A GENERALIZATION THEORY FOR JEPA-BASED WORLD MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Joint Embedding Predictive Architectures (JEPAs) have emerged as a promising paradigm for world modeling by predicting future representations rather than reconstructing observations. Despite their empirical success, the theoretical understanding of JEPA-based world models remains limited. In this paper, we develop the first generalization theory for JEPA-based world models. We formulate JEPA pretraining as a conditional spectral graph learning problem and show that the JEPA objective is equivalent to a low-rank factorization of an action-conditioned co-occurrence matrix. Building on this characterization, we establish a connection between JEPA pretraining error and downstream planning regret, leading to a finite-sample generalization bound for JEPA-based world models. Our analysis reveals an approximation–sample error trade-off governed by latent dimensionality, providing insights into latent- versus input-level prediction. The theoretical trade-off motivates us to investigate a simple Multi-Depth JEPA extension that combines representations from different network depths. In validation experiments, latent-level prediction exhibits greater robustness under unpredictable observations and difficult planning tasks. Multi-depth representations further improve planning success over single-depth representations and a dimensionality-matched control.

1 INTRODUCTION

Recent progress in self-supervised learning has increasingly shifted the focus of representation learning from discriminative objectives toward predictive world modeling. Among the emerging paradigms, Joint Embedding Predictive Architectures (JEPAs) (LeCun et al., 2022) excel by predicting latent-level information while avoiding the inefficiency of input-level generation. This perspective has motivated a rapidly growing line of research on predictive latent world models, including V-JEPA (Assran et al., 2025), VL-JEPA (Chen et al., 2025), LeWM (Maes et al., 2026), etc.

Despite the growing empirical success of JEPA world models, their theoretical understanding remains limited. Existing theoretical studies on JEPAs have primarily focused more on understanding JEPAs as a framework for latent space learning. Littwin et al. (2024) analyzed that JEPAs preferentially learn high-influence semantic features while suppressing noisy or weakly informative signals. Balestriero et al. (2025) reveals that JEPA anti-collapse regularization implicitly performs density estimation of the input observations. Very recently, Klindt et al. (2026) proved the identifiability of the Gaussian-regularized JEPA world model. Nonetheless, these works either studied JEPA in the static representation learning setting or based the theoretical analysis on parametric modeling. To the best of our knowledge, there is currently no theoretical framework explaining how JEPAs generalize as a world model framework in real-world action planning. Since actions are planned in latent space but evaluated in downstream tasks, understanding how JEPA pretraining error affects planning performance is essential. Moreover, it remains unclear how latent dimensionality affects the trade-off between predictive capacity and finite-sample generalization.

In this paper, we establish the first generalization theory for JEPA-based world models. Based on a conditioned spectral graph formulation, we establish the equivalence between the JEPA pretraining risk and matrix factorization of the co-occurrence matrix. Then by establishing the relationship between the downstream action planning regret and the pretrained JEPA risk, we derive the generalization error bound for JEPA-based world models. Our bounds reveal an inherent approximation–sample error trade-off with respect to latent dimensionality. This trade-off provides insights into

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.

the advantages and limitations of latent- versus input-level prediction, while also motivating the design of representations that preserve predictive information without unnecessarily increasing model complexity.

Our contributions are as follows.

- We for the first time establish a spectral graph-based theoretical framework for JEPA-based world models, where we propose a conditioned co-occurrence matrix formulating the co-occurrence probability of the current and next state conditioned on the action, and show that the spectral JEPA objective is equivalent to a low-rank factorization of this matrix.
- We establish a connection between JEPA pretraining error and downstream planning regret in both single- and multi-step settings, leading to finite-sample generalization bounds. Our analysis reveals an inherent approximation–sample error trade-off with respect to latent dimensionality, providing theoretical insights into latent- versus input-level prediction.
- Motivated by the theoretical trade-off, we investigate a simple Multi-Depth JEPA extension that combines representations from different network depths. Through validation experiments, we demonstrate the robustness of latent-level prediction under unpredictable observations and difficult planning tasks, and show that multi-depth representations can improve planning success.

2 Related Works

JEPA and JEPA-based world models. Joint-Embedding Predictive Architectures (JEPAs) learn predictive representations without reconstructing raw observations. I-JEPA (Assran et al., 2023) applies this paradigm to images, while V-JEPA (Bardes et al., 2023) extends it to videos through masked spatio-temporal prediction. Recent advances in world modeling include V-JEPA 2 (Assran et al., 2025), which learns action-conditioned dynamics for robotic planning, DINO-WM (Zhou et al., 2024), which predicts future visual features for zero-shot planning, and LeWorldModel (Maes et al., 2026), which learns a JEPA-style world model directly from pixels. Recent work has systematically examined the effects of architecture, training objectives, and planning algorithms on JEPA-based world models (Terver et al., 2026). Other studies improve latent-space planning through goal-conditioned value learning (Destrade et al., 2025), hierarchical subgoal prediction (Masip et al., 2026), and temporal progress supervision (Bai & Xiong, 2026).

Despite the empirical success, theoretical guarantees of the JEPA-based world models are largely underexplored. Existing theoretical studies have investigated representation learning and regularization in JEPAs, including Gaussian-regularized objectives (Balestriero et al., 2025; Klindt et al., 2026). More recent work has examined the inductive bias of object-level latent masking (Nam et al., 2026) and anti-collapse properties of unified predictive objectives (Lanji et al., 2026). However, these studies do not directly characterize the finite-sample relationship between JEPA pretraining error and downstream action planning regret considered in our work.

Spectral graph theory for representation learning. Spectral graph theory was introduced to self-supervised representation learning by HaoChen et al. (2021), who build generalization guarantees for self-supervised contrastive learning by formulating the similarity of the augmented data through the concept of augmentation graph. The theoretical framework was later extended to other representation learning paradigms, including unsupervised domain adaptation (Shen et al., 2022), multimodal learning (Zhang et al., 2023), weakly supervised learning (Cui et al., 2023), autoregressive and masked self-supervised learning (SSL) (Zhang et al., 2024), and SSL with difficult examples (Zhang et al., 2026). Recently, Balestriero & LeCun (2026) further brought forward a spectral graph theory of SSL that relies on harmonic analysis and spectral graph theory. Despite the theoretical research on representation learning, spectral graph theory has not yet been explored in JEPA-based world models which make predictions in the representation space.

3 Mathematical Formulations

In this section, we formulate JEPA pretraining and latent-space action planning, define downstream planning regret, and introduce the finite-sample setting for our generalization analysis.

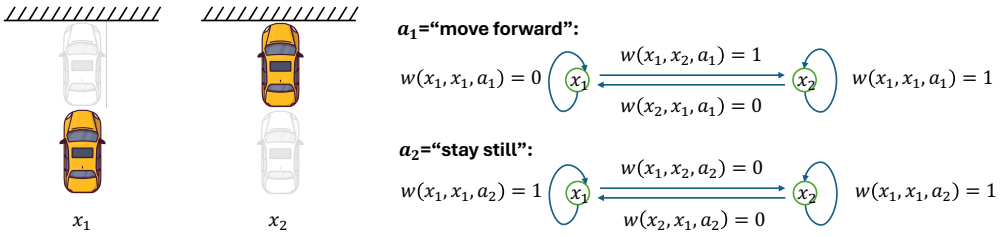

Figure 1: An illustrative example of the graph relationship between world model inputs. (Left) Observations of a car-moving example, where x_1 and x_2 represent two possible states of car locations. (Right) The transition probabilities from x_i to x_j, $i, j \in \{1, 2\}$ are conditioned on the actions a_1 and a_2. The conditioned co-occurrence matrix $M(a)$ is formed by assembling $w(x_i, x_j, a)$.

3.1 Pretraining

JEPA pretraining objective. In the pretraining stage of JEPA, world models are learned by forecasting future latent representations instead of reconstructing raw observations. Given a current observation $x \in \mathcal{X} := \mathbb{R}^d$ and action $a \in \mathcal{A}$, an encoder $f : \mathbb{R}^d \to \mathbb{R}^k$ maps the observation into a latent state $z = f(x)$, while a predictor $g : \mathbb{R}^k \times \mathcal{A} \to \mathbb{R}^k$ estimates the latent representation of a future observation $x^+ \in \mathcal{X}$. The training objective minimizes the discrepancy between the predicted latent representation $\hat{z}^+ = g(f(x), a)$ and the target latent embedding $z^+ = f(x^+)$, i.e.,

$$\mathcal{L}_{\text{JEPA}}(x, x^+; f, g, a) = \|g(f(x), a) - f(x^+)\|^2. \tag{1}$$

Note that the pretraining loss in equation 1 alone would leads to representation collapse. To prevent this, the JEPA-based models either use uniformity regularizations (Balestriero & LeCun, 2025; Maes et al., 2026) or techniques such as stop gradient and exponential moving average (Assran et al., 2025).

