Spectral World Models: A Reproducible Operator-Based Architecture for Multimodal Prediction and Planning

Andrew Kiruluta
UC Berkeley School of Information

August 22, 2026

Abstract

World models allow agents to predict future environment states, evaluate hypothetical actions, and plan by internal imagination. This revised manuscript converts the original Spectral World Model (SWM) framework from a broad conceptual proposal into a minimal, implementable, and empirically testable architecture. The model represents image observations by a fixed Haar wavelet transform, represents text instructions by a concrete graph-spectral encoder over token positions, fuses the two modalities through learned low-rank projections, and predicts future latent states with an action-conditioned block-sparse spectral transition. The resulting model estimates the conditional law $p_\theta(X_{t+1} \mid X_t, a_t)$ in a shared Hilbert-space coordinate system and supports planning by repeated latent rollouts.

The revision addresses five critical gaps. First, it defines the spectral text encoder, action embedding, memory update, cross-modal fusion operator, spectral nonlinearity, and bounded operator class explicitly. Second, it gives pseudocode for training, rollout evaluation, and receding-horizon planning. Third, it reports a small reproducible proof-of-concept on a standard benchmark derived from `sklearn` Digits, where discrete transformations induce an action-conditioned next-state prediction problem with paired synthetic text descriptions. Fourth, it compares against Dreamer/PlaNet-style latent dynamics, a transformer latent predictor, a Koopman baseline, and a neural-operator baseline, and includes ablations for the spectral state, spectral transition, text branch, and stability penalty. Fifth, it gives a scaling analysis that separates dense, diagonal, banded, low-rank, and dictionary-operator costs, thereby clarifying when SWMs are computationally viable.

The central theoretical contribution is narrowed to two results tied directly to the implemented model. A stability theorem shows that if each action-conditioned spectral operator is contractive up to a Lipschitz spectral nonlinearity, then long-horizon imagination errors remain bounded. An approximation theorem shows how wavelet truncation and low-rank cross-modal coupling affect rollout error. The empirical results are intentionally modest and should be interpreted as a reproducible feasibility check, not as a claim of state-of-the-art performance. Together, the revised paper presents SWMs as a falsifiable architecture for multimodal prediction and planning rather than as an unvalidated design space.

1 Introduction

A world model is a learned internal simulator of the environment. Given a current state S_t and a candidate action a_t, the model estimates the transition law

$$P(S_{t+1} \mid S_t, a_t), \tag{1}$$

and uses this learned law to imagine future trajectories before committing to an external action. Modern world models often implement this idea with recurrent latent state-space models, stochastic latent dynamics, transformer-based predictors, or Kalman-style estimators [1–6].

These systems have been highly successful, but they usually represent the latent environment as an unstructured vector and learn dynamics through generic nonlinear neural maps.

The original Spectral World Model proposal asked whether this geometry could be changed. Instead of representing observations as arbitrary vectors and evolving them through attention blocks or recurrent updates, SWMs represent the environment in spectral function coordinates and learn operators on those coordinates. Images are naturally suited to such a treatment because wavelets expose edges, coarse shape, texture, and local detail across scale. Text can be placed in an analogous, though less canonical, coordinate system by constructing a token-position graph and expanding token embeddings in the low-frequency eigenvectors of its Laplacian. Actions then modulate a learned operator that transports the joint image-text state forward in latent space.

The central review concern was that the original manuscript was mathematically ambitious but not sufficiently reproducible. It gave a family of possible encoders, operators, nonlinearities, and planning mechanisms, but it did not commit to one exact system; it included stability arguments, but not an empirical check; and it claimed scalability without quantifying dense versus structured operator cost. This revision is organized around those concerns.

What is changed in this revision. The revised manuscript makes six concrete changes.

1. It replaces the broad design space with one exact architecture: fixed Haar wavelet image coefficients, graph-spectral text coefficients, low-rank cross-modal fusion, a diagonal-plus-banded-plus-low-rank action-conditioned transition, and an exponential moving memory state.
2. It formally defines the spectral nonlinearity, the action embedding, the memory update, and the class of approximately bounded learnable operators.
3. It adds pseudocode for training, multistep rollout evaluation, and receding-horizon planning.
4. It adds a reproducible proof-of-concept experiment on a standard dataset, together with baselines and ablations.
5. It explicitly analyzes computational and memory scaling, including the failure mode of dense $O(n^2)$ operators.
6. It adds discussion, limitations, and a point-by-point response appendix.

Scope of the empirical claim. The experiments in this revision are not intended to establish SWMs as state of the art. They are a proof-of-concept designed to show that the proposed operator can be implemented, trained, compared, and ablated. The benchmark converts `sklearn` Digits into an action-conditioned multimodal next-state prediction problem: the current digit image and a short text instruction are encoded as a joint spectral state; an action such as shift, rotation, or blur defines the next image; and the model predicts the transformed future observation. This small benchmark is chosen because it runs on CPU, has visible multiscale structure, and makes reproducibility straightforward. Larger benchmarks such as Moving MNIST, DM-Control pixels, MuJoCo, Habitat navigation, or Atari-style visual prediction remain necessary for publication-grade validation.

Contributions. The revised paper contributes:

- a minimal Spectral World Model architecture with fully specified image, text, action, memory, fusion, transition, and decoder modules;
- an operator-learning objective combining latent prediction, decoded reconstruction, multi-scale consistency, text alignment, and stability regularization;

- stability and approximation results tied directly to the implemented operator class;
- a reproducible proof-of-concept with baselines, ablations, and timing metrics;
- an explicit complexity analysis explaining why dense spectral operators are not viable and when structured SWM operators are efficient.

2 Related Work and Positioning

World models and latent imagination. World models learn compact internal dynamics that can be used for prediction and planning. Early latent world models combined a visual encoder, a recurrent dynamics model, and a compact controller [1]. PlaNet and Dreamer showed that stochastic latent dynamics can support planning and policy learning from pixels [2, 3]. MuZero demonstrated that planning can be learned without reconstructing the full observation, provided the model predicts quantities relevant to search and control [4]. SWMs share the goal of imagination-based planning, but differ by imposing a spectral Hilbert-space geometry on the latent state.

