UNDERSTANDING EARLY COLLAPSE IN PREDICTIVE WORLD-MODEL PRETRAINING

Sofiane Ennadir*
King AI Labs, Microsoft Gaming
sofiane.ennadir@king.com

Levente Zólyomi*
NXAI GmbH
levente.zolyomi@nx-ai.com

Oleg Smirnov*
King AI Labs, Microsoft Gaming
oleg.smirnov@microsoft.com

ABSTRACT

JEPA-style predictive models are promising foundations for world models, yet they exhibit a surprising early-training pathology: the prediction loss drops rapidly while learned representations remain useless for downstream tasks. We call this the *collapse phase*. It arises because the exponential moving average (EMA) keeps the target encoder too close to the main encoder, making prediction trivial and allowing the model to minimize its objective without extracting meaningful structure. We derive an upper bound showing that collapse risk depends on the interplay between momentum dynamics and masking strategy, and that escape requires the encoder's updates to outpace the EMA's smoothing. Empirically, we show the collapse phase appears across images and time-series sensor data, lasting thousands of steps in each case. Our analysis provides a diagnostic metric for detecting collapse during training and explains why certain hyperparameter regimes prolong it. This reframes collapse not as an objective flaw but as a transient regime with predictable onset and recovery, offering practitioners a tool to monitor and understand early training dynamics in predictive world models.

1 INTRODUCTION

Learning to predict latent representations of unobserved content from partial context is central to building world models that can anticipate, plan, and reason about dynamic environments. Joint-Embedding Predictive Architectures (JEPAs) have emerged as a promising framework for this goal, operating entirely in a learned latent space rather than reconstructing in input space (LeCun, 2022). By predicting embeddings of masked regions from visible context, JEPAs suppress low-level noise and focus on semantically meaningful structure. Instantiations of this paradigm have shown strong results across images (Assran et al., 2023), video (Bardes et al., 2024), and time series (Ennadir et al., 2025) domains, where predictive world models are essential.

Yet JEPA training exhibits a puzzling behavior. Figure 1 illustrates this pattern on CIFAR-10: the reconstruction loss drops rapidly in early epochs while downstream accuracy remains near chance. The model appears to be learning, yet its representations are uninformative. Only after a prolonged period does performance begin to recover, eventually reaching competitive levels. This pattern, which we term the *collapse phase*, recurs across modalities and architectures, suggesting a systematic phenomenon rather than an implementation artifact.

Understanding this collapse matters for both theoretical and practical reasons. On the theoretical side, it reveals a gap in our understanding of how EMA-based objectives shape representation learning. The standard intuition holds that the slow-moving target encoder provides stable supervision, preventing the trivial solutions that plague naive self-prediction (Grill et al., 2020; Chen & He, 2021). But if this mechanism is so effective, why does training pass through a phase where representations carry no information? On the practical side, the collapse phase represents wasted computation. For expensive

*Equal contribution.

pretraining runs common in video world models, this inefficiency compounds. Without understanding what governs collapse duration, we cannot reliably shorten it.

In this work, we provide a principled analysis of this collapse phase. We argue that the phenomenon arises when the encoder and EMA encoder produce outputs that are too similar, rendering the prediction task trivial. The predictor can then minimize loss through a near-identity mapping, achieving low reconstruction error without learning meaningful structure. We formalize this intuition by tracking the discrepancy between encoder and EMA encoder outputs throughout training, deriving bounds that connect collapse risk to the EMA decay parameter and the masking ratio. Our analysis reveals that collapse is not a failure of the JEPA objective but a transient regime induced by conservative EMA updates during early training. This perspective complements recent theoretical work on JEPA regularization (Balestriero & LeCun, 2025), which focuses on preventing complete collapse through distributional constraints rather than characterizing the transient dynamics studied in this paper. We finally validate these insights empirically where we show the validity of the derived insights. Overall, our contributions can be summarized in the following points:

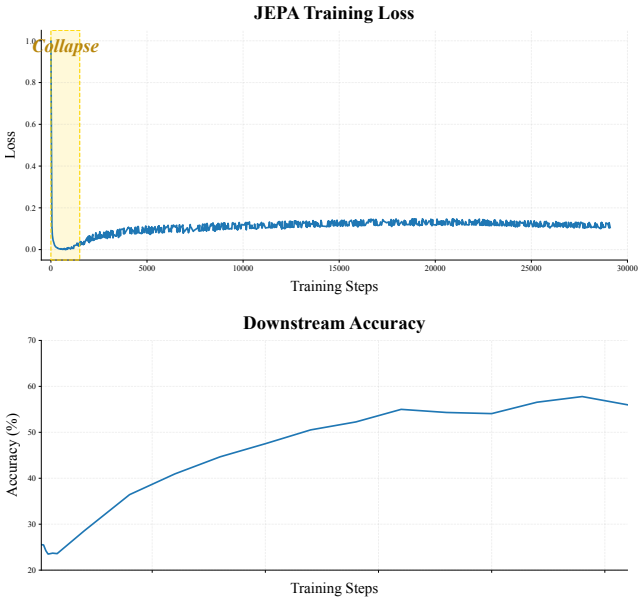

Figure 1: **The collapse phase in JEPA training.** A JEPA trained on CIFAR-10 exhibits a characteristic pattern: the training loss drops rapidly in early epochs (yellow region), yet downstream linear classification accuracy remains near chance, indicating uninformative representations. Only after several thousand steps does accuracy begin to recover.

- We provide a formal characterization of representation collapse in JEPA, defining collapse through the discrepancy between encoder and EMA encoder outputs.
- We derive theoretical bounds connecting collapse risk to the EMA decay parameter and masking ratio, explaining when and why collapse occurs.
- We empirically validate these insights across modalities, confirming the link between collapse dynamics and model hyperparameters.

2 Related Work

Self-supervised learning has evolved through several paradigms, with recent interest converging on predictive architectures as foundations for world models. Contrastive methods such as SimCLR (Chen et al., 2020) and MoCo (He et al., 2020) learn by pulling augmented views together while pushing apart different images, relying on large batches or memory banks for negative samples. Non-contrastive approaches emerged to eliminate this dependence: BYOL (Grill et al., 2020) uses an asymmetric predictor, SimSiam (Chen & He, 2021) demonstrates that stop-gradients suffice, and Barlow Twins (Zbontar et al., 2021) decorrelates representation dimensions.