In this paper, for the ease of theoretical analysis, we adopt a uniformity term inspired by spectral contrastive learning (HaoChen et al., 2021) [1], and define the pretraining loss as

$$\mathcal{L}_{\text{JEPA}} = \|g(f(x), a) - f(x^+)\|^2 + \left[g(f(x), a)^\top f(x')\right]^2. \tag{2}$$

For population risk, we assume the (x, x^+) pairs are generated as follows: 1) sample $x \sim \mathrm{P}_X$ from $\mathbb{R}^d$, and 2) given x and $a \in \mathcal{A}$, sample $x^+ \sim \mathrm{P}(\cdot|x, a)$. Then given action a, we define the pretraining risk as

$$\mathcal{R}_{\text{JEPA}}(f, g, a) = \mathbb{E}_{x\sim \mathrm{P}_X} \mathbb{E}_{x^+\sim \mathrm{P}(\cdot|x,a)} \|g(f(x), a) - f(x^+)\|^2 + \mathbb{E}_{\substack{x\sim \mathrm{P}_X \\ x'\sim \mathrm{P}_{X|a}}} \left[g(f(x), a)^\top f(x')\right]^2. \tag{3}$$

Then given action $a \in \mathcal{A}$, we denote the optimal encoder and predictor under the population pretraining risk as

$$(f^*, g^*) = \arg\min_{f,g} \mathcal{R}_{\text{JEPA}}(f, g, a). \tag{4}$$

Conditional co-occurrence matrix and matrix factorization. Inspired by previous works that introduced spectral graph theory to representation learning (HaoChen et al., 2021; Zhang et al., 2024), we here introduce a conditional co-occurrence matrix that formulates the co-occurrence relationship between (x, x^+) pairs conditioned on the actions.

Given action $a \in \mathcal{A}$, for $x, x^+ \in \mathbb{R}^d$, we denote $w(x, x^+, a) := \mathrm{P}(x, x^+|a) \geq 0$ and denote the co-occurrence matrix

$$M(a) := \left(w(x, x^+, a)\right)_{x,x^+\in\mathbb{R}^d}. \tag{5}$$

[1] We empirically evaluate the effect of the additional uniformity term in Appendix A.3, and observe modest but consistent improvements across all different environments.

Note that different from the adjacency matrix formulated in HaoChen et al. (2021), our conditional co-occurrence matrix $M(a)$ is asymmetric because typically we have $w(x, x^+, a) \neq w(x^+, x, a)$. Also, compared with the co-occurrence matrix proposed in Zhang et al. (2024), our conditional co-occurrence matrix has an additional conditional relationship on the action variable $a \in \mathcal{A}$. In Figure 1, we illustrate how a conditioned co-occurrence matrix $M(a)$ is formulated by a two-state car-moving toy example.

We denote the marginal probabilities as $w(x) := \sum_{x^+ \in \mathcal{X}} w(x, x^+, a)$ and $w(x^+|a) := \sum_{x \in \mathcal{X}} w(x, x^+, a)$. Then according to the data generation process, we have $w(x) = \mathrm{P}_X(x)$ and $w(x^+|a) = \mathrm{P}_{X|a}(x^+)$. Moreover, we denote the normalized co-occurance matrix as $\bar{M}(a) := D^{-1/2} M(a) D_+^{-1/2}(a)$, where $D := \mathrm{diag}(w(x))_{x \in \mathcal{X}}$ and $D_+(a) := \mathrm{diag}(w(x^+|a))_{x^+ \in \mathcal{X}}$.

In Theorem 3.1, we derive the equivalence between JEPA risk and the matrix factorization of $\bar{M}(a)$.

Theorem 3.1 (Equivalence between JEPA and matrix factorization)**.** *Given $a \in \mathcal{A}$, for normalized embedding functions $f(x)$ and $g(f(x), a)$, we have*

$$\mathcal{R}_{\mathrm{JEPA}}(f, g, a) = \|\bar{M}(a) - G(F, a)^\top F\|^2 + \mathrm{const.}, \tag{6}$$

where $F := \left[\sqrt{w(x|a)} f(x)\right]_{x \in \mathcal{X}}$ and $G(F, a) := \left[\sqrt{w(x)} g(f(x), a)\right]_{x \in \mathcal{X}}$.

In the following sections, for notational simplicity, we denote the matrix factorization risk as $\mathcal{R}_{\text{S-JEPA}} := \|\bar{M}(a) - G(F, a)^\top F\|^2$.

3.2 Action Planning

In the stage of action planning, we perform trajectory optimization in the latent space. For a T-step planning, given an initial observation x_0, we randomly initialize a candidate action sequence $(a_0, \ldots, a_{T-1})$ and iteratively rollout predicted latent states up to a planning horizon. Action planning is performed by optimizing the action sequence to minimize the difference between the predicted latent and the latent of goal observation $z_g = f(x_g)$. Specifically, for $t = 0, \ldots, T-1$, given predictor g, we denote

$$\hat{z}_{t+1} = g(\hat{z}_t, a_t) \tag{7}$$

and

$$\mathcal{L}_{\mathrm{plan}}(a_0, \ldots, a_{T-1}; f, g, x_0, x_g) = \|\hat{z}_T - z_g\|^2 = \|g\left(\cdots g(f(x_0), a_0) \ldots, a_{T-1}\right) - f(x_g)\|^2. \tag{8}$$

Then given the optimal encoder f^* and predictor g^*, for an initial observation x_0, we denote the optimal action sequence as

$$(a_0^*, \ldots, a_{T-1}^*) = \arg \min_{a_0, \ldots, a_{T-1}} \mathcal{L}_{\mathrm{plan}}(a_0, \ldots, a_{T-1}; f^*, g^*, x_0, x_g). \tag{9}$$

3.3 Evaluation

Recall that in pretraining and action planning, the prediction is conducted in the latent space. By contrast, the planned actions are to be utilized in the input-space downstream tasks. Therefore, we evaluate the planned actions by measuring the predicted goal observation in the input space. Specifically, given initial observation x_0, we obtain the T-step interatively rollout prediction $\hat{x}_T$ by $\mathrm{P}(\hat{x}_{t+1}|\hat{x}_t, a)$ for $t = 1, \ldots, T-1$ and $\mathrm{P}(\hat{x}_1|x_0, a)$. By Bayes formula, we have $\mathrm{P}(\hat{x}_{t+1}|\hat{x}_t, a) = w(\hat{x}_t, \hat{x}_{t+1}, a)/w(\hat{x}_t)$. Then we have

$$\begin{aligned} V(a_0, \ldots, a_{T-1}) &= \mathrm{P}(\hat{x}_T = x_g | x_0, a_0, \ldots, a_{T-1}) \\ &= \sum_{x_1, \ldots, x_{T-1}} \mathrm{P}(\hat{x}_T = x_g | x_{T-1}, a_{T-1}) \cdots \mathrm{P}(\hat{x}_1 = x_1 | x_0, a_0) \\ &= \sum_{x_1, \ldots, x_{T-1}} \frac{w(x_{T-1}, x_g, a_{T-1})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, a_0)}{w(x_0)}. \end{aligned} \tag{10}$$

By denoting $a^{**} = \arg\max_a V(a)$, we define the pointwise regret as

$$r(a_0, \ldots, a_{T-1}) = V(a^{**}) - V(a). \tag{11}$$

Then we denote the expected planning regret as

$$\mathcal{E}(a) = \mathbb{E}_{x_0, x_g} r(a_0, \ldots, a_{T-1}). \tag{12}$$

3.4 FINITE-SAMPLE OPTIMA

To derive a finite-sample error bound for the JEPA framework, given $a \in \mathcal{A}$, we define the finite-sample optima of the encoder f and predictor g as

$$\hat{f}, \hat{g} = \arg\min_{f,g} \hat{\mathcal{R}}_S(f, g, a), \tag{13}$$

where $\hat{\mathcal{R}}_S(f, g, a)$ is an unbiased estimator of $\mathcal{R}(f, g, a)$ with details specified in Appendix A.1. Then given the initial observation x_0 and the goal observation x_g, we define the finite-sample optimal action sequence as

$$(\hat{a}_0, \ldots, \hat{a}_{T-1}) = \arg\min_{a_0,\ldots,a_{T-1}} \mathcal{L}_{\text{plan}}(a_0, \ldots, a_{T-1}; \hat{f}, \hat{g}, x_0, x_g). \tag{14}$$

4 GENERALIZATION ERROR BOUNDS

In this section, we derive the generalization error bounds for the downstream action planning error of JEPA-based world models.

4.1 EVALUATION ERROR FOR ACTION PLANNING

First, we study the evaluation error for single-step and multi-step action planning. We note that Theorems 4.1 and 4.2 are among the key theoretical results of this paper, revealing why an action series planned in the latent space generalizes well in the downstream evaluation on input-space samples.

Theorem 4.1 (Single-Step Planning)**.** *When $T = 1$, if we assume that $\mathbb{E}_{x_g} w(x_g|a) = \mathbb{E}_{x_g} w(x_g|a')$ for $a, a' \in \mathcal{A}$. For arbitrary f and g, if we define*

$$\tilde{a} = \arg\min_{a} \mathcal{L}_{\text{plan}}(a; f, g), \tag{15}$$

then we have

$$\mathcal{E}(\tilde{a}) \le 2c_0 \cdot \max_a \sqrt{\mathcal{R}_{\text{S-JEPA}}(f, g, a)}, \tag{16}$$

where $c_0 := \mathbb{E}_{x_0,x_g} \sqrt{\max_a w(x_g|a)/w(x_0)}$.

Theorem 4.1 shows that for single-step planning, the expected planning regret is upper bounded by the pretraining risk of S-JEPA. That is, if an encoder f and predictor g makes the pretraining risk small enough, then the action $\tilde{a}$ optimized under f and g is good enough for the downstream action planning task.

