On the Information Bottleneck of VJEPA

Anonymous authors
Paper under double-blind review

Abstract

Joint Embedding Predictive Architectures (JEPAs) learn representations by predicting latent targets rather than reconstructing high-dimensional, pixel-level observations. Variational JEPA (VJEPA) extends this idea by replacing deterministic target regression with a probabilistic predictive model and a target-side KL regularizer. This paper provides a theoretical analysis of VJEPA through the lenses of *Information Bottleneck* (IB) and *Predictive Information Bottleneck* (PIB). We analyze two forms of VJEPA: the *context–target* form, naturally associated with IB, and the *temporal* form, naturally associated with PIB. We show that the current VJEPA objective is a *partial* bottleneck objective: its latent negative log-likelihood implements a Barber–Agakov lower bound on predictive mutual information, while its target KL regularizes the target-side latent distribution, rather than directly compressing the context or current state. We then derive two completed objectives: a *full IB-VJEPA*, which introduces a *stochastic context encoder* and a *KL-to-prior penalty* that upper-bounds the context information $\mathrm{I}(X_C; Z_C)$, and a *full PIB-VJEPA* as its *temporal* specialization, which introduces a *stochastic current-state encoder* and a *KL-to-prior penalty* that upper-bounds the state information $\mathrm{I}(X_{\leq t}; Z_t)$. The resulting analysis separates three information-theoretic roles that are conflated in standard JEPA-style objectives: *predictive information maximization*, *target-side regularization*, and *explicit context/state compression*. This closes the objective-level gap between probabilistic latent prediction and explicit information bottleneck control, providing a principled route to compression-controlled, uncertainty-aware VJEPA models.

1 Introduction

Joint Embedding Predictive Architectures (JEPAs) learn by predicting latent representations of missing or future observations, rather than reconstructing the observations themselves Assran et al. (2023); LeCun (2022). This design has two important consequences. First, prediction is carried out in a learned representation space instead of the raw observation space. As a result, JEPA avoids specifying likelihoods over high-dimensional, high-entropy observations such as pixels, video frames, or tokens, where a large amount of modeling capacity may otherwise be spent on high-frequency detail, background variation, or irreducible noise. Second, the target encoder can act as an implicit[1] representation bottleneck. By mapping the raw target observation to a latent target representation, the training objective focuses on predicting abstract, stable, and semantically meaningful structure, rather than every low-level nuisance factor present in the observation. In this sense, the target encoder has the potential to filter out unpredictable or task-irrelevant variation before the prediction loss is applied.

Variational JEPA (VJEPA) Huang (2026), being the first probabilistic formulation of JEPA, generalizes this deterministic framework by replacing point predictions with probabilistic predictive distributions. In this design, a context encoder maps an observation X_C to a deterministic latent Z_C, a target encoder defines an approximate target distribution $q_{\bar{\theta}}(Z_T \mid X_T)$, and a probabilistic predictor $p_\phi(Z_T \mid Z_C, \xi)$ is trained via a latent negative log-likelihood (NLL) alongside a target Kullback-Leibler (KL) regularizer[2]. While

[1] In standard JEPA, this "bottleneck" is architectural and implicit: it is induced by encoder capacity, masking, target-encoder design, exponential moving average dynamics, and strategic architectural choices, rather than by an explicit information-theoretic penalty.

[2] ξ represents side information which can be useful for e.g. planning and control.

this explicitly models uncertainty to support Bayesian filtering, casting the model in probabilistic terms also invites a rigorous analysis through the lens of the *Predictive Information Bottleneck* (PIB) Bialek et al. (2001); Still (2014). PIB dictates that *a temporal state should compress the past while preserving information about the future* Bialek et al. (2001).

Using the temporal VJEPA form Huang (2026) as an example, the current VJEPA objective can be written as

$$\begin{aligned}\mathcal{L}_{\text{temp-VJEPA}} =& \mathbb{E}_{(X_{\le t}, X_{t+\Delta}, \xi)\sim p_{\text{data}}} \mathbb{E}_{Z_{t+\Delta}\sim q_{\bar\theta}(\cdot\mid X_{t+\Delta})} \left[-\log p_\phi(Z_{t+\Delta} \mid f_\theta(X_{\le t}), \xi)\right] \\ &+ \beta\, \mathbb{E}_{X_{t+\Delta}\sim p_{\text{data}}} D_{\text{KL}}\left(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0(Z_{t+\Delta})\right).\end{aligned} \tag{1}$$

Equivalently, writing $Z_t = f_\theta(X_{\le t})$, the predictive term is the latent NLL: $-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$.

Under the PIB framework, this latent NLL provides a *variational upper bound* on the conditional entropy $H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$. Let $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$ denote the induced *population* conditional distribution over future target latents, determined by the training data distribution, the temporal context–future sampling process, and the target encoder. Then[3]

$$\begin{aligned}&\mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\big[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\big] \\ &= H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) + \mathbb{E}_{q^\star(Z_t,\xi)} D_{\text{KL}}\left(q^\star(Z_{t+\Delta} \mid Z_t, \xi) \| p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right).\end{aligned} \tag{2}$$

The first term is the conditional entropy of the induced population conditional, and the second term is the expected conditional KL divergence between this induced conditional and the parameterized predictor. Since the KL term is nonnegative, the expected latent NLL upper-bounds the conditional entropy:

$$H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) \le \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\big[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\big]. \tag{3}$$

Combining this with the conditional mutual-information identity $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) = H_{q^\star}(Z_{t+\Delta} \mid \xi) - H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$ yields the conditional Barber–Agakov lower bound Barber & Agakov (2003):

$$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) \ge H_{q^\star}(Z_{t+\Delta} \mid \xi) + \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})} \log p_\phi(Z_{t+\Delta} \mid Z_t, \xi). \tag{4}$$

Therefore, provided the marginal entropy of the future latent $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is prevented from collapsing[4], minimizing the latent NLL maximizes a Barber–Agakov variational lower bound on the future-latent predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$.

Thus, the current VJEPA design explicitly targets the predictive half of the PIB objective through a variational lower bound. However, it lacks a direct information-theoretic penalty on $\mathrm{I}(X_{\le t}; Z_t)$, the amount of information the current state stores about the history. As there is no formal compression term for the history-to-state map $X_{\le t} \mapsto Z_t$, state compression remains implicit; it arises from architectural capacity, masking, EMA dynamics, and target regularization rather than from a strict variational bottleneck.

This gap motivates a progression from analyzing the existing VJEPA objective to deriving explicit bottleneck-controlled IB/PIB variants. Accordingly, this work makes four contributions:

1. We provide a unified IB/PIB analysis of VJEPA by distinguishing its *context–target* form and its *temporal* form. In the context–target form, VJEPA is naturally related to IB through the variables (X_C, Z_C, Z_T); in the temporal form, it is naturally related to PIB through $(X_{\le t}, Z_t, Z_{t+\Delta})$.

[3] This is the standard conditional cross-entropy decomposition obtained by adding and subtracting $\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)$ inside the expected NLL:

$$\begin{aligned}&\mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right] \\ &= \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[-\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right] + \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[\log \frac{q^\star(Z_{t+\Delta} \mid Z_t, \xi)}{p_\phi(Z_{t+\Delta} \mid Z_t, \xi)}\right].\end{aligned}$$

The first term is the conditional entropy $H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$. For each fixed (Z_t, ξ), the inner expectation in the second term is the KL divergence between $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$ and $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$. Averaging this conditional KL over $q^\star(Z_t, \xi)$ gives $\mathbb{E}_{q^\star(Z_t,\xi)} D_{\text{KL}}\left(q^\star(Z_{t+\Delta} \mid Z_t, \xi) \| p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right)$.

[4] This stability can be encouraged in VJEPA through its exponential moving average (EMA) target encoder, target diversity, and target KL regularization.

2. We show that the current VJEPA objective is a *partial bottleneck objective*. Its latent NLL implements a Barber–Agakov variational lower bound on predictive mutual information, namely $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$ in the context–target case and $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ in the temporal case. However, it lacks an explicit information penalty on the context/state compression terms $\mathrm{I}(X_C; Z_C)$ and $\mathrm{I}(X_{\leq t}; Z_t)$.
3. We clarify the role of the target KL regularizer. By the KL-to-prior decomposition, the target KL upper-bounds target-side information, such as $\mathrm{I}(X_T; Z_T)$ or $\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta})$, and stabilizes the target latent space. It is therefore distinct from the IB/PIB compression terms on the context or current state.
4. We derive two completed objectives: *full IB-VJEPA*, which introduces a stochastic context encoder $q_\theta(Z_C \mid X_C)$ and a KL-to-prior penalty upper-bounding $\mathrm{I}(X_C; Z_C)$; and *full PIB-VJEPA*, its temporal specialization, which introduces a stochastic current-state encoder $q_\theta(Z_t \mid X_{\leq t})$ and a KL-to-prior penalty upper-bounding $\mathrm{I}(X_{\leq t}; Z_t)$.

This work provides **the first formal and systematic IB/PIB analysis of VJEPA**. It identifies the distinct roles of the objective components: which part realizes predictive information maximization, which part performs target-side regularization, and what must be added to obtain explicit IB/PIB bottleneck control. This analysis provides guidance for the future use of VJEPA and offers a template for analyzing other JEPA variants.

2 Preliminaries: IB and PIB

The generic IB variables (X, Z, Y) are used only to state the classical IB principle. In the VJEPA setting, we instantiate them in two ways. For context–target prediction, $X = X_C$, $Z = Z_C$, and the relevance variable is the target latent Z_T induced by X_T. For temporal PIB, $X = X_{\leq t}$, $Z = Z_t$, and the relevance variable is the future latent $Z_{t+\Delta}$. Thus, $X - Z - Y$ denotes the abstract IB structure, while $X_C - Z_C - Z_T - X_T$ and $X_{\leq t} - Z_t - Z_{t+\Delta} - X_{t+\Delta}$ denote the concrete VJEPA/PIB variables used in the model. A summary of notation is provided in Appendix A.

2.1 Information Bottleneck

Let X be an input, Y a relevant target, and Z a representation sampled from an encoder $q(z \mid x)$. The classical Information Bottleneck (IB) principle formulates representation learning as relevance-preserving compression: one seeks a stochastic representation Z that retains as little information as possible about X while preserving information useful for predicting Y Tishby et al. (1999). The resulting Lagrangian objective is

$$\min_{q(z|x)} \mathcal{L}_{\mathrm{IB}}(q) = \min_{q(z|x)} \left[\mathrm{I}(X; Z) - \beta \mathrm{I}(Z; Y)\right], \tag{5}$$

where $\beta > 0$ controls the relevance–compression trade-off. The term $\mathrm{I}(X; Z)$ penalizes *representation complexity*, measuring how much information about the input is retained in the representation, while $\mathrm{I}(Z; Y)$ rewards *relevance*, measuring how much information about the prediction target survives in Z. Equivalently, IB searches for a representation that is compressed with respect to X but sufficient, or approximately sufficient, for predicting Y Tishby et al. (1999); Shwartz-Ziv & Tishby (2017); Alemi et al. (2017).

2.2 Predictive Information Bottleneck

For temporal data, let $X_{\leq t} = X_{\leq t}$ denote the history and let $Z_t = Z_t$ denote the learned current state. PIB specializes equation 5 to future prediction:

$$\min_{q(Z_t|X_{\leq t})} \mathcal{L}_{\mathrm{PIB}}(q) = \min_{q(Z_t|X_{\leq t})} \left[\mathrm{I}(X_{\leq t}; Z_t) - \lambda \mathrm{I}(Z_t; Z_{t+\Delta})\right], \tag{6}$$

where $\lambda > 0$ controls the prediction–compression trade-off. In this work, $Z_{t+\Delta} = Z_{t+\Delta}$ denotes a future latent representation induced by the target encoder. The first term compresses, or summarizes, the past; the second term preserves information useful for predicting the future. When side information ξ is present, as in VJEPA, the predictive term is naturally replaced by the conditional quantity $\mathrm{I}(Z_t; Z_{t+\Delta} \mid \xi)$. Figure 1 summarizes the relationship between IB and PIB.

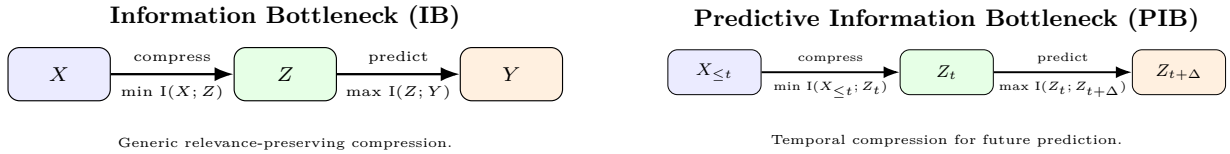

Figure 1: Comparison between the classical Information Bottleneck and the Predictive Information Bottleneck. IB compresses an input X into a representation Z while preserving information about a relevance variable Y. PIB specializes this principle to temporal prediction by compressing the history $X_{\leq t}$ into a state Z_t while preserving information about the future $Z_{t+\Delta}$.

2.3 A useful KL decomposition

We will repeatedly use the following *KL-to-prior decomposition*. Let X be an input random variable with density $p(x)$, and let Z be generated by a stochastic encoder $q(z \mid x)$. Define the *aggregated* posterior

$$q(z) = \int p(x)q(z \mid x)\, dx. \tag{7}$$

For any prior $p_0(z)$ with compatible support,

$$\mathbb{E}_{p(x)} D_{\mathrm{KL}}\big(q(z \mid x) \| p_0(z)\big) = \mathrm{I}(X; Z) + D_{\mathrm{KL}}\big(q(z) \| p_0(z)\big). \tag{8}$$

This identity shows that an expected KL-to-prior penalty controls two quantities: the information $\mathrm{I}(X; Z)$ retained by the representation and the mismatch between the aggregated posterior $q(z)$ and the chosen prior $p_0(z)$. Consequently, since KL divergence is nonnegative, we have

$$\mathrm{I}(X; Z) \leq \mathbb{E}_{p(x)} D_{\mathrm{KL}}\big(q(z \mid x) \| p_0(z)\big). \tag{9}$$

In full IB-VJEPA, this identity is applied with $X = X_C$, $Z = Z_C$, $q(z \mid x) = q_\theta(Z_C \mid X_C)$, and $p_0(z) = p_0^C(Z_C)$. In full PIB-VJEPA, it is applied with $X = X_{\leq t}$, $Z = Z_t$, $q(z \mid x) = q_\theta(Z_t \mid X_{\leq t})$, and $p_0(z) = p_0^S(Z_t)$. The full derivation and these specializations are given in Appendix B.7.

3 Current VJEPA design

VJEPA can be viewed in two closely related forms Huang (2026). The first is the *context–target* form, inherited from JEPA: a context input X_C is encoded into Z_C, and the model predicts a target latent Z_T associated with a target input X_T. This is the form used in this section to illustrate the architecture and VJEPA objective. The second is the *temporal* form, or VJEPA as a dynamical system, used for the PIB interpretation: the context is a history $X_{\leq t}$, the current latent state is Z_t, and the model predicts a future latent $Z_{t+\Delta}$. Thus, the temporal notation is obtained by the specialization $X_C = X_{\leq t}$, $Z_C = Z_t$, $X_T = X_{t+\Delta}$, and $Z_T = Z_{t+\Delta}$.