Joint-Embedding Predictive Architectures offer a conceptually distinct approach, predicting the representation of one signal from a compatible signal in latent space (LeCun, 2022). This formulation sidesteps input-space reconstruction and can leverage spatial, temporal, or cross-modal structure. I-JEPA (Assran et al., 2023) instantiates this for images by predicting embeddings of masked regions from visible context. V-JEPA (Bardes et al., 2024) extends the approach to video through spatiotemporal masking, demonstrating that latent prediction scales to dynamic environments central to world modeling. The paradigm has since been adapted to audio (Fei et al., 2023), point clouds (Saito et al., 2025), time-series (Ennadir et al., 2025), and tabular data (Thimonier et al., 2025).

Despite strong empirical results, JEPAs present a fundamental theoretical challenge: understanding the conditions that prevent the encoder from collapsing inputs to a trivial constant representation. Recent work has clarified both why the JEPA paradigm merits study and how permanent collapse can be avoided. Van Assel et al. (2025) compare joint embedding and reconstruction objectives through closed-form analysis, proving that joint embedding impose strictly weaker alignment conditions when irrelevant features have large magnitude, as is typical in natural images. This theoretical advantage motivates investment in understanding JEPA training, but does not itself preclude collapse. Mo & Tong (2024) demonstrates empirically and theoretically that the EMA mechanism alone is insufficient, integrating variance-covariance regularization to address both complete collapse and deficiencies in learning mean representations. LeJEPA (Balestriero & LeCun, 2025) proves that isotropic Gaussian embeddings minimize downstream prediction risk and introduces a regularizer enforcing this distribution, eliminating the need for momentum-based target networks entirely. Radial-VCReg (Kuang et al., 2025) takes an alternative approach, augmenting variance-covariance constraints with a term aligning embedding norms capturing higher-order structure. Littwin et al. (2024a) analyze implicit bias in deep linear networks, revealing that JEPAs preferentially learn influential features with high regression coefficients, which manifests only with sufficient depth. EC-IJEPA (Littwin et al., 2024b) shows that providing encoders with explicit spatial information about context and target positions alleviates collapse across wider hyperparameter ranges. Balestriero et al. (2025) further reveal that trained JEPAs implicitly estimate data density, connecting representation learning to probabilistic modeling. Recently, SALT (Li et al., 2026) investigated the EMA framework and propose to replace it by using a frozen encoder pretrained via pixel reconstruction, then training a student to predict the teacher's masked latent representations, yielding a simpler and more scalable JEPA training scheme.

A common thread runs through this work: emphasis on collapse prevention at convergence rather than studying how collapse emerges and resolves during optimization. Tian et al. (2021) characterize eigenspace alignment, Jing et al. (2022) identify dimensional collapse, and He et al. (2024) propose weight-level regularization. These analyses provide steady-state properties but leave open the transient behavior of early training. For world model applications, where pretraining is expensive and early representations may be deployed for adaptation or planning, understanding these dynamics is critical. Our work addresses this gap by studying the collapse phase as a dynamical phenomenon, analyzing how EMA decay and masking ratio jointly govern early-training collapse.

3 PRELIMINARIES

We briefly provide elements regarding the JEPA framework, focusing on components central to our analysis. While we present notation using images as the running example; the formulation extends naturally to other modalities by replacing spatial masking with the appropriate structure.

Let $\mathcal{X} = \{x_1, \ldots, x_N\}$ denote a dataset of N samples. The goal of representation learning is to learn a mapping $f : \mathcal{X} \to \mathcal{H}$ from inputs to a representation space $\mathcal{H} \subseteq \mathbb{R}^d$, where d is the representation dimension. In self-supervised learning, this mapping is learned without access to labels; once trained, the representations can be evaluated on downstream tasks such as classification. A JEPA approaches this problem by partitioning each input into observed and masked regions, then training an encoder to predict latent representations of masked content from observed context. Let $\mathcal{N}$ (resp. $\mathcal{M}$) denote the index sets of observed (resp. masked) patches. Three components define the architecture.

Encoder f_θ. The encoder maps observed patches to latent representations. Given unmasked patches $\{p_i\}_{i\in\mathcal{N}}$, it produces embeddings $z_\mathcal{N} = f_\theta(\{p_i\}_{i\in\mathcal{N}}) \in \mathbb{R}^{|\mathcal{N}|\times d}$, where d is the representation dimension. Depending on the application, the encoder may be instantiated as a multilayer perceptron, convolutional network, or Transformer.

EMA (Target) Encoder $g_{\bar\theta}$. The target encoder shares the main encoder's architecture but follows a distinct update rule. Rather than receiving gradients directly, its parameters $\bar\theta$ evolve as an exponential moving average of the main encoder's weights:

$$\bar\theta^{(t)} \leftarrow \beta\,\bar\theta^{(t-1)} + (1-\beta)\,\theta^{(t)}, \tag{1}$$

where $\beta \in [0, 1]$ is the EMA decay rate. Values close to 1 yield a slowly-moving target, providing stable supervision that has proven essential for self-supervised learning without negative samples.

Throughout this analysis, we assume that the encoder and EMA encoder are initialized identically:

$$\theta^{(0)} = \bar{\theta}^{(0)} =: \theta_0$$

Predictor P_ϕ. The predictor transforms context representations to align with the target encoder's outputs at masked positions. Given encoded context, it produces predictions $\hat{z}_\mathcal{M} = P_\phi(z_\mathcal{N}) \in \mathbb{R}^{|\mathcal{M}| \times d}$ for the masked patches.

Training Objective. Training minimizes the discrepancy between predictions and targets in latent space:

$$\mathcal{L} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \|\hat{z}_i - \bar{z}_i\|^2, \tag{2}$$

where $\bar{z}_i = g_{\bar{\theta}}(p_i)$ denotes the target representation of masked patch p_i. This objective encourages the encoder and predictor to learn structure sufficient for reconstructing masked content without operating in pixel space.

Throughout, we use $\|\cdot\|$ to denote the ℓ_2 norm for vectors and the spectral norm for matrices.

4 On the Collapse Phase of JEPA

As in Figure 1, JEPA models trained with an EMA-based target encoder often exhibit a collapse phase during training. This phase typically occurs early, within the first epochs (often between 10 and 100 epochs). Empirically, it is characterized by a rapid decrease in the reconstruction loss, followed by an increase or stagnation effect, while downstream performance remains low (or degrades). This mismatch between reconstruction loss and downstream performance suggests that the learned representations are temporarily uninformative. In this section, we aim to provide a theoretical understanding and explanation for this phenomenon in the context of EMA-based JEPA training.

