K-JEPA: FROM POINT TO DISTRIBUTIONAL PREDICTION BY KERNELIZED JOINT-EMBEDDING PREDICTIVE ARCHITECTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

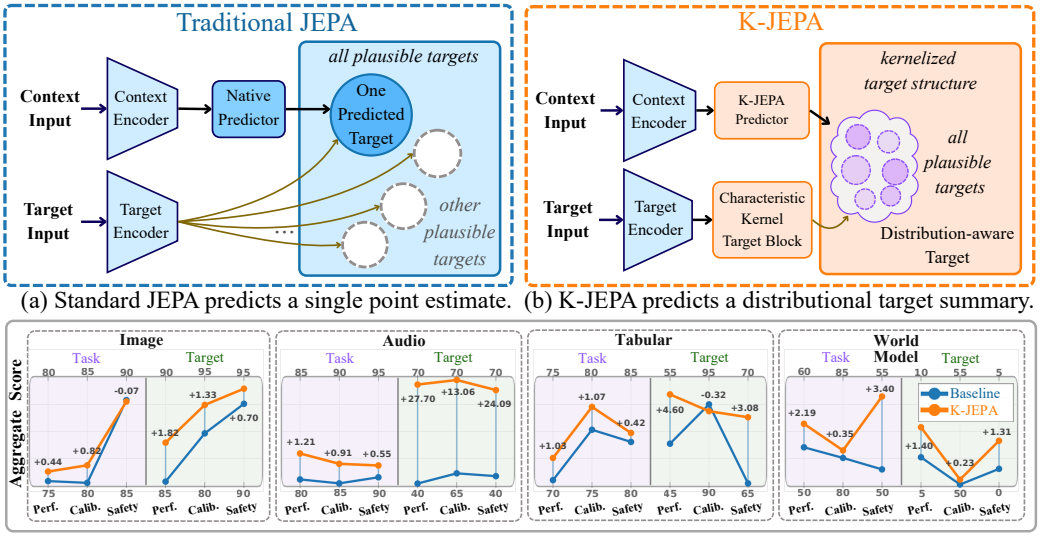
Joint-Embedding Predictive Architectures (JEPA) learn representations by predicting latent target features from a visible context. Although effective, standard JEPA objectives treat prediction as a point-estimation problem: the population-optimal predictor recovers a single point summary of the conditional target distribution. When a visible context x_c admits multiple plausible masked or future targets, this point estimate *retains insufficient distributional information to identify* the underlying conditional target distribution. We propose K-JEPA, a novel, simple, kernelized extension that applies a characteristic-kernel transformation to JEPA latent targets while preserving the architecture and prediction setup. Rather than predicting only the latent target z_t, K-JEPA predicts its kernel representation $\phi(z_t)$. We prove this changes the population target from a point summary of the latent target distribution to its conditional kernel mean embedding, which *uniquely identifies the corresponding conditional target distribution when the kernel is characteristic*. Empirically, we evaluate this distributional behavior in controlled settings and introduce a systematic target-retrieval comparison for direct evaluation of JEPA target prediction across heterogeneous formulations and modalities. We complement this evaluation with probabilistic scoring, calibration, selective prediction, prediction-set efficiency, and downstream transfer analyses. Our results characterize when K-JEPA improves target prediction and how these changes relate to downstream performance and reliability. Code is made available to the reviewers during the discussion process.

1 INTRODUCTION

Joint-Embedding Predictive Architectures (JEPAs) are a prominent family of non-generative self-supervised learning methods that learn representations by predicting latent targets from a visible context (Assran et al., 2023; Bardes et al., 2024). This principle has been instantiated across modalities such as image and video (Assran et al., 2023; Bardes et al., 2024), raw audio (Yuksel et al., 2025), and structured tabular data (Thimonier et al., 2025). However, JEPAs are trained to match a single observed target embedding. For example, I-JEPA and T-JEPA minimize squared distances between predicted and target representations (Assran et al., 2023; Thimonier et al., 2025). For a fixed target representation, the population minimizer of squared prediction loss is the conditional mean (Gneiting, 2011; Song et al., 2026). When the conditional target contains multiple plausible outcomes, this single point cannot enumerate its conditional support and may instead lie between valid target modes (Song et al., 2026) (**Observation 1 – O1**).

A second limitation concerns how JEPA models are evaluated. Native target retrieval can expose predictive failures missed by standard representation diagnostics (Rabby & Auer, 2026), while other studies directly examine what the JEPA predictive pathway infers from incomplete context (Zhou et al., 2026; Nguyen & Nguyen, 2026). Yet, established JEPA work has predominantly assessed learned representations through downstream task performance (Assran et al., 2023; Bardes et al., 2024; Thimonier et al., 2025; Yuksel et al., 2025), and existing native predictor evaluations remain method and setting-specific. To our knowledge, no common native target-retrieval benchmarking has been done for evaluating predictive quality across heterogeneous JEPA formulations and modalities. Beyond retrieval, calibration, selective prediction, and prediction-set coverage provide com-

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



(a) Standard JEPA predicts a single point estimate. (b) K-JEPA predicts a distributional target summary.

(c) Cross-modality effect of K-JEPA on task- and target-level metrics, showing performance & distributional benefits (calibration, safety).

Figure 1: **Motivation of K-JEPA.** (a) Standard JEPA methods predict a single latent target, which can obscure uncertainty and multimodality in the set of plausible future representations. (b) K-JEPA augments latent prediction with a kernel-based distributional target and allows the predictor to capture richer conditional structures. (c) K-JEPA consistently improves or performs on par with the baseline JEPA across image, audio, tabular, and world-model settings in terms of both performance (purple panels) and target distributional qualities (green panel). Results are aggregate scores over the corresponding evaluated models & datasets within each modality so that higher aggregate scores are better. The aggregate-score construction is described in Appendix B.1.

plementary measures of predictive reliability (Guo et al., 2017; Geifman & El-Yaniv, 2019; Romano et al., 2020). Consequently, weaknesses in the predictor can remain obscured even when the learned representations perform well on downstream tasks (**O2**). Further, better predictive quality during JEPA training does not by itself guarantee better downstream representations, nor does it establish that downstream predictions are more reliable under distribution shift (Ovadia et al., 2019) or across relevant subpopulations (Lee et al., 2021; Hébert-Johnson et al., 2018) (**O3**).

Recent work has begun moving JEPA beyond deterministic point prediction. Methods such as VJEPA (Huang, 2026), Var-JEPA (Gögl & Yau, 2026), D-JEPA (Chen et al., 2025), and Branch-JEPA (Song et al., 2026) modify the assumptions or inference structure of standard JEPA through variational inference, generative dynamics, or finite-support prediction (**O4**). In addition, these methods are distributional in different ways. VJEPA and VJE parameterize predictive latent distributions (Huang, 2026; Oji & Fieguth, 2026); Var-JEPA introduces a latent-variable distribution (Gögl & Yau, 2026); D-JEPA models conditional distributions through diffusion or flow matching (Chen et al., 2025); and Branch-JEPA represents a finite set of possible outcomes (Song et al., 2026), whereas K-JEPA represents the conditional target law through its kernel mean embedding. Their evaluations are also largely method- and setting-specific and do not provide a clear analysis of JEPA reliability across formulations and modalities (**O4**). Table 1 summarizes these differences in distribution representation, assumptions, prediction procedure, and predictive object.

Research Questions (RQs). Based on **O1–O4**, we identify four main research questions. **RQ1:** Can the point-prediction limitation of canonical JEPA objectives be formally characterized in settings where one context admits multiple plausible latent targets? **RQ2:** Can JEPA target prediction be systematically analyzed for predictive quality and reliability to complement the predominant focus on downstream representation evaluation? **RQ3:** Do improvements in JEPA target prediction quality consistently translate into better downstream representations and more reliable task predictions? **RQ4:** Can JEPA move beyond point prediction *without* substantial objective changes, explicit parametric distributional assumptions, and additional inference or multi-prediction mechanisms introduced by existing extensions?

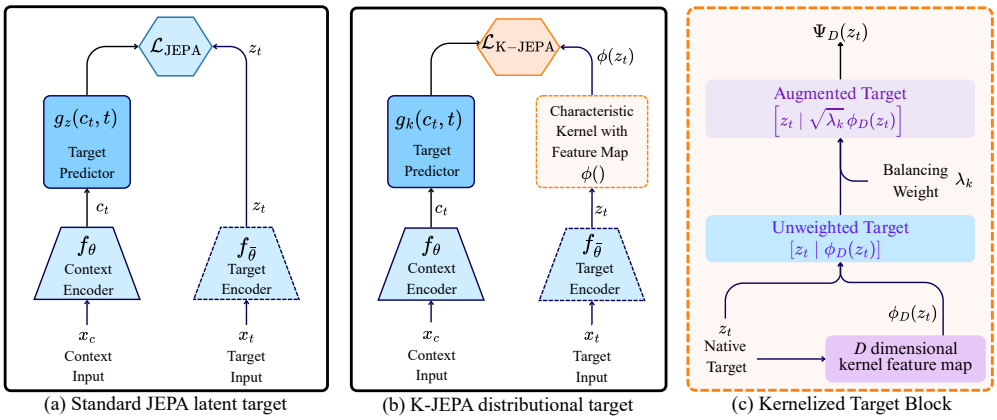
Contributions. To address these RQs, we introduce K-JEPA, a kernelized distributional extension that applies a simple kernel feature transformation, ϕ, to JEPA latent targets while retaining the surrounding JEPA architecture and prediction structure. Our key contributions (KCs) are as follows. **KC1 (addresses RQ1):** We provide a unified formalization of the point-prediction limitation in canonical JEPA objectives and design controlled ambiguous-target analyses that expose when a single predicted embedding fails to represent multiple plausible targets. **KC2 (addresses RQ1 and RQ4):** We formulate K-JEPA as conditional target-distribution prediction, and show that transforming JEPA targets through a characteristic-kernel feature map changes the population target from a point summary to a conditional kernel mean embedding that uniquely identifies the conditional target law. **KC3 (addresses RQ4):** We introduce K-JEPA, a lightweight implementation of this principle using fixed random kernel features and a single prediction procedure, without requiring an explicit conditional distribution, latent sampling, generative decoder, or finite set of prediction branches. **KC4 (addresses RQ2):** We systematically evaluate JEPA target-prediction quality and reliability, covering target retrieval, probabilistic scoring, calibration, selective prediction, prediction-set efficiency, and difficult-target performance. **KC5 (addresses RQ3):** Across multiple JEPA formulations and modalities, we show that K-JEPA substantially improves these predictive properties, with corresponding gains in downstream **representation quality and reliability, including calibration, selective risk, and groupwise behavior**, together with fairness and prediction-set reliability gains in controlled setups.

Table 1: **Comparison with existing works.** K-JEPA assumes a characteristic parent kernel and finite random features to approximate the embedding. The represented distributional predictive objects differ across methods: prior approaches use explicit probabilistic, generative, or finite-support representations, while K-JEPA uses a conditional kernel mean embedding. Abbreviations: Mod.: modality, Eval.: evaluated properties, VIS: visual, AUD: audio, TAB: tabular, STATE: state/trajectory; DQ: distribution quality, UQ: uncertainty quantification, CAL: calibration, SP: selective prediction, PSR: prediction-set reliability, ROB: robustness, FAIR: fairness.

Method	Distribution representation	Distribution assumptions	Native JEPA	Prediction procedure	Predictive object	Mod.	Eval.
V-JEPA	Predictive latent distribution	Parametric family	No	Distribution prediction / sampling	Predictive latent distribution	STATE	ROB
B-JEPA	Product-of-experts	Parametric + prior	No	PoE inference / sampling	Posterior predictive distribution	STATE	ROB
Var-JEPA	Latent-variable distribution	Gaussian latent family	No	Variational inference / sampling	Latent-variable distribution	TAB	UQ, SP
VJE	Conditional latent distribution	Student-t + Gaussian posterior	No	Variational inference / sampling	Conditional latent distribution	VIS	UQ
D-JEPA	Diffusion / flow matching	Score/flow-based	Partial	Iterative generation	Full conditional distribution	VIS, AUD	DQ
Branch-JEPA	K-atomic distribution	Fixed support size	No	Weighted predictions	Best K-atomic approximation	STATE VIS	DQ, CAL
K-JEPA (Ours)	**Kernel mean embedding**	**Non-parametric**	**Yes**	**Single forward prediction**	**Conditional kernel mean embedding identifying the unique conditional law**	**VIS, AUD TAB STATE**	**DQ, UQ CAL, SP PSR, FAIR**

2 Related Work

Joint-embedding predictive learning: JEPAs build on earlier masked latent prediction approaches, including data2vec (Baevski et al., 2022), I-JEPA (Assran et al., 2023) and V-JEPA (Bardes et al., 2024). Related formulations adapt the predictive unit and representation regularization to structured feature subsets, raw audio waveforms, and different architectural settings (Thimonier et al., 2025; Yuksel et al., 2025; Balestriero & LeCun, 2025). Table 1 summarizes how K-JEPA differs from existing work. Existing methods differ in their masking policies, target construction, encoder coupling, and collapse-prevention mechanisms, but their native prediction objectives generally align one predicted representation with an observed target representation. K-JEPA substantially differs from existing methods as it preserves the underlying JEPA formulation and augments its point-valued target with an auxiliary distributional target.