Theorem 4.2 (Multi-Step Planning)**.** *When $T > 1$, if we assume that $\mathbb{E}_{x_g} w(x_g|a) = \mathbb{E}_{x_g} w(x_g|a')$ for $a, a' \in \mathcal{A}$, and $M(a)$ is deterministic, i.e., given $a \in \mathcal{A}$ and $x \in \mathcal{X}$, for $x^+ \in \mathcal{X}$ there exists $w(x, x^+, a) = 1$ and $w(x, x', a) = 0$ for $x' \neq x^+$. Then for arbitrary f and g, by defining*

$$\tilde{a}_0, \ldots, \tilde{a}_{T-1} = \arg\min_{a_0,\ldots a_{T-1}} \mathcal{L}_{\text{plan}}(a_0, \ldots, a_{T-1}; f, g), \tag{17}$$

we have

$$\mathcal{E}(\tilde{a}_0, \ldots, \tilde{a}_T) \le 2Tc_3 \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)}, \tag{18}$$

where $c_3 := \max_{a,x_0,x_g} \sqrt{w(x_g|a)/w(x_0)}$.

In Theorem 4.2, we derive the evaluation error bound for multi-step action planning. Compared with Theorem 4.1, the T-step planning regret is approximately T times the single-step planning regret, explaining that longer planning horizon leads to higher planning error due to error accumulation. Also note that compared with Theorem 4.1, Theorem 4.2 requires a slightly stronger assumption that $\bar{M}(a)$ is deterministic. This assumption adheres to real world dynamic systems, where given a prior observation x_t and action a_t, the next observation x_{t+1} should be deterministically determined.

4.2 Error Analysis for JEPA Pretraining Risk

In the following, we analyze the approximation error and sample error for the S-JEPA pretraining risk respectively.

Theorem 4.3 (Approximation Error for Spectral JEPA Risk)**.** *Given $a \in \mathcal{A}$, the optimal population risk equals*

$$\mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) = \sum_{i>k} \sigma_i^2(a), \tag{19}$$

where $\sigma_i(a)$ is the i-th largest singular value of $\bar{M}(a)$.

In Theorem 4.3, we show that as the approximation error of S-JEPA risk is determined by the singular values of the conditioned co-occurrence matrix. As the latent dimension k increases, the approximation error decreases.

Theorem 4.4 (Sample Error Bound for Spectral JEPA Risk)**.** *Assume that $\|f\|_\infty < \kappa$ and $\|g\|_\infty < \kappa$ for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. Fix $a \in \mathcal{A}$. Then, with probability at least $1-\delta$, we have*

$$\begin{aligned}&\mathcal{R}_{\text{S-JEPA}}(\hat{f}, \hat{g}, a) - \mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) \\ &\le c_1 \left[\mathfrak{R}_{n/2}(\mathcal{G} \circ \mathcal{F}) + \mathfrak{R}_{n/2}(\mathcal{F})\right] + c_2 \left(\sqrt{\frac{\log(2/\delta)}{n}} + \delta\right),\end{aligned} \tag{20}$$

where $c_1 := 32k^2\kappa^3 + 32k\kappa$, $c_2 := 8k\kappa^2 + 2k^2\kappa^4$, and

$$\mathfrak{R}_n(\mathcal{G} \circ \mathcal{F}) = \max_{(z_i, z_i', z_i^+)_{i=1}^n} \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F},\, g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \sigma_i \Big(\big(g(f(z_i), a)^\top f(z_i')\big)^2 - 2g(f(z_i), a)^\top f(z_i^+) \Big)\right],$$

$$\mathfrak{R}_n(\mathcal{F}) = \max_k \max_{\{z_i\}_{i=1}^n} \mathbb{E}_\sigma \left[\sup_{f\in\mathcal{F},\, g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \sigma_i f_k(z_i)\right].$$

In Theorem 4.4, we show the sample error bound of S-JEPA risk, which depends on the Rademacher complexities $\mathfrak{R}_n(\mathcal{G} \circ \mathcal{F})$ and $\mathfrak{R}_n(\mathcal{F})$. Note that as the latent dimension k increases, both the output dimensions of $g \circ f \in \mathcal{G} \circ \mathcal{F}$ and $f \in \mathcal{F}$ increase, so typically we have larger Rademacher complexities and accordingly larger sample error.

4.3 Finite-Sample Error Bounds for Action Planning

Based on Theorems in the previous sections, we can now derive the finite-sample generalization error bound for action planning in Theorems 4.5 and 4.6.

Theorem 4.5 (Finite-Sample Error Bound for Single-Step Planning)**.** *For some $\kappa > 0$, assume $\|g(f(x), a)\|_\infty \le \kappa$ and $\|f(x)\|_\infty \le \kappa$ for all f, g, and a. Assume that all $\mathcal{R}_{n/2}(\mathcal{G} \circ \mathcal{F})$ are of the same order regardless of a. Then for $T = 1$, with probability at least $1-\delta$, we have*

$$\mathcal{E}(\hat{a}) \le 2c_0 \cdot \sqrt{\underbrace{\max_a \sum_{i>k} \sigma_i^2(a)}_{Approximaition\ Error} + \underbrace{c_1 \left[\mathfrak{R}_n(\mathcal{G} \circ \mathcal{F})\mathfrak{R}_n(\mathcal{F})\right] + c_2 \cdot \left(\sqrt{\frac{\log 2/\delta}{n}} + \delta\right)}_{Sample\ Error} + c}, \tag{21}$$

where $c_3 := \max_{a,x_0,x_g} \sqrt{w(x_g|a)/w(x_0)}$, $c_1 := 16k^2\kappa^2 + 16k^2\kappa$, $c_2 := 8k\kappa^2 + 2k^2\kappa^4$, and $c := 2 - \mathbb{E}_{x_0,x_g} \max_a \frac{w(x_0,x_g,a)^2}{w(x_0)w(x_g|a)}$.

In Theorem 4.5, we show the finite-sample error bound for single-step planning. Given sample size n, there is a trade-off between the approximation error term and sample error term with respect to the latent dimension k. That is, larger k leads to smaller $\max_a \sum_{i>k} \sigma_i^2(a)$ but larger c_1, c_2, and $\mathfrak{R}_n(\mathcal{G} \circ \mathcal{F}) + \mathfrak{R}_n(\mathcal{F})$. Moreover, as input-level predictive models can be viewed as special cases of the latent level ones, where $k = n$ and f degenerates to an identity mapping. In this case, we have zero approximation error but the largest sample error.

Theorem 4.6 (Finite-Sample Error Bound for Multi-Step Planning). *For some $\kappa > 0$, assume $\|g(f(x), a)\|_\infty \le \kappa$ and $\|f(x)\|_\infty \le \kappa$ for all f, g, and a, and $M(a)$ is deterministic. Assume that all $\mathcal{R}_{n/2}(\mathcal{G} \circ \mathcal{F})$ are of the same order regardless of a. Then for $T > 1$, with probability at least $1 - \delta$, we have*

$$\mathcal{E}(\hat{a}) \le 2c_3 T \cdot \sqrt{\underbrace{\max_a \sum_{i>k} \sigma_i^2(a)}_{Approximaition\ Error} + \underbrace{c_1 \left[\mathfrak{R}_n(\mathcal{G} \circ \mathcal{F}) + \mathfrak{R}_n(\mathcal{F})\right] + c_2 \cdot \left(\sqrt{\frac{\log 2/\delta}{n}} + \delta\right)}_{Sample\ Error}} + c, \tag{22}$$

where $c_0 := \mathbb{E}_{x_0, x_g} \sqrt{\max_a w(x_g|a)/w(x_0)}$, $c_1 := 16k^2\kappa^2 + 16k^2\kappa$, $c_2 := 8k\kappa^2 + 2k^2\kappa^4$, and $c := 2 - \mathbb{E}_{x_0, x_g} \max_a \frac{w(x_0, x_g, a)^2}{w(x_0) w(x_g|a)}$.

In Theorem 4.6, we show the finite-sample error bound for multi-step action planning. The error trade-off w.r.t. the latent dimension k is similar to the single-step planning case.

4.4 Approximation and Sample Error Trade-off

The above bounds reveal an intrinsic trade-off between the *Approximation Error* and the *Sample Error*, governed by the latent dimension k. On the one hand, a smaller latent dimension imposes a stronger information bottleneck, which reduces the complexity of the learned encoder and predictor and therefore leads to a smaller Sample Error. However, such compression may discard action-relevant transition information, resulting in a larger Approximation Error. On the other hand, increasing k allows the latent model to preserve more singular components of the action-conditioned co-occurrence matrix, thereby reducing the Approximation Error. This benefit comes at the cost of a larger hypothesis space and higher sample complexity, which increases the Sample Error. Therefore, latent predictive models are most advantageous when a moderate-dimensional representation can retain the task-relevant dynamics while filtering out nuisance or unpredictable input-level variations. In contrast, input-level predictive models can be viewed as an extreme case with minimal Approximation Error but potentially maximal Sample Error, since they are required to model all observation dimensions, including those irrelevant to downstream planning.

4.5 A Theory-Motivated Extension: Multi-Depth JEPA

The analysis above highlights the importance of the latent representation in balancing approximation and sample errors. While increasing the latent dimension can reduce the approximation error by preserving more components of the action-conditioned co-occurrence matrix, simply enlarging the representation does not necessarily introduce additional action-relevant information. This motivates us to investigate whether the representational capacity of a latent world model can be increased by incorporating complementary information from different depths of the model.