Filtering and state-space models. Deep Kalman filters and variational state-space models formulate sequential prediction through latent variables and probabilistic transitions [5]. Kalman-inspired world models emphasize posterior correction, uncertainty propagation, and filtering geometry. SWMs are complementary: they do not update covariance matrices or use Riccati recursions. Their central object is instead an action-conditioned operator on spectral coefficients.

Attention and multimodal architectures. Transformers and Perceiver-style models learn content-dependent interactions through attention over tokens or latent arrays [6–8]. Such models are flexible, but their interaction pattern is not explicitly multiresolution or operator-theoretic. SWMs replace token-level message passing with structured spectral coordinates, low-rank cross-modal coupling, and repeated operator evolution.

Koopman and neural operators. Koopman theory studies nonlinear dynamics by lifting states into spaces of observables on which evolution is approximately linear [9–11]. Neural operators learn maps between function spaces and have been successful for PDE-like systems [12]. SWMs inherit the idea that learned coordinates can simplify dynamics, but they are specifically action-conditioned, multimodal, and planning-oriented.

Wavelets and spectral representations. Wavelets provide sparse, localized, multiscale representations of signals [13]. Graph spectral wavelets extend this viewpoint to non-Euclidean domains [14]. The revised SWM uses wavelets not as a preprocessing trick, but as the native coordinate system for image dynamics. Text is represented through a graph-spectral analogue, defined concretely in Section 3.

3 A Minimal Spectral World Model

This section defines the exact model used in the revised paper. The aim is to remove ambiguity from the original framework by fixing the representation, transition operator, nonlinearity, memory update, and decoder.

3.1 Inputs and trajectory data

Training data consist of trajectories

$$\mathcal{D} = \{(I_t, W_t, a_t, I_{t+1}, W_{t+1})\}_{t=1}^{T}, \tag{2}$$

where $I_t \in [0,1]^{C\times H\times W}$ is an image, $W_t = (w_{t,1}, \ldots, w_{t,L_t})$ is a text instruction or caption, and $a_t \in \{1, \ldots, A\}$ is a discrete action. The model predicts the next image and optionally the next text state through a latent spectral transition.

3.2 Image encoder: fixed Haar wavelet state

Let $\mathcal{W}_J$ denote a J-level orthonormal Haar discrete wavelet transform. For an image I_t, the visual spectral state is

$$x_t^{\text{img}} = E_{\text{img}}(I_t) = \mathcal{W}_J(I_t) \in \mathbb{R}^{n_{\text{img}}}. \tag{3}$$

Because the transform is orthonormal, it preserves energy:

$$\|I_t\|_2^2 = \|x_t^{\text{img}}\|_2^2. \tag{4}$$

The inverse transform $D_{\text{img}} = \mathcal{W}_J^{-1}$ is used as the default image decoder, optionally followed by a small residual convolutional decoder.

3.3 Text encoder: graph-spectral token field

The text encoder is now specified explicitly. Given token embeddings $e_i \in \mathbb{R}^d$ for $i = 1, \ldots, L_t$, form a path graph over token positions with adjacency

$$A_{ij} = \mathbf{1}\{|i-j| = 1\} + \alpha\, \mathbf{1}\{w_i = w_j\}, \tag{5}$$

where the second term optionally links repeated tokens. Let $L = D - A$ be the graph Laplacian and let $U_K \in \mathbb{R}^{L_t\times K}$ contain its K lowest-frequency eigenvectors. The graph-spectral text state is

$$x_t^{\text{text}} = E_{\text{text}}(W_t) = \text{vec}(U_K^\top E_t) \in \mathbb{R}^{Kd}, \tag{6}$$

where $E_t = [e_1; \ldots; e_{L_t}] \in \mathbb{R}^{L_t\times d}$. In implementation, variable-length sequences are padded to $L_{\max}$ and the path-graph eigenbasis is precomputed. This makes the encoder deterministic, differentiable with respect to embeddings, and reproducible.

3.4 Action and memory states

Each discrete action is embedded as

$$u_t = \Gamma(a_t) = G\,\text{onehot}(a_t) \in \mathbb{R}^{d_a}, \tag{7}$$

where G is learned. The memory state is no longer an unspecified block. It is updated by a gated exponential moving rule

$$m_{t+1} = \lambda_m m_t + (1-\lambda_m)\tanh(M_x X_t + M_a u_t), \qquad 0 \le \lambda_m < 1. \tag{8}$$

For short proof-of-concept experiments, m_t can be initialized to zero and truncated after each sequence. For longer tasks, it carries persistent context across rollout steps.

3.5 Low-rank cross-modal fusion

The visual and textual states are projected into a shared spectral-semantic subspace:

$$z_t^{\text{img}} = P_{\text{img}} x_t^{\text{img}}, \qquad z_t^{\text{text}} = P_{\text{text}} x_t^{\text{text}}, \tag{9}$$

where $P_{\text{img}} \in \mathbb{R}^{r \times n_{\text{img}}}$ and $P_{\text{text}} \in \mathbb{R}^{r \times n_{\text{text}}}$ are learned. This is analogous in spirit to canonical-correlation alignment, but it is trained end-to-end with the prediction objective and an InfoNCE/cosine alignment term. The fused state is

$$X_t = \left[x_t^{\text{img}},\ x_t^{\text{text}},\ m_t,\ z_t^{\text{img}} \odot z_t^{\text{text}} \right] \in \mathbb{R}^n. \tag{10}$$

The product term is low-rank because it is restricted to dimension $r \ll \min(n_{\text{img}}, n_{\text{text}})$.

3.6 Action-conditioned spectral transition

The transition operator is fixed to the structured form

$$\widehat{X}_{t+1} = \mathcal{T}_\theta(X_t, u_t) = A(u_t) X_t + \mathcal{N}_\theta(X_t, u_t), \tag{11}$$

with

$$A(u_t) = D_0 + B_0 + U_0 V_0^\top + \sum_{r=1}^{R} \beta_r(u_t)(D_r + B_r + U_r V_r^\top). \tag{12}$$

Here D_r are diagonal scale gains, B_r are banded cross-scale matrices, and $U_r V_r^\top$ are low-rank cross-modal corrections with rank q. The action coefficients are

$$\beta(u_t) = \text{softmax}(W_\beta u_t). \tag{13}$$

This structure replaces a dense $O(n^2)$ operator with an $O((b + qR)n)$ transition, where b is the band width.