(a) Standard JEPA latent target prediction

(b) K-JEPA distributional target prediction

(c) Kernelized Target Block Realization

Figure 2: **Overview of K-JEPA.** (a) A standard JEPA predicts a point-valued latent target from the context representation. (b) K-JEPA augments the native target with random features of a characteristic kernel. (c) A practical realization of kernelization that also preserves the original JEPA objective while adding a distribution-identifying prediction signal.

Ambiguous and probabilistic latent prediction: Ambiguous prediction has traditionally been addressed using multiple-hypothesis outputs (Rupprecht et al., 2017) and stochastic latent-variable models (Babaeizadeh et al., 2018; Denton & Fergus, 2018). Multiple-hypothesis prediction represents several plausible outcomes through parallel predictions and winner-takes-all-style training (Rupprecht et al., 2017), while stochastic video-prediction models introduce latent variables to generate diverse futures conditioned on the same observed context (Babaeizadeh et al., 2018; Denton & Fergus, 2018). Recent JEPA methods extend the above: variational JEPA and Var-JEPA learn probabilistic latent predictions through variational objectives (Huang, 2026; Gögl & Yau, 2026); D-JEPA combines joint-embedding learning with diffusion or flow-matching objectives (Chen et al., 2025); and Branch-JEPA represents stochastic successors using a finite collection of specialized prediction branches (Song et al., 2026). TC-JEPA instead supplies textual information to reduce uncertainty at masked visual positions, making the target more predictable rather than representing the residual conditional uncertainty (Huang et al., 2026). In contrast, K-JEPA represents the conditional target through a single nonparametric kernel embedding, *without* introducing an explicit latent likelihood, a sampling-based decoder, or a predefined number of output branches.

3 Methodology

K-JEPA uses neural-kernel conditional embeddings (Shimizu et al., 2024) to latent targets with JEPA training: its auxiliary predictor estimates the conditional mean of fixed kernel features while the original JEPA objective continues to learn the underlying representation. Random Fourier features further provide a scalable explicit approximation to shift-invariant kernels (Rahimi & Recht, 2007).

3.1 From latent point prediction to distributional target prediction

JEPAs learn by predicting latent target representations from a visible context. Let x denote an input, x_c a visible context, and x_t a masked target region. A context encoder f_θ maps the visible context x_c to context representations c_t, while a target encoder $f_{\bar{\theta}}$, typically updated as an exponential moving average of f_θ, produces latent target representation z_t. For a target position t, we write

$$c_t = f_\theta(x_c), \qquad z_t = f_{\bar{\theta}}(x_t). \tag{1}$$

A standard JEPA predictor g_z is trained to recover z_t from c_t:

$$\mathcal{L}_{\text{JEPA}} = \|g_z(c_t, t) - \text{sg}(z_t)\|_2^2, \tag{2}$$

where $\text{sg}(\cdot)$ denotes stop-gradient through the target branch.

To characterize this objective at the population level, let T denote the target position or masked-region index, C_T the random context representation available for that target, and Z_T the corresponding random target representation. Also let h be a generic measurable latent-space predictor.

Proposition 1 (Latent JEPA regression estimates a conditional mean)**.** *For squared latent regression, the population minimizer of* $\mathbb{E}_{C_T,T,Z_T}\left[\|h(C_T,T)-Z_T\|_2^2\right]$ *over measurable latent-space predictors* h *is* $h^\star(C_T,T)=\mathbb{E}_{Z_T|C_T,T}\left[Z_T\right]$.

[See Section A.2.1 for the proof.]

The conditional mean provides a single point summary of the target law and does not, in general, identify the distribution from which that target is drawn. K-JEPA augments this point-valued target representation with a kernelized one. Let z, z' denote latent target representations and let k be a characteristic kernel with feature map ϕ into the reproducing-kernel Hilbert space $\mathcal{H}$, such that

$$k(z,z')=\langle\phi(z),\phi(z')\rangle_{\mathcal{H}}. \tag{3}$$

We apply ϕ directly to the JEPA target representation and let g_k denote the corresponding $\mathcal{H}$-valued kernel-feature predictor, with objective

$$\mathcal{L}_k=\|g_k(c_t,t)-\mathrm{sg}(\phi(z_t))\|_{\mathcal{H}}^2. \tag{4}$$

Thus, K-JEPA changes the representation in which the JEPA target is predicted while leaving the underlying context and target representations unchanged.

By the Hilbert-space analogue of Proposition 1, the population minimizer of kernel-feature regression is $\mathbb{E}_{Z_T|C_T,T}[\phi(Z_T)]$, the conditional kernel mean embedding; see Proposition 2 in Appendix A. Because k is characteristic, kernel mean embeddings are injective over probability distributions; see Lemma 1 in Appendix A.

Corollary 1 (Distributional target of K-JEPA)**.** *For a characteristic kernel, the conditional kernel mean embedding*

$$\mu_{Z_T|C_T,T}=\mathbb{E}_{Z_T|C_T,T}\left[\phi(Z_T)\right] \tag{5}$$

uniquely identifies the corresponding conditional target distribution at the population level.

[See Section A.2.4 for the proof.]

The resulting distinction is therefore

$$\underbrace{\mathbb{E}_{Z_T|C_T,T}\left[Z_T\right]}_{\text{point target}} \xrightarrow{\text{K-JEPA}} \underbrace{\mathbb{E}_{Z_T|C_T,T}\left[\phi(Z_T)\right]}_{\text{distribution-identifying target}}. \tag{6}$$

The distinction in Eq. 6 shows that standard squared JEPA regression recovers a conditional point estimate in latent space, whereas characteristic-kernel regression recovers a representation that identifies the conditional latent target law. In this sense, K-JEPA extends JEPA learning from point prediction to distributional prediction through a kernel transformation of its latent targets.

3.2 Finite-dimensional realization and training objective

The characteristic-kernel formulation in Eq. 3 is defined in the reproducing-kernel Hilbert space $\mathcal{H}$, which may be high- or infinite-dimensional. We grid-search over four characteristic shift-invariant kernels, RBF, Matérn-$1/2$, Matérn-$3/2$, and rational quadratic (RQ), and approximate the selected kernel's feature map with D random Fourier features ϕ_D, satisfying

$$\phi_D(z)^\top\phi_D(z')\approx k(z,z'). \tag{7}$$

The map ϕ_D in Eq. 7 is fixed during training and is applied directly to the target representation produced by the existing JEPA target encoder.

As illustrated in Fig. 2(c), K-JEPA retains the native JEPA objective alongside the kernelized term. At the population level, the characteristic-kernel branch already identifies the conditional target law and therefore implicitly determines its conditional mean, so the native term is not required for distribution identification. We retain it in the finite-dimensional realization as an explicit first-moment anchor. While the random-feature branch predicts a finite set of nonlinear kernel features, the native squared objective directly constrains the conditional mean $\mathbb{E}[Z_T \mid C_T, T]$ in latent space. Equivalently, the augmented target map adds a linear-kernel component to the characteristic-kernel component and for any $\rho > 0$ and $\lambda_k > 0$, the resulting sum kernel remains characteristic, so the distribution-identification result of Corollary 1 is preserved.

We now define the finite kernel-feature loss $\mathcal{L}_{k,D} = \|g_k(c_t, t) - \mathrm{sg}(\phi_D(z_t))\|_2^2$, which is the finite-dimensional counterpart of $\mathcal{L}_k$ in Eq. 4. We reuse g_k for its practical D-dimensional realization. Now we define $\lambda_k \geq 0$ as the adaptive balancing factor and $\rho > 0$ as the kernelized-loss weight. In the finite realization, we keep the native latent coordinates together with the kernel features through the target map

$$\Psi_D(z) = \begin{bmatrix} z \\ \sqrt{\rho\lambda_k}\,\phi_D(z) \end{bmatrix}, \tag{8}$$

and corresponding predictor

$$G(c_t, t) = \begin{bmatrix} g_z(c_t, t) \\ \sqrt{\rho\lambda_k}\,g_k(c_t, t) \end{bmatrix}. \tag{9}$$

Training is then expressed as the single target-prediction objective

$$\mathcal{L}_{\text{K-JEPA}} = \|G(c_t, t) - \mathrm{sg}(\Psi_D(z_t))\|_2^2. \tag{10}$$

Expanding Eq. 10 gives

$$\mathcal{L}_{\text{K-JEPA}} = \underbrace{\|g_z(c_t, t) - \mathrm{sg}(z_t)\|_2^2}_{\mathcal{L}_{\text{JEPA}}} + \rho\lambda_k \underbrace{\|g_k(c_t, t) - \mathrm{sg}(\phi_D(z_t))\|_2^2}_{\mathcal{L}_{k,D}}. \tag{11}$$

Thus, Eq. 11 retains the native JEPA prediction and implements the kernelized target transformation within the same objective. We denote the corresponding native and kernel gradient-signal strengths at the shared context representation by $s_{\text{JEPA}} = \|\nabla_{c_t}\mathcal{L}_{\text{JEPA}}\|_2$ and $s_\phi = \|\nabla_{c_t}\mathcal{L}_{k,D}\|_2$.

Because the two loss terms can induce substantially different gradient magnitudes across datasets, we use these non-negative quantities to balance their update scales and prevent the kernel term from either dominating or becoming negligible relative to the native JEPA signal. The kernel weight is then set as $\lambda_k = \frac{s_{\text{JEPA}}}{s_\phi + \varepsilon}$, where $\varepsilon > 0$ is a numerical stabilizer.

For any fixed $\lambda_k > 0$, the corresponding exact direct-sum construction remains characteristic and therefore preserves the distribution-identification result of Corollary 1; see Appendix A.1.1. Random Fourier features provide a scalable finite-dimensional approximation to the kernel formulation in Eq. 7 (Rahimi & Recht, 2007). The exact injectivity statement applies to the parent characteristic kernel, while the finite feature map approximates its geometry with controllable dimensionality (Sutherland & Schneider, 2015).

4 Experimental Settings

Datasets and baselines. We evaluate K-JEPA across audio, image, world-model, and tabular JEPA settings. **Audio:** We use AudioSet (Gemmeke et al., 2017) for pre-training and evaluate on Speech Commands (Warden, 2018) and ESC-50 (Piczak, 2015) with Audio-JEPA (Tuncay et al., 2025), WavJEPA (Yuksel et al., 2025), and PupuJEPA (Gu et al., 2026). **Image:** We evaluate I-JEPA (Assran et al., 2023), LeJEPA (Balestriero & LeCun, 2025), and Rectified LpJEPA (LpJEPA) (Kuang et al., 2026) on ImageNet-100 (Deng et al., 2009), with Copydays (Jégou et al., 2008) for external zero-shot retrieval. **World Models:** We use Stochastic Moving MNIST (Denton & Fergus, 2018) with V-JEPA (Bardes et al., 2024), VJE (Oji & Fieguth, 2026), D-JEPA (Chen et al., 2025), B-JEPA (Huang, 2026), and C-JEPA (Nam et al., 2026). **Tabular:** We use HIGGS (Baldi et al., 2014), ALOI (Geusebroek et al., 2005), and Jannis (Guyon et al., 2019) with T-JEPA (Thimonier et al., 2025), and additionally evaluate LpJEPA, KerJEPA (Zimmermann et al., 2025), and T-JEPA on MIMIC-IV (Johnson et al., 2023). Each baseline is compared with its corresponding K-JEPA variant under the same backbone and evaluation protocol.

