A Family of Joint Embedding Predictive Architectures

From Deterministic Prediction to Probabilistic, Symmetric, Symbolic, and Joint-Density Models

Yongchao Huang (yongchao.huang@abdn.ac.uk)

22/03/2026

Contents

The evolution of JEPA: what each variant fixes

Classic JEPA moved self-supervised learning from pixel reconstruction to latent prediction:

$$z_c = e_\theta(x_c), \quad z_t^\star = e_{\bar{\theta}}(x_t), \qquad \hat{z}_t = g_\phi(z_c, \xi).$$

Limitations of classic JEPA

- **Deterministic:** returns a point prediction, not uncertainty.
- **One-way:** usually trains $x_c \to x_t$, leaving inverse structure unused.
- **Purely statistical:** can learn shortcuts when human/domain rules matter.
- **Architectural anti-collapse:** often relies on EMA, stop-gradient, or asymmetry, e.g. using a trainable online encoder for x_c and a slowly updated/frozen target encoder for x_t rather than two identical jointly updated branches.

Family-level response

- **VJEPA:** predictive distributions and variational latent dynamics.
- **BJEPA:** modular priors via Product-of-Experts belief fusion.
- **BiJEPA:** forward/backward consistency plus norm control.
- **RiJEPA:** symbolic rules as energy constraints.
- **GJE/GMJE:** joint density $p(z_c, z_t)$ and covariance/mixture geometry.

Notations

We use the following symbols in this talk; even when the individual papers use different ones.

- Observation pair: context x_c, target x_t.
- Encoders: online/context encoder e_θ, target encoder $e_{\bar{\theta}}$.
- Latents: $z_c = e_\theta(x_c)$ and $z_t \sim q_{\bar{\theta}}(\cdot \mid x_t)$.
- Structural or task side-information: ξ (mask pattern, temporal offset, action suffix, goal, rule descriptor, physics, control, etc.).
- Predictor: either deterministic $\hat{z}_t = g_\phi(z_c, \xi)$ or probabilistic $p_\phi(z_t \mid z_c, \xi)$.

Deterministic JEPA template

$$\mathcal{L}_{\text{JEPA}} = \|g_\phi(z_c, \xi) - e_{\bar{\theta}}(x_t)\|_2^2.$$

Probabilistic template

$$\mathcal{L} = \mathbb{E}_{z_t \sim q_{\bar{\theta}}(\cdot \mid x_t)}[-\log p_\phi(z_t \mid z_c, \xi)] + \text{regularization}.$$

Primary sources: VJEPA (Jan. 2026), BiJEPA (Feb. 2026), RiJEPA (Feb. 2026), GJE/GMJE (Mar. 2026), and PIB-VJEPA (upcoming).

What changes across the family?

Model	What is predicted?	Main mathematical move	Why it matters
JEPA	point latent	deterministic regression	semantic prediction without reconstruction
VJEPA	distribution over future latent	variational latent NLL + KL	uncertainty, predictive-state semantics, control
BiJEPA	two directional predictions	forward/backward cycle-consistency + norm control	symmetric structure, inverse reasoning
RiJEPA	latent + logic energy landscape	JEPA loss + energy-based rule constraints	zero-shot logical alignment, abductive reasoning
GJE	joint Gaussian $p(z_c, z_t)$	generative joint modeling	closed-form conditional inference, covariance geometry
GMJE	Gaussian mixture joint density	multimodal latent density	resolves conditional averaging, richer uncertainty

Baseline: classic JEPA as latent prediction

Architecture

$$z_c = e_\theta(x_c), \qquad z_t^\star = e_{\bar{\theta}}(x_t), \qquad \hat{z}_t = g_\phi(z_c, \xi).$$
$$\mathcal{L}_{\text{JEPA}} = \|\hat{z}_t - z_t^\star\|_2^2.$$

- Learns predictable structure rather than pixels.
- EMA target network helps avoid collapse.
- But the objective is still *deterministic*: it yields a point estimate, not a belief state.

Limitation of MSE

If z_t is genuinely multimodal given z_c, then the Bayes-optimal deterministic predictor is ▸ proof

$$\hat{z}_t^\star(z_c) = \mathbb{E}[z_t \mid z_c],$$

which can lie *between* valid modes instead of on any plausible mode.

Motivation

A world model for planning should carry *predictive uncertainty*, not only a mean.

VJEPA: probabilistic JEPA

Latent predictive model + target inference model

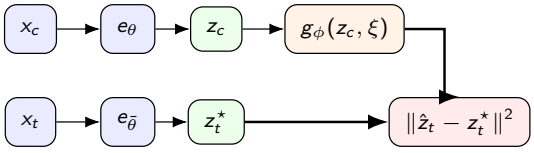
$$z_c = e_\theta(x_c), \qquad p_\phi(z_t \mid z_c, \xi), \qquad q_{\bar{\theta}}(z_t \mid x_t).$$

Variational JEPA objective

details | JEPA limit

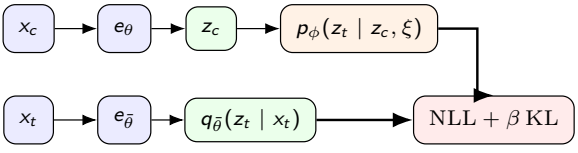
$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}_{z_t \sim q_{\bar{\theta}}(\cdot \mid x_t)}\big[-\log p_\phi(z_t \mid z_c, \xi)\big] + \beta\, \mathrm{KL}\Big(q_{\bar{\theta}}(z_t \mid x_t) \,\|\, p_0(z_t)\Big).$$

Classic JEPA: point prediction



$$\mathcal{L}_{\mathrm{JEPA}} = \|g_\phi(z_c, \xi) - e_{\bar{\theta}}(x_t)\|_2^2.$$

VJEPA: predictive distribution



$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}[-\log p_\phi(z_t \mid z_c, \xi)] + \beta \mathrm{KL}(q_{\bar{\theta}} \| p_0).$$

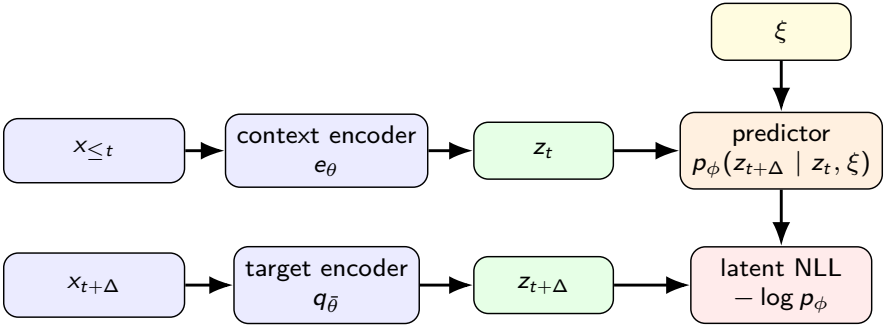
Main difference

Classic JEPA predicts a point target; VJEPA predicts a distribution and regularizes the target posterior.

Temporal VJEPA architecture

Our original VJEPA learns a predictive distribution over the future target latent:

$$q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}), \qquad p_{\phi}(z_{t+\Delta} \mid z_t, \xi).$$



$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}[-\log p_{\phi}(z_{t+\Delta} \mid z_t, \xi)] + \beta \mathrm{KL}\Big(q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta})\Big).$$

Key point

Original VJEPA targets future-latent prediction, but does not explicitly penalize $\mathrm{I}(x_{\leq t}; z_t)$. ▸ PIB comparison

The predictive model: $p_\phi(z_C, \varepsilon)$

In our implementation,

$$p_\phi(z_T \mid z_C, \xi)$$

was **not** a Bayesian neural network. It was a deterministic neural network that outputs parameters of a conditional latent distribution:

$$p_\phi(z_T \mid z_C, \xi) = \mathcal{N}\Big(z_T; \mu_\phi(z_C, \xi), \Sigma_\phi(z_C, \xi)\Big).$$

The parameters ϕ are point-estimated; uncertainty appears in the predicted distribution over z_T.

Comparison

Classic JEPA	$\hat{z}_T = g_\phi(z_C, \xi)$	point prediction
VJEPA implementation	$p_\phi(z_T \mid z_C, \xi) = \mathcal{N}(z_T; \mu_\phi, \Sigma_\phi)$	distributional prediction
BNN version, not used	$p(z_T \mid z_C, \xi, \mathcal{D}) = \int p_\phi(z_T \mid z_C, \xi) p(\phi \mid \mathcal{D})\, d\phi$	parameter uncertainty

Takeaway

VJEPA is a probabilistic neural predictor, but not a Bayesian neural network.

The target inference model: $q_{\bar{\theta}}$

In our VJEPA implementation, the target inference model

$$q_{\bar{\theta}}(z_T \mid x_T)$$

was chosen as a simple Gaussian distribution in latent space. Given a target input x_T, the target encoder outputs a latent mean:

$$\mu_{\bar{\theta}}(x_T) = e_{\bar{\theta}}(x_T).$$

We then define

$$\boxed{q_{\bar{\theta}}(z_T \mid x_T) = \mathcal{N}\big(z_T; \mu_{\bar{\theta}}(x_T), \sigma_q^2 I\big).}$$

Here $\sigma_q^2 I$ is an isotropic covariance matrix.

Interpretation

The target encoder does not output an arbitrary posterior. It defines a Gaussian cloud around the target embedding, with uncertainty controlled by the scalar variance σ_q^2.

Special case

As $\sigma_q^2 \to 0$, the target posterior approaches a Dirac delta: $q_{\bar{\theta}}(z_T \mid x_T) \to \delta(z_T - e_{\bar{\theta}}(x_T))$, recovering the deterministic JEPA target.

The KL term for isotropic Gaussian $q_{\bar{\theta}}$

Assume

$$q_{\bar{\theta}}(z_T \mid x_T) = \mathcal{N}\big(z_T; \mu_{\bar{\theta}}(x_T), \sigma_q^2 I\big),$$

and use the standard Gaussian prior

$$p_0(z_T) = \mathcal{N}(0, I).$$

For latent dimension d, the Gaussian KL is

$$\mathrm{KL}(q_{\bar{\theta}}(z_T \mid x_T) \| p_0(z_T)) = \frac{1}{2}\left[\mathrm{tr}(\sigma_q^2 I) + \|\mu_{\bar{\theta}}(x_T)\|_2^2 - d - \log|\sigma_q^2 I|\right].$$

Since

$$\mathrm{tr}(\sigma_q^2 I) = d\sigma_q^2, \qquad \log|\sigma_q^2 I| = d\log\sigma_q^2,$$

we get

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

If σ_q^2 is fixed

The variance-dependent part is constant, so the KL mainly penalizes the target latent norm:

$$\mathrm{KL}(q_{\bar{\theta}} \| p_0) = \frac{1}{2}\|\mu_{\bar{\theta}}(x_T)\|_2^2 + \mathrm{const}.$$

Trainable parameters and update rules in our VJEPA

The implementation contains three main parameter groups:

$$\theta, \qquad \bar{\theta}, \qquad \phi.$$

Component	Parameters	How updated
Context encoder	$e_\theta(x_C)$	Trainable by backpropagation through the VJEPA loss.
Predictive model	$p_\phi(z_T \mid z_C, \xi)$, e.g. $\mu_\phi(z_C, \xi), \Sigma_\phi(z_C, \xi)$	Trainable by backpropagation through the latent NLL term.
Target encoder	$e_{\bar{\theta}}(x_T)$ or $q_{\bar{\theta}}(z_T \mid x_T)$	Not directly optimized by ordinary gradient steps; updated as a slow-moving EMA copy of the online encoder.
Target variance	$\sigma_q^2 I$ in $q_{\bar{\theta}}(z_T \mid x_T)$	Fixed hyperparameter in the simple isotropic-Gaussian implementation, unless explicitly made learnable.
Prior and weights	$p_0(z_T)$, β	Fixed design choices / hyperparameters.

The online parameters are updated by gradient descent:

$$(\theta, \phi) \leftarrow (\theta, \phi) - \eta \nabla_{\theta,\phi} \mathcal{L}_{\mathrm{VJEPA}}.$$

The target encoder is updated by exponential moving average (EMA):

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

Why VJEPA is more than "JEPA + Gaussian noise"

1. **Predictive semantics become explicit.** The model learns $p_\phi(z_t \mid z_c, \xi)$ rather than only a regressor.
2. **Belief-state interpretation.** The latent becomes a distributional predictive state, making Bayesian filtering ▸ appendix and MPC ▸ appendix natural.
3. **Uncertainty propagation.** One can sample latent trajectories:

$$z_{t+k+1}^{(m)} \sim p_\phi\big(z_{t+k+1} \mid z_{t+k}^{(m)}, \xi_{t+k+1}, u_{t+k}\big).$$

4. **Multimodal futures are allowed.** In principle p_ϕ can be Gaussian, mixture-based, or flow-based.

Distributional planning objective ▸ mean/MAP case

$$\min_{u_{t:t+H-1}} \mathbb{E}\Big[\sum_{k=0}^{H-1} c\big(z_{t+k+1}, u_{t+k}\big) \;\Big|\; z_t, u_{t:t+H-1}\Big].$$

Secs. 4.6 and 6 of VJEPA.

BJEPA: Product-of-Experts priors on top of VJEPA

BJEPA asks: how can a VJEPA belief be steered by goals, safety, physics, or task constraints *without retraining the dynamics*?

Belief factorization

$$p_{\text{post}}(z_t \mid z_c, \eta) \propto p_{\text{dyn}}(z_t \mid z_c)\, p_{\text{prior}}(z_t \mid \eta),$$

where p_{dyn} is the learned VJEPA dynamics expert and p_{prior} is a modular structural, goal, or constraint expert.

- Inference becomes latent-space belief fusion rather than pixel reconstruction.
- Priors can encode safety, stationarity, goals, or known physics.
- This gives a route to zero-shot task adaptation by changing the prior expert.

Gaussian PoE intuition

▸ formula ▸ derivation

If both experts are Gaussian, precisions add:

$$\Sigma_{\text{post}}^{-1} = \Sigma_{\text{dyn}}^{-1} + \Sigma_{\text{prior}}^{-1}.$$

High-confidence experts contribute more information.

VJEPA/BJEPA Sec. 8; the Gaussian precision formula is expanded in the appendix.