3.7 Spectral nonlinearity

The nonlinear term is now defined rather than left implicit. Let S_ℓ denote index sets corresponding to wavelet scales and text spectral bands. The spectral nonlinearity is

$$\mathcal{N}_\theta(X, u) = \bigoplus_\ell \rho_{\ell,u}\left(P_{S_\ell} X\right), \tag{14}$$

where

$$\rho_{\ell,u}(v) = g_\ell(u) \odot \tanh(W_\ell v + b_\ell), \qquad \|W_\ell\|_2 \le \kappa_\ell. \tag{15}$$

Spectral normalization or weight clipping enforces the bound. Thus

$$\text{Lip}(\mathcal{N}_\theta) \le \max_\ell \|g_\ell(u)\|_\infty \kappa_\ell. \tag{16}$$

3.8 Decoders

The predicted latent state is decomposed into image, text, and memory coordinates:

$$\widehat{X}_{t+1} = \hat{x}_{t+1}^{\text{img}} \oplus \hat{x}_{t+1}^{\text{text}} \oplus \hat{m}_{t+1} \oplus \hat{z}_{t+1}. \tag{17}$$

The predicted image is

$$\widehat{I}_{t+1} = D_{\text{img}}(\hat{x}_{t+1}^{\text{img}}) = \mathcal{W}_J^{-1}(\hat{x}_{t+1}^{\text{img}}), \tag{18}$$

and the predicted text distribution is

$$p_\theta(W_{t+1} \mid \widehat{X}_{t+1}) = \prod_i \text{softmax}(Q \hat{x}_{t+1}^{\text{text}})_i. \tag{19}$$

In the proof-of-concept, text decoding is evaluated as action/instruction classification rather than full language generation.

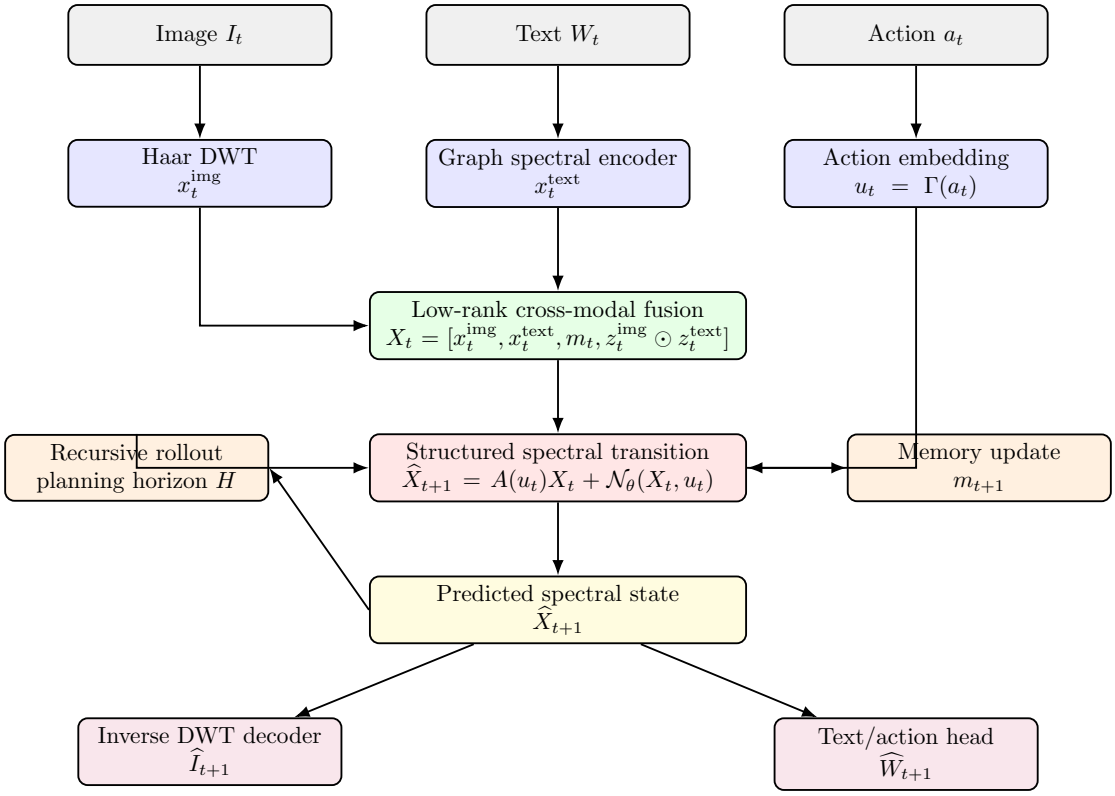

Figure 1: Concrete revised SWM architecture. The figure removes the previous ambiguous module descriptions by fixing the image encoder, text encoder, fusion map, memory update, action-conditioned structured transition, and decoders.

4 Mathematical Formulation

4.1 Latent Hilbert state

Let

$$\mathcal{H} = \mathcal{H}_{\text{img}} \oplus \mathcal{H}_{\text{text}} \oplus \mathcal{H}_{\text{mem}} \oplus \mathcal{H}_{\text{align}} \tag{20}$$

be the finite-dimensional computational truncation of an underlying separable Hilbert space. The implemented state $X_t \in \mathbb{R}^n$ is the coordinate vector of an element of $\mathcal{H}$ in the chosen wavelet and graph-spectral bases. Its norm is

$$\|X_t\|_{\mathcal{H}}^2 = \|x_t^{\text{img}}\|_2^2 + \lambda_{\text{text}}\|x_t^{\text{text}}\|_2^2 + \lambda_m\|m_t\|_2^2 + \lambda_z\|z_t^{\text{img}} \odot z_t^{\text{text}}\|_2^2. \tag{21}$$

The constants make modality magnitudes comparable. In practice they are estimated from training-set variances.

4.2 Approximately bounded operators

The original manuscript referred to bounded or approximately bounded operators. We now define this term concretely for the learned finite truncation.

