Sat-JEPA-Diff: Bridging Self-Supervised Learning and Generative Diffusion for Remote Sensing

Kursat Komurcu, Linas Petkevicius *
Vilnius University
Faculty of Mathematics and Informatics
Institute of Computer Science
Artificial Intelligence Methods Lab
Vilnius, LT-03225, Lithuania
{kursat.komurcu,linas.petkevicius}@mif.vu.lt

Abstract

Predicting satellite imagery requires a balance between structural accuracy and textural detail. Standard deterministic methods like PredRNN or SimVP minimize pixel-based errors but suffer from the "regression to the mean" problem, producing blurry outputs that obscure subtle geographic-spatial features. Generative models provide realistic textures but often misleadingly reveal structural anomalies. To bridge this gap, we introduce Sat-JEPA-Diff, which combines Self-Supervised Learning (SSL) with Latent Diffusion Models (LDM). An IJEPA module predicts stable semantic representations, which then route a frozen Stable Diffusion backbone via a lightweight cross-attention adapter. This ensures that the synthesized high-accuracy textures are based on absolutely accurate structural predictions. Evaluated on a global Sentinel-2 dataset, Sat-JEPA-Diff excels at resolving sharp boundaries. It achieves leading perceptual scores (GSSIM: 0.8984, FID: 0.1475) and significantly outperforms deterministic baselines, despite standard autoregressive stability limits. The code and dataset are publicly available on GitHub.

1 Introduction

Continuous Earth observation is crucial for environmental monitoring but hindered by frequent cloud cover. Therefore, spatiotemporal forecasting ($t \rightarrow t+1$) becomes an indispensable "virtual sensor" that helps bridge the gap. Current forecasting models face a fundamental trade-off. Deterministic methods, such as PredRNN Wang et al. (2022) and SimVP Gao et al. (2022), focus on optimizing pixel-wise error (MSE). This, however, results in "regression toward the mean" with blurry images that lack spectral detail. On the contrary, generative forecasting approaches, like Denoising Diffusion Probabilistic Models (DDPMs) Rombach et al. (2022), excel in reproducing plausible textures. Nevertheless, these are often "hallucinations" that generate plausible but false structures, especially without adequate semantic guidance.

In this paper, we propose Sat-JEPA-Diff, a novel spatiotemporal forecasting model that integrates the benefits of self-supervised learning with generative capabilities. We adapt IJEPA Assran et al. (2023) to forecast pre-computed Alpha Earth Foundation Model Brown et al. (2025) embeddings, providing robust SOTA semantic guidance to a frozen Latent Diffusion Model via a custom cross-attention adapter. Details on the dataset are provided in Appendix A and Figure 3.

*This project was funded by the European Union (project No S-MIP-23-45) under the agreement with the Research Council of Lithuania (LMTLT).

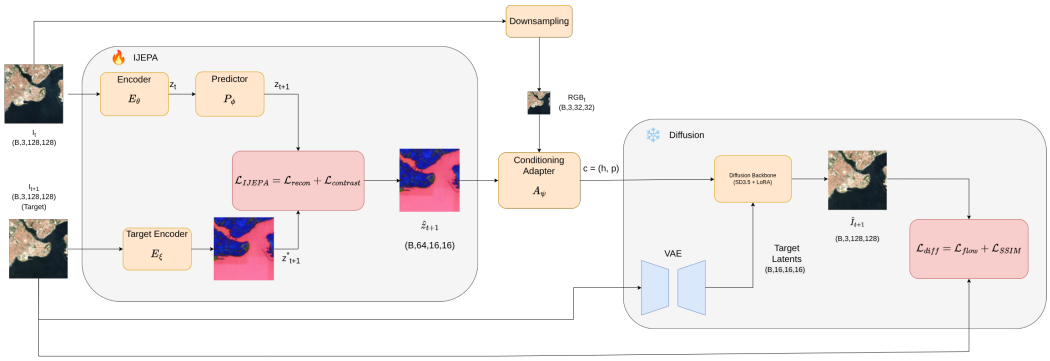

Figure 1: Overview of Sat-JEPA-Diff. The IJEPA module (left) predicts future semantic embeddings $\hat{z}_{t+1}$ from input I_t. These embeddings, combined with coarse spatial structure, condition a frozen SD3.5 backbone via a learned adapter to generate $\hat{I}_{t+1}$.