4.1 Characterizing Collapse

To analyze this behavior, we first need a precise way to characterize and measure collapse. Empirically, collapse manifests as a regime in which the reconstruction objective improves, yet the learned representations perform no better than random features on downstream tasks. We hypothesize that this behavior arises from the early-stage training dynamics of the encoder and the EMA encoder.

Specifically, during the initial epochs, the parameters of the encoder and the EMA encoder remain very close. As a result, their outputs are nearly identical. In this regime, the prediction task becomes trivial: the predictor can minimize the reconstruction loss by learning a near-identity or constant mapping. While this leads to a low reconstruction error, it produces representations that lack semantic structure and are therefore ineffective for downstream tasks. We can formalize this intuition with the following assumption.

> **Main Assumption.** If the output of the EMA encoder g is close to that of the encoder f, then the prediction task becomes easy for the predictor P. In this case, the predictor can converge to a trivial solution (e.g., an identity or constant mapping), leading to representation collapse.

Motivated by this assumption, we focus on tracking the discrepancy between the outputs of the encoder and the EMA encoder during training. Given two views x and $\hat{x}$ sampled from the underlying data distribution $\mathcal{D}_\mathcal{X}$, we define the following collapse quantity:

$$\mathcal{C}_{f,g} = \mathbb{E}_{x,\hat{x} \sim \mathcal{D}_\mathcal{X}} \left[d_\mathcal{Y}\big(f(x), g(\hat{x})\big) \right], \tag{3}$$

where $d_\mathcal{Y}$ denotes a distance metric in the representation space. In our analysis, we will be using the ℓ_2 norm as our distance, although the framework naturally extends to other distances depending on the application. We note that we validated this assumption empirically (Section 5.2) to better show that it's valid in practical setting and that it could be a way of quantifying the collapse phase.

While we focus on masking-based strategies, this formulation also applies to augmented-view settings. In this case, x and $\hat{x}$ may correspond to different crops, augmentations, or masked versions of the

same input. We assume that such pairs are close in the input space, and we formalize this by

$$d_{\mathcal{X}}(x, \hat{x}) = \|x - \hat{x}\| \leq \epsilon,$$

where $d_{\mathcal{X}}$ is a distance metric in the input space and $\epsilon > 0$ controls the similarity between the two views.

This neighborhood-based formulation does not explicitly encode the masking mechanism but instead captures it through a Lipschitz-style assumption. In practice, the value of ϵ depends on the masking strategy and the data distribution. We revisit this connection later when analyzing how masking influences collapse behavior.

The quantity $\mathcal{C}_{f,g}$ measures the discrepancy between the outputs of the encoder and the EMA encoder. Under the introduced assumption, small values of $\mathcal{C}_{f,g}$ during the early stages of training indicate that the prediction task becomes trivial, thereby increasing the risk of representation collapse. This motivates the following formal definition of the collapse phase of JEPA when considering EMA regularization.

Definition 4.1 (Representation Collapse)**.** We say that the encoder–EMA pair is in a collapse phase at epoch t with risk level $\gamma^{(t)}$ if

$$\mathcal{C}_{f,g}^{(t)}(\epsilon) \leq \gamma^{(t)}.$$

Directly computing $\mathcal{C}_{f,g}$ is generally intractable without strong assumptions. Therefore, in practice, we rely on upper bounding this quantity. A smaller bound $\gamma^{(t)}$ indicates a smaller discrepancy between the encoder and EMA encoder outputs, and thus a higher likelihood of collapse at epoch t.

4.2 Connecting Collapse to Hyperparameters

Building on the collapse characterization introduced in the previous section, we now derive an upper bound on the collapse risk γ and analyze its dependence on the model hyperparameters. While Definition 4.1 is general and applies to a wide range of encoders, we focus on a simplified setting that enables a tractable theoretical analysis.

Problem Setup. For simplicity, we consider an encoder instantiated as a single-layer MLP with a 1-Lipschitz activation function. This assumption holds for commonly used nonlinearities such as ReLU and TanH (Virmaux & Scaman, 2018). Although simplified, this setting captures the essential behavior of JEPA training and the insights naturally extend and can easily be extended to more expressive architectures, such as Vision Transformers.

Following this setup, we analyze the collapse quantity. For two views x and $\hat{x}$, we write the introduced formulation as:

$$\mathcal{C} = \|f(x) - g(\hat{x})\| \leq \underbrace{\|f(x) - f(\hat{x})\|}_{\text{input mismatch}} + \underbrace{\|f(\hat{x}) - g(\hat{x})\|}_{\text{EMA-effect mismatch}} \tag{4}$$

$$\leq L_f^{(t)} \epsilon + \|f(\hat{x}) - g(\hat{x})\|, \tag{5}$$

where $L_f^{(t)}$ denotes an upper bound on the Lipschitz constant of the encoder f at epoch t, and ϵ controls the similarity between the two input views. The first term captures the effect of input perturbations, while the second term reflects the discrepancy between the encoder and the EMA encoder induced by their parameter dynamics (which is the main cause of the collapse). Upper bounds on $L_f^{(t)}$ are well known for MLPs and Transformers, and can be directly incorporated into the analysis. By further analyzing the parameter mismatch term, which depends on the EMA update rule, we derive an explicit upper bound on the collapse risk. The detailed derivation is provided in the appendix, and leads to the following result characterizing the collapse phase.

Theorem 4.2. *Let f be an encoder following the considered problem setup, then the encoder–EMA pair is in a γ–collapse phase, with*

$$\gamma = L_f^{(t)} \epsilon + \left\| \beta \Delta^{(t)} - (1 - \beta) \sum_{\ell=1}^{t-1} \beta^{t-\ell} \Delta^{(\ell)} \right\| \|\hat{x}\|,$$

where $\Delta^{(t)}$ denotes the encoder weight update (the difference to original weights) at epoch t, and $\beta \in [0, 1]$ is the EMA decay parameter.

The upper bound γ in Theorem 4.2 reveals a direct connection between representation collapse and the EMA update dynamics. In particular, collapse is governed by the discrepancy between the current encoder update and the exponentially weighted average of past updates. The bound depends on the difference between two terms: the current deviation scaled by β, namely $\beta\Delta^{(t)}$, and the historical EMA contribution $(1-\beta)\sum_{\ell=1}^{t-1}\beta^{t-\ell}\Delta^{(\ell)}$.