3.1 Context–Target VJEPA

The context–target form of VJEPA Huang (2026) keeps the JEPA context–target structure but replaces deterministic target regression with probabilistic prediction (Fig. 2). Given a context–target pair (X_C, X_T) and side information ξ, VJEPA uses

$$Z_C = f_\theta(X_C), \qquad q_{\bar{\theta}}(Z_T \mid X_T), \qquad p_\phi(Z_T \mid Z_C, \xi). \tag{10}$$

The target distribution $q_{\bar{\theta}}$ is typically parameterized by a target encoder whose shared parameters are updated by EMA[5], and the predictor p_ϕ is a deterministic neural network that outputs the parameters of a conditional distribution, e.g.

$$p_\phi(Z_T \mid Z_C, \xi) = \mathcal{N}\left(Z_T; \mu_\phi(Z_C, \xi), \Sigma_\phi(Z_C, \xi)\right). \tag{11}$$

[5]EMA stands for *exponential moving average* Mnih et al. (2015; 2016); Lillicrap et al. (2019); Grill et al. (2020). When the online/context encoder and target encoder have compatible parameterizations, the target encoder parameters $\bar{\theta}$ are not updated

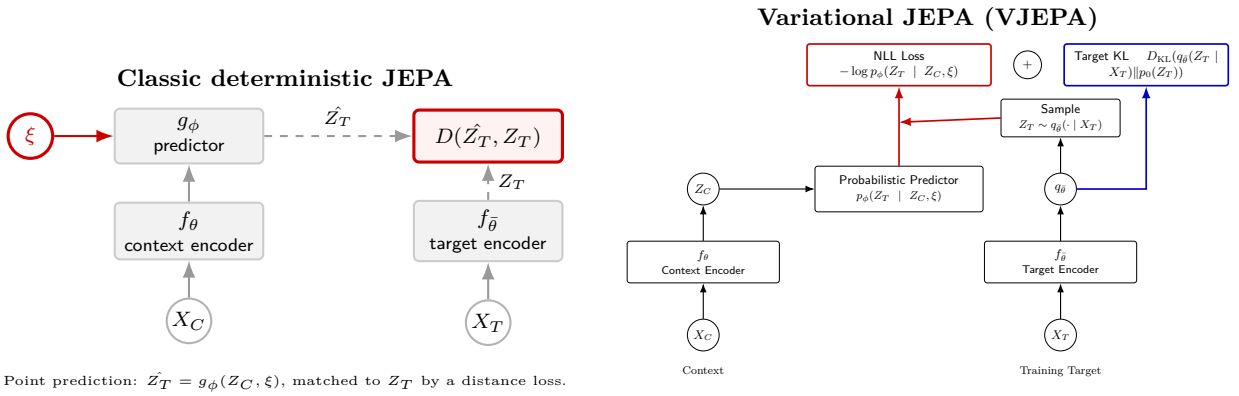

Figure 2: Comparison between classic deterministic JEPA and VJEPA. Classic JEPA predicts a point target embedding $\hat{Z}_T = g_\phi(Z_C, \xi)$ and matches it to Z_T through a distance loss. VJEPA retains the context–target structure but replaces point prediction with a conditional predictive distribution $p_\phi(Z_T \mid Z_C, \xi)$ and uses a stochastic target distribution $q_{\bar{\theta}}(Z_T \mid X_T)$ together with a target KL regularizer $D_{\mathrm{KL}}(q_{\bar{\theta}}(Z_T \mid X_T) \| p_0(Z_T))$. Note that VJEPA still commonly uses an EMA / stop-gradient-style target branch for stability, but its collapse-avoidance mechanism is not based solely on this asymmetry Huang (2026). In our view, JEPA and its variants are information-flow templates rather than fixed architectures: the specific encoder and predictor architecture choices are task- and modality-dependent.

This is a probabilistic neural predictor, but *not* a Bayesian neural network over the predictor weights ϕ.

The VJEPA objective[6] is

$$\begin{aligned}\mathcal{L}_{\mathrm{VJEPA}} =& \mathbb{E}_{(X_C, X_T, \xi) \sim p_{\mathrm{data}}} \mathbb{E}_{Z_T \sim q_{\bar{\theta}}(\cdot \mid X_T)} \left[-\log p_\phi(Z_T \mid f_\theta(X_C), \xi) \right] \\ &+ \beta \, \mathbb{E}_{X_T \sim p_{\mathrm{data}}} D_{\mathrm{KL}}\big(q_{\bar{\theta}}(Z_T \mid X_T) \| p_0(Z_T)\big).\end{aligned} \tag{12}$$

The first term is the latent *negative log-likelihood* (NLL). It trains the context encoder f_θ and the probabilistic predictor p_ϕ, since the predicted distribution is conditioned on the encoded context $Z_C = f_\theta(X_C)$. Under the usual EMA/stop-gradient target-branch design, gradients are not propagated through the target encoder in this term. The second term regularizes the target distribution, e.g. towards $p_0(Z_T) = \mathcal{N}(0, I)$. If $q_{\bar{\theta}}$ collapses to a Dirac delta and p_ϕ is a fixed-variance isotropic Gaussian centered at a deterministic predictor, then equation 12 reduces to squared-loss JEPA up to constants Huang (2026).

3.2 Temporal VJEPA

The context–target form above can be specialized to temporal prediction by identifying the context input with the history and the target input with a future observation: $X_C = X_{\le t}$ and $X_T = X_{t+\Delta}$. Correspondingly, the context latent becomes the current predictive state, $Z_C = Z_t$, and the target latent becomes the future latent, $Z_T = Z_{t+\Delta}$. Thus, the temporal VJEPA form is

$$Z_t = f_\theta(X_{\le t}), \qquad q_{\bar{\theta}}(Z_{t+\Delta} \mid X_{t+\Delta}), \qquad p_\phi(Z_{t+\Delta} \mid Z_t, \xi). \tag{13}$$

directly by backpropagation. Instead, they are updated as a slow-moving average of the online/context encoder parameters:

$$\bar{\theta} \leftarrow \tau \bar{\theta} + (1 - \tau)\theta, \qquad 0 < \tau < 1.$$

In the common case (e.g. I-JEPA Assran et al. (2023)), this is implemented by using the same encoder architecture for both branches. This makes the target branch evolve more slowly than the online branch, providing a more stable training target and helping avoid representational collapse.

[6]Here p_{data} denotes the empirical training distribution together with the context–target partitioning and side-information sampling process, such as masking patterns, temporal offsets, or other structural descriptors. It is a training-data/masking distribution, *not* a learned generative model of observations. In the existing VJEPA objective we keep the original notation $p_0(Z_T)$ for the target prior. In the IB/PIB analysis below, we write this target prior as $p_0^T(Z_T)$ whenever it is necessary to distinguish it from the context prior $p_0^C(Z_C)$ and the state prior $p_0^S(Z_t)$.

In this form, the history $X_{\leq t}$ plays the same architectural role as the context input X_C: it is encoded by the online/context encoder into a latent state. The future observation $X_{t+\Delta}$ plays the same role as the target input X_T: it is encoded by the target encoder into a distribution over future latents. The predictor then models a conditional distribution p_ϕ over future latents $Z_{t+\Delta}$ given the current latent state Z_t and side information ξ.

The temporal VJEPA objective corresponding to equation 12 is

$$\begin{aligned}\mathcal{L}_{\text{temp-VJEPA}} =& \mathbb{E}_{(X_{\leq t}, X_{t+\Delta}, \xi)\sim p_{\text{data}}} \mathbb{E}_{Z_{t+\Delta}\sim q_{\bar{\theta}}(\cdot|X_{t+\Delta})} \left[-\log p_\phi(Z_{t+\Delta} \mid f_\theta(X_{\leq t}), \xi)\right] \\ &+ \beta\, \mathbb{E}_{X_{t+\Delta}\sim p_{\text{data}}} D_{\text{KL}}\big(q_{\bar{\theta}}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0(Z_{t+\Delta})\big).\end{aligned} \tag{14}$$

Equivalently, since $Z_t = f_\theta(X_{\leq t})$, the predictive term can be written as an NLL of the form $-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$. This is the form used in the PIB analysis below: the NLL controls future-latent predictive information, while the current state Z_t remains deterministic unless an explicit stochastic state encoder is introduced. In the remainder of the paper, the context–target VJEPA form is used for the general IB/context-compression analysis, while the temporal VJEPA form is used for the PIB analysis, where $X_{\leq t}$ is compressed into Z_t to predict $Z_{t+\Delta}$.

4 IB analysis of context-target VJEPA

We first analyze the context–target form of VJEPA through the classical IB lens. In this setting, the IB variables are instantiated as follows: the input is the context X_C, the representation is the context latent Z_C, and the relevance variable is the target latent Z_T induced by the target input X_T. Thus, the relevant IB-like relation is $X_C \to Z_C \rightsquigarrow Z_T$, where Z_T is sampled from the target encoder distribution $q_{\bar{\theta}}(Z_T \mid X_T)$ and predicted from Z_C through $p_\phi(Z_T \mid Z_C, \xi)$. The side information ξ is treated as a conditioning variable.

4.1 Predictive part: latent NLL induces a context–target MI lower bound

The predictive term in the context–target VJEPA objective equation 12 is

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{(X_C, X_T, \xi)\sim p_{\text{data}}} \mathbb{E}_{Z_T\sim q_{\bar{\theta}}(\cdot|X_T)} \left[-\log p_\phi(Z_T \mid f_\theta(X_C), \xi)\right]. \tag{15}$$

Since $Z_C = f_\theta(X_C)$, the sampling process in equation 15 induces a joint distribution over (Z_C, ξ, Z_T) by sampling $(X_C, X_T, \xi) \sim p_{\text{data}}$, setting $Z_C = f_\theta(X_C)$, and drawing $Z_T \sim q_{\bar{\theta}}(\cdot \mid X_T)$. We denote this induced joint distribution by $q^\star(Z_C, \xi, Z_T)$. In this section, $q^\star$ refers to the context–target induced population distribution; in the later temporal PIB section, the same symbol will denote the analogous induced distribution over $(Z_t, \xi, Z_{t+\Delta})$. Its marginal over (Z_C, ξ) is denoted by $q^\star(Z_C, \xi)$, and its conditional distribution of Z_T given (Z_C, ξ) is denoted by $q^\star(Z_T \mid Z_C, \xi)$.

Therefore, equation 15 can equivalently be written as

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{q^\star(Z_C, \xi, Z_T)} \left[-\log p_\phi(Z_T \mid Z_C, \xi)\right] = \mathbb{E}_{q^\star(Z_C, \xi)} \mathbb{E}_{q^\star(Z_T|Z_C, \xi)} \left[-\log p_\phi(Z_T \mid Z_C, \xi)\right]. \tag{16}$$

Here $q^\star(Z_T \mid Z_C, \xi)$ is not a learned generative model; it is the conditional component of the induced population distribution $q^\star(Z_C, \xi, Z_T)$, and is the ideal target conditional approximated by $p_\phi(Z_T \mid Z_C, \xi)$.

By the conditional cross-entropy decomposition in Appendix B.6,

$$\begin{aligned}\mathcal{L}_{\text{pred}} &= H_{q^\star}(Z_T \mid Z_C, \xi) + \mathbb{E}_{q^\star(Z_C, \xi)} D_{\text{KL}}\left(q^\star(Z_T \mid Z_C, \xi) \| p_\phi(Z_T \mid Z_C, \xi)\right) \\ &\geq H_{q^\star}(Z_T \mid Z_C, \xi).\end{aligned} \tag{17}$$

The connection to context–target predictive mutual information follows from the conditional mutual-information identity $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi) = H_{q^\star}(Z_T \mid \xi) - H_{q^\star}(Z_T \mid Z_C, \xi)$. Together with equation 17, this gives

$$\mathcal{L}_{\text{pred}} \geq H_{q^\star}(Z_T \mid \xi) - \mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi).$$

Thus, when the target-latent marginal entropy $H_{q^\star}(Z_T \mid \xi)$ is controlled, minimizing $\mathcal{L}_{\text{pred}}$ corresponds to maximizing a variational lower bound on the context–target predictive information $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$.

Equivalently, the Barber–Agakov form of the same lower bound is

$$\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi) \geq H_{q^\star}(Z_T \mid \xi) + \mathbb{E}_{q^\star(Z_C,\xi,Z_T)}\left[\log p_\phi(Z_T \mid Z_C, \xi)\right]. \tag{18}$$

Hence, the context–target VJEPA NLL targets the *relevance* or *predictive mutual-information* term of an IB objective, with Z_T playing the role of the relevance variable.

4.2 Missing part: no explicit IB penalty on context compression

The classical IB objective in equation 5 contains an explicit information penalty $\mathrm{I}(X; Z)$. Under the context–target VJEPA instantiation $X = X_C$ and $Z = Z_C$, this becomes $\mathrm{I}(X_C; Z_C)$, which measures how much information about the context input is retained in the context latent representation. Thus, although the existing VJEPA learns a predictive context representation, the objective does *not* explicitly control how much information from X_C is stored in Z_C.

To complete the IB objective equation 5 in the context–target setting, one would replace the deterministic context encoder $Z_C = f_\theta(X_C)$ with a *stochastic context encoder* $q_\theta(Z_C \mid X_C)$[7]. This makes the context representation Z_C an explicit random variable induced by X_C, so that the IB compression term $\mathrm{I}(X_C; Z_C)$ becomes directly controllable. The corresponding tractable compression penalty is obtained by regularizing this encoder against a context prior $p_0^C(Z_C)$. This gives

$$\mathbb{E}_{X_C \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big). \tag{19}$$

By the KL-to-prior decomposition in Appendix B.7,

$$\mathbb{E}_{X_C \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big) = \mathrm{I}(X_C; Z_C) + D_{\mathrm{KL}}\big(q_\theta(Z_C) \| p_0^C(Z_C)\big), \tag{20}$$

where $q_\theta(Z_C) = \int p_{\text{data}}(X_C) q_\theta(Z_C \mid X_C)\, dX_C$ is the aggregated context posterior. Since KL divergence is nonnegative, equation 20 implies

$$\mathrm{I}(X_C; Z_C) \leq \mathbb{E}_{X_C \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big). \tag{21}$$

Thus, the stochastic encoder supplies the missing IB conditional representation distribution $q_\theta(Z_C \mid X_C)$, while the KL-to-prior regularizer supplies the explicit variational compression penalty.

The existing VJEPA objective equation 12 contains *no* such context-compression term. Therefore, context–target VJEPA should be interpreted as *predictive-IB-like* or as a *partial IB objective*: it targets the relevance term $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$ through the latent NLL, while compression of X_C into Z_C is imposed only indirectly by architectural capacity, masking, latent dimension, target-encoder design, EMA stabilization, and target-side regularization.

4.3 Target KL is not context compression

The target KL in equation 12 regularizes the target distribution $q_{\bar\theta}(Z_T \mid X_T)$ against the target prior $p_0^T(Z_T)$. Let $q_{\bar\theta}(Z_T) = \int p_{\text{data}}(X_T) q_{\bar\theta}(Z_T \mid X_T)\, dX_T$ denote the aggregated target posterior. As discussed above, the target KL can be interpreted as an upper bound on the target-side information $\mathrm{I}(X_T; Z_T)$, since

$$\mathbb{E}_{X_T \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_T \mid X_T) \| p_0^T(Z_T)\big) = \mathrm{I}(X_T; Z_T) + D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_T) \| p_0^T(Z_T)\big). \tag{22}$$

This target-side regularization can stabilize the target latent space and influence the marginal entropy term $H_{q^\star}(Z_T \mid \xi)$ in the context–target predictive MI identity $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi) = H_{q^\star}(Z_T \mid \xi) - H_{q^\star}(Z_T \mid Z_C, \xi)$. However, it is *not* the IB compression term $\mathrm{I}(X_C; Z_C)$, because it controls how much information the target latent Z_T stores about the target input X_T, not how much information the context latent Z_C stores about the context input X_C.

