mathbb{R}^{T \times D}$. Using the orthogonal decomposition $h_t - h_0 = \alpha_t \hat{v} \|v_{\text{geo}}\|_2 + e_t$:

$$H \;=\; \mathbf{1} h_0^\top + \alpha \big(\hat{v} \|v_{\text{geo}}\|_2\big)^\top + E, \tag{20}$$

where $\alpha = (\alpha_1, \ldots, \alpha_T)^\top \in \mathbb{R}^T$ and $E \in \mathbb{R}^{T \times D}$ has rows e_t. The logit matrix then is $L = HU^\top$ (where U is the unembedding matrix). We define the logits of rank at most two:

$$\tilde{H} \;:=\; \mathbf{1} h_0^\top + \alpha \big(\hat{v} \|v_{\text{geo}}\|_2\big)^\top, \quad \tilde{L} \;:=\; \tilde{H} U^\top \tag{21}$$

Then $\mathrm{rank}(\tilde{L}) \leq 2$, since $\tilde{H}$ is a sum of two rank-one terms and right multiplication cannot increase rank.

We bound the distance of the next-token distributions from a low-dimensional family in Proposition 4.1, this distance is bounded by an explicit, continuous, monotone function of δ.

Proposition 4.1 (Spectral capacity bound)**.** *Suppose $\|e_t\|_2 \leq \delta$ for all t. Then*

$$\frac{1}{T} \sum_{j>2} \sigma_j \, (L)^2 \leq \delta^2 \, |U|_{op}^2, \tag{22}$$

and consequently there is a smooth manifold $\mathcal{N}$ of dimension at most 2 in the probability simplex with

$$\frac{1}{T} \sum_{t=1}^{T} \mathrm{dist}_{\ell_2}(p_\theta(.|x_{<t}), \mathcal{N})^2 \leq \frac{1}{4} \delta^2 \, |U|_{op}^2. \tag{23}$$

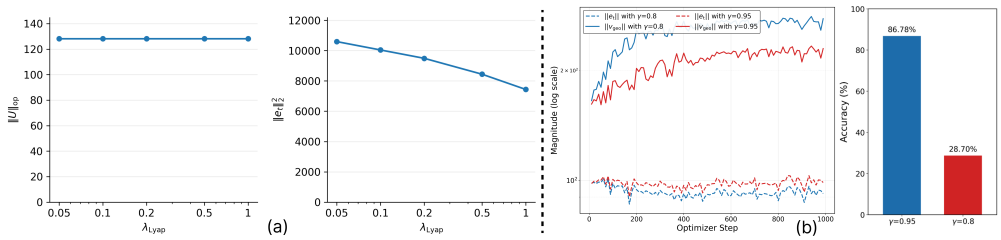


Figure 2: (a) Unembedding norm and transversal energy across λ_{Lyap}. (b) Impact of contraction rate γ on geometric components and downstream accuracy.

The formal proof is deferred to Appendix A.1. The key message is that any regularizer that drives $\|e_t\|_2$ too small destroys representational capacity as a side effect, regardless of how well it aligns the trajectory with the geodesic. The bound depends on the unembedding norm: it involves $|U|_{op}$, so it decreases with δ only if the unembedding norm does not grow to compensate. We therefore report measured $|U|_{op}$ along side δ across the λ_{Lyap} sweep in Figure 2(a), so that the product $\delta^2 |U|_{op}^2$ governing (23) can be read off directly rather than assumed.

4.2 The Partial-ISS Regime

Full ISS satisfaction leads to the same potential destruction. If $\mathcal{L}_{\text{Lyap}}(\theta^*) \approx 0$, the ISS condition holds at every step, and iterating from $V_0 = 0$ gives:

$$V_t \;\leq\; \frac{\tau}{1-\gamma}, \tag{24}$$

so $\|e_t\|_2 \leq \sqrt{\tau/(1-\gamma)}\,\|v_{\text{geo}}\|_2$. For small τ this again forces $\|e_t\|_2$ small, making the low-rank approximation bound in Proposition 4.1 small, provided $\|U\|$ and the reference scale are controlled. Hence, driving $\mathcal{L}_{\text{Lyap}}$ to approximately zero loss is likely to be *incompatible* with good NTP performance.

The practical training regime therefore necessarily satisfies $\mathcal{L}_{\text{Lyap}}(\theta) \gg 0$ (Figure 4(b)): the ISS condition is violated at some steps, keeping $\|e_t\|_2$ strictly positive and preserving the representational capacity stored in E. We call this the *partial-ISS regime*. Figure 2(b) confirms empirically that well-performing checkpoints always reside in this regime.

The question then becomes: what does adding Lyapunov Tube Loss actually achieve in the partial-ISS regime? The answer is the stability of the *temporal profile* of V_t rather than suppressing its level, which is the content of the next theorem. Specifically, we measure profile stability by the temporal variance $\sigma^2 := \text{Var}_t[V_t]$; a small σ^2 means the trajectory maintains a *uniform* distance from the geodesic across all token positions, rather than alternating between near-collapse and large excursions.

4.3 Transversal Profile Characterization

We first fix notation for the partial-ISS regime. At each step, define the **signed innovation**

$$\varepsilon_t \;:=\; V_{t+1} - \gamma V_t - \tau, \tag{25}$$

so that $\varepsilon_t \leq 0$ when the ISS condition is satisfied and $\varepsilon_t > 0$ when it is violated. The Lyapunov Tube Loss with the softplus activation satisfies

$$\mathcal{L}_{\text{Lyap}}(\theta) \;=\; \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \log\!\big(1 + e^{\varepsilon_t}\big). \tag{26}$$

Let $M := \max_{t\in\mathcal{T}} \max(\varepsilon_t, 0)$ denote the largest ISS violation along the trajectory, and $n := \frac{M^2}{\log(1+e^M)}$, so that $\frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \max(\varepsilon_t,0)^2 \leq n\,\mathcal{L}_{\text{Lyap}}(\theta)$ (Appendix A.2).

Theorem 4.2 (Transversal Profile Characterization)**.** *Let $\{V_t\}_{t=0}^T$ be the Lyapunov sequence with $V_0 = 0$, and let $\{\varepsilon_t\}$ be the innovations defined in (25). Assume $\gamma \in (0,1)$ and $\tau \geq 0$. Then the following hold for every trajectory, without any distributional assumption on $\{\varepsilon_t\}$:*

(i) Mean confinement. *The time-averaged Lyapunov value satisfies*

$$\bar{V}_T \;\leq\; \frac{\tau}{1-\gamma} \;+\; \frac{\mathcal{L}_{\text{Lyap}}(\theta)}{1-\gamma}. \tag{27}$$

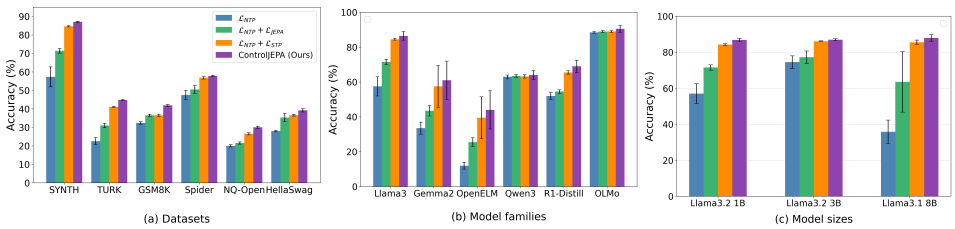


Figure 3: Accuracy and generalization performance of $\mathcal{L}_{NTP}$, LLM-JEPA, STP, and ControlJEPA across (a) Datasets, (b) Model families, and (c) Model sizes.

