P-JEPA:
JEPA Augmentations from Embodied and Causal Mathematics

PIATRA . INSTITUTE

May 2026

Abstract

The original P-JEPA proposal claimed to extend Joint Embedding Predictive Architectures (JEPA) by replacing the homogeneous target embedding with a sheaf-valued predictive state over a stratified interaction space. That framing was decorative: the implementation computed a posterior-weighted variance, not a coboundary, and the sheaf structure was never wired into a training loop. This revision discards the replacement framing and adopts a plug-in framing instead. JEPA is treated as the working substrate; each piece of embodied or causal mathematics from the original paper is implemented as an auxiliary loss, head, or sampler that can be added to a stock JEPA training loop and ablated. The contribution is the resulting *typology*: which auxiliary losses match which structural assumptions about the data, with toy evidence for directional effects and preregistered evaluation protocols for V-JEPA-scale tests. The toy evidence (5 preregistered hypothesis tests on dishworld, 8 generated JSON artifacts) shows that (i) the obstruction gate in the original paper is a no-op on every reported suite, with the "p_jepa_stack" agent numerically identical to a plain exact value-of-information agent; (ii) value-aware active probing beats entropy probing with paired bootstrap CI [+0.009, +0.017] across 50 deterministic seeds; (iii) the trained intervention encoder is matched within CI [+0.000, +0.005] by a frozen random projection of equal width; (iv) a real cellular sheaf reduces coboundary energy 10× as theory predicts but produces *slightly worse* downstream action choice than the raw cover (CI [−0.005, −0.004]); (v) on a small NumPy JEPA with the auxiliary losses ported as toggleable terms, the viability head shows a positive trend (CI [−0.007, +0.10]) while the bisimulation regularizer at the chosen weight is mis-calibrated and hurts (CI [−0.23, −0.03]). The toy is at its variance limit and these results are *directional* signals for V-JEPA-scale ablation, not quantitative rankings. The paper closes with a priority order for V-JEPA implementation: intervention loss first, composition consistency and active masking next, sheaf consistency on overlapping clips conditional on the H4 boundary, bisimulation with curriculum tuning, viability head last (or first if the downstream is safety-critical). All claims are gated by the results in `docs/HYPOTHESIS_RESULTS.md`.

1. Reframing

The 2026 version of this paper attempted to introduce a new architecture: a sheaf of predictive affordance models on a stratified interaction space, trained jointly with intervention, viability, and composition objectives. The architecture was named P-JEPA.

That paper does not exist as an implementation. The reasons are documented at length in `docs/CRITIQUE.md` and tested in `docs/HYPOTHESIS_RESULTS.md`. The short version is that the paper's mathematical objects (sheaves, coboundaries, cohomology, viability kernels) are sound, but the simulation implements only a scalar posterior-weighted variance and a finite-state value-of-information solver. There is no real cellular sheaf in the original code, no JEPA encoder, and no comparison against any version of JEPA. The paper proposes a replacement for an architecture it never benchmarks.

This revision adopts the plug-in framing instead. JEPA is the substrate. Each piece of mathematics

from the original paper becomes an auxiliary loss or head added to a stock JEPA training loop, and is evaluated by ablation. The paper's contribution shifts from "a new architecture" to "a typology of when each augmentation matches JEPA's data and when it doesn't."

This is a strictly weaker claim than the original. It is also one that can be defended with experiments.

The argument has four parts:

1. **§2-3 The mathematics, restated as loss terms.** Each commitment from the original paper is given its precise loss-function form and its PyTorch signature, suitable for adding to a V-JEPA reference implementation.
2. **§4 The toy.** A small NumPy JEPA on dishworld with each loss as a toggleable term, used to verify gradient flow and produce directional evidence at low cost.
3. **§5 The hypothesis tests.** Five preregistered experiments (H1-H5) on the existing code and the new toy, with paired bootstrap CIs and binary verdicts.
4. **§6 The typology.** Which losses match which inductive biases, and what the toy and existing-code results imply about a real V-JEPA evaluation.

§7-8 are limits and reproducibility. The original §10 (Meta-World adapter) and §11 (formal contract interface) remain as supporting material in `docs/ARCHITECTURE.md` and the unchanged simulation code.