To this end, we introduce *Multi-Depth JEPA (MD-JEPA)*, which constructs the latent representation by concatenating intermediate representations from multiple layers of the world model. Let $f_l(x) \in \mathbb{R}^{k_l}$ denote the representation of observation x at layer l. Given a set of selected layers $\mathcal{L}$, we define the multi-depth latent representation as

$$f_{\mathcal{L}}(x) = [f_l(x)]_{l \in \mathcal{L}} \in \mathbb{R}^{k_{\mathcal{L}}}, \qquad k_{\mathcal{L}} = \sum_{l \in \mathcal{L}} k_l. \tag{23}$$

The action-conditioned predictor then operates on the concatenated representation and predicts the corresponding future representation:

$$\hat{z}^+_{\mathcal{L}} = g_{\mathcal{L}}\left(f_{\mathcal{L}}(x), a\right), \tag{24}$$

where the prediction target is given by $f_{\mathcal{L}}(x^+)$. Accordingly, the latent prediction objective becomes

$$\mathcal{L}^{\mathcal{L}}_{\mathrm{MD-JEPA}} = \left\| g_{\mathcal{L}}\left(f_{\mathcal{L}}(x), a\right) - f_{\mathcal{L}}(x^+) \right\|_2^2, \tag{25}$$

together with the same anti-collapse regularization as in the standard JEPA formulation.

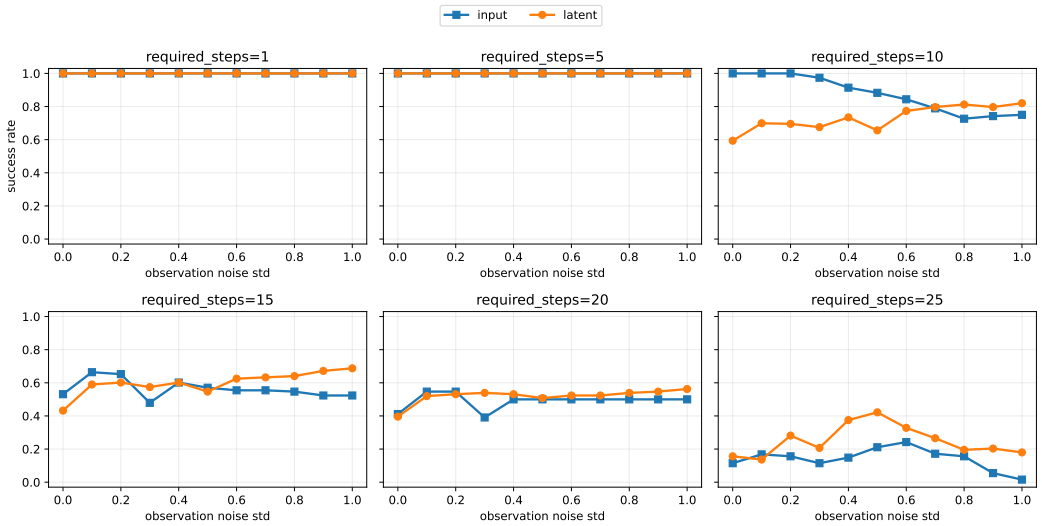

Figure 2: Success rates of latent- and input-level predictive models under different noise levels and required planning steps. Both approaches perform similarly on short and easy planning tasks, while latent-level prediction becomes more robust against increasingly unpredictable observations and more difficult long-horizon planning.

MD-JEPA provides an alternative way to increase the effective latent dimension. Rather than expanding a representation within a single layer, it incorporates features from different depths of the world model, which may contain complementary information relevant to predicting future states. Importantly, increasing the representation dimension alone does not imply that additional information is obtained.

5 Validation Experiments

We conduct two complementary experiments to validate the implications of our theoretical analysis. First, we use a fully controlled continuous-control environment to compare latent-level prediction with input-level reconstruction, and examine how their relative performance changes with observation noise and planning difficulty. Second, motivated by our analysis of latent representation capacity, we investigate whether incorporating representations from multiple model depths can further improve planning performance.

5.1 Latent- vs. Input-Level Prediction

Experimental setup. We construct a controlled 2D point-mass environment to compare latent-level prediction with input-level reconstruction. The underlying state is defined by the 2D position and velocity of the point mass, $s_t = [p_x, p_y, v_x, v_y]$, and the action $a_t = [a_x, a_y]$ specifies the applied acceleration along the two spatial dimensions. The model does not directly observe this state, but instead receives a higher-dimensional observation composed of task-relevant state features, nonlinear nuisance features, and random noise features that are independent of the action and downstream objective.

Both methods use the same action-conditioned predictive backbone and differ only in the prediction target: the latent-level model predicts the future representation z_{t+H}, whereas the input-level model predicts the future observation x_{t+H}. We use a fixed prediction horizon of $H = 3$. For downstream evaluation, we consider a goal-reaching task with receding-horizon CEM planning. At the beginning of each episode, we sample a start state and a goal state, where the goal specifies the target position to be reached and is provided to the planner through its corresponding observation x_{goal}. For latent-level planning, the goal is encoded as $z_{\text{goal}} = \text{Enc}(x_{\text{goal}})$. The latent-level planner evaluates candidate action sequences in latent space, whereas the input-level planner evaluates them in observation space.

Table 1: Planning success rates (%) of different latent representations at different horizons. Layer 12 $\times 2$ serves as a dimensionality-matched control for MD-JEPA.

Representation	$H = 25$	$H = 50$	$H = 75$
Layer 12	64.0	24.7	6.7
Layer 12 $\times$ 2	68.0	23.3	8.0
MD-JEPA $[6, 12]$	74.7	**33.3**	**14.0**
MD-JEPA $[3, 12]$	**77.3**	24.7	10.0
MD-JEPA $[9, 12]$	66.7	28.0	10.0

In Figure 2, we vary two factors. The *noise level* controls the contribution of unpredictable noise dimensions to the input reconstruction objective, while the *required planning steps* indicates approximately how many closed-loop planning decisions are needed to reach the goal. Larger values therefore correspond to stronger interference from unpredictable observations and more difficult long-horizon planning, respectively. Additional environment, training, and planning details are provided in Appendix A.4.

Results. Figure 2 shows that the two approaches perform similarly on short and easy planning tasks. As the required number of planning steps and the influence of unpredictable observation noise increase, latent-level prediction becomes increasingly more robust. The difference is most pronounced at 25 required planning steps, where input-level prediction degrades substantially while latent-level prediction maintains noticeably higher success rates.

These observations are consistent with our theoretical analysis: latent prediction can suppress unpredictable observation components that are irrelevant to downstream control, and this advantage becomes more important as prediction errors accumulate over repeated long-horizon planning steps.

5.2 Empirical Evaluation of Multi-Depth Representations

Experimental setup. We evaluate MD-JEPA introduced in Section 4.5 on PushT using DINO-WM with a ViT backbone, considering different combinations of intermediate representations. We use Layer 12 as the standard single-depth baseline and consider three multi-depth variants, $3, 12$, $6, 12$, and $9, 12$. To control for the increase in representation dimensionality, we additionally include Layer 12 $\times 2$, which duplicates the Layer-12 representation without introducing information from another depth. All other training and planning configurations are kept unchanged. We evaluate planning performance at horizons $H = 25$, 50, and 75.

Results. As shown in Table 1, MD-JEPA improves planning performance over the standard Layer-12 representation across different horizons. At $H = 25$, the $[3, 12]$ combination achieves the highest success rate of 77.3%, compared with 64.0% for Layer 12. At longer horizons, $[6, 12]$ performs best, improving the success rate from 24.7% to 33.3% at $H = 50$ and from 6.7% to 14.0% at $H = 75$.

The dimensionality-matched control further shows that the gains cannot be attributed solely to a larger latent dimension. Duplicating Layer 12 provides only limited improvement, whereas incorporating representations from different depths yields substantially larger gains, especially at longer horizons. These results suggest that representations from different depths provide complementary predictive information that cannot be obtained by simply enlarging the final-layer representation.

6 Conclusion

We established a generalization theory for JEPA-based world models by showing that the spectral JEPA objective is equivalent to a low-rank factorization of an action-conditioned co-occurrence matrix. We further derived finite-sample planning regret bounds, revealing an approximation–sample error trade-off with respect to latent dimensionality. Experiments demonstrate the robustness of latent-level prediction under unpredictable observations and difficult planning tasks. Motivated by our theory, we also investigated Multi-Depth JEPA, which improves planning success over single-depth representations and a dimensionality-matched control, illustrating how theoretical insights can inform representation design.

AI USE STATEMENT

In this work, we used generative AI tools for none of the tasks with required disclosure.

We have not used generative AI tools for developing the theoretical framework, formulating mathematical claims or proofs, proposing or refining hypotheses, designing the methodology or experiments, implementing the proposed methods, processing experimental data, or interpreting the results, and the rest of the required disclosure tasks are not applicable to this work.

Additionally, we used generative AI tools for grammar checking and language polishing of selected paragraphs.

We have reviewed all AI-assisted work. In particular, all AI-assisted edits were manually checked by the authors to ensure that they did not alter the technical meaning, scientific claims, mathematical content, or experimental conclusions of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We provide complete proofs of our theoretical results in the Appendix. For the empirical evaluation, we provide detailed experimental settings for the controlled synthetic environment, including the environment dynamics, observation construction, predictive model architecture, training objectives, noise configurations, planning procedure, and evaluation criterion. We also specify the configurations used for Multi-Depth JEPA, including the selected intermediate layers, the dimensionality-matched control, and the evaluated planning horizons. For the uniformity-regularization ablation, we follow the original LeWM experimental protocol and report results over three random seeds with 50 evaluation episodes per seed. These details are provided to facilitate reproduction of both the theoretical results and empirical evaluations.