[7] Here θ denotes the trainable parameters of the stochastic context encoder, for example the neural-network weights producing the mean and covariance of $q_\theta(Z_C \mid X_C)$.

Proposition 1 (Context–target VJEPA is predictive-IB-like but not full IB)**.** *Assume the target-latent marginal entropy $H_{q^\star}(Z_T \mid \xi)$ is controlled. Then minimizing the context–target predictive NLL maximizes the Barber–Agakov lower bound induced by p_ϕ on the relevance term $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$. If p_ϕ is expressive, this lower bound can become tight. However, unless an explicit stochastic context encoder $q_\theta(Z_C \mid X_C)$ and a context-compression penalty of the form $\mathbb{E}_{X_C \sim p_{\mathrm{data}}} D_{\mathrm{KL}}(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C))$ are added, the objective does not directly minimize the IB compression term $\mathrm{I}(X_C; Z_C)$.*

Proof. The first statement follows from the conditional cross-entropy decomposition in equation 17 and the Barber–Agakov lower bound in equation 18. These show that the predictive NLL controls $H_{q^\star}(Z_T \mid Z_C, \xi)$ and therefore maximizes a variational lower bound on $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$ when $H_{q^\star}(Z_T \mid \xi)$ is controlled. Expressivity of p_ϕ determines how tightly the variational predictor can approximate the induced conditional $q^\star(Z_T \mid Z_C, \xi)$.

The second statement follows by inspection of the existing VJEPA objective equation 12. Its KL regularizer is applied to the target-latent distribution $q_{\bar{\theta}}(Z_T \mid X_T)$, not to a stochastic context encoder $q_\theta(Z_C \mid X_C)$. Therefore, equation 12 contains no context-compression term of the form $\mathbb{E}_{X_C \sim p_{\mathrm{data}}} D_{\mathrm{KL}}(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C))$, which would upper-bound $\mathrm{I}(X_C; Z_C)$. Hence the full IB compression term is absent from existing context–target VJEPA. □

5 PIB analysis of temporal VJEPA

We now specialize the preceding analysis to the temporal form of VJEPA and show that it is a partial predictive information bottleneck.

5.1 Predictive part: latent NLL induces a temporal MI lower bound

For temporal VJEPA, write $Z_t = f_\theta(X_{\le t})$ and $Z_{t+\Delta} \sim q_{\bar{\theta}}(\cdot \mid X_{t+\Delta})$. The predictive term in equation 14 is

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}_{(X_{\le t}, X_{t+\Delta}, \xi) \sim p_{\mathrm{data}}} \mathbb{E}_{Z_{t+\Delta} \sim q_{\bar{\theta}}(\cdot \mid X_{t+\Delta})} \left[-\log p_\phi(Z_{t+\Delta} \mid f_\theta(X_{\le t}), \xi) \right]. \tag{23}$$

The sampling process in equation 23 induces a joint distribution over $(Z_t, \xi, Z_{t+\Delta})$ by sampling

$$(X_{\le t}, X_{t+\Delta}, \xi) \sim p_{\mathrm{data}}, \qquad Z_t = f_\theta(X_{\le t}), \qquad Z_{t+\Delta} \sim q_{\bar{\theta}}(\cdot \mid X_{t+\Delta}).$$

We denote this induced joint distribution by $q^\star(Z_t, \xi, Z_{t+\Delta})$. Its marginal over (Z_t, ξ) is denoted by $q^\star(Z_t, \xi)$, and its conditional distribution of $Z_{t+\Delta}$ given (Z_t, ξ) is denoted by $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$. Thus, all uses of $q^\star$ below refer to the same induced *population* distribution, either through its joint, marginal, or conditional form.

Therefore, equation 23 can equivalently be written as

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}_{q^\star(Z_t, \xi, Z_{t+\Delta})} \left[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi) \right] = \mathbb{E}_{q^\star(Z_t, \xi)} \mathbb{E}_{q^\star(Z_{t+\Delta} \mid Z_t, \xi)} \left[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi) \right]. \tag{24}$$

Here $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$ is not a learned generative model; it is the conditional component of the induced population distribution $q^\star(Z_t, \xi, Z_{t+\Delta})$, and is the ideal target conditional approximated by $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$.

By the conditional cross-entropy decomposition[8],

$$\begin{aligned}\mathcal{L}_{\text{pred}} &= H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) + \mathbb{E}_{q^\star(Z_t,\xi)} D_{\text{KL}}\left(q^\star(Z_{t+\Delta} \mid Z_t, \xi) \| p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right) \\ &\geq H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi).\end{aligned} \tag{25}$$

The connection to predictive mutual information follows from the *conditional mutual-information* identity:

$$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) = H_{q^\star}(Z_{t+\Delta} \mid \xi) - H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi).$$

Together with equation 25, this gives

$$\mathcal{L}_{\text{pred}} \geq H_{q^\star}(Z_{t+\Delta} \mid \xi) - \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi).$$

Thus, when the marginal future-latent entropy $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled, minimizing $\mathcal{L}_{\text{pred}}$ corresponds to maximizing a variational lower bound on the predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$.

This can also be derived from an equivalent Barber–Agakov variational bound[9]:

$$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) \geq H_{q^\star}(Z_{t+\Delta} \mid \xi) + \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. \tag{26}$$

Hence, the current VJEPA NLL targets the *predictive mutual-information term* of the PIB objective in Eq. equation 6.

5.2 Missing part: no explicit PIB penalty on state compression

The full PIB objective in equation 6 contains an explicit information penalty $\mathrm{I}(X_{\leq t}; Z_t)$, which measures how much information about the history is retained in the current latent state. The existing temporal VJEPA does compress $X_{\leq t}$ into Z_t through the deterministic encoder $Z_t = f_\theta(X_{\leq t})$ and learns a predictive state, but this compression is not explicitly penalized in the objective, i.e. the objective does *not* explicitly control how much information from $X_{\leq t}$ is stored in Z_t.

To obtain an explicit PIB-style compression penalty, one would introduce a *stochastic current-state encoder* $q_\theta(Z_t \mid X_{\leq t})$ and regularize it against a state prior:

$$\mathbb{E}_{p(X_{\leq t})} D_{\text{KL}}\big(q_\theta(Z_t \mid X_{\leq t}) \| p_0^S(Z_t)\big). \tag{27}$$

Here $p_0^S(Z_t)$ denotes a prior over the current latent state. This term is needed because it provides a variational upper bound on the missing PIB compression term. Specifically, by the PIB-VJEPA *KL-to-prior decomposition* proved in Appendix C.2,

$$\mathbb{E}_{p(X_{\leq t})} D_{\text{KL}}\big(q_\theta(Z_t \mid X_{\leq t}) \| p_0^S(Z_t)\big) = \mathrm{I}(X_{\leq t}; Z_t) + D_{\text{KL}}\big(q_\theta(Z_t) \| p_0^S(Z_t)\big). \tag{28}$$

[8]The decomposition follows by adding and subtracting $\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)$ inside the expected NLL in equation 24:

$$\begin{aligned}&\mathbb{E}_{q^\star(Z_t,\xi)} \mathbb{E}_{q^\star(Z_{t+\Delta} \mid Z_t,\xi)}\left[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right] \\ &= \mathbb{E}_{q^\star(Z_t,\xi)} \mathbb{E}_{q^\star(Z_{t+\Delta} \mid Z_t,\xi)}\left[-\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right] \\ &\quad + \mathbb{E}_{q^\star(Z_t,\xi)} \mathbb{E}_{q^\star(Z_{t+\Delta} \mid Z_t,\xi)}\left[\log \frac{q^\star(Z_{t+\Delta} \mid Z_t, \xi)}{p_\phi(Z_{t+\Delta} \mid Z_t, \xi)}\right].\end{aligned}$$

The first double expectation is the conditional entropy

$$H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) = \mathbb{E}_{q^\star(Z_t,\xi)}\left[-\mathbb{E}_{q^\star(Z_{t+\Delta} \mid Z_t,\xi)} \log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right].$$

For each fixed (Z_t, ξ), the inner expectation in the second term is

$$D_{\text{KL}}\left(q^\star(Z_{t+\Delta} \mid Z_t, \xi) \| p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right).$$

Averaging this conditional KL over $q^\star(Z_t, \xi)$ gives the expected conditional KL term. The general cross-entropy identity is given in Appendix B.6.

[9]The Barber–Agakov bound gives a variational lower bound on mutual information Barber & Agakov (2003); in this setting, all expectations and information quantities are taken under the induced population distribution $q^\star(Z_t, \xi, Z_{t+\Delta})$. See Appendix C.3.

Since KL divergence is nonnegative, this implies

$$\mathrm{I}(X_{\le t}; Z_t) \le \mathbb{E}_{p(X_{\le t})} D_{\mathrm{KL}}\big(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t)\big). \tag{29}$$

The temporal VJEPA objective equation 14, equivalently the current VJEPA objective equation 12 under the temporal specialization, contains *no* such state-compression term. Therefore, temporal VJEPA should be interpreted as *predictive-PIB-like* or as a *partial PIB objective*: it targets future-predictive latent information through the NLL, while compression of $X_{\le t}$ into Z_t is imposed only indirectly by architectural bottlenecks, masking, latent dimension, target-encoder design, EMA stabilization, and target-side KL regularization.

Proposition 2 (Temporal VJEPA is predictive-PIB-like but not full PIB)**.** *Assume the future-latent marginal entropy $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled. Then minimizing the temporal predictive NLL maximizes the Barber–Agakov lower bound induced by p_ϕ on the predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$. If p_ϕ is expressive, this lower bound can become tight. However, unless an explicit stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$ and a state-compression penalty of the form $\mathbb{E}_{p(X_{\le t})} D_{\mathrm{KL}}(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t))$ are added, the objective does not directly minimize the PIB compression term $\mathrm{I}(X_{\le t}; Z_t)$.*

Proof. The first statement follows from the conditional cross-entropy decomposition in equation 25 and the Barber–Agakov lower bound in equation 26. These show that the predictive NLL controls $H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$ and therefore maximizes a variational lower bound on $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ when $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled. Expressivity of p_ϕ determines how tightly the variational predictor can approximate the induced conditional $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$.

The second statement follows by inspection of the temporal VJEPA objective equation 14, or equivalently equation 12 under the temporal specialization. Its KL regularizer is applied to the future target-latent distribution $q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$, not to a stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$. Therefore, equation 14 contains no state-compression term of the form $\mathbb{E}_{p(X_{\le t})} D_{\mathrm{KL}}\big(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t)\big)$, which, by Appendix C.2, would upper-bound $\mathrm{I}(X_{\le t}; Z_t)$. Hence the full PIB compression term is absent from temporal VJEPA. □

5.3 Target KL is not the same as PIB compression

The target KL in the temporal VJEPA objective equation 14 can itself be interpreted as an upper bound on the *future target-side* information $\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta})$. Define the aggregated future-target posterior $q_{\bar\theta}(Z_{t+\Delta}) = \int p_{\mathrm{data}}(X_{t+\Delta})\, q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})\, dX_{t+\Delta}$. Then, by the same *KL-to-prior decomposition* in Appendix B.7,

$$\mathbb{E}_{X_{t+\Delta} \sim p_{\mathrm{data}}} D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big) = \mathrm{I}(X_{t+\Delta}; Z_{t+\Delta}) + D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big). \tag{30}$$

Therefore,

$$\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta}) \le \mathbb{E}_{X_{t+\Delta} \sim p_{\mathrm{data}}} D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big). \tag{31}$$

This regularizes the future-target representation. It can prevent future latents from becoming uncontrolled or degenerate because the KL term keeps $q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$ close to the target prior $p_0^T(Z_{t+\Delta})$ and thereby shapes the aggregated future-latent distribution $q_{\bar\theta}(Z_{t+\Delta})$. This can affect the spread and stability of the future-latent marginal, and hence the marginal entropy term $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ in the predictive MI identity $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) = H_{q^\star}(Z_{t+\Delta} \mid \xi) - H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$. However, it is *not* the PIB compression term $\mathrm{I}(X_{\le t}; Z_t)$, because it controls how much information the future latent $Z_{t+\Delta}$ stores about the future target input $X_{t+\Delta}$, not how much information the current state Z_t stores about the history $X_{\le t}$. This distinction is important for a mathematically faithful PIB interpretation of temporal VJEPA.

6 Full IB-VJEPA: adding an explicit context-compression penalty

The preceding IB analysis showed that context–target VJEPA already targets *the relevance term* of IB through the latent NLL, but it does not *explicitly* penalize the information retained by the context latent Z_C about the context input X_C. As mentioned before, a direct recipe for completing the IB objective is therefore to replace the deterministic context encoder $Z_C = f_\theta(X_C)$ with a stochastic encoder and add a KL-to-prior compression term.

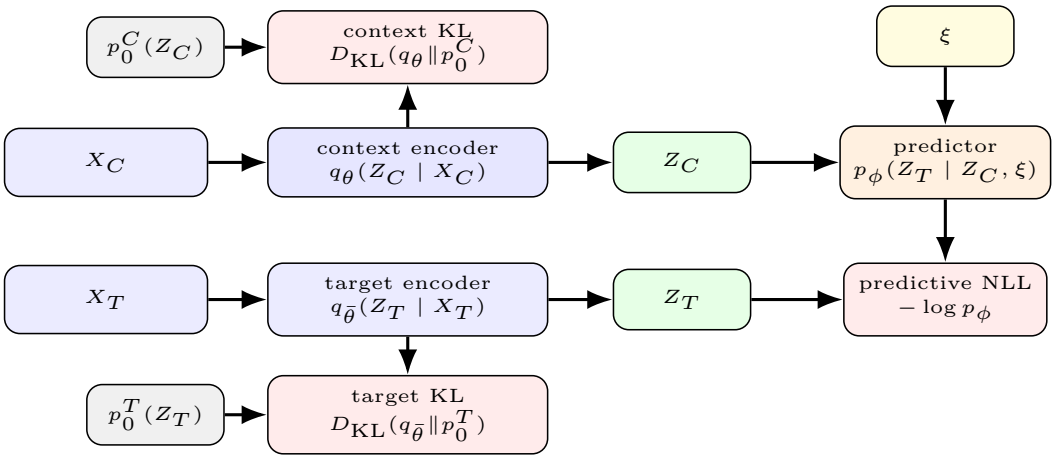

Figure 3: Full IB-VJEPA architecture. The deterministic context encoder is replaced by a stochastic context encoder $q_\theta(Z_C \mid X_C)$. The context KL against $p_0^C(Z_C)$ supplies explicit IB-style compression by upper-bounding $\mathrm{I}(X_C; Z_C)$; the target KL against $p_0^T(Z_T)$ regularizes the target latent; and the predictive NLL trains $p_\phi(Z_T \mid Z_C, \xi)$ to preserve context–target predictive information.

6.1 Stochastic context encoder

To obtain a full IB-style objective, replace the deterministic context encoder with a *stochastic context encoder*

$$Z_C \sim q_\theta(Z_C \mid X_C). \tag{32}$$

For example, one may use a Gaussian encoder

$$q_\theta(Z_C \mid X_C) = \mathcal{N}\left(Z_C; \mu_\theta(X_C), \mathrm{diag}(\sigma_\theta^2(X_C))\right). \tag{33}$$

This makes the context latent Z_C an explicit random variable induced by X_C, so that the representation information $\mathrm{I}(X_C; Z_C)$ becomes directly regularizable. The full IB-VJEPA architecture is shown in Fig.3.