Evaluation. Metrics: We perform evaluations across standard performance metrics as well as metrics targeting the distributional qualities of the representations. *Performance* uses classification, future-state, retrieval, and tail-risk metrics: accuracy/Top-1, macro-F1, AUROC, AUPRC, MAE, RMSE, CVaR95, worst-10% error, R@1/R@10, and MRR (Davis & Goadrich, 2006; Rockafellar & Uryasev, 2000). *Uncertainty calibration* uses ECE, Brier score, NLL (Gneiting & Raftery, 2007; Guo et al., 2017). *Selective prediction and uncertainty safety* use Error AUROC, AURC, risk at fixed coverage (Geifman & El-Yaniv, 2019; Traub et al., 2024). *Conformal reliability* uses 95% coverage and prediction-set size, which are directly interpretable in the controlled SM-MNIST target space (Romano et al., 2020; Zhou & Sesia, 2024). *Fairness* is evaluated using aggregated fairness (Cheong et al., 2026; Liu et al., 2025). **Protocol:** Full details are within Appendix B.1.

Table 2: **Audio downstream classification results.** Acc., Macro-F1, and AUPRC measure standard performance; Brier, NLL, and ECE measure uncertainty calibration; Err-AUC and AURC measure selective prediction quality.

		Performance			Uncertainty Calibration			Selective Safety	
Dataset	**Method**	**Acc. ↑**	**Macro-F1 ↑**	**AUPRC ↑**	**Brier ↓**	**NLL ↓**	**ECE ↓**	**Err-AUC ↑**	**AURC ↓**
SpeechCommands	Audio-JEPA	54.1	52.7	53.5	0.61	1.74	4.30	78.4	24.1
	+ K-JEPA	**56.3**	**55.1**	**57.3**	**0.58**	**1.63**	**3.15**	**79.2**	**21.8**
	WavJEPA	92.7	92.1	96.9	0.11	0.27	0.75	94.9	0.7
	+ K-JEPA	**93.0**	**92.5**	**97.2**	**0.10**	**0.26**	**0.72**	94.9	**0.6**
	PupuJEPA	92.2	91.4	96.8	0.11	0.27	1.45	94.5	0.8
	+ K-JEPA	**92.8**	**92.1**	**97.0**	0.11	**0.25**	**1.16**	94.5	**0.7**
ESC-50	Audio-JEPA	54.0	52.7	59.6	0.61	1.73	6.52	76.3	25.9
	+ K-JEPA	**55.9**	**54.5**	**60.7**	**0.60**	**1.69**	**5.82**	**76.8**	**23.9**
	WavJEPA	86.9	86.5	93.6	0.19	0.47	**4.23**	88.6	2.7
	+ K-JEPA	**89.2**	**88.9**	**95.2**	**0.16**	**0.38**	4.30	**90.0**	**1.9**
	PupuJEPA	87.3	87.0	94.1	0.19	0.44	3.59	**89.5**	2.7
	+ K-JEPA	**88.4**	**88.1**	**95.0**	**0.17**	**0.40**	**3.53**	89.0	**2.3**

Table 3: **Image downstream classification results on ImageNet-100.** Metrics follow Table 2.

		Performance			Uncertainty Calibration			Selective Safety	
Dataset	**Method**	**Acc. ↑**	**Macro-F1 ↑**	**AUPRC ↑**	**Brier ↓**	**NLL ↓**	**ECE ↓**	**Err-AUC ↑**	**AURC ↓**
ImageNet-100	I-JEPA	62.2	61.8	**67.2**	0.51	**1.40**	3.85	**83.6**	**15.1**
	+ K-JEPA	**63.3**	**62.0**	67.0	**0.49**	1.41	**3.52**	83.3	15.2
	LpJEPA	69.5	69.5	76.6	0.42	1.10	3.07	**86.5**	9.9
	+ K-JEPA	**70.5**	**70.4**	**77.2**	**0.41**	**1.08**	**2.57**	85.7	**9.6**
	LeJEPA	**88.0**	**88.0**	93.9	0.17	0.42	2.74	90.8	2.1
	+ K-JEPA	87.9	87.9	**94.5**	**0.16**	0.42	**2.65**	**91.3**	2.1

5 RESULTS

We conduct both (i) performance evaluation using standard retrieval and downstream performance and (ii) reliability-oriented evaluations to analyze the benefits of distribution estimation through uncertainty calibration, selective prediction and uncertainty safety, conformal reliability and fairness. Additional results on fairness benefits of K-JEPA and full statistical results for future-state prediction, stochastic-future prediction, and the controlled future-prediction evaluation on SM-MNIST are presented in Sections D.1 and D.2 respectively within the Appendix.

5.1 PERFORMANCE EVALUATION

Audio: Across SpeechCommands and ESC-50, K-JEPA consistently improves standard classification performance, increasing accuracy, Macro-F1, and AUPRC for all three audio JEPA baselines (Table 2). **Image:** On ImageNet-100, K-JEPA improves the overall performance–reliability trade-off across the three visual JEPA formulations (Tables 3 and 4). **Tabular:** We see from Table 6 that performance improved consistently across most performance metrics (top-1, R@1, MRR), settings (downstream task and target-prediction quality), and datasets (Jannis, HIGGS, ALOI). **World Model:** On Stochastic Moving MNIST, K-JEPA improves downstream future prediction across most of the evaluated JEPA formulations (Table 7), while the target-prediction evaluation provides a complementary view (Table 8) of the consistent gains. B-JEPA also shows target-prediction gains, but its downstream changes are more modest. Conversely, C-JEPA remains comparatively stable at the target-prediction level. Figure 3 summarizes the target-to-task transfer across modalities.

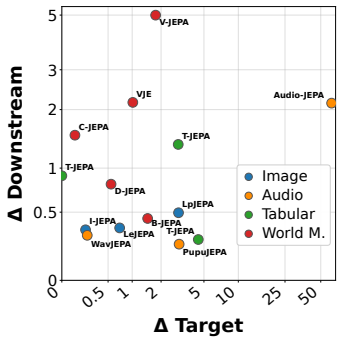


Figure 3: **Target-to-task transfer.** Gains in target retrieval are accompanied by consistent downstream improvements across JEPA variants.

5.2 RELIABILITY-ORIENTED EVALUATION

Audio: We see from Table 2 that calibration generally improves through lower Brier, NLL, and ECE, while AURC decreases for every baseline on both the SpeechCommands and ESC-50 datasets. The direct target-prediction evaluation on SpeechCommands (Table 4) exposes larger method-dependent differences. Audio-JEPA shows the clearest target-level gain, with R@1 increasing from 14.1 to

Table 4: **Target-prediction quality and reliability across modalities.** R@1 and MRR measure prediction performance; Brier, NLL, and ECE measure uncertainty calibration; AURC and Risk50 measure selective prediction safety.

		Performance		Uncertainty Calibration			Selective Safety	
Dataset	**Method**	**R@1 ↑**	**MRR ↑**	**Brier ↓**	**NLL ↓**	**ECE ↓**	**AURC ↓**	**Risk50 ↓**
SpeechCommands	Audio-JEPA	14.1	21.9	0.94	3.67	6.08	67.55	76.92
	+ K-JEPA	**95.3**	**95.9**	**0.12**	**1.63**	**5.88**	**4.25**	**7.69**
	WavJEPA	97.8	98.8	**0.00**	**0.09**	**0.45**	**0.32**	0.12
	+ K-JEPA	**99.4**	**99.6**	0.01	0.10	0.53	0.36	**0.08**
	PupuJEPA	1.6	8.5	0.98	4.16	**1.56**	100.00	100.00
	+ K-JEPA	**4.7**	**14.0**	0.98	4.16	4.32	**95.69**	**92.31**
ImageNet-100	I-JEPA	98.4	99.0	0.02	0.08	1.59	0.02	0.00
	+ K-JEPA	**99.0**	**99.3**	0.02	**0.05**	**1.28**	0.02	0.00
	LpJEPA	60.6	71.5	0.54	1.60	9.86	16.04	11.65
	+ K-JEPA	**65.6**	**74.7**	**0.50**	**1.51**	**8.48**	**13.95**	**9.71**
	LeJEPA	89.1	92.4	0.25	1.16	**13.06**	5.65	**3.88**
	+ K-JEPA	**89.8**	**93.5**	**0.22**	**0.91**	13.79	**4.48**	4.85

Table 5: **Zero-shot retrieval on Copydays.** Zero-shot source retrieval from partially observed queries. Each entry reports baseline → +K-JEPA. Results at 20% missing are omitted because performance is near ceiling. R@1, MRR, ECE, AURC, and Risk50 are reported in percent.

		Performance		Uncertainty Calibration			Selective Safety	
Model	**Missing**	**R@1 ↑**	**MRR ↑**	**Brier ↓**	**NLL ↓**	**ECE ↓**	**AURC ↓**	**Risk50 ↓**
LpJEPA	40%	84.5 → **86.2**	90.1 → **91.2**	0.31 → **0.27**	3.92 → **3.59**	15.40 → **13.70**	13.20 → **10.70**	14.00 → **10.70**
	60%	67.0 → **69.2**	76.9 → **78.3**	0.66 → **0.61**	8.95 → **7.99**	33.00 → **30.60**	30.50 → **25.00**	32.30 → **27.00**
	80%	36.9 → **38.0**	50.4 → 51.1	1.26 → **1.24**	17.38 → **16.93**	62.90 → 62.00	57.90 → 58.00	59.80 → 61.10
LeJEPA	40%	92.9 → **94.3**	95.6 → **96.7**	0.13 → **0.10**	0.64 → **0.42**	5.50 → **3.50**	1.90 → 1.40	1.80 → 1.00
	60%	80.4 → **82.2**	87.0 → **88.2**	0.33 → **0.31**	1.89 → **1.79**	14.50 → **12.60**	6.60 → **5.50**	4.80 → **3.80**
	80%	47.8 → 48.5	60.2 → **61.3**	0.93 → 0.92	8.72 → **8.15**	43.90 → 43.50	30.50 → **28.70**	31.60 → **30.30**
I-JEPA	40%	98.9 → 99.1	99.3 → 99.4	0.02 → 0.02	0.28 → **0.24**	1.10 → 0.90	1.20 → 1.40	1.30 → 1.50
	60%	95.2 → 95.4	96.7 → 97.0	0.10 → 0.09	1.16 → **1.05**	4.80 → 4.50	4.20 → 4.30	4.80 → 4.80
	80%	80.5 → 80.8	86.5 → 86.8	0.39 → 0.38	4.81 → **4.67**	19.20 → 19.00	16.20 → **14.00**	19.60 → **17.30**

95.3 while Brier, NLL, AURC, and selective risk decrease. WavJEPA and PupuJEPA show smaller target-retrieval and selective-risk gains alongside modest downstream improvements.

Image: LpJEPA shows broad target-prediction improvements in retrieval, probabilistic scoring, calibration, and selective risk together with gains in downstream accuracy, Macro-F1, AUPRC, calibration, and AURC. I-JEPA further improves target NLL and ECE, with more modest downstream gains in accuracy and calibration. LeJEPA improves target retrieval, probabilistic scoring, and AURC.

External retrieval validation: Frozen models were evaluated on the Copydays benchmark, where a partially observed query must retrieve its full source image (Table 5). As missing visual content increases, K-JEPA improves retrieval while also producing better calibrated and safer uncertainty estimates, with consistent gains for LpJEPA and LeJEPA. Importantly, *improvements are obtained **without** training a retrieval head or further model adaptation*.

Tabular results. Across Jannis, HIGGS, and ALOI, K-JEPA improves downstream performance while generally improving calibration and selective prediction safety (Table 6). The strongest results appear on HIGGS. On ALOI, K-JEPA improves target retrieval and AURC. Across *MIMIC-IV* (Table 10 in Appendix D.1), we further see that K-JEPA improves the majority of performance, calibration, selective-safety, and fairness metrics across the evaluated hosts.

World-model results. From Tables 7 and 8, we see V-JEPA improves future-target retrieval and selective risk together with strong gains in future-state and stochastic-future prediction. VJE and D-JEPA improve target retrieval, together with several calibration and selective-prediction measures with smaller but consistent downstream improvements. Conformal results are strongest for C-JEPA and V-JEPA, where K-JEPA moves coverage closer to 95% while reducing set size; VJE also improves coverage with a small size trade-off.