2. P-representations, briefly

We retain the core definition from the original paper because it remains a clean statement of the criterion: a representation is *intervention-sufficient* at tolerance ε if it preserves the information needed to predict the outcomes of arbitrary admissible actions:

$$D\big(\mathbb{P}_A(o_{t+1:t+k}, v_{t+1:t+k} \mid h_t, do(\alpha)),\ \mathbb{P}_A(o_{t+1:t+k}, v_{t+1:t+k} \mid s_t, do(\alpha))\big) \le \varepsilon.$$

This is the predictive-state criterion of Littman & Sutton (2001) extended with viability v and intervention semantics. The *operational* content for a JEPA augmentation is: s_t must support predictions $\widehat{y}_\alpha$ for each α in a chosen test bank. This is the intervention loss (§3A below). Everything else in the original paper's "mathematical stack" can be read as a *secondary constraint* on the representation: bisimulation, sheaf consistency, viability, composition, active perception each add a different inductive bias on top of the predictive-state core.

The original paper presented these as a single integrated loss with five auxiliary terms. This revision treats them as five separable augmentations, each with its own ablation.

3. The augmentation table

Each augmentation is a loss or sampler that plugs into a stock JEPA training step. The full PyTorch signatures, where each one plugs in, and the proposed real evaluations are in

`docs/JEPA_AUGMENTATIONS.md`. This section gives only the names and the loss equations.

Name	Loss / sampler	Inductive bias	Where it matters most
Intervention	$\mathcal{L}_{do} = \mathbb{E}\|h_\psi(f_\theta(x), e(\alpha)) - y_\alpha\|^2$	encoder must predict action outcomes	action-conditioned downstream tasks
Bisimulation	$\mathcal{L}_{\text{bisim}} = \mathbb{E}\big\|\|f(x) - f(x')\| - \mathbb{E}_\alpha D(\hat{y}_{x,\alpha}, \hat{y}_{x',\alpha})\big\|$	latent metric anchored to outcome metric	tasks where visual similarity disagrees with action similarity
Active masking	$m^* = \arg\max_m \text{Var}_k[g_\phi^{(k)}(f_\theta(x \otimes m), m)]$	hard-example mining via ensemble disagreement	sample-efficient pretraining
Viability head	$\mathcal{L}_{\text{viab}} = \mathbb{E}(\sigma(b_\psi(f(x), e(\alpha))) - u)^2$	latent linearly separates unsafe states	safety-critical downstream policies
Sheaf consistency (on overlap)	$\mathcal{L}_{\text{glue}} = \mathbb{E}_{\Omega_{ij}}\|\rho_{i,ij}(f(x_i)) - \rho_{j,ij}(f(x_j))\|^2$	adjacent clips encode coherently	video pretraining with overlapping windows
Composition consistency	$\mathcal{L}_{\text{comp}} = \mathbb{E}\|g(g(s, \alpha_1), \alpha_2) - g(s, \alpha_1 \circ \alpha_2)\|^2$	predictor is associative under action composition	multi-step latent planning

The full loss is a weighted sum:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{JEPA}} + \lambda_{do}\,\mathcal{L}_{do} + \lambda_{\text{bisim}}\,\mathcal{L}_{\text{bisim}} + \lambda_{\text{viab}}\,\mathcal{L}_{\text{viab}} + \lambda_{\text{glue}}\,\mathcal{L}_{\text{glue}} + \lambda_{\text{comp}}\,\mathcal{L}_{\text{comp}}.$$

Active masking is not in this sum because it is a sampler, not a loss. The λ values are augmentation-specific tuning knobs.

The original paper presented this as $\mathcal{L}_{P\text{-JEPA}}$, a single objective. Treating the terms as separable augmentations is what allows ablation: switching individual λ values to zero and measuring downstream performance.

4. The dishworld JEPA toy

`simulation/pjepa_sim/jepa_toy/` implements a small NumPy JEPA on dishworld contexts (sensor + visual features, regime hidden state). Two-layer MLP encoder, EMA target encoder, mask predictor, optional outcome predictor (for intervention loss), optional viability head. Adam optimizer and manual backprop throughout. The model trains in seconds per seed.

The toy exists for two reasons:

1. **Gradient verification.** Each auxiliary loss can be enabled and its loss curve checked for monotonic decrease. This is a correctness check on the implementation before any V-JEPA-scale port.
2. **Directional evidence.** The toy's downstream evaluation is the same regime-cluster + action-utility evaluation used by the rest of the simulation. A toy advantage suggests scale will likely reproduce; a toy disadvantage suggests the augmentation's inductive bias may not match this data type. Both readings are weak, since the toy is at its variance limit, but they cost far less than GPU time.