6.2 Objective

The proposed full IB-VJEPA objective is

$$\begin{aligned}\mathcal{L}_{\text{IB-VJEPA}} =& \mathbb{E}_{(X_C, X_T, \xi) \sim p_{\text{data}}} \mathbb{E}_{Z_C \sim q_\theta(\cdot \mid X_C)} \mathbb{E}_{Z_T \sim q_{\bar{\theta}}(\cdot \mid X_T)} \left[-\log p_\phi(Z_T \mid Z_C, \xi)\right] \\ &+ \gamma_C \, \mathbb{E}_{X_C \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big) \\ &+ \beta_T \, \mathbb{E}_{X_T \sim p_{\text{data}}} D_{\mathrm{KL}}\big(q_{\bar{\theta}}(Z_T \mid X_T) \| p_0^T(Z_T)\big).\end{aligned} \tag{34}$$

The first term is the context–target latent NLL, now averaged over samples from both the stochastic context encoder $q_\theta(Z_C \mid X_C)$ and the stochastic target encoder $q_{\bar{\theta}}(Z_T \mid X_T)$. The second term is the new context-compression penalty, with $\gamma_C > 0$ controlling the strength of $X_C \to Z_C$ compression. The third term retains the target-side regularization of VJEPA, with $\beta_T > 0$ controlling the strength of target regularization, and is written using the target prior p_0^T to distinguish it from the context prior p_0^C.

The first two terms are the terms that correspond to the IB structure at the variational-objective level: the NLL term targets the relevance term $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$ through a *variational lower bound*, while the context KL term *upper-bounds* the compression term $\mathrm{I}(X_C; Z_C)$. The third term is not part of the classical IB objective; it is a *target-side* regularizer. By the same KL-to-prior decomposition, it upper-bounds the target-side information $\mathrm{I}(X_T; Z_T)$ and helps stabilize the target latent distribution.

6.3 Relation to IB

Define the aggregated context posterior $q_\theta(Z_C) = \int p_{\text{data}}(X_C) q_\theta(Z_C \mid X_C)\, dX_C$. By the KL-to-prior decomposition in Appendix B.7,

$$\mathbb{E}_{X_C \sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big) = \text{I}(X_C; Z_C) + D_{\text{KL}}\big(q_\theta(Z_C) \| p_0^C(Z_C)\big). \tag{35}$$

Thus, the new context KL regularization term *upper-bounds* the IB compression term $\text{I}(X_C; Z_C)$.

For the predictive part, the NLL term induces a joint distribution $q^\star(Z_C, \xi, Z_T)$ by sampling $(X_C, X_T, \xi) \sim p_{\text{data}}$, $Z_C \sim q_\theta(\cdot \mid X_C)$, and $Z_T \sim q_{\bar\theta}(\cdot \mid X_T)$. Under this induced distribution, the Barber–Agakov lower bound gives

$$\text{I}_{q^\star}(Z_C; Z_T \mid \xi) \geq H_{q^\star}(Z_T \mid \xi) + \mathbb{E}_{q^\star(Z_C, \xi, Z_T)} \left[\log p_\phi(Z_T \mid Z_C, \xi)\right]. \tag{36}$$

Therefore, when the target-latent marginal entropy $H_{q^\star}(Z_T \mid \xi)$ is controlled, minimizing the NLL term corresponds to maximizing a variational lower bound on the relevance term $\text{I}_{q^\star}(Z_C; Z_T \mid \xi)$.

Combining equation 35 and equation 36, the full IB-VJEPA objective can be interpreted as *a variational surrogate* for

$$\gamma_C\, \text{I}(X_C; Z_C) - \text{I}_{q^\star}(Z_C; Z_T \mid \xi), \tag{37}$$

up to the nonnegative aggregated-posterior mismatch $D_{\text{KL}}(q_\theta(Z_C) \| p_0^C(Z_C))$, the target-side KL regularizer, and the predictive variational gap between $p_\phi(Z_T \mid Z_C, \xi)$ and the induced population conditional $q^\star(Z_T \mid Z_C, \xi)$.

The target-side KL regularizer admits its own KL-to-prior decomposition:

$$\mathbb{E}_{X_T \sim p_{\text{data}}} D_{\text{KL}}\big(q_{\bar\theta}(Z_T \mid X_T) \| p_0^T(Z_T)\big) = \text{I}(X_T; Z_T) + D_{\text{KL}}\big(q_{\bar\theta}(Z_T) \| p_0^T(Z_T)\big), \tag{38}$$

where $q_{\bar\theta}(Z_T) = \int p_{\text{data}}(X_T) q_{\bar\theta}(Z_T \mid X_T)\, dX_T$ is the aggregated target posterior. Thus, this term controls the information retained by the target latent Z_T about the target input X_T. It is useful for stabilizing the target representation and controlling the target-latent marginal, but it is separate from the IB pair $\gamma_C \text{I}(X_C; Z_C) - \text{I}_{q^\star}(Z_C; Z_T \mid \xi)$.

In summary, the three terms in equation 34 play distinct information-theoretic roles: the context KL upper-bounds $\text{I}(X_C; Z_C)$, the latent NLL maximizes a Barber–Agakov lower bound on $\text{I}_{q^\star}(Z_C; Z_T \mid \xi)$, and the target KL upper-bounds $\text{I}(X_T; Z_T)$. Therefore, the full IB-VJEPA objective is not exactly the classical IB objective alone; rather, it is a variational IB surrogate $\gamma_C \text{I}(X_C; Z_C) - \text{I}_{q^\star}(Z_C; Z_T \mid \xi)$ augmented with target-side regularization that controls the information and stability of the target latent Z_T. In implementation, full IB-VJEPA can retain the usual JEPA/VJEPA target-branch asymmetry: the target encoder $q_{\bar\theta}(Z_T \mid X_T)$ may be treated as an EMA-stabilized teacher, with gradients stopped through the target branch when optimizing the context encoder $q_\theta(Z_C \mid X_C)$ and predictor p_ϕ. Thus, the added IB context-compression penalty changes the online/context side of the objective, while the target encoder continues to provide stable latent targets.

Proposition 3 (IB interpretation of full IB-VJEPA)**.** *Assume the target-latent marginal entropy $H_{q^\star}(Z_T \mid \xi)$ is controlled. Then minimizing equation 34 minimizes a variational surrogate of the IB objective $\gamma_C\, \text{I}(X_C; Z_C) - \text{I}_{q^\star}(Z_C; Z_T \mid \xi)$, up to nonnegative prior-matching, target-regularization, and predictive-approximation terms. If p_ϕ is expressive, the predictive variational gap can become small.*

Proof. The compression part follows from equation 35: the context KL equals $\text{I}(X_C; Z_C)$ plus the nonnegative aggregated-posterior mismatch $D_{\text{KL}}(q_\theta(Z_C) \| p_0^C(Z_C))$. The predictive part follows from the Barber–Agakov lower bound in equation 36, which shows that the NLL term maximizes a variational lower bound on $\text{I}_{q^\star}(Z_C; Z_T \mid \xi)$ when $H_{q^\star}(Z_T \mid \xi)$ is controlled. Expressivity of p_ϕ determines how tightly the variational predictor can approximate the induced conditional $q^\star(Z_T \mid Z_C, \xi)$. The remaining target KL regularizes $q_{\bar\theta}(Z_T \mid X_T)$ against $p_0^T(Z_T)$ and does not implement context compression. □

Algorithm 1 Full IB-VJEPA Training

Require: Dataset $\{x^{(i)}\}_{i=1}^{N}$, stochastic context encoder $q_\theta(Z_C \mid X_C)$, EMA target encoder $q_{\bar{\theta}}(Z_T \mid X_T)$, predictive model $p_\phi(Z_T \mid Z_C, \xi)$, context prior $p_0^C(Z_C)$, target prior $p_0^T(Z_T)$, EMA rate τ, context-compression weight γ_C, target-regularization weight β_T

1: **for** each minibatch $\{x\}$ **do**
2: Sample a context–target partition (X_C, X_T) and corresponding side information ξ
3: Compute context distribution $q_\theta(Z_C \mid X_C)$
4: Sample context latent $Z_C \sim q_\theta(Z_C \mid X_C)$, e.g. using the reparameterization trick
5: Compute target distribution $q_{\bar{\theta}}(Z_T \mid X_T)$
6: Sample target latent $Z_T \sim q_{\bar{\theta}}(Z_T \mid X_T)$, e.g. using the reparameterization trick
7: Evaluate predictive distribution $p_\phi(Z_T \mid Z_C, \xi)$
8: Compute the full IB-VJEPA loss (Eq. equation 34)

$$\mathcal{L} = -\log p_\phi(Z_T \mid Z_C, \xi) + \gamma_C D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big) + \beta_T D_{\mathrm{KL}}\big(q_{\bar{\theta}}(Z_T \mid X_T) \| p_0^T(Z_T)\big)$$

9: Update θ and ϕ by gradient descent on $\mathcal{L}$, stopping gradients through the EMA target branch if using the standard JEPA/VJEPA target-teacher design; any target-branch parameters not updated by EMA may be fixed or updated according to the chosen implementation
10: Update shared target-encoder parameters via EMA, when the online and target encoders have compatible parameterizations:

$$\bar{\theta} \leftarrow \tau\bar{\theta} + (1-\tau)\theta$$

11: **end for**

6.4 Why the stochastic context encoder matters

For continuous variables, a deterministic encoder $Z_C = f_\theta(X_C)$ can make $\mathrm{I}(X_C; Z_C)$ ill-behaved or infinite under common modeling assumptions. A stochastic encoder makes the context bottleneck operational by introducing conditional uncertainty in the representation. It also enables reparameterized training, calibrated context uncertainty, and direct comparison to variational IB objectives Alemi et al. (2017); Kingma & Welling (2014).

6.5 Training algorithm

Algorithm 1 summarizes a minimal training procedure for full IB-VJEPA. The key difference from the original VJEPA training algorithm is that the context branch now outputs a distribution $q_\theta(Z_C \mid X_C)$ rather than a deterministic embedding $Z_C = f_\theta(X_C)$, and the objective includes an explicit context-compression penalty.

When the context and target encoders have compatible parameterizations, the EMA update transfers the online stochastic encoder parameters θ into the target encoder parameters $\bar{\theta}$. If the target encoder uses additional parameters not shared with the context encoder, such as fixed or separately learned variance parameters, the EMA update can be applied only to the shared components while the remaining parameters are held fixed or updated according to the chosen implementation.

6.6 Computational overhead

Let d_C denote the dimension of the context latent Z_C and let d_T denote the dimension of the target latent Z_T. The original VJEPA training step (Algorithm 1 in Huang (2026)) requires a forward pass through the context encoder, a forward pass through the target encoder, evaluation of $p_\phi(Z_T \mid Z_C, \xi)$, and evaluation of the target KL. Full IB-VJEPA keeps these operations but replaces the deterministic context embedding with a stochastic encoder $q_\theta(Z_C \mid X_C)$ and adds the context KL: $D_{\mathrm{KL}}(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C))$.

For a diagonal-Gaussian context encoder, the additional arithmetic cost per sample is $O(d_C)$ for sampling Z_C by reparameterization and $O(d_C)$ for evaluating the closed-form KL to $p_0^C(Z_C)$, assuming a standard

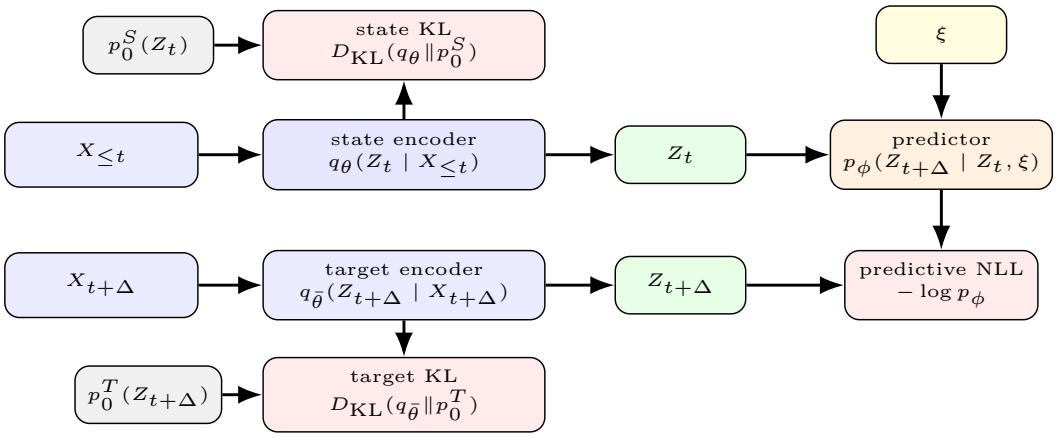

Figure 4: Full PIB-VJEPA architecture. The deterministic temporal state $Z_t = f_\theta(X_{\le t})$ is replaced by a stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$. The state KL against $p_0^S(Z_t)$ supplies explicit PIB-style compression by upper-bounding $\mathrm{I}(X_{\le t}; Z_t)$; the target KL against $p_0^T(Z_{t+\Delta})$ regularizes the future target latent; and the predictive NLL trains $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$ to preserve future-latent predictive information.

Gaussian or diagonal Gaussian prior (see Appendix C.5). The target-side cost remains the same as in VJEPA. Hence, if the encoder and predictor networks dominate computation, full IB-VJEPA has the same asymptotic complexity as VJEPA, with only a small constant-factor overhead from the additional context-variance head, reparameterized sampling, and context KL evaluation.

7 Full PIB-VJEPA: adding an explicit state-compression penalty

Full PIB-VJEPA is parallel to, and can be viewed as the temporal specialization of full IB-VJEPA under the identifications $X_C = X_{\le t}$, $Z_C = Z_t$, $X_T = X_{t+\Delta}$, and $Z_T = Z_{t+\Delta}$. Under this specialization, the stochastic context encoder $q_\theta(Z_C \mid X_C)$ becomes the stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$, and the context prior $p_0^C(Z_C)$ becomes the state prior $p_0^S(Z_t)$. Thus, full IB-VJEPA provides the general context–target recipe, while full PIB-VJEPA provides its temporal predictive-state specialization.

The preceding PIB analysis showed that temporal VJEPA already targets future-predictive information through the latent NLL, but it does not explicitly penalize the information retained by the current state Z_t about the history $X_{\le t}$. The PIB recipe is therefore to replace the deterministic state encoder $Z_t = f_\theta(X_{\le t})$ with a stochastic encoder and add a KL-to-prior state-compression term.

7.1 Stochastic current-state encoder

To obtain a full PIB-style objective, replace the deterministic current-state encoder with a *stochastic current-state encoder*

$$Z_t \sim q_\theta(Z_t \mid X_{\le t}). \tag{39}$$

For example, one may use a Gaussian state encoder

$$q_\theta(Z_t \mid X_{\le t}) = \mathcal{N}\left(Z_t; \mu_\theta(X_{\le t}), \mathrm{diag}(\sigma_\theta^2(X_{\le t}))\right). \tag{40}$$

This turns the current latent state into an explicit random variable induced by the history, making the representation information $\mathrm{I}(X_{\le t}; Z_t)$ directly regularizable. The full PIB-VJEPA architecture is shown in Fig.4.

7.2 Objective

The proposed full PIB-VJEPA objective is

$$\begin{aligned}\mathcal{L}_{\text{PIB-VJEPA}} =& \mathbb{E}_{(X_{\le t}, X_{t+\Delta}, \xi)\sim p_{\text{data}}} \mathbb{E}_{Z_t \sim q_\theta(\cdot \mid X_{\le t})} \mathbb{E}_{Z_{t+\Delta}\sim q_{\bar\theta}(\cdot \mid X_{t+\Delta})} \left[-\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right] \\ &+ \gamma_S \, \mathbb{E}_{X_{\le t}\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t)\big) \\ &+ \beta_T \, \mathbb{E}_{X_{t+\Delta}\sim p_{\text{data}}} D_{\text{KL}}\big(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big).\end{aligned} \tag{41}$$

The first term is the temporal latent NLL, now averaged over samples from both the stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$ and the stochastic target encoder $q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$. The second term is the new state-compression penalty, with $\gamma_S > 0$ controlling the strength of $X_{\le t} \to Z_t$ compression. The third term retains the target-side regularization of VJEPA, with $\beta_T > 0$ controlling the strength of target regularization, and is written using the target prior p_0^T to distinguish it from the state prior p_0^S.