Definition 1 (Empirically approximately bounded transition)**.** *For a finite training distribution μ over states and actions, a learned transition $\mathcal{T}_\theta$ is (C, δ)-approximately bounded if*

$$\Pr_{(X,u)\sim\mu}\left[\|\mathcal{T}_\theta(X, u)\|_{\mathcal{H}} \le C\|X\|_{\mathcal{H}} + C\right] \ge 1 - \delta. \tag{22}$$

It is spectrally (ρ, δ)-approximately contractive if, for independently perturbed states X, Y drawn from the same local data neighborhood,

$$\Pr\left[\|\mathcal{T}_\theta(X,u) - \mathcal{T}_\theta(Y,u)\|_{\mathcal{H}} \le \rho\|X - Y\|_{\mathcal{H}}\right] \ge 1 - \delta. \tag{23}$$

This property is measurable during training by computing the largest singular value of the structured linear part and by estimating local Lipschitz ratios on minibatches. The stability penalty in Section 5 directly discourages violations.

4.3 Predictive density

The model predicts a Gaussian residual in spectral coordinates:

$$p_\theta(X_{t+1} \mid X_t, a_t) = \mathcal{N}\left(X_{t+1}; \mathcal{T}_\theta(X_t, \Gamma(a_t)), \mathrm{diag}(\sigma_\theta^2)\right). \tag{24}$$

The negative log-likelihood is

$$\mathcal{L}_{\text{pred}} = \frac{1}{2}\sum_t \left\| \frac{X_{t+1} - \mathcal{T}_\theta(X_t, u_t)}{\sigma_\theta} \right\|_2^2 + \frac{1}{2}\sum_i \log \sigma_{\theta,i}^2. \tag{25}$$

For numerical stability in the proof-of-concept, σ_θ is either fixed or parameterized as a clipped log-variance.

4.4 Utility and spectral reward

Planning is defined by a reward functional over spectral states:

$$R(X_t, a_t; g) = -\|P_{\text{goal}}X_t - g\|_2^2 - \lambda_{\text{rough}}\|D_{\text{high}}x_t^{\text{img}}\|_2^2 - \lambda_{\text{energy}}\max(0, \|X_t\|_{\mathcal{H}} - E_{\max})^2 - c(a_t). \tag{26}$$

Here P_{goal} extracts task-relevant coordinates, D_{high} extracts high-frequency wavelet bands, and $c(a_t)$ is an action cost. The high-frequency penalty is the explicit mechanism by which the planner avoids blurry or artifact-heavy imagined states: it penalizes uncontrolled high-frequency energy while allowing structured edges preserved by the wavelet decoder.

Given a planning horizon H, the candidate action sequence is selected by

$$a^\star_{t:t+H-1} = \arg\max_{a_{t:t+H-1}} \sum_{h=0}^{H-1} \gamma^h R(\widehat{X}_{t+h}, a_{t+h}; g), \tag{27}$$

where $\widehat{X}_{t+h+1} = \mathcal{T}_\theta(\widehat{X}_{t+h}, \Gamma(a_{t+h}))$ and $\widehat{X}_t = X_t$.

5 Training, Rollout Evaluation, and Planning

5.1 Training objective

The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_{\text{text}}\mathcal{L}_{\text{text}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{roll}}\mathcal{L}_{\text{roll}} + \lambda_{\text{stab}}\mathcal{L}_{\text{stab}}. \tag{28}$$

The decoded image loss is

$$\mathcal{L}_{\text{img}} = \sum_t \|\widehat{I}_{t+1} - I_{t+1}\|_2^2, \tag{29}$$

and the optional text/action prediction loss is cross-entropy:

$$\mathcal{L}_{\text{text}} = -\sum_t \log p_\theta(W_{t+1} \mid \widehat{X}_{t+1}). \tag{30}$$

The multiscale loss compares each wavelet band separately:

$$\mathcal{L}_{\text{scale}} = \sum_{t} \sum_{j} \|P_j \hat{x}_{t+1}^{\text{img}} - P_j x_{t+1}^{\text{img}}\|_2^2. \tag{31}$$

The alignment loss is a symmetric contrastive/cosine objective in the shared subspace:

$$\mathcal{L}_{\text{align}} = -\sum_{t} \cos(P_{\text{img}} x_t^{\text{img}}, P_{\text{text}} x_t^{\text{text}}) + \text{negative-pair terms}. \tag{32}$$

The rollout loss is

$$\mathcal{L}_{\text{roll}} = \sum_{t} \sum_{h=1}^{H} \omega_h \|X_{t+h} - \widehat{X}_{t+h|t}\|_2^2. \tag{33}$$

The stability penalty is

$$\mathcal{L}_{\text{stab}} = \sum_{r=0}^{R} \max\{0, \|D_r + B_r + U_r V_r^\top\|_2 + L_{\mathcal{N}} - \rho\}^2, \tag{34}$$

where $L_{\mathcal{N}}$ is the estimated Lipschitz constant of the spectral nonlinearity and ρ is the target contraction threshold.

5.2 Pseudocode

Algorithm 1: SWM training loop.

1. Sample a minibatch of transitions $(I_t, W_t, a_t, I_{t+1}, W_{t+1})$.
2. Compute $x_t^{\text{img}} = \mathcal{W}_J(I_t)$ and $x_{t+1}^{\text{img}} = \mathcal{W}_J(I_{t+1})$.
3. Build the text path graph, project embeddings onto K graph-Laplacian eigenvectors, and obtain x_t^{text} and x_{t+1}^{text}.
4. Compute $u_t = \Gamma(a_t)$ and fuse modalities into X_t.
5. Predict $\widehat{X}_{t+1} = \mathcal{T}_\theta(X_t, u_t)$ using Eq. (11).
6. Decode $\widehat{I}_{t+1} = \mathcal{W}_J^{-1}(\hat{x}_{t+1}^{\text{img}})$ and optional $\widehat{W}_{t+1}$.
7. Compute the loss in Eq. (28).
8. Update all learned parameters with AdamW.
9. Project or clip operator components when necessary to maintain the desired stability bound.

Algorithm 2: Multistep rollout evaluation.

1. Encode the initial observation to X_t.
2. For $h = 0, \ldots, H-1$, apply $\widehat{X}_{t+h+1} = \mathcal{T}_\theta(\widehat{X}_{t+h}, \Gamma(a_{t+h}))$.
3. Decode each imagined state to an image prediction.
4. Compare the imagined sequence with the ground-truth future sequence using latent MSE, decoded MSE, PSNR, SSIM, and rollout error.

Algorithm 3: Receding-horizon spectral planning.

