THE JEPA PARADOX IN LANGUAGE: THE GEOMETRY OF LINGUISTIC ALTERNATIVES

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint-Embedding Predictive Architectures (JEPAs) are effective for images, video, and audio, yet deterministic JEPA-style latent prediction has not become a standard objective for text encoders. We argue that this gap reflects a mismatch between squared-error latent prediction and the conditional structure of language. The key requirement is conditional concentration: given a context and target location, the target representation should lie near a single meaningful point. Local image prediction often satisfies this through spatial continuity, whereas masked text can admit multiple valid token or span completions whose representations need not share a coherent center. We formalize this mismatch through three conditions—predictability, non-collapse, and low conditional variance—and show how their failure creates centroid degeneracy and collapse pressure in text. Matched I-JEPA and T-JEPA experiments reveal the predicted sequence: mutual-information saturation and elevated target variance precede train–validation instability, effective-rank degeneration, cosine collapse, and poor downstream transfer. The same pattern appears across five independent data seeds, indicating that it is not a sampling artifact. These results do not rule out predictive learning for language; they show that text-compatible JEPA objectives must preserve multiple plausible completions rather than compress them into a single latent point.

1 INTRODUCTION

Self-supervised learning often succeeds by predicting missing information from context. In language, this principle is usually implemented as prediction over discrete symbols or probability distributions, as in masked language modeling (MLM) (Devlin et al., 2019), causal language modeling (CLM) (Radford et al., 2019; Brown et al., 2020), denoising pretraining (Lewis et al., 2020; Raffel et al., 2020), and replaced-token detection (Clark et al., 2020). In vision, Joint-Embedding Predictive Architectures instead avoid reconstructing raw inputs: a context encoder observes a masked view, a target encoder embeds the unmasked signal, and a predictor learns to match the target representation in latent space (LeCun, 2022; Assran et al., 2023). This latent-prediction paradigm has achieved strong results on continuous modalities, including images, video, and audio (Assran et al., 2023; Bardes et al., 2024; Tuncay et al., 2025). Yet deterministic JEPA-style latent prediction has not become a standard recipe for text, where the dominant objectives remain distributional.

We argue that this gap reflects a statistical mismatch. JEPA is well suited to targets that are *conditionally concentrated*: given a context and a target position, the target representation should lie near a single geometrically meaningful value. Natural images often satisfy this condition locally because nearby pixels and patches are correlated, and masked regions are constrained by surrounding content (Feige, 2015; He et al., 2022; Assran et al., 2023). Token-level language does not. The same textual context can admit many valid continuations, reflected in the persistent entropy and perplexity of language even under strong models (Shannon, 1951; Brown et al., 2020). As illustrated in Figure 1, an image patch typically has a low-variance conditional target, whereas a masked token may correspond to multiple valid lexical or semantic continuations whose embeddings occupy different directions. A squared-error JEPA predictor must return a single vector, and therefore regresses toward a centroid over heterogeneous continuations rather than a coherent linguistic target.

We formalize this mismatch through three necessary conditions for useful JEPA learning: **predictability**: the target must have low conditional entropy given the context (spatial smoothness

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

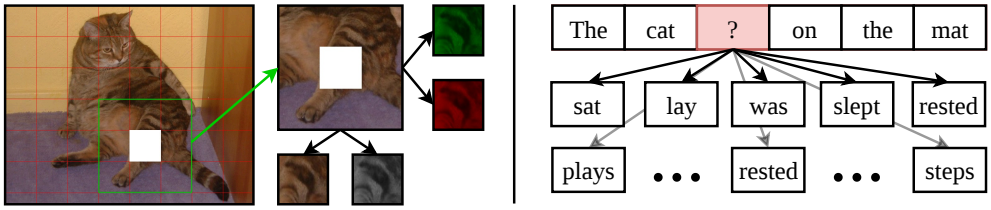


Figure 1: Conditional concentration explains the image-text divide in JEPA. In images, spatial continuity constrains masked patches to a low-variance conditional distribution. In text, masking leaves many valid continuations whose embeddings occupy distinct directions in representation space, forcing a squared-error predictor to regress toward a centroid over heterogeneous continuations rather than a coherent target.

supports this in images; lexical ambiguity imposes an irreducible floor in text); **non-collapse**: the objective must force distinct inputs to occupy distinct regions of representation space (visual diversity creates such pressure, whereas text permits low-rank collapse without geometric penalty); and **low conditional variance**: the reducible loss component must dominate the irreducible one (in images, residual uncertainty is limited to local texture; in text, multiple valid continuations dominate the loss and drive the predictor toward centroid degeneracy).

We validate these predictions by constructing a language analogue of JEPA, which we denote T-JEPA[1], and comparing it with I-JEPA (Assran et al., 2023) under matched training protocols, tracking mutual-information proxies, effective rank, pairwise cosine similarity, train-validation loss divergence, irreducible target variance, and downstream transfer. T-JEPA exhibits early information saturation followed by unstable optimization, rank divergence, cosine-similarity collapse, and poor transfer, while I-JEPA maintains increasing information, non-degenerate rank, and stable generalization – suggesting that T-JEPA fails not from insufficient capacity or tuning, but because deterministic squared-error latent prediction is misaligned with the conditional geometry of token-level language.

Our contributions are threefold. First, we introduce *conditional concentration* as a condition under which deterministic latent prediction yields useful representations. Second, we identify centroid degeneracy as a failure mode of squared-error latent prediction under ambiguous text continuations. Third, we empirically diagnose T-JEPA through information, rank, cosine-similarity, variance, and downstream metrics, showing that information saturation precedes representation collapse and motivating distributional, contrastive, mixture-based, or semantic-level alternatives. Our results do not imply that predictive representation learning is impossible for language. Rather, they identify the missing inductive bias: language prediction must preserve the multimodal structure of valid continuations instead of compressing them into a single latent centroid.

2 Related Work

Self-supervised representation learning. Self-supervised learning has developed along several complementary paradigms. Contrastive methods such as CPC/InfoNCE (van den Oord et al., 2018), SimCLR (Chen et al., 2020), and MoCo (He et al., 2020) learn by distinguishing compatible views from negative examples. Non-contrastive methods such as BYOL (Grill et al., 2020), Barlow Twins (Zbontar et al., 2021), and VICReg (Bardes et al., 2022) remove explicit negatives through architectural asymmetry, stop-gradient mechanisms, or variance–covariance regularization. Joint-Embedding Predictive Architectures (LeCun, 2022) form a third family: instead of reconstructing inputs or contrasting views, they predict target representations in latent space from a masked context.

JEPA-style latent prediction across modalities. I-JEPA demonstrated that masked latent prediction can learn strong visual representations by predicting the embeddings of masked image regions

[1]We use T-JEPA only as a local abbreviation for our text-based JEPA instantiation. It is unrelated to prior methods with the same name for trajectory similarity computation and tabular representation learning (Li et al., 2024; Thimonier et al., 2025).

from visible context (Assran et al., 2023). Subsequent work has extended JEPA-style objectives to video and world modeling (Bardes et al., 2024; Assran et al., 2025; Maes et al., 2026), audio and speech (Fei et al., 2023; Tuncay et al., 2025), and multimodal text-image settings (Vo et al., 2025; Chen et al., 2025). Recent variants also study stability, regularization, and collapse prevention in JEPA-like objectives (Balestriero & LeCun, 2025; Mo & Tong, 2024). This growing family shows that latent prediction is not limited to images. However, most successful JEPA-style applications involve continuous, spatial, temporal, or cross-modal structure, where masked targets are often constrained by nearby context. Our work asks why the same deterministic latent-prediction recipe has not become a standard general-purpose objective for training text encoders.

Representation collapse and variance preservation. Collapse is a central failure mode in self-supervised learning, where representations become constant or occupy a low-dimensional subspace (Jing et al., 2022). Existing methods avoid collapse through negatives, predictor asymmetry, EMA targets, redundancy reduction, or explicit variance constraints (Chen et al., 2020; He et al., 2020; Grill et al., 2020; Zbontar et al., 2021; Bardes et al., 2022). Among these, SIGReg (Balestriero & LeCun, 2025) is a particularly principled and effective collapse-prevention mechanism, but because it constrains only the marginal representation distribution rather than any context-conditional quantity, our analysis in App. K shows it is not, by itself, guaranteed to yield predictable or non-degenerate representations when applied to text. JEPA-style methods inherit stabilization from stop-gradient and EMA target encoders, but this stabilization is most effective when the prediction task supplies a structured target signal. We identify a complementary route to collapse: when the conditional target distribution is multimodal, squared-error latent prediction can reduce loss by moving toward a conditional mean or by allowing the encoder to merge distinctions among plausible targets. Thus, our analysis connects collapse to the geometry of the target distribution, not only to optimization dynamics or architectural symmetry.

Language pretraining preserves uncertainty. Language pretraining has instead converged on objectives that preserve uncertainty over discrete symbols. Masked language modeling (Devlin et al., 2019), causal language modeling (Radford et al., 2019; Brown et al., 2020), denoising sequence-to-sequence pretraining (Lewis et al., 2020; Raffel et al., 2020), and replaced-token detection (Clark et al., 2020) predict token distributions, corrupted-token labels, or reconstructed sequences rather than a single continuous target. data2vec also predicts contextualized latent teacher representations from masked inputs and was evaluated across speech, vision, and language (Baevski et al., 2022), making it an important precursor to modality-general latent prediction. More recent text-involving JEPA variants, such as task-specific Text-JEPA for NL-to-FOL conversion and multimodal TI-JEPA/VL-JEPA models (Le et al., 2026; Vo et al., 2025; Chen et al., 2025), show that JEPA-style objectives can incorporate language in specific settings. Two very recent efforts, LLM-JEPA (Huang et al., 2026) and DLLM-JEPA (Nam, 2026), apply a JEPA-style loss directly to text, but in both cases the JEPA term is trained only as an auxiliary addition on top of a distributional generative anchor rather than observed in isolation, so neither resolves why a purely latent-prediction objective transfers from continuous modalities to text at all; we return to a detailed comparison with both systems in App. J. They do not remove the central question studied here: when is a masked textual target itself well represented as a single deterministic latent point?

3 JEPA AS DETERMINISTIC LATENT PREDICTION

We study JEPAs as deterministic latent predictors. A JEPA observes a masked context, encodes the unmasked target with a separate target encoder, and trains a predictor to match the target representation in latent space. This captures the standard image JEPA setting (Assran et al., 2023) and defines the text analogue analyzed in this work.

Let $x = (x^{(1)}, \ldots, x^{(N)})$ be an input decomposed into N units. For images, units are non-overlapping patches on a two-dimensional grid; for text, units are discrete tokens in a sequence. A masking procedure samples disjoint context and target sets $C, T \subseteq [N]$. The context view x_C is passed through a context encoder f_θ, producing

$$z_C = f_\theta(x_C). \tag{1}$$

The full unmasked input x is passed through an EMA target encoder $f_{\bar{\theta}}$, and the target representation at position $j \in T$ is

$$z_T^{(j)} = f_{\bar{\theta}}(x)^{(j)} \in \mathbb{R}^d. \tag{2}$$

The target encoder is updated by exponential moving average,

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta, \text{ with } \tau \in (0,1), \tag{3}$$

and is not differentiated through. A predictor g_ϕ receives the context representation and a positional query p_j, then predicts

$$\hat{z}_T^{(j)} = g_\phi(z_C, p_j). \tag{4}$$

The training objective is the squared-error latent prediction loss

$$\mathcal{L}_{\text{JEPA}}(\theta,\phi) = \mathbb{E}_{x,C,T}\left[\frac{1}{|T|}\sum_{j\in T}\left\| g_\phi(f_\theta(x_C), p_j) - \text{sg}\Big(f_{\bar{\theta}}(x)^{(j)}\Big)\right\|_2^2\right], \tag{5}$$

where $\text{sg}(\cdot)$ denotes stop-gradient. Stop-gradient and EMA stabilize training, but the objective does not explicitly require the representation to remain informative, high-rank, or semantically separated.

The key property of Eq. 5 is that it is a point-prediction objective. For fixed (z_C, p_j), the squared-error optimal predictor is the conditional mean

$$g_\phi^\star(z_C, p_j) = \mathbb{E}\Big[z_T^{(j)} \mid z_C, p_j\Big]. \tag{6}$$

Thus, deterministic JEPA is most suitable when the conditional mean is a representative target. This requires *conditional concentration*: given the context and target location, valid target representations should cluster around a single meaningful point. When this holds, squared-error prediction provides a stable semantic learning signal. Images often satisfy this condition because neighboring patches, object boundaries, textures, and spatial continuity strongly constrain masked regions.

Masked text is less likely to be conditionally concentrated. The same context may admit many valid lexical or semantic completions, while the positional query specifies only where the missing content belongs. The resulting target distribution can therefore be broad or multimodal, making its conditional mean a centroid over incompatible alternatives rather than a coherent completion. Preserving these distinctions weakens point prediction; averaging them away encourages indistinct representations and, ultimately, collapse.