(ii) Variance control. *The temporal variance of the Lyapunov sequence satisfies*

$$\mathrm{Var}_t[V_t] \;\le\; \frac{2n}{(1-\gamma)^2}\left[1+\frac{4\gamma^2}{(1-\gamma)^2}\right]\mathcal{L}_{\mathrm{Lyap}}(\theta) \;+\; \frac{4\tau^2}{(1-\gamma)^2}\left[1+\frac{2\gamma^2}{(1-\gamma)^2}\right] \;+\; \mathcal{R}_T, \tag{28}$$

where $\mathcal{R}_T$ is an explicit boundary remainder with $\mathcal{R}_T = O(\frac{1}{T})$ whenever M is bounded uniformly in T.

(iii) Practical tube width. *By Jensen's inequality,*

$$\mathbb{E}_t[\|e_t\|_2] \;\le\; \sqrt{\bar{V}_T}\cdot\|v_{\mathrm{geo}}\|_2 \;\le\; \|v_{\mathrm{geo}}\|_2\cdot\sqrt{\frac{\tau+\mathcal{L}_{\mathrm{Lyap}}(\theta)}{1-\gamma}}. \tag{29}$$

The full proof is in Appendix A.2. Together, (27) and (28) relate the mean transversal level and its temporal fluctuations to the Lyapunov Tube Loss, the partial-ISS parameters (γ, τ), and the trajectory-dependent violation factor n. Under additional uniform control of the maximal positive innovation M, the variance bound also yields a horizon-independent control of the transversal fluctuations.

5 Experimental Results

In this section, we empirically validate ControlJEPA against standard fine-tuning and geodesic regularizers, showing that ISS-based trajectory control improves generalization, sample efficiency, and structural stability beyond heuristic geometric methods.

5.1 Experimental Setup

We evaluate ControlJEPA following Semantic Tube Prediction [19] across diverse tasks, architectures, and scales. Benchmarks include NL-RX-SYNTH, NL-RX-TURK [25], GSM8K [8], Spider [45], NQ-Open [24], and HellaSwag [46]. Evaluated models span Llama-3.1 (8B), Llama-3.2 (1B/3B) [10], Gemma-2-2b-it [38], OpenELM-1.1B [27], OLMo-2-1B [35], Qwen3-1.7B [44], and DeepSeek-R1-Distill-Qwen-1.5B [12]. We compare against standard fine-tuning, LLM-JEPA [18], and Semantic Tube Prediction [19]. All methods share an identical codebase, splits, decoding, validation rules, and 5 random seeds (82, 23, 37, 84, 4). Baselines use their published, dataset-tuned optimal epochs and learning rates; ControlJEPA fixes these shared hyperparameters and tunes only its stability parameters (Lyapunov penalty λ, contraction rate γ, ISS slack τ), conservatively favoring the baselines in tuning budget. All experiments ran on two NVIDIA H100 80GB GPUs.

5.2 Accuracy and Generalization

As shown in Figure 3, ControlJEPA consistently outperforms all baselines across datasets, architectures, and model scales. Its gains are especially pronounced on complex reasoning tasks such as GSM8K and NQ-Open, where heuristic methods like STP suffer from unconstrained error magnitudes. By enforcing the Lyapunov condition $V_{t+1} \le \gamma V_t + \tau$, ControlJEPA bounds transversal drift and keeps hidden states anchored to the semantic geodesic over long contexts. Its latent-manifold formulation also generalizes across Llama-3, Gemma-2, OpenELM, Qwen3, R1-Distill, and OLMo, and scales smoothly across Llama-3 models from 1B to 8B, with the dimensionless Lyapunov candidate preventing both scale collapse and representational bottlenecks.

5.3 Data Efficiency and Signal-to-Noise Ratio

A key benefit of trajectory noise filtering is improved sample efficiency. As shown in Figure 4(a) (Left), ControlJEPA outperforms standard fine-tuning and STP in moderate data regimes and remains competitive under severe subsampling, matching full-dataset $\mathcal{L}_{NTP}$ performance with only a fraction

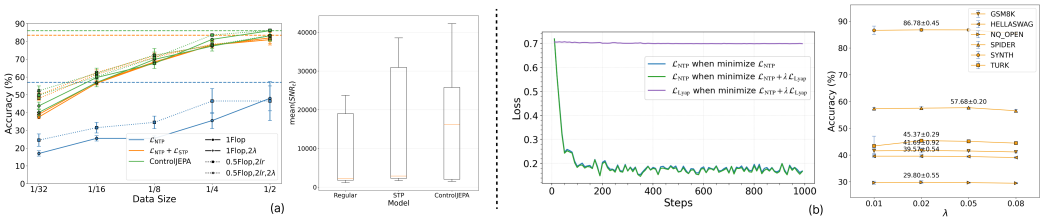


Figure 4: (a) Data Efficiency and Signal-to-Noise ratio. (b) Training Dynamics and Hyperparameter Robustness.

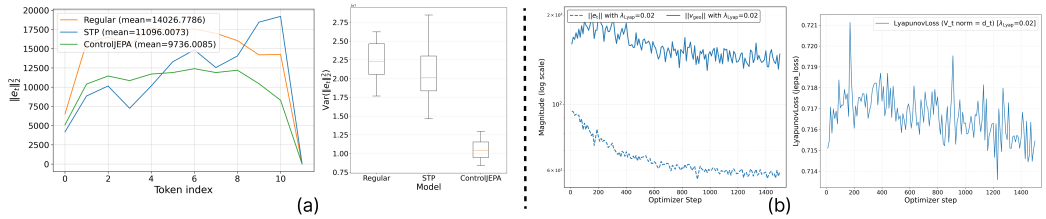


Figure 5: (a) Inference-time trajectory of squared transversal error $\|e_t\|_2^2$ across generated tokens in NL-RX-SYNTH dataset. (b) Without normalization, the representation scale collapses, so both $\|e_t\|$ and $\|v_{\text{geo}}\|$ decrease over time. **(Right)** Normalizing by local displacement introduces severe heteroscedastic instability and volatile loss dynamics.

of the data. This efficiency is consistent with an improved signal-to-noise ratio: by constraining orthogonal deviations through the ISS condition with slack τ, ControlJEPA can extract semantic alignment signals more effectively. Using the previously defined displacement decomposition, we quantify this effect on greedy generations as

$$\text{SNR}_t = \frac{|\langle d_t, \hat{v}\rangle \hat{v}|_2^2}{||e_t||_2^2} = \frac{\langle d_t, \hat{v}\rangle^2}{||e_t||_2^2}, \tag{30}$$

which measures the ratio of aligned displacement energy to orthogonal residual energy. Figure 4(a) (Right) supports this interpretation, showing that ControlJEPA achieves a substantially higher median SNR than both baselines.