This characterization provides a clear interpretation of the collapse behavior. **(a)** Early in training, when t is small, the updates $\Delta^{(\ell)}$ are typically small, making both terms small and their difference negligible. In this regime, the encoder and EMA encoder remain close, corresponding to the collapse phase. **(b)** Escape from collapse occurs when the current update $\Delta^{(t)}$ grows sufficiently faster than its EMA history, inducing a larger discrepancy between the encoder and EMA encoder outputs.

Corollary 4.3 (Effect of EMA Parameter)**.** *Under the results and conditions of Theorem 4.2, increasing the EMA decay parameter β prolongs the collapse phase by reducing the discrepancy between the encoder and the EMA encoder during early training.*

Corollary 4.3 follows directly from the structure of the derived upper-bound in Theorem 4.2. Specifically, when β is close to one, the EMA encoder closely tracks the encoder parameters, causing the weighted historical quantity term $(1-\beta)\sum_{\ell=1}^{t-1}\beta^{t-\ell}\Delta^{(\ell)}$ to evolve slowly. As a result, the difference between the current update $\beta\Delta^{(t)}$ and its EMA history remains small during the early epochs. This keeps the bound γ small, increasing the likelihood of collapse during the first training phase. The resulting insights and analysis actually suggests that the collapse phase is not an inherent failure of JEPA training, but rather a transient regime induced by conservative EMA updates.

4.3 Connecting Collapse to Masking

In the previous sections, we established a connection between representation collapse and the EMA dynamics, highlighting the role of the decay parameter β. We now investigate how the masking strategy used during JEPA pre-training influences the collapse behavior. In particular, we study how masking affects the input mismatch term introduced in our earlier decomposition of the collapse quantity.

Without loss of generality, we focus on the I-JEPA setting. Let $\hat{x} \in \mathbb{R}^{N\times d}$ denote the full sequence of token embeddings extracted from an input image, where each row $\hat{x}_i \in \mathbb{R}^d$ corresponds to a token. We assume that token embeddings are uniformly bounded, i.e., $\|\hat{x}_i\|_2 \leq B \quad$ for all $i \in [0, N]$, which is realistic due to input normalization and bounded positional encodings.

In I-JEPA, masking is applied by selecting contiguous blocks of tokens, with an expected masking ratio $p \in [0,1]$. Let $M \in \{0,1\}^{N\times N}$ denote the diagonal masking operator, where masked tokens are retained and visible tokens are removed. Under this formulation, the masked input can be written as $x = M\hat{x}$, while $\hat{x}$ represents the unmasked input. This induces an input-space discrepancy that depends explicitly on the masking ratio p. According to this formulation, we can further refine the collapse bound provided previously by incorporating the effect of masking.

Proposition 4.4. *Under the assumptions of Theorem 4.2, , for any input $\hat{x}$ and masking with expected coverage p, the encoder–EMA pair is in a γ-collapse phase at epoch t, with*

$$\gamma = L_f^{(t)}\, B\sqrt{pN} + \left\| \beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1}\beta^{t-\ell}\Delta^{(\ell)} \right\| \|\hat{x}\|.$$

This bound highlights two distinct control of the masking on the collapse mechanism. The first term arises from the input mismatch induced by masking and scales with the masking ratio p, the number of tokens N, and the Lipschitz constant $L_f^{(t)}$. The second term corresponds to the parameter mismatch between the encoder and the EMA encoder, as characterized previously. Specifically, the bound implies:

- Increasing the masking ratio p amplifies the input mismatch term, which can delay escape from the collapse phase by increasing the overall collapse bound.
- Contiguous block masking, as used in I-JEPA, introduces spatial correlations between masked tokens, increasing the variance of the input mismatch term and leading to variability in collapse dynamics across runs.

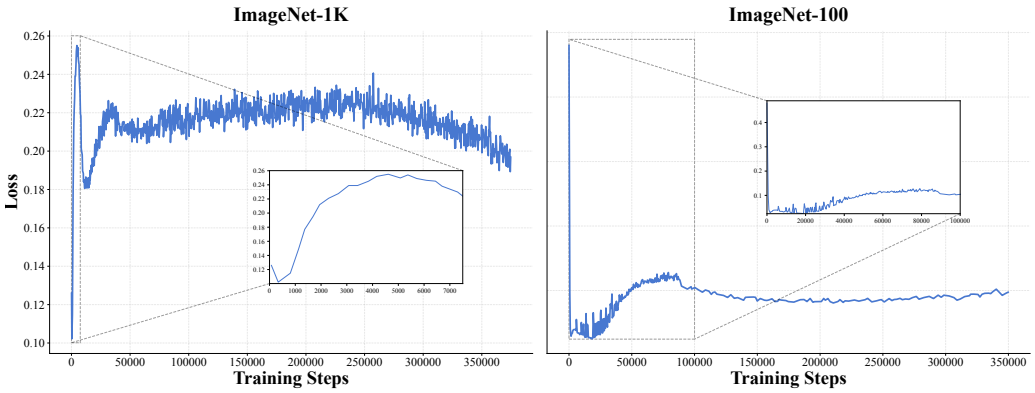

Figure 2: **JEPA prediction loss on ImageNet-1K and ImageNet-100.** Both datasets exhibit the characteristic collapse-and-recovery pattern observed on CIFAR-10. Insets highlight the early collapse phase where loss drops rapidly before stabilizing. The collapse phase is not an artifact of small datasets; it persists at ImageNet scale, confirming this is a general feature of EMA-based update.

Overall, the relative influence of masking and EMA dynamics depends on the interplay between the masking ratio p, the EMA decay parameter β, and the encoder smoothness $L_f^{(t)}$. This suggests that masking strategy and EMA design act as complementary controls over collapse behavior, with potentially different effects during early and late training.

Remark. In practice, Transformer-based architectures such as I-JEPA and TS-JEPA operate only on visible tokens by extracting them into a shorter sequence, rather than explicitly forming the zero-padded representation $(I - M)\hat{x}$ used in our analysis. However, masking operators are non-expansive, i.e. we have the following:

$$\|(I - M)v\| \leq \|v\| \quad \text{for any } v,$$

which ensures that our bound on $\|M\hat{x}\|$ provides a valid upper bound on the discrepancy between masked and full representations, even if it is not tight in this setting.

5 Experimental Validation

We now aim to validate empirically the previously derived insights. We specifically show: (i) the validity of our introduced collapse assumption and metric, (ii) the validity of the our insights regarding the EMA parameters and their link to collapse.