4 Theoretical Analysis: When Does JEPA Work?

We analyze deterministic JEPA through three necessary conditions: predictability (P), non-collapse (NC), and low conditional variance (LV). These conditions are naturally supported in local image prediction, where spatial smoothness and content-dependent visual structure constrain masked targets. In masked text, however, the same context may admit multiple valid token or span completions. If these completions are represented as distinct latent targets, the conditional distribution is not concentrated around a single point; if they are mapped close together, the representation risks collapse. Formal statements and proofs appear in App. B.1- B.3; metric definitions appear in App. C.

The three conditions are not independent failure modes. They describe a single mechanism. If the context does not concentrate the target, the squared-error predictor learns a conditional mean over alternatives. If the representation preserves those alternatives, the loss contains irreducible conditional variance; if the representation removes them, the model reduces loss by merging distinctions. Thus, ambiguity creates pressure toward either noisy point prediction or representational collapse.

4.1 Argument I: Predictability Requires Conditional Concentration

JEPA can reduce its squared-error loss only when the target representation is predictable from the context. By the Data Processing Inequality, latent representations cannot contain more predictive information than the raw inputs from which they are computed. Thus, the usefulness of deterministic latent prediction is limited by the conditional structure of the data.

For images, local smoothness makes nearby patches informative about one another. Under a Lipschitz target encoder, nearby image patches have nearby target representations:

$$\mathbb{E}\left[\left\| f_{\bar{\theta}}(x)^{(j)} - f_{\bar{\theta}}(x)^{(i)} \right\|_2^2\right] \leq L^2 d(i,j)^2. \tag{7}$$

When the context contains patches near j, the masked target is therefore conditionally constrained by visible content. For text, a masked token or span can have several valid completions under the same context. This ambiguity does not automatically imply large latent variance, since an encoder could collapse distinct completions. The key trade-off is therefore: preserving textual distinctions makes the target less predictable as a single point, while removing them reduces uncertainty by discarding information.

Validation. We track the InfoNCE mutual-information proxy $\hat{I}(z_C; z_T) \geq \log N - \mathcal{L}_{\text{InfoNCE}}$ (Poole et al., 2019; van den Oord et al., 2018) together with effective rank. A saturated MI proxy with non-trivial rank suggests limited predictive information; simultaneous rank collapse suggests that the model is reducing uncertainty by merging distinctions.

4.2 Argument II: Non-Collapse Requires a Variance-Preserving Signal

Stop-gradient and EMA target encoders stabilize JEPA training, but the squared-error objective does not explicitly enforce non-degenerate covariance. Whether collapse is discouraged depends on whether the prediction task itself penalizes collapsed representations. In images, a collapsed context representation cannot adapt to content-dependent target variation. If visual targets retain nonzero variation beyond position, then a constant context representation incurs prediction error bounded below by the target variance. This gives image JEPA a natural anti-collapse signal: different visual contexts are needed to predict different masked regions. Text admits a different route. A text encoder can reduce the conditional variance of ambiguous completions by mapping distinct but plausible targets into a lower-dimensional representation. This can preserve low JEPA loss while weakening the distinctions the representation should encode. Thus, collapse in text can arise not only from optimization dynamics, but from the objective's incentive to make multimodal targets easier to predict as a single point.

Validation. We track effective rank,

$$\operatorname{erank}(\Sigma_z) = \exp\left(-\sum_i p_i \log p_i\right) \tag{8}$$

where

$$p_i = \lambda_i / \sum_j \lambda_j \tag{9}$$

along with mean pairwise cosine similarity, its 95th percentile, and its standard deviation over held-out representations (see App. E). Collapse is indicated by decreasing effective rank, increasing cosine similarity, and shrinking cosine dispersion.

4.3 Argument III: Squared-Error Prediction Learns a Conditional Mean

The JEPA loss decomposes into reducible approximation error and irreducible conditional variance:

$$\mathbb{E}\left[\left\| g_\phi(z_C, p_j) - z_T^{(j)} \right\|_2^2\right] = \mathbb{E}\left[\left\| g_\phi(z_C, p_j) - \mathbb{E}[z_T^{(j)} \mid z_C, p_j] \right\|_2^2\right] + \mathbb{E}\left[\operatorname{Var}(z_T^{(j)} \mid z_C, p_j)\right]. \tag{10}$$

The first term can be reduced by improving the predictor; the second is irreducible for fixed representations. Deterministic JEPA is therefore well matched to conditionally concentrated targets.

For images, the residual uncertainty of a masked patch is often limited to local texture and fine detail, so the conditional variance remains small. For text, let S be a masked span and let $s \sim p(s \mid x_C)$ be a candidate completion. If $h_S(s; x_C)$ denotes the target representation induced by completing the span with s, then the squared-error optimal predictor is

$$g_\phi^\star(z_C, S) = \sum_s p(s \mid x_C) h_S(s; x_C). \tag{11}$$

Thus, the predictor returns a centroid over plausible span completions. If multiple completions have non-negligible probability and separated target representations, this centroid need not correspond to any coherent completion. The issue is not autoregression: even when all tokens in a span are predicted in parallel, deterministic squared-error prediction compresses a multimodal conditional distribution into a single Euclidean point.

Validation. We estimate $\widehat{\mathrm{Var}}(z^* \mid z_C, S)$ using $K = 16$ candidate completions per context. For text, completions are sampled from a frozen masked-language-model oracle and passed through the target encoder. For images, target variability is estimated from augmented variants of the masked region. Estimates are averaged over held-out contexts on both training and validation splits.

5 Empirical Evidence

5.1 Experimental Setup

We compare image and text instantiations of the same deterministic latent-prediction framework. The goal is not to optimize either modality independently, but to evaluate whether the diagnostics predicted by our theory appear under matched JEPA-style training.

I-JEPA. For the image setting, we train an I-JEPA model (Assran et al., 2023) on 100K ImageNet-1K images (Deng et al., 2009), using a 90/10 train-validation split and seed 42 . Images are resized to a short side of 256 and randomly cropped to 224×224. Both context and target encoders use a ViT-H/16 backbone (Dosovitskiy et al., 2021) with hidden dimension $d = 1{,}280$, 32 layers, and 16 attention heads. The target encoder is updated by EMA ($\tau = 0.996$). We follow block-wise spatial masking, using encoder scale 0.85-1.0, predictor scale 0.15-0.2, and 4 target blocks per image.

T-JEPA. For the text setting, we construct a direct text analogue, T-JEPA, trained on 100K English C4 sentences (Raffel et al., 2020), using a 90/10 train–validation split and maximum sequence length 256. Main-text diagnostics use seed 42. To test robustness, we repeat T-JEPA training with five independent seeds, each resampling its own 100K-sentence subset; the resulting MI, variance, rank, loss, and cosine trajectories are reported in App. D. The context and target encoders use a BERT-Large backbone (Devlin et al., 2019) with hidden dimension $d = 1{,}024$ and 24 layers. The target encoder is updated by EMA with $\tau = 0.996$. The predictor is a 6-layer Transformer with hidden dimension 384 and 8 attention heads. We apply random non-overlapping span masking with 1–5 spans per sentence and span lengths of 1–5 tokens.

Training and diagnostics. Both models are trained for 15 epochs with AdamW, using a 10-epoch linear warmup from 2×10^{-4} to 10^{-3} followed by cosine decay to 10^{-6}, gradient clipping at 0.3, and weight decay increasing from 0.04 to 0.4. This schedule was chosen to make collapse dynamics observable: fixed learning rates or shorter warmups caused T-JEPA to collapse before the diagnostics could resolve the transition. We monitor five quantities predicted by Sec. 4 – the InfoNCE mutual-information proxy $\widehat{I}(z_C; z_T)$ (MoCo queue of 2,048, temperature 0.1), effective rank, pairwise cosine similarity, train–validation JEPA loss, and conditional target variance – which serve as complementary lenses on a single collapse dynamic, so their joint convergence on the same transition corroborates one event rather than five coincidences. Table 1 summarizes the predicted temporal failure signature: T-JEPA does not collapse first and then lose predictability; rather, limited predictability and elevated conditional variance appear before the representation degenerates.

5.2 Argument I Validation

Figure 2 compares the evolution of context–target predictability and representation rank for I-JEPA and T-JEPA. Faint curves denote training metrics and bold curves denote validation metrics; our interpretation focuses on the validation split, which is disjoint from training.

Predictability saturates before collapse. T-JEPA's validation MI proxy remains below 0.35 nats throughout training, indicating that the masked target representation is only weakly predictable from the visible context. This behavior matches the mechanism in Consequence I (App. B.1): masked

Table 1: Temporal failure signature predicted for T-JEPA. The ordering links the theory to the empirical diagnostics: limited predictability appears before collapse, suggesting that collapse is a response to ambiguous targets rather than the original cause.

Stage	Diagnostic	Interpretation
1	MI proxy saturates	Context does not concentrate the target
2	Conditional variance remains high	Plausible completions remain separated
3	Train-validation loss diverges	Optimization fits unstable signal
4	Effective rank degenerates	Encoder merges distinctions among targets
5	Downstream transfer fails	Representation loses useful text structure

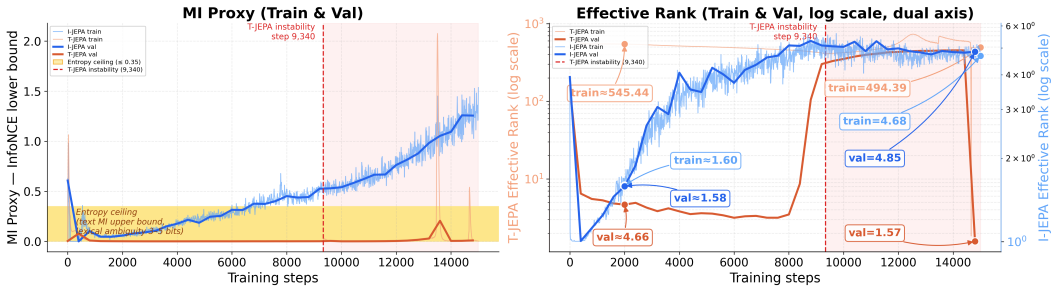


Figure 2: Context–target predictability and representation rank for I-JEPA and T-JEPA. Left: InfoNCE mutual-information proxy $\widehat{I}(z_C;\, z_T)$. Right: effective rank of the representation covariance Σ_z. Faint curves show training metrics; bold curves show validation metrics. The red dashed line marks T-JEPA's instability point at step 9,340.

text can leave several plausible completions, so the target representation need not concentrate around a single latent point. I-JEPA shows the opposite pattern. Its validation MI increases steadily, consistent with local spatial structure making masked image targets more predictable from nearby context.

The ordering of the two diagnostics is critical. T-JEPA's MI proxy saturates early, while its effective rank is still non-trivial ($\approx$4.66 at step 2,000); collapse occurs only later, near the instability point at step 9,340. Thus, low predictability is not an artifact of an already-collapsed representation. Instead, predictability fails first, and representation degeneration follows. This supports the central causal chain of our analysis: lack of conditional concentration weakens the latent prediction signal, creating pressure toward later collapse.

5.3 Argument II Validation

MSE Loss. Figure 3 (left) shows a clear divergence between the two models. I-JEPA's losses decrease smoothly in tandem, reflecting the geometric anti-collapse tension of Prop. B.8. T-JEPA's losses co-decrease initially, then diverge catastrophically at step 9,340: validation loss spikes while training loss becomes erratic; a pattern of sustained instability rather than clean descent, consistent with the degenerate solution of Prop. B.10: starved of a meaningful gradient signal by the irreducible entropy floor, the optimiser memorises batch-specific noise while generalisation collapses entirely.

Effective Rank. Figure 3 (right) reveals the representational mechanism underlying the loss divergence. I-JEPA's training and validation ranks grow jointly and converge near 4.7-4.85, confirming that the encoder learns to spread representations across multiple directions without collapse. T-JEPA's trajectory is starkly different: the *validation* rank stabilises near 4.66 from step $\sim$2,000, then undergoes a sharp upward spike at step 9,340, briefly plateaus, before collapsing catastrophically to 1.57, while the *training* rank artificially inflates to 494.39-545.44. This train-val divergence directly instantiates $\lambda_{\min}(\Sigma_z) \to 0$ on the generalisation distribution as predicted by Prop. B.10: the training objective is simultaneously gamed by overfitting to noise, while the encoder loses all representa-

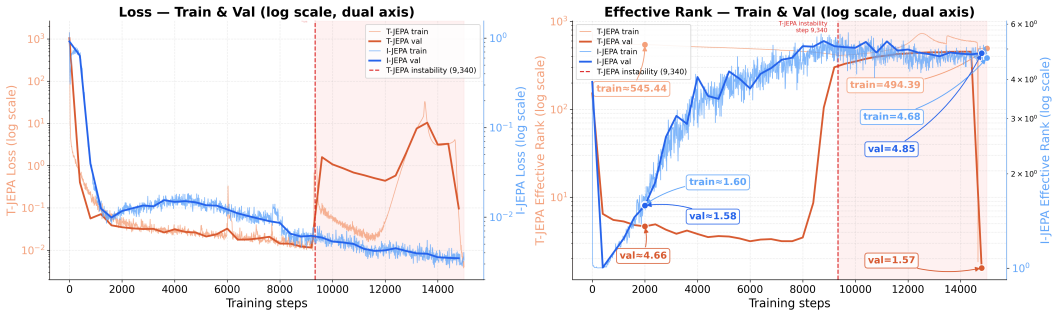


Figure 3: Loss and Effective Rank for I-JEPA and T-JEPA (training and validation, log scale). *Left:* JEPA training and validation loss curves. *Right:* Effective rank of Σ_z. Red dashed line marks T-JEPA's instability point (step 9,340).