5.4 Training Dynamics and Hyperparameter Sensitivity

As shown in Figure 4(b) (Left), ControlJEPA is both optimization-stable and robust to hyperparameters. The $\mathcal{L}$NTP curves overlap when trained alone or jointly with ControlJEPA, indicating that the geometric constraint improves trajectory structure without harming language modeling. The Lyapunov loss stabilizes near 0.69, matching the softplus value at zero innovation $\ln(1+e^0) \approx 0.693$ once the ISS condition $V_{t+1} - \gamma V_t - \tau \leq 0$ is satisfied. Moreover, ControlJEPA maintains consistently high accuracy across $\lambda \in [0.01, 0.08]$, providing a broad stability region and reducing the need for exhaustive tuning as shown in Figure 4(b) (Right).

6 Empirical Analysis

6.1 Implicit Tube Preservation at Inference Time

At generation time, the Lyapunov loss is deactivated. To test whether the learned geometry generalizes, Figure 5(a) plots the unnormalized squared transversal error (Left) and its variance (Right) on unseen trajectories. Standard fine-tuning shows substantial drift, while STP reduces the mean error but remains highly oscillatory. In contrast, ControlJEPA achieves the lowest and most stable transversal error, keeping hidden states tightly confined within the semantic tube even without an active penalty. This suggests that ISS-based training reshapes the representation manifold rather than merely memorizing training trajectories.

6.2 Preventing Dimensionality Collapse and Heteroscedastic Instability

The dimensionless Lyapunov candidate $V_t = \frac{\|e_t\|_2^2}{\|v_{\text{geo}}\|_2^2}$ is designed to avoid two key failure modes in trajectory optimization: scale collapse and heteroscedastic instability.

Preventing Scale Collapse. Penalizing the unnormalized transversal energy, $V_t = |e_t|_2^2$, encourages a trivial solution that shrinks the entire representation space, with $d_t \to 0$ and $v_{\text{geo}} \to 0$. As

shown in Figure 5(b) (Left), this collapses the chord length $|v_{\text{geo}}|_2^2$, destroys manifold capacity, and reduces accuracy to 6.3%. Normalizing by a length scale makes V_t scale-invariant and removes the regularizer's direct incentive for global shrinking.

Preventing Heteroscedastic Instability. Normalizing by the local displacement, $V_t = \frac{\|e_t\|_2^2}{\|d_t\|_2^2}$, destabilizes the ISS condition because $\|d_t\|_2^2$ varies with token position, creating a moving denominator that assigns inconsistent penalties across positions. This heteroscedasticity leads to volatile loss dynamics and suboptimal convergence, as shown in Figure 5(b) (Right). In contrast, the global chord length $\|v_{\text{geo}}\|_2^2$ provides a sequence-level invariant for uniformly measuring Lyapunov violations.

6.3 The Necessity of Orthogonal Slack and Joint Equilibrium

To validate Proposition 4.1 and the Partial-ISS Regime discussed in Section 4.2, we ablate the contraction rate γ and track the orthogonal slack $|e_t|$ and macroscopic chord $|v_{\text{geo}}|$ during training with $\lambda_{\text{Lyap}} = 0.05$, as shown in Figure 2(b).

Capacity Collapse under Strict Constraints. With a strong contraction, $\gamma = 0.8$, ControlJEPA approximates the strict geodesic constraint of STP, driving $|e_t|$ toward zero. As predicted by Proposition 4.1, removing this orthogonal slack collapses the hidden-state rank, restricts expressivity, and reduces downstream accuracy to 28.7%, exposing the limitation of static geodesic interpolation.
Geometric Linearization via Chord Expansion. Under optimal soft regularization, $\gamma = 0.95$, the transversal error $\|e_t\|_2$ stabilizes at a positive value, preserving representational capacity and yielding 86.78% accuracy, while $|v_{\text{geo}}|$ grows substantially. This confirms that ControlJEPA achieves geometric linearization not by eliminating microscopic slack, but by expanding the global reference frame under the NTP objective.

7 Related Work

Joint-Embedding Predictive Architectures. JEPA [23] learns representations by predicting target embeddings from context, and has been extended to vision [2, 6, 3, 29, 17, 30], language and multimodal learning [18, 7], sequential and signal-based settings [40, 9, 11], and world models and energy-based models [39, 26]. While many methods address dimensional collapse through embedding-space regularization [21, 5, 4, 14, 22, 28], STP [19] instead regularizes hidden-state trajectories via cosine-based geodesic alignment. ControlJEPA replaces cosine alignment with a one-sided, scale-invariant Lyapunov contraction constraint on transversal deviation, avoiding collapse through partial enforcement and normalization.

Geometric Regularization of Hidden-State Trajectories. Trajectory straightening has been observed in both vision [15] and language [16], motivating methods such as Temporal Straightening [42] and STP [19]. However, these cosine-based regularizers control only the *direction* of motion, not the *magnitude* of deviation. ControlJEPA penalizes violations of $V_{t+1} \leq \gamma V_t + \tau$, bounding deviation where satisfied and its mean otherwise.

Control-Theoretic Methods for Deep Learning. Control-theoretic methods enforce stability through either architectural design [13, 32, 33, 1, 31, 34] or auxiliary losses [36, 41]. ControlJEPA differs in two ways: it uses a discrete-time practical-ISS contraction condition suited to step-wise hidden-state dynamics, and enforces it as a soft trajectory-level penalty rather than an architectural constraint.

8 Concluding Remarks

We presented ControlJEPA, a trajectory regularization framework built on a discrete-time ISS-type contraction condition. Its Lyapunov Tube Loss penalizes violations of this condition on normalized transversal deviation. Wherever the condition holds, it implies an explicit tube-width bound. In the partial-ISS regime used in practice, the effective width is governed by the penalty strength and the contraction rate. Theoretical analysis identifies the partial-ISS regime as the balance point between geometric regularity and representational capacity. Empirically, ControlJEPA consistently improves accuracy and data efficiency over standard fine-tuning and prior trajectory regularizers across six benchmarks and six model families. A smoother transversal profile also persists at inference time, suggesting that the method reshapes the representation manifold rather than memorizing trajectories.

Limitations and future work. Our guarantees apply only to trajectories observed during training; inference-time stability remains empirical, experiments are limited to models up to 8B parameters, and semantic anchors depend on dataset-provided segmentation. Future work includes vision and multimodal extensions and Lyapunov-constrained decoding.

Acknowledgments and Disclosure of Funding

This paper is dedicated to Professor Richard G. Baraniuk on the occasion of his 60th birthday.

This research is supported by the National Research Foundation Singapore under the AI Singapore Programme (AISG Award No: AISG2-TC-2023-012-SGIL). This research is also supported by the Ministry of Education, Singapore, under the Academic Research Fund Tier 1 (FY2023) (A-8002040-00-00, A-8002039-00-00), the NUS Presidential Young Professorship Award (A-0009807-01-00), the NUS Artificial Intelligence Institute–Seed Funding (A-8003062-00-00), and the Cross Faculty Grant 2025, CFG25-012 (A-8004460-00-00).

References

[1] Laziz Abdullaev and Tan Nguyen. Transformer meets twicing: Harnessing unattended residual information. In *International Conference on Learning Representations*, volume 2025, pages 57583–57606, 2025.

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15619–15629, 2023.

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

[4] Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *URL https://arxiv. org/abs/2511.08544*, 10, 2025.

[5] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=xm6YD62D1Ub.

[6] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: latent video prediction for visual representation learning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum.