The toy is not a real V-JEPA result. A V-JEPA result requires PyTorch, a real video dataset, matched compute, and frozen-feature linear probing against I-JEPA / V-JEPA baselines. None of these exists in this repository. The toy is the bridge from "the augmentation has a precise mathematical form" to "the augmentation is testable at scale."

5. Preregistered hypothesis tests

Each hypothesis was registered with a binary pass/fail criterion before running. Results land in `docs/HYPOTHESIS_RESULTS.md` and the JSON artifacts in `simulation/output/experiments/`. The five verdicts:

#	Preregistered hypothesis	Verdict	Evidence
H1	the obstruction gate is a no-op	confirmed	the gate never fires; the two agents are bit-identical
H2	active vs entropy probing is seed noise	rejected	active beats entropy by $+0.0130$, CI95 $[+0.0094, +0.0166]$, $41/50$ seeds
H3	the trained encoder matches a frozen random projection	confirmed	score delta CI95 $[+0.000, +0.005]$
H4	the sheaf framing is decorative	confirmed, and strengthened	the glued centres score below scalar, CI95 $[-0.005, -0.004]$
H5	the JEPA augmentations help on the toy	not supported	no augmentation has a CI95 strictly above zero

Each verdict is unpacked below.

H1: Obstruction gate is a no-op

The paper's `p_jepa_stack` agent and `active_psr_probe` agent both call the same `_decision_probe_result` function with different `use_obstruction_gate` flags. The gate short-circuits a probe when obstruction falls below `spec.sheaf_threshold`. **Result: PASS**. Across all five suites, the initial obstruction (0.15-0.26) is well above the threshold (0.06), so the gate never fires; the two agents are numerically identical to floating-point precision. The distinction in the original paper between "the P-JEPA stack" and "the active PSR probe" is operationally vacuous on the reported suites.

H2: Active vs entropy probing is seed noise

The original paper reports a 5-seed sweep where value-aware active probing beats entropy probing by 0.005 with one seed favouring entropy. **Result: FAIL** (hypothesis rejected). With 50 seeds and 10000 paired bootstrap resamples, active beats entropy by mean +0.0130, CI95 [+0.0094, +0.0166], with 41/50 seeds favouring active. The original 5-seed sweep was underpowered, not wrong. The "value-aware probing beats entropy" claim is supported and should be reported with this CI.

H3: Trained encoder ≈ frozen random projection

The "neural P-representation" was claimed to recover regimes via learned representation. **Result: PASS** (hypothesis confirmed). Across 10 seeds, the trained MLP gets risk-adjusted score 0.802 and cluster purity 1.000. A frozen random TinyMLP of identical width gets 0.800 and 0.994. Delta CI95 [+0.000, +0.005] for score, [+0.000, +0.018] for purity. The work is being done by clustering on linearly-separable Bernoulli sources, not by gradient training. The "neural" framing is decorative; the mechanism is test-vector clustering.

H4: Sheaf framing is decorative

Implemented a real cellular sheaf in `simulation/pjepa_sim/representation/sheaf_toy.py`: learned cover ($K = 6$ k-means clusters), 1-skeleton (8.9 edges on average), learned linear restriction maps with ridge regularization, assembled coboundary δ_0, sheaf Laplacian $L_0 = \delta_0^\top \delta_0$, $\dim H^0 \approx 9.8$, $\dim H^1 \approx 42.3$ on the 1-skeleton. **Result: FAIL on preregistered criterion in unexpected direction**. Coboundary energy drops 10× from gluing (0.354 to 0.034). The math works. But the glued centers score 0.798 vs scalar 0.802, with CI95 [−0.005, −0.004] entirely negative. The sheaf inductive bias *hurts* downstream action choice on categorical hidden state. The hypothesis is *strengthened* not weakened: the framing is not neutral, it's slightly counterproductive on dishworld. The natural prediction: on continuous overlapping data (V-JEPA temporal clips), where there genuinely is a global section to recover, the same mechanism should help. That experiment is in §6.