2 Related Work

Spatiotemporal Forecasting. Early studies used RNNs like ConvLSTM Shi et al. (2015) and PredRNN Wang et al. (2022) for temporal modeling, but these are computationally intensive. CNN-based models like SimVP Gao et al. (2022) improved efficiency via spatial convolutions. However, these deterministic models optimize pixel-wise L1/L2 losses, biasing predictions towards the mean and producing blurry outputs lacking high-frequency detail Mathieu et al. (2015).

Generative Models in Remote Sensing. While GANs Goodfellow et al. (2014) mitigate blurring, they suffer from mode collapse. Denoising Diffusion Probabilistic Models (DDPMs) and Latent Diffusion Models (LDMs) have emerged as promising alternatives to GANs for texture synthesis tasks. However, using LDMs for satellite time-series images is not an easy task, as these models are prone to "hallucinating" features such as road topologies or buildings, etc., without any semantic constraints Saharia et al. (2022).

Self-Supervised Representation Learning. Effective structure-texture bridging requires semantic guidance. While MAE He et al. (2022) and SatMAE Cong et al. (2022) learn representations by reconstructing pixels, they overfit on noise. IJEPA Assran et al. (2023) avoids this by learning in abstract space. Concurrently, geo foundation models such as Alpha Earth Brown et al. (2025), Panopticon Waldmann et al. (2025), and TerraMind Jakubik et al. (2025) provide powerful pre-trained EO embeddings. Sat-JEPA-Diff leverages IJEPA with such embeddings for diffusion conditioning, avoiding spatial collapse of prior hybrid approaches.

3 Methodology

Our framework, illustrated in Figure 1, consists of two synergistic modules: (1) an IJEPA-based temporal predictor that forecasts future semantic embeddings, and (2) a conditioned diffusion generator that synthesizes high-fidelity RGB imagery guided by these predictions. Full architectural specifications and hyperparameters are provided in Appendix B.

Problem Formulation. Given a satellite image $I_t \in \mathbb{R}^{3\times H\times W}$ at time t, our goal is to predict the corresponding image I_{t+1} at time $t+1$. We decompose this into two stages: (i) predicting the semantic representation $\hat{z}_{t+1}$ of the future frame, and (ii) generating the RGB output conditioned on this prediction.

IJEPA Temporal Prediction. We adapt the Joint-Embedding Predictive Architecture (IJEPA) Assran et al. (2023) for temporal forecasting. Unlike masked autoencoders that reconstruct pixels, IJEPA operates entirely in latent space, learning representations robust to sensor noise and atmospheric variations.

Encoder. A Vision Transformer E_θ processes the input image I_t into patch embeddings:

$$z_t = E_\theta(I_t) \in \mathbb{R}^{N \times D}$$

where $N = (H/p)^2$ is the number of patches with patch size p, and D is the embedding dimension.

Predictor. A transformer-based predictor P_ϕ forecasts future embeddings from the encoded representation:

$$\hat{z}_{t+1} = P_\phi(z_t) \in \mathbb{R}^{N \times D}$$

Target Encoder. Following IJEPA, we maintain an exponential moving average (EMA) copy E_ξ of the encoder to produce stable target embeddings $z^*_{t+1} = E_\xi(I_{t+1})$.