[7] Delong Chen, Mustafa Shukor, Théo Moutakanni, Willy Chung, Lei Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=tjimrqc2BU.

[8] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[9] Hafez Ghaemi, Eilif Benjamin Muller, and Shahab Bakhtiari. Seq-JEPA: Autoregressive predictive learning of invariant-equivariant world models. In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024. URL https://openreview.net/forum?id=MO1OLAKcsJ.

[10] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[11] Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-jepa: Towards seamless cross-dataset transfer through dynamic spatial attention. *arXiv preprint arXiv:2403.11772*, 2024.

[12] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

[13] Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse problems*, 34(1):014004, 2018.

[14] Max Hartman and Lav Varshney. Sparsejepa: Sparse representation learning of joint embedding predictive architectures. *arXiv preprint arXiv:2504.16140*, 2025.

[15] Olivier J Hénaff, Robbe LT Goris, and Eero P Simoncelli. Perceptual straightening of natural videos. *Nature neuroscience*, 22(6):984–991, 2019.

[16] Eghbal Hosseini and Evelina Fedorenko. Large language models implicitly learn to straighten neural sentence trajectories to construct a predictive representation of natural language. *Advances in Neural Information Processing Systems*, 36:43918–43930, 2023.

[17] Naiwen Hu, Haozhe Cheng, Yifan Xie, Shiqi Li, and Jihua Zhu. 3d-jepa: A joint embedding predictive architecture for 3d self-supervised representation learning. *arXiv preprint arXiv:2409.15803*, 2024.

[18] Hai Huang, Yann LeCun, and Randall Balestriero. LLM-JEPA: Large language models meet joint embedding predictive architectures. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=GbXKPo9QfH.

[19] Hai Huang, Yann LeCun, and Randall Balestriero. Semantic tube prediction: Beating LLM data efficiency with JEPA. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=M8l3MmWaQB.

[20] Zhong-Ping Jiang and Yuan Wang. Input-to-state stability for discrete-time nonlinear systems. *Automatica*, 37(6):857–869, 2001.

[21] Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=YevsQ05DEN7.

[22] Yilun Kuang, Yash Dagade, Tim GJ Rudner, Randall Balestriero, and Yann LeCun. Rectified lpjepa: Joint-embedding predictive architectures with sparse and maximum-entropy representations. *arXiv preprint arXiv:2602.01456*, 2026.

[23] Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

[24] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 6086–6096, 2019.

[25] Nicholas Locascio, Karthik Narasimhan, Eduardo DeLeon, Nate Kushman, and Regina Barzilay. Neural generation of regular expressions from natural language with minimal domain knowledge. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1918–1923, 2016.

[26] Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. In *Neural information processing systems*, 2026.

[27] Sachin Mehta, Mohammad Hossein Sekhavat, Qingqing Cao, Maxwell Horton, Yanzi Jin, Chenfan Sun, Iman Mirzadeh, Mahyar Najibi, Dmitry Belenko, Peter Zatloukal, et al. Openelm: An efficient language model family with open training and inference framework. *arXiv preprint arXiv:2404.14619*, 2024.

[28] Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. *Advances in neural information processing systems*, 37: 2348–2377, 2024.

[29] Lorenzo Mur-Labadia, Matthew Muckley, Amir Bar, Mido Assran, Koustuv Sinha, Mike Rabbat, Yann LeCun, Nicolas Ballas, and Adrien Bardes. V-jepa 2.1: Unlocking dense features in video self-supervised learning. *arXiv preprint arXiv:2603.14482*, 2026.

[30] Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent masking. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VMAHQDOtjp.

[31] Dung Viet Nguyen, Yen Pham, Hieu Vu, Lei Zhang, and Tan Nguyen. Activation steering with a feedback controller. In *International Conference on Learning Representations*, volume 2026, pages 154471–154505, 2026.

[32] Tam Nguyen, Tan Nguyen, and Richard Baraniuk. Mitigating over-smoothing in transformers via regularized nonlocal functionals. *Advances in Neural Information Processing Systems*, 36: 80233–80256, 2023.

[33] Tam Minh Nguyen, César A Uribe, Tan Minh Nguyen, and Richard Baraniuk. Pidformer: Transformer meets control theory. In *Forty-first International Conference on Machine Learning*, 2024.

[34] Thai Khanh Nguyen, Vo Ngoc Bich Uyen, Thieu Vo, Tan Minh Nguyen, and Cuong Pham. MuonSSM: Orthogonalizing state space models for sequence modeling. In *Forty-third International Conference on Machine Learning*, 2026.

[35] Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

[36] Ivan Dario Jimenez Rodriguez, Aaron Ames, and Yisong Yue. Lyanet: A lyapunov framework for training neural odes. In *International conference on machine learning*, pages 18687–18703. PMLR, 2022.

[37] Florian Schmidt. Generalization in generation: A closer look at exposure bias. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 157–167, 2019.

[38] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[39] Basile Terver, Randall Balestriero, Megi Dervishi, David Fan, Quentin Garrido, Tushar Nagarajan, Koustuv Sinha, Wancong Zhang, Mike Rabbat, Yann LeCun, et al. A lightweight library for energy-based joint-embedding predictive architectures. *arXiv preprint arXiv:2602.03604*, 2026.

[40] Hugo Thimonier, José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. T-jepa: Augmentation-free self-supervised learning for tabular data. *arXiv preprint arXiv:2410.05016*, 2024.

[41] Hiroyasu Tsukamoto and Soon-Jo Chung. Neural contraction metrics for robust estimation and control: A convex optimization approach. *IEEE Control Systems Letters*, 5(1):211–216, 2020.

[42] Ying Wang, Oumayma Bounou, Gaoyue Zhou, Randall Balestriero, Tim G. J. Rudner, Yann LeCun, and Mengye Ren. Temporal straightening for latent planning. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Ik1mKtUYlZ.

[43] Stephen Wiggins. *Introduction to applied nonlinear dynamical systems and chaos*. Springer, 2003.

[44] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[45] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3911–3921, 2018.

[46] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4791–4800, 2019.

Supplement to "ControlJEPA: Principled Trajectory Regularization via Lyapunov Tube Loss"

Table of Contents

A Theoretical Proofs and Supplementary Analysis **15**

A.1 Proof of Proposition 4.1 15

A.1.1 Setup and notations 15

A.1.2 Two lemmas 15

A.1.3 Proof of Proposition 4.1 16

A.2 Proof of Theorem 4.2 16

B Additional Experiment Results **19**

B.1 Implicit Tube Preservation at Inference Time 19

B.2 Hyperparameter Sensitivity: Contraction Rate and Orthogonal Slack 19

B.3 Hardware Settings 20

C Broader Societal Impacts **20**

A Theoretical Proofs and Supplementary Analysis

A.1 Proof of Proposition 4.1

A.1.1 Setup and notations

Throughout, $\|\cdot\|_2$ denotes the Euclidean norm on vectors, $\|\cdot\|_F$ the Frobenius norm, $\|A\|_{\mathrm{op}} = \sigma_1(A)$ the operator norm, and $\sigma_1(A) \geq \sigma_2(A) \geq \cdots$ the singular values in decreasing order, indexed from one.