ETHICS STATEMENT

This work is a theoretical and methodological study of JEPA-based world models and action planning. Our experiments are conducted in controlled synthetic environments and existing simulated benchmark environments. The work does not involve human-subject experiments, the collection of new personal or sensitive data, or deployment in real-world decision-making systems.

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 *CVPR*, 2023.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Jiaxin Bai and Jiaxuan Xiong. Temporal-distance jepa: Plan-aware representation learning for latent world model predictive control. *arXiv preprint arXiv:2607.25337*, 2026.

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

Randall Balestriero and Yann LeCun. Spectral graph theory: The mathematics of self-supervised learning [special issue on the mathematics of deep learning]. *IEEE Signal Processing Magazine*, 43(3):8–20, 2026.

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

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-jepa: Latent video prediction for visual representation learning. 2023.

Delong Chen, Mustafa Shukor, Theo Moutakanni, Willy Chung, Jade Yu, Tejaswi Kasarla, Yejin Bang, Allen Bolourchi, Yann LeCun, and Pascale Fung. Vl-jepa: Joint embedding predictive architecture for vision-language. *arXiv preprint arXiv:2512.10942*, 2025.

Jingyi Cui, Weiran Huang, Yifei Wang, and Yisen Wang. Rethinking weak supervision in helping contrastive learning. In *ICML*, 2023.

Matthieu Destrade, Oumayma Bounou, Quentin Le Lidec, Jean Ponce, and Yann LeCun. Value-guided action planning with jepa world models. *arXiv preprint arXiv:2601.00844*, 2025.

Jeff Z HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. *NeurIPS*, 2021.

David Klindt, Yann LeCun, and Randall Balestriero. When does lejepa learn a world model? *arXiv preprint arXiv:2605.26379*, 2026.

An Lanji, Dawei Liu, Jin Li, Haoran Xu, Mei Chen, and Yu Tian. Unijepa: A unified joint-embedding predictive architecture for task-agnostic visual world modeling. *arXiv preprint arXiv:2608.07409*, 2026.

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

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How jepa avoids noisy features: The implicit bias of deep linear self distillation networks. *NeurIPS*, 2024.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Sergi Masip, Jonathan Swinnen, Yutong Hu, Renaud Detry, and Tinne Tuytelaars. Ff-jepa: Long-horizon planning in world models with latent planners. *arXiv preprint arXiv:2606.09311*, 2026.

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

Kendrick Shen, Robbie M Jones, Ananya Kumar, Sang Michael Xie, Jeff Z HaoChen, Tengyu Ma, and Percy Liang. Connect, not collapse: Explaining contrastive learning for unsupervised domain adaptation. In *ICML*, 2022.

Basile Terver, Tsung-Yen Yang, Jean Ponce, Adrien Bardes, and Yann LeCun. What drives success in physical planning with joint-embedding predictive world models? *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856.

Qi Zhang, Yifei Wang, and Yisen Wang. On the generalization of multi-modal contrastive learning. In *ICML*, 2023.

Qi Zhang, Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Look ahead or look around? a theoretical comparison between autoregressive and masked pretraining. In *ICML*, 2024.

Yi-Ge Zhang, Jingyi Cui, Qiran Li, and Yisen Wang. Difficult examples hurt unsupervised contrastive learning: A theoretical perspective. In *ICLR*, 2026.

Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

A Appendix

A.1 Detailed Finite-Sample Formulation

We define the empirical JEPA risk in Definition A.1.

Definition A.1 (Empirical JEPA risk)**.** Consider a dataset $\hat{\mathcal{X}} = \{x_1, \ldots, x_n\}$ containing n data points i.i.d. sampled from P_X. Let $\hat{\mathrm{P}}_X$ be the uniform distribution over $\hat{\mathcal{X}}$. Let $\hat{\mathrm{P}}_{x,x'}$ be the uniform distribution over data pairs (x_i, x_j) where $i \neq j$. We define the empirical spectral JEPA loss of a feature extractor f as

$$\begin{aligned}\hat{\mathcal{R}}_n(f, g, a) := &-2\mathbb{E}_{x\in\hat{\mathrm{P}}_X, x^+\sim\mathrm{P}(\cdot|x,a)}\left[g(f(x), a)^\top f(x')\right] \\ &+ \mathbb{E}_{x,x^-\in\hat{\mathrm{P}}_{x,x^-}, x'\sim\mathrm{P}(\cdot|x^-,a)}\left[g(f(x), a)^\top f(x')\right]^2. \end{aligned} \tag{26}$$

Lemma A.2 shows that $\hat{\mathcal{R}}_n(f)$ is an unbiased estimator of population spectral JEPA loss.

Lemma A.2. *$\hat{\mathcal{R}}_n(f, g, a)$ is an unbiased estimator of $\mathcal{R}(f, g, a)$, i.e.,*

$$\mathbb{E}_{\hat{\mathcal{X}}}\left[\hat{\mathcal{R}}_n(f, g, a)\right] = \mathcal{R}(f, g, a). \tag{27}$$

Definition A.3. Given dataset $\hat{\mathcal{X}}$, we sample a subset of tuples as follows: first sample a permutation $\pi : [n] \to [n]$. Then given $a \in \mathcal{A}$, we sample tuple $S = \{(z_i, z_i^+, z_i')\}_{i=1}^{n/2}$ as follows:

$$z_i = x_{\pi_{2i-1}}, \qquad z_i^+ \sim \mathrm{P}(\cdot|z_i, a), \qquad z_i' \sim \mathrm{P}(\cdot|x_{\pi_{2i}}, a). \tag{28}$$

We define the following loss on S:

$$\hat{\mathcal{R}}_S(f, g, a) = \frac{1}{n/2}\sum_{i=1}^{n/2}\left[\left(g(f(z_i), a)^\top f(z_i')\right)^2 - 2g(f(z_i), a)^\top f(z_i^+)\right]. \tag{29}$$

In Lemma A.4, we see that $\hat{\mathcal{R}}_S(f, g, a)$ is an unbiased estimator of $\hat{\mathcal{R}}_n(f, g, a)$.

Lemma A.4. *Given $\hat{\mathcal{X}}$, we have*

$$\mathbb{E}_S\hat{\mathcal{R}}_S(f, g, a) = \hat{\mathcal{R}}_n(f, g, a). \tag{30}$$

A.2 Proofs

Proof of Theorem 3.1. By the definition of JEPA risk in equation 3, given $a \in \mathcal{A}$, we have

$$\begin{aligned}&\mathcal{R}_{\mathrm{JEPA}}(f, g, a) \\ &= \mathbb{E}_{x\sim\mathrm{P}_X}\mathbb{E}_{x^+\sim\mathrm{P}(\cdot|x,a)}\|g(f(x), a) - f(x^+)\|^2 + \mathbb{E}_{\substack{x\sim\mathrm{P}_X \\ x'\sim\mathrm{P}_{X|a}}}\left[g(f(x), a)^\top f(x')\right]^2 \\ &= \sum_{x,x^+} w(x, x^+, a)\|g(f(x), a) - f(x^+)\|^2 + \sum_{x,x'} w(x)w(x'|a)\left[g(f(x), a)^\top f(x')\right]^2 \\ &= 2 - 2\sum_{x,x^+} w(x, x^+, a)g(f(x), a)^\top f(x^+) + \sum_{x,x^+} w(x)w(x^+|a)\left[g(f(x), a)^\top f(x^+)\right]^2 \\ &= \sum_{x,x^+}\left[\frac{w(x, x^+, a)}{\sqrt{w(x)w(x^+|a)}} - \left[\sqrt{w(x)}g(f(x), a)\right]^\top\left[\sqrt{w(x^+|a)}f(x^+)\right]\right]^2 + 2 - \frac{w(x, x^+, a)^2}{w(x)w(x^+|a)}\end{aligned} \tag{31}$$

where the third equation holds because $\|g(f(x), a)\|^2 = \|f(x^+)\|^2 = 1$. Then by denoting $F := \left[\sqrt{w(x|a)}f(x)\right]_{x\in\mathcal{X}}$ and $G(F, a) := \left[\sqrt{w(x)}g(f(x), a)\right]_{x\in\mathcal{X}}$, we have

$$\mathcal{R}_{\mathrm{JEPA}}(f, g, a) = \|\bar{M}(a) - G(F, a)^\top F\|^2 + \mathrm{const}. \tag{32}$$