The first two terms are the terms that correspond to the PIB structure at the variational-objective level: the NLL term targets the predictive information term $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ through a *variational lower bound*, while the state KL term *upper-bounds* the compression term $\mathrm{I}(X_{\le t}; Z_t)$. The third term is not part of the classical PIB objective; it is a *target-side* regularizer. By the same KL-to-prior decomposition, it upper-bounds the target-side information $\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta})$ and helps stabilize the future-latent distribution.

7.3 Relation to PIB

Define the aggregated state posterior $q_\theta(Z_t) = \int p_{\text{data}}(X_{\le t}) q_\theta(Z_t \mid X_{\le t}) \, dX_{\le t}$. By the PIB-VJEPA KL-to-prior decomposition proved in Appendix C.2,

$$\mathbb{E}_{X_{\le t}\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t)\big) = \mathrm{I}(X_{\le t}; Z_t) + D_{\text{KL}}\big(q_\theta(Z_t) \| p_0^S(Z_t)\big). \tag{42}$$

Thus, the new state KL regularization term *upper-bounds* the PIB compression term $\mathrm{I}(X_{\le t}; Z_t)$.

For the predictive part, the NLL term induces a joint distribution $q^\star(Z_t, \xi, Z_{t+\Delta})$ by sampling $(X_{\le t}, X_{t+\Delta}, \xi) \sim p_{\text{data}}$, $Z_t \sim q_\theta(\cdot \mid X_{\le t})$, and $Z_{t+\Delta} \sim q_{\bar\theta}(\cdot \mid X_{t+\Delta})$. Under this induced distribution, the Barber–Agakov lower bound gives

$$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) \ge H_{q^\star}(Z_{t+\Delta} \mid \xi) + \mathbb{E}_{q^\star(Z_t, \xi, Z_{t+\Delta})} \left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. \tag{43}$$

Therefore, when the future-latent marginal entropy $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled, minimizing the NLL term corresponds to maximizing a variational lower bound on the predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$.

Combining equation 42 and equation 43, the full PIB-VJEPA objective can be interpreted as *a variational surrogate* for

$$\gamma_S \, \mathrm{I}(X_{\le t}; Z_t) - \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi), \tag{44}$$

up to the nonnegative aggregated-posterior mismatch $D_{\text{KL}}(q_\theta(Z_t) \| p_0^S(Z_t))$, the target-side KL regularizer, and the predictive variational gap between $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$ and the induced population conditional $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$.

The target-side KL regularizer admits its own KL-to-prior decomposition:

$$\mathbb{E}_{X_{t+\Delta}\sim p_{\text{data}}} D_{\text{KL}}\big(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big) = \mathrm{I}(X_{t+\Delta}; Z_{t+\Delta}) + D_{\text{KL}}\big(q_{\bar\theta}(Z_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big). \tag{45}$$

where $q_{\bar\theta}(Z_{t+\Delta}) = \int p_{\text{data}}(X_{t+\Delta}) q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \, dX_{t+\Delta}$ is the aggregated future-target posterior. Thus, this term controls the information retained by the future latent $Z_{t+\Delta}$ about the future target input $X_{t+\Delta}$. It is useful for stabilizing the target/future representation and controlling the future-latent marginal, but it is separate from the PIB pair $\gamma_S \mathrm{I}(X_{\le t}; Z_t) - \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$.

In summary, the three terms in equation 41 play distinct information-theoretic roles: the state KL upper-bounds $\mathrm{I}(X_{\le t}; Z_t)$, the temporal NLL maximizes a Barber–Agakov lower bound on $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$, and the target KL upper-bounds $\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta})$. Therefore, the full PIB-VJEPA objective is not exactly the classical

PIB objective alone; rather, it is a variational PIB surrogate $\gamma_S \mathrm{I}(X_{\le t}; Z_t) - \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ augmented with target-side regularization that controls the information and stability of the future latent $Z_{t+\Delta}$. In implementation, full PIB-VJEPA can retain the usual JEPA/VJEPA target-branch asymmetry: the target encoder $q_{\bar{\theta}}(Z_{t+\Delta} \mid X_{t+\Delta})$ may be treated as an EMA-stabilized teacher, with gradients stopped through the target branch when optimizing the stochastic state encoder $q_\theta(Z_t \mid X_{\le t})$ and predictor p_ϕ. Thus, the added PIB state-compression penalty changes the online/state side of the objective, while the target encoder continues to provide stable future-latent targets.

Proposition 4 (PIB interpretation of full PIB-VJEPA)**.** *Assume the future-latent marginal entropy $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled. Then minimizing equation 41 minimizes a variational surrogate of the PIB objective $\gamma_S \, \mathrm{I}(X_{\le t}; Z_t) - \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$, up to nonnegative prior-matching, target-regularization, and predictive-approximation terms. If p_ϕ is expressive, the predictive variational gap can become small.*

Proof. The compression part follows from equation 42: the state KL equals $\mathrm{I}(X_{\le t}; Z_t)$ plus the nonnegative aggregated-posterior mismatch $D_{\mathrm{KL}}(q_\theta(Z_t) \| p_0^S(Z_t))$. The predictive part follows from the Barber–Agakov lower bound in equation 43, which shows that the temporal NLL maximizes a variational lower bound on $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ when $H_{q^\star}(Z_{t+\Delta} \mid \xi)$ is controlled. Expressivity of p_ϕ determines how tightly the variational predictor can approximate the induced conditional $q^\star(Z_{t+\Delta} \mid Z_t, \xi)$. The remaining target KL regularizes $q_{\bar{\theta}}(Z_{t+\Delta} \mid X_{t+\Delta})$ against $p_0^T(Z_{t+\Delta})$ and does not implement state compression. □

7.4 Why the stochastic state encoder matters

For continuous variables, a deterministic encoder $Z_t = f_\theta(X_{\le t})$ can make $\mathrm{I}(X_{\le t}; Z_t)$ ill-behaved or infinite under common modeling assumptions. A stochastic encoder makes the temporal bottleneck operational by introducing conditional uncertainty in the current state. It also enables reparameterized training, calibrated state uncertainty, and direct comparison to variational PIB objectives Alemi et al. (2017); Kingma & Welling (2014).

7.5 Training algorithm

Algorithm 2 summarizes a minimal training procedure for full PIB-VJEPA. The key difference from temporal VJEPA is that the online/state branch now outputs a distribution $q_\theta(Z_t \mid X_{\le t})$ rather than a deterministic state $Z_t = f_\theta(X_{\le t})$, and the objective includes an explicit state-compression penalty.

When the state encoder and future-target encoder have compatible parameterizations, the EMA update transfers the online stochastic state-encoder parameters θ into the target encoder parameters $\bar{\theta}$. If the target encoder uses additional parameters not shared with the state encoder, such as fixed or separately learned variance parameters, the EMA update can be applied only to the shared components while the remaining parameters are held fixed or updated according to the chosen implementation.

7.6 Computational overhead

Let d_S denote the dimension of the current state Z_t and let d_T denote the dimension of the future latent $Z_{t+\Delta}$. The temporal VJEPA training step requires a forward pass through the state/context encoder, a forward pass through the target encoder, evaluation of $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$, and evaluation of the target KL. Full PIB-VJEPA keeps these operations but replaces the deterministic current state with a stochastic encoder $q_\theta(Z_t \mid X_{\le t})$ and adds the state KL: $D_{\mathrm{KL}}(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t))$.

For a diagonal-Gaussian state encoder, the additional arithmetic cost per sample is $O(d_S)$ for sampling Z_t by reparameterization and $O(d_S)$ for evaluating the closed-form KL to $p_0^S(Z_t)$, assuming a standard Gaussian or diagonal Gaussian prior (see Appendix C.5, with Z_C replaced by Z_t and p_0^C replaced by p_0^S). The target-side cost remains the same as in temporal VJEPA. Hence, if the encoder and predictor networks dominate computation, full PIB-VJEPA has the same asymptotic complexity as temporal VJEPA, with only a small constant-factor overhead from the additional state-variance head, reparameterized sampling, and state KL evaluation.

Algorithm 2 Full PIB-VJEPA Training

Require: Dataset $\{x^{(i)}\}_{i=1}^{N}$, stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$, EMA target encoder $q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$, predictive model $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$, state prior $p_0^S(Z_t)$, target prior $p_0^T(Z_{t+\Delta})$, EMA rate τ, state-compression weight γ_S, target-regularization weight β_T

1: **for** each minibatch $\{x\}$ **do**
2: Sample a temporal context–future pair $(X_{\le t}, X_{t+\Delta})$ and corresponding side information ξ
3: Compute current-state distribution $q_\theta(Z_t \mid X_{\le t})$
4: Sample current state $Z_t \sim q_\theta(Z_t \mid X_{\le t})$, e.g. using the reparameterization trick
5: Compute future target distribution $q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$
6: Sample future latent $Z_{t+\Delta} \sim q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})$, e.g. using the reparameterization trick
7: Evaluate predictive distribution $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$
8: Compute the full PIB-VJEPA loss (Eq. equation 41)

$$\mathcal{L} = -\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi) + \gamma_S D_{\mathrm{KL}}\big(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t)\big) + \beta_T D_{\mathrm{KL}}\big(q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta}) \| p_0^T(Z_{t+\Delta})\big)$$

9: Update θ and ϕ by gradient descent on $\mathcal{L}$, stopping gradients through the EMA target branch if using the standard JEPA/VJEPA target-teacher design; any target-branch parameters not updated by EMA may be fixed or updated according to the chosen implementation
10: Update shared target-encoder parameters via EMA, when the online and target encoders have compatible parameterizations:

$$\bar\theta \leftarrow \tau\bar\theta + (1-\tau)\theta$$

11: **end for**

8 Discussion

This paper is primarily a theoretical and objective-level analysis of VJEPA. Rather than introducing a new empirical benchmark, we clarify what the existing context–target and temporal VJEPA objectives already optimize, what information-theoretic terms are missing from full IB/PIB interpretations, and how these missing terms can be added through variational KL-to-prior penalties.

What current VJEPA already optimizes. The current VJEPA objective already captures the predictive part of an IB/PIB principle. In the context–target form, the latent NLL targets a variational lower bound on the context–target relevance term $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$. In the temporal form, the same NLL targets a variational lower bound on the future-predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$. Thus, VJEPA is not merely a probabilistic regression objective: its predictive likelihood has a precise information-theoretic interpretation through the Barber–Agakov bound.

However, the current VJEPA objective does not *explicitly* penalize how much information the context latent Z_C retains about X_C, or how much information the temporal state Z_t retains about $X_{\le t}$. These quantities are the compression terms $\mathrm{I}(X_C; Z_C)$ for IB and $\mathrm{I}(X_{\le t}; Z_t)$ for PIB. Existing VJEPA may still compress through architecture, finite latent dimension, masking, EMA dynamics, and target-side regularization, but this compression remains implicit rather than an explicit term in the optimization objective.

Why the target KL is not the bottleneck compression term. A key clarification is that the target KL in VJEPA has a different role from the IB/PIB compression term. By the KL-to-prior decomposition, the target KL upper-bounds target-side information such as $\mathrm{I}(X_T; Z_T)$ or, in the temporal case, $\mathrm{I}(X_{t+\Delta}; Z_{t+\Delta})$. This can stabilize the target latent space, discourage uncontrolled target representations, and influence marginal entropy terms such as $H_{q^\star}(Z_T \mid \xi)$ or $H_{q^\star}(Z_{t+\Delta} \mid \xi)$. But it does *not* directly control the context/state compression terms $\mathrm{I}(X_C; Z_C)$ or $\mathrm{I}(X_{\le t}; Z_t)$.

This distinction is important as it separates three different roles: the latent NLL promotes predictive relevance, the target KL regularizes the target representation, and the missing context/state KL supplies explicit

bottleneck compression. Without this separation, it is easy to mis-interpret the target KL as a full IB/PIB compression mechanism.

What full IB-VJEPA adds. Full IB-VJEPA completes the context–target IB interpretation by replacing the deterministic context encoder $Z_C = f_\theta(X_C)$ with a stochastic encoder $q_\theta(Z_C \mid X_C)$ and adding the KL-to-prior penalty $\mathbb{E}_{X_C \sim p_{\text{data}}} D_{\text{KL}}(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C))$. This term upper-bounds $\text{I}(X_C; Z_C)$ up to the aggregated-posterior mismatch $D_{\text{KL}}(q_\theta(Z_C) \| p_0^C(Z_C))$. Therefore, full IB-VJEPA turns context compression from an *implicit* architectural effect into an *explicit* variational penalty.

The resulting objective is not exactly the classical IB objective alone. It is a variational IB surrogate $\gamma_C \text{I}(X_C; Z_C) - \text{I}_{q^\star}(Z_C; Z_T \mid \xi)$ augmented with target-side regularization[10]. This extra target-side term remains useful because the target encoder defines the representation space in which prediction is evaluated.

What full PIB-VJEPA adds. Full PIB-VJEPA is the temporal specialization of full IB-VJEPA. Under the identifications $X_C = X_{\le t}$, $Z_C = Z_t$, $X_T = X_{t+\Delta}$, and $Z_T = Z_{t+\Delta}$, the stochastic context encoder becomes a stochastic current-state encoder $q_\theta(Z_t \mid X_{\le t})$, and the context prior $p_0^C(Z_C)$ becomes a state prior $p_0^S(Z_t)$. The added state KL $\mathbb{E}_{X_{\le t} \sim p_{\text{data}}} D_{\text{KL}}(q_\theta(Z_t \mid X_{\le t}) \| p_0^S(Z_t))$ upper-bounds $\text{I}(X_{\le t}; Z_t)$ and therefore supplies the missing PIB compression term.

This changes the interpretation of temporal VJEPA. Current temporal VJEPA is predictive-PIB-like: it promotes future-predictive latent information, but only *implicitly* compresses history. Full PIB-VJEPA makes the trade-off *explicit*: the state should retain enough information to predict $Z_{t+\Delta}$, but not so much information that it memorizes nuisance variation in $X_{\le t}$.

Design implications. The analysis suggests several design principles. *First*, the target encoder matters because it defines what the model is asked to predict. If the target latent preserves nuisance variation, then the predictive objective may also preserve nuisance information. *Second*, the compression weights γ_C (Eq. 34) and γ_S (Eq. 41) control the strength of the context/state bottleneck: weak compression may leak nuisance information, while excessive compression may remove predictive or control-relevant information from the input. *Third*, the expressivity of p_ϕ determines whether the model can represent uncertain or multimodal futures; a simple Gaussian, especially an isotropic or diagonal one, may be insufficient when several future latents are plausible. *Finally*, multi-horizon prediction can strengthen the temporal bottleneck by encouraging representations that remain predictive beyond immediate next-step correlations.

Scope and limitations. The present work is theoretical. It establishes objective-level correspondences, variational bounds, and principled extensions of VJEPA, but we do not yet claim empirical superiority of full IB-VJEPA or full PIB-VJEPA over current VJEPA. The practical benefit of the added bottleneck terms will depend on model class, target-encoder design, the amount and type of nuisance variation, and how the compression weights are tuned. For example, if the task requires retaining fine-grained information from the context or history, an overly strong KL penalty may degrade performance. Conversely, in noisy or distractor-rich settings, explicit compression may improve robustness, calibration, and downstream control.