Information Bottleneck (IB)

Given an input X, a representation Z, and a relevant target Y, the classical information bottleneck solves

▸ IB appendix ▸ PIB appendix

$$\min_{p(z|x)} \mathcal{L}_{\mathrm{IB}} = \min_{p(z|x)} \Big[\mathrm{I}(X;Z) - \beta\, \mathrm{I}(Z;Y) \Big].$$

- $\mathrm{I}(X;Z)$ penalizes **complexity**: how much of the input is retained.
- $\mathrm{I}(Z;Y)$ rewards **relevance**: how much information about the task or target survives.
- The optimum is a *minimal sufficient statistic*: retain what matters for Y, discard the unrelevant.

Predictive Information Bottleneck (PIB)

For temporal/self-supervised prediction, let $X = x_{\leq t}$ (past/history) and $Y = x_{t+\Delta}$ (future observation) or $Y = z_{t+\Delta}$ (future representation). Then PIB asks for a compressed summary of the past that is maximally predictive of the future.

Key intuition

Compression is not the goal by itself; it is a mechanism for removing nuisance variability that does not help predict the future.

VJEPA explicitly connects its objective to the predictive information bottleneck in Sec. 7.2.

From mutual information to the VJEPA objective

We prove a Barber–Agakov style lower bound ▸ proof in appendix:

$$\mathrm{I}(z_t; z_{t+\Delta}) \geq \mathbb{E}\Big[\log p_\phi(z_{t+\Delta} \mid z_t)\Big] + H(z_{t+\Delta}).$$

Hence minimizing latent NLL is equivalent to maximizing a lower bound on predictive mutual information.

Variational MI lower bound

$$\mathcal{L}_{\mathrm{VJEPA}} \approx H(z_{t+\Delta} \mid z_t) = H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).$$

If the future-latent marginal is prevented from becoming degenerate by target diversity, EMA/asymmetry, and suitably balanced KL regularization, then

$$\min \mathcal{L}_{\mathrm{VJEPA}} \iff \max \mathrm{I}(z_t; z_{t+\Delta}).$$

- This is the precise information-theoretic meaning of "predict the future in latent space".
- The optimization focuses on *predictive* information, not necessarily all information in the observation.

Theorem 4 and Sec. 7.1 of VJEPA.

Information bottleneck view: reconstruction objective

Assume observations decompose as

$$x_t = (S_t, N_t),$$

where S_t is predictive signal and N_t is nuisance variation.

Autoregressive / reconstruction objective

derivation

Here "AR" denotes an autoregressive generative objective that predicts or reconstructs the future observation $x_{t+\Delta}$:

$$\mathcal{L}_{\mathrm{AR}} \approx H(x_{t+\Delta} \mid z_t).$$

Using the mutual-information identity,

$$H(x_{t+\Delta} \mid z_t) = H(x_{t+\Delta}) - \mathrm{I}(x_{t+\Delta}; z_t).$$

Therefore,

$$\mathcal{L}_{\mathrm{AR}} \approx H(x_{t+\Delta}) - \mathrm{I}(x_{t+\Delta}; z_t).$$

Interpretation

Minimizing reconstruction uncertainty about $x_{t+\Delta}$ is equivalent to maximizing how much information z_t contains about the raw future observation.

Information bottleneck view: VJEPA objective

Since

$$x_{t+\Delta} = (S_{t+\Delta}, N_{t+\Delta}),$$

the chain rule gives ▸ chain-rule proof

$$\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

Thus a reconstruction objective can reward encoding both signal and nuisance.

VJEPA objective ▸ comparison

VJEPA predicts the future latent rather than the raw future observation:

$$\mathcal{L}_{\mathrm{VJEPA}} \approx H(z_{t+\Delta} \mid z_t) = H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).$$

If the target encoder removes nuisance,

$$z_{t+\Delta} \approx f_{\bar{\theta}}(S_{t+\Delta}),$$

then $\mathcal{L}_{\mathrm{VJEPA}} \approx -\mathrm{I}(z_t; S_{t+\Delta})$ up to marginal-entropy terms.

Bottom line

VJEPA behaves like a predictive information bottleneck because nuisance is filtered before the prediction loss is applied.

PIB: what VJEPA claims, and what it does not claim

Careful statement

VJEPA does *not* magically discard every nuisance variable. It biases learning toward information that is predictable in the target latent and useful for future latent dynamics.

Why generative world models can keep nuisance

▸ decomposition

Observation reconstruction rewards information about all predictable/reconstructable aspects of $x_{t+\Delta}$:

$$\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

If N is high-variance, a decoder may spend capacity on it.

Why VJEPA can filter nuisance

▸ comparison

The prediction loss is applied after the target encoder:

$$x_t \mapsto q_{\bar{\theta}}(z_t \mid x_t).$$

If $q_{\bar{\theta}}$ removes nuisance and preserves predictive structure, VJEPA maximizes information about future latent states rather than pixels.

Design implication

The bottleneck quality depends on target encoder design, β balancing, multi-horizon targets, and the expressivity of p_ϕ: the encoder decides what information enters the target latent; β controls how strongly latent information is regularized; multi-horizon prediction encourages temporally robust features by forcing z_t to predict not only the immediate next latent, but also longer-term future latents, so the representation must retain stable dynamics rather than short-lived noise; and an expressive p_ϕ is needed to represent uncertain or multi-modal futures rather than collapsing to a mean.

Why the bottleneck matters in practice I

The bottleneck is not only a theoretical idea. It changes what the representation is encouraged to store.

Efficiency

Allocate capacity to future-relevant structure, not irrelevant high-entropy nuisance variation such as sensor noise or background clutter.

A predictive bottleneck instead encourages the latent state to preserve information useful for predicting future structure:

z_t should retain future-relevant information, not everything in x_t.

Control

For planning, the agent does not need a photorealistic decoder. It needs a state that predicts action consequences:

$$p(z_{t+1} \mid z_t, u_t).$$

Thus, a good control representation should preserve information relevant to future costs and transitions, not necessarily all pixel-level details.

Why the bottleneck matters in practice II

Robustness

Nuisance variables can be high-variance but task-irrelevant:

$$x_t = (S_t, N_t),$$

where S_t is predictive signal and N_t is nuisance.
If the representation strongly encodes N_t, then small irrelevant changes in the observation can cause large changes in the latent state:

$$N_t \text{ changes} \implies z_t \text{ changes} \implies \text{unstable predictions.}$$

Filtering nuisance therefore improves robustness by making the latent dynamics depend more on stable predictive factors.

Uncertainty

Compression does not mean throwing away uncertainty.
VJEPA remains probabilistic because it predicts a distribution:

$$p_\phi(z_{t+\Delta} \mid z_t, \xi).$$

Why the bottleneck matters in practice III

Main design tension

The bottleneck must be neither too weak nor too strong.

$$\text{Too weak} \quad \Rightarrow \quad \text{noisy targets and nuisance leakage.}$$

$$\text{Too strong} \quad \Rightarrow \quad \text{loss of control-relevant detail.}$$

This creates a trade-off:

$$\text{target diversity} \quad \leftrightarrow \quad \text{regularization strength } \beta.$$

A larger β enforces stronger regularization, but if it is too large, useful predictive information may be removed.

Concrete improvements

- **State-dependent or multi-horizon bottlenecks:** different states and prediction horizons may need different amounts of information.
- **Mixture / flow heads for p_ϕ:** preserve multimodal futures instead of collapsing them to a single Gaussian mean.
- **Information-geometric regularization:** regularize covariance or precision in distribution space, not only latent means in Euclidean space.

Predictive sufficiency for control

VJEPA formalizes the latent state as an amortized predictive information state. Here "amortized" means that the encoder learns to perform inference in one forward pass. Given the history $h_t = (x_{1:t}, u_{1:t-1})$, the deterministic version directly computes $z_t = e_\theta(h_t)$, or, in a stochastic version, $z_t \sim q_\theta(z_t \mid h_t)$. Thus we avoid recomputing the posterior state by iterative filtering for each new history.

Predictive sufficiency ▸ proof

For any action sequence $u_{t:t+H-1}$,

$$p(z_{t+1:t+H} \mid h_t, u_{t:t+H-1}) = p(z_{t+1:t+H} \mid z_t, u_{t:t+H-1}),$$

with $h_t = (x_{1:t}, u_{1:t-1})$.

Control consequence ▸ proof

If future cost is measurable in latent space, then

$$p(C_{t:t+H-1} \mid h_t, u_{t:t+H-1}) = p(C_{t:t+H-1} \mid z_t, u_{t:t+H-1}),$$

so there exists an optimal policy of the form

$$\pi^\star(u_t \mid h_t) = \pi^\star(u_t \mid z_t).$$

Interpretation

The latent need not reconstruct observations; it must preserve information needed to predict future cost-relevant consequences of actions.

Collapse avoidance: statement

Theorem (informal) ▸ proof idea

Assume target diversity and context–target predictive dependence. That is, the target distributions

$$q_{\bar{\theta}}(z_t \mid x_t)$$

are not all identical, and the context contains information that can reduce uncertainty about z_t beyond the side information ξ.
If the predictive family $p_\phi(z_t \mid z_c, \xi)$ is expressive enough to exploit this dependence, then a globally optimal VJEPA solution cannot have

$$e_\theta(x_c) \equiv c$$

for all contexts.

Intuition

If $e_\theta(x_c)$ is constant, the predictor becomes unconditional: $p_\phi(z_t \mid z_c, \xi) = p_\phi(z_t \mid c, \xi)$. Such a predictor cannot use context to reduce target uncertainty. Hence collapse is suboptimal whenever context-dependent prediction achieves strictly lower NLL.

Collapse avoidance: proof idea I

Assume the context encoder collapses:

$$e_\theta(x_c) \equiv c.$$

Then every context has the same representation, so the predictor cannot use information from x_c:

$$p_\phi(z_t \mid z_c, \xi) = p_\phi(z_t \mid c, \xi).$$

Thus the best collapsed predictor is an unconditional predictor of the target latent, conditioned only on side information ξ:

$$p^\star(z_t \mid c, \xi) = q(z_t \mid \xi).$$

The minimum achievable collapsed NLL is therefore

$$\inf_{p_\phi} \mathbb{E}[-\log p_\phi(z_t \mid c, \xi)] = H_q(z_t \mid \xi).$$

Meaning

If $e_\theta(x_c)$ is constant, the model can only learn the average target-latent distribution given ξ, not context-specific target predictions.

Collapse avoidance: proof idea II

Now compare the collapsed predictor with an ideal context-dependent predictor.
If the context contains predictive information about the target latent, then: $\mathrm{I}_q(z_t; x_c \mid \xi) > 0$.
Using the conditional mutual-information identity,

$$\mathrm{I}_q(z_t; x_c \mid \xi) = H_q(z_t \mid \xi) - H_q(z_t \mid x_c, \xi).$$

Rearranging gives

$$H_q(z_t \mid \xi) = H_q(z_t \mid x_c, \xi) + \mathrm{I}_q(z_t; x_c \mid \xi).$$

Therefore, if $\mathrm{I}_q(z_t; x_c \mid \xi) > 0$, then

$$H_q(z_t \mid \xi) > H_q(z_t \mid x_c, \xi).$$

So unconditional collapsed prediction has a strict excess-loss gap:

$$\boxed{\text{excess loss} = \mathrm{I}_q(z_t; x_c \mid \xi).}$$

Conclusion

A collapsed context encoder cannot be globally optimal whenever the context contains information that can reduce uncertainty about the target latent.

Theorem 1 and Remark 1 in VJEPA. ▸ appendix version

BiJEPA: symmetric prediction and representation explosion

Forward + backward objectives

$$\hat{z}_t = g_\phi^{\rightarrow}(z_c), \qquad \hat{z}_c = g_\phi^{\leftarrow}(z_t),$$
$$\mathcal{L}_{\mathrm{BiJEPA}} = \alpha \|\hat{z}_t - z_t^\star\|_2^2 + (1-\alpha)\|\hat{z}_c - z_c^\star\|_2^2.$$

- Adds the inverse relation instead of only context $\rightarrow$ target.
- Improves semantic coherence and inverse reasoning.
- But introduces a new instability: **representation explosion**.

Stability mechanism

▸ feedback proof

BiJEPA identifies norm control as structurally necessary:

$$z \leftarrow \frac{f(x)}{\|f(x)\|_2} \quad \text{(hard constraint)} \qquad \text{or} \qquad \text{LayerNorm} + \text{weight decay} \quad \text{(soft constraint)}.$$

The paper argues that bidirectional feedback can behave like a linear loop $h_{k+1} \approx Wh_k$. If the spectral radius satisfies $\rho(W) > 1$, then some latent directions are amplified as $W^k h_0$, causing representation explosion unless norms are controlled.

Sec. 3.2–3.3 and Sec. 5.3 of BiJEPA.

BiJEPA: what the symmetric constraint buys

Semantic consistency

Forward prediction asks whether z_c contains enough information for z_t. Backward prediction asks whether z_t contains enough information for z_c:

$$\hat{z}_t = g_\phi^{\rightarrow}(z_c), \qquad \hat{z}_c = g_\phi^{\leftarrow}(z_t).$$

This discourages one-way shortcut features and encourages globally coherent latents.

Inference modes

- **Representation probing:** freeze e_θ and train a lightweight decoder/classifier.
- **Forward imputation:** use $g^{\rightarrow}$ for future/missing targets.
- **Backward inference:** use $g^{\leftarrow}$ for likely causes or missing contexts.

Empirical role

The BiJEPA experiments use synthetic periodic signals, Lorenz trajectories, and MNIST to demonstrate stable convergence, improved semantic consistency, and sharper left/right digit completion under norm-controlled training.

How BiJEPA relates to the bottleneck story

- BiJEPA is still deterministic, so it does not directly implement a variational information bottleneck.
- Nevertheless, the backward constraint acts as a *structural regularizer*: the representation must preserve enough information for both directions.
- In the paper's context, this combats *shortcut features*, i.e. superficial one-way cues that reduce loss without capturing the true context–target structure, and induces more globally coherent latents.