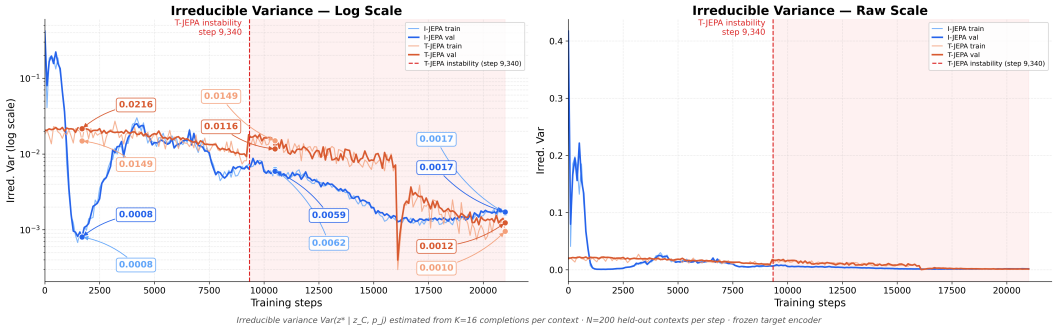


Figure 4: Irreducible variance $\widehat{\mathrm{Var}}(z^* \mid z_C, p_j)$ for I-JEPA and T-JEPA (log and raw scale). T-JEPA's irreducible variance remains persistently higher than I-JEPA's throughout training and collapses to near-zero only after representational collapse at step $\sim$15,000. The red dashed line marks T-JEPA's instability point (step 9,340).

tional structure on held-out data. Loss and rank together constitute the collapse signature: training loss *decreases* precisely *as* validation rank collapses, confirming the degenerate solution is loss-free.

5.4 Argument III Validation

I-JEPA's irreducible variance initially falls sharply to 0.0008 (on both train and val splits) near step 1,500 (Figure 4). However, instead of a monotonic decline, it rebounds and plateaus alongside T-JEPA until roughly step 7,000, before decisively dropping again to establish a clear gap. By step 9,340, I-JEPA's variance falls to 0.0062 (train) and 0.0059 (val), eventually stabilising near 0.0017 for both splits - a small residual consistent with $\sigma^2_{\text{texture}}$ (Prop. B.14). T-JEPA's variance, meanwhile, remains elevated at 0.0149 (train) and 0.0116 (val) at the middle step. This represents roughly double I-JEPA's level at this step, reflecting the persistent irreducible floor imposed by lexical ambiguity (Prop. B.16) that no amount of encoder updates can reduce. After step 9,340, T-JEPA's variance drops sharply, contracting to 0.0010 (train) and 0.0012 (val) between steps 17,500 and 21,000 - not because ambiguity resolves, but because the encoder degenerates and compresses all token embeddings into a narrow region, consistent with the rank collapse reported in Sec. 5.3. This clear gap during the stable phase, followed by artificial compression at collapse, provides direct empirical support for Consequence III (App. B.3). The same qualitative ordering is reproduced across five independent T-JEPA runs: predictability remains limited before effective-rank degeneration, while the late reduction in target variance coincides with directional collapse rather than improved target concentration (App. D).

Table 2: Downstream performance of T-JEPA and self-supervised baselines across classification and retrieval benchmarks. **Bold**: best result; underline: best among SSL methods. Retrieval metrics are reported in %. ^{M}: mask augmentation; ^{R}: replace augmentation. Mean ± std are reported over 5 independent runs.

Model	IMDB		SNLI		MTEB/FEVER		MTEB/MSMARCO	
	Acc.	F1	Acc.	F1	nDCG@10	R@100	nDCG@10	R@100
BERT	**81.91** ± 1.35	**81.89** ± 2.43	**59.72** ± 1.25	**59.11** ± 2.31	**14.43** ± 1.17	**38.82** ± 1.24	**9.494** ± 0.93	**33.672** ± 1.37
Barlow Twins^{M}	64.34 ± 2.26	64.99 ± 1.29	45.55 ± 1.31	45.99 ± 2.59	0 ± 0.00	0 ± 0.00	0.1082 ± 0.009	0.419 ± 0.096
VICReg^{M}	63.12 ± 2.13	68.85 ± 2.25	43.93 ± 2.27	41.76 ± 1.14	0 ± 0.00	0 ± 0.00	0.009 ± 0.021	0.115 ± 0.035
BYOL^{M}	50.23 ± 0.72	66.63 ± 1.08	34.28 ± 1.12	17.02 ± 0.99	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00
Barlow Twins^{R}	62.02 ± 1.33	63.85 ± 1.36	43.87 ± 1.18	43.10 ± 2.21	0 ± 0.00	0 ± 0.00	0.009 ± 0.032	0.201 ± 0.042
VICReg^{R}	59.10 ± 1.22	66.56 ± 2.12	41.59 ± 1.41	41.14 ± 1.31	0 ± 0.00	0 ± 0.00	0.004 ± 0.001	0.006 ± 0.002
BYOL^{R}	61.77 ± 1.17	67.39 ± 1.24	43.99 ± 2.28	44.04 ± 1.23	0 ± 0.00	0 ± 0.00	0.021 ± 0.004	0.358 ± 0.012
T-JEPA	50.02 ± 0.13	66.68 ± 0.41	33.34 ± 0.12	16.69 ± 0.62	0 ± 0.00	0 ± 0.00	0 ± 0.00	0 ± 0.00

5.5 Downstream Transfer Across NLP Tasks

We next test whether the intrinsic failures observed above translate into poor downstream representations. We compare T-JEPA with BERT and three non-contrastive self-supervised baselines: Barlow Twins, VICReg, and BYOL. All models are pretrained on 3 million English C4 sentences using the `bert-base-uncased` tokenizer with maximum sequence length 256, we also use this model as backbone for all baselines. After pretraining, each encoder is frozen and used as a fixed feature extractor. We evaluate on two classification benchmarks, IMDB and SNLI (Maas et al., 2011; Bowman et al., 2015), and two retrieval benchmarks from MTEB, FEVER and MSMARCO (Muennighoff et al., 2023). Full hyperparameters and protocol details are reported in App. F.

Because text lacks the augmentation diversity of images, we evaluate the SSL baselines under two corruption schemes: *Mask*, where both views are independently span-masked at 15–20% of tokens, and *Replace*, where one view is the original sentence and the other has 15–20% of tokens replaced via a frozen pretrained BERT proposal distribution (constrained to differ from the originals). T-JEPA uses the same 15–20% span-masking budget.

Table 2 shows that BERT, trained with MLM, substantially outperforms all self-supervised baselines across classification and retrieval. This is consistent with our hypothesis: distributional token prediction preserves uncertainty over plausible completions, whereas deterministic latent regression compresses that uncertainty into a single target. Among the SSL baselines, Barlow Twins under masking performs best on most classification metrics, suggesting that explicit redundancy reduction is comparatively robust to span corruption. In contrast, T-JEPA reaches chance-level classification and zero retrieval performance, matching the intrinsic collapse diagnostics in Sections 5.2–4.3.

The contrast between BYOL^{M} and BYOL^{R} further supports the role of conditional ambiguity. BYOL with independent masking collapses similarly to T-JEPA, whereas BYOL with replacement corruption remains competitive with VICReg and Barlow Twins. Replacement preserves most of the sentence and keeps the two views globally aligned, reducing the need to resolve multiple missing lexical completions. Masking, by contrast, exposes the predictor to the same multi-hypothesis ambiguity that drives centroid degeneracy in T-JEPA. The activation-landscape visualizations in App. H provide a qualitative counterpart: BERT retains rich inter-token and intra-vector variation; VICReg, Barlow Twins, and BYOL^{R} retain moderate structure; BYOL^{M} is flatter; and T-JEPA is nearly uniform and low-amplitude.

6 Conclusion

Deterministic JEPA-style latent prediction works when masked targets have a meaningful conditional center. Images often satisfy this through spatial smoothness, whereas masked text admits multiple valid completions, forcing T-JEPA toward either centroid prediction or representational collapse. Our experiments show that limited predictability and elevated variance precede rank degeneration, cosine collapse, and poor transfer. Predictive learning for language must therefore preserve multiple plausible completions; related LLM-JEPA, DLLM-JEPA variants avoid this issue by retaining a generative objective alongside JEPA (App. J). Limitation is discussed in Appendix M.

AI USE STATEMENT

Generative AI tools were used to assist with drafting and editing the manuscript. All AI-assisted suggestions and outputs were reviewed by the authors for correctness, relevance, and originality. The authors take full responsibility for the final methodology, experiments, claims, and presentation.

REFERENCES

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*, pp. 15619–15629, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning, 2025.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1298–1312. PMLR, 2022.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics, 2025.

Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. Featured Certification.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, and Jian Su (eds.), *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Rich Caruana. Multitask learning. *Machine Learning*, 28(1):41–75, July 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Allen Bolourchi, Yann LeCun, and Pascale Fung. VL-JEPA: Joint embedding predictive architecture for vision-language, 2025.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*, 2020.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-JEPA: Joint-embedding predictive architecture can listen, 2023.

Uriel Feige. Why are images smooth? In *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, pp. 229–236. Association for Computing Machinery, 2015.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21271–21284, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9726–9735, 2020.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

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.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *International Conference on Learning Representations*, 2017.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. In *International Conference on Learning Representations*, 2022.

Trong Le, Phat Thai, Sang Nguyen, Minh Hua, Ngan Pham, Thang Bui, Tho Quan, and Tuan Bui. Text-JEPA: A joint embedding predictive architecture for the conversion of natural language into first-order logic. In *Computational Collective Intelligence*, volume 16138 of *Lecture Notes in Computer Science*, pp. 200–214. Springer, 2026.

Yann LeCun. A path towards autonomous machine intelligence. Technical report, Meta AI, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7871–7880, Online, 2020. Association for Computational Linguistics.

Lihuan Li, Hao Xue, Yang Song, and Flora Salim. T-jepa: A joint-embedding predictive architecture for trajectory similarity computation. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, SIGSPATIAL '24, pp. 569–572, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711077. doi: 10.1145/3678717.3691271. URL https://doi.org/10.1145/3678717.3691271.

Nicholas Locascio, Karthik Narasimhan, Eduardo DeLeon, Nate Kushman, and Regina Barzilay. Neural generation of regular expressions from natural language with minimal domain knowledge. In Jian Su, Kevin Duh, and Xavier Carreras (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1918–1923, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1197.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels, 2026.

Shentong Mo and Shengbang Tong. Connecting joint-embedding predictive architecture with contrastive self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 2014–2037, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.148.

Sangdae Nam. DLLM-JEPA: Joint embedding predictive architectures for masked diffusion language models. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 2019.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

C. E. Shannon. Prediction and entropy of printed english. *The Bell System Technical Journal*, 30 (1):50–64, 1951.

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. In *The Thirteenth International Conference on Learning Representations*, 2025.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-JEPA: Joint-embedding predictive architecture for audio representation learning. In *IEEE International Conference on Multimedia and Expo*, 2025.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2018.

Khang H. N. Vo, Duc P. T. Nguyen, Thong T. Nguyen, and Tho T. Quan. TI-JEPA: An innovative energy-based joint embedding strategy for text-image multimodal systems. In *Information and Communication Technology*, volume 2350 of *Communications in Computer and Information Science*, pp. 141–154. Springer, Singapore, 2025.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12310–12320. PMLR, 2021.

A Appendix Overview

Appendix	Contents	Page
A	Appendix Overview (this section)	14
B	Formal Proofs	14
B.1	Proofs of Argument I: The Predictability Bound	14
B.2	Proofs of Argument II: Representation Collapse	17
B.3	Proofs of Argument III: Bias-Variance Decomposition	19
C	Empirical Metrics: Definitions and Rationale	21
D	Multi-Seed Robustness Of The Collapse Diagnostics	23
E	Argument II's Supplementary Metric: Pairwise Cosine Similarity	24
F	Experimental Details for Section 5.5	25
G	Visualization: The Phenomenon of Centroid Degeneration	26
H	Visualization: Representation Activation Heatmaps Across Self-Supervised Models	26
I	Visualization: Patterns of T-JEPA's Representation Collapse	28
J	Discussion: JEPA on Top of LLMs versus JEPA Directly on Text	28
K	Discussion: SIGReg Prevents Collapse But Does Not Resolve Predictability or Centroid Degeneracy	30
L	Mathematical Notations used throughout the paper and appendices	31

Table 3: Structure of the appendix.