Experimental Setting. For our experimental results, we used a ViT-Base, and the evaluation is done through linear probing, where the encoder is frozen and only a linear classification head is trained. This follows the typical evaluation that was done in previous works I-JEPA and TS-JEPA. Additional details about the hyper-parameters and implementation details are provided in Appendix C.

5.1 On the Existence of the Collapse

While in Figure 1, we showed the existence of the collapse existence in the case of CIFAR-10, similar insights are also observed in the case of larger datasets such as ImageNet-1K and ImageNet-100. Figure 2 provides the corresponding JEPA pre-training loss for these two datasets, where we see a similar loss profile of collapse as the one observed on the CIFAR-10 dataset.

Critically, the collapse phase extends to domains where JEPAs serve as explicit dynamics models. We observe identical behavior when applying TS-JEPA to sensor time-series from physical systems, such as FordA (automotive engine monitoring) and FaultDetectionA (electromechanical drive diagnostics). These datasets involve predicting future sensor states from partial temporal context, precisely the predictive modeling task central to world models.

Figure 3 illustrates this behavior. The left panel reports the evolution of the pre-training loss, while the right panel shows the corresponding downstream accuracy. Together, these results confirm the

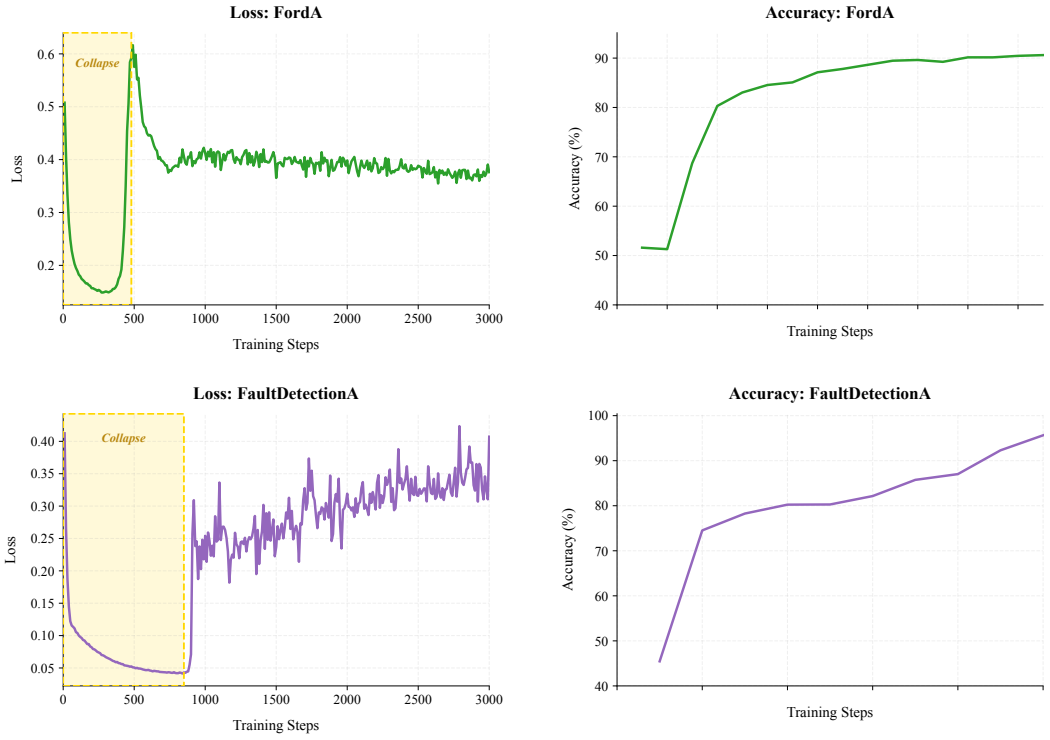

Figure 3: **The collapse phase generalizes beyond images to sequential tasks.** TS-JEPA trained on sensor data from physical systems (FordA and FaultDetectionA) exhibits the same pattern: prediction loss initially decreases (yellow) while downstream accuracy remains poor, followed by recovery.

presence of a collapse phase in time-series JEPA models, indicating that the phenomenon is not modality-specific but rather a general feature of EMA-based JEPA training.

5.2 On the Collapse Hypothesis

As discussed in Section 4.1, our analysis assumes that representation collapse occurs when the encoder and EMA encoder outputs become overly similar, making the prediction task trivial and allowing the predictor to minimize the pre-training loss without learning meaningful representations. We start by empirically validating this central assumption.

Figure 4 shows the evolution of the JEPA pre-training loss, downstream accuracy, and the ℓ_2 distance between encoder and EMA outputs. Early in training, the pre-training loss decreases while downstream performance degrades, indicating a collapsed regime. During this phase, the encoder–EMA discrepancy remains small. As training progresses, the loss increases, downstream accuracy improves, and the encoder-EMA difference grows, signaling the emergence of informative representations. These results support our hypothesis that a small discrepancy is a key indicator of representation collapse in JEPA.

5.3 On the Effect of Hyperparameters

Building on the empirical validation of the collapse hypothesis, we now study the practical implications of our analysis by focusing on the EMA update parameter β. As indicated by Corollary 4.3, β directly controls the similarity between the encoder and the EMA encoder and thus governs the onset and duration of the collapse phase. We empirically evaluate this effect by varying β and analyzing the resulting training dynamics, with particular attention to collapse onset, duration, and recovery. Figure 5 reports the JEPA pre-training loss and downstream accuracy. Different values of β induce collapse at different training stages and lead to substantial differences in downstream performance,

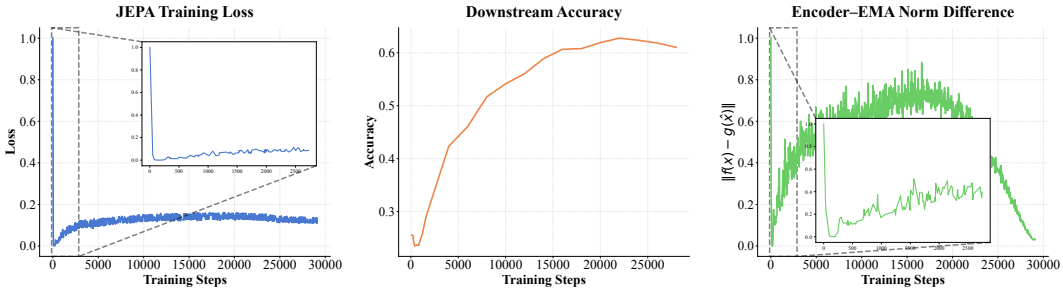

Figure 4: **Validating the collapse hypothesis.** Left: prediction loss drops early while downstream accuracy (middle) degrades. Right: the encoder-EMA output discrepancy remains small during collapse, then grows as useful representations emerge.

highlighting the strong sensitivity of JEPA training to the EMA update rate. These results show that an appropriate choice of β can mitigate collapse and improve representation quality.