Practical comparison to VJEPA

Property	BiJEPA	VJEPA
Uncertainty model	none (point prediction)	explicit $p_\phi(z_t \mid z_c, \xi)$
Main regularizer	norm control / cycle consistency	KL + predictive conditioning
Best use-case	reversible structure, inverse maps	stochastic futures, planning, filtering
Potential fusion	bidirectional VJEPA with stochastic forward/backward heads	yes

RiJEPA: rule-informed latent geometry

Hybrid loss

$$\mathcal{L}_{\mathrm{RiJEPA}} = \mathcal{L}_{\mathrm{JEPA}} + \beta\, \mathcal{L}_{\mathrm{EBC}}.$$

For a rule antecedent A and consequent C, define energy

$$E(A, C) = \|g(f_c(A)) - f_t(C)\|_2^2.$$

Then

$$\mathcal{L}_{\mathrm{EBC}} = \sum_{(A,C)\in\mathcal{V}} E(A, C) + \lambda \sum_{(A,C^-)\in\mathcal{I}} \max\{0, m - E(A, C^-)\}.$$

- Valid rules become low-energy basins; invalid rules are repelled beyond a margin.
- Raw data encoders and rule encoders are bridged by a shared predictor into one semantic latent space.
- Enables zero-shot logical inference: compare a data-derived latent directly to symbolic poles or rule consequents.

Secs. 3.2 and 4 of RiJEPA.

RiJEPA: two advantages of rule-shaped latent geometry

RiJEPA uses symbolic rules to shape the latent energy landscape: $E(A, C) = \|g(f_c(A)) - f_t(C)\|_2^2$.

1. Zero-shot logical alignment

Valid rules are trained to become low-energy transitions:

$$A \Rightarrow C \quad \Longrightarrow \quad E(A, C) \text{ small.}$$

Thus, a new data point can be matched to symbolic consequents without requiring supervised examples for every possible query. The model can align neural latents with rule semantics at test time.

2. Abductive reasoning

RiJEPA can also reason backwards from a consequence to plausible causes. Given a target/consequent latent z_t, search for a context/antecedent latent: $z_c^\star = \arg\min_{z_c} E(z_c, z_t)$. This corresponds to inferring possible explanations or causes from an observed effect.

Takeaway

Rule constraints turn the latent space into a logic-aware energy landscape: valid symbolic transitions become attractors, while invalid transitions are repelled.

RiJEPA as a differentiable rule manifold

RiJEPA converts discrete rule search into continuous energy minimization. ▸ rule supervision

Joint energy model

Let $z = [z_c^\top, z_t^\top]^\top$.

$$p(z_c, z_t) = \frac{1}{Z} \exp(-E(z_c, z_t)), \qquad E(z_c, z_t) = \|g(z_c) - z_t\|_2^2.$$

Gradient-guided Langevin discovery

$$z^{(k+1)} = z^{(k)} - \eta \nabla_z E\left(z^{(k)}\right) + \sqrt{2\eta T}\, \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I).$$

Three modes follow immediately:

1. joint discovery: update both (z_c, z_t),
2. forward inference: fix z_c, update z_t,
3. abductive reasoning: fix z_t, update z_c.

Interpretation

RiJEPA equips JEPA with a *symbolically structured energy geometry*. This is different from VJEPA's probabilistic bottleneck, but complementary to it.

Secs. 4.1–4.2 of RiJEPA.

GJE: joint-density move

Core move

Instead of directly learning $p(z_t \mid z_c)$ through a black-box predictor, model the joint density

$$p(z_c, z_t).$$

By the chain rule,

$$\log p(z_c, z_t) = \log p(z_t \mid z_c) + \log p(z_c),$$

so joint modeling adds an explicit marginal geometry term missing from plain VJEPA.

Primal Gaussian joint NLL

For $z = [z_c^\top, z_t^\top]^\top \sim \mathcal{N}(\mu, \Sigma)$,

$$\mathcal{L}_{\mathrm{GJE}} = \frac{1}{2N} \sum_{i=1}^{N} (z_i - \mu)^\top \Sigma^{-1} (z_i - \mu) + \frac{1}{2} \log |\Sigma| + \mathrm{const}.$$

Careful claim

GJE can reduce reliance on EMA/stop-gradient by modeling covariance geometry; it does not automatically remove every anti-collapse requirement.

GJE: closed-form conditional inference

Assume a block Gaussian model

$$\begin{bmatrix} z_c \\ z_t \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} C_{cc} & C_{ct} \\ C_{tc} & C_{tt} \end{bmatrix}\right).$$

Then predictive inference is closed-form ▸ proof:

$$p(z_t \mid z_c) = \mathcal{N}(\mu_{t|c}, \Sigma_{t|c}),$$

$$\mu_{t|c} = C_{tc} C_{cc}^{-1} z_c, \qquad \Sigma_{t|c} = C_{tt} - C_{tc} C_{cc}^{-1} C_{ct}.$$

- The predictor is induced by covariance blocks rather than a separate deterministic head.
- Uncertainty appears as the Schur complement covariance

$$\Sigma_{t|c} = C_{tt} - C_{tc} C_{cc}^{-1} C_{ct}.$$

  This is the residual covariance of z_t after linearly conditioning on z_c: the term C_{tt} is the marginal target uncertainty, while $C_{tc} C_{cc}^{-1} C_{ct}$ is the part explained by the context.

Secs. 2–3 of GJE/GMJE.

GMJE: mixture joint density for multimodal futures

GJE is unimodal. GMJE extends the joint model to a mixture:

$$p(z_c, z_t) = \sum_{k=1}^{K} \pi_k \, \mathcal{N}\left(\begin{bmatrix} z_c \\ z_t \end{bmatrix} ; \mu_k, \Sigma_k \right).$$

Conditioning gives a mixture of conditional Gaussians:

$$p(z_t \mid z_c) = \sum_{k=1}^{K} \tilde{\pi}_k(z_c) \, \mathcal{N}(\mu_{k,t|c}(z_c), \Sigma_{k,t|c}).$$

GMJE provides a richer predictive family than JEPA by preserving multiple plausible futures rather than collapsing them to a conditional mean. ▸ why

Secs. 4.1–4.3 of GJE/GMJE.

The Mahalanobis Trace Trap

In primal GJE, naively optimizing the empirical Gaussian joint NLL causes an exact cancellation:

$$\frac{1}{2N}\sum_{i=1}^{N} z_i^\top C^{-1} z_i = \frac{1}{2}\,\mathrm{Tr}(C^{-1}C) = \frac{d}{2}.$$

Thus the data-fit term becomes a constant, leaving only the log-determinant term, which then drives collapse.

Interpretation

This is a concrete example of an information-geometric pitfall: if the geometry of the empirical covariance is handled incorrectly, the objective ceases to carry informative gradients about data alignment.

- The paper proposes prototype-based, MDN-based, topology-adaptive, and SMC-based GMJE remedies.
- For the JEPA family, the lesson is broad: probabilistic structure helps only if its geometry is optimized correctly.

Sec. 3.2 and Appendix O of GJE/GMJE.

Information geometry: the statistical manifold view

Once JEPA becomes probabilistic, the model family is no longer just a set of points in Euclidean parameter space; it is a manifold of distributions.

Local KL geometry

derivation

For nearby parameters ϑ and $\vartheta + d\vartheta$,

$$\mathrm{KL}\big(p_{\vartheta+d\vartheta} \,\|\, p_{\vartheta}\big) = \frac{1}{2} d\vartheta^{\top} G(\vartheta) d\vartheta + o(\|d\vartheta\|^2),$$

where $G(\vartheta)$ is the Fisher information matrix.

- $G(\vartheta)$ is the intrinsic metric of the family of predictive distributions.
- Natural gradient follows geodesics of this metric, not raw Euclidean directions. GD view
- In VJEPA/BJEPA/GMJE, the objects we manipulate are exactly Gaussian or mixture distributions over latents, so this geometry is directly relevant.

Where information geometry enters these models

1. **VJEPA.** The KL term $\mathrm{KL}(q_{\bar{\theta}}(z_t \mid x_t) \| p_0)$ compares two distributions on a statistical manifold; changing covariance/precision changes local geometry, not just scale.
2. **BJEPA.** Product-of-Experts fusion of Gaussians adds precisions:

$$\Sigma_{\text{post}}^{-1} = \Sigma_{\text{dyn}}^{-1} + \Sigma_{\text{aux}}^{-1},$$

   which is the natural exponential-family way of intersecting information from two sources. ▸ derivation
3. **GJE/GMJE.** Covariance is the central geometric object; mixture modeling gives a stratified manifold rather than a single ellipsoid.
4. **RiJEPA.** RiJEPA does not define a Fisher manifold in the usual probabilistic sense. Instead, its energy constraints reshape the latent space: valid symbolic transitions are pulled into low-energy basins, while invalid transitions are pushed into high-energy regions. In this sense, the latent geometry is *warped* to align with logical rules. ▸ details

Note

JEPA variants differ not only in their losses, but in the *geometry* they impose on latent predictive states: Euclidean regression, variational KL geometry, PoE precision geometry, and energy-based rule geometry.

Information-geometric improvements for VJEPA

- **Natural-gradient training** for Gaussian mean/covariance heads, especially if p_ϕ is parameterized in precision space. ▸ derivation
- **Riemannian trust regions** using local KL balls ▸ derivation:

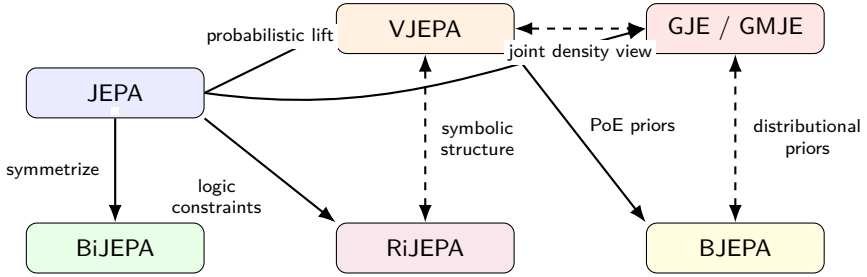
$$\min_{\Delta\vartheta} \nabla\mathcal{L}^\top \Delta\vartheta \quad \text{s.t.} \quad \Delta\vartheta^\top G(\vartheta)\Delta\vartheta \le \epsilon.$$

- **Geodesic interpolation** between latent beliefs during multi-step rollout instead of naive Euclidean interpolation.
- **Structured priors** on covariance or precision manifolds, e.g. low-rank + diagonal, temporal transport priors, or curvature penalties on uncertainty evolution.

High-level message

The next generation of JEPA world models may be limited less by network size than by how well we optimize and regularize the geometry of predictive distributions.

The family



Synthesis

All variants answer the same question — *what latent information should be preserved?* — but they answer it with different mathematics:

prediction / distribution / symmetry / logic / joint density.

Take-home messages

1. VJEPA turns JEPA into a **probabilistic predictive state model**: latent NLL + KL, explicit uncertainty, and control-relevant predictive sufficiency.
2. The **information bottleneck** story is central: VJEPA aims to maximize future-relevant mutual information while discarding nuisance variability before prediction loss is applied.
3. BiJEPA, RiJEPA, and GJE/GMJE each contribute a distinct axis of generalization: symmetry, symbolic structure, and generative joint density.
4. Information geometry provides a common language for the whole family: Fisher/KL geometry, PoE precision geometry, covariance geometry, and energy-shaped manifolds.
5. The most promising next step is likely a **hybrid**: VJEPA with mixture-based predictive heads, information-geometric training, and optional rule/goal priors.

Appendix: deterministic JEPA is a special case of VJEPA

Take

$$q_{\bar{\theta}}(z_t \mid x_t) = \delta(z_t - e_{\bar{\theta}}(x_t)), \qquad p_\phi(z_t \mid z_c, \xi) = \mathcal{N}\big(g_\phi(z_c, \xi), \sigma^2 I\big).$$

Then

$$\mathbb{E}_{z_t \sim q_{\bar{\theta}}}[-\log p_\phi(z_t \mid z_c, \xi)] = \frac{1}{2\sigma^2} \|g_\phi(z_c, \xi) - e_{\bar{\theta}}(x_t)\|_2^2 + \text{const}.$$

If $q_{\bar{\theta}}$ is deterministic, the KL term becomes either constant or removable by design, so VJEPA reduces to standard squared-loss JEPA.

back to VJEPA

Appendix: MI lower bound derivation

Start from the mutual-information identity:

$$\mathrm{I}(z_t; z_{t+\Delta}) = H(z_{t+\Delta}) - H(z_{t+\Delta} \mid z_t).$$

The conditional entropy is

$$H(z_{t+\Delta} \mid z_t) = -\mathbb{E} \log p(z_{t+\Delta} \mid z_t).$$

Therefore,

$$\mathrm{I}(z_t; z_{t+\Delta}) = H(z_{t+\Delta}) + \mathbb{E} \log p(z_{t+\Delta} \mid z_t).$$

Introduce a variational predictive model

$$p_\phi(z_{t+\Delta} \mid z_t).$$

By non-negativity of conditional KL,

$$\mathrm{KL}\big(p(\cdot \mid z_t) \| p_\phi(\cdot \mid z_t)\big) \geq 0.$$

Hence

$$\mathbb{E} \log p(z_{t+\Delta} \mid z_t) \geq \mathbb{E} \log p_\phi(z_{t+\Delta} \mid z_t).$$

Thus,

$$\boxed{\mathrm{I}(z_t; z_{t+\Delta}) \geq H(z_{t+\Delta}) + \mathbb{E} \log p_\phi(z_{t+\Delta} \mid z_t).}$$

This is the Barber–Agakov variational lower bound used in the VJEPA information-theoretic interpretation.

back

Appendix: predictive bottleneck picture in one equation

If observations split into signal and nuisance, $x_{t+\Delta} = (S_{t+\Delta}, N_{t+\Delta})$, then:

$$\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

For a pixel-level generative model,

$$\min H(x_{t+\Delta} \mid z_t) \iff \max \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

For VJEPA, if the target encoder discards nuisance so that $z_{t+\Delta} \approx f_{\bar{\theta}}(S_{t+\Delta})$, then substitution directly yields:

$$\min H(z_{t+\Delta} \mid z_t) \iff \max \mathrm{I}(z_t; S_{t+\Delta}),$$

without rewarding nuisance capture.