6 Discussion and Conclusion

Summary. In the paper, we prove that standard JEPA training objectives can recover only a single point summary of the conditional target distribution. To address this gap, we proposed K-JEPA as a novel kernelized extension that applies a characteristic-kernel transformation to JEPA latent targets. We proved that this extension changes the population target from a point summary of the latent target

Table 6: **Tabular downstream and target-prediction results.**

		Downstream Task						Target-Prediction Quality						
		Perf.	Uncertainty Cal.			Selective Safety		Perf.		Uncertainty Cal.			Selective Safety	
Dataset	Method	Top-1 ↑	Brier ↓	NLL ↓	ECE ↓	AURC ↓	Risk50 ↓	R@1 ↑	MRR ↑	Brier ↓	NLL ↓	ECE ↓	AURC ↓	Risk50 ↓
Jannis	T-JEPA	63.3	0.48	0.85	2.1	21.5	23.3	**36.2**	**41.5**	0.01	1.29	**34.2**	64.9	65.5
	+ K-JEPA	**64.5**	**0.47**	**0.84**	**1.7**	**21.0**	**22.2**	36.1	40.9	0.01	1.29	36.0	**63.5**	**64.4**
HIGGS	T-JEPA	61.3	0.46	0.66	3.8	30.0	**29.9**	42.0	47.0	0.01	1.14	**0.1**	31.4	32.0
	+ K-JEPA	**63.0**	**0.45**	**0.64**	**1.1**	**28.4**	30.5	**46.3**	**50.9**	0.01	1.14	0.2	**27.6**	**26.4**
ALOI	T-JEPA	86.1	0.24	1.22	10.6	**2.6**	**0.6**	61.4	65.1	0.01	0.66	0.1	11.8	3.9
	+ K-JEPA	**86.3**	**0.23**	**1.20**	**10.3**	2.6	0.7	**72.4**	**74.2**	0.01	0.66	0.1	**6.6**	**2.5**

Table 7: **SM-MNIST downstream future-prediction results.** Future-state prediction reports average normalized state error together with tail-risk measures. Stochastic-future prediction reports standard classification metrics.

		Future-State Prediction				Stochastic-Future Prediction						
		Performance		Tail Risk		Performance			Uncertainty Cal.			Selective Safety	
Host	Method	MAE ↓	RMSE ↓	CVaR95 ↓	Worst10 ↓	Acc. ↑	Macro-F1 ↑	AUPRC ↑	Brier ↓	NLL ↓	ECE ↓	Err-AUC ↑	AURC ↓
C-JEPA	Baseline	0.38	0.48	0.69	0.65	75.0	42.9	26.3	0.37	0.56	0.5	51.4	24.4
	+ K-JEPA	**0.36**	**0.45**	**0.66**	**0.63**	75.0	42.9	**28.8**	0.37	0.56	0.7	**55.6**	**21.8**
V-JEPA	Baseline	0.40	0.48	0.66	0.64	75.0	43.1	26.5	0.37	0.56	1.7	51.4	24.3
	+ K-JEPA	**0.36**	**0.46**	**0.65**	**0.62**	74.9	**51.9**	**41.0**	**0.34**	**0.52**	**1.4**	**70.0**	**14.0**
D-JEPA	Baseline	0.36	0.45	0.64	0.62	76.3	57.1	46.7	0.33	0.50	**1.1**	70.9	12.9
	+ K-JEPA	**0.34**	**0.44**	0.64	**0.61**	76.4	**58.8**	**48.3**	**0.32**	**0.49**	1.4	**72.2**	**12.3**
B-JEPA	Baseline	0.40	0.49	0.66	0.64	75.0	42.9	25.8	0.37	0.56	1.5	51.5	23.5
	+ K-JEPA	0.40	**0.48**	0.66	0.64	75.0	**43.5**	**27.6**	0.37	0.56	1.7	**53.7**	**22.6**
VJE	Baseline	0.40	0.49	0.66	0.64	75.0	42.9	24.9	0.37	0.56	1.5	50.0	25.4
	+ K-JEPA	**0.39**	**0.48**	0.66	**0.63**	75.0	**43.6**	**32.9**	0.37	**0.55**	1.6	**60.8**	**18.9**

Table 8: **Controlled SM-MNIST future-prediction results.** Retrieval performance is measured by R@10 and MRR; Brier, NLL, and ECE measure uncertainty calibration; AURC and Risk25 measure selective prediction safety; Cov95 and Size95 report 95% conformal coverage and the corresponding average prediction-set size, respectively. Coverage closer to 95% and smaller sets are preferred.

		Performance		Uncertainty Calibration			Selective Safety		Conformal	
Host	Method	R@10 ↑	MRR ↑	Brier ↓	NLL ↓	ECE ↓	AURC ↓	Risk25 ↓	Cov95 (→95)	Size95 ↓
C-JEPA	Baseline	9.2	4.8	0.99	4.85	0.4	98.6	98.2	96.9	125.4
	+ K-JEPA	9.3	**5.1**	0.99	4.85	0.5	98.5	98.0	**95.2**	**122.8**
V-JEPA	Baseline	8.6	4.6	0.99	4.85	**0.2**	98.4	97.9	98.4	126.2
	+ K-JEPA	**10.9**	**6.0**	0.99	**4.84**	0.5	**95.3**	**93.8**	**95.8**	**123.8**
D-JEPA	Baseline	6.7	3.6	0.99	4.94	1.7	100.0	100.0	**97.1**	**124.4**
	+ K-JEPA	**8.0**	**4.6**	0.99	**4.88**	**0.9**	99.8	100.0	97.9	125.4
B-JEPA	Baseline	6.7	4.5	0.99	4.87	1.7	99.6	100.0	**99.5**	**126.8**
	+ K-JEPA	**10.7**	**6.2**	0.99	**4.84**	**0.4**	**98.8**	**99.0**	99.7	127.2
VJE	Baseline	9.2	5.1	0.99	4.86	0.7	99.7	100.0	91.6	**118.9**
	+ K-JEPA	**10.7**	**5.5**	0.99	4.86	**0.1**	**98.2**	**97.9**	**92.7**	120.0

distribution to its conditional kernel mean embedding, which *uniquely identifies the corresponding conditional target distribution when the kernel is characteristic*.

Discussion. In the paper, we explored the four research questions (RQ1-4) defined in Section 1: We *theoretically* and *empirically* illustrated the **point-prediction limitation (RQ1)**; namely that standard JEPA regression has the conditional mean of the latent target as its population optimum, and thus returns a single point that does not identify a multimodal conditional target law. Direct evaluation of the JEPA predictor exposes behavior that downstream representation metrics alone can obscure, while our proposed **K-JEPA gives clear target-retrieval and reliability gains across different modalities (RQ2)**. Results on Copydays further show that these *gains extend to external held-out retrieval*. Moreover, across settings, improvements in JEPA target prediction correspond to preserved or **improved downstream representation quality and reliability (RQ3; Fig. 3)**. Results further show that target-prediction evaluation and downstream representation evaluation capture related but distinct aspects of JEPA behavior. Our proposed **K-JEPA can move beyond point prediction *without* requiring substantial objective changes (RQ4)**. We do so by transforming latent targets with characteristic-kernel features *without* requiring an explicit parametric conditional distribution, iterative generation, or a fixed set of prediction branches.

Limitations. The exact distribution-identification guarantee applies to the population characteristic kernel. The practical method instead uses a finite random-feature approximation and remains subject to finite data, model capacity, and optimization error. Direct multi-target evaluation is also limited to settings where valid alternatives can be constructed or observed. Future work can examine larger feature approximations and broader ambiguous-target benchmarks.

AI USE STATEMENT

Generative AI tools were used to check human-written mathematical proofs for potential errors. The proofs were otherwise developed by the authors, except for Proposition 4, for which generative AI provided limited assistance in developing the proof. Generative AI was not used to implement the methods or write the experimental code; implementations were based on the official codebases of the corresponding methods with the modifications required for our experiments. Generative AI was additionally used to assist literature search, primarily to identify potentially relevant methods that may have been missed for citation or benchmarking. Identified works were manually verified by the authors. We did not use generative AI for other tasks requiring disclosure. We take responsibility for the final content of this work.

ETHICS STATEMENT

Our experiments use publicly available datasets and do not involve direct interaction with human participants. Permission for use has been obtained in all instances. We further report methodological limitations and potential failures to support responsible interpretation and use of our findings.

REPRODUCIBILITY STATEMENT

We support full reproducibility by providing detailed descriptions of our methodology, experimental setup, datasets, evaluation protocols, and implementation details in the main paper and appendix. The proposed method and training objective are described in Section 3, while the datasets, baselines, and experimental setting are described in Section 4. Our theoretical claims are further supported by detailed proofs presented in Appendix A. The complete evaluation protocol is presented in Appendix B.1, while configuration, training details, pretrained initialization, and hyperparameters are reported in Appendix B.2.1–B.2.3 and Table 9. In addition, the anonymous source code will be made available. Following ICLR's instructions, we will provide reviewers and area chairs with a link to our anonymous repository after the discussion forums open. The code will also be released publicly upon acceptance.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023. doi: 10.1109/CVPR52729.2023.01499.

Mohammad Babaeizadeh, Chelsea Finn, Dumitru Erhan, Roy H. Campbell, and Sergey Levine. Stochastic variational video prediction. In *International Conference on Learning Representations*, 2018.

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

Pierre Baldi, Peter Sadowski, and Daniel Whiteson. Searching for exotic particles in high-energy physics with deep learning. *Nature Communications*, 5:4308, 2014. doi: 10.1038/ncomms5308.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. *arXiv preprint arXiv:2511.08544*, 2025. doi: 10.48550/arXiv.2511.08544.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. arXiv:2404.08471.

Dengsheng Chen, Jie Hu, Xiaoming Wei, and Enhua Wu. Denoising with a joint-embedding predictive architecture. In *The Thirteenth International Conference on Learning Representations*, 2025. arXiv:2410.03755.

Jiaee Cheong, Abtin Mogharabin, Paul Pu Liang, Hatice Gunes, and Sinan Kalkan. Fairssl: Fair multimodal self-supervised learning. In *Forty-third International Conference on Machine Learning*, 2026.

Jesse Davis and Mark Goadrich. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 233–240. ACM, 2006. doi: 10.1145/1143844.1143874.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Emily Denton and Rob Fergus. Stochastic video generation with a learned prior. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1174–1183. PMLR, 2018.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780, 2017. doi: 10.1109/ICASSP.2017.7952261.

Jan-Mark Geusebroek, Gertjan J. Burghouts, and Arnold W. M. Smeulders. The amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005. doi: 10.1023/B:VISI.0000042993.50813.60.

Tilmann Gneiting. Making and evaluating point forecasts. *Journal of the American Statistical Association*, 106(494):746–762, 2011. doi: 10.1198/jasa.2011.r10138.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Moritz Gögl and Christopher Yau. Var-JEPA: A variational formulation of the joint-embedding predictive architecture—bridging predictive and generative self-supervised learning. In *Forty-third International Conference on Machine Learning*, 2026.

Steffen Grünewälder, Guy Lever, Luca Baldassarre, Sam Patterson, Arthur Gretton, and Massimiliano Pontil. Conditional mean embeddings as regressors. In *Proceedings of the 29th International Conference on Machine Learning*, pp. 1823–1830. Omnipress, 2012.

Yicheng Gu, Junan Zhang, Jerry Li, Zhizheng Wu, and Lauri Juvela. Frequency-aware self-supervised music representation learning. *arXiv:2606.25713*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017.

Isabelle Guyon, Lisheng Sun-Hosoya, Marc Boullé, Hugo Jair Escalante, Sergio Escalera, Zhengying Liu, Damir Jajetic, Bisakha Ray, Mehreen Saeed, Michèle Sebag, Alexander Statnikov, Wei-Wei Tu, and Evelyne Viegas. Analysis of the automl challenge series 2015–2018. In *Automated Machine Learning*, pp. 177–219. Springer, 2019. doi: 10.1007/978-3-030-05318-5_10.

Ursula Hébert-Johnson, Michael P. Kim, Omer Reingold, and Guy N. Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1939–1948. PMLR, 2018.

Chen Huang, Xianhang Li, Vimal Thilak, Etai Littwin, and Josh Susskind. Text-conditional JEPA for learning semantically rich visual representations. In *Forty-third International Conference on Machine Learning*, 2026.