□

Proof of Theorem 4.1. If $T = 1$, by definition, we have

$$\tilde{a} = \arg\min_a \|g(f(x_0), a) - f(x_g)\|^2 = \arg\max_a g(f(x_0), a)^\top f(x_g), \tag{33}$$

and

$$a^{**} = \arg\max_a V(a) = \arg\max_a w(x_g, x_0, a). \tag{34}$$

Then for given x_0 and x_g, we have

$$\begin{aligned}
&r(\tilde{a}; x_0, x_g) \\
&= \frac{w(x_0, x_g, a^{**})}{w(x_0)} - \frac{w(x_0, x_g, \tilde{a})}{w(x_0)} \\
&= \frac{1}{w(x_0)} \left[w(x_0, x_g, a^{**}) - w(x_0, x_g, \tilde{a})\right] \\
&= \frac{1}{w(x_0)} \big[w(x_0, x_g, a^{**}) - w(x_0) w(x_g|a^{**}) g(f(x_0), a^{**})^\top f(x_g) \\
&\qquad + w(x_0) w(x_g|a^{**}) g(f(x_0), a^{**})^\top f(x_g) - w(x_0) w(x_g|\tilde{a}) g(f(x_0), \tilde{a})^\top f(x_g) \\
&\qquad + w(x_0) w(x_g|\tilde{a}) g(f(x_0), \tilde{a})^\top f(x_g) - w(x_0, x_g, \tilde{a})\big]. \\
&= \sqrt{\frac{w(x_g|a^{**})}{w(x_0)}} \left[\frac{w(x_0, x_g, a^{**})}{\sqrt{w(x_0) w(x_g|a^{**})}} - \sqrt{w(x_0) w(x_g|a^{**})} g(f(x_0), a^{**})^\top f(x_g)\right] && (35) \\
&+ \big[w(x_g|a^{**}) - w(x_g|\tilde{a})\big] g(f(x_0), a^{**})^\top f(x_g) && (36) \\
&+ w(x_g|\tilde{a}) \big[g(f(x_0), a^{**})^\top f(x_g) - g(f(x_0), \tilde{a})^\top f(x_g)\big] && (37) \\
&- \sqrt{\frac{w(x_g|\tilde{a})}{w(x_0)}} \left[\frac{w(x_0, x_g, \tilde{a})}{\sqrt{w(x_0) w(x_g|\tilde{a})}} - \sqrt{w(x_0) w(x_g|\tilde{a})} g(f(x_0), \tilde{a})^\top f(x_g)\right]. && (38)
\end{aligned}$$

Denote $\delta(x_0, x_g, a) := \frac{w(x_0, x_g, a)}{\sqrt{w(x_0) w(x_g|a)}} - \sqrt{w(x_0) w(x_g|a)} g(f(x_0), a)^\top f(x_g)$, then we have

$$(35) = \sqrt{w(x_g|a^{**})/w(x_0)}\, \delta(x_0, x_g, a^{**}) \tag{39}$$

and

$$(38) = -\sqrt{w(x_g|\tilde{a})/w(x_0)}\, \delta(x_0, x_g, \tilde{a}). \tag{40}$$

Recall that $\mathcal{R}_{\text{S-JEPA}}(f, g, a) = \sum_{x_0, x_g} \delta(x_0, x_g, a)^2$. Then we have

$$\begin{aligned}
r(\tilde{a}; x_0, x_g) &\le \sqrt{w(x_g|a^{**})/w(x_0)} \cdot \delta(x_0, x_g, a^{**}) + \sqrt{w(x_g|\tilde{a})/w(x_0)} \cdot \delta(x_0, x_g, \tilde{a}) \\
&\quad + \left[w(x_g|a^{**}) - w(x_g|\tilde{a})\right] g(f(x_0), a^{**})^\top f(x_g) \\
&\le \sqrt{w(x_g|a^{**})/w(x_0) \cdot \mathcal{R}_{\text{S-JEPA}}(f, g, a^{**})} + \sqrt{w(x_g|\tilde{a})/w(x_0) \cdot \mathcal{R}_{\text{S-JEPA}}(f, g, \tilde{a})} \\
&\quad + w(x_g|a^{**}) - w(x_g|\tilde{a}) \\
&\le 2\sqrt{\max_a w(x_g|a)/w(x_0) \cdot \mathcal{R}_{\text{S-JEPA}}(f, g, a)} + w(x_g|a^{**}) - w(x_g|\tilde{a}) \\
&\le 2\sqrt{\max_a w(x_g|a)/w(x_0)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)} + w(x_g|a^{**}) - w(x_g|\tilde{a}). \qquad (41)
\end{aligned}$$

Then if $\mathbb{E}_{x_g} w(x_g|a) = \mathbb{E}_{x_g} w(x_g|a')$ for $a, a' \in \mathcal{A}$, we have

$$\mathcal{E}(a^*) = \mathbb{E}_{x_0, x_g} r(\tilde{a}; x_0, x_g) \le 2\mathbb{E}_{x_0, x_g} \sqrt{\max_a w(x_g|a)/w(x_0)} \cdot \max_a \sqrt{\mathcal{R}_{\text{S-JEPA}}(f, g, a)}. \tag{42}$$

□

Proof of Theorem 4.2. By definition, given x_0 and x_g, we have

$$
\begin{aligned}
&r(\tilde{a}_0, \dots, \tilde{a}_{T-1}) \\
&= V(a^{**}) - V(\tilde{a}) \\
&= \mathrm{P}(\hat{x}_T = x_g | x_0, a_0^{**}, \dots, a_{T-1}^{**}) - \mathrm{P}(\hat{x}_T = x_g | x_0, \tilde{a}_0, \dots, \tilde{a}_{T-1}) \\
&= \sum_{x_1, \dots, x_{T-1}} \frac{w(x_{T-1}, x_g, a_{T-1}^{**})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, a_0^{**})}{w(x_0)} - \frac{w(x_{T-1}, x_g, \tilde{a}_{T-1})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, \tilde{a}_0)}{w(x_0)} \\
&= \sum_{x_1, \dots, x_{T-1}} \frac{w(x_{T-1}, x_g, a_{T-1}^{**})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, a_0^{**})}{w(x_0)} - \frac{w(x_{T-1}, x_g, a_{T-1}^{**})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, \tilde{a}_0)}{w(x_0)} \\
&\qquad + \cdots \\
&\qquad + \frac{w(x_{T-1}, x_g, \hat{a}_{T-1}^{*})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, \hat{a}_0^{*})}{w(x_0)} - \frac{w(x_{T-1}, x_g, \hat{a}_{T-1})}{w(x_{T-1})} \cdots \frac{w(x_0, x_1, \hat{a}_0)}{w(x_0)} \\
&= \sum_{x_1, \dots, x_{T-1}} \prod_{t=0}^{T-1} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} - \prod_{t=0}^{T-1} \frac{w(x_t, x_{t+1}, \tilde{a}_t)}{w(x_t)} \\
&= \sum_{x_1, \dots, x_{T-1}} \sum_{\ell=0}^{T-1} \prod_{t=0}^{\ell} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} \cdot \prod_{t=\ell+1}^{T-1} \frac{w(x_t, x_{t+1}, \tilde{a}_t)}{w(x_t)} - \prod_{t=0}^{\ell-1} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} \cdot \prod_{t=\ell}^{T-1} \frac{w(x_t, x_{t+1}, \tilde{a}_t)}{w(x_t)} \\
&= \sum_{x_1, \dots, x_{T-1}} \sum_{\ell=0}^{T-1} \prod_{t=0}^{\ell-1} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} \cdot \left[\frac{w(x_\ell, x_{\ell+1}, a_\ell^{**})}{w(x_\ell)} - \frac{w(x_\ell, x_{\ell+1}, \tilde{a}_\ell)}{w(x_\ell)} \right] \cdot \prod_{t=\ell+1}^{T-1} \frac{w(x_t, x_{t+1}, \tilde{a}_t)}{w(x_t)}.
\end{aligned} \tag{43}
$$

By Theorem 4.1, we have for $\ell = 0, \dots, T-1$

$$
\frac{w(x_\ell, x_{\ell+1}, a_\ell^{**})}{w(x_\ell)} - \frac{w(x_\ell, x_{\ell+1}, \tilde{a}_\ell)}{w(x_\ell)} \le 2\sqrt{\max_a w(x_{\ell+1}|a)/w(x_\ell)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)}. \tag{44}
$$

Then if $M(a)$ is deterministic, i.e., given $a \in \mathcal{A}$ and $x \in \mathcal{X}$, for $x^+ \in \mathcal{X}$ there exists $w(x, x^+, a) = 1$ and $w(x, x', a) = 0$ for $x' \neq x^+$, we have

$$
\begin{aligned}
&r(\tilde{a}_0, \dots, \tilde{a}_{T-1}) \\
&\le \sum_{x_1, \dots, x_{T-1}} \sum_{\ell=0}^{T-1} \prod_{t=0}^{\ell-1} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} \cdot \prod_{t=\ell+1}^{T-1} \frac{w(x_t, x_{t+1}, \tilde{a}_t)}{w(x_t)} \\
&\qquad \cdot 2\sqrt{\max_a w(x_{\ell+1}|a)/w(x_\ell)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)} \\
&= \sum_{\ell=0}^{T-1} \sum_{x_1, \dots, x_{T-1}} \prod_{t=0}^{\ell-2} \frac{w(x_t, x_{t+1}, a_t^{**})}{w(x_t)} \cdot \prod_{t=\ell+1}^{T-1} \frac{w(x_t, x_{t+2}, \tilde{a}_t)}{w(x_t)} \\
&\cdot \frac{w(x_{\ell-1}, x_\ell, a_{\ell-1}^{**})}{w(x_{\ell-1})} \cdot \frac{w(x_{\ell+1}, x_{\ell+2}, a_{\ell+1}^{**})}{w(x_{\ell+1})} \cdot 2\sqrt{\max_a w(x_{\ell+1}|a)/w(x_\ell)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)} \\
&\le \sum_{\ell=0}^{T-1} \sum_{x_\ell} \mathrm{P}(x_\ell | x_0, a_0^{**}, \dots, a_{\ell-1}^{**}) \mathrm{P}(x_g | x_\ell, \tilde{a}_\ell, \dots, \tilde{a}_{T-1}) \\
&\qquad \cdot 2 \max_\ell \sqrt{\max_a w(x_{\ell+1}|a)/w(x_\ell)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)} \\
&\le 2T \cdot \max_{a, x, x'} \sqrt{w(x_g|a)/w(x)} \cdot \sqrt{\max_a \mathcal{R}_{\text{S-JEPA}}(f, g, a)}.
\end{aligned} \tag{45}
$$