◂ back to VJEPA bottleneck ◂ back to PIB claims

Appendix: proof skeleton for no collapsed optimum

Assume collapse: $e_\theta(x_c) \equiv c$. Then the predictor ignores context and becomes unconditional, hence best achievable NLL is the cross-entropy with the aggregated target distribution:

$$\inf_p \mathbb{E}_{x_t|\xi}\mathbb{E}_{z_t\sim q(\cdot|x_t)}[-\log p(z_t)] = H_q(z_t \mid \xi).$$

Use the chain rule:

$$H_q(z_t \mid \xi) = H_q(z_t \mid x_t, \xi) + \mathrm{I}_q(z_t; x_t \mid \xi).$$

If target diversity holds, then for some ξ,

$$\mathrm{I}_q(z_t; x_t \mid \xi) > 0,$$

so unconditional prediction has a strict excess-loss gap versus a non-collapsed context-dependent predictor.

back

Appendix: when is mean/MAP planning optimal?

If

$$z_{t+1} \mid z_t, u_t \sim \mathcal{N}\big(\mu_\phi(z_t, u_t), \Sigma\big),$$

with action-independent Σ, and cost is quadratic,

$$c(z_{t+1}, u_t) = (z_{t+1} - z^\star)^\top Q(z_{t+1} - z^\star) + r(u_t),$$

then

$$\mathbb{E}[c(z_{t+1}, u_t) \mid z_t, u_t] = (\mu_\phi - z^\star)^\top Q(\mu_\phi - z^\star) + \mathrm{tr}(Q\Sigma) + r(u_t).$$

The trace term is action-independent, so

$$\arg\min_{u_t} \mathbb{E}[c(\cdot)] = \arg\min_{u_t} c(\mu_\phi(z_t, u_t), u_t).$$

Thus mean planning equals MAP planning in this regime.

back

Appendix: BJEPA in formulas

BJEPA fuses a learned dynamics expert and an auxiliary prior expert:

$$p(z_t \mid z_c, \eta) \propto p_{\text{like}}(z_t \mid z_c)\, p_{\text{prior}}(z_t \mid \eta).$$

If both are Gaussian,

$$p_{\text{like}} = \mathcal{N}(\mu_{\text{dyn}}, \Sigma_{\text{dyn}}), \qquad p_{\text{prior}} = \mathcal{N}(\mu_{\text{aux}}, \Sigma_{\text{aux}}),$$

then the posterior is Gaussian with

$$\Sigma_{\text{post}}^{-1} = \Sigma_{\text{dyn}}^{-1} + \Sigma_{\text{aux}}^{-1},$$
$$\mu_{\text{post}} = \Sigma_{\text{post}}\left(\Sigma_{\text{dyn}}^{-1}\mu_{\text{dyn}} + \Sigma_{\text{aux}}^{-1}\mu_{\text{aux}}\right).$$

VJEPA is the special case with an uninformative prior $p_{\text{prior}} \propto 1$.

◂ back to BJEPA ◂ back to geometry

Appendix: BiJEPA instability as a feedback system

Informally, the forward/backward predictor-encoder loop behaves like a coupled linear update

$$h_{k+1} \approx W h_k.$$

If the spectral radius $\rho(W) > 1$, feature norms amplify over iterations/training steps.

Why the issue is worse than in one-way JEPA

Forward perturbations amplify the target, and the backward path feeds the amplified signal back again. Hence the two-way coupling can create an expanding loop unless norms are controlled.

Soft normalization (LayerNorm + decay) preserves expressive magnitude better than hard projection to the unit sphere, while still keeping $\rho(W)$ effectively under control.

back

Appendix: RiJEPA versus ordinary predictive supervision

Ordinary JEPA supervision only says:

"predict the paired target."

RiJEPA adds:

"make valid symbolic transitions low-energy and invalid ones high-energy."

So its latent space is not merely predictive; it is also *normatively shaped*.

Practical consequence

A downstream classifier is no longer strictly necessary for some logical queries. One can classify by geometric proximity to symbolic anchors in latent space.

◂ back to RiJEPA ◂ back to geometry

Appendix: GJE conditional inference from a joint Gaussian

If

$$\begin{bmatrix} z_c \\ z_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mu_c \\ \mu_t \end{bmatrix}, \begin{bmatrix} \Sigma_{cc} & \Sigma_{ct} \\ \Sigma_{tc} & \Sigma_{tt} \end{bmatrix} \right),$$

then

$$z_t \mid z_c \sim \mathcal{N}\big(\mu_t + \Sigma_{tc}\Sigma_{cc}^{-1}(z_c - \mu_c), \Sigma_{tt} - \Sigma_{tc}\Sigma_{cc}^{-1}\Sigma_{ct}\big)\,.$$

Thus predictive inference is closed-form and uncertainty-aware, with no separate neural predictor required.

back

Appendix: why GMJE fixes the conditional-mean problem

For a deterministic predictor trained by MSE,

$$\hat{z}_t^{\star}(z_c) = \mathbb{E}[z_t \mid z_c].$$

If the conditional target law has multiple modes, the conditional mean may fall in a low-density region between them.
GMJE instead uses

$$p(z_t \mid z_c) = \sum_{k=1}^{K} \tilde{\pi}_k(z_c)\, \mathcal{N}\big(z_t; \mu_{k|c}(z_c), \Sigma_{k|c}\big),$$

so the model can preserve multiple plausible futures rather than average them away.

back

Appendix: future hybrid model

A natural next synthesis is:

Bi-directional, mixture-based, rule-informed VJEPA with PoE priors

with ingredients:

- VJEPA variational predictive-state objective,
- GMJE-style multimodal heads for $p_\phi(z_t \mid z_c, \xi)$,
- BiJEPA forward/backward consistency,
- RiJEPA energy constraints for symbolic/task structure,
- BJEPA PoE priors for goals and safety constraints,
- information-geometric optimization for covariance/precision parameters.

This would unify uncertainty, symmetry, logic, and multimodality in one predictive-state architecture.

Appendix: entropy

Let X be a random variable with mass function or density $p(x)$.

For a discrete random variable,

$$H(X) = -\sum_x p(x) \log p(x).$$

For a continuous random variable,

$$H(X) = -\int p(x) \log p(x)\, dx = -\mathbb{E}_{X\sim p}\big[\log p(X)\big].$$

In the latent-variable setting, $H(\cdot)$ should be read as differential entropy when the latent variables are continuous.

Interpretation

Entropy measures the uncertainty, spread, or average code length associated with a random variable.

Appendix: conditional entropy

The conditional entropy of Y given X is

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

Equivalently,

$$H(Y \mid X) = \mathbb{E}_{p(x)}\big[H(Y \mid X = x)\big].$$

Expanding the expectation,

$$H(Y \mid X) = -\int p(x,y) \log p(y \mid x)\, dx\, dy.$$

Interpretation

$H(Y)$ is the uncertainty about Y before observing X, while $H(Y \mid X)$ is the remaining uncertainty after observing X.

Appendix: KL divergence

For two distributions p and q over the same variable,

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

Equivalently,

$$\mathrm{KL}(p\|q) = \int p(x) \log \frac{p(x)}{q(x)}\, dx.$$

Interpretation

$\mathrm{KL}(p\|q)$ measures the information loss incurred when q is used to approximate p.

It is asymmetric:

$$\mathrm{KL}(p\|q) \neq \mathrm{KL}(q\|p)$$

in general.

Appendix: non-negativity of KL

Using the convexity of $-\log(\cdot)$,

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[-\log\frac{q(X)}{p(X)}\right].$$

By Jensen's inequality,

$$\mathbb{E}_p\left[-\log\frac{q(X)}{p(X)}\right] \geq -\log\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right].$$

But

$$\mathbb{E}_p\left[\frac{q(X)}{p(X)}\right] = \int p(x)\frac{q(x)}{p(x)}\,dx = \int q(x)\,dx = 1.$$

Therefore,

$$\mathrm{KL}(p\|q) \geq -\log 1 = 0.$$

Hence

$$\boxed{\mathrm{KL}(p\|q) \geq 0.}$$

Equality holds if and only if $p = q$ almost everywhere.

Appendix: mutual information definition

The mutual information between X and Y is

$$\mathrm{I}(X;Y) = \mathrm{KL}\big(p(x,y) \,\|\, p(x)p(y)\big).$$

Therefore,

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

Using

$$p(x,y) = p(y \mid x)p(x),$$

we obtain

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

Interpretation

Mutual information measures how much knowing X reduces uncertainty about Y.

Appendix: mutual information and entropy

Starting from

$$\mathrm{I}(X;Y) = \mathbb{E}_{p(x,y)}\left[\log p(y \mid x) - \log p(y)\right],$$

we split the expectation:

$$\mathrm{I}(X;Y) = \mathbb{E}_{p(x,y)}[\log p(y \mid x)] - \mathbb{E}_{p(y)}[\log p(y)].$$

Using

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

and

$$H(Y) = -\mathbb{E}_{p(y)}[\log p(y)],$$

we obtain

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

By symmetry,

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

Appendix: conditional mutual information

The conditional mutual information between X and Y given Z is

$$\mathrm{I}(X; Y \mid Z) = \mathbb{E}_{p(x,y,z)} \left[\log \frac{p(x, y \mid z)}{p(x \mid z) p(y \mid z)} \right].$$

Using

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

we get

$$\mathrm{I}(X; Y \mid Z) = \mathbb{E} \left[\log \frac{p(y \mid x, z)}{p(y \mid z)} \right].$$

Therefore,

$$\mathrm{I}(X; Y \mid Z) = \mathbb{E}[\log p(Y \mid X, Z)] - \mathbb{E}[\log p(Y \mid Z)].$$

Appendix: conditional MI and entropy

Recall

$$H(Y \mid Z) = -\mathbb{E}[\log p(Y \mid Z)]$$

and

$$H(Y \mid X, Z) = -\mathbb{E}[\log p(Y \mid X, Z)].$$

Substituting these into the previous expression gives

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

Equivalently,

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

Interpretation

$\mathrm{I}(X; Y \mid Z)$ measures the information shared by X and Y after the information already explained by Z has been removed.

Appendix: chain rule for mutual information

For three variables A, B, C, the chain rule is

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

To derive it, start from

$$\mathrm{I}((A, B); C) = H(A, B) - H(A, B \mid C).$$

Use entropy chain rules:

$$H(A, B) = H(A) + H(B \mid A),$$
$$H(A, B \mid C) = H(A \mid C) + H(B \mid A, C).$$

Hence

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

Therefore,

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

Appendix: information bottleneck

Let X be the input, Z the representation, and Y the target.
The classical information bottleneck objective is

$$\boxed{\mathcal{L}_{\mathrm{IB}} = \mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y).}$$

The term

$$\mathrm{I}(X;Z)$$

penalizes how much information Z retains about X.
The term

$$\mathrm{I}(Z;Y)$$

rewards how much information Z preserves about the target Y.

Thus minimizing

$$\mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y)$$

encourages a representation that is compressed with respect to X but predictive of Y.

back

Appendix: predictive information bottleneck

For temporal prediction, let

$$X = x_{\leq t}, \qquad Y = x_{t+\Delta}, \qquad Z = z_t.$$

Then the predictive information bottleneck objective is

$$\mathcal{L}_{\mathrm{PIB}} = \mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; x_{t+\Delta}).$$

If the future target is represented in latent space,

$$Y = z_{t+\Delta},$$

then

$$\boxed{\mathcal{L}_{\mathrm{PIB}} = \mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; z_{t+\Delta}).}$$

Interpretation

z_t should be a compressed summary of the past that keeps information useful for predicting the future.

back

Appendix: KL as an information bottleneck I

Assume a stochastic encoder

$$z \sim q_\theta(z \mid x),$$

with data distribution $p(x)$.
The joint distribution induced by the encoder is

$$q_\theta(x, z) = p(x) q_\theta(z \mid x).$$

The aggregated posterior is the marginal distribution of Z:

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

By definition, the mutual information between X and Z is

$$\mathrm{I}(X; Z) = \mathrm{KL}\big(q_\theta(x, z) \,\|\, p(x) q_\theta(z)\big).$$

Expanding the KL divergence,

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

Canceling $p(x)$ gives

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

Appendix: KL as an information bottleneck II

From the previous slide,

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

Equivalently,

$$\boxed{\mathrm{I}(X;Z) = \mathbb{E}_{p(x)} \left[\mathrm{KL}\big(q_\theta(z \mid x) \| q_\theta(z)\big) \right].}$$

In practice, the aggregated posterior

$$q_\theta(z) = \int p(x) q_\theta(z \mid x) \, dx$$

is usually difficult to compute.
So variational objectives often regularize against a simple prior $p_0(z)$:

$$\mathbb{E}_{p(x)} \left[\mathrm{KL}\big(q_\theta(z \mid x) \| p_0(z)\big) \right].$$

Now expand this regularizer:

$$\mathbb{E}_{p(x)} \mathrm{KL}\big(q_\theta(z \mid x) \| p_0(z)\big) = \int p(x) q_\theta(z \mid x) \log \frac{q_\theta(z \mid x)}{p_0(z)} \, dz \, dx.$$

Insert the aggregated posterior $q_\theta(z)$:

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

Appendix: KL as an information bottleneck III

Using

$$\log \frac{q_\theta(z \mid x)}{p_0(z)} = \log \frac{q_\theta(z \mid x)}{q_\theta(z)} + \log \frac{q_\theta(z)}{p_0(z)},$$

we have

$$\mathbb{E}_{p(x)} \mathrm{KL}\big(q_\theta(z \mid x) \| p_0(z)\big) = A + B,$$

where

$$A = \int p(x) q_\theta(z \mid x) \log \frac{q_\theta(z \mid x)}{q_\theta(z)} \, dz \, dx = \mathrm{I}(X; Z),$$

and

$$B = \int p(x) q_\theta(z \mid x) \log \frac{q_\theta(z)}{p_0(z)} \, dz \, dx.$$

Since

$$\int p(x) q_\theta(z \mid x) \, dx = q_\theta(z),$$

we get

$$B = \int q_\theta(z) \log \frac{q_\theta(z)}{p_0(z)} \, dz = \mathrm{KL}(q_\theta(z) \| p_0(z)).$$

Therefore,

$$\boxed{\mathbb{E}_{p(x)} \mathrm{KL}\big(q_\theta(z \mid x) \| p_0(z)\big) = \mathrm{I}(X; Z) + \mathrm{KL}(q_\theta(z) \| p_0(z)).}$$