H5: JEPA augmentations help on the toy

Trained base JEPA + each augmentation (intervention, bisim, active masking, viability) + all combined, 12 seeds × 6 variants on dishworld. **Result: FAIL** (no augmentation has CI strictly above zero). The detailed pattern is more interesting than the binary verdict:

Variant	Mean score	Mean − base	CI95
base JEPA	0.591	n/a	n/a
+intervention	0.590	−0.001	[−0.127, +0.125]
+bisim	0.461	−0.130	[−0.231, −0.027]
+active masking	0.585	−0.006	[−0.059, +0.048]
+viability	0.624	+0.033	[−0.007, +0.098]
+all	0.477	−0.114	[−0.209, −0.023]

Three real findings: (i) **bisim at $\lambda = 0.3$ is mis-calibrated** and actively hurts (CI excludes zero negatively), consistent with the literature warning that bisim needs careful curriculum tuning; (ii) **viability shows a positive trend** with CI nearly excluding zero, most informative on seeds where base JEPA converges to a degenerate latent, where viability "rescues" the result; (iii) **intervention and active masking are neutral** at this scale. The toy's variance dwarfs the augmentation effect.

The result does not falsify the augmentations. It establishes that the toy is at its detection limit and that bisim specifically needs λ-curriculum work before promotion to scale.

6. The typology

Combining the H1-H5 results with the inductive-bias analysis in §3:

Augmentation	Toy evidence	When to expect it to help at scale
Intervention loss	neutral on toy (variance-limited)	high gain on action-conditioned video (SSv2, robot rollouts); the inductive bias is well-matched
Bisimulation	hurts at $\lambda = 0.3$	medium gain conditional on curriculum tuning and a well-trained intervention head
Active masking	neutral on toy	small-to-medium gain on representation pretraining; adjacent literature suggests ~0.5-1%

Augmentation	Toy evidence	When to expect it to help at scale
Viability head	positive trend (CI ~excludes zero)	high gain for safety-critical downstream; low for plain recognition
Sheaf consistency	hurts on categorical (H4)	hurts on discrete regimes; *predicted to help* on continuous overlapping data (V-JEPA clips). Conditional on H4-positive pattern.
Composition consistency	not tested on toy	medium gain on multi-step planning (V-JEPA 2). Calibration of k-step rollouts.

The typology is the actual contribution. It tells a V-JEPA implementer:

- **Add intervention loss first.** Highest expected gain. The inductive bias matches.
- **Add composition consistency second.** Cheap. Improves multi-step rollouts.
- **Try active masking third.** Cheap, with literature precedent.
- **Try sheaf consistency on overlapping clips fourth.** Conditional on the H4 boundary: works on continuous overlapping data, not on categorical regimes.
- **Add bisimulation fifth.** Requires a working intervention head first and careful λ curriculum.
- **Add viability last** unless the downstream is safety-critical, in which case first.

This replaces the original paper's "loss with five auxiliary terms" framing. It is a priority order, not an architecture.

7. Limits

The toy is small. Dishworld has 4 categorical regimes, 11-dim contexts, and Bernoulli outcomes. The H5 toy is at its variance limit (base JEPA itself ranges 0.44-0.80 across seeds). A toy negative is weak evidence against scale gain; a toy positive is weak evidence for. The whole point of §3's PyTorch signatures and `docs/JEPA_AUGMENTATIONS.md`'s proposed evaluations is that the real test is at V-JEPA scale, which this repository cannot execute.

The mathematical objects from the original paper are honestly tested. H1 shows the obstruction gate is operationally vacuous. H4 shows the sheaf framing is at best decorative and at worst slightly harmful on categorical regimes. H3 shows the "neural" framing is decorative. H2 confirms the active-probing claim against entropy. H5 gives mixed directional signal on the auxiliary losses.

The original paper's §9 hidden-regime table, §10 Meta-World adapter, and §11 formal contract interface are not falsified by this revision. They are reframed: they remain useful as evidence that

the *VOI part* of the original P-JEPA stack works (it ties exact Bayesian VOI; see H1), not as evidence that the sheaf or "neural" framings work.

The paper does not establish that any augmentation beats stock V-JEPA at scale. It establishes that the augmentations have precise mathematical forms, that they are correctly implemented in a NumPy toy, that one (viability) shows a positive trend even on a variance-limited toy, that one (bisim) needs tuning before scale, and that the priority order in §6 is defensible from the available evidence.