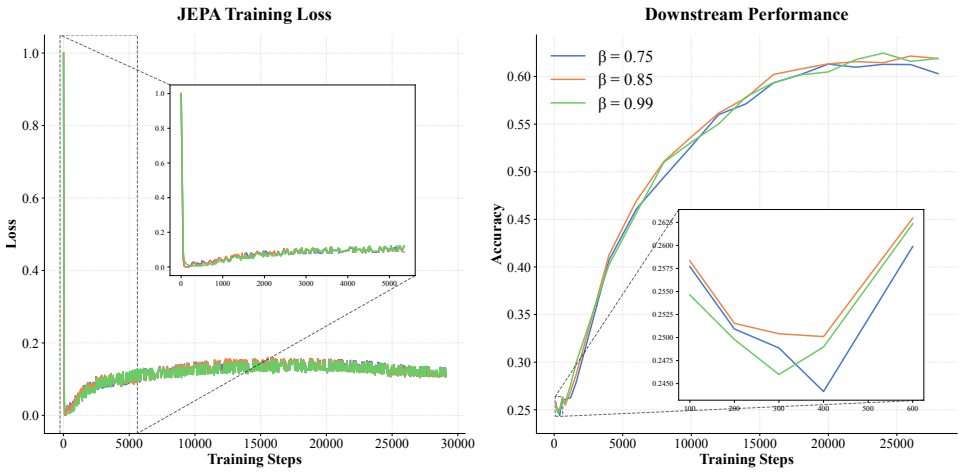

Figure 5: **Effect of EMA decay on collapse dynamics.** Different values of β induce collapse at different training stages and lead to substantial differences in downstream performance. Insets show early training detail.

6 Conclusion

In this work we identified and characterized the *collapse phase* in JEPA training: a transient regime where prediction loss drops while representations remain uninformative for downstream tasks. Our theoretical analysis traces this to tight coupling between encoder and EMA target during early optimization, with bounds that disentangle the roles of momentum and masking ratio. This coupling makes prediction easy for the wrong reasons, letting the model succeed at its proxy task without learning transferable structures. Experiments on ImageNet and time-series data confirm these dynamics are general, rather than modality-specific. For practitioners training JEPA-style world models, our encoder-EMA discrepancy metric provides a diagnostic for monitoring collapse, and our analysis offers guidance on hyperparameter regimes that shorten this unproductive phase.

Several directions merit future investigation. Our framework surfaces masking ratio as a control lever, but we did not explore adaptive schedules. The collapse phenomenon likely generalizes to other momentum-based predictive architectures, and our framework provides a starting point for monitoring training health in large-scale world model pretraining.

Finally, the theory also points toward mitigation strategies: our bound suggests that increasing encoder-EMA divergence, e.g. via controlled perturbations, can break the tight early coupling and accelerate escape from collapse without altering the long-term training objective.

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.

Randall Balestriero and Yann LeCun. Lejepa: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025.

Randall Balestriero, Nicolas Ballas, Mike Rabbat, and Yann LeCun. Gaussian embeddings: How jepas secretly learn your data density. *arXiv preprint arXiv:2510.05949*, 2025.

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. ISSN 2835-8856. URL https://openreview.net/forum?id=QaCCuDfBk2. Featured Certification.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pp. 1597–1607. PmLR, 2020.

Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

Hoang Anh Dau, Anthony Bagnall, Kaveh Kamgar, Chin-Chia Michael Yeh, Yan Zhu, Shaghayegh Gharghabi, Chotirat Ann Ratanamahatana, and Eamonn Keogh. The ucr time series archive. *IEEE/CAA Journal of Automatica Sinica*, 6(6):1293–1305, 2019.

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. URL https://openreview.net/forum?id=YicbFdNTTy.

Sofiane Ennadir, Siavash Golkar, and Leopoldo Sarra. Joint embeddings go temporal. *arXiv preprint arXiv:2509.25449*, 2025.

Sofiane ENNADIR, Levente Zólyomi, Oleg Smirnov, Tianze Wang, John Pertoft, Filip Cornell, and Lele Cao. Pool me wisely: On the effect of pooling in transformer-based models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=8uhXfdSJmA.

Zhengcong Fei, Mingyuan Fan, and Junshi Huang. A-jepa: Joint-embedding predictive architecture can listen. *arXiv preprint arXiv:2311.15830*, 2023.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Junlin He, Jinxiao Du, and Wei Ma. Preventing dimensional collapse in self-supervised learning via orthogonality regularization. *Advances in Neural Information Processing Systems*, 37:95579–95606, 2024.

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. 9729–9738, 2020.

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

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Yilun Kuang, Yash Dagade, Deep Chakraborty, Erik Learned-Miller, Randall Balestriero, Tim GJ Rudner, and Yann LeCun. Radial-vcreg: More informative representation learning through radial gaussianization. In *NeurIPS 2025 Workshop on Symmetry and Geometry in Neural Representations*, 2025.

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

Christian Lessmeier, James Kuria Kimotho, Detmar Zimmer, and Walter Sextro. Condition monitoring of bearing damage in electromechanical drive systems by using motor current signals of electric motors: A benchmark data set for data-driven classification. In *PHM Society European Conference*, volume 3 1, 2016.

Xianhang Li, Chen Huang, Chun-Liang Li, Eran Malach, Joshua M. Susskind, Vimal Thilak, and Etai Littwin. Rethinking JEPA: Compute-efficient video self-supervised learning with frozen teachers. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=3cB9243E9i.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *Advances in Neural Information Processing Systems*, 37:91300–91336, 2024a.

Etai Littwin, Vimal Thilak, and Anand Gopalakrishnan. Enhancing JEPAs with spatial conditioning: Robust and efficient representation learning. In *NeurIPS 2024 Workshop: Self-Supervised Learning - Theory and Practice*, 2024b. URL https://openreview.net/forum?id=IAqwCSv7kI.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

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

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015.