Appendix: VJEPA objective

VJEPA uses a target inference model and a predictive model:

$$q_{\bar{\theta}}(z_t \mid x_t), \qquad p_\phi(z_t \mid z_c, \xi).$$

Its objective is

$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}_{z_t \sim q_{\bar{\theta}}(\cdot \mid x_t)} \left[-\log p_\phi(z_t \mid z_c, \xi) \right] + \beta \mathrm{KL}\Big(q_{\bar{\theta}}(z_t \mid x_t) \| p_0(z_t) \Big).$$

The first term is a latent-space negative log-likelihood:

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}[-\log p_\phi(z_t \mid z_c, \xi)].$$

The second term regularizes the target posterior against a prior.

back

Appendix: VJEPA NLL and conditional entropy

Let the true conditional distribution induced by the data and target encoder be

$$p^\star(z_t \mid z_c, \xi).$$

Then

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{p^\star}\left[-\log p_\phi(z_t \mid z_c, \xi)\right].$$

Add and subtract $\log p^\star(z_t \mid z_c, \xi)$:

$$\mathcal{L}_{\text{pred}} = \mathbb{E}_{p^\star}\left[-\log p^\star(z_t \mid z_c, \xi)\right] + \mathbb{E}_{p^\star}\left[\log \frac{p^\star(z_t \mid z_c, \xi)}{p_\phi(z_t \mid z_c, \xi)}\right].$$

Therefore,

$$\mathcal{L}_{\text{pred}} = H_{p^\star}(z_t \mid z_c, \xi) + \mathbb{E}_{p^\star(z_c,\xi)}\mathrm{KL}\Big(p^\star(z_t \mid z_c, \xi) \| p_\phi(z_t \mid z_c, \xi)\Big).$$

Since KL is nonnegative,

$$\mathcal{L}_{\text{pred}} \geq H_{p^\star}(z_t \mid z_c, \xi).$$

If $p_\phi = p^\star$, then

$$\boxed{\mathcal{L}_{\text{pred}} = H(z_t \mid z_c, \xi).}$$

back

Appendix: deterministic JEPA as a VJEPA limit

Let the target posterior be deterministic:

$$q_{\bar{\theta}}(z_t \mid x_t) = \delta\Big(z_t - e_{\bar{\theta}}(x_t)\Big).$$

Let the predictive model be isotropic Gaussian:

$$p_\phi(z_t \mid z_c, \xi) = \mathcal{N}\Big(z_t; g_\phi(z_c, \xi), \sigma^2 I\Big).$$

Then

$$-\log p_\phi(z_t \mid z_c, \xi) = \frac{d}{2}\log(2\pi\sigma^2) + \frac{1}{2\sigma^2}\,\|z_t - g_\phi(z_c, \xi)\|_2^2\,.$$

Taking expectation under the Dirac posterior gives

$$\mathbb{E}_{z_t \sim q_{\bar{\theta}}}[-\log p_\phi(z_t \mid z_c, \xi)] = \frac{1}{2\sigma^2}\,\|e_{\bar{\theta}}(x_t) - g_\phi(z_c, \xi)\|_2^2 + \text{const}.$$

Therefore,

$$\boxed{\mathcal{L}_{\text{VJEPA}} \to \|g_\phi(z_c, \xi) - e_{\bar{\theta}}(x_t)\|_2^2\,.}$$

back

Appendix: Barber–Agakov MI lower bound I

Let

$$z_t = z_t, \qquad z_{t+\Delta} = z_{t+\Delta}.$$

By definition,

$$\mathrm{I}(z_t; z_{t+\Delta}) = H(z_{t+\Delta}) - H(z_{t+\Delta} \mid z_t).$$

Since

$$H(z_{t+\Delta} \mid z_t) = -\mathbb{E}_{p(z_t, z_{t+\Delta})}\big[\log p(z_{t+\Delta} \mid z_t)\big],$$

we have

$$\mathrm{I}(z_t; z_{t+\Delta}) = H(z_{t+\Delta}) + \mathbb{E}_{p(z_t, z_{t+\Delta})}\big[\log p(z_{t+\Delta} \mid z_t)\big].$$

Now introduce a variational predictive model

$$p_\phi(z_{t+\Delta} \mid z_t).$$

back

Appendix: Barber–Agakov MI lower bound II

By non-negativity of conditional KL,

$$\mathrm{KL}\big(p(z_{t+\Delta} \mid z_t) \| p_\phi(z_{t+\Delta} \mid z_t)\big) \geq 0.$$

Therefore,

$$\mathbb{E}_{p(z_{t+\Delta}|z_t)}[\log p(z_{t+\Delta} \mid z_t)] \geq \mathbb{E}_{p(z_{t+\Delta}|z_t)}[\log p_\phi(z_{t+\Delta} \mid z_t)].$$

Taking expectation over z_t,

$$\boxed{\mathrm{I}(z_t; z_{t+\Delta}) \geq H(z_{t+\Delta}) + \mathbb{E}_{p(z_t, z_{t+\Delta})}[\log p_\phi(z_{t+\Delta} \mid z_t)].}$$

Equivalently,

$$\boxed{-\mathbb{E} \log p_\phi(z_{t+\Delta} \mid z_t) \geq H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).}$$

Thus minimizing latent NLL maximizes a lower bound on predictive mutual information.

back

Appendix: latent NLL and predictive MI

The identity

$$\mathrm{I}(z_t; z_{t+\Delta}) = H(z_{t+\Delta}) - H(z_{t+\Delta} \mid z_t)$$

implies

$$\boxed{H(z_{t+\Delta} \mid z_t) = H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).}$$

If the predictive model is well specified, then

$$\mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t)] \approx H(z_{t+\Delta} \mid z_t).$$

Therefore,

$$\boxed{\mathcal{L}_{\mathrm{VJEPA}} \approx H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).}$$

If $H(z_{t+\Delta})$ is prevented from degenerating, then minimizing the latent predictive NLL approximately maximizes

$$\mathrm{I}(z_t; z_{t+\Delta}).$$

back

Appendix: signal–nuisance decomposition

Assume the future observation decomposes as

$$x_{t+\Delta} = (S_{t+\Delta}, N_{t+\Delta}),$$

where $S_{t+\Delta}$ is predictive signal and $N_{t+\Delta}$ is nuisance variation.
Using the chain rule,

$$\mathrm{I}((S_{t+\Delta}, N_{t+\Delta}); z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

Since

$$x_{t+\Delta} = (S_{t+\Delta}, N_{t+\Delta}),$$

we obtain

$$\boxed{\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).}$$

Thus pixel reconstruction may reward both signal and nuisance information.

back

Appendix: reconstruction versus VJEPA

For reconstruction,

$$H(x_{t+\Delta} \mid z_t) = H(x_{t+\Delta}) - \mathrm{I}(x_{t+\Delta}; z_t).$$

Thus minimizing reconstruction uncertainty maximizes

$$\mathrm{I}(x_{t+\Delta}; z_t).$$

Using the signal–nuisance decomposition,

$$\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).$$

So reconstruction can reward nuisance information.
For VJEPA,

$$H(z_{t+\Delta} \mid z_t) = H(z_{t+\Delta}) - \mathrm{I}(z_{t+\Delta}; z_t).$$

If the target encoder removes nuisance,

$$z_{t+\Delta} \approx f_{\bar{\theta}}(S_{t+\Delta}),$$

then the objective focuses on future latent signal rather than raw observation detail.

◂ back to reconstruction ◂ back to VJEPA ◂ back to PIB

Appendix: MSE gives the conditional mean

For a deterministic predictor $g(z_c)$ trained with squared loss,

$$\mathcal{L}(g) = \mathbb{E}\left[\|z_t - g(z_c)\|_2^2\right].$$

Conditioning on a fixed z_c, minimize

$$\mathbb{E}[\|z_t - a\|_2^2 \mid z_c]$$

with respect to a.
Expanding,

$$\mathbb{E}[\|z_t - a\|_2^2 \mid z_c] = \mathbb{E}[\|z_t\|_2^2 \mid z_c] - 2a^\top \mathbb{E}[z_t \mid z_c] + a^\top a.$$

Differentiating with respect to a,

$$-2\mathbb{E}[z_t \mid z_c] + 2a = 0.$$

Hence

$$\boxed{g^\star(z_c) = \mathbb{E}[z_t \mid z_c].}$$

back

Appendix: why the conditional mean can fail

From the previous slide,

$$g^\star(z_c) = \mathbb{E}[z_t \mid z_c].$$

If

$$p(z_t \mid z_c)$$

is unimodal and approximately Gaussian, the conditional mean is usually a plausible prediction.
However, if

$$p(z_t \mid z_c)$$

is multimodal, the conditional mean may lie between modes.
For example, if

$$z_t \mid z_c \in \{-1, +1\}$$

with equal probability, then

$$\mathbb{E}[z_t \mid z_c] = 0.$$

But 0 is not a valid mode.

Connection to GMJE

A mixture predictive model can preserve multiple possible futures instead of averaging them into a single point.

Appendix: predictive sufficiency for control I

Let the history be

$$h_t = (x_{1:t}, u_{1:t-1}).$$

Let z_t be the learned latent state.
Predictive sufficiency means

$$\boxed{p(z_{t+1:t+H} \mid h_t, u_{t:t+H-1}) = p(z_{t+1:t+H} \mid z_t, u_{t:t+H-1}).}$$

This says that, once z_t is known, the full history h_t contains no additional information about future latent trajectories.

Meaning

z_t is a sufficient predictive state for future latent dynamics under planned actions.

back

Appendix: predictive sufficiency for control II

Suppose the future cost is

$$C_{t:t+H-1} = \sum_{k=0}^{H-1} c(z_{t+k+1}, u_{t+k}).$$

Because this cost depends on the future only through the future latent trajectory,

$$z_{t+1:t+H},$$

predictive sufficiency implies

$$\boxed{p(C_{t:t+H-1} \mid h_t, u_{t:t+H-1}) = p(C_{t:t+H-1} \mid z_t, u_{t:t+H-1}).}$$

Therefore the optimal value can be written using z_t rather than h_t:

$$Q^\star(h_t, u_t) = Q^\star(z_t, u_t).$$

Hence an optimal policy can be written as

$$\boxed{\pi^\star(u_t \mid h_t) = \pi^\star(u_t \mid z_t).}$$

back

Appendix: collapse avoidance I

Assume a collapsed context encoder:

$$e_\theta(x_c) \equiv c.$$

Then the predictor cannot use context:

$$p_\phi(z_t \mid z_c, \xi) = p_\phi(z_t \mid c, \xi).$$

The best possible collapsed predictor is therefore the unconditional target distribution given ξ:

$$p^\star(z_t \mid \xi).$$

The minimal achievable collapsed NLL is

$$\inf_{p_\phi} \mathbb{E}[-\log p_\phi(z_t \mid c, \xi)] = H_q(z_t \mid \xi).$$

So a collapsed representation pays the entropy of the aggregated target distribution.

back

Appendix: collapse avoidance II

Use the conditional mutual information identity:

$$\mathrm{I}_q(z_t; x_t \mid \xi) = H_q(z_t \mid \xi) - H_q(z_t \mid x_t, \xi).$$

Rearranging,

$$\boxed{H_q(z_t \mid \xi) = H_q(z_t \mid x_t, \xi) + \mathrm{I}_q(z_t; x_t \mid \xi).}$$

If target diversity holds, then

$$\mathrm{I}_q(z_t; x_t \mid \xi) > 0.$$

Therefore,

$$H_q(z_t \mid \xi) > H_q(z_t \mid x_t, \xi).$$

So the collapsed predictor has strict excess loss

$$\boxed{\text{excess loss} = \mathrm{I}_q(z_t; x_t \mid \xi).}$$

back

Appendix: local KL geometry I

Consider a smooth parametric family of probability densities

$$p_\vartheta(x), \qquad \vartheta \in \mathbb{R}^d.$$

Let

$$d\vartheta$$

be a small perturbation of the parameter.
We study the local KL divergence

$$\mathrm{KL}(p_{\vartheta+d\vartheta} \| p_\vartheta) = \mathbb{E}_{p_{\vartheta+d\vartheta}} \left[\log \frac{p_{\vartheta+d\vartheta}(X)}{p_\vartheta(X)} \right].$$

For small $d\vartheta$, the key claim is

$$\boxed{\mathrm{KL}(p_{\vartheta+d\vartheta} \| p_\vartheta) = \frac{1}{2} d\vartheta^\top G(\vartheta) d\vartheta + o(\|d\vartheta\|^2).}$$

The matrix $G(\vartheta)$ is the Fisher information matrix. This means KL divergence locally behaves like a quadratic distance on the statistical manifold.

back

Appendix: local KL geometry II

It is slightly cleaner to expand

$$\mathrm{KL}(p_\vartheta \| p_{\vartheta+d\vartheta}) = \mathbb{E}_{p_\vartheta}\left[\log p_\vartheta(X) - \log p_{\vartheta+d\vartheta}(X)\right].$$

Taylor expand the log-density:

$$\log p_{\vartheta+d\vartheta}(x) = \log p_\vartheta(x) + d\vartheta^\top \nabla_\vartheta \log p_\vartheta(x) + \frac{1}{2} d\vartheta^\top \nabla^2_\vartheta \log p_\vartheta(x) d\vartheta + o(\|d\vartheta\|^2).$$

Substituting into the KL gives

$$\mathrm{KL}(p_\vartheta \| p_{\vartheta+d\vartheta}) = -d\vartheta^\top \mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] - \frac{1}{2} d\vartheta^\top \mathbb{E}_{p_\vartheta}[\nabla^2_\vartheta \log p_\vartheta(X)] d\vartheta + o(\|d\vartheta\|^2).$$

Appendix: local KL geometry III

From the Taylor expansion, the first-order term contains the score expectation:

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)].$$

We now show that this term is zero.
Using

$$\nabla_\vartheta \log p_\vartheta(x) = \frac{\nabla_\vartheta p_\vartheta(x)}{p_\vartheta(x)},$$

we have

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] = \int p_\vartheta(x) \frac{\nabla_\vartheta p_\vartheta(x)}{p_\vartheta(x)}\, dx.$$

Canceling $p_\vartheta(x)$ gives

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] = \int \nabla_\vartheta p_\vartheta(x)\, dx.$$

Appendix: local KL geometry IV

From the previous slide,

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] = \int \nabla_\vartheta p_\vartheta(x)\, dx.$$

We use the regularity condition

$$\int \nabla_\vartheta p_\vartheta(x)\, dx = \nabla_\vartheta \int p_\vartheta(x)\, dx.$$

Typical sufficient conditions are:

- the support of $p_\vartheta(x)$ is fixed in ϑ;
- $p_\vartheta(x)$ is differentiable with respect to ϑ;
- $\nabla_\vartheta p_\vartheta(x)$ is dominated by an integrable function;
- boundary terms vanish if the support has boundaries or is unbounded;
- the Fisher information is finite.