Future empirical validation. A natural next step is to empirically validate the proposed objectives. Since the present paper focuses on theory and objective design, we leave empirical validation to follow-up work. Several directions are particularly relevant: *first*, noisy-distractor sequence benchmarks can test whether explicit context/state compression improves robustness when nuisance entropy increases. One can compare deterministic JEPA, current VJEPA, full IB-VJEPA, and full PIB-VJEPA using predictive NLL, calibration, linear-probe recovery of predictive factors, and sensitivity to nuisance variables. *Second*, multimodal future benchmarks can test whether probabilistic predictors with sufficient expressivity improve over deterministic or diagonal-Gaussian heads. Branching dynamical systems, stochastic video patches, or controlled latent systems with multiple plausible futures would be suitable testbeds. *Third*, latent model-predictive-control benchmarks can evaluate whether explicit state compression improves planning robustness under nuisance perturbations. In this setting, the predictor may be action-conditioned, for example $p_\phi(Z_{t+\Delta} \mid Z_t, u_t, \xi)$, and

[10] Removing the target-side KL regularizer yields the core variational IB-VJEPA objective, but no longer retains the target-side regularization inherited from VJEPA.

the key question is whether full PIB-VJEPA learns more stable predictive states for downstream planning. *Finally*, information diagnostics can directly visualize the predicted trade-off. Variational upper bounds on $\mathrm{I}(X_{\leq t}; Z_t)$ and Barber–Agakov lower bounds on $\mathrm{I}(Z_t; Z_{t+\Delta} \mid \xi)$ can be tracked during training to determine whether full PIB-VJEPA realizes the intended compression–prediction frontier.

9 Related works

Joint-embedding predictive architectures. JEPA was proposed as a latent predictive architecture in which learning proceeds by predicting representations of missing, masked, or future observations rather than reconstructing observations in pixel space LeCun (2022). I-JEPA instantiated this idea for images by predicting the representations of target blocks from a context block within the same image, using masking strategies designed to encourage semantic representation learning Assran et al. (2023). V-JEPA extended feature prediction to video and demonstrated that visual representations can be learned from video through feature prediction alone, without relying on pixel reconstruction, text supervision, negative examples, or pretrained image encoders Bardes et al. (2024). These methods provide the architectural foundation for the present work. Our focus, however, is different: rather than proposing a new JEPA architecture or empirical pretraining recipe, we analyze the information-theoretic content of the recently proposed VJEPA[11] Huang (2026) and identify which objective terms correspond to predictive information, target-side regularization, and explicit bottleneck compression.

Contrastive and non-contrastive self-supervised learning. JEPA is closely related to the broader family of self-supervised representation learning methods. Contrastive Predictive Coding learns representations by predicting future latent variables using a contrastive loss and negative samples van den Oord et al. (2019), while SimCLR studies contrastive view prediction in vision and shows the importance of augmentations, projection heads, batch size, and training duration Chen et al. (2020). Non-contrastive methods avoid explicit negative samples but require mechanisms to prevent collapse. BYOL uses online and target networks, with the target network updated by a slow-moving average of the online network Grill et al. (2020). Barlow Twins avoids collapse by matching cross-correlations to the identity and reducing redundancy between embedding dimensions Zbontar et al. (2021). VICReg combines invariance, variance, and covariance regularization to avoid collapse through explicit representation regularizers Bardes et al. (2022). In contrast to these works, our analysis does not introduce a new collapse-prevention regularizer; instead, we study VJEPA as a probabilistic latent predictive objective and show how explicit IB/PIB compression penalties can be added to control the information retained by the context or state.

Masked reconstruction versus latent prediction. Masked autoencoders reconstruct missing pixels from visible patches and have shown strong scalability in vision pretraining He et al. (2022). JEPA-style methods instead predict latent target representations, which avoids placing the main learning objective directly on high-dimensional, potentially noisy observation reconstruction. This distinction is central to the VJEPA formulation Huang (2026): *the probabilistic predictor models uncertainty in latent space rather than defining a likelihood over raw observations.* The present work complements reconstruction-based SSL by asking what information-theoretic objective is implicitly optimized when prediction is performed in latent space, and what extra terms are needed to obtain an explicit bottleneck objective.

Information Bottleneck and variational bottlenecks. The Information Bottleneck principle formalizes representation learning as the search for a compressed representation Z of an input X that preserves information about a relevance variable Y Tishby et al. (1999). Deep Variational Information Bottleneck constructs a tractable variational objective for the IB Lagrangian[12] using a stochastic encoder, a variational decoder for the relevance term, and an upper bound on $I(X; Z)$ obtained by replacing the intractable

[11] The name VJEPA is used without a hyphen to emphasize it as a general probabilistic JEPA framework, rather than task-specific variants.

[12] Unlike our explicit KL-to-prior decomposition, which separates an expected KL term into $\mathrm{I}(X_C; Z_C)$ plus an aggregated-posterior mismatch, Deep VIB uses a variational marginal $r(z)$ to upper-bound $I(X; Z)$ and thereby obtains a tractable KL regularizer.

marginal $p(z)$ with a tractable variational marginal $r(z)$ Alemi et al. (2017). Others have studied information-plane interpretations of deep networks Shwartz-Ziv & Tishby (2017), deterministic variants of IB Strouse & Schwab (2017), nonlinear and neural approximations to IB Kolchinsky et al. (2019), and practical issues in estimating or optimizing information-theoretic objectives Poole et al. (2019). Our full IB-VJEPA objective follows this variational tradition, but applies it specifically to the context–target structure of VJEPA: the stochastic context encoder $q_\theta(Z_C \mid X_C)$ and context KL term provide an explicit upper bound on $\mathrm{I}(X_C; Z_C)$, while the latent NLL preserves target-predictive information.

Predictive Information Bottleneck. Predictive information measures how much the past of a process tells us about its future Bialek et al. (2001). The Predictive Information Bottleneck extends the IB idea to temporal prediction by compressing past observations while preserving information useful for future prediction Still (2014). Related variational formulations connect predictive bottleneck objectives to variational inference and Bayesian updating Alemi (2020). State Predictive Information Bottleneck methods have also been used to learn low-dimensional predictive state variables in dynamical systems, for example in molecular simulation settings Wang & Tiwary (2021). Our full PIB-VJEPA objective can be viewed as a temporal specialization of full IB-VJEPA: it introduces $q_\theta(Z_t \mid X_{\leq t})$ and a state KL term to upper-bound $\mathrm{I}(X_{\leq t}; Z_t)$, while minimizing the temporal latent NLL maximizes a Barber–Agakov lower bound on the predictive information $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$.

Variational mutual-information bounds. Our analysis relies on the standard connection between predictive log-likelihood, conditional entropy, and variational lower bounds on mutual information. The Barber–Agakov bound provides a variational lower bound on mutual information by replacing an intractable conditional distribution with a tractable variational predictor Barber & Agakov (2003). Later work systematized variational mutual-information bounds and clarified their bias-variance and estimation properties Poole et al. (2019). In this paper, the VJEPA latent NLL plays the role of such a variational predictive term: under the induced latent population distribution $q^\star$, minimizing it maximizes Barber–Agakov lower bounds on $\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$ in the context–target case and $\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$ in the temporal case.

10 Conclusion

We provided a formal IB/PIB analysis of VJEPA. The main finding is that current VJEPA already optimizes the predictive side of a bottleneck principle: its latent NLL maximizes a variational lower bound on *context–target* or *future-latent* predictive information. However, its target KL regularizes the target representation rather than explicitly compressing the context or current state. Therefore, current VJEPA design is best understood as a *partial* IB/PIB objective.

We then derived two bottleneck-complete objectives. *Full IB-VJEPA* adds a *stochastic context encoder* and a *KL-to-prior penalty* that upper-bounds the context information $\mathrm{I}(X_C; Z_C)$. *Full PIB-VJEPA* specializes this construction to temporal prediction by adding a *stochastic current-state encoder* and a *KL-to-prior penalty* that upper-bounds the state information $\mathrm{I}(X_{\leq t}; Z_t)$. These objectives separate predictive information, target-side regularization, and explicit bottleneck compression into distinct terms.

Overall, this work turns the informal intuition that VJEPA learns predictive compressed representations into a precise variational statement. We show exactly which part of VJEPA implements predictive information maximization, which part stabilizes the target latent space, and what must be added to obtain explicit IB/PIB-style bottleneck control. This provides a principled foundation for future empirical studies of compression-controlled, uncertainty-aware VJEPA models.

References

Alexander A. Alemi. Variational predictive information bottleneck. In Cheng Zhang, Francisco Ruiz, Thang Bui, Adji Bousso Dieng, and Dawen Liang (eds.), *Proceedings of The 2nd Symposium on Advances in Approximate Bayesian Inference*, volume 118 of *Proceedings of Machine Learning Research*, pp. 1–6. PMLR, 08 Dec 2020. URL https://proceedings.mlr.press/v118/alemi20a.html.

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=HyxQzBceg.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

David Barber and Felix Agakov. The im algorithm: a variational approach to information maximization. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, pp. 201–208, Cambridge, MA, USA, 2003. MIT Press.

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

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.

William Bialek, Ilya Nemenman, and Naftali Tishby. Predictability, complexity, and learning. *Neural Comput.*, 13(11):2409–2463, November 2001. ISSN 0899-7667. doi: 10.1162/089976601753195969. URL https://doi.org/10.1162/089976601753195969.

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*, pp. 1597–1607. PMLR, 2020.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

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.

Yongchao Huang. Avariational joint embedding predictive architectures as probabilistic world models. In *International Conference on Machine Learning*, 2026.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Artemy Kolchinsky, Brendan D. Tracey, and David H. Wolpert. Nonlinear information bottleneck. *Entropy*, 21(12), 2019. ISSN 1099-4300. doi: 10.3390/e21121181. URL https://www.mdpi.com/1099-4300/21/12/1181.

Yann LeCun. A path towards autonomous machine intelligence version 0.9.2, 2022-06-27. *Open Review*, 62 (1):1–62, 2022.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning, 2019. URL https://arxiv.org/abs/1509.02971.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis

Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518:529–533, 2015. doi: 10.1038/nature14236.

Volodymyr Mnih, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Timothy P. Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning, 2016. URL https://arxiv.org/abs/1602.01783.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information, 2017. URL https://arxiv.org/abs/1703.00810.

Susanne Still. Information bottleneck approach to predictive inference. *Entropy*, 16(2):968–989, 2014.

DJ Strouse and David J. Schwab. The deterministic information bottleneck. *Neural Computation*, 29(6): 1611–1630, 06 2017. ISSN 0899-7667. doi: 10.1162/NECO_a_00961. URL https://doi.org/10.1162/NECO_a_00961.

Naftali Tishby, Fernando C. Pereira, and William Bialek. The information bottleneck method. In *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 1999.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019. URL https://arxiv.org/abs/1807.03748.

Dedi Wang and Pratyush Tiwary. State predictive information bottleneck. *The Journal of Chemical Physics*, 154(13):134111, 04 2021. ISSN 0021-9606. doi: 10.1063/5.0038198. URL https://doi.org/10.1063/5.0038198.

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*, pp. 12310–12320. PMLR, 2021.

A Notations

This appendix summarizes the notation used throughout the paper. We use uppercase letters for random variables and lowercase letters for their realizations. For example, $X_{\leq t}$ denotes the random history, while $x_{\leq t}$ denotes a particular observed history. Similarly, Z_t denotes the current latent random variable, while z_t denotes a realization.

Symbol	Meaning
X, Z, Y	Generic variables used only for the classical IB principle: X is the input, Z is the compressed representation, and Y is the relevance variable. In VJEPA, these are instantiated as (X_C, Z_C, Z_T) for context–target prediction or as $(X_{\leq t}, Z_t, Z_{t+\Delta})$ for temporal PIB.
$q(z \mid x)$	Generic stochastic encoder used in the IB formulation.
$\mathrm{I}(X;Z)$	Mutual information between input X and representation Z; interpreted as representation complexity.
$\mathrm{I}(Z;Y)$	Mutual information between representation Z and relevance variable Y; interpreted as predictive relevance.
$\mathcal{L}_{\mathrm{IB}}$	Classical Information Bottleneck objective, $\mathrm{I}(X;Z) - \beta\mathrm{I}(Z;Y)$.

Symbol	Meaning
β	Generic weight. In the classical IB objective, it controls the relevance–compression trade-off; in the original VJEPA objective, it weights the target KL term. In the full IB/PIB-VJEPA objectives, the target-KL weight is denoted by β_T.
Temporal and predictive bottleneck notation	
$X_{\leq t}$, $X_{\leq t}$	History up to time t. This is the past information to be compressed in PIB.
$x_{\leq t}$	A realization of the history random variable $X_{\leq t}$.
$X_{t+\Delta}$, $X_{t+\Delta}$	Future observation, view, or target segment at prediction offset Δ.
Z_t, Z_t	Current latent state or predictive state at time t.
z_t	A realization of the current latent state Z_t.
$Z_{t+\Delta}$, $Z_{t+\Delta}$	Future latent target at offset Δ.
$z_{t+\Delta}$	A realization of the future latent target $Z_{t+\Delta}$.
$\mathcal{L}_{\text{PIB}}$	Predictive Information Bottleneck objective, $\mathrm{I}(X_{\leq t}; Z_t) - \lambda \mathrm{I}(Z_t; Z_{t+\Delta})$.
λ	Weight on predictive mutual information in the PIB objective.
$\mathrm{I}(X_{\leq t}; Z_t)$	Past-to-state representation information; the compression term in PIB.
$\mathrm{I}(Z_t; Z_{t+\Delta})$	Future-latent predictive information; the prediction/relevance term in PIB.
$\mathrm{I}(Z_t; Z_{t+\Delta} \mid \xi)$	Conditional predictive information given side information ξ.
JEPA and VJEPA variables	
X_C, X_C	Context input in JEPA/VJEPA. In temporal settings this may be a history or a masked context.
x_C	A realization of the context input X_C.
X_T, X_T	Target input in JEPA/VJEPA. In temporal settings this may be a future observation, view, or target segment.
x_T	A realization of the target input X_T.
Z_C, Z_C	Context latent representation produced from X_C.
Z_T, Z_T	Target latent representation associated with X_T.
ξ or ξ_T	Side information, such as mask position, temporal offset, action, goal, or structural descriptor.
f_θ	Online/context encoder. In deterministic VJEPA it maps X_C to $Z_C = f_\theta(X_C)$.
$f_{\bar{\theta}}$ or $f_{\theta'}$	Target encoder, often updated by exponential moving average rather than direct backpropagation.
θ	Parameters of the online/context encoder. In full IB/PIB-VJEPA, θ also denotes the parameters of the stochastic context/state encoder.
$\bar{\theta}$	Parameters of the EMA target encoder.
ϕ	Parameters of the predictive model p_ϕ. These are point-estimated neural-network parameters, not Bayesian random weights.
g_ϕ	Deterministic JEPA predictor, e.g. $\hat{Z}_T = g_\phi(Z_C, \xi_T)$.
$p_\phi(Z_T \mid Z_C, \xi)$	VJEPA probabilistic predictive model over target latents conditioned on the context latent and side information.
$p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$	Temporal VJEPA probabilistic predictive model over future latents conditioned on the current latent state and side information.
$q_{\bar{\theta}}(Z_T \mid X_T)$	Target inference distribution produced by the target encoder. In a simple implementation it may be an isotropic or diagonal Gaussian around the target embedding.
$q_{\bar{\theta}}(Z_{t+\Delta} \mid X_{t+\Delta})$	Temporal target inference distribution over future latents.
$\mathcal{L}_{\text{VJEPA}}$	Current context–target VJEPA objective: latent negative log-likelihood plus a target-side KL regularizer.
$\mathcal{L}_{\text{temp-VJEPA}}$	Temporal specialization of the current VJEPA objective, obtained by setting $X_C = X_{\leq t}$, $Z_C = Z_t$, $X_T = X_{t+\Delta}$, and $Z_T = Z_{t+\Delta}$.