Ayumu Saito, Prachi Kudeshia, and Jiju Poovvancheri. Point-jepa: A joint embedding predictive architecture for self-supervised learning on point cloud. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7348–7357. IEEE, 2025.

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. URL https://openreview.net/forum?id=gx3LMRB15C.

Yuandong Tian, Xinlei Chen, and Surya Ganguli. Understanding self-supervised learning dynamics without contrastive pairs. In *International Conference on Machine Learning*, pp. 10268–10278. PMLR, 2021.

Hugues Van Assel, Mark Ibrahim, Tommaso Biancalani, Aviv Regev, and Randall Balestriero. Joint embedding vs reconstruction: Provable benefits of latent space prediction for self supervised learning. *arXiv preprint arXiv:2505.12477*, 2025.

Aladin Virmaux and Kevin Scaman. Lipschitz regularity of deep neural networks: analysis and efficient estimation. *Advances in Neural Information Processing Systems*, 31, 2018.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International conference on machine learning*, pp. 12310–12320. PMLR, 2021.

A PROOF OF THEOREM 4.2

Theorem. *Let f be an encoder following the considered problem setup, then the couple Encoder/EMA are in γ–collapse phase, with:*

$$\gamma = \left\| \beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\Delta^{(\ell)} \right\| \|\hat{x}\|,$$

where $\Delta^{(t)}$ denotes the encoder weight update (the difference to original weights) at epoch t, and $\beta \in [0,1]$ is the EMA decay parameter.

Proof. We start with the simpler case where the encoder is an MLP consisting of one layer followed by an activation function σ that is 1-Lipschitz.

For $f(\hat{x}) = \sigma(\theta^{(t)}\hat{x})$ and $g(\hat{x}) = \sigma(\bar{\theta}^{(t)}\hat{x})$, using the Lipschitz property of σ:

$$\|f(\hat{x}) - g(\hat{x})\| = \|\sigma(\theta^{(t)}\hat{x}) - \sigma(\bar{\theta}^{(t)}\hat{x})\| \leq \|\theta^{(t)}\hat{x} - \bar{\theta}^{(t)}\hat{x}\| \tag{6}$$

Applying the submultiplicative property of the operator norm:

$$\|f(\hat{x}) - g(\hat{x})\| \leq \|\theta^{(t)} - \bar{\theta}^{(t)}\| \|\hat{x}\| \tag{7}$$

EMA unrolling: Consider the EMA update function after epoch t with parameter β. The update can be formulated as:

$$\bar{\theta}^{(t)} = \beta\bar{\theta}^{(t-1)} + (1-\beta)\theta^{(t)}. \tag{8}$$

Recursively unrolling the previous equation gives the following closed form:

$$\bar{\theta}^{(t)} = \beta^t\bar{\theta}^{(0)} + (1-\beta)\sum_{\ell=1}^{t} \beta^{t-\ell}\theta^{(\ell)}. \tag{9}$$

To compute $\theta^{(t)} - \bar{\theta}^{(t)}$, we first separate the $\ell = t$ term from the sum (noting that $\beta^{t-t} = 1$):

$$\begin{aligned}
\theta^{(t)} - \bar{\theta}^{(t)} &= \theta^{(t)} - \beta^t\bar{\theta}^{(0)} - (1-\beta)\sum_{\ell=1}^{t} \beta^{t-\ell}\theta^{(\ell)} \\
&= \theta^{(t)} - \beta^t\bar{\theta}^{(0)} - (1-\beta)\theta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\theta^{(\ell)} \\
&= \beta\theta^{(t)} - \beta^t\bar{\theta}^{(0)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\theta^{(\ell)}
\end{aligned} \tag{10}$$

By assumption $\theta^{(0)} = \bar{\theta}^{(0)} = W^{(0)}$, we define the weight deviation at epoch ℓ as:

$$\Delta^{(\ell)} := \theta^{(\ell)} - W^{(0)}$$

Substituting $\theta^{(\ell)} = W^{(0)} + \Delta^{(\ell)}$ into Equation equation 10:

$$\begin{aligned}
\theta^{(t)} - \bar{\theta}^{(t)} &= \beta(W^{(0)} + \Delta^{(t)}) - \beta^t W^{(0)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}(W^{(0)} + \Delta^{(\ell)}) \\
&= \underbrace{W^{(0)}\left[\beta - \beta^t - (1-\beta)\sum_{\ell=1}^{t-1}\beta^{t-\ell}\right]}_{\text{initialization term}} + \underbrace{\beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1}\beta^{t-\ell}\Delta^{(\ell)}}_{\text{learning dynamics term}}
\end{aligned} \tag{11}$$

The initialization term vanishes: We now show that the coefficient of $W^{(0)}$ equals zero. By applying the change of variables $j = t - \ell$, we obtain:

$$\sum_{\ell=1}^{t-1} \beta^{t-\ell} = \sum_{j=1}^{t-1} \beta^j = \beta \cdot \frac{1-\beta^{t-1}}{1-\beta}$$

Therefore:

$$(1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell} = \beta(1-\beta^{t-1}) = \beta - \beta^t$$

Substituting back into the initialization term coefficient:

$$\beta - \beta^t - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell} = \beta - \beta^t - (\beta - \beta^t) = 0$$

Consequently, we can write the final bound as:

$$\theta^{(t)} - \bar{\theta}^{(t)} = \beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\Delta^{(\ell)} \tag{12}$$

Combining with Equation equation 7 and 4 concludes the proof:

$$\gamma = \left\| \beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\Delta^{(\ell)} \right\| \|\hat{x}\| \tag{13}$$

□

B Proof of Proposition 4.4

Proposition. *Under the assumptions of Theorem 4.2, , for any input $\hat{x}$ and masking with expected coverage p, the encoder–EMA pair is in a γ-collapse phase at epoch t, with*

$$\gamma = L_f^{(t)} B\sqrt{pN} + \left\| \beta\Delta^{(t)} - (1-\beta)\sum_{\ell=1}^{t-1} \beta^{t-\ell}\Delta^{(\ell)} \right\| \|\hat{x}\|.$$

Proof. Let $\hat{x} \in \mathbb{R}^{N\times d}$ denote the full input, where each row $\hat{x}_i \in \mathbb{R}^d$ represents a token embedding. We assume $\|\hat{x}_i\| \le B$ for all tokens $i \in [N]$, which holds due to input normalization and bounded positional encodings commonly used in practice.