Since p_ϑ is a probability density,

$$\int p_\vartheta(x)\, dx = 1.$$

Therefore,

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] = \nabla_\vartheta 1 = 0.$$

Appendix: local KL geometry V

From the Taylor expansion, we had

$$\mathrm{KL}(p_\vartheta \| p_{\vartheta+d\vartheta}) = -d\vartheta^\top \mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] - \frac{1}{2} d\vartheta^\top \mathbb{E}_{p_\vartheta}[\nabla^2_\vartheta \log p_\vartheta(X)] d\vartheta + o(\|d\vartheta\|^2).$$

Since

$$\mathbb{E}_{p_\vartheta}[\nabla_\vartheta \log p_\vartheta(X)] = 0,$$

the first-order term disappears:

$$\mathrm{KL}(p_\vartheta \| p_{\vartheta+d\vartheta}) = -\frac{1}{2} d\vartheta^\top \mathbb{E}_{p_\vartheta}[\nabla^2_\vartheta \log p_\vartheta(X)] d\vartheta + o(\|d\vartheta\|^2).$$

Define

$$\boxed{G(\vartheta) = -\mathbb{E}_{p_\vartheta}[\nabla^2_\vartheta \log p_\vartheta(X)].}$$

Appendix: why Euclidean gradient descent I

Suppose we want to minimize a smooth loss

$$\mathcal{L}(\vartheta).$$

At the current parameter value ϑ, consider a small update

$$\vartheta \leftarrow \vartheta + \Delta\vartheta.$$

Using a first-order Taylor expansion,

$$\mathcal{L}(\vartheta + \Delta\vartheta) \approx \mathcal{L}(\vartheta) + \nabla_\vartheta \mathcal{L}^\top \Delta\vartheta.$$

Since $\mathcal{L}(\vartheta)$ is constant with respect to $\Delta\vartheta$, locally decreasing the loss means minimizing

$$\nabla_\vartheta \mathcal{L}^\top \Delta\vartheta.$$

But this alone is unbounded below: choosing a very large step in the direction

$$-\nabla_\vartheta \mathcal{L}$$

can make the linearized objective arbitrarily negative.

◂ back to improvements ◂ back to geometry

Appendix: why Euclidean gradient descent II

To prevent an arbitrarily large update, we penalize the Euclidean step size:

$$\|\Delta\vartheta\|_2^2.$$

This gives the local quadratic objective

$$\min_{\Delta\vartheta} \quad \nabla_\vartheta \mathcal{L}^\top \Delta\vartheta + \frac{1}{2\eta}\|\Delta\vartheta\|_2^2.$$

Here $\eta > 0$ controls the step size.

Differentiate with respect to $\Delta\vartheta$:

$$\nabla_{\Delta\vartheta}\left(\nabla_\vartheta \mathcal{L}^\top \Delta\vartheta + \frac{1}{2\eta}\Delta\vartheta^\top \Delta\vartheta\right) = \nabla_\vartheta \mathcal{L} + \frac{1}{\eta}\Delta\vartheta.$$

Setting this to zero gives

$$\nabla_\vartheta \mathcal{L} + \frac{1}{\eta}\Delta\vartheta = 0.$$

Therefore,

$$\boxed{\Delta\vartheta = -\eta\nabla_\vartheta \mathcal{L}.}$$

Appendix: why Euclidean gradient descent III

Thus ordinary gradient descent is the solution of

$$\min_{\Delta\vartheta} \quad \nabla_\vartheta \mathcal{L}^\top \Delta\vartheta + \frac{1}{2\eta}\|\Delta\vartheta\|_2^2.$$

Equivalently, it solves:

decrease the linearized loss $+$ stay close in Euclidean distance.

The resulting update is

$$\boxed{\vartheta \leftarrow \vartheta - \eta\nabla_\vartheta \mathcal{L}.}$$

Key point

Euclidean gradient descent assumes that the natural notion of distance between nearby parameters is

$$\|\Delta\vartheta\|_2^2.$$

Information-geometric replacement

Natural gradient replaces this Euclidean distance by the local KL/Fisher distance between distributions.

Appendix: KL trust-region update

A KL trust-region update solves

$$\min_{\Delta\vartheta} \quad \nabla_\vartheta \mathcal{L}^\top \Delta\vartheta$$

subject to

$$\mathrm{KL}(p_{\vartheta+\Delta\vartheta} \| p_\vartheta) \leq \epsilon.$$

Using the local KL approximation,

$$\mathrm{KL}(p_{\vartheta+\Delta\vartheta} \| p_\vartheta) \approx \frac{1}{2} \Delta\vartheta^\top G(\vartheta) \Delta\vartheta.$$

Hence the constraint becomes

$$\Delta\vartheta^\top G(\vartheta) \Delta\vartheta \leq 2\epsilon.$$

The solution points in the natural-gradient direction:

$$\boxed{\Delta\vartheta \propto -G(\vartheta)^{-1} \nabla_\vartheta \mathcal{L}.}$$

back

Appendix: Gaussian Product-of-Experts I

BJEPA fuses a learned dynamics expert and an auxiliary prior:

$$p_{\text{post}}(z_t \mid z_c, \eta) \propto p_{\text{dyn}}(z_t \mid z_c) p_{\text{prior}}(z_t \mid \eta).$$

Assume both experts are Gaussian:

$$p_{\text{dyn}}(z_t \mid z_c) = \mathcal{N}(z_t; \mu_{\text{dyn}}, \Sigma_{\text{dyn}}),$$

$$p_{\text{prior}}(z_t \mid \eta) = \mathcal{N}(z_t; \mu_{\text{prior}}, \Sigma_{\text{prior}}).$$

The posterior log-density is the sum:

$$\log p_{\text{post}} = \log p_{\text{dyn}} + \log p_{\text{prior}} + \text{const}.$$

Thus multiplying Gaussian experts is equivalent to adding their quadratic log-density terms.

back

Appendix: Gaussian Product-of-Experts II

Write the two log densities up to constants:

$$\log p_{\text{dyn}} = -\frac{1}{2}(z_t - \mu_{\text{dyn}})^\top \Sigma_{\text{dyn}}^{-1}(z_t - \mu_{\text{dyn}}),$$

$$\log p_{\text{prior}} = -\frac{1}{2}(z_t - \mu_{\text{prior}})^\top \Sigma_{\text{prior}}^{-1}(z_t - \mu_{\text{prior}}).$$

Adding them gives

$$\log p_{\text{post}} = -\frac{1}{2} z_t^\top \left(\Sigma_{\text{dyn}}^{-1} + \Sigma_{\text{prior}}^{-1}\right) z_t + z_t^\top \left(\Sigma_{\text{dyn}}^{-1}\mu_{\text{dyn}} + \Sigma_{\text{prior}}^{-1}\mu_{\text{prior}}\right) + \text{const}.$$

Therefore,

$$\boxed{\Sigma_{\text{post}}^{-1} = \Sigma_{\text{dyn}}^{-1} + \Sigma_{\text{prior}}^{-1}}$$

and

$$\boxed{\mu_{\text{post}} = \Sigma_{\text{post}} \left(\Sigma_{\text{dyn}}^{-1}\mu_{\text{dyn}} + \Sigma_{\text{prior}}^{-1}\mu_{\text{prior}}\right).}$$

back

Appendix: information-theoretic summary I

The main entropy and mutual-information identities are:

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

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

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

Interpretation

Mutual information measures uncertainty reduction. For example,

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

means that X is informative about Y if observing X reduces uncertainty about Y.

Appendix: information-theoretic summary II: IB

The Information Bottleneck objective is

$$\boxed{\mathcal{L}_{\mathrm{IB}} = \mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y).}$$

Here

$$X = \text{input}, \qquad Z = \text{representation}, \qquad Y = \text{relevance variable / target}.$$

The first term,

$$\mathrm{I}(X;Z),$$

penalizes how much information Z retains about X.
The second term,

$$\mathrm{I}(Z;Y),$$

rewards how much information Z preserves about Y.

Meaning of IB

IB asks for a representation Z that is compressed with respect to X but still predictive of Y.

$$\boxed{\text{IB: compress } X \to Z \quad \text{while preserving information about } Y.}$$

Appendix: information-theoretic summary III: PIB

Predictive Information Bottleneck is the temporal/self-supervised version of IB.
For temporal prediction,

$$X = x_{\leq t}, \qquad Z = z_t, \qquad Y = z_{t+\Delta}.$$

Substituting these into the IB objective gives

$$\boxed{\mathcal{L}_{\mathrm{PIB}} = \mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; z_{t+\Delta}).}$$

The first term,

$$\mathrm{I}(x_{\leq t}; z_t),$$

penalizes how much past information is stored in the latent state.
The second term,

$$\mathrm{I}(z_t; z_{t+\Delta}),$$

rewards information useful for predicting the future latent state.

PIB: compress the past into z_t while preserving information about the future.

Appendix: information-theoretic summary II

The VJEPA predictive MI bound is

$$\boxed{\mathrm{I}(z_t; z_{t+\Delta}) \geq H(z_{t+\Delta}) + \mathbb{E} \log p_\phi(z_{t+\Delta} \mid z_t).}$$

Equivalently,

$$\boxed{H(z_{t+\Delta} \mid z_t) = H(z_{t+\Delta}) - \mathrm{I}(z_t; z_{t+\Delta}).}$$

The signal–nuisance decomposition is

$$\boxed{\mathrm{I}(x_{t+\Delta}; z_t) = \mathrm{I}(S_{t+\Delta}; z_t) + \mathrm{I}(N_{t+\Delta}; z_t \mid S_{t+\Delta}).}$$

The local KL geometry is

$$\boxed{\mathrm{KL}(p_{\vartheta+d\vartheta} \| p_\vartheta) = \frac{1}{2} d\vartheta^\top G(\vartheta) d\vartheta + o(\|d\vartheta\|^2).}$$

Appendix: why information bottleneck was needed

Classical supervised learning asks for a predictor from inputs X to targets Y.

The Information Bottleneck (IB) asks a more structural question:

What is the smallest representation Z of X that still preserves information about Y?

This leads to the trade-off

$$\min_{p(z|x)} \quad \mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y).$$

Core idea

A good representation should be compressed with respect to the input, but sufficient for the target.

Historical motivation

IB generalizes the idea of sufficient statistics beyond classical parametric statistics and connects it to rate–distortion theory.

Tishby, Pereira, and Bialek, 1999/2000; Shannon-style information theory and rate–distortion theory are the conceptual roots.

Appendix: pre-history of IB

IB did not appear in isolation. It sits at the intersection of several older ideas.

Information theory

Shannon introduced entropy and mutual information as quantitative measures of uncertainty and communication.

$$H(X), \qquad \mathrm{I}(X;Y).$$

Rate–distortion theory

Rate–distortion theory studies how much information must be retained to reconstruct a signal under a distortion constraint:

$$\min I(X;\hat{X}) \quad \text{s.t.} \quad \mathbb{E}[d(X,\hat{X})] \leq D.$$

Sufficient statistics

Classical statistics asks for a statistic $Z = T(X)$ that preserves all information about a parameter or target.

Appendix: IB as relevance-based compression

The key conceptual shift in IB is that distortion is not specified manually.

Instead, relevance is defined through another variable Y.

$$X \longrightarrow Z \longrightarrow Y.$$

Here:

- X is the observed input,
- Z is the compressed representation,
- Y is the relevant variable.

The IB objective is

$$\mathcal{L}_{\mathrm{IB}} = \mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y).$$

Meaning

The representation Z should forget parts of X that are irrelevant for predicting Y.

Tishby, Pereira, and Bialek introduced IB as "a short code for X that preserves information about Y."

Appendix: milestone I — original IB method

The original IB method formulated representation learning as the constrained problem

$$\max_{p(z|x)} \quad \mathrm{I}(Z;Y)$$

subject to

$$\mathrm{I}(X;Z) \leq R.$$

Using a Lagrange multiplier gives

$$\boxed{\min_{p(z|x)} \quad \mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y).}$$

The method also produced self-consistent equations for:

$$p(z \mid x), \qquad p(y \mid z), \qquad p(z).$$

Milestone

IB turned representation learning into an information-theoretic compression–prediction trade-off.

Tishby, Pereira, and Bialek, "The Information Bottleneck Method", Allerton 1999 / arXiv 2000.

Appendix: IB self-consistent form

The IB optimum satisfies a soft clustering equation of the form

$$p(z \mid x) = \frac{p(z)}{Z_\beta(x)} \exp\left(-\beta D_{\mathrm{KL}}\left(p(y \mid x) \,\|\, p(y \mid z)\right)\right),$$

where

$$Z_\beta(x) = \sum_z p(z) \exp\left(-\beta D_{\mathrm{KL}}\left(p(y \mid x) \,\|\, p(y \mid z)\right)\right)$$

is a normalizing constant.
The associated marginals are

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

and

$$p(y \mid z) = \frac{1}{p(z)} \sum_x p(x, y) p(z \mid x).$$

Interpretation

Inputs x are clustered together when they induce similar predictive distributions $p(y \mid x)$.

Appendix: milestone II — IB algorithms and clustering

After the original formulation, early work developed practical IB algorithms.

Agglomerative IB

A bottom-up clustering method that merges input clusters while preserving information about Y.

Deterministic annealing view

The parameter β acts like an inverse temperature:

$$\beta \uparrow \implies \text{more target-relevant detail is preserved.}$$

Geometric and clustering interpretations

IB became a principled method for clustering objects by the predictive distributions they induce, rather than by Euclidean distance alone.

Slonim and Tishby, Agglomerative IB; Still, Bialek, and Bottou, geometric clustering using IB.