B Formal Proofs

B.1 Proofs of Argument I: The Predictability Bound

We first formalize the sense in which JEPA requires the masked target representation to be predictable from the visible context. Let

$$z_C = f_\theta(x_C) \tag{12}$$

and

$$z_T^{(j)} = f_{\bar{\theta}}(x)^{(j)} \tag{13}$$

denote the context representation and the target representation at position j. Since the predictor receives only z_C and a positional query p_j, the JEPA loss can be small only when $z_T^{(j)}$ is nearly determined by (z_C, p_j). Equivalently, the conditional uncertainty

$$H\left(z_T^{(j)} \mid z_C, p_j\right) \tag{14}$$

must be small relative to the target uncertainty at that position. By the chain rule of mutual information,

$$I\left(z_C;\, z_T^{(j)} \mid p_j\right) = H\left(z_T^{(j)} \mid p_j\right) - H\left(z_T^{(j)} \mid z_C, p_j\right). \tag{15}$$

Thus, useful deterministic latent prediction requires the context representation to remove a substantial fraction of the target uncertainty.

Lemma B.1 (Data-processing ceiling)**.** *For deterministic encoders f_θ and $f_{\bar{\theta}}$,*

$$I\left(z_C;\, z_T^{(j)}\right) \le I(x_C;\, x). \tag{16}$$

If the target representation at position j is local, in the sense that $z_T^{(j)}$ is a deterministic function of $x^{(j)}$, then

$$I\left(z_C;\, z_T^{(j)}\right) \le I\left(x_C;\, x^{(j)}\right). \tag{17}$$

Proof. Because $z_C = f_\theta(x_C)$ is a deterministic function of x_C and $z_T^{(j)} = f_{\bar{\theta}}(x)^{(j)}$ is a deterministic function of the full input x, the pair $(z_C, z_T^{(j)})$ is obtained from (x_C, x) by deterministic

post-processing. The Data Processing Inequality therefore gives Eq. 16: deterministic encoders cannot increase the mutual information available between the raw context and the raw input.

Under the additional locality assumption, there exists a deterministic map $h_{\bar{\theta},j}$ such that

$$z_T^{(j)} = h_{\bar{\theta},j}\left(x^{(j)}\right). \tag{18}$$

Hence z_C is a deterministic function of x_C, while $z_T^{(j)}$ is a deterministic function of $x^{(j)}$. Applying the Data Processing Inequality to these two post-processings gives

$$I\left(f_\theta(x_C);\, h_{\bar{\theta},j}(x^{(j)})\right) \le I\left(x_C;\, x^{(j)}\right), \tag{19}$$

which is Eq. 17. □

Remark B.2. Lemma B.1 gives an information ceiling, not a collapse result. It says that the latent representations cannot contain more predictive information than is already present in the raw context-target relationship. The image-text difference therefore cannot be resolved by encoder capacity alone: it depends on how strongly the visible context constrains the missing target.

B.1.1 Images: local geometry narrows the target

Natural images exhibit local spatial regularity: nearby pixels and patches tend to be correlated, and neighboring patches often belong to the same object, boundary, or texture region (Feige, 2015). We use this regularity to formalize the sense in which visible image context can geometrically restrict a masked target.

Proposition B.3 (Spatial concentration for images)**.** *Let x be a natural image divided into patches on a two-dimensional grid, and let $d(i,j)$ denote the distance between patch centers. Assume that the image distribution has bounded local variation and that the target representation is locally Lipschitz with respect to patch-level perturbations on the data manifold. Then there exists a constant $L > 0$ such that, for nearby patches i and j,*

$$\mathbb{E}\left[\left\| f_{\bar{\theta}}(x)^{(j)} - f_{\bar{\theta}}(x)^{(i)} \right\|_2^2\right] \le L^2 d(i,j)^2. \tag{20}$$

Proof. Bounded local variation means that nearby image patches have bounded expected discrepancy:

$$\mathbb{E}\left[\left\| x^{(j)} - x^{(i)} \right\|_2^2\right] \le G^2 d(i,j)^2 \tag{21}$$

for some constant $G > 0$. By local Lipschitzness of the target representation on the data manifold, there exists $L_f > 0$ such that

$$\left\| f_{\bar{\theta}}(x)^{(j)} - f_{\bar{\theta}}(x)^{(i)} \right\|_2 \le L_f \left\| x^{(j)} - x^{(i)} \right\|_2. \tag{22}$$

Squaring both sides, taking expectation, and setting $L = L_f G$ gives Eq. 20. □

Remark B.4. Proposition B.3 does not claim that images are deterministic or that every masked patch has a unique completion. It states that local visual context restricts the range of plausible target representations. When visible patches lie near the masked region, the conditional distribution of $z_T^{(j)}$ given (z_C, p_j) is expected to be concentrated up to residual texture-level uncertainty. This is the regime in which squared-error latent prediction is well aligned with the data.

B.1.2 Text: lexical alternatives disperse the target

For text, the target unit $x^{(j)}$ is a discrete token from a vocabulary $\mathcal{V}$. Even when the surrounding context is informative, several continuations may remain valid. This creates a prediction geometry different from images: uncertainty is not merely local noise around one value, but a set of plausible lexical alternatives.

Empirical lexical uncertainty. For a masked token $x^{(j)}$ and context x_C, the conditional uncertainty is

$$H\left(x^{(j)} \mid x_C\right) = -\sum_{v \in \mathcal{V}} p(v \mid x_C) \log_2 p(v \mid x_C). \tag{23}$$

A conditional perplexity of 8-32 corresponds to 3-5 bits of uncertainty per token, and classical estimates of English entropy support the broader claim that natural language retains uncertainty even under strong contextual constraints (Shannon, 1951). The precise value depends on tokenization, domain, and context length; the important point is that the conditional distribution is often not concentrated on a single continuation.

To connect lexical uncertainty to latent-space uncertainty, we require the encoder to preserve distinctions among plausible completions. Without such a condition, an encoder could trivially make the prediction problem easy by mapping distinct tokens to the same vector, which is precisely the collapse mechanism analyzed below.

Assumption 1 (Target separation). For a context x_C, define the set of plausible continuations

$$\mathcal{V}_\epsilon(x_C) = \{v \in \mathcal{V} : p(v \mid x_C) > \epsilon\}. \tag{24}$$

The target encoder preserves distinctions among plausible continuations if there exists $\Delta > 0$ such that, for all distinct $u, v \in \mathcal{V}_\epsilon(x_C)$,

$$\left\| f_{\bar{\theta}}(x_{j \leftarrow u})^{(j)} - f_{\bar{\theta}}(x_{j \leftarrow v})^{(j)} \right\|_2 \geq \Delta, \tag{25}$$

where $x_{j \leftarrow v}$ denotes the sequence obtained by placing token v at position j while keeping the context fixed.

Proposition B.5 (Non-concentration of separated text targets)**.** *Fix a context x_C and suppose there exist at least two plausible continuations $u, v \in \mathcal{V}_\epsilon(x_C)$ whose target representations satisfy Assumption 1. Then the conditional latent target distribution cannot have zero variance around a single point. In particular, any deterministic squared-error predictor must either incur nonzero conditional variance or rely on an encoder that reduces the separation between plausible continuations.*

Proof. For a fixed context, the latent target is the random variable induced by

$$v \sim p(\cdot \mid x_C), \qquad z_T^{(j)} = f_{\bar{\theta}}(x_{j \leftarrow v})^{(j)}. \tag{26}$$

If two plausible continuations have non-negligible probability and their target representations are separated by at least Δ, then $z_T^{(j)}$ places mass on at least two separated points. Therefore it cannot be concentrated at a single vector. Its expected squared error around any deterministic prediction is minimized by the conditional mean,

$$\mathbb{E}\left[z_T^{(j)} \mid z_C, p_j\right], \tag{27}$$

but the residual conditional variance is nonzero whenever separated alternatives retain nonzero mass. The only way for this variance to vanish is for the encoder to map the plausible continuations to the same, or nearly the same, latent vector, which violates the separation condition and removes the distinctions the representation was meant to preserve. □

Consequence I: Predictability Requires Concentration

For images, local spatial structure restricts the masked target to a low-variance set of plausible representations. For token-level text, the same context can leave several valid lexical alternatives. If the encoder preserves these alternatives, the latent target remains non-concentrated; if it removes them, the representation loses distinctions among valid completions. Deterministic JEPA prediction on text therefore faces a trade-off between irreducible prediction variance and loss-reducing collapse.

B.2 Detailed Proofs for Argument II: Representation Collapse

B.2.1 The collapse problem

The JEPA objective in Eq. 5 contains no explicit constraint requiring the representation distribution to remain full-rank. Stop-gradient and EMA stabilize the target branch, but they do not by themselves rule out degenerate representations. The most extreme solution is a constant representation where

$$f_\theta(x_C) = \mathbf{c}, \qquad f_{\bar{\theta}}(x)^{(j)} = \mathbf{c}, \qquad g_\phi(\mathbf{c}, p_j) = \mathbf{c}, \tag{28}$$

for all contexts x_C, inputs x, and target positions j, where $\mathbf{c} \in \mathbb{R}^d$ is fixed. This solution achieves zero JEPA loss while encoding no information about the input. More generally, partial collapse occurs when the encoder maps inputs to a low-dimensional subset of the representation space.

Definition B.6 (Representation covariance and collapse)**.** *Let $z = f_\theta(x_C)$, where x_C is drawn from the data distribution. The representation covariance is*

$$\Sigma_z = \mathbb{E}\big[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^\top\big] \in \mathbb{R}^{d\times d}. \tag{29}$$

We say that the representation collapses along a direction $u \in \mathbb{R}^d$ if

$$\mathrm{Var}(u^\top z) = u^\top \Sigma_z u \to 0. \tag{30}$$

Full collapse corresponds to $\Sigma_z = 0$. Dimensional collapse corresponds to variance vanishing along one or more directions, equivalently to a loss of effective rank in Σ_z.

B.2.2 A sufficient condition for non-collapse

Lemma B.7 (Non-collapse via directional spread)**.** *If there exists $\gamma > 0$ such that for every unit vector $u \in \mathbb{R}^d$,*

$$\mathrm{Var}(u^\top z) = u^\top \Sigma_z u \geq \gamma, \tag{31}$$

then $\lambda_{\min}(\Sigma_z) \geq \gamma$, and the representation is non-collapsed.

Proof. By the variational characterization of eigenvalues,

$$\lambda_{\min}(\Sigma_z) = \min_{\|u\|_2=1} u^\top \Sigma_z u.$$

If Eq. 31 holds for every unit vector u, then this minimum is at least γ. Conversely, if there exists a unit vector u such that $\mathrm{Var}(u^\top z) = 0$, then $u^\top z$ is constant almost surely, so the representation carries no variation along that direction. □

The question is therefore whether the prediction objective supplies pressure for directional spread. We argue that local image prediction provides such pressure through content-dependent targets, whereas token-level text admits a collapse route because many distinct continuations can be made equivalent under the latent loss.

B.2.3 Images: content-dependent prediction discourages collapse

Natural images contain spatially local structure, but they are not determined by position alone. A patch at a fixed location can contain different objects, textures, colors, and boundaries across images. If the context encoder collapses, the predictor receives no image-dependent information from z_C and is left only with the positional query p_j. Such a predictor cannot explain content-dependent variation in the target representation.

Assumption 2 (Non-degenerate visual targets). For each target position j, the target representation has nonzero variation beyond position:

$$\mathbb{E}\left[\left\| z_T^{(j)} - \mathbb{E}\left[z_T^{(j)} \mid p_j\right]\right\|_2^2\right] \geq \sigma_{\mathrm{vis}}^2 \tag{32}$$

for some $\sigma_{\mathrm{vis}}^2 > 0$.

Proposition B.8 (Collapsed image context incurs prediction error)**.** *Suppose Assumption 2 holds. If the context encoder collapses to a constant representation $z_C = \mathbf{c}$, then any JEPA predictor incurs loss at least σ^2_{vis}:*

$$\inf_{g_\phi} \mathbb{E}\left[\left\|g_\phi(\mathbf{c}, p_j) - z_T^{(j)}\right\|_2^2\right] \geq \sigma^2_{\text{vis}}. \tag{33}$$

Proof. If $z_C = \mathbf{c}$ for all inputs, then the predictor can depend only on the positional query p_j. Under squared-error loss, the optimal predictor is the conditional mean

$$g^\star_\phi(\mathbf{c}, p_j) = \mathbb{E}\left[z_T^{(j)} \mid p_j\right]. \tag{34}$$

The minimum achievable loss is therefore

$$\mathbb{E}\left[\left\|z_T^{(j)} - \mathbb{E}\left[z_T^{(j)} \mid p_j\right]\right\|_2^2\right], \tag{35}$$

which is at least σ^2_{vis} by Assumption 2. □

Remark B.9. Proposition B.8 does not claim that image JEPA is immune to collapse in all architectures or optimization regimes. It states that, when target features vary with image content, a collapsed context representation cannot match the prediction loss achievable by a content-aware representation. Spatial locality strengthens this anti-collapse pressure: neighboring visible patches reduce uncertainty about the target, so encoding visual content in z_C can lower the loss beyond what position alone permits.

B.2.4 TEXT: COLLAPSE CAN REDUCE AMBIGUITY

For token-level text, the failure mode is different. A masked context often admits several valid continuations. If the target encoder preserves these distinctions, the predictor faces a non-concentrated target distribution. If the encoder removes them, the target distribution becomes easier to predict, but the representation loses information. The JEPA loss itself does not prevent this second route.