Let $h_t \in \mathbb{R}^D$ denote the hidden state at position t after the final normalization and before the output projection, let $U \in \mathbb{R}^{|\mathcal{V}|\times D}$ be the unembedding matrix, and let $H \in \mathbb{R}^{T\times D}$ have rows $h_t^\top$, so that the logit matrix is $L = HU^\top$ with rows $L_t = Uh_t$ and $p_\theta(\cdot \mid x_{<t}) = \mathrm{softmax}(L_t)$.

Using the orthogonal decomposition $h_t - h_0 = \alpha_t \hat{v}\|v_{\mathrm{geo}}\|_2 + e_t$ with $e_t \perp \hat{v}$, set

$$\widetilde{H} := \mathbf{1}h_0^\top + \alpha\big(\hat{v}\|v_{\mathrm{geo}}\|_2\big)^\top, \qquad \widetilde{L} := \widetilde{H}U^\top, \qquad E := H - \widetilde{H},$$

so E has rows $e_t^\top$ and $L - \widetilde{L} = EU^\top$. Write $\mathcal{S} := \mathrm{span}(h_0, v_{\mathrm{geo}})$, so $\dim \mathcal{S} \leq 2$ and every row $\tilde{h}_t = h_0 + \alpha_t v_{\mathrm{geo}}$ of $\widetilde{H}$ lies in $\mathcal{S}$.

A.1.2 Two lemmas

Lemma A.1 (Rank of the reference logits)**.** $\mathrm{rank}(\widetilde{L}) \leq 2$.

Proof. $\widetilde{H} = \mathbf{1}h_0^\top + \alpha v_{\mathrm{geo}}^\top$ is a sum of two outer products, each of rank at most one, so $\mathrm{rank}(\widetilde{H}) \leq 2$. Since $\mathrm{rank}(AB) \leq \min\{\mathrm{rank}(A), \mathrm{rank}(B)\}$, right multiplication by $U^\top$ cannot increase rank. □

Lemma A.2 (Softmax is $\frac{1}{2}$-Lipschitz in ℓ_2)**.** *For all* $z, z' \in \mathbb{R}^{|\mathcal{V}|}$, $\|\mathrm{softmax}(z) - \mathrm{softmax}(z')\|_2 \leq \frac{1}{2}\|z - z'\|_2$.

Proof. Write $p = \mathrm{softmax}(z)$. Differentiating $p_i = e^{z_i}/\sum_j e^{z_j}$ gives the Jacobian $J(z) = \mathrm{diag}(p) - pp^\top$, which is symmetric. For any x,

$$x^\top J(z)x = \sum_i p_i x_i^2 - \Big(\sum_i p_i x_i\Big)^2 = \mathrm{Var}_p(X),$$

where X takes the value x_i with probability p_i. Hence $J(z)$ is positive semidefinite and, by symmetry and the variational characterization of the largest eigenvalue, $\|J(z)\|_{\mathrm{op}} = \lambda_{\max}(J(z)) = \sup_{\|x\|_2=1} \mathrm{Var}_p(X)$. Since X takes values in $[x_{\min}, x_{\max}]$, Popoviciu's inequality gives $\mathrm{Var}_p(X) \le (x_{\max} - x_{\min})^2/4$, and $(x_{\max} - x_{\min})^2 \le 2(x_{\max}^2 + x_{\min}^2) \le 2\|x\|_2^2$ because the extremes are attained at distinct coordinates. Therefore $\|J(z)\|_{\mathrm{op}} \le \frac{1}{2}$ for every z.

The domain $\mathbb{R}^{|\mathcal{V}|}$ is convex, so the segment $\gamma(s) = z' + s(z - z')$, $s \in [0, 1]$, lies in it. By the chain rule and the fundamental theorem of calculus, $\mathrm{softmax}(z) - \mathrm{softmax}(z') = \int_0^1 J(\gamma(s))(z - z')\, ds$, and taking norms,

$$\big\|\mathrm{softmax}(z) - \mathrm{softmax}(z')\big\|_2 \le \Big(\sup_{s\in[0,1]} \|J(\gamma(s))\|_{\mathrm{op}}\Big)\|z - z'\|_2 \le \tfrac{1}{2}\|z - z'\|_2. \qquad \square$$

A.1.3 Proof of Proposition 4.1

Proposition A.3 (Spectral capacity bound)**.** *Suppose $\|e_t\|_2 \le \delta$ for all t. Then*

$$\frac{1}{T}\sum_{j>2} \sigma_j\,(L)^2 \le \delta^2\,|U|_{op}^2, \tag{22}$$

and consequently there is a smooth manifold $\mathcal{N}$ of dimension at most 2 in the probability simplex with

$$\frac{1}{T}\sum_{t=1}^{T} \mathrm{dist}_{\ell_2}(p_\theta(.|x_{<t}), \mathcal{N})^2 \le \frac{1}{4}\delta^2\,|U|_{op}^2. \tag{23}$$

Proof. By the Eckart-Young-Mirsky theorem, the squared spectral tail of L equals its best rank-two approximation error, and $\tilde{L}$ is one admissible competitor:

$$\sum_{j>2} \sigma_j(L)^2 = \min_{\mathrm{rank}(R)\le 2} |L - R|_F^2 \le |L - \tilde{L}|_F^2 = |EU^\top|_F^2. \tag{31}$$

By submultiplicativity,

$$|EU^\top|_F^2 \le |E|_F^2 |U^\top|_{op}^2 = |E|_F^2 |U|_{op}^2 = (\sum_t \|e_t\|_2^2)|U|_{op}^2 \le T\delta^2 |U|_{op}^2, \tag{32}$$

which gives (22) after division by T. For (23), set $\mathcal{N} := \{\mathrm{softmax}(Uw) : w \in \mathcal{S}\}$, the image of $\mathcal{S}$ under the smooth map $w \mapsto \mathrm{softmax}(Uw)$; since $\dim \mathcal{S} \le 2$, this set is smoothly parametrized by at most two parameters. Because $\tilde{h}_t \in \mathcal{S}$, the point $\mathrm{softmax}(\widetilde{L}_t)$ lies in $\mathcal{N}$ and is admissible in the distance. Row-wise, $L_t - \widetilde{L}_t = U(h_t - \tilde{h}_t) = Ue_t$, so $\|L_t - \widetilde{L}_t\|_2 \le \|U\|_{\mathrm{op}}\delta$, and Lemma A.2 gives $\mathrm{dist}_{\ell_2}(p_\theta(\cdot \mid x_{<t}), \mathcal{N}) \le \frac{1}{2}\|U\|_{\mathrm{op}}\delta$. Squaring and averaging over t gives (23). $\square$

A.2 Proof of Theorem 4.2

Theorem 4.2 (Transversal Profile Characterization)**.** *Let $\{V_t\}_{t=0}^T$ be the Lyapunov sequence with $V_0 = 0$, and let $\{\varepsilon_t\}$ be the innovations defined in (25). Assume $\gamma \in (0, 1)$ and $\tau \ge 0$. Then the following hold for every trajectory, without any distributional assumption on $\{\varepsilon_t\}$:*

(i) Mean confinement. *The time-averaged Lyapunov value satisfies*

$$\bar{V}_T \;\le\; \frac{\tau}{1-\gamma} \;+\; \frac{\mathcal{L}_{\mathrm{Lyap}}(\theta)}{1-\gamma}. \tag{27}$$

(ii) Variance control. *The temporal variance of the Lyapunov sequence satisfies*

$$\mathrm{Var}_t[V_t] \;\le\; \frac{2n}{(1-\gamma)^2}\left[1 + \frac{4\gamma^2}{(1-\gamma)^2}\right]\mathcal{L}_{\mathrm{Lyap}}(\theta) \;+\; \frac{4\tau^2}{(1-\gamma)^2}\left[1 + \frac{2\gamma^2}{(1-\gamma)^2}\right] \;+\; \mathcal{R}_T, \tag{28}$$

where $\mathcal{R}_T$ is an explicit boundary remainder with $\mathcal{R}_T = O(\frac{1}{T})$ whenever M is bounded uniformly in T.