Appendix: milestone III — deterministic IB

The original IB allows stochastic encoders:

$$p(z \mid x).$$

Later deterministic variants considered representations of the form

$$z = f(x).$$

This leads to deterministic information bottleneck objectives, where compression is imposed through the structure or cardinality of f.

Why this matters

Many machine-learning representations are deterministic maps, especially in neural networks:

$$x \mapsto f_\theta(x).$$

Conceptual shift

Deterministic IB connects the classical IB principle more directly to modern representation learning and clustering.

Strouse and Schwab, "The Deterministic Information Bottleneck", UAI 2016.

Appendix: milestone IV — IB and deep learning

A major revival of IB came from interpreting deep neural networks as a sequence of representations:

$$X \to T_1 \to T_2 \to \cdots \to T_L \to Y.$$

Each layer can be studied in the information plane:

$$\big(\mathrm{I}(X;T_\ell), \mathrm{I}(T_\ell;Y)\big).$$

Deep-learning IB hypothesis

Good hidden layers should discard irrelevant input information while preserving label-relevant information.

Caveat

The strong claim that ordinary deep networks always undergo an explicit compression phase is debated; nevertheless, the IB viewpoint remains influential.

Tishby and Zaslavsky, 2015; Shwartz-Ziv and Tishby, 2017; later critiques refined the interpretation.

Appendix: milestone V — variational and deep IB

For high-dimensional continuous variables, exact mutual information terms are usually intractable.

The Variational Information Bottleneck replaces the exact IB objective with a tractable variational bound. A typical neural form is

$$z \sim q_\theta(z \mid x), \qquad p_\phi(y \mid z).$$

The objective becomes

$$\mathcal{L}_{\mathrm{VIB}} = \mathbb{E}_{q_\theta(z|x)}[-\log p_\phi(y \mid z)] + \beta \mathrm{KL}(q_\theta(z \mid x) \| p_0(z)).$$

Milestone

VIB made IB scalable to neural networks by combining variational inference, stochastic encoders, and reparameterized gradients.

Alemi, Fischer, Dillon, and Murphy, "Deep Variational Information Bottleneck", 2016/2017.

Appendix: from IB to Predictive IB

Classical IB uses an externally specified target Y.

Predictive Information Bottleneck replaces this with the future.

$$X = \text{past}, \qquad Y = \text{future}.$$

For a time series,

$$X = x_{\leq t}, \qquad Y = x_{t+\Delta} \quad \text{or} \quad Y = x_{>t}.$$

The representation is

$$Z = z_t.$$

The objective becomes

$$\boxed{\mathcal{L}_{\mathrm{PIB}} = \mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; x_{>t}).}$$

Interpretation

PIB asks for the smallest state representation of the past that remains maximally predictive of the future.

Appendix: milestone VI — predictive information

Before PIB became a learning objective, predictive information was studied as a measure of temporal structure.

For a sequence, define past and future:

$$X_{\text{past}} = x_{\leq t}, \qquad X_{\text{future}} = x_{>t}.$$

The predictive information is

$$\boxed{\mathrm{I}(X_{\text{past}}; X_{\text{future}}).}$$

Meaning

Predictive information measures how much the past tells us about the future.

Connection to complexity

In complex dynamical systems, the scaling of predictive information can reflect the richness of the underlying process.

Bialek and collaborators (1999) studied predictive information as a measure of structure and complexity in time series.

Appendix: milestone VII — past–future IB

Past–future IB makes the predictive-information idea operational.

Let

$$X_{\text{past}} = x_{\leq t}, \qquad X_{\text{future}} = x_{>t}, \qquad Z = z_t.$$

Then the objective is

$$\min_{p(z_t|x_{\leq t})} \quad \mathrm{I}(X_{\text{past}}; Z) - \beta \mathrm{I}(Z; X_{\text{future}}).$$

Equivalently,

compress the past, preserve the future.

Milestone

This reframed model construction as efficient predictive representation learning.

Creutzig, Globerson, and Tishby (2009) developed past–future information bottleneck formulations for dynamical systems.

Appendix: milestone VIII — predictive inference

Predictive inference views modelling as efficient communication between the past and the future.

The model state Z is useful only insofar as it helps predict future observations.

$$X_{\text{past}} \to Z \to X_{\text{future}}.$$

The PIB objective is

$$\mathcal{L}_{\text{PIB}} = \mathrm{I}(X_{\text{past}}; Z) - \beta \mathrm{I}(Z; X_{\text{future}}).$$

Interpretation

A model should not memorize the past; it should store only the information from the past that is useful for future prediction.

Connection to world models

This is exactly the representation-learning problem faced by predictive state models, latent dynamics models, and JEPA-style architectures.

Still, "Information Bottleneck Approach to Predictive Inference", Entropy 2014.

Appendix: milestone IX — state predictive IB

Recent work applies PIB to learned state representations.

The aim is to learn a low-dimensional state variable

$$z_t = f_\theta(x_{\leq t})$$

that is maximally predictive of the future while discarding irrelevant fluctuations.
The objective has the same information-theoretic form:

$$\mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; x_{>t}).$$

Example use

In molecular dynamics, State Predictive Information Bottleneck methods learn reaction coordinates by preserving future-predictive state information.

Broader relevance

This connects PIB to modern self-supervised learning, latent-state discovery, and control.

State Predictive Information Bottleneck work applies PIB-style objectives to high-dimensional dynamical systems.

Appendix: from VJEPA objective to PIB I

VJEPA uses a target inference model

$$q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta})$$

and a latent predictive model

$$p_{\phi}(z_{t+\Delta} \mid z_t, \xi).$$

The variational VJEPA objective is

$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}_{z_{t+\Delta} \sim q_{\bar{\theta}}(\cdot \mid x_{t+\Delta})} \left[-\log p_{\phi}(z_{t+\Delta} \mid z_t, \xi) \right] + \beta \mathrm{KL} \left(q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta}) \right).$$

The first term is the predictive latent negative log-likelihood:

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}[-\log p_{\phi}(z_{t+\Delta} \mid z_t, \xi)].$$

The second term regularizes the target latent distribution.

Appendix: from VJEPA objective to PIB II

To connect VJEPA to PIB, focus first on the predictive term:

$$\mathcal{L}_{\text{pred}} = \mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)].$$

Let the true conditional distribution of future latents be

$$p^\star(z_{t+\Delta} \mid z_t, \xi).$$

Then the expected negative log-likelihood decomposes as

$$\mathcal{L}_{\text{pred}} = H_{p^\star}(z_{t+\Delta} \mid z_t, \xi) + \mathbb{E}_{p^\star(z_t,\xi)}\mathrm{KL}\left(p^\star(z_{t+\Delta} \mid z_t, \xi) \| p_\phi(z_{t+\Delta} \mid z_t, \xi)\right).$$

Since KL is nonnegative,

$$\mathcal{L}_{\text{pred}} \geq H(z_{t+\Delta} \mid z_t, \xi).$$

At the optimum, if

$$p_\phi(z_{t+\Delta} \mid z_t, \xi) = p^\star(z_{t+\Delta} \mid z_t, \xi),$$

then

$$\boxed{\mathcal{L}_{\text{pred}} = H(z_{t+\Delta} \mid z_t, \xi).}$$

Appendix: from VJEPA objective to PIB III

Now use the mutual-information identity

$$\mathrm{I}(z_t; z_{t+\Delta} \mid \xi) = H(z_{t+\Delta} \mid \xi) - H(z_{t+\Delta} \mid z_t, \xi).$$

Rearranging,

$$H(z_{t+\Delta} \mid z_t, \xi) = H(z_{t+\Delta} \mid \xi) - \mathrm{I}(z_t; z_{t+\Delta} \mid \xi).$$

From the previous slide,

$$\mathcal{L}_{\mathrm{pred}} \approx H(z_{t+\Delta} \mid z_t, \xi).$$

Therefore,

$$\boxed{\mathcal{L}_{\mathrm{pred}} \approx H(z_{t+\Delta} \mid \xi) - \mathrm{I}(z_t; z_{t+\Delta} \mid \xi).}$$

If the marginal future-latent entropy

$$H(z_{t+\Delta} \mid \xi)$$

is prevented from collapsing, minimizing the latent NLL approximately maximizes

$$\mathrm{I}(z_t; z_{t+\Delta} \mid \xi).$$

Appendix: from VJEPA objective to PIB IV

The Predictive Information Bottleneck objective is

$$\mathcal{L}_{\mathrm{PIB}} = \mathrm{I}(x_{\leq t}; z_t) - \lambda \mathrm{I}(z_t; z_{t+\Delta}).$$

Its two forces are:

$$\mathrm{I}(x_{\leq t}; z_t) \quad \text{compress the past into } z_t,$$

and

$$\mathrm{I}(z_t; z_{t+\Delta}) \quad \text{preserve future-predictive information.}$$

VJEPA implements the second force through latent prediction:

$$\mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)] \approx H(z_{t+\Delta} \mid z_t, \xi).$$

Because

$$H(z_{t+\Delta} \mid z_t, \xi) = H(z_{t+\Delta} \mid \xi) - \mathrm{I}(z_t; z_{t+\Delta} \mid \xi),$$

minimizing VJEPA's predictive NLL maximizes a lower bound on future-latent predictive information.

Takeaway

VJEPA is a variational, neural, latent-space realization of PIB: it learns a compressed state z_t whose main value is predictive information about future latents.

Appendix: PIB and VJEPA: careful comparison

The full Predictive Information Bottleneck objective is

$$\mathcal{L}_{\mathrm{PIB}} = \underbrace{\mathrm{I}(x_{\leq t}; z_t)}_{\text{compression of the past}} - \lambda \underbrace{\mathrm{I}(z_t; z_{t+\Delta})}_{\text{future-predictive information}} .$$

VJEPA directly targets the second term through latent prediction:

$$\mathcal{L}_{\mathrm{pred}} = \mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)].$$

If p_ϕ is expressive and the future-latent marginal is nondegenerate, then

$$\mathcal{L}_{\mathrm{pred}} \approx H(z_{t+\Delta} \mid z_t, \xi) = H(z_{t+\Delta} \mid \xi) - \mathrm{I}(z_t; z_{t+\Delta} \mid \xi).$$

Thus minimizing latent NLL approximately maximizes future-predictive information.

Careful point

VJEPA does not necessarily minimize $\mathrm{I}(x_{\leq t}; z_t)$ explicitly. Compression is usually imposed indirectly through architecture, masking, latent dimension, stochasticity, KL regularization, or target-encoder design.

back

Appendix: towards a full PIB-VJEPA objective I (undergoing work: information geometry of VJEPA)

The original VJEPA objective is

$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}_{z_{t+\Delta}\sim q_{\bar{\theta}}(\cdot|x_{t+\Delta})}\left[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)\right] + \beta \mathrm{KL}\left(q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta})\right).$$

The predictive NLL term satisfies, approximately,

$$\mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)] \approx H(z_{t+\Delta} \mid z_t, \xi).$$

Using

$$H(z_{t+\Delta} \mid z_t, \xi) = H(z_{t+\Delta} \mid \xi) - \mathrm{I}(z_t; z_{t+\Delta} \mid \xi),$$

minimizing the NLL encourages large future-predictive information:

$$\mathrm{I}(z_t; z_{t+\Delta} \mid \xi).$$

Careful point

This mainly implements the predictive term of PIB, but does not explicitly minimize

$$\mathrm{I}(x_{\leq t}; z_t).$$

Appendix: towards a full PIB-VJEPA objective II

A more explicit PIB-style extension would add a compression penalty on the current latent state:

$$z_t \sim q_\theta(z_t \mid x_{\leq t}).$$

The proposed objective is

$$\begin{aligned}\mathcal{L}_{\text{PIB-VJEPA}} =& \mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)] \\ &+ \gamma \mathrm{KL}\big(q_\theta(z_t \mid x_{\leq t}) \| p_0(z_t)\big) \\ &+ \beta \mathrm{KL}\big(q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta})\big).\end{aligned}$$

The new term is

$$\gamma \mathrm{KL}\big(q_\theta(z_t \mid x_{\leq t}) \| p_0(z_t)\big).$$

Role of the new term

It penalizes how much information the current latent state z_t carries about the whole past $x_{\leq t}$.

Appendix: why the new KL term compresses the past I

Let the past/history be denoted by

$$h_t := x_{\leq t}.$$

Assume a stochastic current-state encoder

$$z_t \sim q_\theta(z_t \mid h_t).$$

Together with the data distribution $p(h_t)$, this defines the joint distribution

$$q_\theta(h_t, z_t) = p(h_t) q_\theta(z_t \mid h_t).$$

The aggregated current-latent distribution is the marginal

$$q_\theta(z_t) = \int q_\theta(h_t, z_t)\, dh_t = \int p(h_t) q_\theta(z_t \mid h_t)\, dh_t.$$

The representation information is

$$\mathrm{I}(h_t; z_t) = \mathbb{E}_{p(h_t)} \mathrm{KL}\big(q_\theta(z_t \mid h_t) \| q_\theta(z_t)\big).$$

Equivalently,

$$\mathrm{I}(h_t; z_t) = \int p(h_t) q_\theta(z_t \mid h_t) \log \frac{q_\theta(z_t \mid h_t)}{q_\theta(z_t)}\, dz_t\, dh_t.$$

Appendix: why the new KL term compresses the past II

Now consider the proposed KL-to-prior compression term:

$$\mathbb{E}_{p(h_t)}\mathrm{KL}\big(q_\theta(z_t \mid h_t)\|p_0(z_t)\big).$$

Expanding the KL,

$$\mathbb{E}_{p(h_t)}\mathrm{KL}\big(q_\theta(z_t \mid h_t)\|p_0(z_t)\big) = \int p(h_t)q_\theta(z_t \mid h_t)\log\frac{q_\theta(z_t \mid h_t)}{p_0(z_t)}\,dz_t\,dh_t.$$

Insert the aggregated posterior $q_\theta(z_t)$ inside the logarithm:

$$\log\frac{q_\theta(z_t \mid h_t)}{p_0(z_t)} = \log\frac{q_\theta(z_t \mid h_t)}{q_\theta(z_t)} + \log\frac{q_\theta(z_t)}{p_0(z_t)}.$$

Therefore,

$$\mathbb{E}_{p(h_t)}\mathrm{KL}\big(q_\theta(z_t \mid h_t)\|p_0(z_t)\big) = A + B,$$

where

$$A = \int p(h_t)q_\theta(z_t \mid h_t)\log\frac{q_\theta(z_t \mid h_t)}{q_\theta(z_t)}\,dz_t\,dh_t,$$

and

$$B = \int p(h_t)q_\theta(z_t \mid h_t)\log\frac{q_\theta(z_t)}{p_0(z_t)}\,dz_t\,dh_t.$$

Appendix: why the new KL term compresses the past III

Recall from the previous slide:

$$\mathbb{E}_{p(h_t)}\mathrm{KL}\big(q_\theta(z_t \mid h_t)\|p_0(z_t)\big) = A + B.$$

The first term was

$$A = \int p(h_t)q_\theta(z_t \mid h_t)\log\frac{q_\theta(z_t \mid h_t)}{q_\theta(z_t)}\,dz_t\,dh_t.$$

By the definition of mutual information,

$$\mathrm{I}(h_t; z_t) = \int p(h_t)q_\theta(z_t \mid h_t)\log\frac{q_\theta(z_t \mid h_t)}{q_\theta(z_t)}\,dz_t\,dh_t.$$

Therefore,

$$\boxed{A = \mathrm{I}(h_t; z_t).}$$

Meaning

This is exactly the information that the latent state z_t carries about the past/history $h_t = x_{\leq t}$.