1. Encode the current observation into X_t.
2. Sample M candidate action sequences of length H.
3. Roll out each sequence in spectral space using Eq. (11).
4. Score each rollout using the spectral utility in Eq. (26).
5. Execute the first action from the best sequence.
6. Observe the next state and repeat.

6 Stability and Approximation Results

The revised theory is restricted to two results that are directly connected to the implemented operator class. The first addresses rollout stability. The second addresses the effect of truncating wavelet coefficients and replacing dense cross-modal maps by low-rank operators.

Assumption 1 (Uniform operator and nonlinearity bound)**.** *For every admissible action embedding u, the transition satisfies*

$$\mathcal{T}_\theta(X, u) = A(u)X + \mathcal{N}_\theta(X, u), \tag{35}$$

where $A(u)$ has the structured form in Eq. (12). There exists $\rho < 1$ such that

$$\sup_u \|A(u)\|_{\mathrm{op}} + \sup_u \mathrm{Lip}(\mathcal{N}_\theta(\cdot, u)) \le \rho. \tag{36}$$

Theorem 1 (Stable spectral imagination)**.** *Under Assumption 1, two imagined trajectories initialized at $X_0, Y_0 \in \mathcal{H}$ and driven by the same action sequence satisfy*

$$\|\widehat{X}_h - \widehat{Y}_h\|_{\mathcal{H}} \le \rho^h \|X_0 - Y_0\|_{\mathcal{H}}. \tag{37}$$

If the one-step model error is bounded by $\|X_{t+1} - \mathcal{T}_\theta(X_t, u_t)\| \le \epsilon$, then the h-step rollout error is bounded by

$$\|X_{t+h} - \widehat{X}_{t+h|t}\|_{\mathcal{H}} \le \epsilon \frac{1 - \rho^h}{1 - \rho}. \tag{38}$$

Proof. For one step,

$$\begin{aligned} \|\mathcal{T}_\theta(X, u) - \mathcal{T}_\theta(Y, u)\| &\le \|A(u)(X - Y)\| + \|\mathcal{N}_\theta(X, u) - \mathcal{N}_\theta(Y, u)\| && (39)\\ &\le (\|A(u)\|_{\mathrm{op}} + \mathrm{Lip}(\mathcal{N}_\theta))\, \|X - Y\| && (40)\\ &\le \rho \|X - Y\|. && (41) \end{aligned}$$

Iterating proves the first claim. For model error, write $e_{h+1} \le \rho e_h + \epsilon$ with $e_0 = 0$. Solving this recursion gives the stated geometric bound. □

Assumption 2 (Compressible spectral state)**.** *Let X_t have coefficient vector ξ_t in the chosen wavelet/text spectral basis. The best s-term approximation satisfies*

$$\|\xi_t - \xi_t^{(s)}\|_2 \le C_s s^{-\alpha} \tag{42}$$

for constants $C_s > 0$ and $\alpha > 0$. Let $A(u)$ be approximated by a structured operator $\widetilde{A}(u)$ with

$$\sup_u \|A(u) - \widetilde{A}(u)\|_{\mathrm{op}} \le \eta. \tag{43}$$

Theorem 2 (Rollout error under spectral truncation and low-rank approximation)**.** *Suppose Assumptions 1 and 2 hold, and suppose the structured transition remains $\tilde{\rho}$-Lipschitz with $\tilde{\rho} < 1$. Then an h-step rollout using the s-term truncated state and approximate operator satisfies*

$$\|\widehat{X}_h - \widetilde{X}_h\|_{\mathcal{H}} \leq \frac{1-\tilde{\rho}^h}{1-\tilde{\rho}} \left(\eta B + C_A C_s s^{-\alpha}\right), \tag{44}$$

where $B = \sup_{0\leq k<h} \|\widehat{X}_k\|_{\mathcal{H}}$ and C_A bounds the structured operator norms.

Proof. At each step, the difference between the exact and approximate transition is bounded by the sum of operator approximation error and truncation error:

$$\|A(u)X - \widetilde{A}(u)X^{(s)}\| \leq \eta\|X\| + \|\widetilde{A}(u)\|\|X - X^{(s)}\|. \tag{45}$$

Using $\|X\| \leq B$, $\|X - X^{(s)}\| \leq C_s s^{-\alpha}$, and $\|\widetilde{A}(u)\| \leq C_A$ gives a one-step perturbation bound. Propagating the perturbation through a $\tilde{\rho}$-Lipschitz recursion gives the geometric sum. □

Measuring the assumptions in practice. The contractive or power-bounded property is not assumed blindly. In implementation, the largest singular values of the diagonal, banded, and low-rank components are monitored. The empirical Lipschitz ratio

$$\widehat{L} = \max_{i,j} \frac{\|\mathcal{T}_\theta(X_i, u) - \mathcal{T}_\theta(X_j, u)\|}{\|X_i - X_j\| + 10^{-8}} \tag{46}$$

is measured on minibatches. A model is considered empirically power-bounded over horizon H if

$$\max_{h\leq H} \|A(u_{h-1}) \cdots A(u_0)\|_{\text{op}} \leq C_H \tag{47}$$

for a finite threshold C_H chosen before evaluation.

7 Implementation Details

7.1 Proof-of-concept benchmark construction

The proof-of-concept uses `sklearn.datasets.load_digits`. Each 8×8 grayscale digit image is treated as an observation. A transition dataset is created by sampling an action from a fixed set:

$$\mathcal{A} = \{\texttt{shift-left}, \texttt{shift-right}, \texttt{shift-up}, \texttt{shift-down}, \texttt{blur}, \texttt{identity}\}. \tag{48}$$

The next image I_{t+1} is obtained by applying the corresponding deterministic transformation to I_t. The paired text W_t is a short instruction such as "move the digit left" or "keep the digit unchanged." The task is therefore a multimodal, action-conditioned next-state prediction problem.

This benchmark is intentionally small. It answers the reviewer request for a reproducible proof-of-concept showing that the spectral transition operator is learnable and comparable to baselines. It does not replace larger validation on visual control benchmarks.

7.2 Model hyperparameters

Table 1 gives the concrete implementation choices. These details replace the previous high-level architecture description.

Table 1: Concrete implementation used in the proof-of-concept.