(iii) Practical tube width. *By Jensen's inequality,*

$$\mathbb{E}_t[\|e_t\|_2] \;\le\; \sqrt{\bar{V}_T}\cdot\|v_{\mathrm{geo}}\|_2 \;\le\; \|v_{\mathrm{geo}}\|_2 \cdot \sqrt{\frac{\tau + \mathcal{L}_{\mathrm{Lyap}}(\theta)}{1-\gamma}}. \tag{29}$$

Proof. Since $e_0 = 0$ by construction ($h_{t_{\text{start}}}$ is the starting anchor and $d_0 = h_{t_{\text{start}}} - h_{t_{\text{start}}} = 0$), we have $V_0 = \|e_0\|_2^2/\|v_{\text{geo}}\|_2^2 = 0$. The exact recurrence holds at every step:

$$V_{t+1} \;=\; \gamma V_t + \tau + \varepsilon_t, \qquad t = 0, 1, \ldots, T-1, \tag{33}$$

where ε_t is defined in (25).

Softplus lower bound. For every $x \in \mathbb{R}$,

$$\log(1 + e^x) \;\geq\; x^+ := \max(x, 0), \tag{34}$$

since $\log(1 + e^x) \geq \log e^x = x$ and $\log(1 + e^x) \geq \log 1 = 0$.

Proof of part (i): Mean confinement.

Sum (33) over $t = 0, \ldots, T-1$:

$$\sum_{t=0}^{T-1} V_{t+1} \;=\; \gamma \sum_{t=0}^{T-1} V_t + T\tau + \sum_{t=0}^{T-1} \varepsilon_t. \tag{35}$$

Write $\sum_{t=0}^{T-1} V_{t+1} = \sum_{t=1}^{T} V_t = \sum_{t=0}^{T-1} V_t - V_0 + V_T = \sum_{t=0}^{T-1} V_t + V_T$ (using $V_0 = 0$). Substituting into (35):

$$\sum_{t=0}^{T-1} V_t + V_T \;=\; \gamma \sum_{t=0}^{T-1} V_t + T\tau + \sum_{t=0}^{T-1} \varepsilon_t. \tag{36}$$

Rearranging and using $V_T \geq 0$:

$$(1-\gamma) \sum_{t=0}^{T-1} V_t \;=\; -V_T + T\tau + \sum_{t=0}^{T-1} \varepsilon_t \;\leq\; T\tau + \sum_{t=0}^{T-1} \varepsilon_t. \tag{37}$$

We construct the bound for the violation sum. Since $\varepsilon_t \leq \varepsilon_t^+ := \max(\varepsilon_t, 0)$ and the softplus lower bound (34) gives $\log(1 + e^{\varepsilon_t}) \geq \varepsilon_t^+$:

$$\sum_{t=0}^{T-1} \varepsilon_t \;\leq\; \sum_{t=0}^{T-1} \varepsilon_t^+ \;\leq\; \sum_{t=0}^{T-1} \log(1 + e^{\varepsilon_t}) \;=\; T\,\mathcal{L}_{\text{Lyap}}(\theta). \tag{38}$$

Substituting (38) into (37) and dividing by $T(1-\gamma)$:

$$\bar{V}_T \;=\; \frac{1}{T} \sum_{t=0}^{T-1} V_t \;\leq\; \frac{\tau}{1-\gamma} + \frac{\mathcal{L}_{\text{Lyap}}(\theta)}{1-\gamma}. \tag{39}$$

Note that the V_0/T transient term present in the general case vanishes here because $V_0 = 0$. $\square_{(i)}$

Proof of part (ii): Variance control.

Throughout, the trajectory is a single contiguous segment $\mathcal{T} = \{0, \ldots, T-1\}$ with $V_t \geq 0$, and we write $x^+ := \max(x, 0)$, $x^- := \max(-x, 0)$, so that $\varepsilon_t = \varepsilon_t^+ - \varepsilon_t^-$ and $\varepsilon_t^2 = (\varepsilon_t^+)^2 + (\varepsilon_t^-)^2$. Recall from Section 4.3 that $M = \max_t \varepsilon_t^+$ is the largest ISS violation and $n = \frac{M^2}{\log(1+e^M)}$. Let $\bar{\varepsilon} := \frac{1}{T}\sum_{t=0}^{T-1} \varepsilon_t$ and

$$P := \frac{1}{T} \sum_{t=0}^{T-1} (\varepsilon_t^+)^2, \qquad D := \frac{1}{T} \sum_{t=0}^{T-1} (\varepsilon_t^-)^2. \tag{40}$$

Dividing the equality in (37) by T gives

$$(1-\gamma)\bar{V}_T = \tau + \bar{\varepsilon} - \frac{V_T}{T}. \tag{41}$$

Let $\widetilde{V}_t := V_t - \bar{V}_T$ and $\delta_t := \varepsilon_t - \bar{\varepsilon}$. Subtracting $\bar{V}_T$ from (33) and using (41) yields $\widetilde{V}_{t+1} = \gamma \widetilde{V}_t + \delta_t + V_T/T$. Unrolling from $\widetilde{V}_0 = -\bar{V}_T$,

$$\widetilde{V}_t = S_t + R_t, \qquad S_t := \sum_{s=0}^{t-1} \gamma^{t-1-s} \delta_s, \qquad R_t := -\gamma^t \bar{V}_T + \frac{V_T(1-\gamma^t)}{T(1-\gamma)}. \tag{42}$$

By $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathrm{Var}_t[V_t] = \frac{1}{T}\sum_{t=0}^{T-1} \widetilde{V}_t{}^2 \le \frac{2}{T}\sum_{t=0}^{T-1} S_t^2 + \mathcal{R}_T, \qquad \mathcal{R}_T := \frac{2}{T}\sum_{t=0}^{T-1} R_t^2. \tag{43}$$

The sequence S_t is the convolution of $(\gamma^k)_{k\ge 0}$, whose ℓ^1-norm is $1/(1-\gamma)$, with $(\delta_s)_{s=0}^{T-1}$. Young's inequality $\|a * b\|_{\ell^2} \le \|a\|_{\ell^1}\|b\|_{\ell^2}$ gives

$$\frac{1}{T}\sum_{t=0}^{T-1} S_t^2 \le \frac{1}{(1-\gamma)^2}\,\frac{1}{T}\sum_{t=0}^{T-1}\delta_t^2 = \frac{1}{(1-\gamma)^2}\Big(\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_t^2 - \bar{\varepsilon}^2\Big) \le \frac{P+D}{(1-\gamma)^2}. \tag{44}$$