Appendix: why the new KL term compresses the past IV

Now consider the second term:

$$B = \int p(h_t) q_\theta(z_t \mid h_t) \log \frac{q_\theta(z_t)}{p_0(z_t)} \, dz_t \, dh_t.$$

The logarithm term

$$\log \frac{q_\theta(z_t)}{p_0(z_t)}$$

depends only on z_t, not on h_t.
So we can marginalize over h_t:

$$B = \int \left[\int p(h_t) q_\theta(z_t \mid h_t) \, dh_t \right] \log \frac{q_\theta(z_t)}{p_0(z_t)} \, dz_t.$$

By definition of the aggregated posterior,

$$\int p(h_t) q_\theta(z_t \mid h_t) \, dh_t = q_\theta(z_t).$$

Appendix: why the new KL term compresses the past V

Substituting

$$\int p(h_t) q_\theta(z_t \mid h_t)\, dh_t = q_\theta(z_t)$$

into the expression for B, we get

$$B = \int q_\theta(z_t) \log \frac{q_\theta(z_t)}{p_0(z_t)}\, dz_t.$$

This is precisely the KL divergence

$$\boxed{B = \mathrm{KL}(q_\theta(z_t) \| p_0(z_t)).}$$

Therefore,

$$\boxed{\mathbb{E}_{p(h_t)} \mathrm{KL}(q_\theta(z_t \mid h_t) \| p_0(z_t)) = \mathrm{I}(h_t; z_t) + \mathrm{KL}(q_\theta(z_t) \| p_0(z_t)).}$$

Since KL divergence is nonnegative,

$$\boxed{\mathrm{I}(h_t; z_t) \leq \mathbb{E}_{p(h_t)} \mathrm{KL}(q_\theta(z_t \mid h_t) \| p_0(z_t)).}$$

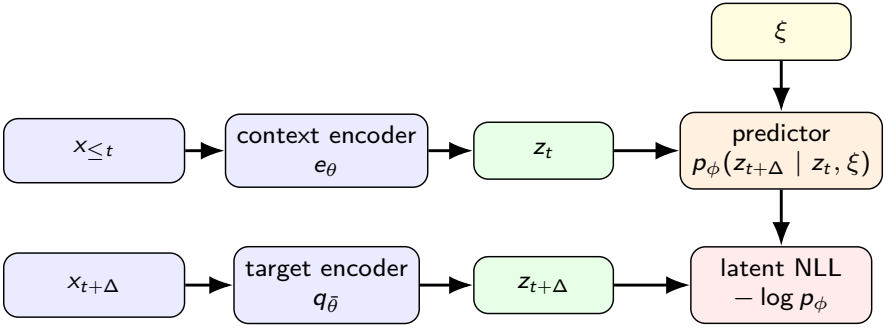
Conclusion

The new KL term minimizes an upper bound on the past-to-state information.

Appendix: original VJEPA architecture

Original VJEPA learns a predictive distribution over the future target latent:

$$q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}), \qquad p_\phi(z_{t+\Delta} \mid z_t, \xi).$$



$$\mathcal{L}_{\mathrm{VJEPA}} = \mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)] + \beta \mathrm{KL}\Big(q_{\bar{\theta}}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta})\Big).$$

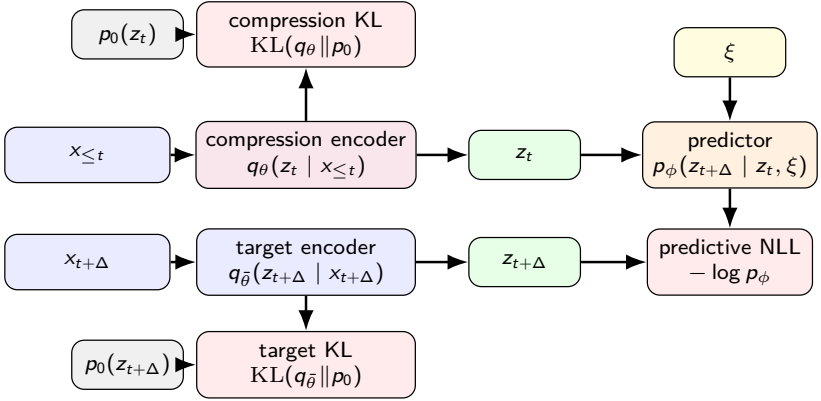
Key point

Original VJEPA targets future-latent prediction, but does not explicitly penalize $\mathrm{I}(x_{\le t}; z_t)$.

Appendix: proposed PIB-VJEPA architecture

To make the PIB compression term explicit, introduce a stochastic current-state encoder:

$$z_t \sim q_\theta(z_t \mid x_{\le t}).$$



$$\mathcal{L}_{\text{PIB-VJEPA}} = \mathbb{E}[-\log p_\phi(z_{t+\Delta} \mid z_t, \xi)] + \gamma \mathbb{E}_{p(x_{\le t})} \mathrm{KL}\Big(q_\theta(z_t \mid x_{\le t}) \| p_0(z_t)\Big) + \beta \mathbb{E}_{p(x_{t+\Delta})} \mathrm{KL}\Big(q_{\bar\theta}(z_{t+\Delta} \mid x_{t+\Delta}) \| p_0(z_{t+\Delta})\Big).$$

Appendix: historical timeline of IB and PIB

Period	Milestone	Contribution
1948	Shannon information theory	Entropy and mutual information formalize uncertainty and communication.
1950s–70s	Rate–distortion theory	Compression is studied under distortion constraints.
1999/2000	Information Bottleneck	Tishby, Pereira, and Bialek define relevant compression: $\mathrm{I}(X;Z) - \beta \mathrm{I}(Z;Y)$.
Early 2000s	IB clustering algorithms	Agglomerative, annealing, and geometric interpretations make IB practical for clustering.
2000s	Predictive information	Past–future mutual information is used to quantify temporal structure.
2009–2014	Predictive IB / predictive inference	The past is compressed into a state that preserves future information.
2015–2017	IB and deep learning	Hidden layers are studied via the information plane.
2016–2017	Deep Variational IB	Variational neural objectives make IB scalable to deep models.
2020s	State predictive IB	PIB-style objectives are used for dynamical systems, state discovery, and scientific ML.

Appendix: conceptual evolution

The evolution can be summarized as:

compression $\longrightarrow$ relevant compression $\longrightarrow$ predictive compression.

Rate–distortion

Compress X while preserving reconstruction quality.

Information Bottleneck

Compress X while preserving information about a relevant target Y.

Predictive Information Bottleneck

Compress the past while preserving information about the future.

Connection to JEPA

JEPA-style models fit naturally into the final stage: they learn representations by predicting future or missing latent variables rather than reconstructing raw observations.

Appendix: takeaway

The Information Bottleneck began as a theory of relevance-based compression:

$$\min \mathrm{I}(X; Z) - \beta \mathrm{I}(Z; Y).$$

Predictive IB specializes this idea to temporal and self-supervised learning:

$$\min \mathrm{I}(x_{\leq t}; z_t) - \beta \mathrm{I}(z_t; x_{>t}).$$

Main lesson

A representation should not preserve everything. It should preserve what is useful.

For VJEPA

The useful information is future-predictive latent information:

$$\mathrm{I}(z_t; z_{t+\Delta}).$$

High-level message

IB explains representation learning as compression with relevance; PIB explains world-model learning as compression with prediction.

Bayesian filtering: belief over hidden states

Bayesian filtering recursively estimates a hidden state z_t from observations

$$x_{1:t} = (x_1, \ldots, x_t).$$

The central object is the filtering posterior:

$$\boxed{p(z_t \mid x_{1:t})}$$

which means:

belief about the current hidden state given all observations so far.

At each time step, Bayesian filtering has two stages:

$$\boxed{\text{prediction} \quad + \quad \text{update/correction.}}$$

1. Prediction

Use the dynamics model to propagate the previous belief:

$$p(z_t \mid x_{1:t-1}) = \int p(z_t \mid z_{t-1}) p(z_{t-1} \mid x_{1:t-1})\, dz_{t-1}.$$

Bayesian filtering: update and VJEPA connection

After predicting the current state, use the new observation x_t to correct the belief.

2. Update / correction

By Bayes' rule,

$$p(z_t \mid x_{1:t}) = \frac{p(x_t \mid z_t)p(z_t \mid x_{1:t-1})}{p(x_t \mid x_{1:t-1})}.$$

The denominator is the normalizing constant:

$$p(x_t \mid x_{1:t-1}) = \int p(x_t \mid z_t)p(z_t \mid x_{1:t-1})\, dz_t.$$

Connection to VJEPA

VJEPA provides a latent predictive model

$$p_\phi(z_{t+\Delta} \mid z_t, \xi),$$

which can serve as the prediction step of a learned latent Bayesian filter.

Model Predictive Control (MPC): objective

Model Predictive Control chooses actions by repeatedly solving a finite-horizon planning problem.
At time t, given the current state s_t, MPC solves

$$u^{\star}_{t:t+H-1} = \underset{u_{t:t+H-1}}{\arg\min}\ \mathbb{E}\left[\sum_{k=0}^{H-1} c(s_{t+k}, u_{t+k}) + c_H(s_{t+H}) \,\middle|\, s_t\right].$$

Meaning

MPC searches for the best future action sequence

$$u_t, u_{t+1}, \ldots, u_{t+H-1}$$

by predicting the future consequences of those actions over a horizon H.

back

Model Predictive Control (MPC): symbols

In the MPC objective,

$$u^{\star}_{t:t+H-1} = \underset{u_{t:t+H-1}}{\arg\min}\, \mathbb{E}\left[\sum_{k=0}^{H-1} c(s_{t+k}, u_{t+k}) + c_H(s_{t+H}) \,\middle|\, s_t\right],$$

the symbols mean:

- s_t: current state of the system.
- u_t: action chosen at time t.
- $u_{t:t+H-1}$: planned action sequence over horizon H.
- $c(s_{t+k}, u_{t+k})$: step cost at future time $t+k$.
- $c_H(s_{t+H})$: terminal cost at the end of the horizon.
- $\mathbb{E}[\cdot \mid s_t]$: expectation over future trajectories predicted from s_t.
- $u^{\star}_{t:t+H-1}$: optimal planned action sequence.

back

MPC with a VJEPA latent dynamics model

VJEPA provides a latent predictive model:

$$p_\phi(z_{t+\Delta} \mid z_t, \xi).$$

For control, let the conditioning include actions:

$$p_\phi(z_{t+1} \mid z_t, u_t, \xi_t).$$

Then MPC can be performed in latent space:

$$u^\star_{t:t+H-1} = \underset{u_{t:t+H-1}}{\arg\min}\ \mathbb{E}\left[\sum_{k=0}^{H-1} c(z_{t+k}, u_{t+k}) + c_H(z_{t+H}) \,\middle|\, z_t\right].$$

Future latent rollouts are generated by

$$z^{(m)}_{t+k+1} \sim p_\phi(z_{t+k+1} \mid z^{(m)}_{t+k}, u_{t+k}, \xi_{t+k}).$$

Connection to VJEPA

VJEPA supplies a probabilistic latent dynamics model, allowing MPC to plan over uncertain latent futures rather than raw observations.

back

How does the action enter VJEPA for control?

For control, the latent predictive model must be action-conditioned.
The original VJEPA predictive model is

$$p_\phi(z_{t+\Delta} \mid z_t, \xi_t),$$

where ξ_t denotes side information such as mask position, temporal offset, or task context.
For control, we include the action as part of the conditioning:

$$\boxed{p_\phi(z_{t+1} \mid z_t, u_t, \xi_t).}$$

Equivalently, define an augmented side-information variable

$$\tilde{\xi}_t = (\xi_t, u_t),$$

so that

$$p_\phi(z_{t+1} \mid z_t, u_t, \xi_t) = p_\phi(z_{t+1} \mid z_t, \tilde{\xi}_t).$$

Interpretation

The action u_t tells the latent dynamics model which future transition to predict. Different actions induce different predictive distributions over z_{t+1}.

Action conditioning: simple input or separate encoder?

There are two common implementation choices.

Option 1: concatenate action with latent state

For low-dimensional continuous or discrete actions, use

$$h_t = [z_t, u_t, \xi_t],$$

and feed it directly into the predictor: $p_\phi(z_{t+1} \mid z_t, u_t, \xi_t) = \mathcal{N}\big(z_{t+1}; \mu_\phi(h_t), \Sigma_\phi(h_t)\big)$.

Option 2: encode the action separately

For structured actions, use an action encoder

$$a_t = r_\psi(u_t),$$

then predict with

$$p_\phi(z_{t+1} \mid z_t, a_t, \xi_t) = \mathcal{N}\big(z_{t+1}; \mu_\phi(z_t, a_t, \xi_t), \Sigma_\phi(z_t, a_t, \xi_t)\big).$$

Practical choice

If u_t is a simple vector, concatenation is enough. If u_t is text, symbolic, graph-structured, or high-dimensional, a separate action encoder is cleaner.