Consider the sentences

"The cat sat on the mat."

and

"The dog lay on the rug."

They differ lexically, but share a coarse semantic pattern. A representation that preserves lexical identity must separate them; a representation that keeps only coarse semantics may map them nearby or even identically. The problem is not that such invariance is always bad. The problem is that a deterministic squared-error objective has no intrinsic mechanism for deciding which distinctions should be preserved.

Proposition B.10 (Low-rank text encodings can minimize JEPA loss)**.** *Let $\pi : \mathcal{X} \to \{1, \ldots, K\}$ be a coarse partition of text inputs into equivalence classes, with $K < d$. Suppose the context encoder factors through this partition,*

$$f_\theta(x_C) = \mathbf{e}_{\pi(x_C)}, \tag{36}$$

where $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\} \subset \mathbb{R}^d$. Then the representation covariance of $z_C = f_\theta(x_C)$ has rank at most $K - 1$, and is therefore rank-deficient whenever $K < d$. Moreover, if the target representation also depends only on the same class and target position,

$$f_{\bar\theta}(x)^{(j)} = \mathbf{t}_{\pi(x_C), j}, \tag{37}$$

then there exists a predictor g_ϕ that attains zero JEPA loss on this representation.

Proof. Since $f_\theta(x_C)$ takes values only in the finite set $\{\mathbf{e}_1, \ldots, \mathbf{e}_K\}$, the support of z_C lies in the affine span of at most K points. Therefore,

$$\operatorname{rank}(\Sigma_z) \leq K - 1. \tag{38}$$

If $K < d$, the covariance is rank-deficient.

Now assume the target representation is determined by the same equivalence class and target position. Define the predictor on each class by

$$g_\phi(\mathbf{e}_k, p_j) = \mathbf{t}_{k,j}. \tag{39}$$

Then, for every example,

$$g_\phi(f_\theta(x_C), p_j) = f_{\bar{\theta}}(x)^{(j)}. \tag{40}$$

Hence the JEPA loss is zero, despite the context representation being supported on at most K points. If the factorization holds only approximately, the same construction yields low loss while the representation remains low-rank. □

Remark B.11. Proposition B.10 does not claim that semantic compression is always undesirable. Coarse invariance can be useful for many tasks. The issue is that token-level language contains lexical, syntactic, and semantic distinctions that may all matter, while the JEPA loss only rewards predictability of the target vector. When merging alternatives reduces conditional variance, squared-error latent prediction can favor a lower-rank representation even if it discards useful linguistic structure.

Consequence II: Collapse as a Low-Variance Escape Route

For images, a collapsed context representation leaves the predictor with position alone and incurs content-prediction error whenever target features vary with visual content. For token-level text, the objective can reduce ambiguity by mapping several valid continuations or contexts into a lower-dimensional equivalence class. Thus, collapse is not merely an optimization accident: it is a low-variance escape route made available by deterministic latent prediction.

B.3 Detailed Proofs for Argument III: Bias–Variance Decomposition

B.3.1 Decomposition of the JEPA loss

For a fixed context representation and target position, write

$$z_C = f_\theta(x_C), \qquad z^* = z_T^{(j)} = f_{\bar{\theta}}(x)^{(j)}. \tag{41}$$

Even after conditioning on (z_C, p_j), the target z^* may remain random because the visible context need not uniquely determine the masked content. Define the conditional mean target

$$\bar{z}^{(j)}(z_C, p_j) = \mathbb{E}[z^* \mid z_C, p_j]\,. \tag{42}$$

Proposition B.12 (Bias–variance decomposition of JEPA loss)**.** *For any deterministic predictor g_ϕ, the per-target JEPA loss decomposes as*

$$\begin{aligned}\mathbb{E}\Big[\|g_\phi(z_C, p_j) - z^*\|_2^2\Big] = \mathbb{E}\Big[\Big\|g_\phi(z_C, p_j) - \bar{z}^{(j)}(z_C, p_j)\Big\|_2^2\Big] \\ + \mathbb{E}\Big[\Big\|z^* - \bar{z}^{(j)}(z_C, p_j)\Big\|_2^2\Big]\,.\end{aligned} \tag{43}$$

The first term is reducible approximation error. The second term is the irreducible conditional variance of the target representation.

Proof. Add and subtract the conditional mean:

$$g_\phi(z_C, p_j) - z^* = g_\phi(z_C, p_j) - \bar{z}^{(j)}(z_C, p_j) + \bar{z}^{(j)}(z_C, p_j) - z^*. \tag{44}$$

Expanding the squared norm gives two squared-norm terms and a cross-term. Conditioning on (z_C, p_j), the cross-term vanishes because $g_\phi(z_C, p_j)$ and $\bar{z}^{(j)}(z_C, p_j)$ are deterministic functions of (z_C, p_j), while

$$\mathbb{E}\Big[\bar{z}^{(j)}(z_C, p_j) - z^* \mid z_C, p_j\Big] = 0. \tag{45}$$

Taking expectation over (z_C, p_j) gives Eq. 43. □

Remark B.13. Eq. 43 is the central reason squared-error latent prediction requires conditional concentration. The predictor can reduce the first term by approximating the conditional mean, but the second term remains whenever the target distribution is not concentrated around that mean. Thus, a deterministic JEPA objective is well aligned only when the conditional mean is representative of the target distribution.

B.3.2 IMAGES: LOCAL CONTEXT YIELDS CONCENTRATED TARGETS

In local image prediction, visible neighboring patches strongly constrain the missing target. The remaining uncertainty is not zero, but it is often confined to visually compatible texture, color, or boundary details. Thus, the conditional mean remains a meaningful target rather than an average over unrelated alternatives.

Assumption 3 (Image target concentration). For natural images under the masking scheme considered, the target representation satisfies

$$\mathbb{E}\left[\mathrm{Var}\left(z_T^{(j)} \mid z_C, p_j\right)\right] \le \sigma_{\mathrm{img}}^2, \tag{46}$$

where σ_{img}^2 is small relative to the reducible prediction error during the informative phase of training.

Proposition B.14 (Image JEPA has a useful reducible signal)**.** *Under Assumption 3, the irreducible component of the image JEPA loss is bounded by σ_{img}^2. Consequently, whenever*

$$\mathbb{E}\left[\left\|g_\phi(z_C, p_j) - \bar{z}^{(j)}(z_C, p_j)\right\|_2^2\right] \gg \sigma_{\mathrm{img}}^2, \tag{47}$$

loss reduction is dominated by the reducible approximation term, so improving the predictor and context representation yields a meaningful training signal.

Proof. By Proposition B.12, the JEPA loss decomposes into a reducible approximation term and an irreducible conditional-variance term. Assumption 3 bounds the irreducible term by σ_{img}^2. When the approximation term is much larger than this bound, changes that improve $g_\phi(z_C, p_j)$ as an estimate of $\bar{z}^{(j)}(z_C, p_j)$ dominate changes in the total loss. □

Remark B.15. Assumption 3 does not require deterministic reconstruction of the masked patch. It only requires that visible context restrict the target representation enough that the conditional mean is representative. Spatial smoothness and local continuity make this plausible for images: nearby visible patches narrow the set of compatible target representations, leaving residual uncertainty mostly in fine texture or detail.

B.3.3 TEXT: MULTI-HYPOTHESIS CONDITIONAL TARGETS

For token-level text, a fixed context can admit several valid continuations. This is not a problem for distributional language modeling: masked or causal language models predict a distribution over the vocabulary, so multiple valid tokens can all receive probability mass. The issue arises for deterministic latent prediction, where the model must output a single vector under squared-error loss.

Let $x_{j \leftarrow v}$ denote the sequence obtained by inserting token v at position j while keeping the observed context fixed. Define the target representation induced by continuation v as

$$h_j(v; x_C) = f_{\bar{\theta}}(x_{j \leftarrow v})^{(j)}. \tag{48}$$

Under the conditional token distribution $v \sim p(\cdot \mid x_C)$, the latent target is the random variable

$$z^* = h_j(v; x_C). \tag{49}$$

Proposition B.16 (Conditional variance under multiple valid continuations)**.** *For a fixed context x_C, the irreducible variance of the text target is*

$$\mathrm{Var}(z^* \mid z_C, p_j) = \sum_{v \in \mathcal{V}} p(v \mid x_C) \left\|h_j(v; x_C) - \bar{z}^{(j)}(z_C, p_j)\right\|_2^2, \tag{50}$$

where

$$\bar{z}^{(j)}(z_C, p_j) = \sum_{v \in \mathcal{V}} p(v \mid x_C) h_j(v; x_C) \tag{51}$$

is the conditional mean. If the distribution assigns nonzero mass to multiple separated continuations, then this variance is nonzero.

Proof. The expression follows from the definition of conditional variance for the random variable $z^* = h_j(v; x_C)$ with $v \sim p(\cdot \mid x_C)$. If there exist two continuations u and v with $p(u \mid x_C) > 0$, $p(v \mid x_C) > 0$, and

$$\|h_j(u; x_C) - h_j(v; x_C)\|_2 > 0, \tag{52}$$

then the conditional distribution of z^* is not a point mass. Therefore its variance around the conditional mean is strictly positive. □

A quantitative lower bound follows under a separation condition.

Assumption 4 (Separated plausible continuations). For a context x_C, suppose there exist $m \geq 2$ plausible continuations $v_1, \ldots, v_m$ such that

$$p(v_i \mid x_C) \geq \alpha \qquad \text{for all } i, \tag{53}$$

and their target representations are pairwise separated:

$$\|h_j(v_i; x_C) - h_j(v_\ell; x_C)\|_2 \geq \Delta \qquad \text{for } i \neq \ell. \tag{54}$$

Proposition B.17 (Lower bound from separated alternatives)**.** *Under Assumption 4,*

$$\mathrm{Var}(z^* \mid z_C, p_j) \geq \frac{\alpha^2 m(m-1)}{2}\Delta^2. \tag{55}$$

Proof. For any random vector Z with an independent copy Z',

$$\mathrm{Var}(Z) = \frac{1}{2}\mathbb{E}\left[\|Z - Z'\|_2^2\right], \tag{56}$$

where $\mathrm{Var}(Z) = \mathbb{E}\|Z - \mathbb{E}Z\|_2^2$. Applying this identity to $Z = z^*$ and restricting the expectation to the ordered pairs (v_i, v_ℓ) with $i \neq \ell$ gives

$$\mathrm{Var}(z^* \mid z_C, p_j) \geq \frac{1}{2}\sum_{i \neq \ell} p(v_i \mid x_C)p(v_\ell \mid x_C)\Delta^2. \tag{57}$$

Since each $p(v_i \mid x_C) \geq \alpha$, the bound follows. □

The squared-error optimal predictor is therefore

$$g_\phi^\star(z_C, p_j) = \bar{z}^{(j)}(z_C, p_j) = \sum_{v \in \mathcal{V}} p(v \mid x_C)h_j(v; x_C). \tag{58}$$

For text, this point is a centroid over plausible continuations. Unless the continuations are already close in representation space, the centroid need not correspond to any actual continuation. This clarifies why language modeling objectives avoid the same failure mode: they preserve uncertainty by predicting a distribution over tokens, whereas deterministic latent MSE prediction compresses that distribution into one conditional mean. Reducing the variance requires either a distributional or mixture-valued predictor, or an encoder that brings distinct continuations closer together. The latter route is precisely the collapse mechanism analyzed in Sec. B.2.

Consequence III: Centroid Degeneracy

Squared-error latent prediction learns a conditional mean. This is appropriate when the target distribution is conditionally concentrated, as in local image prediction. For token-level text, plausible continuations can be separated in representation space. The conditional mean then becomes a centroid over alternatives rather than a representation of a single coherent target. Thus, deterministic latent prediction faces a bias–variance trade-off: preserve distinctions and incur irreducible variance, or reduce variance by collapsing the distinctions the representation should encode.

C Empirical Metrics: Definitions and Rationale

Each of the three arguments in Sec. 4 is validated by one or more diagnostic metrics tracked throughout training on held-out validation batches. We define each metric precisely below and explain why it constitutes evidence for the corresponding argument.

C.1 Metrics for Argument I: Mutual Information Proxy

InfoNCE MI proxy. Direct estimation of $I(z_C; z_T)$ is intractable in high-dimensional representation spaces. We therefore use the InfoNCE objective as a relative proxy for context–target predictability (Poole et al., 2019; van den Oord et al., 2018). For a batch or queue of N candidate targets, define

$$\widehat{I}_{\mathrm{NCE}}(z_C; z_T) = \log N - \mathcal{L}_{\mathrm{InfoNCE}}, \tag{59}$$

where

$$\mathcal{L}_{\mathrm{InfoNCE}} = -\,\mathbb{E}\left[\log \frac{\exp(s(z_C, z_T)/\tau)}{\frac{1}{N}\sum_{k=1}^{N} \exp\Big(s(z_C, z_T^{(k)})/\tau\Big)}\right]. \tag{60}$$

Here $s(\cdot,\cdot)$ is the similarity function; in our experiments we use the dot product, $s(z_C, z_T) = z_C^\top z_T$. Negatives $z_T^{(k)}$ are drawn from a MoCo-style queue of size $N = 2{,}048$, with temperature $\tau = 0.1$.