Symbol	Meaning
$p_{\text{data}}(X_C, X_T, \xi)$	Training distribution over context–target–side-information triples. This includes the empirical data distribution and the masking/partitioning process.
$p_{\text{data}}(X_{\le t}, X_{t+\Delta}, \xi)$	Training distribution over temporal history–future–side-information triples. This includes the empirical data distribution and the temporal sampling or masking process.
$p_0(Z_T)$	Prior over target latents in the current VJEPA objective when no ambiguity with other priors exists. Usually chosen as $\mathcal{N}(0, I)$.
Full IB-VJEPA notation	
$q_\theta(Z_C \mid X_C)$	Stochastic context encoder introduced for full IB-VJEPA.
$p_0^C(Z_C)$	Prior over context latents. The superscript C denotes context.
$p_0^T(Z_T)$	Prior over target latents when it is useful to distinguish the target prior from context or state priors.
γ_C	Weight on the context-compression KL in full IB-VJEPA.
β_T	Weight on the target-side KL in full IB-VJEPA and full PIB-VJEPA.
$\mathcal{L}_{\text{IB-VJEPA}}$	Full IB-VJEPA objective with context–target NLL, context-compression KL, and target-side KL regularization.
$q_\theta(Z_C)$	Aggregated context posterior, $q_\theta(Z_C) = \int p_{\text{data}}(X_C) q_\theta(Z_C \mid X_C)\, dX_C$.
$D_{\text{KL}}(q_\theta(Z_C \mid X_C) \Vert p_0^C(Z_C))$	KL-to-prior compression penalty for the context latent. Its expectation upper-bounds $\text{I}(X_C; Z_C)$.
$\text{I}(X_C; Z_C)$	Context information retained by the compressed context latent. This is upper-bounded by the expected context KL term.
Full PIB-VJEPA notation	
$q_\theta(Z_t \mid X_{\le t})$	Stochastic current-state encoder introduced for full PIB-VJEPA.
$p_0^S(Z_t)$	Prior over the current latent state Z_t used in the PIB compression KL. The superscript S denotes state.
$p_0^T(Z_{t+\Delta})$	Prior over future target latents in the temporal target KL.
γ_S	Weight on the current-state compression KL in full PIB-VJEPA.
$\mathcal{L}_{\text{PIB-VJEPA}}$	Full PIB-VJEPA objective with temporal NLL, current-state compression KL, and target-side KL regularization.
$q_\theta(Z_t)$	Aggregated state posterior, $q_\theta(Z_t) = \int p_{\text{data}}(X_{\le t}) q_\theta(Z_t \mid X_{\le t})\, dX_{\le t}$.
$D_{\text{KL}}(q_\theta(Z_t \mid X_{\le t}) \Vert p_0^S(Z_t))$	KL-to-prior compression penalty for the current latent state. Its expectation upper-bounds $\text{I}(X_{\le t}; Z_t)$.
$\text{I}(X_{\le t}; Z_t)$	History-to-state information retained by the current latent state. This is upper-bounded by the expected state KL term.
Induced population distributions	
$q^\star(Z_C, \xi, Z_T)$	Induced context–target population distribution obtained by sampling $(X_C, X_T, \xi) \sim p_{\text{data}}$, encoding Z_C from X_C, and sampling $Z_T \sim q_{\bar\theta}(\cdot \mid X_T)$. In full IB-VJEPA, Z_C is sampled from $q_\theta(\cdot \mid X_C)$.
$q^\star(Z_T \mid Z_C, \xi)$	Conditional target-latent distribution induced by $q^\star(Z_C, \xi, Z_T)$; this is the ideal conditional approximated by $p_\phi(Z_T \mid Z_C, \xi)$.
$q^\star(Z_t, \xi, Z_{t+\Delta})$	Induced temporal population distribution obtained by sampling $(X_{\le t}, X_{t+\Delta}, \xi) \sim p_{\text{data}}$, encoding Z_t from $X_{\le t}$, and sampling $Z_{t+\Delta} \sim q_{\bar\theta}(\cdot \mid X_{t+\Delta})$. In full PIB-VJEPA, Z_t is sampled from $q_\theta(\cdot \mid X_{\le t})$.
$q^\star(Z_{t+\Delta} \mid Z_t, \xi)$	Conditional future-latent distribution induced by $q^\star(Z_t, \xi, Z_{t+\Delta})$; this is the ideal conditional approximated by $p_\phi(Z_{t+\Delta} \mid Z_t, \xi)$.
$q_{\bar\theta}(Z_T)$	Aggregated target posterior, $q_{\bar\theta}(Z_T) = \int p_{\text{data}}(X_T) q_{\bar\theta}(Z_T \mid X_T)\, dX_T$.
$q_{\bar\theta}(Z_{t+\Delta})$	Aggregated future-target posterior, $q_{\bar\theta}(Z_{t+\Delta}) = \int p_{\text{data}}(X_{t+\Delta}) q_{\bar\theta}(Z_{t+\Delta} \mid X_{t+\Delta})\, dX_{t+\Delta}$.
Information-theoretic quantities	

Symbol	Meaning
$H(X)$	Differential entropy of a continuous random variable X.
$H(Y \mid X)$	Conditional entropy of Y given X.
$H_{q^\star}(Z_T \mid Z_C, \xi)$	Conditional entropy of target latents under the induced context–target population distribution $q^\star(Z_C, \xi, Z_T)$.
$H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi)$	Conditional entropy of future latents under the induced temporal population distribution $q^\star(Z_t, \xi, Z_{t+\Delta})$.
$D_{\mathrm{KL}}(p\|q)$	Kullback–Leibler divergence from distribution p to distribution q.
$\mathrm{I}(X;Y)$	Mutual information between X and Y. For continuous variables, $\mathrm{I}(X;Y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)}\, dx\, dy$.
$\mathrm{I}(X;Y \mid Z)$	Conditional mutual information between X and Y given Z.
$\mathrm{I}_{q^\star}(Z_C; Z_T \mid \xi)$	Context–target predictive mutual information under the induced context–target population distribution.
$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi)$	Future-latent predictive mutual information under the induced temporal population distribution.
$p_\phi(y \mid x)$	Generic variational predictive distribution used in Barber–Agakov bounds.
Gaussian and prior notation	
$\mathcal{N}(\mu, \Sigma)$	Gaussian distribution with mean μ and covariance Σ.
$\mu_\phi(Z_C, \xi)$	Mean output of the context–target probabilistic predictor.
$\Sigma_\phi(Z_C, \xi)$	Covariance output of the context–target probabilistic predictor.
$\mu_\phi(Z_t, \xi)$	Mean output of the temporal probabilistic predictor.
$\Sigma_\phi(Z_t, \xi)$	Covariance output of the temporal probabilistic predictor.
$\mu_{\bar\theta}(X_T)$	Mean of the target distribution produced by the target encoder.
$\mu_{\bar\theta}(X_{t+\Delta})$	Mean of the future target distribution produced by the target encoder in temporal VJEPA.
$\sigma_q^2 I$	Isotropic covariance used in a simple Gaussian target distribution.
$\mu_\theta(X_C)$	Mean of the stochastic context encoder in full IB-VJEPA.
$\mathrm{diag}(\sigma_\theta^2(X_C))$	Diagonal covariance of the stochastic context encoder in full IB-VJEPA.
$\mu_\theta(X_{\le t})$	Mean of the stochastic current-state encoder in full PIB-VJEPA.
$\mathrm{diag}(\sigma_\theta^2(X_{\le t}))$	Diagonal covariance of the stochastic current-state encoder in full PIB-VJEPA.
Auxiliary notation	
Δ	Prediction horizon or temporal offset.
u_t	Action at time t, when the model is used in control or latent MPC.
$c(\cdot)$	Cost function used in planning/control discussions, if such a downstream planning objective is considered.
η	Generic learning rate or optimization step size, depending on context.
τ	EMA coefficient for updating the target encoder, $\bar\theta \leftarrow \tau\bar\theta + (1-\tau)\theta$.
$\mathbb{E}$	Expectation operator.
$\arg\min$, $\arg\max$	Operators denoting the argument that minimizes or maximizes an objective.

B Information quantities

For completeness[13], we review the standard information-theoretic quantities and identities used throughout the paper. Let X, Y, and Z be continuous random variables with a joint probability density $p(x, y, z)$.

[13] We include this review to make the paper self-contained.

B.1 Entropy and Conditional Entropy

The differential entropy of X measures the uncertainty or spread of the random variable and is defined as

$$H(X) = -\mathbb{E}_{p(x)}[\log p(x)]. \tag{46}$$

The conditional entropy of Y given X measures the expected remaining uncertainty in Y after X has been observed:

$$H(Y \mid X) = -\mathbb{E}_{p(x,y)}[\log p(y \mid x)]. \tag{47}$$

B.2 Kullback–Leibler Divergence

The Kullback–Leibler (KL) divergence measures the discrepancy incurred when a distribution $q(x)$ is used to approximate or represent a reference distribution $p(x)$:

$$D_{\mathrm{KL}}(p\|q) = \mathbb{E}_{p(x)}\left[\log \frac{p(x)}{q(x)}\right]. \tag{48}$$

By *Jensen's inequality*, the KL divergence is nonnegative, i.e. $D_{\mathrm{KL}}(p\|q) \geq 0$, with equality holding if and only if $p = q$ almost everywhere.

B.3 Mutual Information

Mutual information (MI) measures how much the observation of one variable reduces the uncertainty about another. It is defined as the KL divergence between the joint distribution and the product of their marginals:

$$\mathrm{I}(X;Y) = D_{\mathrm{KL}}(p(x,y)\|p(x)p(y)) = \mathbb{E}_{p(x,y)}\left[\log \frac{p(y \mid x)}{p(y)}\right]. \tag{49}$$

Mutual information can be decomposed into entropy terms:

$$\mathrm{I}(X;Y) = H(Y) - H(Y \mid X) = H(X) - H(X \mid Y). \tag{50}$$

B.4 Conditional Mutual Information and the Chain Rule

The conditional mutual information between X and Y given Z is the information shared by X and Y after conditioning on Z:

$$\mathrm{I}(X;Y \mid Z) = H(Y \mid Z) - H(Y \mid X, Z). \tag{51}$$

The chain rule for mutual information allows us to decompose the information shared between a joint variable (A, B) and a target C:

$$\mathrm{I}((A,B);C) = \mathrm{I}(A;C) + \mathrm{I}(B;C \mid A). \tag{52}$$

B.5 Variational Bounds

Exact conditional entropy is often intractable to compute when the true conditional distribution $p(y \mid x)$ is unknown. We typically introduce a parameterized variational predictive distribution $p_\phi(y \mid x)$.

Using the nonnegativity of the conditional KL divergence gives the entropy bound pointwise for each fixed x. For any x,

$$D_{\mathrm{KL}}\left(p(\cdot \mid x) \,\|\, p_\phi(\cdot \mid x)\right) = \mathbb{E}_{p(y|x)}\left[\log \frac{p(y \mid x)}{p_\phi(y \mid x)}\right] \geq 0.$$

Expanding the KL gives

$$\mathbb{E}_{p(y|x)}[\log p(y \mid x)] - \mathbb{E}_{p(y|x)}[\log p_\phi(y \mid x)] \geq 0.$$

Rearranging,

$$-\mathbb{E}_{p(y|x)}[\log p(y \mid x)] \leq -\mathbb{E}_{p(y|x)}[\log p_\phi(y \mid x)].$$

Since $H(Y \mid X = x) = -\mathbb{E}_{p(y|x)}[\log p(y \mid x)]$, we obtain

$$H(Y \mid X = x) \leq -\mathbb{E}_{p(y|x)}[\log p_\phi(y \mid x)].$$

Finally, averaging over $p(x)$ yields

$$H(Y \mid X) \leq -\mathbb{E}_{p(x,y)}[\log p_\phi(y \mid x)]. \tag{53}$$

Substituting equation 53 into the mutual information identity equation 50 yields the Barber–Agakov lower bound:

$$\mathrm{I}(X;Y) \geq H(Y) + \mathbb{E}_{p(x,y)}[\log p_\phi(y \mid x)]. \tag{54}$$

This bound connects predictive representation learning to information theory: *maximizing the expected latent predictive log-likelihood maximizes a variational lower bound on the mutual information between the context and the target.*

B.6 Cross-Entropy Decomposition

A basic identity used throughout the paper is that expected negative log-likelihood decomposes into an intrinsic conditional entropy term plus a conditional KL mismatch term.

Let (X, Y) have joint density $q^\star(x, y)$, with conditional density $q^\star(y \mid x)$. Let $p_\phi(y \mid x)$ be a variational predictive distribution. Then

$$\mathbb{E}_{q^\star(x,y)}\left[-\log p_\phi(y \mid x)\right] = H_{q^\star}(Y \mid X) + \mathbb{E}_{q^\star(x)} D_{\mathrm{KL}}\left(q^\star(y \mid x) \| p_\phi(y \mid x)\right). \tag{55}$$

To see this, add and subtract $\log q^\star(y \mid x)$:

$$\mathbb{E}_{q^\star(x,y)}[-\log p_\phi(y \mid x)] = \mathbb{E}_{q^\star(x,y)}[-\log q^\star(y \mid x)] \tag{56}$$

$$+ \mathbb{E}_{q^\star(x,y)}\left[\log \frac{q^\star(y \mid x)}{p_\phi(y \mid x)}\right]. \tag{57}$$

The first term is the conditional entropy:

$$\mathbb{E}_{q^\star(x,y)}[-\log q^\star(y \mid x)] = H_{q^\star}(Y \mid X).$$

For the second term, write the expectation explicitly:

$$\mathbb{E}_{q^\star(x,y)}\left[\log \frac{q^\star(y \mid x)}{p_\phi(y \mid x)}\right] \tag{58}$$

$$= \int q^\star(x) \left[\int q^\star(y \mid x) \log \frac{q^\star(y \mid x)}{p_\phi(y \mid x)}\, dy\right] dx \tag{59}$$

$$= \mathbb{E}_{q^\star(x)} D_{\mathrm{KL}}\left(q^\star(y \mid x) \| p_\phi(y \mid x)\right). \tag{60}$$

Therefore,

$$\mathbb{E}_{q^\star(x,y)}[-\log p_\phi(y \mid x)] \geq H_{q^\star}(Y \mid X),$$

with equality if and only if $p_\phi(y \mid x) = q^\star(y \mid x)$ for $q^\star(x)$-almost every x.

In the context–target VJEPA analysis, this identity is applied with $X = (Z_C, \xi)$, $Y = Z_T$, and $q^\star(y \mid x) = q^\star(Z_T \mid Z_C, \xi)$. In the temporal VJEPA analysis, it is applied with $X = (Z_t, \xi)$, $Y = Z_{t+\Delta}$, and $q^\star(y \mid x) = q^\star(Z_{t+\Delta} \mid Z_t, \xi)$.