Yongchao Huang. VJEPA: Variational joint embedding predictive architectures as probabilistic world models. In *Forty-third International Conference on Machine Learning*, 2026.

Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Computer Vision – ECCV 2008*, pp. 304–317. Springer, 2008. doi: 10.1007/978-3-540-88682-2_24.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

Yilun Kuang, Yash Dagade, Tim G. J. Rudner, Randall Balestriero, and Yann LeCun. Rectified lpjepa: Joint-embedding predictive architectures with sparse and maximum-entropy representations. In *Forty-third International Conference on Machine Learning*, 2026.

Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2796–2804. PMLR, 2018.

Joshua K. Lee, Yuheng Bu, Deepta Rajan, Prasanna Sattigeri, Rameswar Panda, Subhro Das, and Gregory W. Wornell. Fair selective classification via sufficiency. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6076–6086. PMLR, 2021.

Hongbin Liu, Jinyuan Jia, Wenjie Qu, and Neil Zhenqiang Gong. EncoderMI: Membership inference against pre-trained encoders in contrastive learning. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2081–2095. Association for Computing Machinery, 2021. doi: 10.1145/3460120.3484749.

Qinyi Liu, Oscar Deho, Farhad Vadiee, Mohammad Khalil, Srecko Joksimovic, and George Siemens. Can synthetic data be fair and private? a comparative study of synthetic data generation and fairness algorithms. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, pp. 591–600, 2025.

Heejeong Nam, Quentin Le Lidec, Lucas Maes, Yann LeCun, and Randall Balestriero. Causal-JEPA: Learning world models through object-level latent interventions. *arXiv preprint arXiv:2602.11389*, 2026.

William Nguyen and Christopher Nguyen. The JEPA predictor: A transferable operator for occluded feature completion, 2026.

Amin Oji and Paul W. Fieguth. Joint embedding variational bayes. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. arXiv:2602.05639.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, pp. 13991–14002. Curran Associates, Inc., 2019.

Junhyung Park and Krikamol Muandet. A measure-theoretic approach to kernel conditional mean embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21247–21259. Curran Associates, Inc., 2020.

Karol J. Piczak. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pp. 1015–1018. ACM, 2015. doi: 10.1145/2733373.2806390.

Gollam Rabby and Sören Auer. When Graph-JEPA learns the wrong thing: Diagnosing and repairing category-conditional collapse, 2026.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20, pp. 1177–1184. Curran Associates, Inc., 2007.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *The Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/JOR.2000.038.

Yaniv Romano, Matteo Sesia, and Emmanuel J. Candès. Classification with valid and adaptive coverage. In *Advances in Neural Information Processing Systems*, volume 33, pp. 3581–3591. Curran Associates, Inc., 2020.

Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D. Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3611–3620, 2017. doi: 10.1109/ICCV.2017.388.

Eiki Shimizu, Kenji Fukumizu, and Dino Sejdinovic. Neural-kernel conditional mean embeddings. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45040–45059. PMLR, 2024.

Le Song, Jonathan Huang, Alex J. Smola, and Kenji Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968. Association for Computing Machinery, 2009. doi: 10.1145/1553374.1553497.

Zhi Song, Ximing Xing, Zhenchao Tang, Hanbo Huang, Jiehui Huang, Weilong Yan, Tianxu Lv, Minghao Yang, Zhongzheng Niu, Bing He, Lusheng Wang, and Jianhua Yao. Branch-JEPA: Finite-support predictive distributions for JEPA world models. *arXiv preprint arXiv:2607.05238*, 2026. doi: 10.48550/arXiv.2607.05238.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(50):1517–1561, 2010.

Danica J. Sutherland and Jeff Schneider. On the error of random fourier features. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pp. 862–871. AUAI Press, 2015.

Hugo Thimonier, José Lucas De Melo Costa, Fabrice Popineau, Arpad Rimmel, and Bich-Liên Doan. T-JEPA: Augmentation-free self-supervised learning for tabular data. In *The Thirteenth International Conference on Learning Representations*, 2025.

Jeremias Traub, Till J. Bungert, Carsten T. Lüth, Michael Baumgartner, Klaus H. Maier-Hein, Lena Maier-Hein, and Paul F. Jäger. Overcoming common flaws in the evaluation of selective classification systems. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0076.

Wei-Cheng Tseng, Wei-Tsung Kao, and Hung-yi Lee. Membership inference attacks against self-supervised speech models. In *Interspeech 2022*, pp. 5040–5044, 2022. doi: 10.21437/Interspeech.2022-11245.

Ludovic Tuncay, Etienne Labbé, Emmanouil Benetos, and Thomas Pellegrini. Audio-JEPA: Joint-Embedding Predictive Architecture for Audio Representation Learning. In *ICME 2025*, Nantes, France, 2025.

Pete Warden. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*, 2018.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, Xuefeng Du, Kaiyang Zhou, Wayne Zhang, Dan Hendrycks, Yixuan Li, and Ziwei Liu. OpenOOD: Benchmarking generalized out-of-distribution detection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 32598–32611. Curran Associates, Inc., 2022.

Goksenin Yuksel, Pierre Guetschel, Michael Tangermann, Marcel van Gerven, and Kiki van der Heijden. WavJEPA: Semantic learning unlocks robust audio foundation models for raw waveforms. *arXiv preprint arXiv:2509.23238*, 2025. doi: 10.48550/arXiv.2509.23238.

Yanfei Zhou and Matteo Sesia. Conformal classification with equalized coverage for adaptively selected groups. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Zihan Zhou, Qifu Wen, and Xi Zeng. SR-JEPA: Learning predictive latent state in 3d scenes, 2026.

Jie Zhu, Jirong Zha, Ding Li, and Leye Wang. A unified membership inference method for visual self-supervised encoder via part-aware capability. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 1241–1255. Association for Computing Machinery, 2024. doi: 10.1145/3658644.3690202.

Eric Zimmermann, Harley Wiltzer, Justin Szeto, David Alvarez-Melis, and Lester Mackey. KerJEPA: Kernel discrepancies for euclidean self-supervised learning. *arXiv preprint arXiv:2512.19605*, 2025. doi: 10.48550/arXiv.2512.19605.

TABLE OF CONTENTS: APPENDIX

A Additional Theory and Proofs for Methodological Claims **15**

A.1 Additional Theoretical Results 15

A.2 Proofs 17

B Additional Experimental Settings Details **19**

B.1 Evaluation protocol 19

B.2 Configuration and Training Details 20

C Extended Related Work **20**

D Additional Results **21**

D.1 MIMIC-IV Classification and Fairness 21

D.2 SM-MNIST Statistical Uncertainty and Significance 22

A ADDITIONAL THEORY AND PROOFS FOR METHODOLOGICAL CLAIMS

This appendix gives additional theoretical results and proofs for the formal claims in the method section. We write $C = (C_T, T)$ for the context and target-position pair, and $Z = Z_T$ for the corresponding random target representation produced by the EMA target encoder. All expectations are assumed to exist, and predictors are assumed to be measurable and square-integrable whenever required.

A.1 ADDITIONAL THEORETICAL RESULTS

Proposition 2 (Kernel regression estimates a conditional **kernel** mean embedding)**.** *For squared regression in a reproducing-kernel feature space, the population minimizer of*

$$\mathbb{E}_{C_T,T,Z_T}\left[\|h(C_T,T)-\phi(Z_T)\|_{\mathcal{H}}^2\right] \tag{12}$$

over measurable $\mathcal{H}$-valued predictors h is

$$h^\star(C_T,T)=\mathbb{E}_{Z_T|C_T,T}\left[\phi(Z_T)\right]. \tag{13}$$

Equivalently, for a realized context representation c_t and target index t,

$$h^\star(c_t,t)=\mathbb{E}_{Z_T|C_T=c_t,T=t}\left[\phi(Z_T)\right]. \tag{14}$$

This is the conditional kernel mean embedding of the target distribution.

[See Section A.2.2 for the proof.]

Lemma 1 (Characteristic kernels identify distributions)**.** *If $k(z,z')=\langle\phi(z),\phi(z')\rangle_{\mathcal{H}}$ is characteristic, then for probability distributions P and Q over latent targets, the kernel mean embeddings*

$$\mu_P=\mathbb{E}_{Z\sim P}[\phi(Z)], \qquad \mu_Q=\mathbb{E}_{Z\sim Q}[\phi(Z)] \tag{15}$$

are injective in the corresponding distributions. Therefore,

$$\mu_P=\mu_Q \implies P=Q. \tag{16}$$

[See Section A.2.3 for the proof.]

A.1.1 CHARACTERISTICNESS UNDER POSITIVE ADAPTIVE BALANCING

The distribution-identification result in Corollary 1 is carried by the characteristic-kernel component of the exact target representation. We next show that positive rescaling of this component does not remove characteristicness, and that the same conclusion holds when the realized balancing coefficient changes across optimization steps.

Proposition 3 (Positive balancing preserves characteristicness of the exact augmented target)**.** *Let k be a characteristic kernel with feature map $\phi : \mathcal{Z} \to \mathcal{H}$. For any scalar $\lambda > 0$, define the exact augmented feature map*

$$\Psi_\lambda(z) = \begin{bmatrix} z \\ \sqrt{\lambda}\,\phi(z) \end{bmatrix}, \tag{17}$$

which induces the kernel

$$K_\lambda(z, z') = \langle z, z' \rangle + \lambda k(z, z'). \tag{18}$$

Then K_λ is characteristic for every $\lambda > 0$.

More generally, let $\{\lambda_r\}_{r\geq 1} \subset (0, \infty)$ be any sequence of positive scalars, where r indexes optimization steps. Then every kernel K_{λ_r} is characteristic. Thus, characteristicness of the exact augmented target does not require the same positive balancing coefficient to be used throughout optimization.

In particular, whenever $s_{JEPA} > 0$, the balancing rule

$$\lambda_k = \frac{s_{JEPA}}{s_\phi + \varepsilon}, \qquad \varepsilon > 0,$$

produces $\lambda_k > 0$, and the corresponding exact augmented construction retains the distribution-identifying characteristic-kernel component.

[See Section A.2.5 for the proof.]

A.1.2 Finite random-feature approximation of distributional discrepancy

The exact injectivity result in Corollary 1 applies to the parent characteristic kernel. The practical realization instead uses the finite random-feature map ϕ_D. We therefore characterize how closely the distributional discrepancy induced by ϕ_D approximates that of the parent kernel.

Let $\mathcal{S} \subseteq \mathcal{Z}$ be a set containing the supports of the target distributions under consideration, and define the finite-feature kernel

$$k_D(z, z') = \phi_D(z)^\top \phi_D(z'). \tag{19}$$

For a fixed realization of the random features, define its uniform kernel approximation error on $\mathcal{S}$ by

$$\eta_D = \sup_{z,z' \in \mathcal{S}} |k_D(z, z') - k(z, z')|\,. \tag{20}$$

For a probability distribution P over latent targets, let

$$\mu_{P,D} = \mathbb{E}_{Z\sim P}[\phi_D(Z)], \tag{21}$$

and define

$$\mathrm{MMD}_k^2(P, Q) = \|\mu_P - \mu_Q\|_{\mathcal{H}}^2, \qquad \mathrm{MMD}_{k_D}^2(P, Q) = \|\mu_{P,D} - \mu_{Q,D}\|_2^2. \tag{22}$$

Proposition 4 (Finite random features preserve characteristic-kernel separation up to approximation error)**.** *Let P and Q be probability distributions supported on $\mathcal{S}$. For the finite-feature kernel k_D in Eq. 19 and approximation error η_D in Eq. 20,*

$$\left|\mathrm{MMD}_{k_D}^2(P, Q) - \mathrm{MMD}_k^2(P, Q)\right| \leq 4\eta_D. \tag{23}$$

Consequently, whenever

$$\mathrm{MMD}_k^2(P, Q) > 4\eta_D, \tag{24}$$

the finite random-feature embeddings remain distinct:

$$\mu_{P,D} \neq \mu_{Q,D}.$$

The same statement applies pointwise to conditional target laws $P(\cdot \mid c)$ and $Q(\cdot \mid c)$ for every conditioning value c for which their supports are contained in $\mathcal{S}$. Hence, although a finite feature map need not be injective over all probability distributions, any pair of conditional target laws whose parent-kernel discrepancy exceeds the finite-feature approximation error remains distinguishable by their D-dimensional kernel mean embeddings.