□

Proof of Theorem 4.3. By Theorem 3.1, given $a \in \mathcal{A}$, we have

$$\mathcal{R}_{\text{JEPA}}(f, g, a) = \|\bar{M}(a) - G(F, a)^\top F\|_F^2 + \text{const.} \tag{46}$$

Since $B := G(F, a)^\top F$ has rank at most k minimizing the JEPA population risk is equivalent to solving

$$\min_{\text{rank}(B) \le k} \|\bar{M}(a) - B\|_F^2. \tag{47}$$

Let

$$\bar{M}(a) = U(a)\Sigma(a)V(a)^\top \tag{48}$$

be the singular value decomposition of $\bar{M}(a)$. By the Eckart–Young–Mirsky theorem, the unique best rank-k approximation of $\bar{M}(a)$ under the Frobenius norm is

$$B^* := G^*(F, a)^\top F^* = U_k(a)\Sigma_k(a)V_k(a)^\top. \tag{49}$$

Then one valid factorization is obtained by distributing the singular values equally between the two factors, i.e.,

$$F^* = \Sigma_k(a)^{1/2} V_k(a)^\top, \tag{50}$$

and

$$G^*(F, a) = \Sigma_k(a)^{1/2} U_k(a)^\top. \tag{51}$$

In this case, we have

$$\|\bar{M}(a) - G^*(F, a)^\top F^*\|_F^2 = \sum_{i>k} \sigma_i^2(a), \tag{52}$$

where $\sigma_i(a)$ is the i-th largest singular value of $\bar{M}(a)$ for $i \in \{1, \dots, \#|\mathcal{X}|\}$. Hence we have

$$\mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) = \min_{f,g} \mathcal{R}_{\text{S-JEPA}}(f, g, a) = \sum_{i>k} \sigma_i^2(a). \tag{53}$$

□

Proof of Theorem 4.4. Fix $a \in \mathcal{A}$. Define the composed hypothesis class

$$\mathcal{H}_a := \{h_{f,g,a} := g \circ f \circ a : f \in \mathcal{F},\ g \in \mathcal{G}\},$$

where $g \circ f \circ a(z, z^+, z') := \left(g(f(z_i), a)^\top f(z_i')\right)^2 - 2g(f(z_i), a)^\top f(z_i^+)$. For simplicity, write $h(z) := g(f(z), a)$. The population spectral JEPA risk can be written as

$$\mathcal{R}_{\text{S-JEPA}}(f, g, a) = \mathbb{E}_{(z,z^+,z^-)}\left[\left(h(z)^\top f(z')\right)^2 - 2h(z)^\top f(z^+)\right],$$

where (z, z^+) is a positive pair and z' is a negative sample. Its empirical counterpart based on $m = n/2$ independent tuples is

$$\hat{\mathcal{R}}_m(f, g, a) = \frac{1}{m}\sum_{i=1}^{m}\left[\left(h(z_i)^\top f(z_i')\right)^2 - 2h(z_i)^\top f(z_i^+)\right].$$

Since $\|f\|_\infty \le \kappa$ and $\|g\|_\infty \le \kappa$, we have

$$\|h(z)\|_\infty \le \kappa, \qquad \|f(z)\|_\infty \le \kappa.$$

Therefore,

$$\left|h(z)^\top f(z')\right| \le \sum_{\ell=1}^{k} |h_\ell(z)|\,|f_\ell(z')| \le k\kappa^2.$$

Consequently, the tuple loss

$$\ell_{f,g,a}(z, z^+, z^-) := \left(h(z)^\top f(z^-)\right)^2 - 2h(z)^\top f(z^+)$$

is uniformly bounded as

$$\ell_{f,g,a}(z, z^+, z^-) \in \left[-2k\kappa^2,\ k^2\kappa^4 + 2k\kappa^2\right].$$

We now control the Rademacher complexity of the tuple-loss class

$$\mathcal{L}_a := \{\ell_{f,g,a} : f \in \mathcal{F},\ g \in \mathcal{G}\}.$$

We now give the detailed Rademacher complexity bound for the tuple-loss class. Recall that

$$\ell_{f,g,a}(z, z^+, z^-) := \left(h(z)^\top f(z^-)\right)^2 - 2h(z)^\top f(z^+), \qquad h := g \circ f \circ a.$$

Let

$$\mathcal{H}_a := \{h_{f,g,a} = g \circ f \circ a :\ f \in \mathcal{F},\ g \in \mathcal{G}\}.$$

Since a is fixed, for any sample $S = \{(z_i, z_i^+, z_i^-)\}_{i=1}^m$, we have

$$\hat{\mathfrak{R}}_S(\mathcal{H}_a) = \hat{\mathfrak{R}}_{a(S)}(\mathcal{G} \circ \mathcal{F}),$$

where $a(S) := \{(a, z_i)\}_{i=1}^m$. If $\hat{\mathfrak{R}}_m(\mathcal{G} \circ \mathcal{F})$ denotes the worst-case empirical Rademacher complexity over all samples of size m, then

$$\hat{\mathfrak{R}}_S(\mathcal{H}_a) \le \hat{\mathfrak{R}}_m(\mathcal{G} \circ \mathcal{F}).$$

Let $S = \{(z_i, z_i^+, z_i^-)\}_{i=1}^m$ be a fixed sample of m tuples. We need to bound

$$\hat{\mathfrak{R}}_S(\mathcal{L}_a) := \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F},\ g\in\mathcal{G}} \frac{1}{m}\sum_{i=1}^m \sigma_i \ell_{f,g,a}(z_i, z_i^+, z_i^-)\right],$$

where $\sigma_1, \ldots, \sigma_m$ are i.i.d. Rademacher random variables. By subadditivity of the supremum,

$$\hat{\mathfrak{R}}_S(\mathcal{L}_a) \le \underbrace{\mathbb{E}_\sigma\left[\sup_{f,g} \frac{1}{m}\sum_{i=1}^m \sigma_i \left(h(z_i)^\top f(z_i^-)\right)^2\right]}_{T_1} + 2\underbrace{\mathbb{E}_\sigma\left[\sup_{f,g} \frac{1}{m}\sum_{i=1}^m \sigma_i h(z_i)^\top f(z_i^+)\right]}_{T_2}. \tag{54}$$

We first control T_1. Since

$$|h(z_i)^\top f(z_i^-)| \le k\kappa^2,$$

the map $u \mapsto u^2$ is $2k\kappa$-Lipschitz on $[-k\kappa^2, k\kappa^2]$. Therefore, by Talagrand's contraction lemma,

$$T_1 \le 2k\kappa^2\, \mathbb{E}_\sigma\left[\sup_{f,g} \frac{1}{m}\sum_{i=1}^m \sigma_i h(z_i)^\top f(z_i^-)\right]. \tag{55}$$

Expanding the inner product gives

$$T_1 \le 2k\kappa^2 \sum_{\ell=1}^k \mathbb{E}_\sigma\left[\sup_{h_\ell\in\mathcal{H}_{a,\ell},\ f_\ell\in\mathcal{F}_\ell} \frac{1}{m}\sum_{i=1}^m \sigma_i h_\ell(z_i) f_\ell(z_i^-)\right]. \tag{56}$$

For two scalar function classes $\mathcal{U}$ and $\mathcal{V}$ uniformly bounded by κ, the following standard product-class bound holds:

$$\mathbb{E}_\sigma\left[\sup_{u\in\mathcal{U},\ v\in\mathcal{V}} \frac{1}{m}\sum_{i=1}^m \sigma_i u(x_i) v(y_i)\right] \le 4\kappa\left[\hat{\mathfrak{R}}_m(\mathcal{U}) + \hat{\mathfrak{R}}_m(\mathcal{V})\right]. \tag{57}$$

Indeed, using the identity

$$uv = \frac{1}{2}\left[(u+v)^2 - u^2 - v^2\right],$$

or equivalently the polarization identity, and then applying Talagrand's contraction lemma to the square function on $[-2\kappa, 2\kappa]$, one obtains equation 57. Applying this bound with

$$\mathcal{U} = \mathcal{H}_{a,\ell}, \qquad \mathcal{V} = \mathcal{F}_\ell,$$

we get

$$T_1 \le 2k\kappa^2 \sum_{\ell=1}^{k} 4\kappa \left[\hat{\mathfrak{R}}_m(\mathcal{H}_{a,\ell}) + \hat{\mathfrak{R}}_m(\mathcal{F}_\ell)\right] \le 8k^2\kappa^3 \left[\hat{\mathfrak{R}}_m(\mathcal{H}_a) + \hat{\mathfrak{R}}_m(\mathcal{F})\right]. \tag{58}$$

We next control T_2. Expanding the inner product gives

$$\begin{aligned} T_2 &= \mathbb{E}_\sigma \left[\sup_{f,g} \frac{1}{m} \sum_{i=1}^{m} \sigma_i \sum_{\ell=1}^{k} h_\ell(z_i) f_\ell(z_i^+)\right] \\ &\le \sum_{\ell=1}^{k} \mathbb{E}_\sigma \left[\sup_{h_\ell \in \mathcal{H}_{a,\ell},\ f_\ell \in \mathcal{F}_\ell} \frac{1}{m} \sum_{i=1}^{m} \sigma_i h_\ell(z_i) f_\ell(z_i^+)\right]. \end{aligned} \tag{59}$$