To bound P, let $q(x) := \frac{x^2}{\log(1+e^x)}$ for $x \ge 0$, so that $n = q(M)$. With $\sigma(x) := \frac{1}{1+e^{-x}}$, the derivative of $\log(1+e^x)$,

$$q'(x) = \frac{x\big(2\log(1+e^x) - x\,\sigma(x)\big)}{\log^2(1+e^x)} \ge 0,$$

since $\log(1+e^x) \ge x \ge x\,\sigma(x)$ by (34). Hence q is nondecreasing on $[0,\infty)$. Whenever $\varepsilon_t > 0$, we have $0 < \varepsilon_t \le M$ and therefore $(\varepsilon_t^+)^2 = q(\varepsilon_t)\log(1+e^{\varepsilon_t}) \le q(M)\log(1+e^{\varepsilon_t}) = n\log(1+e^{\varepsilon_t})$; the same inequality is trivial when $\varepsilon_t \le 0$. Averaging,

$$P \le n\,\mathcal{L}_{\mathrm{Lyap}}(\theta). \tag{45}$$

To bound D, note that $V_{t+1} \ge 0$ implies $\varepsilon_t^- \le \gamma V_t + \tau$, so $D \le 2\gamma^2\,\frac{1}{T}\sum_{t=0}^{T-1} V_t^2 + 2\tau^2$. From $V_{t+1} \le \gamma V_t + \tau + \varepsilon_t^+$ and $V_0 = 0$,

$$0 \le V_t \le \frac{\tau}{1-\gamma} + \sum_{s=0}^{t-1}\gamma^{t-1-s}\varepsilon_s^+.$$

Squaring and applying Young's inequality as in (44) gives $\frac{1}{T}\sum_{t=0}^{T-1} V_t^2 \le \frac{2(\tau^2+P)}{(1-\gamma)^2}$. Therefore

$$D \le 2\tau^2 + \frac{4\gamma^2(\tau^2+P)}{(1-\gamma)^2}. \tag{46}$$

Combining (43), (44) and (46),

$$\mathrm{Var}_t[V_t] \le \frac{2P}{(1-\gamma)^2}\Big[1 + \frac{4\gamma^2}{(1-\gamma)^2}\Big] + \frac{4\tau^2}{(1-\gamma)^2}\Big[1 + \frac{2\gamma^2}{(1-\gamma)^2}\Big] + \mathcal{R}_T,$$

and (45) yields (28).

Finally, from (42), using $\sum_{t\ge 0}\gamma^{2t} = 1/(1-\gamma^2)$ and $(1-\gamma^t)^2 \le 1$,

$$\mathcal{R}_T \le \frac{4\bar{V}_T^2}{T(1-\gamma^2)} + \frac{4V_T^2}{T^2(1-\gamma)^2}. \tag{47}$$

If $M \le M_0$ for all T, then unrolling $V_{t+1} \le \gamma V_t + \tau + M_0$ gives $V_t \le (\tau + M_0)/(1-\gamma)$ for all t. Hence $\bar{V}_T$ and V_T are bounded uniformly in T, and (47) gives $\mathcal{R}_T = O(1/T)$. $\square_{(ii)}$

Proof of part (iii): Practical tube width.

By Jensen's inequality applied to the concave function $\sqrt{\cdot}$:

$$\mathbb{E}_t[\|e_t\|_2] \;=\; \mathbb{E}_t\Big[\sqrt{V_t}\,\|v_{\mathrm{geo}}\|_2\Big] \;\le\; \sqrt{\mathbb{E}_t[V_t]}\cdot\|v_{\mathrm{geo}}\|_2 \;=\; \sqrt{\bar{V}_T}\cdot\|v_{\mathrm{geo}}\|_2. \tag{48}$$

Substituting (27) (with $V_0 = 0$):

$$\mathbb{E}_t[\|e_t\|_2] \;\le\; \|v_{\mathrm{geo}}\|_2\cdot\sqrt{\frac{\tau + \mathcal{L}_{\mathrm{Lyap}}(\theta)}{1-\gamma}}. \tag{49}$$

$\square_{(iii)}$

$\square$

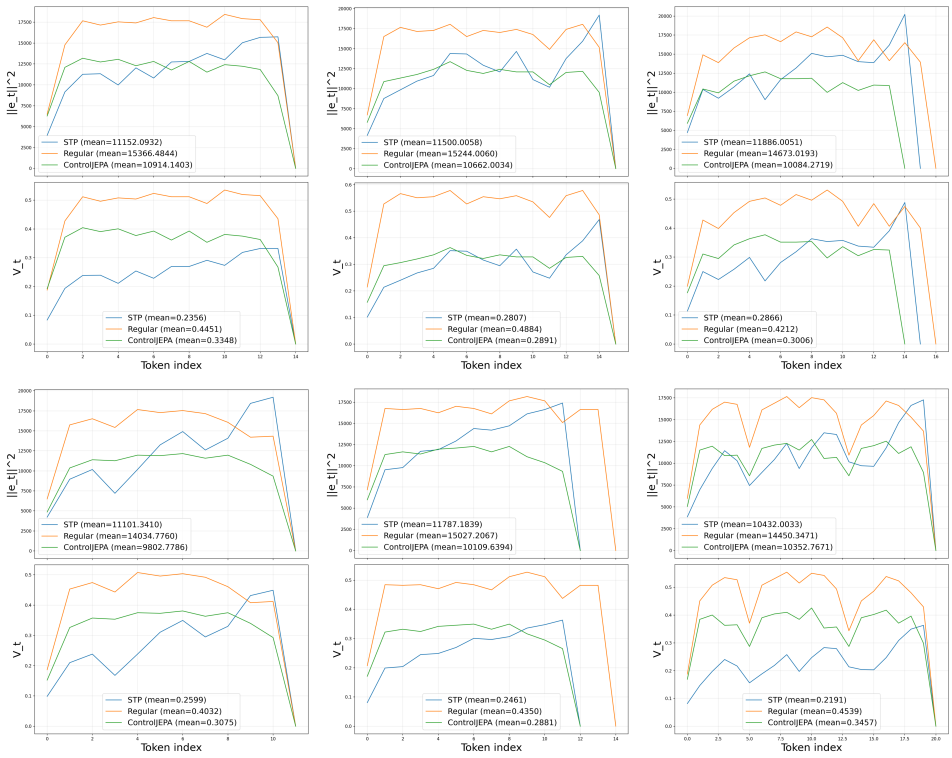


Figure 6: Step-by-step evolution of the unnormalized transversal error $\|e_t\|_2^2$ (top panels) and the dimensionless Lyapunov candidate V_t (bottom panels) across multiple generated sequences. With the Lyapunov loss deactivated during generation, ControlJEPA (green) uniquely maintains a tight, stable semantic tube. In contrast, regular fine-tuning (orange) exhibits severe geometric drift, and STP (blue) suffers from high trajectory volatility.

B Additional Experiment Results

B.1 Implicit Tube Preservation at Inference Time

To further investigate the intrinsic geometric structure learned by the models, we visualize the trajectory dynamics during autoregressive generation (Figure 6). Crucially, during inference, all regularization losses are deactivated. The hidden states are propelled solely by the vector field learned by the network weights.

Baseline Drift: The standard fine-tuned model (Regular, orange) exhibits substantial transversal drift across all samples, confirming the susceptibility of unconstrained models to inference-time Brownian motion and mode collapse.