B.7 General KL-to-Prior Decomposition

Another standard identity connects an expected KL-to-prior penalty to mutual information. Let X have density $p(x)$, and let Z be generated by a stochastic encoder $q(z \mid x)$. The induced joint density is

$$q(x, z) = p(x)q(z \mid x),$$

and the aggregated posterior is

$$q(z) = \int p(x)q(z \mid x)\, dx.$$

For any prior density $p_0(z)$ with compatible support,

$$\mathbb{E}_{p(x)} D_{\mathrm{KL}}(q(z \mid x) \| p_0(z)) = \mathrm{I}(X; Z) + D_{\mathrm{KL}}(q(z) \| p_0(z)). \tag{61}$$

Intuitively, the expected KL-to-prior penalty decomposes into two parts: the information Z retains about X, namely $\mathrm{I}(X; Z)$, and an additional prior-matching term that measures how far the aggregated latent distribution $q(z)$ is from the chosen prior $p_0(z)$.

Indeed, by definition,

$$\mathbb{E}_{p(x)} D_{\mathrm{KL}}(q(z \mid x) \| p_0(z)) = \iint p(x)q(z \mid x) \log \frac{q(z \mid x)}{p_0(z)}\, dx\, dz. \tag{62}$$

Insert the aggregated posterior $q(z)$ inside the logarithm:

$$\log \frac{q(z \mid x)}{p_0(z)} = \log \frac{q(z \mid x)}{q(z)} + \log \frac{q(z)}{p_0(z)}.$$

Therefore,

$$\mathbb{E}_{p(x)} D_{\mathrm{KL}}(q(z \mid x) \| p_0(z)) = \iint p(x)q(z \mid x) \log \frac{q(z \mid x)}{q(z)}\, dx\, dz \tag{63}$$

$$+ \iint p(x)q(z \mid x) \log \frac{q(z)}{p_0(z)}\, dx\, dz. \tag{64}$$

The first term is exactly

$$\mathrm{I}(X; Z) = \iint q(x, z) \log \frac{q(x, z)}{p(x)q(z)}\, dx\, dz = \iint p(x)q(z \mid x) \log \frac{q(z \mid x)}{q(z)}\, dx\, dz.$$

For the second term, the logarithm does not depend on x, so

$$\iint p(x)q(z \mid x) \log \frac{q(z)}{p_0(z)}\, dx\, dz \tag{65}$$

$$= \int \left[\int p(x)q(z \mid x)\, dx \right] \log \frac{q(z)}{p_0(z)}\, dz \tag{66}$$

$$= \int q(z) \log \frac{q(z)}{p_0(z)}\, dz = D_{\mathrm{KL}}(q(z) \| p_0(z)). \tag{67}$$

This proves equation 61. Since KL divergence is nonnegative,

$$\mathrm{I}(X; Z) \leq \mathbb{E}_{p(x)} D_{\mathrm{KL}}(q(z \mid x) \| p_0(z)).$$

In full IB-VJEPA, this identity is applied with (see Appendix C.1)

$$X = X_C, \qquad Z = Z_C, \qquad q(z \mid x) = q_\theta(Z_C \mid X_C), \qquad p_0(z) = p_0^C(Z_C).$$

Hence,

$$\mathbb{E}_{X_C\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_C \mid X_C)\|p_0^C(Z_C)\big) = \text{I}(X_C; Z_C) + D_{\text{KL}}\big(q_\theta(Z_C)\|p_0^C(Z_C)\big).$$

Therefore, the context KL term in full IB-VJEPA upper-bounds the IB context-compression term $\text{I}(X_C; Z_C)$.

In full PIB-VJEPA, this identity is applied with (see Appendix C.2)

$$X = X_{\le t} = X_{\le t}, \qquad Z = Z_t, \qquad q(z \mid x) = q_\theta(Z_t \mid X_{\le t}), \qquad p_0(z) = p_0^S(Z_t).$$

Hence,

$$\mathbb{E}_{X_{\le t}\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_t \mid X_{\le t})\|p_0^S(Z_t)\big) = \text{I}(X_{\le t}; Z_t) + D_{\text{KL}}\big(q_\theta(Z_t)\|p_0^S(Z_t)\big).$$

Therefore, the state KL term in full PIB-VJEPA upper-bounds the PIB state-compression term $\text{I}(X_{\le t}; Z_t)$.

C Additional derivations

C.1 IB-VJEPA KL-to-Prior Decomposition

This subsection specializes the general KL-to-prior identity from Appendix B.7 to the stochastic context encoder used in full IB-VJEPA. Specifically, we take

$$X = X_C, \qquad Z = Z_C, \qquad q(z \mid x) = q_\theta(Z_C \mid X_C), \qquad p_0(z) = p_0^C(Z_C).$$

This gives

$$\mathbb{E}_{X_C\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_C \mid X_C)\|p_0^C(Z_C)\big) = \text{I}(X_C; Z_C) + D_{\text{KL}}\big(q_\theta(Z_C)\|p_0^C(Z_C)\big), \tag{68}$$

where the aggregated context posterior is

$$q_\theta(Z_C) = \int p_{\text{data}}(X_C) q_\theta(Z_C \mid X_C)\, dX_C.$$

Since KL divergence is nonnegative, equation 68 implies

$$\text{I}(X_C; Z_C) \le \mathbb{E}_{X_C\sim p_{\text{data}}} D_{\text{KL}}\big(q_\theta(Z_C \mid X_C)\|p_0^C(Z_C)\big). \tag{69}$$

Thus, the context KL term in full IB-VJEPA is a variational upper bound on the IB context-compression term $\text{I}(X_C; Z_C)$.

C.2 PIB-VJEPA KL-to-Prior Decomposition

This subsection specializes the general KL-to-prior identity from Appendix B.7 to the stochastic current-state encoder used in full PIB-VJEPA. Specifically, we take

$$X = X_{\le t}, \qquad Z = Z_t, \qquad q(z \mid x) = q_\theta(Z_t \mid x_{\le t}), \qquad p_0(z) = p_0^S(Z_t).$$

This gives

$$\mathbb{E}_{p(x_{\le t})} D_{\text{KL}}\big(q_\theta(Z_t \mid x_{\le t})\|p_0^S(Z_t)\big) = \text{I}(X_{\le t}; Z_t) + D_{\text{KL}}\big(q_\theta(Z_t)\|p_0^S(Z_t)\big). \tag{70}$$

We provide the full derivation below for completeness.

Proof. By the definition of KL divergence,

$$\mathbb{E}_{p(x_{\le t})} D_{\text{KL}}\big(q_\theta(Z_t \mid x_{\le t})\|p_0^S(Z_t)\big) = \int p(x_{\le t}) \left[\int q_\theta(Z_t \mid x_{\le t}) \log \frac{q_\theta(Z_t \mid x_{\le t})}{p_0^S(Z_t)}\, dZ_t\right] dx_{\le t} \tag{71}$$

$$= \iint p(x_{\le t}) q_\theta(Z_t \mid x_{\le t}) \log \frac{q_\theta(Z_t \mid x_{\le t})}{p_0^S(Z_t)}\, dx_{\le t}\, dZ_t. \tag{72}$$

Now multiply and divide inside the logarithm by the aggregated posterior $q_\theta(Z_t)$:

$$\frac{q_\theta(Z_t \mid x_{\le t})}{p_0^S(Z_t)} = \frac{q_\theta(Z_t \mid x_{\le t})}{q_\theta(Z_t)} \cdot \frac{q_\theta(Z_t)}{p_0^S(Z_t)}.$$

Therefore,

$$\log \frac{q_\theta(Z_t \mid x_{\leq t})}{p_0^S(Z_t)} = \log \frac{q_\theta(Z_t \mid x_{\leq t})}{q_\theta(Z_t)} + \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)}.$$

Substituting this into equation 72 gives

$$\mathbb{E}_{p(x_{\leq t})} D_{\mathrm{KL}}\big(q_\theta(Z_t \mid x_{\leq t}) \| p_0^S(Z_t)\big) = \iint p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \log \frac{q_\theta(Z_t \mid x_{\leq t})}{q_\theta(Z_t)} \, dx_{\leq t} \, dZ_t \tag{73}$$

$$+ \iint p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)} \, dx_{\leq t} \, dZ_t. \tag{74}$$

We now identify the first term. By the definition of mutual information for jointly continuous random variables,

$$\mathrm{I}(X_{\leq t}; Z_t) = \iint q_\theta(x_{\leq t}, Z_t) \log \frac{q_\theta(x_{\leq t}, Z_t)}{p(x_{\leq t}) q_\theta(Z_t)} \, dx_{\leq t} \, dZ_t.$$

Since

$$q_\theta(x_{\leq t}, Z_t) = p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}),$$

we have

$$\frac{q_\theta(x_{\leq t}, Z_t)}{p(x_{\leq t}) q_\theta(Z_t)} = \frac{p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t})}{p(x_{\leq t}) q_\theta(Z_t)} = \frac{q_\theta(Z_t \mid x_{\leq t})}{q_\theta(Z_t)}.$$

Hence

$$\mathrm{I}(X_{\leq t}; Z_t) = \iint p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \log \frac{q_\theta(Z_t \mid x_{\leq t})}{q_\theta(Z_t)} \, dx_{\leq t} \, dZ_t.$$

Therefore, the first term in equation 73 is exactly $\mathrm{I}(X_{\leq t}; Z_t)$.

For the second term, note that

$$\log \frac{q_\theta(Z_t)}{p_0^S(Z_t)}$$

does not depend on $x_{\leq t}$. Thus

$$\iint p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)} \, dx_{\leq t} \, dZ_t \tag{75}$$

$$= \int \left[\int p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \, dx_{\leq t} \right] \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)} \, dZ_t. \tag{76}$$

By the definition of the aggregated posterior:

$$\int p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \, dx_{\leq t} = q_\theta(Z_t).$$

Therefore,

$$\iint p(x_{\leq t}) q_\theta(Z_t \mid x_{\leq t}) \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)} \, dx_{\leq t} \, dZ_t = \int q_\theta(Z_t) \log \frac{q_\theta(Z_t)}{p_0^S(Z_t)} \, dZ_t \tag{77}$$

$$= D_{\mathrm{KL}}\big(q_\theta(Z_t) \| p_0^S(Z_t)\big). \tag{78}$$

Combining the two identified terms gives

$$\mathbb{E}_{p(x_{\leq t})} D_{\mathrm{KL}}\big(q_\theta(Z_t \mid x_{\leq t}) \| p_0^S(Z_t)\big) = \mathrm{I}(X_{\leq t}; Z_t) + D_{\mathrm{KL}}\big(q_\theta(Z_t) \| p_0^S(Z_t)\big).$$

Finally, since KL divergence is nonnegative,

$$D_{\mathrm{KL}}\big(q_\theta(Z_t) \| p_0^S(Z_t)\big) \geq 0,$$

we obtain

$$\mathrm{I}(X_{\leq t}; Z_t) \leq \mathbb{E}_{p(x_{\leq t})} D_{\mathrm{KL}}\big(q_\theta(Z_t \mid x_{\leq t}) \| p_0^S(Z_t)\big).$$

□

C.3 Barber–Agakov bound with side information

For completeness, we derive equation 26. All information quantities below are computed under the induced population distribution

$$q^\star(Z_t, \xi, Z_{t+\Delta}) = q^\star(Z_t, \xi)\, q^\star(Z_{t+\Delta} \mid Z_t, \xi).$$

By the definition of conditional mutual information,

$$\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) = H_{q^\star}(Z_{t+\Delta} \mid \xi) - H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi). \tag{79}$$

The second entropy term is

$$\begin{aligned} H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) &= -\mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right] && (80)\\ &= -\mathbb{E}_{q^\star(Z_t,\xi)}\mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right]. && (81)\end{aligned}$$

For each fixed (Z_t, ξ), non-negativity of KL divergence gives

$$D_{\mathrm{KL}}\left(q^\star(Z_{t+\Delta} \mid Z_t, \xi) \| p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right) \geq 0. \tag{82}$$

Equivalently,

$$\mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log \frac{q^\star(Z_{t+\Delta} \mid Z_t, \xi)}{p_\phi(Z_{t+\Delta} \mid Z_t, \xi)}\right] \geq 0, \tag{83}$$

which implies

$$\mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right] \geq \mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. \tag{84}$$

Multiplying by -1 reverses the inequality:

$$-\mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log q^\star(Z_{t+\Delta} \mid Z_t, \xi)\right] \leq -\mathbb{E}_{q^\star(Z_{t+\Delta}\mid Z_t,\xi)}\left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. \tag{85}$$

Averaging over $q^\star(Z_t, \xi)$ yields

$$H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) \leq -\mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. \tag{86}$$

Substituting equation 86 into equation 79 gives

$$\begin{aligned} \mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) &= H_{q^\star}(Z_{t+\Delta} \mid \xi) - H_{q^\star}(Z_{t+\Delta} \mid Z_t, \xi) && (87)\\ &\geq H_{q^\star}(Z_{t+\Delta} \mid \xi) + \mathbb{E}_{q^\star(Z_t,\xi,Z_{t+\Delta})}\left[\log p_\phi(Z_{t+\Delta} \mid Z_t, \xi)\right]. && (88)\end{aligned}$$

Therefore,

$$\boxed{\mathrm{I}_{q^\star}(Z_t; Z_{t+\Delta} \mid \xi) \geq H_{q^\star}(Z_{t+\Delta} \mid \xi) + \mathbb{E}_{q^\star} \log p_\phi(Z_{t+\Delta} \mid Z_t, \xi).} \tag{89}$$

This is the Barber–Agakov lower bound with side information ξ.

C.4 Gaussian target KL

If

$$q_{\bar\theta}(Z_T \mid X_T) = \mathcal{N}(Z_T; \mu_{\bar\theta}(X_T), \sigma_q^2 I), \qquad p_0^T(Z_T) = \mathcal{N}(0, I), \tag{90}$$

then in dimension d,

$$D_{\mathrm{KL}}(q_{\bar\theta} \| p_0^T) = \frac{1}{2}\left[\|\mu_{\bar\theta}(X_T)\|_2^2 + d(\sigma_q^2 - 1 - \log \sigma_q^2)\right]. \tag{91}$$

If σ_q^2 is fixed, the variance-dependent term is constant and the KL mainly penalizes the target latent norm.

C.5 Context Gaussian KL

If

$$q_\theta(Z_C \mid X_C) = \mathcal{N}\big(Z_C; \mu_\theta(X_C), \mathrm{diag}(\sigma_\theta^2(X_C))\big)\,, \qquad p_0^C(Z_C) = \mathcal{N}(0, I), \tag{92}$$

then

$$D_{\mathrm{KL}}\big(q_\theta(Z_C \mid X_C) \| p_0^C(Z_C)\big) = \frac{1}{2}\sum_{j=1}^{d}\left[\mu_{\theta,j}^2(X_C) + \sigma_{\theta,j}^2(X_C) - 1 - \log \sigma_{\theta,j}^2(X_C)\right]. \tag{93}$$

This is the standard variational-bottleneck penalty for the compressed context.

C.6 Comparison with reconstruction objectives

Suppose the future observation decomposes as $X^+ = (S^+, N^+)$, where S^+ is predictive signal and N^+ is nuisance. A reconstruction objective approximates

$$\mathcal{L}_{\mathrm{rec}} \approx H(X^+ \mid Z) = H(X^+) - \mathrm{I}(X^+; Z). \tag{94}$$

By the chain rule,

$$\mathrm{I}(X^+; Z) = \mathrm{I}(S^+; Z) + \mathrm{I}(N^+; Z \mid S^+). \tag{95}$$

Thus reconstructing X^+ rewards encoding both signal and nuisance. In contrast, if the target encoder maps X^+ to $Z^+ \approx f(S^+)$, then the VJEPA prediction loss approximates Huang (2026)

$$H(Z^+ \mid Z) = H(Z^+) - \mathrm{I}(Z; Z^+) \approx H(f(S^+)) - \mathrm{I}(Z; S^+), \tag{96}$$

with no direct reward for encoding N^+.