Again applying the product-class bound equation 57,

$$T_2 \le \sum_{\ell=1}^{k} 4\kappa \left[\hat{\mathfrak{R}}_m(\mathcal{H}_{a,\ell}) + \hat{\mathfrak{R}}_m(\mathcal{F}_\ell)\right] \le 4k\kappa \left[\hat{\mathfrak{R}}_m(\mathcal{H}_a) + \hat{\mathfrak{R}}_m(\mathcal{F})\right]. \tag{60}$$

Combining equation 58 and equation 60, we obtain

$$\begin{aligned} \hat{\mathfrak{R}}_S(\mathcal{L}_a) &\le T_1 + 2T_2 \\ &\le \left(8k^2\kappa^3 + 8k\kappa\right) \left[\hat{\mathfrak{R}}_m(\mathcal{H}_a) + \hat{\mathfrak{R}}_m(\mathcal{F})\right] \\ &\le \left(8k^2\kappa^3 + 8k\kappa\right) \left[\hat{\mathfrak{R}}_m(\mathcal{G} \circ \mathcal{F}) + \hat{\mathfrak{R}}_m(\mathcal{F})\right]. \end{aligned} \tag{61}$$

Taking $m = n/2$ gives the desired Rademacher complexity control for the tuple-loss class.

Here $\hat{\mathfrak{R}}_m(\mathcal{G} \circ \mathcal{F})$ controls the online branch $h = g \circ f \circ a$, while $\hat{\mathfrak{R}}_m(\mathcal{F})$ controls the target branch f. Since a is fixed, the composition with a does not introduce an additional trainable function class and is absorbed into $\mathcal{G} \circ \mathcal{F}$.

By the standard Rademacher uniform convergence bound, with probability at least $1 - \delta$, uniformly over $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\begin{aligned} &\left|\mathcal{R}_{\text{S-JEPA}}(f, g, a) - \hat{\mathcal{R}}_m(f, g, a)\right| \\ &\le \left(16k^2\kappa^3 + 16k\kappa\right) \left[\hat{\mathfrak{R}}_m(\mathcal{G} \circ \mathcal{F}) + \hat{\mathfrak{R}}_m(\mathcal{F})\right] + \left(4k\kappa^2 + k^2\kappa^4\right) \left(\sqrt{\frac{\log(2/\delta)}{n}} + \delta\right). \end{aligned} \tag{62}$$

Taking $m = n/2$, define

$$\epsilon_n := \left(16k^2\kappa^3 + 16k\kappa\right) \left[\hat{\mathfrak{R}}_{n/2}(\mathcal{G} \circ \mathcal{F}) + \hat{\mathfrak{R}}_{n/2}(\mathcal{F})\right] + \left(4k\kappa^2 + k^2\kappa^4\right) \left(\sqrt{\frac{\log(2/\delta)}{n}} + \delta\right). \tag{63}$$

Then, with probability at least $1 - \delta$, uniformly over $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\left|\mathcal{R}_{\text{S-JEPA}}(f, g, a) - \hat{\mathcal{R}}_{n/2}(f, g, a)\right| \le \epsilon_n.$$

Using this uniform deviation bound and the empirical optimality of $(\hat{f}, \hat{g})$, we have

$$\begin{aligned} &\mathcal{R}_{\text{S-JEPA}}(\hat{f}, \hat{g}, a) - \mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) \\ &\le \hat{\mathcal{R}}_{n/2}(\hat{f}, \hat{g}, a) + \epsilon_n - \mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) \\ &\le \hat{\mathcal{R}}_{n/2}(f^*, g^*, a) + \epsilon_n - \mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) \\ &\le 2\epsilon_n. \end{aligned} \tag{64}$$

Substituting the definition of ϵ_n gives

$$\begin{aligned}
&\mathcal{R}_{\text{S-JEPA}}(\hat{f}, \hat{g}, a) - \mathcal{R}_{\text{S-JEPA}}(f^*, g^*, a) \\
&\leq \left(32k^2\kappa^3 + 32k\kappa\right) \left[\hat{\mathfrak{R}}_{n/2}(\mathcal{G} \circ \mathcal{F}) + \hat{\mathfrak{R}}_{n/2}(\mathcal{F})\right] \\
&\quad + \left(8k\kappa^2 + 2k^2\kappa^4\right) \left(\sqrt{\frac{\log(2/\delta)}{n}} + \delta\right).
\end{aligned} \tag{65}$$

This proves the desired sample error bound. □

A.3 Effect of the Additional Uniformity Regularization

Our theoretical formulation introduces an additional uniformity regularization term to the original JEPA objective, which enables the spectral characterization used in our analysis. Since this term is introduced primarily for theoretical analysis and is not part of the original JEPA objective, a natural question is whether it changes the empirical performance of the underlying world model.

To examine this, we conduct an additional experiment based on the representative JEPAL-based world model LeWM (Maes et al., 2026) and follow its original experimental setting, including the model architecture, training protocol, and downstream planning procedure. The original LeWM objective consists of the latent prediction MSE loss together with SIGReg. We additionally incorporate the uniformity term used in our theoretical formulation:

$$\mathcal{L} = \mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{SIGReg}} + \mathcal{L}_{\text{uni}}, \tag{66}$$

Table 2: Effect of the additional uniformity regularization used in our theoretical formulation. We follow the original LeWM setting and only add the uniformity term. Each result is the mean value evaluated over three seeds with 50 episodes per seed.

Objective	PushT	Cube	TwoRoom
LeWM	90.7	66.7	84.7
LeWM $+\mathcal{L}_{\text{uni}}$	**91.3**	**69.3**	**85.3**

We evaluate the resulting models on PushT, Cube, and TwoRoom following the original LeWM evaluation protocol with a planning horizon of $H = 25$. Each result is evaluated over three seeds with 50 episodes per seed, resulting in 150 episodes in total. As shown in Table 2, adding the uniformity term does not degrade performance on any of the three environments. Instead, we observe small improvements from 90.7% to 91.3% on PushT, 66.7% to 69.3% on Cube, and 84.7% to 85.3% on TwoRoom. These results suggest that the additional uniformity regularization introduced in our theoretical formulation is empirically benign and can even provide modest performance improvements.

A.4 Details of the Synthetic Control Environment

We provide additional implementation details for the controlled experiment in Section 5.1.

Environment. We use a 2D point-mass system with underlying state $s_t = [p_x, p_y, v_x, v_y]$ and action $a_t = [a_x, a_y]$. The dynamics are given by

$$v_{t+1} = \text{damping} \cdot v_t + \text{action_scale} \cdot a_t, \qquad p_{t+1} = p_t + dt \cdot v_{t+1}. \tag{67}$$

The model does not directly observe the underlying state. Instead, its input is a higher-dimensional observation

$$x_t = [x_t^{\text{state}}, x_t^{\text{nuisance}}, x_t^{\text{noise}}], \tag{68}$$

where x_t^{state} contains features derived from the position and velocity, x_t^{nuisance} contains additional nonlinear projections of the state, and x_t^{noise} consists of random features independent of both the action and task objective.

Predictive models. Both latent- and input-level models use the same predictive backbone, consisting of an observation encoder $x_t \mapsto z_t$, an action encoder, and a GRU-based dynamics predictor. An additional decoder $z \mapsto x$ is used for input-level reconstruction.

Given the current observation x_t and an action sequence $a_{t:t+H-1}$, the model predicts the state after H steps. We use $H = 3$ throughout the experiment. The latent-level model is trained to predict the encoded future observation,

$$\mathcal{L}_{\text{latent}} = \text{MSE}\left(\hat{z}_{t+H}, \text{stopgrad}(\text{Enc}(x_{t+H}))\right), \tag{69}$$

together with a lightweight variance regularization term. The input-level model decodes the predicted latent and minimizes the reconstruction error

$$\mathcal{L}_{\text{input}} = \text{MSE}\left(\hat{x}_{t+H}, x_{t+H}\right). \tag{70}$$

To control the influence of unpredictable observation components, we vary the reconstruction weight assigned to the random-noise dimensions. This quantity corresponds to the *noise level* reported in Figure 2. Increasing this weight places greater emphasis on reconstructing information that is inherently unpredictable from the current state and action.

Planning. For each evaluation episode, we sample a start state and a goal state. The corresponding goal observation x_{goal} is provided to the planner, and in the latent setting it is encoded as $z_{\text{goal}} = \text{Enc}(x_{\text{goal}})$.

We use receding-horizon planning with the Cross-Entropy Method (CEM). At each decision step, CEM optimizes a candidate action sequence using the learned predictive model, the first action is executed, and the agent replans from the resulting observation. The latent-level planner minimizes the distance between the predicted latent and z_{goal}, whereas the input-level planner minimizes the corresponding distance in observation space.

The *required planning steps* in Figure 2 controls the approximate number of such closed-loop decisions needed to reach the goal. It is therefore distinct from the fixed training prediction horizon $H = 3$. Larger values correspond to tasks requiring more repeated model predictions and replanning steps.

All methods are evaluated using the same ground-truth positional criterion. A rollout is counted as successful when

$$\|p_{\text{current}} - p_{\text{goal}}\|_2 \leq 0.08. \tag{71}$$