Moreover, if a random-feature approximation guarantee gives

$$\mathbb{P}(\eta_D \le \bar{\eta}_D(\delta)) \ge 1 - \delta,$$

then, with probability at least $1 - \delta$, Eq. 23 holds simultaneously for all P, Q supported on $\mathcal{S}$, with η_D replaced by $\bar{\eta}_D(\delta)$. Standard random-feature approximation results provide such control under their corresponding assumptions (Sutherland & Schneider, 2015).

[See Section A.2.6 for the proof.]

A.2 Proofs

A.2.1 Proof of Proposition 1

Proof. Using the tower property of conditional expectation, the population risk can be written as

$$\mathbb{E}\left[\|h(C) - Z\|_2^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|h(C) - Z\|_2^2 \mid C\right]\right].$$

It is therefore sufficient to minimize the inner conditional risk for each fixed value $C = c$. For any vector a,

$$\mathbb{E}\left[\|a - Z\|_2^2 \mid C = c\right] = \|a - \mathbb{E}[Z \mid C = c]\|_2^2 + \mathbb{E}\left[\|Z - \mathbb{E}[Z \mid C = c]\|_2^2 \mid C = c\right].$$

The second term does not depend on a, while the first is minimized by $a = \mathbb{E}[Z \mid C = c]$. Hence

$$h^\star(C) = \mathbb{E}[Z \mid C]$$

almost surely. Substituting $C = (C_T, T)$ and $Z = Z_T$ gives

$$h^\star(C_T, T) = \mathbb{E}[Z_T \mid C_T, T].$$

□

A.2.2 Proof of Proposition 2

Proof. The proof is the Hilbert-space analogue of the squared-regression result. By the tower property,

$$\mathbb{E}\left[\|h(C) - \phi(Z)\|_{\mathcal{H}}^2\right] = \mathbb{E}\left[\mathbb{E}\left[\|h(C) - \phi(Z)\|_{\mathcal{H}}^2 \mid C\right]\right].$$

For a fixed value $C = c$ and any $a \in \mathcal{H}$,

$$\mathbb{E}\left[\|a - \phi(Z)\|_{\mathcal{H}}^2 \mid C = c\right] = \|a - \mathbb{E}[\phi(Z) \mid C = c]\|_{\mathcal{H}}^2 + \mathbb{E}\left[\|\phi(Z) - \mathbb{E}[\phi(Z) \mid C = c]\|_{\mathcal{H}}^2 \mid C = c\right].$$

The second term is independent of a, so the conditional risk is minimized by

$$a = \mathbb{E}[\phi(Z) \mid C = c].$$

Therefore

$$h^\star(C) = \mathbb{E}[\phi(Z) \mid C]$$

almost surely. Substituting $C = (C_T, T)$ and $Z = Z_T$ yields

$$h^\star(C_T, T) = \mathbb{E}[\phi(Z_T) \mid C_T, T],$$

which is the conditional kernel mean embedding of the target distribution. □

A.2.3 Proof of Lemma 1

Proof. A kernel k is characteristic when the kernel mean embedding map

$$P \mapsto \mu_P, \qquad \mu_P = \mathbb{E}_{Z \sim P}[\phi(Z)],$$

is injective over the considered class of probability distributions. Therefore,

$$\mu_P = \mu_Q \implies P = Q,$$

which proves the claim. □

A.2.4 Proof of Corollary 1

Proof. For each fixed conditioning value $C = c$, the conditional kernel mean embedding is

$$\mu_{P(\cdot|c)} = \mathbb{E}_{Z\sim P(\cdot|c)}[\phi(Z)] = \mathbb{E}[\phi(Z) \mid C = c].$$

Suppose two conditional laws $P(\cdot \mid c)$ and $Q(\cdot \mid c)$ have equal conditional kernel mean embeddings for almost every c. For each such c,

$$\mu_{P(\cdot|c)} = \mu_{Q(\cdot|c)}.$$

Since k is characteristic, Lemma 1 implies

$$P(\cdot \mid c) = Q(\cdot \mid c)$$

for almost every c. Thus equality of conditional kernel mean embeddings implies equality of the corresponding conditional target laws. Applying this with $C = (C_T, T)$ shows that

$$\mathbb{E}[\phi(Z_T) \mid C_T, T]$$

is distribution-identifying at the population level. □

A.2.5 Proof of Proposition 3

Proof. Fix any $\lambda > 0$. For a probability distribution P over latent targets, the kernel mean embedding induced by the exact augmented feature map in Eq. 17 is

$$\mu_{P,\lambda} = \mathbb{E}_{Z\sim P}[\Psi_\lambda(Z)] = \begin{bmatrix} \mathbb{E}_{Z\sim P}[Z] \\ \sqrt{\lambda}\,\mathbb{E}_{Z\sim P}[\phi(Z)] \end{bmatrix} = \begin{bmatrix} \mathbb{E}_{Z\sim P}[Z] \\ \sqrt{\lambda}\,\mu_P \end{bmatrix}.$$

Suppose two distributions P and Q have equal augmented kernel mean embeddings:

$$\mu_{P,\lambda} = \mu_{Q,\lambda}.$$

Equality of the second components then gives

$$\sqrt{\lambda}\,\mu_P = \sqrt{\lambda}\,\mu_Q.$$

Since $\lambda > 0$,

$$\mu_P = \mu_Q.$$

Because k is characteristic, Lemma 1 implies

$$P = Q.$$

Therefore the kernel K_λ induced by Eq. 17 is characteristic for every $\lambda > 0$.

The argument is pointwise in λ. Hence, for any sequence of positive scalars $\{\lambda_r\}_{r\geq 1}$, the same argument applies separately to every λ_r, and each corresponding kernel K_{λ_r} is characteristic. No assumption that λ_r remains constant across optimization steps is required for this characteristicness statement.

Finally, since $s_\phi \geq 0$ and $\varepsilon > 0$, the denominator $s_\phi + \varepsilon$ is strictly positive. Whenever $s_{\text{JEPA}} > 0$, the balancing rule therefore gives

$$\lambda_k = \frac{s_{\text{JEPA}}}{s_\phi + \varepsilon} > 0,$$

so the preceding result applies to the corresponding exact augmented construction. □

A.2.6 Proof of Proposition 4

Proof. For any kernel q, the squared maximum mean discrepancy between distributions P and Q can be written as

$$\mathrm{MMD}_q^2(P, Q) = \mathbb{E}_{Z,Z'\sim P}[q(Z, Z')] + \mathbb{E}_{Z,Z'\sim Q}[q(Z, Z')] - 2\mathbb{E}_{Z\sim P, Z'\sim Q}[q(Z, Z')],$$

where the paired variables in the first two expectations are independent draws from the corresponding distribution.

Define

$$e_D(z, z') = k_D(z, z') - k(z, z').$$

By Eq. 20,

$$|e_D(z, z')| \le \eta_D$$

for all $z, z' \in \mathcal{S}$. Subtracting the two MMD expansions gives

$$\begin{aligned}&\mathrm{MMD}^2_{k_D}(P, Q) - \mathrm{MMD}^2_k(P, Q)\\ &= \mathbb{E}_{Z,Z'\sim P}[e_D(Z, Z')] + \mathbb{E}_{Z,Z'\sim Q}[e_D(Z, Z')] - 2\mathbb{E}_{Z\sim P,Z'\sim Q}[e_D(Z, Z')].\end{aligned}$$

Taking absolute values and applying the triangle inequality,

$$\begin{aligned}&\left|\mathrm{MMD}^2_{k_D}(P, Q) - \mathrm{MMD}^2_k(P, Q)\right|\\ &\le |\mathbb{E}_{Z,Z'\sim P}[e_D(Z, Z')]| + |\mathbb{E}_{Z,Z'\sim Q}[e_D(Z, Z')]| + 2\,|\mathbb{E}_{Z\sim P,Z'\sim Q}[e_D(Z, Z')]|\\ &\le \eta_D + \eta_D + 2\eta_D = 4\eta_D.\end{aligned}$$

This proves Eq. 23.

By construction of k_D,

$$\mathrm{MMD}^2_{k_D}(P, Q) = \|\mathbb{E}_{Z\sim P}[\phi_D(Z)] - \mathbb{E}_{Z\sim Q}[\phi_D(Z)]\|_2^2 = \|\mu_{P,D} - \mu_{Q,D}\|_2^2.$$

Therefore, if

$$\mathrm{MMD}^2_k(P, Q) > 4\eta_D,$$

Eq. 23 gives

$$\mathrm{MMD}^2_{k_D}(P, Q) \ge \mathrm{MMD}^2_k(P, Q) - 4\eta_D > 0.$$

Hence

$$\mu_{P,D} \ne \mu_{Q,D},$$

which establishes the finite-feature separation result.

For conditional target laws, fix any conditioning value $C = c$ and apply the same argument with

$$P = P(\cdot \mid c), \qquad Q = Q(\cdot \mid c).$$

Thus the result holds pointwise for every conditioning value for which the corresponding conditional supports lie in $\mathcal{S}$.

Finally, suppose that the random-feature construction satisfies

$$\mathbb{P}(\eta_D \le \bar{\eta}_D(\delta)) \ge 1 - \delta.$$

On this event, the deterministic argument above applies with $\eta_D \le \bar{\eta}_D(\delta)$. Therefore, with probability at least $1 - \delta$,

$$\left|\mathrm{MMD}^2_{k_D}(P, Q) - \mathrm{MMD}^2_k(P, Q)\right| \le 4\bar{\eta}_D(\delta)$$

simultaneously for all distributions P, Q supported on $\mathcal{S}$, completing the proof. □

B Additional Experimental Settings Details

B.1 Evaluation protocol

We evaluate K-JEPA at both the self-supervised target-prediction and downstream representation level. First, we measure held-out kernel target-prediction quality using the same masked-target setting in which the JEPA is trained. In controlled ambiguous-target settings, where multiple valid targets $z_t^{(1)}, \ldots, z_t^{(M)}$ are available for the same context, we estimate the conditional kernel mean by: $\widehat{\mu}_{Z|c_t,t} = \frac{1}{M}\sum_{m=1}^{M} \phi_D\left(z_t^{(m)}\right)$, and measure the discrepancy between this empirical conditional embedding and model prediction. This directly evaluates how faithfully the predictor represents the conditional target law when a single context admits multiple plausible targets.

Next, we evaluate the learned representations using a common frozen-encoder protocol. The pretrained encoder is kept fixed and the same downstream evaluator is applied to every representation. We report task performance together with probabilistic scoring, calibration, selective prediction, and groupwise counterparts where demographic attributes are available; controlled SM-MNIST additionally includes conformal coverage and prediction-set efficiency.

For Fig. 1, metrics are expressed on a 0–100 higher-is-better scale, with lower-is-better scores transformed as $100 - s$, and equally averaged within each performance, calibration, and safety group over the evaluated model–dataset pairs in each modality.

B.2 Configuration and Training Details

B.2.1 Training protocol and K-JEPA selection

For each host method, we preserve its faithful training configuration from the corresponding paper, including the native optimizer, learning-rate schedule, weight decay, EMA behavior, masking strategy, and method-specific objective parameters. Baseline and K-JEPA variants therefore use the same host architecture and optimization configuration, with K-JEPA adding only the kernel-feature prediction branch described in Section 3. The corresponding host methods are described in their original works (Assran et al., 2023; Kuang et al., 2026; Balestriero & LeCun, 2025; Zimmermann et al., 2025; Tuncay et al., 2025; Yuksel et al., 2025; Gu et al., 2026; Thimonier et al., 2025; Bardes et al., 2024; Oji & Fieguth, 2026; Chen et al., 2025; Huang, 2026; Nam et al., 2026).

K-JEPA-specific model selection uses validation set aggregated score (described in Appendix B.1) over the kernel family. For the image, tabular, and SM-MNIST experiments, we consider RBF, Matérn-$1/2$, Matérn-$3/2$, and rational-quadratic kernels. The audio experiments use the fixed RBF configuration due to computational limits. All remaining K-JEPA-specific parameters are held fixed: the random-feature dimension is $D = 256$, the weight of the kernelized loss term is $\rho = 0.01$, the kernel length scale is $\ell = \sqrt{d}$, the rational-quadratic shape parameter is $\alpha = 1$, and the random-feature seed is 0. For image and audio models, we use the released pretrained weights specified below. We report the fixed epoch-5 checkpoint for the image experiments, use a fixed three-epoch adaptation budget for the audio experiments, and report the fixed epoch-20 checkpoint for the world model and tabular experiments. The SM-MNIST experiments use a five-frame context, one-step prediction horizon, and random seed 42.