The proxy is high when z_C can reliably identify its paired target z_T among many distractors, and low when the context contains limited information about the target representation. We use it only as a comparative diagnostic across training: a sustained plateau indicates that context–target predictability has saturated. Under our hypothesis, T-JEPA should plateau early because masked text often admits several plausible continuations, whereas I-JEPA should continue improving as spatial context increasingly constrains the masked image target.

C.2 Metrics for Argument II: Effective Rank and Cosine Similarity

Effective rank. Given the representation covariance

$$\Sigma_z = \mathbb{E}\big[(z - \mathbb{E}[z])(z - \mathbb{E}[z])^\top\big], \tag{61}$$

let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0$ be its eigenvalues. We define the effective rank as the exponential of the spectral entropy:

$$\mathrm{erank}(\Sigma_z) = \exp\left(-\sum_{i=1}^{d} p_i \log p_i\right), \qquad p_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j}. \tag{62}$$

When the spectrum is spread across many directions, $\mathrm{erank}(\Sigma_z)$ is large; when most variance is concentrated in a few directions, it is small. A drop in effective rank is therefore a spectral signature of dimensional collapse, as predicted by Proposition B.10. In the degenerate case where all eigenvalues vanish, the representation is constant; in practice, we compute effective rank with standard numerical stabilization. Effective rank is computed from full-sequence context representations every 10 training steps.

Pairwise cosine similarity. As a complementary geometry-level collapse indicator, we track pairwise cosine similarities over a random subsample of $N = 2{,}048$ held-out representations:

$$\bar{s} = \frac{2}{N(N-1)} \sum_{i<j} \frac{z_i^\top z_j}{\|z_i\|_2 \, \|z_j\|_2}. \tag{63}$$

We also report the 95th percentile s_{95} and standard deviation σ_s of the pairwise cosine distribution. Non-collapsed representations occupy multiple directions, so $\bar{s}$ remains bounded away from 1 and the distribution retains nontrivial spread. Under directional collapse, representations align, driving $\bar{s} \to 1$, $s_{95} \to 1$, and $\sigma_s \to 0$. Together, effective rank and cosine statistics distinguish low-dimensional spectral collapse from mere changes in representation scale. Full results are reported in App. E.

C.3 Metrics for Argument III: Irreducible Variance Estimator

Estimating conditional target variance. We estimate the irreducible component in Proposition B.12 by sampling multiple plausible targets for the same context and measuring their dispersion in target-representation space. For each held-out context and target position, let

$$z_1^*, \ldots, z_K^* \tag{64}$$

denote K target representations obtained from alternative completions or perturbations that are compatible with the same visible context. We estimate

$$\widehat{\mathrm{Var}}(z^* \mid z_C, p_j) = \sum_{k=1}^{K} w_k \left\| z_k^* - \bar{z}^* \right\|_2^2, \qquad \bar{z}^* = \sum_{k=1}^{K} w_k z_k^*, \tag{65}$$

where $w_k = 1/K$ for uniformly sampled variants and $\sum_k w_k = 1$ for importance-weighted completions. We use $K = 16$ variants per context, average over $N = 200$ held-out contexts, and log the estimate every 100 training steps on both train and validation splits.

Images. For images, alternative targets are generated by applying small pixel-space perturbations to the same image while keeping the mask fixed:

$$x_k = \mathrm{clip}(x + \varepsilon_k,\, -1,\, 1)\,, \qquad \varepsilon_k \sim \mathcal{N}(0, \sigma_{\mathrm{aug}}^2 I), \tag{66}$$

with $\sigma_{\mathrm{aug}} = 0.2$, together with sparse sign-flip perturbations. Each perturbed image is passed through the EMA target encoder $f_{\bar{\theta}}$, and the target embeddings at the masked positions are mean-pooled to obtain $z_k^* \in \mathbb{R}^{1280}$. We then apply Eq. 65 with uniform weights.

Text. For text, alternative targets are sampled from a frozen masked-language-model oracle. Given the visible context x_C, the oracle defines a proposal distribution over span completions. We sample $K = 16$ completions whose token probabilities exceed $\varepsilon_{\min} = 0.001$, fill the masked span to obtain completed sequences $x_{S \leftarrow s_k}$, and pass each completed sequence through the EMA target encoder:

$$z_k^* = \frac{1}{|S|} \sum_{j \in S} f_{\bar{\theta}}(x_{S \leftarrow s_k})^{(j)}. \tag{67}$$

Completions are weighted by their oracle probabilities,

$$w_k = \frac{q(s_k \mid x_C)}{\sum_{\ell=1}^{K} q(s_\ell \mid x_C)}, \qquad q(s_k \mid x_C) = \prod_{j \in S} p(v_k^{(j)} \mid x_C). \tag{68}$$

We then compute Eq. 65 using these normalized weights.

A persistently larger text variance during the stable phase indicates that multiple plausible completions remain separated in target-representation space. A later drop in this variance is not by itself evidence that ambiguity has disappeared; when accompanied by rank degeneration and cosine collapse, it indicates that the encoder has reduced variance by compressing the representations themselves.

D Multi-Seed Robustness Of The Collapse Diagnostics

To verify that the collapse signature reported for T-JEPA under seed 42 in Sec. 5.1 is not an artifact of a single data sample, we repeat *T-JEPA-only* training with five independent seeds (0–4), each resampling its own 100K-sentence subset from C4 (each seed has it's own 100K sample sentences); no I-JEPA comparison is included in this appendix, as the goal here is solely to check within-model stability of the diagnostics across data draws. All five T-JEPA runs reproduce the same qualitative failure chain observed for seed 42: the MI proxy stays near zero, well below the entropy ceiling, throughout training (Figure 5, left); effective rank stays near-degenerate for the first five epochs, rises sharply and jointly across seeds between epochs 6–9, then collapses back to near 1 by epoch 12–15 (Figure 5, right; Figure 6, right); validation loss shows sharp transient spikes in the same epoch 9–11 window before settling to a low, degenerate value once rank collapses (Figure 6, left); irreducible variance declines monotonically with a common late-training contraction coinciding with rank collapse rather than resolved ambiguity (Figure 7); and pairwise cosine similarity stays pinned near 1.0, with only a brief seed-dependent dip around epochs 9–12 (Figure 8). These results confirm that the predictability-saturation-before-collapse ordering established for T-JEPA in the main text is a property of it's objective itself, not of any particular data draw.

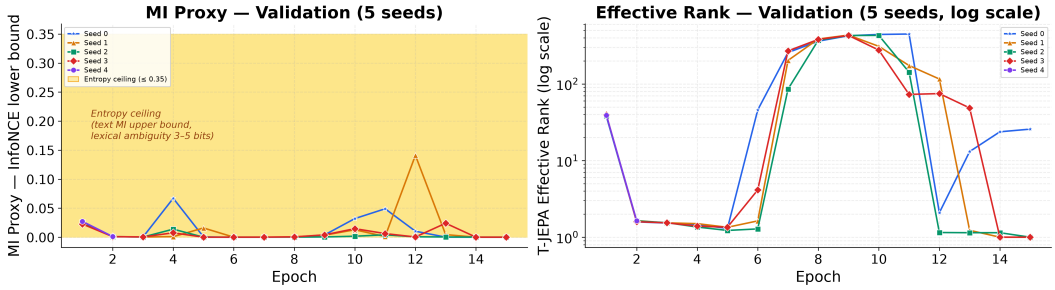


Figure 5: T-JEPA validation MI proxy (left) and effective rank (right, log scale) across 5 seeds (0–4), each independently re-sampling 100K C4 sentences, confirming early predictability saturation is not a sampling artifact.

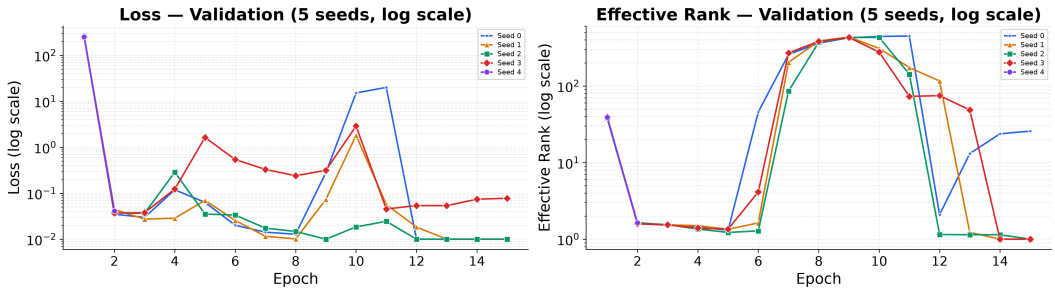


Figure 6: T-JEPA validation loss (left) and effective rank (right, both log scale) across 5 seeds (0–4), each independently re-sampling 100K C4 sentences, confirming the loss–rank collapse signature is robust to sampling variation.

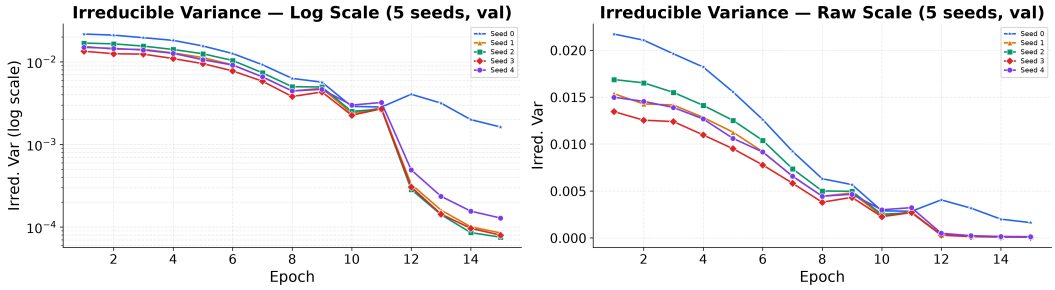


Figure 7: Irreducible variance $\widehat{\mathrm{Var}}(z^* \mid z_C, p_j)$ for T-JEPA (log scale, left; raw scale, right) across 5 seeds (0–4), each independently re-sampling 100K C4 sentences, showing the elevated conditional-variance floor is not seed-specific.

E Supplementary Metric for Argument II (Seed 42)

Cosine similarity. Figure 9 provides a geometry-level confirmation of the rank-collapse diagnostics. I-JEPA maintains a broad pairwise-cosine distribution, with substantial dispersion throughout training, indicating that different inputs continue to occupy diverse directions in representation space. T-JEPA behaves differently: after the instability point at step 9,340, its mean pairwise cosine

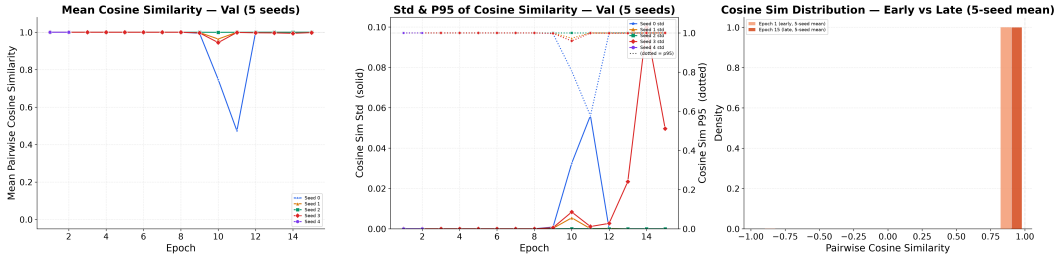


Figure 8: Pairwise cosine-similarity statistics for T-JEPA across 5 seeds (0–4, each independently resampling 100K C4 sentences): mean similarity (left), std/p95 (center), and early-vs-late histogram averaged over seeds (right), confirming directional collapse is robust to sampling variation.

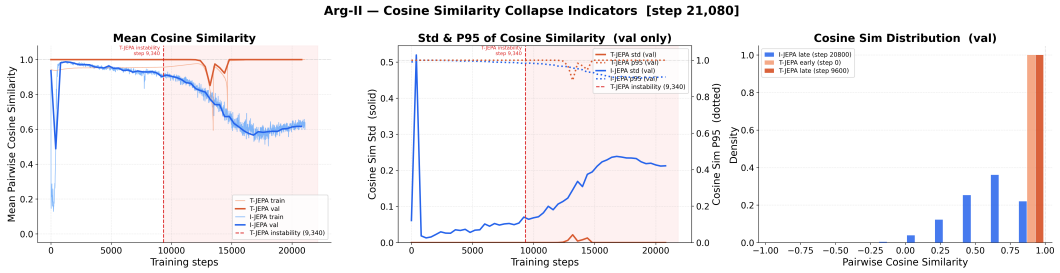


Figure 9: **Pairwise cosine similarity of I-JEPA and T-JEPA representations.** Left: mean pairwise cosine similarity over training. T-JEPA rises to ≈ 1.0 after the instability point at step 9,340, while I-JEPA remains bounded away from 1. Centre: standard deviation and 95th percentile of validation cosine similarities. After collapse, T-JEPA has vanishing dispersion and near-unit upper quantiles. Right: late-training histogram of pairwise cosine similarities. T-JEPA concentrates at 1.0, indicating directional collapse, while I-JEPA maintains a broad distribution.

similarity approaches 1, its dispersion vanishes, and its late-training histogram concentrates at the right edge. Thus, the drop in effective rank is not merely a change in representation scale; held-out inputs are mapped to nearly the same direction in latent space.