Let $M = \text{diag}(m_1, \ldots, m_N)$ be a binary masking operator, where $m_i = 1$ for masked (removed) tokens and $m_i = 0$ for visible (retained) tokens. The visible input is then given by:

$$x = (I - M)\hat{x}$$

This operation retains the visible tokens in their original positions while zeroing the masked tokens.

Bounding the input mismatch: The distance between the visible and full inputs is:

$$\|x - \hat{x}\| = \|(I-M)\hat{x} - \hat{x}\| = \|M\hat{x}\| \tag{14}$$

For the masked portion $M\hat{x}$, we can derive the following bounds. First, note that:

$$\|M\hat{x}\|_F^2 = \sum_{i=1}^{N} m_i \|\hat{x}_i\|^2 \le |M| B^2 \tag{15}$$

where $|M| = \sum_{i=1}^{N} m_i$ is the number of masked tokens. Using the relationship between spectral and Frobenius norms, we obtain:

$$\|x - \hat{x}\| \le \|M\hat{x}\|_F \le B\sqrt{|M|} \tag{16}$$

This provides a deterministic upper bound on the input mismatch term in our collapse analysis.

Combining input and parameter mismatch: For a fixed input $\hat{x}$, we take expectation over the masking randomness only. By the triangle inequality:

$$\mathbb{E}_M[\|f(x) - g(\hat{x})\| \mid \hat{x}] \leq L_f^{(t)} \mathbb{E}_M[\|x - \hat{x}\|] + \|f(\hat{x}) - g(\hat{x})\|. \tag{17}$$

Note that the second term has no expectation since $\hat{x}$ is fixed.

For the *input mismatch term*, since $L_f^{(t)}$ is deterministic at epoch t, applying Jensen's inequality to Equation equation 16:

$$L_f^{(t)} \mathbb{E}_M[\|x - \hat{x}\|] \leq L_f^{(t)} B \sqrt{\mathbb{E}[|M|]} = L_f^{(t)} B \sqrt{pN}, \tag{18}$$

where $\mathbb{E}[|M|] = pN$ holds for both Bernoulli and block masking.

For the *EMA mismatch term*, since $\hat{x}$ is fixed, Theorem 4.2 gives deterministically:

$$\|f(\hat{x}) - g(\hat{x})\| \leq \left\| \beta \Delta^{(t)} - (1-\beta) \sum_{\ell=1}^{t-1} \beta^{t-\ell} \Delta^{(\ell)} \right\| \|\hat{x}\|. \tag{19}$$

Combining yields $\mathbb{E}_M[\|f(x) - g(\hat{x})\| \mid \hat{x}] \leq \gamma(\hat{x})$ with

$$\gamma(\hat{x}) = L_f^{(t)} B \sqrt{pN} + \left\| \beta \Delta^{(t)} - (1-\beta) \sum_{\ell=1}^{t-1} \beta^{t-\ell} \Delta^{(\ell)} \right\| \|\hat{x}\|. \tag{20}$$

□

Remark B.1 (Bernoulli masking). When each token is masked independently with probability p (as in TS-JEPA), we have $m_i \sim \text{Bernoulli}(p)$ independently. The expected number of masked tokens is:

$$\mathbb{E}[|M|] = pN, \qquad \text{Var}(|M|) = Np(1-p)$$

By Hoeffding's inequality, with probability at least $1 - \delta$:

$$|M| \leq pN + \sqrt{\frac{N}{2} \log \frac{2}{\delta}} \tag{21}$$

Therefore, with high probability:

$$\epsilon = \|x - \hat{x}\| \leq B \sqrt{pN + \mathcal{O}_p(\sqrt{N})}$$

Remark B.2 (Block masking). When masking uses contiguous blocks with expected coverage p (as in I-JEPA), the mean bound holds:

$$\mathbb{E}[\|x - \hat{x}\|] \leq B \sqrt{pN}$$

However, block masking introduces positive correlations among masked indices, selecting spatially or temporally correlated tokens. This creates higher variance in $\|M\hat{x}\|$ compared to independent Bernoulli masking, with values more frequently approaching the extremes of the deterministic bound $B\sqrt{|M|}$.

C Implementation Details

C.1 Datasets

For our empirical validation, we focused on two main modalities, namely Images and Time Series. For the first modality, in terms of pre-training, we used Imagenet-1K and ImageNet-10, which are two subset of the larger ImageNet dataset (Russakovsky et al., 2015). For the evaluation, we focused additionally on the CIFAR-10 and CIFAR-100 datasets (Krizhevsky et al., 2009). For these specific results, we used an average pooling during linear probing phase, other pooling could also be used and investigated (ENNADIR et al., 2025).

For the time series related tasks, we focused on the classification task where we consider the FordA, FordB (Dau et al., 2019), FaultDetectionA and FaultDetectionB (Lessmeier et al., 2016), all of which comprise outputs from various sensors. Additional statistics about these datasets are provided in Table 1.

Table 1: Statistics of the classification datasets used in our experiments.

DATASET	#TRAINING POINT	#TEST POINTS	#LENGTH	#CLASSES
FORDA	3601	1320	500	2
FORDB	3636	810	500	2
FAULTDETECTIONA	10912	2728	5120	3
FAULTDETECTIONB	10912	2728	5120	3

C.2 ARCHITECTURE AND PARAMETERS DETAILS

All computer vision related experiments were based on using a Transformer backbone (ViT-base (Dosovitskiy et al., 2021)). We used the same predictor and masking strategy as the one used in the original I-JEPA.

Optimization For our experiments, we followed the same setting as the original I-JEPA here as well. Specifically, all the experiments were optimized using AdamW (Loshchilov & Hutter, 2019), to produce both the encoder and predictor weights. For the ImageNet-100, we used a batch size of 512, while for the ImageNet-1K, we used a batch size of 2048, and the learning rate is linearly increased from 10^{-4} to 10^{-3} during the first 15 epochs of pretraining, and decayed to 10^{-6} following a cosine schedule thereafter. All experiments were conducted on NVIDIA A100 and H100 GPUs.

Time-Series Experiments. For this specific part, we have followed again a similar setup as the original TS-JEPA paper, where the encoder and decoder are a transformer with 2 attention heads and an embedding dimension of 128. For all the considered datasets and experiments, we segment each time series into 10 patches and employ a batch size of 32. For TS-JEPA, we apply a masking ratio of 70%. Similar to the image-based tasks, we trained all the models using the AdamW optimizer (Loshchilov & Hutter, 2019). We used a learning rate as the ones that were used in the original work, namely within the range 10^{-4}.