B.2.2 Pretrained initialization

The image and audio hosts are initialized from publicly released pretrained weights rather than trained from scratch. For I-JEPA (Assran et al., 2023), we use the Birder checkpoint `birder-project/vit_reg4_m16_rms_avg_i-jepa` (`vit_reg4_m16_rms_avg_i-jepa.pt`). For Rectified LpJEPA (Kuang et al., 2026), we use the released $p = 1$, $\mu = 0.25$, 1000-epoch checkpoint from `YashDagade/rectified-lp-jepa`. For LeJEPA (Balestriero & LeCun, 2025), we use the 300-epoch ViT-B/16 checkpoint `OK-AI/lejepa-vitb16-pretrain-in1k`. For Audio-JEPA (Tuncay et al., 2025), we use the released `JEPA.ckpt` from `ltuncay/Audio-JEPA`. For WavJEPA (Yuksel et al., 2025), we use `labhamlet/wavjepa-base` (`model.safetensors`). For PupuJEPA (Gu et al., 2026), we use the released PupuJEPA-Base checkpoint `pupujepaV2_25hz_base` from `spellbrush/PupuJEPA`. This was the reason for the shorter training budgets of image and audio experiments. The training included controlled continuation/adaptation of pretrained representations rather than training from scratch. Exact checkpoint download locations and loading commands are provided with the supplementary code.

B.2.3 Host training configurations

Table 9 reports the host-specific training configurations used in the experiments. These settings are shared between each baseline and its corresponding K-JEPA variant.

C Extended Related Work

Kernel representations of conditional distributions. Kernel mean embeddings represent probability measures as elements of a reproducing kernel Hilbert space, with characteristic kernels yielding injective population embeddings of probability distributions (Sriperumbudur et al., 2010). Conditional mean embeddings extend this representation to conditional laws (Song et al., 2009; Grünewälder et al., 2012). Later measure-theoretic treatments define conditional mean embeddings directly as RKHS-valued conditional expectations (Park & Muandet, 2020) and neural-kernel conditional embeddings further combine learned representations with kernel conditional-distribution estimation to improve expressiveness and scalability (Shimizu et al., 2024). K-JEPA specifically applies this mechanism to latent targets within JEPA training: its auxiliary predictor estimates the conditional mean of fixed kernel features while the original JEPA objective continues to learn the

Table 9: **Host training configurations.** K-JEPA preserves the corresponding host configuration. Only the additional kernel-feature branch and the fixed K-JEPA settings described above are added.

Baseline	Training configuration
I-JEPA	Optimizer: AdamW; batch: 256; LR: $2\times10^{-4} \rightarrow 8\times10^{-4} \rightarrow 10^{-4}$; scheduler: warmup + cosine; warmup: 2.5%; WD: $0.04 \rightarrow 0.4$; EMA: $0.996 \rightarrow 1.0$; precision: BF16
Rectified LpJEPA	Optimizer: LARS; batch: 256; LR: $3\times10^{-5} \rightarrow 0.165 \rightarrow 0$; scheduler: warmup + cosine; WD: 10^{-4}; p: 1; μ: 0.25; random projections: 8192; invariance weight: 25; RDMReg weight: 125
LeJEPA	Optimizer: AdamW; batch: 256; LR: $0 \rightarrow 5\times10^{-4} \rightarrow 10^{-6}$; scheduler: linear warmup + cosine; warmup: 10%; WD: 0.05; layer decay: 0.9; patch LR multiplier: 0.2; gradient clip: 3; precision: BF16
KerJEPA	Optimizer: AdamW; batch: 256; peak LR: 5×10^{-4}; scheduler: cosine; WD: 0.05; views: 4; projector dim: 128; slices: 1024; quadrature points: 21; precision: BF16
T-JEPA	Optimizer: AdamW; batch: 512; LR: paper-tuned per dataset (10^{-5}–10^{-3}); scheduler: cosine to 0; EMA: $0.996 \rightarrow 1.0$; prediction masks: 4; architecture/masking: paper-tuned per dataset
Audio-JEPA	Optimizer: AdamW; betas: $(0.9, 0.95)$; batch: 256; LR: $10^{-6} \rightarrow 10^{-3} \rightarrow 0$; WD: 10^{-6}; EMA: $0.996 \rightarrow 1.0$; mask ratio: 0.4–0.6; precision: BF16
WavJEPA	Optimizer: AdamW; betas: $(0.9, 0.98)$; physical batch: 32; crops/audio: 8; effective batch: 256; LR: $0 \rightarrow 2\times10^{-4} \rightarrow 0$; scheduler: warmup + cosine; WD: 0.04; EMA: $0.999 \rightarrow 0.99999$; gradient clip: 5; precision: BF16
PupuJEPA	Optimizer: AdamW; betas: $(0.9, 0.95)$; batch: 256; LR: $10^{-6} \rightarrow 10^{-3} \rightarrow 10^{-6}$; WD: 0.05; EMA: $0.99995 \rightarrow 0.99999$; mask ratio: 0.8; gradient clip: 1; precision: BF16
V-JEPA	Optimizer: Adam; LR: 10^{-3}; WD: 0; batch: 256; EMA: 0.99; β: 10^{-2}; target variance: learned
VJE	Optimizer: SGD + Nesterov; LR: 5×10^{-2}; WD: 5×10^{-4}; scheduler: warmup + cosine; warmup: 10 epochs; ν: 1; β: 1; bottleneck ratio: 0.25; EMA: off
D-JEPA	Optimizer: AdamW; betas: $(0.9, 0.95)$; ϵ: 10^{-6}; global batch: 2048; LR: 8×10^{-4}; WD: $0.02 \rightarrow 0.2$; EMA: 0.9999; gradient clip: 3; SmoothL1 β: 2; diffusion multiplier: 4; precision: BF16
B-JEPA	Optimizer: Adam; LR: 10^{-3}; WD: 0; batch: 256; EMA: 0.99; β: 10^{-2}; γ: 10^{-1}; target variance: learned
C-JEPA	Predictor LR: 5×10^{-4}; action-encoder LR: 5×10^{-4}; proprio-encoder LR: 10^{-4}; WD: 10^{-2}; gradient accumulation: 1

underlying representation. Random Fourier features provide a scalable explicit approximation to shift-invariant kernels (Rahimi & Recht, 2007), although finite feature maps introduce approximation error and should not be conflated with the exact injectivity of the population characteristic kernel (Sutherland & Schneider, 2015).

Reliability-oriented evaluation. We utilize holistic reliability-oriented evaluation as probabilistic representation quality cannot be summarized by downstream accuracy or average prediction error alone. Proper scoring rules should evaluate the quality of an entire predictive distribution (Gneiting & Raftery, 2007) and calibration measures should be used to assess whether predictive probabilities agree with empirical outcomes (Guo et al., 2017). Continuous-target settings additionally require regression-specific calibration of predictive distributions or intervals (Kuleshov et al., 2018). Selective prediction evaluates whether uncertainty can identify unreliable examples through the risk–coverage trade-off (Geifman & El-Yaniv, 2019). Conformal prediction complements these evaluations by constructing prediction sets with finite-sample marginal coverage (Romano et al., 2020). Because aggregate reliability can conceal subgroup failures, prior work also studies subgroup disparities in selective decisions and conformal coverage (Lee et al., 2021; Zhou & Sesia, 2024). Under distribution shift, uncertainty and calibration may degrade substantially, motivating standardized OOD evaluations with common datasets and scoring protocols (Ovadia et al., 2019; Yang et al., 2022). Finally, membership-inference studies show that self-supervised visual and speech encoders can leak information about their pretraining examples (Liu et al., 2021; Tseng et al., 2022; Zhu et al., 2024). We use this literature to define a unified audit of K-JEPA; it provides evaluation criteria rather than fairness, robustness, or privacy guarantees implied by kernelization itself.

D Additional Results

D.1 MIMIC-IV Classification and Fairness

We additionally evaluate the JEPA variants on MIMIC-IV classification. To remain consistent with the downstream evaluation reported for the tabular experiments in the main paper, we report classification performance together with uncertainty calibration and selective prediction safety. Specifically, we report accuracy, NLL, Brier score, ECE, AURC, and risk at 50% coverage (Risk50). Most

Table 10: **MIMIC-IV classification and fairness results.** Accuracy measures downstream predictive performance; NLL, Brier, and ECE measure uncertainty calibration; AURC and Risk50 measure selective prediction safety. Agg. Age, Agg. Gender, Agg. Race, and Agg. Inter. denote the aggregated fairness score evaluated with respect to each protected attribute or their intersection. Arrows indicate the preferred direction.

		Performance	Uncertainty Calibration			Selective Safety		Fairness			
Model	**Method**	**Acc. ↑**	**NLL ↓**	**Brier ↓**	**ECE ↓**	**AURC ↓**	**Risk50 ↓**	**Agg. Age ↑**	**Agg. Gender ↑**	**Agg. Race ↑**	**Agg. Inter. ↑**
Rectified LpJEPA	Baseline	89.9	0.29	0.17	1.1	4.8	4.3	0.44	0.68	0.32	0.28
	+ K-JEPA	**90.2**	**0.28**	**0.16**	**0.6**	**3.8**	**3.1**	**0.85**	**0.81**	**0.34**	**0.29**
KerJEPA	Baseline	90.7	0.26	0.15	0.7	3.0	2.3	0.66	0.93	**0.49**	**0.33**
	+ K-JEPA	**90.8**	**0.25**	0.15	**0.6**	**2.8**	**2.1**	**0.72**	**0.94**	0.48	0.32
T-JEPA	Baseline	91.0	0.25	0.14	1.2	2.9	2.1	**0.81**	0.93	0.46	0.32
	+ K-JEPA	**91.1**	0.25	0.14	**1.1**	**2.7**	**1.9**	0.71	**0.94**	**0.57**	**0.34**

Table 11: **SM-MNIST downstream future-prediction results (future-state prediction; full statistical results).** The full future-state evaluation underlying Table 7 is reported separately for normalized state, position, and velocity. Panel (a) reports MAE and RMSE, while panel (b) reports CVaR95 and Worst10 tail-risk measures. Lower values are better throughout.

(a) Performance

		Normalized State		Position		Velocity	
Host	**Method**	**MAE ↓**	**RMSE ↓**	**MAE ↓**	**RMSE ↓**	**MAE ↓**	**RMSE ↓**
C-JEPA	Baseline	0.38 [0.38, 0.39]	0.48 [0.47, 0.48]	8.16 [8.11, 8.22]	9.56 [9.41, 9.85]	2.03 [2.01, 2.04]	2.39 [2.37, 2.41]
	+ K-JEPA	**0.36 [0.36, 0.36]*****	**0.45 [0.45, 0.45]*****	**7.34 [7.29, 7.39]*****	**8.69 [8.63, 8.74]*****	**1.93 [1.91, 1.94]*****	**2.30 [2.28, 2.31]*****
V-JEPA	Baseline	0.40 [0.40, 0.40]	0.48 [0.48, 0.49]	8.51 [8.47, 8.57]	9.73 [9.68, 9.77]	2.08 [2.07, 2.10]	2.44 [2.43, 2.45]
	+ K-JEPA	**0.36 [0.35, 0.36]*****	**0.46 [0.46, 0.46]*****	**6.26 [6.19, 6.30]*****	**7.67 [7.59, 7.71]*****	**2.03 [2.02, 2.05]*****	**2.40 [2.39, 2.41]*****
D-JEPA	Baseline	0.36 [0.35, 0.36]	0.45 [0.45, 0.45]	7.00 [6.95, 7.05]	8.32 [8.27, 8.38]	1.94 [1.93, 1.95]	2.30 [2.29, 2.31]
	+ K-JEPA	**0.34 [0.34, 0.35]*****	**0.44 [0.44, 0.44]*****	**6.59 [6.54, 6.64]*****	**7.95 [7.90, 8.01]*****	**1.89 [1.88, 1.91]*****	**2.26 [2.25, 2.27]*****
B-JEPA	Baseline	0.40 [0.40, 0.40]	0.49 [0.48, 0.49]	8.51 [8.46, 8.56]	9.72 [9.68, 9.77]	2.08 [2.07, 2.10]	2.44 [2.43, 2.45]
	+ K-JEPA	**0.40 [0.39, 0.40]*****	**0.48 [0.48, 0.49]*****	**8.47 [8.42, 8.53]*****	**9.69 [9.65, 9.74]*****	**2.08 [2.07, 2.10]*****	**2.44 [2.43, 2.45]*****
VJE	Baseline	0.40 [0.40, 0.40]	0.49 [0.48, 0.49]	8.55 [8.50, 8.60]	9.76 [9.72, 9.81]	2.08 [2.07, 2.10]	2.44 [2.43, 2.45]
	+ K-JEPA	**0.39 [0.39, 0.39]*****	**0.48 [0.48, 0.48]*****	**8.12 [8.07, 8.17]*****	**9.33 [9.28, 9.38]*****	**2.08 [2.07, 2.09]*****	**2.44 [2.42, 2.45]*****