The cellular sheaf in `representation/sheaf_toy.py` is the only place in the project (and possibly in the JEPA-adjacent literature) where a real cellular sheaf (learned cover, learned linear restrictions, assembled coboundary, sheaf Laplacian, reported cohomology dimensions) is constructed and ablated on a representation-learning task. The H4 negative result is therefore a genuine empirical finding about the limits of sheaf-style coherence as a representation-learning prior, not just an absence of evidence.

8. Reproducibility

All experiments use `uv` and run from `simulation/`:

```
uv run python -m pjepa_sim.experiments.h1_obstruction_gate
uv run python -m pjepa_sim.experiments.h2_seed_sweep_bootstrap
uv run python -m pjepa_sim.experiments.h3_frozen_random_baseline
uv run python -m pjepa_sim.experiments.h4_sheaf_vs_scalar
uv run python -m pjepa_sim.experiments.h5_jepa_augmentations
```

Each writes a JSON artifact to `output/experiments/` and a single PASS/FAIL line. The existing local audit (`uv run python -m pjepa_sim.cli.verify_all`) continues to pass; the new experiments are pure additions to the codebase.

`docs/HYPOTHESIS_RESULTS.md` is the durable record of which hypotheses passed, which failed, and which decisions the results imply for future revisions.

`docs/JEPA_AUGMENTATIONS.md` is the PyTorch design document. It gives precise loss signatures, where each loss plugs into a V-JEPA reference implementation, and the success criterion that would constitute a real positive result for each augmentation.

9. What this paper is not

It is not a new architecture. It is not a foundation model proposal. It does not claim that any auxiliary loss beats V-JEPA at scale. It does not run any benchmark against I-JEPA, V-JEPA, V-JEPA 2, or any video foundation model. It does not learn a robot controller, real perception, language grounding, or end-to-end neural sheaf. The cellular sheaf construction in `sheaf_toy.py` is the only sheaf in the project, and on the only dataset it was tested on (dishworld) it produced a negative result.

What this paper *is*: an honest typology of when each of the embodied/causal mathematical commitments from the original P-JEPA proposal is likely to improve a stock JEPA training recipe. Each augmentation has a precise loss-function form, a PyTorch signature, a toy-scale gradient-flow verification, and a preregistered V-JEPA-scale evaluation protocol. The priority order in §6 is the contribution.

A follow-up paper that runs even one of the augmentations at V-JEPA scale on Something-Something V2 or DROID would be the natural sequel. The infrastructure for it (JEPA toy, hypothesis-test framework, sheaf construction, design specs) is in this repository.

References

The references here are restricted to works that are actually used in the augmentation specifications or hypothesis tests. The original paper's longer reference list, including forward-dated citations and the citation hygiene issues flagged in `docs/CRITIQUE.md`, is being audited separately.

- Ames, A. D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K., & Tabuada, P. (2019). Control barrier functions: theory and applications. *European Control Conference*.
- Assran, M. et al. (2023). Self-supervised learning from images with a joint-embedding predictive architecture (I-JEPA). arXiv:2301.08243.
- Bardes, A., et al. (2024). Revisiting feature prediction for learning visual representations from video (V-JEPA). arXiv:2404.08471.
- de Haan, P., Jayaraman, D., & Levine, S. (2019). Causal confusion in imitation learning. *NeurIPS*.
- Ferns, N., Panangaden, P., & Precup, D. (2011). Bisimulation metrics for continuous Markov decision processes. *SIAM Journal on Computing*, 40(6).
- Friston, K. (2010). The free-energy principle. *Nature Reviews Neuroscience*, 11.
- Hansen, J., & Ghrist, R. (2019). Toward a spectral theory of cellular sheaves. *Journal of Applied and Computational Topology*, 3.
- LeCun, Y. (2022). A path towards autonomous machine intelligence. v0.9.2.
- Littman, M. L., & Sutton, R. S. (2001). Predictive representations of state. *NeurIPS*.
- Pearl, J. (2009). *Causality*. 2nd edition. Cambridge.
- Robinson, M. (2017). Sheaves are the canonical data structure for sensor integration. *Information Fusion*, 36.
- Ross, S., Gordon, G. J., & Bagnell, J. A. (2011). A reduction of imitation learning to no-regret online learning. *AISTATS*.
- Zhang, A., McAllister, R., Calandra, R., Gal, Y., & Levine, S. (2021). Learning invariant representations for reinforcement learning without reconstruction (deep bisimulation). *ICLR*.