F Experimental Details for Section 5.5

Dataset and benchmarks. All models are pretrained on 3M English C4 sentences with maximum sequence length 256, using the `bert-base-uncased` tokenizer. After pretraining, encoders are frozen and evaluated as fixed feature extractors. We report classification performance on IMDB and SNLI using accuracy and F1, and retrieval performance on MTEB/FEVER and MTEB/MSMARCO using nDCG@10 and Recall@100. Mean $\pm$ std are reported over 5 independent runs.

Baselines and corruptions. We compare T-JEPA against BERT, VICReg, Barlow Twins, and BYOL. BERT is trained with standard masked language modeling. The two-view SSL baselines are trained under two text corruption schemes. In the Mask setting ($^{\mathrm{M}}$), both views are independently corrupted by random non-overlapping span masks covering 15–20% of tokens, with span lengths between 1 and 5. In the Replace setting ($^{\mathrm{R}}$), the first view is the original sentence and the second replaces 15–20% of tokens using a pretrained `bert-base-uncased` proposal model, with replacements constrained to differ from the original tokens. T-JEPA uses random span masking with 1–5 non-overlapping spans per sentence, span lengths between 1 and 5, and the same 15–20% masking budget.

Optimization. All downstream-comparison models are trained for 10 epochs with AdamW, initial learning rate 10^{-4}, and weight decay 10^{-2}. We use 100 warmup steps followed by linear learning-

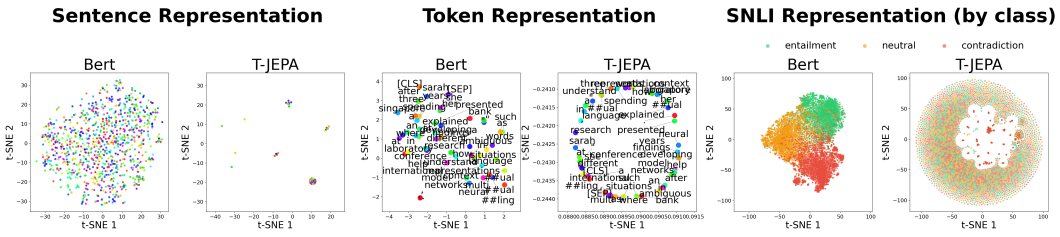


Figure 10: t-SNE visualization of representation geometry for BERT and T-JEPA. *Left:* sentence-level embeddings from the raw pretrained last hidden state, computed over 10K held-out C4 sentences. *Middle:* token-level embeddings for a single representative sentence, with sequential tokens connected by edges. *Right:* sentence embeddings after fine-tuning on SNLI, colored by label (entailment / neutral / contradiction).

rate decay to zero. VICReg uses loss weights $\lambda = 25$, $\mu = 25$, and $\nu = 1$; Barlow Twins uses redundancy-reduction coefficient $\lambda = 0.005$; BYOL uses EMA target momentum $\tau_{\text{base}} = 0.996$ annealed to 1 with cosine scheduling and its standard predictor architecture. T-JEPA uses EMA momentum $\tau = 0.9996$.

Relation to the diagnostic setup. The downstream protocol in Sec. 5.5 and the diagnostic protocol in Sec. 5.1 serve different purposes and should not be read as competing configurations. The downstream comparison is deliberately *standardized*: T-JEPA and every baseline share the same training budget, warmup length, and optimizer schedule, so no method is given a configuration advantage. The diagnostic setup instead uses a longer, 15-epoch warmup for a different reason: not to induce a failure that would otherwise not occur, but to *slow down its onset* enough that the internal sequence of events – MI saturation, elevated conditional variance, rank degeneration, and cosine collapse – can be resolved as distinct, separately measurable stages rather than a single abrupt event. Critically, T-JEPA does not avoid collapse under the shorter, standardized schedule used in Sec. 5.5; if anything, it collapses *faster* there, compressing the same five-stage sequence into a window too narrow to dissect on its own (Table 2). The two setups therefore probe the same underlying failure at two different temporal resolutions: the standardized schedule confirms that the failure occurs under realistic, no-advantage training conditions, while the extended-warmup schedule is a diagnostic instrument used to slow the failure down enough to explain *why* it happens. The longer warmup is a measurement choice, not a cause of the collapse.

G Visualization: The Phenomenon of Centroid Degeneration

Figure 10 gives a qualitative counterpart to the quantitative collapse diagnostics of Sec. 5.2–5.4. On raw pretrained C4 sentences (left), BERT embeddings spread broadly and continuously across the t-SNE plane, whereas T-JEPA embeddings collapse into a handful of tight, well-isolated clusters – a low-rank, centroid-like geometry consistent with Prop. B.10. The same degeneration is visible at the token level (middle): BERT places tokens at distinct, semantically organized positions, while T-JEPA compresses all tokens of the same sentence into a band spanning a t-SNE range of $\sim$0.003, indicating near-total loss of token-level distinctions even before any downstream adaptation. Fine-tuning on SNLI (right) does not undo this geometry: BERT's embeddings reorganize into class-conditional regions that separate entailment, neutral, and contradiction examples, whereas T-JEPA's embeddings remain confined to a narrow, ring-like manifold with all three classes fully intermixed. This shows that the representational collapse identified during pretraining is not corrected by fine-tuning, and directly explains T-JEPA's chance-level SNLI performance in Table 2.

H Visualization: Representation Activation Heatmaps

Figures 11 and 12 visualize last hidden-state activations of each context encoder on the same held-out sentence. These plots are intended as qualitative diagnostics of representation geometry, complementing the quantitative rank and cosine-similarity metrics in Sec. 4.2 and App. E.

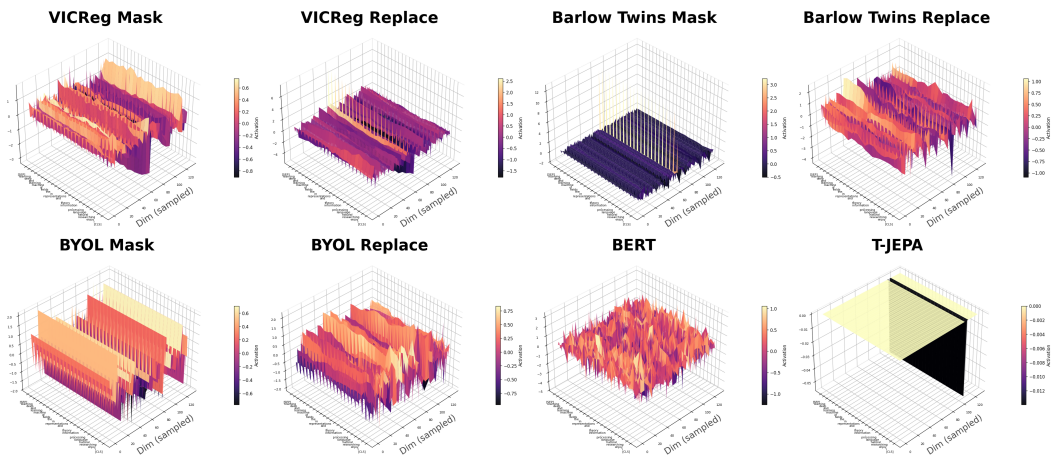


Figure 11: **3D activation surfaces across self-supervised text encoders.** Each surface shows last hidden-state activations for a held-out sentence, with token index on one axis and sampled hidden dimensions on the other. BERT exhibits strong token-wise and dimension-wise variation. VICReg, Barlow Twins, and BYOL retain visible activation structure, with variation across both tokens and dimensions. T-JEPA produces a nearly flat, low-amplitude surface, consistent with the collapse diagnostics in App. B.2.

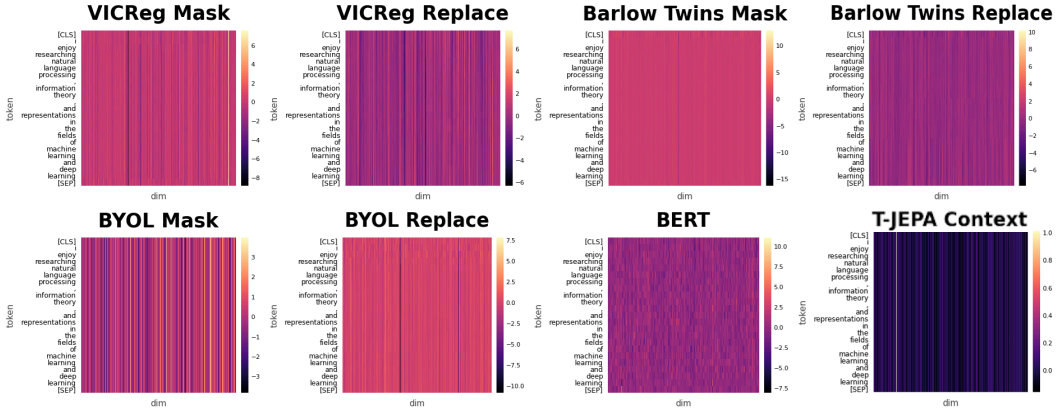


Figure 12: **Activation heatmaps across self-supervised text encoders.** Each heatmap shows last hidden-state activations as a token-by-dimension matrix for the same held-out sentence. BERT and the non-JEPA SSL baselines retain structured variation across tokens and dimensions, whereas T-JEPA is nearly uniform and low-amplitude, matching the effective-rank and cosine-collapse results.

Qualitative structure. A non-degenerate token representation should vary across tokens and use multiple hidden dimensions. BERT shows the strongest activation structure, with clear token-dependent patterns and substantial variation across dimensions. VICReg, Barlow Twins, and BYOL retain moderate structure, suggesting that their objectives preserve some token-level and dimension-level diversity under the same evaluation protocol.

T-JEPA collapse. T-JEPA is qualitatively different: its activations are nearly uniform and low-amplitude across both tokens and hidden dimensions. This visual pattern is consistent with the quantitative collapse signatures reported earlier: low effective rank, cosine-similarity saturation, and poor downstream transfer. The heatmaps therefore provide an intuitive view of the same failure mode: T-JEPA does not merely learn weaker features, but compresses the representation geometry itself.

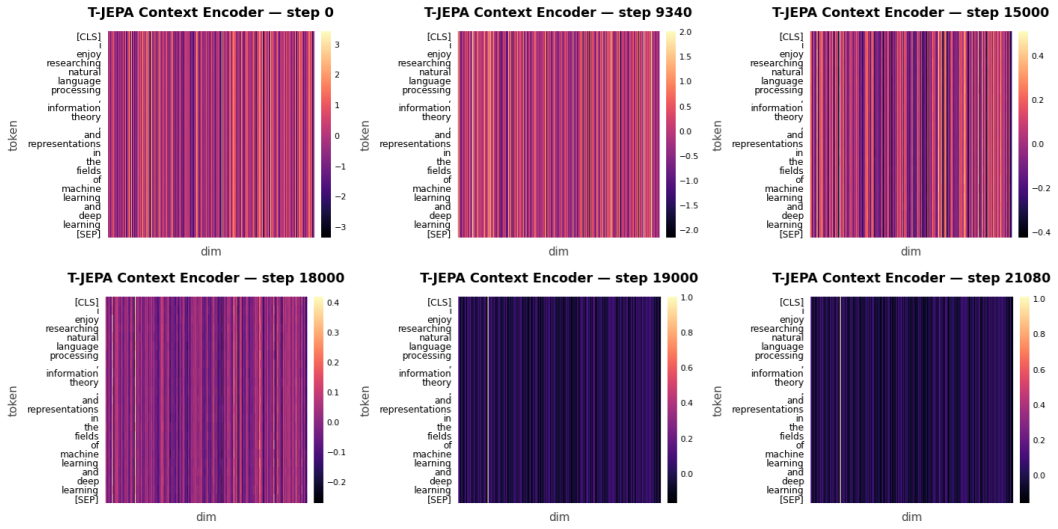


Figure 13: Per-step representation heatmaps of the T-JEPA context encoder's last hidden state for the sentence *"I enjoy researching natural language processing, information theory, and representations in the fields of machine learning and deep learning"* at steps 0, 9,340, 15,000, 18,000, 19,000, and 21,080. Rows denote tokens and columns denote hidden dimensions.

I Visualization: T-JEPA Collapse Dynamics

Figure 13 visualizes the last hidden-state activations of the T-JEPA context encoder at several training steps for the same held-out sentence. Rows correspond to tokens and columns to hidden dimensions.

Token-wise and dimensional compression. The heatmaps show two qualitative trends. First, token-specific variation is weak throughout training: different tokens share similar activation patterns, with only coarse dimension-wise structure visible. Second, training progressively compresses the activation landscape. The overall activation magnitude shrinks over time, and the remaining variation becomes concentrated in a small subset of hidden dimensions while most dimensions flatten.