(b) Tail Risk

		Normalized State		Position		Velocity	
Host	**Method**	**CVaR95 ↓**	**Worst10 ↓**	**CVaR95 ↓**	**Worst10 ↓**	**CVaR95 ↓**	**Worst10 ↓**
C-JEPA	Baseline	0.69 [0.67, 0.71]	0.65 [0.64, 0.66]	14.26 [13.73, 15.03]	13.31 [13.04, 13.74]	3.61 [3.52, 3.67]	3.41 [3.36, 3.45]
	+ K-JEPA	**0.66 [0.65, 0.66]*****	**0.63 [0.62, 0.63]*****	**13.21 [13.07, 13.39]*****	**12.53 [12.42, 12.65]*****	**3.50 [3.47, 3.53]*****	**3.32 [3.30, 3.35]*****
V-JEPA	Baseline	0.66 [0.66, 0.66]	0.64 [0.63, 0.64]	13.47 [13.38, 13.55]	12.98 [12.91, 13.05]	3.48 [3.46, 3.51]	3.35 [3.34, 3.37]
	+ K-JEPA	**0.65 [0.64, 0.65]*****	**0.62 [0.62, 0.62]*****	**12.81 [12.67, 12.93]*****	**12.02 [11.91, 12.11]*****	3.49 [3.46, 3.51]	**3.34 [3.32, 3.36]*****
D-JEPA	Baseline	0.64 [0.64, 0.65]	0.62 [0.61, 0.62]	12.84 [12.73, 12.96]	12.18 [12.09, 12.27]	3.44 [3.41, 3.47]	3.28 [3.26, 3.30]
	+ K-JEPA	0.64 [0.64, 0.65]	**0.61 [0.61, 0.62]*****	**12.70 [12.59, 12.82]*****	**12.02 [11.92, 12.11]*****	3.44 [3.41, 3.47]	**3.27 [3.25, 3.29]****
B-JEPA	Baseline	0.66 [0.66, 0.66]	0.64 [0.63, 0.64]	13.48 [13.39, 13.56]	12.99 [12.92, 13.06]	3.49 [3.46, 3.51]	3.36 [3.34, 3.37]
	+ K-JEPA	0.66 [0.66, 0.66]	**0.64 [0.63, 0.64]***	**13.44 [13.36, 13.53]*****	**12.96 [12.89, 13.03]*****	3.48 [3.46, 3.51]	3.35 [3.33, 3.37]
VJE	Baseline	0.66 [0.66, 0.66]	0.64 [0.63, 0.64]	13.49 [13.41, 13.58]	13.01 [12.94, 13.08]	3.49 [3.46, 3.51]	3.36 [3.34, 3.37]
	+ K-JEPA	**0.66 [0.65, 0.66]*****	**0.63 [0.63, 0.64]*****	**13.21 [13.12, 13.29]*****	**12.67 [12.60, 12.75]*****	**3.48 [3.46, 3.51]****	**3.35 [3.33, 3.37]*****

importantly, we report demographic fairness using aggregated fairness scores on age, gender, race, and their intersection. Higher values are better for accuracy and all aggregated fairness scores, while lower values are better for NLL, Brier, ECE, AURC, and Risk50.

D.2 SM-MNIST Statistical Uncertainty and Significance

Here, we provide the full statistical results for future-state prediction, stochastic-future prediction, and the controlled future-prediction evaluation on SM-MNIST. The main-text SM-MNIST results in Tables 7 and 8 report the corresponding point estimates.

Each entry reports the point estimate followed by its 95% bootstrap confidence interval in brackets. Confidence intervals are computed using 1,000 bootstrap resamples. Statistical significance of the paired K-JEPA improvement is assessed using one-sided paired permutation tests with 999 permutations. Significance marks denote favorable K-JEPA effects: $^{***}p < 0.01$, $^{**}p < 0.05$, and $^{*}p < 0.10$.

Table 12: **SM-MNIST downstream future-prediction results (stochastic-future prediction; full statistical results).** This table expands the stochastic-future portion of Table 7. Accuracy, Macro-F1, and AUPRC measure standard predictive performance; Brier and ECE measure uncertainty calibration; and Err-AUC and AURC measure selective prediction quality. Arrows indicate the preferred direction.

		Performance			Uncertainty Calibration		Selective Safety	
Host	**Method**	**Acc.** ↑	**Macro-F1** ↑	**AUPRC** ↑	**Brier** ↓	**ECE** ↓	**Err-AUC** ↑	**AURC** ↓
C-JEPA	Baseline	75.0 [74.2, 76.0]	42.9 [42.6, 43.2]	26.3 [24.9, 27.8]	0.37 [0.37, 0.38]	0.5 [0.2, 1.5]	51.4 [49.9, 52.8]	24.4 [23.0, 25.6]
	+ K-JEPA	75.0 [74.1, 76.0]	42.9 [42.6, 43.2]	**28.8 [27.5, 30.8]*****	**0.37 [0.36, 0.38]*****	0.7 [0.4, 1.8]	**55.6 [54.3, 57.2]*****	**21.8 [20.6, 23.1]*****
V-JEPA	Baseline	75.0 [74.1, 76.0]	43.1 [42.7, 43.4]	26.5 [25.3, 27.7]	0.37 [0.36, 0.38]	1.7 [0.7, 2.6]	51.4 [50.4, 52.4]	24.3 [22.7, 25.3]
	+ K-JEPA	74.9 [74.0, 75.9]	**51.9 [50.8, 53.1]*****	**41.0 [39.0, 43.2]*****	**0.34 [0.33, 0.35]*****	1.4 [1.2, 2.8]	**70.0 [68.8, 71.4]*****	**14.0 [13.1, 14.9]*****
D-JEPA	Baseline	76.3 [75.4, 77.2]	57.1 [55.5, 58.2]	46.7 [44.7, 49.4]	0.33 [0.32, 0.33]	1.1 [0.9, 2.2]	70.9 [70.0, 72.6]	12.9 [11.8, 13.6]
	+ K-JEPA	**76.4 [75.8, 77.6]****	**58.8 [58.0, 60.6]*****	**48.3 [46.5, 51.3]*****	**0.32 [0.31, 0.33]*****	1.4 [1.1, 2.5]	72.2 [70.5, 73.1]	**12.3 [11.4, 13.1]*****
B-JEPA	Baseline	75.0 [74.2, 76.0]	42.9 [42.6, 43.2]	25.8 [24.7, 26.9]	0.37 [0.37, 0.39]	1.5 [0.6, 2.5]	51.5 [50.7, 52.3]	23.5 [22.1, 24.7]
	+ K-JEPA	75.0 [74.1, 76.0]	**43.5 [43.1, 44.0]*****	**27.6 [26.2, 29.1]*****	**0.37 [0.36, 0.38]*****	1.7 [0.9, 2.8]	**53.7 [52.4, 54.8]****	22.6 [21.3, 23.7]
VJE	Baseline	75.0 [74.2, 76.0]	42.9 [42.6, 43.2]	24.9 [24.0, 25.9]	0.37 [0.37, 0.39]	1.5 [0.5, 2.5]	50.0 [50.0, 50.0]	25.4 [23.5, 26.3]
	+ K-JEPA	75.0 [74.2, 76.0]	**43.6 [43.2, 44.3]*****	**32.9 [31.2, 34.9]*****	**0.37 [0.36, 0.37]*****	1.6 [1.0, 2.7]	**60.8 [59.4, 62.3]*****	**18.9 [17.7, 20.1]*****

Table 13: **Controlled SM-MNIST future-prediction results (full statistical results).** This table expands Table 8. R@10 and MRR measure retrieval performance; Brier, NLL, and ECE measure uncertainty calibration; AURC and Risk25 measure selective prediction safety; and Cov95 measures conformal coverage. Each entry reports the point estimate followed by its 95% bootstrap confidence interval in brackets. Significance marks denote favorable K-JEPA effects under the one-sided paired permutation test: $^{***}p < 0.01$, $^{**}p < 0.05$, and $^{*}p < 0.10$. Arrows indicate the preferred direction.

		Performance		Uncertainty Calibration			Selective Safety		Conformal
Host	**Method**	**R@10** ↑	**MRR** ↑	**Brier** ↓	**NLL** ↓	**ECE** ↓	**AURC** ↓	**Risk25** ↓	**Cov95 (→95)**
C-JEPA	Baseline	9.2 [7.9, 10.2]	4.8 [4.3, 5.2]	0.99 [0.99, 0.99]	4.85 [4.85, 4.85]	0.4 [0.0, 0.9]	98.6 [97.9, 99.3]	98.2 [97.2, 99.2]	96.9 [96.1, 97.6]
	+ K-JEPA	9.3 [8.1, 10.4]	5.1 [4.6, 5.6]	0.99 [0.99, 0.99]	4.85 [4.85, 4.85]	0.5 [0.1, 1.0]	98.5 [97.8, 99.1]	98.0 [96.8, 99.2]	95.2 [94.4, 96.3]
V-JEPA	Baseline	8.6 [6.1, 11.1]	4.6 [3.6, 5.8]	0.99 [0.99, 0.99]	4.85 [4.85, 4.85]	0.2 [0.0, 1.3]	98.4 [96.2, 99.9]	97.9 [94.8, 100.0]	98.4 [97.1, 99.5]
	+ K-JEPA	10.9 [8.4, 14.3]	**6.0 [4.7, 7.5]****	**0.99 [0.99, 0.99]****	4.84 [4.81, 4.87]	0.5 [0.0, 1.8]	95.3 [91.0, 98.6]	93.8 [88.5, 97.9]	95.8 [93.2, 97.4]
D-JEPA	Baseline	6.7 [4.8, 9.4]	3.6 [3.2, 5.0]	0.99 [0.99, 0.99]	4.94 [4.89, 4.96]	1.7 [0.4, 1.7]	100.0 [98.3, 100.0]	100.0 [96.9, 100.0]	97.1 [96.1, 99.0]
	+ K-JEPA	8.0 [5.9, 10.9]	4.6 [3.6, 5.5]	**0.99 [0.99, 0.99]***	**4.88 [4.86, 4.91]****	0.9 [0.1, 1.4]	99.8 [97.4, 100.0]	100.0 [96.9, 100.0]	97.9 [96.6, 99.5]
B-JEPA	Baseline	6.7 [4.6, 9.0]	4.5 [3.5, 5.5]	0.99 [0.99, 0.99]	4.87 [4.86, 4.88]	1.7 [0.9, 2.3]	99.6 [98.9, 100.0]	100.0 [100.0, 100.0]	99.5 [98.7, 100.0]
	+ K-JEPA	**10.7 [8.0, 13.4]****	**6.2 [5.0, 7.6]****	**0.99 [0.99, 0.99]*****	**4.84 [4.83, 4.85]*****	**0.4 [0.1, 1.8]***	98.8 [97.4, 99.9]	99.0 [96.9, 100.0]	99.7 [99.0, 100.0]
VJE	Baseline	9.2 [6.7, 11.9]	5.1 [4.1, 6.2]	0.99 [0.99, 0.99]	4.86 [4.84, 4.89]	0.7 [0.1, 1.2]	99.7 [99.1, 100.0]	100.0 [100.0, 100.0]	91.6 [89.0, 94.0]
	+ K-JEPA	10.7 [8.0, 13.4]	5.5 [4.3, 6.8]	**0.99 [0.99, 0.99]****	**4.86 [4.83, 4.88]****	0.1 [0.1, 1.4]	**98.2 [96.1, 99.8]****	97.9 [94.8, 100.0]	92.7 [90.1, 95.0]