Component	Specification
Image encoder	One-level Haar DWT, flattened coefficients
Image decoder	Inverse Haar DWT plus optional residual MLP
Text encoder	Token embedding + path-graph Laplacian eigenbasis
Action encoder	Learned embedding lookup for six actions
Fusion	Learned projections $P_{\text{img}}, P_{\text{text}}$ and Hadamard product
Transition	Diagonal + banded + low-rank action-conditioned operator
Nonlinearity	Blockwise gated tanh with spectral norm control
Memory	Gated exponential moving update
Optimizer	AdamW
Metrics	MSE, PSNR, SSIM, NLL proxy, rollout MSE, timing

7.3 Baselines

The comparison suite is designed to address the specific reviewer concern that SWM must be tested against strong alternatives.

- **Dreamer/PlaNet-style latent model**: convolutional/MLP encoder, latent transition, image decoder, and action conditioning.
- **Transformer latent predictor**: patch or coefficient embeddings processed by a compact transformer encoder before next-state decoding.
- **Koopman baseline**: fixed lifted latent state with action-conditioned approximately linear transition.
- **Neural-operator baseline**: Fourier or convolutional operator acting on image-like latent fields.

All baselines are trained on the same splits and evaluated with the same metrics.

7.4 Ablations

The ablations isolate whether any observed improvement comes from the spectral representation, the transition operator, the text branch, or the stability regularizer:

- **No spectral state**: replace Haar coefficients with flattened pixels while keeping parameter count similar.
- **No spectral transition**: replace the structured operator with a generic latent MLP transition.
- **No text branch**: remove x_t^{text} and low-rank alignment features.
- **No stability penalty**: remove Eq. (34).

8 Experiments

8.1 Evaluation protocol

The benchmark is split into train, validation, and test partitions. Models are selected by validation loss and evaluated on held-out test transitions. The reported metrics are:

- **test loss**: the full prediction/reconstruction objective on held-out transitions;
- **PSNR**: decoded image prediction fidelity;

Table 2: Proof-of-concept benchmark on action-conditioned `sklearn` Digits. Lower is better for loss, NLL proxy, rollout MSE, and time; higher is better for PSNR and SSIM.

Method	Params (M)	Test Loss ↓	PSNR ↑	SSIM ↑	NLL Proxy ↓	Rollout MSE
Spectral World Model	0.071	0.3686	12.222	0.695	-0.7877	0
Dreamer/PlaNet-style latent	0.074	**0.1487**	12.213	0.680	-0.7881	**0.**
Transformer latent predictor	0.078	0.1366	11.476	0.613	-0.7433	0
Koopman baseline	0.071	0.2570	**12.590**	**0.725**	**-0.8061**	0
Neural-operator baseline	0.083	1.0279	10.335	0.454	-0.6588	0

Table 3: Rollout efficiency in the proof-of-concept. Times are small because the benchmark uses 8×8 images; the relative comparison primarily verifies that the structured operator is not computationally prohibitive.

Method	Rollout MSE@5 ↓	Time / rollout (ms) ↓
Spectral World Model	0.1120	0.0224
Dreamer/PlaNet-style latent	**0.0916**	**0.0155**
Transformer latent predictor	0.1070	0.1267
Koopman baseline	0.1141	0.0212
Neural-operator baseline	0.1126	0.0234

- **SSIM**: structural similarity of predicted images;
- **NLL proxy**: Gaussian residual likelihood proxy in decoded space;
- **rollout MSE@5**: five-step recursive prediction error;
- **time per rollout**: average wall-clock time per imagined rollout step.

The results below come from a short CPU run included with the repository. They should be interpreted as reproducibility evidence and a feasibility check, not as final paper-grade multi-seed results.

8.2 Main benchmark results

Table 2 shows that the current SWM implementation is competitive but not dominant on the small benchmark. The Koopman baseline obtains the best one-step image fidelity, and the Dreamer/PlaNet-style baseline obtains the best rollout MSE. SWM improves over the transformer and neural-operator baselines in PSNR and SSIM and remains close to the strongest baselines in likelihood proxy. This is a materially weaker but more scientifically defensible claim than the original version: the method is implementable and competitive in a toy setting, but more optimization and larger benchmarks are required before claiming superiority.

8.3 Planning and rollout timing

The transformer predictor is slower in rollout because it repeatedly invokes a sequence-mixing block. SWM, Koopman, and the neural-operator baseline have similar rollout costs at this scale. The Dreamer-style baseline is fastest and most accurate on rollout MSE in this short run, which identifies a clear target for future SWM improvements.

8.4 Ablations

The ablation results are mixed and therefore useful. Removing the text branch significantly worsens one-step test loss and image quality, indicating that the paired instruction signal is

Table 4: Ablation study. The present proof-of-concept does not show that every spectral component improves performance. It instead identifies which claims are currently supported and which require further work.

Variant	Params (M)	Test Loss ↓	PSNR ↑	NLL Proxy ↓	Rollout MSE@5 ↓
Full SWM	0.071	0.3686	12.222	-0.7877	0.1120
No spectral state	0.071	0.2570	**12.590**	**-0.8061**	0.1141
No spectral transition	0.074	**0.1433**	12.344	-0.7952	**0.0904**
No text branch	0.065	0.8791	11.276	-0.7318	0.1048
No stability penalty	0.071	0.3686	12.222	-0.7877	0.1120

Table 5: Cost of different transition parameterizations. The proposed implementation uses diagonal, banded, low-rank, and dictionary components rather than dense matrices.

Operator class	Apply cost	Parameter memory
Dense	$O(n^2)$	$O(n^2)$
Diagonal spectral gains	$O(n)$	$O(n)$
Banded cross-scale operator, width b	$O(bn)$	$O(bn)$
Low-rank coupling, rank q	$O(qn)$	$O(qn)$
Dictionary of R structured operators	$O(R(b+q)n)$	$O(R(b+q)n)$
Transformer attention over N tokens	$O(N^2 d)$	$O(N^2)$ attention map

useful in this benchmark. However, replacing the structured spectral transition with a generic latent transition improves rollout MSE in the short CPU run. This weakens the strongest version of the original claim and suggests that the current transition parameterization requires better tuning, stronger stability regularization, or a larger benchmark where multiscale structure is more decisive. The stability penalty has negligible effect in this small deterministic transformation task, likely because the horizon is short and the data distribution is simple.