This pattern complements the quantitative diagnostics in Sec. 4.2 and App. E. Effective-rank degeneration shows that the covariance spectrum loses dimensions; cosine-similarity saturation shows that different inputs align in representation space; the heatmaps provide a token-level view of the same process. T-JEPA does not merely produce lower-amplitude features. It progressively reduces both token-wise diversity and dimensional usage, consistent with the low-rank collapse predicted by Proposition B.10.

Why EMA is insufficient. The visualization also clarifies why stop-gradient and EMA do not by themselves prevent this failure mode. EMA stabilizes the target branch by making it track the context encoder slowly, but it does not impose a constraint on the variance or covariance of the representation distribution. If both branches drift toward low-rank activations, the target remains stable while still becoming less informative. Avoiding this failure requires an additional mechanism that preserves variance across examples or dimensions, such as contrastive separation, redundancy reduction, or explicit variance regularization.

J Comparison with JEPA-Augmented Language Models

Our analysis shows that a deterministic, squared-error latent prediction objective is mismatched to token-level text: it collapses ambiguous completions into a centroid corresponding to no real token. This might seem to conflict with two recent systems, LLM-JEPA and DLLM-JEPA (Huang et al.,

Table 4: Comparison of the three text-oriented JEPA studies. T-JEPA is the direct standalone baseline: JEPA applied to text with the same backbone and recipe as I-JEPA, with no auxiliary generative loss.

Property	T-JEPA (Our)	LLM-JEPA	DLLM-JEPA
Encoder (BERT-style) backbone	✓	×	×
Autoregressive decoder backbone	×	✓	×
Masked diffusion LM backbone	×	×	✓
JEPA is sole training objective	✓	×	×
JEPA is auxiliary to a generative loss	×	✓	✓
Views from masking a single input	✓	×	✓
Requires external paired dataset	×	✓	×
EMA target encoder + stop-gradient	✓	×	✓
Generative loss resolves token ambiguity	×	✓	✓
Downstream performance improves	×	✓	✓

2026; Nam, 2026), which report gains from adding a JEPA-style loss to language model training. The conflict is only apparent: in both systems the JEPA term never trains alone, and the reported ablations that remove or weaken the generative objective are themselves evidence for our thesis.

LLM-JEPA cannot drop the language-modelling loss. Huang et al. (2026) sweep the generative-loss weight γ down to zero on NL-RX-SYNTH (Locascio et al., 2016) while holding the JEPA weight $\lambda = 1$: at $\gamma = 0$ (JEPA loss alone) accuracy is 0.00%, rising monotonically with γ to 70.42% at $\gamma = 1.0$ versus 57.29% for cross-entropy-only. The authors also report that minimizing $\mathcal{L}_{\mathrm{LLM}}$ alone does not implicitly minimize the JEPA loss. This matches our theory: the generative loss keeps the representation informative about token identity, while JEPA adds structure on top. It also matches what LLM-JEPA's "views" are: `Text` and `Code` are fully-specified sequences, not a masked span with several live completions, so the token-level ambiguity of Proposition B.5 is absorbed by $\mathcal{L}_{\mathrm{LLM}}$ before JEPA ever touches the representation.

DLLM-JEPA cannot drop the diffusion loss or its asymmetric views. DLLM-JEPA's generative term $\mathcal{L}_{\mathrm{diff}}$ is itself a masked, distributional objective, structurally akin to masked language modelling, and is never removed in any reported configuration; Nam (2026) instead ablate only the JEPA side. On LLaDA-8B GSM8K (Cobbe et al., 2021), the diffusion-only baseline reaches $39.5\%/42.6\%$ (0-/4-shot). Removing the asymmetric JEPA view construction drops this slightly to $38.9\%/42.5\%$; keeping asymmetric views but replacing the predictor with identity yields only $43.2\%/44.4\%$; the full method reaches $44.9\%/61.3\%$. $\mathcal{L}_{\mathrm{diff}}$ is never weakened in this ablation — it is the JEPA-side components that prove individually insufficient. Consistently, the paper's own diagnostics show the pooled JEPA embeddings' effective rank and cosine diversity are essentially unchanged from the pre-trained base, since $\mathcal{L}_{\mathrm{diff}}$ already resolves ambiguity on its own.

T-JEPA is the missing "JEPA-only" cell. T-JEPA (Section 5) is the natural counterfactual: a pure latent-prediction loss over masked spans with no distributional term, trained under the same recipe as I-JEPA. It exhibits the full failure chain, falling to chance-level classification and zero retrieval — effectively the $\gamma = 0$ cell that Huang et al. (2026) show collapses to empty output, and the "no diffusion loss" cell that Nam (2026) never actually train.

A general pattern, not one specific to JEPA. An auxiliary latent-prediction loss helps once a primary objective already supplies task-relevant structure, but fails, or is never tried, on its own — a pattern also seen in multi-task learning (Caruana, 1997) and in auxiliary losses added to reinforcement-learning rewards (Jaderberg et al., 2017). LLM-JEPA and DLLM-JEPA fit the same template: the generative loss resolves ambiguity and keeps condition (P) satisfied, and JEPA then adds structure on top of an already well-formed representation. Table 4 summarizes this: the two properties that correlate with downstream success — a distributional generative anchor and a non-standalone JEPA objective — are exactly what T-JEPA lacks.

K Discussion: SIGReg Prevents Collapse But Does Not Resolve Predictability or Centroid Degeneracy

SIGReg (Balestriero & LeCun, 2025) is a distribution-matching regularizer that penalizes deviation of the pooled representation distribution from an isotropic Gaussian, proposed as a principled, heuristic-free alternative to stop-gradient asymmetry or explicit negatives for preventing collapse. We show it resolves Argument II but not Arguments I or III, because it constrains a *marginal* statistic while both remaining arguments are properties of the *conditional* target distribution given a single context.

SIGReg resolves Argument II. If the SIGReg penalty is minimized, $\Sigma_z \approx \sigma^2 I_d$ for some $\sigma^2 > 0$, so $u^\top \Sigma_z u \approx \sigma^2$ for every unit vector u and Lemma B.7 gives $\lambda_{\min}(\Sigma_z) > 0$: the low-rank equivalence-class solution of Prop. B.10 is ruled out.

A marginal constraint cannot certify a conditional property. Let $\mu(x_C) = \mathbb{E}[z_T^{(j)} \mid x_C]$. By the law of total variance,

$$\Sigma_z = \underbrace{\mathbb{E}_{x_C}\Big[\mathrm{Cov}(z_T^{(j)} \mid x_C)\Big]}_{\text{irreducible (Arg. III)}} + \underbrace{\mathrm{Cov}_{x_C}(\mu(x_C))}_{\text{context signal (Arg. I)}}. \tag{69}$$

SIGReg constrains only the sum. It has no mechanism that separately bounds either term, so satisfying it is compatible with any split between them.

Proposition K.1 (Isotropic marginal, vanishing context signal)**.** *Let $z_T^{(j)} = z_{\text{sig}}(x_C) + z_{\text{nui}}$ with $z_{\text{nui}} \sim \mathcal{N}(0, \sigma^2 I_d)$ independent of x_C and $\|z_{\text{sig}}(x_C)\|_2 \le \epsilon$ a.s. Then $\Sigma_z \to \sigma^2 I_d$ as $\epsilon \to 0$ (SIGReg penalty vanishes), while $I(z_C; z_T^{(j)}) \to 0$.*

Proof. As $\epsilon \to 0$, the between-context term in Eq. 69 vanishes while the within-context term saturates at $\sigma^2 I_d$, so $z_T^{(j)}$ becomes independent of x_C and $I(x_C; z_T^{(j)}) \to 0$; Lemma B.1 then gives $I(z_C; z_T^{(j)}) \to 0$. □

This is exactly the failure the MI proxy of Sec. C.1 is designed to catch, and SIGReg supplies no defense against it: Argument I remains open.

Argument III is also untouched. Even granting nonzero context signal, the within-context term of Eq. 69 is exactly the quantity Prop. B.17 bounds below under Assumption 4, and SIGReg places no constraint on it. The optimal predictor is still the centroid of Eq. 58: SIGReg forbids resolving this term by merging continuations together, but supplies no alternative mechanism for representing them as anything other than an averaged point.

Consequence IV: Non-Collapse Is Necessary; Not Sufficient

SIGReg guarantees $\lambda_{\min}(\Sigma_z) > 0$, closing the collapse route of Prop. B.10. It gives no corresponding guarantee for predictability or centroid degeneracy: an encoder can satisfy SIGReg exactly while remaining unpredictable from context (Prop. K.1) and while its Bayes-optimal predictor still collapses distinct continuations onto a single non-existent centroid (Prop. B.17). Non-collapse is necessary but, on text, not sufficient.

Future work. We predict that T-JEPA trained with an added SIGReg penalty would keep effective rank and cosine similarity healthy throughout training, yet still show early MI-proxy saturation, elevated irreducible variance, and downstream transfer well below BERT, i.e. a representation that passes geometric collapse tests without becoming predictable or semantically faithful. Testing this with the full five-metric protocol of Sec. 5 is a natural next step.

L MATHEMATICAL NOTATION

Symbol	Description	Symbol	Description
Architecture and encoders			
f_θ	Context encoder; $z_C = f_\theta(x_C)$	$f_{\bar{\theta}}$	EMA target encoder
g_ϕ	Predictor; $(z_C, p_j) \mapsto \hat{z}_T^{(j)}$	p_j	Positional query for target position j
τ	EMA momentum; $\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$	d	Latent representation dimension
$\mathrm{sg}(\cdot)$	Stop-gradient operator	$\theta, \bar{\theta}, \phi$	Context, target, and predictor parameters
Data and masking			
x	Full unmasked input	x_C	Masked context view
$x^{(j)}$	Unit at position j; patch or token	C, T	Context and target index sets
N	Number of input units or contrastive candidates	$\mathcal{V}$	Token vocabulary
$\mathcal{X}$	Input space	$d(i,j)$	Distance between image patch centers
S	Masked text span	$x_{j \leftarrow v}$	Sequence with token v placed at position j
$x_{S \leftarrow s}$	Sequence with span completion s placed in S	$\mathcal{V}_\epsilon(x_C)$	Plausible token set given context x_C
Representations and objectives			
z_C	Context representation	$z_T^{(j)}$	Target representation at position j
z^*	Target random variable	$\bar{z}^{(j)}(z_C, p_j)$	Conditional mean target
$\mathcal{L}_{\mathrm{JEPA}}$	JEPA squared-error objective	$\mathcal{L}_{\mathrm{InfoNCE}}$	InfoNCE contrastive loss
$h_j(v; x_C)$	Target representation induced by token v	$h_S(s; x_C)$	Target representation induced by span completion s
Information-theoretic quantities			
$H(X)$	Shannon entropy	$H(X \mid Y)$	Conditional entropy
$I(X;Y)$	Mutual information	$\widehat{I}_{\mathrm{NCE}}(z_C; z_T)$	InfoNCE predictability proxy
Collapse and spectral quantities			
Σ_z	Representation covariance matrix	λ_i	Eigenvalue of Σ_z
$\mathrm{erank}(\Sigma_z)$	Effective rank of Σ_z	Δ	Minimum separation between plausible targets
$\bar{s}$	Mean pairwise cosine similarity	s_{95}, σ_s	95th percentile and standard deviation of cosine similarity
Variance quantities			
$\mathrm{Var}(z^* \mid z_C, p_j)$	Irreducible conditional target variance	σ^2_{img}	Bound on image conditional target variance
σ^2_{vis}	Nonzero visual target variation beyond position	$\widehat{\mathrm{Var}}(z^* \mid z_C, p_j)$	Empirical target-variance estimate
K	Number of sampled variants or completions	w_k	Importance weight for sampled completion k
General notation			
$\mathbb{R}^d$	d-dimensional Euclidean space	$\mathbb{E}[\cdot]$	Expectation
$\mathcal{N}(0, \sigma^2 I)$	Isotropic Gaussian distribution	$[N]$	Index set $\{1, \ldots, N\}$

Table 5: Notation used throughout the paper and appendices.

M LIMITATIONS

While our study clearly outlines why standard JEPA struggles with language, there are a few practical boundaries to our work to keep in mind. First, we deliberately used an artificially long training schedule - specifically stretching our warmup to 10 epochs - just so we could watch the model's collapse in slow motion and measure each stage clearly. Under standard, shorter training conditions, this failure actually happens much faster and is harder to unpack. Second, we ran our tests using

standard, mid-sized models like BERT-Large trained on up to three million sentences, meaning we cannot guarantee how these dynamics might shift if scaled up to massive, multi-billion parameter sizes. We also focused strictly on the basic, deterministic version of JEPA to definitively prove why it misfires on ambiguous, token-level text. Because of this strict focus, we did not try to invent a working alternative, such as building new mixture-based predictors or masking at the continuous semantic level, though our findings strongly point to those as the next logical steps. Ultimately, our work is meant to serve as a diagnostic tool that pinpoints the root cause of the problem, leaving the development of a complete cure to future research.