STP Volatility: While Semantic Tube Prediction (STP, blue) generally reduces the mean error compared to the baseline, its generation trajectories are highly erratic. This empirical volatility is a direct consequence of its heuristic formulation: by solely penalizing local angular deviations without imposing global magnitude bounds, STP fails to construct a monotone structural boundary, allowing the transversal error to oscillate wildly.

ControlJEPA Stability: In stark contrast, ControlJEPA (green) consistently traces the smoothest trajectories with the lowest mean transversal errors. Despite the absence of an active Lyapunov penalty during decoding, the hidden states remain tightly confined within the semantic tube. This provides compelling evidence that our Input-to-State Stability (ISS) optimization does not merely memorize training paths; rather, it fundamentally reshapes the representation manifold by embedding a global attractor that implicitly corrects transversal drift step-by-step at inference time.

B.2 Hyperparameter Sensitivity: Contraction Rate and Orthogonal Slack

To comprehensively evaluate the robustness and behavior of the decoupled Lyapunov hyperparameters, we perform a grid search over the contraction rate γ and the orthogonal slack τ. As illustrated in Figure 7, ControlJEPA exhibits a broad region of high measured accuracy across most combinations, freeing practitioners from brittle hyperparameter tuning. Crucially, an analysis of the peak performance

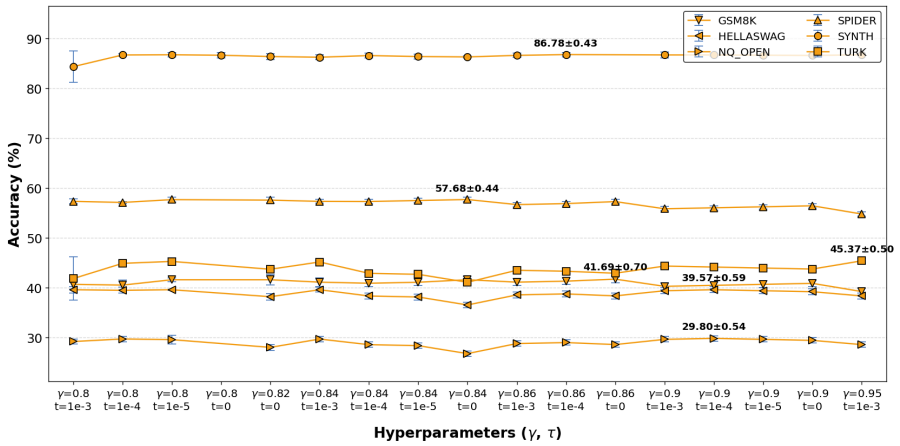


Figure 7: The accuracy landscape across diverse datasets demonstrates a broad stability region. Tasks with rigid, rule-based structures (e.g., SYNTH) peak at tighter tube configurations ($\gamma \in [0.84, 0.9]$), while semantically diverse open-domain tasks (e.g., NQ-Open) require a more relaxed envelope ($\gamma \geq 0.9$ with larger τ) to accommodate complex manifold variations.

regions reveals a profound connection between the optimal tube geometry and the intrinsic semantic diversity of the downstream task:

Rigid Geodesics for Rule-Based Tasks: For tasks governed by strict syntactic rules or deterministic logic with low semantic variance (e.g., SYNTH), accuracy peaks at a tighter configuration ($\gamma \in [0.84, 0.9]$ combined with smaller τ). Because the ground-truth geodesic for these tasks is highly rigid, a faster contraction rate aggressively squashes transversal noise and forces strict adherence to the target trajectory, yielding maximal accuracy.

Relaxed Envelopes for Semantic Diversity: Conversely, for open-domain tasks characterized by high semantic diversity, complex reasoning, and linguistic nuance (e.g., NQ-Open, HellaSwag), the optimal configuration systematically shifts toward a more relaxed boundary ($\gamma \in [0.9, 0.95]$ coupled with larger τ). High-diversity tasks naturally possess meandering semantic trajectories. Enforcing an overly aggressive contraction would over-regularize the representations, leading to the dimensionality collapse discussed in Section 6.1. A larger γ (slower contraction) and a larger τ (greater slack) provide the necessary geometric volume to accommodate these rich manifold variations while still preventing catastrophic trajectory drift.

B.3 Hardware Settings

We conducted all experiments on two NVIDIA H100 80GB GPUs.

C Broader Societal Impacts

This work is primarily methodological and theoretical: it introduces ControlJEPA, a trajectory regularization framework based on discrete-time Input-to-State Stability for improving the geometry, stability, and data efficiency of hidden-state trajectories in language models. Its societal impact is therefore indirect. On the positive side, by providing explicit and interpretable control over trajectory deviation, ControlJEPA may help researchers design more reliable and sample-efficient models, reduce trial-and-error fine-tuning, and better diagnose representation-level failure modes such as trajectory drift, collapse, or instability. However, improved language-model performance and data efficiency may also amplify broader risks associated with capable generative systems, including misinformation, misuse, overreliance, and deployment in high-stakes settings without adequate safeguards. Our method does not directly address alignment, fairness, privacy, factuality, robustness to adversarial inputs, or harmful-content prevention; the theoretical guarantees apply to the controlled hidden-state geometry rather than to safe or truthful model behavior. Thus, ControlJEPA should be viewed as a foundation for improving representation stability, not as a standalone safety guarantee for deployed language models.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's three main contributions: the ISS-based trajectory regularization framework, the Lyapunov Tube Loss, and the theoretical/empirical analysis of stability and representational capacity. These claims are supported by theoretical derivations in Sections 3–4 and empirical validation in Sections 5– 6.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly discusses limitations in the "Limitations" paragraph of Section 8, including lack of guarantees for out-of-distribution trajectories, inference-time stability remaining empirical, scalability only tested up to 8B models, and dependence on dataset-provided semantic anchors.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theoretical statements are formally presented with numbered definitions, propositions, and theorems. Assumptions are explicitly stated and complete proofs are provided in Appendix A.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper specifies datasets, model families, baseline methods, training objectives, and hyperparameters (λ, γ, τ) in Section 5.1. Additional implementation and proof details are included in the appendix and supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide anonymized code, datasets, training scripts, and reproduction instructions in the Supplementary Material. All experimental assets required to reproduce the main results are included in the released repository.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper details datasets, model scales, baselines, and tuned hyperparameters in Section 5.1. It also explains the optimization behavior and hyperparameter sensitivity in Section 5.4.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars for the main experimental results in Figures 3–5. The reported error bars reflect variability across experimental runs and are described alongside the corresponding figures in Sections 5 and 6

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have provided information about computing resources needed in Appendix B.3

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research involves theoretical analysis and fine-tuning of publicly available language models and datasets, with no human subject experiments or unsafe deployment claims. The work appears consistent with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper discusses both positive and negative societal impacts in Appendix C. It highlights improved reliability and data efficiency for language models, while acknowledging broader risks such as misuse, misinformation, and unsafe deployment of capable generative systems.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper proposes a training-time regularization loss and does not release new models, datasets, or tools that pose risks for misuse.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All models, benchmarks, and baseline methods are properly cited with their original references throughout the paper and in the bibliography.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper releases an implementation of ControlJEPA together with documentation, configuration files, and reproduction instructions in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing or research with human subjects, so IRB approval is not applicable.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, a declaration is not required.

Answer: [N/A]

Justification: LLMs were used only as a writing assistant for editing and polishing the manuscript (grammar, clarity, concision).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.