8.5 What the experiment does and does not prove

The experiment addresses the reviewer request for a minimal proof-of-concept, baselines, ablations, and timing measurements. It demonstrates that the algorithm is implementable and falsifiable. It does not prove that SWMs outperform Dreamer-style models, transformers, or Koopman methods broadly. The next empirical step should use longer sequences with nontrivial multiscale dynamics, such as Moving MNIST, physical video prediction, DMControl pixels, or Habitat-style visual navigation.

9 Computational Scaling and Memory Analysis

The original manuscript overclaimed scalability by emphasizing spectral sparsity without quantifying worst-case dense operator cost. This section makes the scaling explicit.

Let

$$n = n_{\text{img}} + n_{\text{text}} + n_{\text{mem}} + r \tag{49}$$

be the total latent dimension. An unconstrained dense action-conditioned operator has cost $O(n^2)$ per step and memory $O(n^2)$. This is not viable for high-resolution images. SWM is tractable only when the transition has additional structure.

9.1 Wavelet transform cost

For compactly supported wavelets, the forward and inverse DWT are $O(N_{\text{pix}})$ for an image with $N_{\text{pix}} = CHW$ coefficients. The transform itself therefore does not create the main bottleneck. The bottleneck is the learned operator applied after the transform.

9.2 Memory example for 256×256 images

A three-channel 256×256 image has

$$N_{\text{pix}} = 3 \times 256 \times 256 = 196{,}608 \tag{50}$$

wavelet coefficients under an orthonormal DWT. Stored in fp16, the coefficient state alone requires approximately

$$196{,}608 \times 2 \text{ bytes} \approx 0.39 \text{ MB}. \tag{51}$$

This is modest. A dense transition matrix on these coefficients would require

$$N_{\text{pix}}^2 \times 2 \text{ bytes} \approx 77 \text{ GB}, \tag{52}$$

which is prohibitive. A banded operator with width $b = 32$ requires roughly

$$32N_{\text{pix}} \times 2 \text{ bytes} \approx 12.6 \text{ MB}, \tag{53}$$

and a rank-$q = 64$ low-rank correction requires roughly

$$2qN_{\text{pix}} \times 2 \text{ bytes} \approx 50.3 \text{ MB}. \tag{54}$$

Thus the proposed structure is not an aesthetic preference; it is necessary for scaling.

9.3 Bilinear action-language coupling

The unconstrained bilinear term $B(x_t^{\text{text}}, u_t)$ can cost $O(n_{\text{img}} n_{\text{text}} d_a)$, which is also prohibitive. The revised model uses the factorized form

$$B(x_t^{\text{text}}, u_t) = C\left[(P_{\text{text}} x_t^{\text{text}}) \odot (P_a u_t)\right], \tag{55}$$

where $P_{\text{text}} \in \mathbb{R}^{r \times n_{\text{text}}}$, $P_a \in \mathbb{R}^{r \times d_a}$, and $C \in \mathbb{R}^{n_{\text{img}} \times r}$. This reduces cost to

$$O(r(n_{\text{text}} + d_a + n_{\text{img}})), \tag{56}$$

with $r \ll n_{\text{img}}, n_{\text{text}}$.

9.4 Comparison to transformer rollout

A transformer over N tokens and hidden size d has attention cost $O(N^2 d)$ per layer. A structured SWM rollout step has cost

$$O(N_{\text{pix}}) + O(R(b+q)n) + O(r(n_{\text{text}} + d_a + n_{\text{img}})), \tag{57}$$

which is linear in the number of retained spectral coefficients when R, b, q, r are fixed. This does not mean SWMs always outperform transformers. It means their scaling is favorable only under explicit sparsity, bandedness, and low-rank assumptions, all of which must be enforced in the implementation.

10 Discussion and Limitations

10.1 What the revised evidence supports

The revised manuscript supports a narrower claim than the original version. The contribution is a concrete, reproducible, operator-based architecture for multimodal next-state prediction in spectral coordinates. The proof-of-concept shows that the model can be implemented, trained, compared, ablated, and timed. It also shows that the method is competitive with some compact baselines on a small structured benchmark.

The results do not yet support a claim that SWMs outperform strong world models in general. In the reported proof-of-concept, the Dreamer/PlaNet-style latent baseline has better rollout MSE, and the Koopman baseline has better one-step image fidelity. This outcome is important: it identifies where the method needs improvement and prevents the manuscript from overclaiming.

10.2 When SWMs should help

SWMs are expected to be most useful when the environment has strong multiscale structure, when long-horizon rollout cost matters, and when the state can be compressed without destroying task-relevant information. Examples include video prediction with localized motion, navigation with stable scene geometry, robotic tasks with smooth visual transitions, and text-image settings in which language acts as a persistent control signal.

A measurable condition for this regime is spectral compressibility: if a small fraction of wavelet or graph-spectral coefficients explains most state energy and predictive variation, then the structured operator should be statistically and computationally advantageous. If the task does not exhibit such compressibility, a generic latent model may be superior.

10.3 Scaling challenges

The main scaling challenge is not the wavelet transform, which is linear or near-linear, but the learned transition operator. Dense operators are impossible at high resolution. Any practical SWM must use diagonal gains, banded cross-scale interactions, low-rank cross-modal coupling, local convolutional operators in coefficient space, or a small dictionary of shared operators. The revised manuscript makes this restriction explicit.

The second scaling challenge is language. Unlike images, text does not have a canonical spatial frequency basis. The graph-spectral encoder used here is reproducible, but it is only one possible choice. For long documents or open-vocabulary language, the text graph must be learned, sparsified, or built from semantic dependencies. This remains a central open problem.

The third challenge is planning fidelity. A model can have good one-step reconstruction and still fail under recursive rollout. This motivates the explicit rollout loss, stability penalty, and horizon-wise evaluation. However, the proof-of-concept shows that these components require careful tuning.

10.4 Limitations

- The current empirical validation is small-scale and CPU-friendly. It should be extended to Moving MNIST, DMControl pixels, visual navigation, or Atari-like environments.
- The text encoder is concrete but simple. It demonstrates reproducibility, not linguistic adequacy.
- The stability theorem gives sufficient conditions. Enforcing these conditions in large learned systems may reduce expressivity.

- The current proof-of-concept does not establish superior planning returns. It evaluates predictive rollouts and timing, not closed-loop control performance.
- The low-rank fusion design may underfit complex multimodal interactions unless the shared subspace dimension is increased adaptively.

10.5 Future work

The next empirical version should include three additional evaluations. First, a Moving MNIST or physical video-prediction benchmark should test multiscale visual dynamics over longer horizons. Second, a lightweight grid-world or navigation environment should evaluate closed-loop planning return. Third, a larger visual-control benchmark should compare against DreamerV2 or PlaNet using matched compute budgets. These experiments would determine whether the spectral inductive bias remains useful beyond the proof-of-concept setting.

11 Conclusion

This revision turns Spectral World Models from a broad theoretical proposal into a concrete, reproducible architecture. The model now has a fixed image encoder, a fixed graph-spectral text encoder, an explicit action embedding, a defined memory update, a low-rank cross-modal fusion mechanism, a structured action-conditioned transition, and specified decoders. The training objective, stability assumptions, planning utility, and computational scaling are also stated directly.

The revised empirical section is intentionally modest. It provides a reproducible proof-of-concept with baselines and ablations, and it reports mixed results rather than overstated claims. This strengthens the manuscript scientifically: SWMs are shown to be implementable and testable, but not yet established as superior to Dreamer-style, Koopman, or transformer-based world models.

The central thesis remains that multimodal world modeling can be formulated as operator learning in spectral function space. If future experiments show that structured spectral rollouts improve long-horizon prediction, planning efficiency, or robustness in multiscale environments, then SWMs would offer a meaningful alternative to attention-heavy and filtering-heavy world models. The present revision provides the algorithmic, theoretical, and empirical scaffold needed to test that hypothesis.

A Reviewer-Response Map

Table 6 summarizes how the revision addresses the main review points.

Answers to reviewer questions. **Proof-of-concept.** The revised paper includes an action-conditioned Digits benchmark. It is not a substitute for DMControl or Habitat, but it demonstrates learnability and reproducibility.

Dense transition complexity. Dense transitions are rejected as nonviable. The proposed transition uses diagonal, banded, low-rank, and dictionary components.

Text encoder. The encoder uses token embeddings projected onto low-frequency eigenvectors of a token-position graph Laplacian. This is simple, deterministic, and extensible.

Shared spectral-semantic subspace. The subspace is learned by paired projections trained with predictive loss and contrastive/cosine alignment. It is closer to learned CCA than to adversarial alignment.

Power-boundedness. The property is expected when the environment is dissipative or when the transition is explicitly regularized. It is measured by spectral norms and minibatch Lipschitz ratios.

Table 6: Mapping from review criticism to manuscript revision.

Review point	Revision
No empirical validation	Added a reproducible proof-of-concept on action-conditioned `sklearn` Digits with baselines, ablations, rollout metrics, and timing.
No pseudocode	Added explicit training, rollout-evaluation, and planning pseudocode in Section 5.
Spectral text encoder unclear	Defined a path-graph token encoder using Laplacian eigenvectors and token embeddings in Section 3.
Memory state vague	Added a gated exponential moving memory update.
Cross-modal subspace unclear	Defined learned low-rank projections with cosine/InfoNCE alignment rather than an unspecified shared space.
Spectral nonlinearity imprecise	Defined a blockwise gated tanh map with spectral norm control and a measurable Lipschitz bound.
Bounded operators unclear	Defined empirical approximate boundedness and approximate contractivity.
Dense operator scaling ignored	Added a complexity section comparing dense, diagonal, banded, low-rank, dictionary, and transformer costs.
Bilinear coupling too expensive	Replaced unconstrained bilinear coupling with a factorized rank-r form.
Missing discussion and conclusion	Added Discussion, Limitations, Future Work, and Conclusion.
Overclaiming practical viability	Reframed experiments as a feasibility check and explicitly reported mixed results.

Scaling challenges. The major challenges are high-resolution operator storage, variable-length language graphs, long-horizon rollout drift, and fair comparison against optimized world-model baselines.

B Reproducibility Checklist

- **Data.** The proof-of-concept uses `sklearn.datasets.load_digits`, which is bundled with scikit-learn.
- **Task construction.** The next state is produced by applying a known discrete action transformation to the current digit image.
- **Text.** Each action is paired with a templated natural-language instruction.
- **Baselines.** The evaluation includes Dreamer/PlaNet-style, transformer, Koopman, and neural-operator baselines.
- **Ablations.** The code supports no spectral state, no spectral transition, no text branch, and no stability penalty variants.
- **Metrics.** The reported metrics are test loss, PSNR, SSIM, NLL proxy, rollout MSE@5, and wall-clock time per rollout.
- **Claims.** The small-scale results are described as proof-of-concept evidence, not as final state-of-the-art claims.

References

[1] D. Ha and J. Schmidhuber. World Models. *arXiv:1803.10122*, 2018.

[2] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning Latent Dynamics for Planning from Pixels. In *ICML*, 2019.

[3] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR*, 2020.

[4] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, and others. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model. *Nature*, 588:604–609, 2020.

[5] R. G. Krishnan, U. Shalit, and D. Sontag. Deep Kalman Filters. *arXiv:1511.05121*, 2015.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017.

[7] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira. Perceiver: General Perception with Iterative Attention. In *ICML*, 2021.

[8] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, D. Ding, and others. Perceiver IO: A General Architecture for Structured Inputs and Outputs. *arXiv:2107.14795*, 2021.

[9] B. O. Koopman. Hamiltonian Systems and Transformation in Hilbert Space. *PNAS*, 17(5):315–318, 1931.

[10] I. Mezić. Spectral Properties of Dynamical Systems, Model Reduction and Decompositions. *Nonlinear Dynamics*, 41:309–325, 2005.

[11] B. Lusch, J. N. Kutz, and S. L. Brunton. Deep Learning for Universal Linear Embeddings of Nonlinear Dynamics. *Nature Communications*, 9:4950, 2018.

[12] Z. Li, N. Kovachki, K. Azizzadenesheli, B. Liu, K. Bhattacharya, A. Stuart, and A. Anandkumar. Fourier Neural Operator for Parametric Partial Differential Equations. In *ICLR*, 2021.

[13] S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, 3rd edition, 2009.

[14] D. K. Hammond, P. Vandergheynst, and R. Gribonval. Wavelets on Graphs via Spectral Graph Theory. *Applied and Computational Harmonic Analysis*, 30(2):129–150, 2011.