IJEPA Loss. We employ a hybrid loss combining reconstruction and contrastive objectives (see Appendix B.8 for component justification):

$$\mathcal{L}_{\text{IJEPA}} = \lambda_1 \|\hat{z}_{t+1} - z^*_{t+1}\|_1 + \lambda_2(1 - \cos(\hat{z}_{t+1}, z^*_{t+1})) + \lambda_3 \mathcal{L}_{\text{spatial}} + \lambda_4 \mathcal{L}_{\text{contrast}}$$

where $\mathcal{L}_{\text{spatial}}$ penalizes variance mismatch between predicted and target embeddings to prevent spatial collapse, and $\mathcal{L}_{\text{contrast}}$ is an InfoNCE loss ensuring global discriminability.

Conditioned Diffusion Generation. We leverage Stable Diffusion 3.5 Esser et al. (2024) as our generative backbone, keeping the core transformer frozen and training only a lightweight conditioning adapter with LoRA Hu et al. (2022).

Conditioning Adapter. The adapter A_ψ transforms IJEPA embeddings into cross-attention conditioning signals:

$$c = (h, p) = A_\psi(\hat{z}_{t+1}, I^c_t)$$

where $h \in \mathbb{R}^{M \times 4096}$ provides token-level conditioning via cross-attention, $p \in \mathbb{R}^{2048}$ provides global conditioning, and I^c_t is a coarse 32×32 downsampled version of I_t that preserves low-frequency spatial structure.

The adapter employs a learned fusion gate α to balance semantic (IJEPA) and structural (coarse RGB) signals:

$$h = \alpha \cdot h_{\text{semantic}} + (1 - \alpha) \cdot h_{\text{coarse}}$$

Flow Matching Objective. Following rectified flow formulation Liu et al. (2022), we train with velocity prediction. Given target latents x_0 from VAE-encoded I_{t+1} and noise $\epsilon \sim \mathcal{N}(0, I)$, we construct:

$$x_\sigma = (1 - \sigma)x_0 + \sigma\epsilon, \quad v^* = \epsilon - x_0$$

The diffusion backbone learns to predict this velocity:

$$\mathcal{L}_{\text{diff}} = \|v_\theta(x_\sigma, \sigma, c) - v^*\|^2 + \lambda_{\text{ssim}} \mathcal{L}_{\text{SSIM}}$$

Total Objective. The complete training loss combines both modules:

$$\mathcal{L} = \mathcal{L}_{\text{IJEPA}} + \lambda \mathcal{L}_{\text{diff}}$$

4 Results

As presented in Table 1, deterministic baselines (PredRNN, SimVP) achieve high PSNR and SSIM scores. However, these metrics are pixel-wise and encourage blurred predictions that "regress towards the mean." In contrast, Sat-JEPA-Diff achieves a significantly higher GSSIM score of 0.8984, which is an increase of over 11% over the best baseline. GSSIM evaluates the preservation of edges and structural gradients. Our results indicate that semantic guidance successfully maintains sharp geospatial features such as roads and urban areas avoiding the blurring typical of deterministic models. Furthermore, swapping the ViT encoder for Panopticon Waldmann et al. (2025) produces similar perceptual results, demonstrating that our approach is not dependent on a single encoder type. Although SSIM drops slightly, this decrease reflects the well-known perception-distortion trade-off Blau & Michaeli (2018), as the model prioritizes realistic texture synthesis over exact pixel averaging.

Table 1: Quantitative comparison on the test set. Arrows indicate whether lower (↓) or higher (↑) values are better. Best results are highlighted in bold.

Model	L1 ↓	MSE ↓	PSNR ↑	SSIM ↑	GSSIM ↑	LPIPS ↓	FID ↓
Deterministic Baselines							
Default	0.0131	0.0008	37.52	0.9361	0.7858	0.0708	0.6959
PredRNN Wang et al. (2022)	0.0117	0.0005	38.38	0.9476	0.7836	0.0726	9.9720
SimVP v2 Gao et al. (2022)	0.0131	0.0006	37.63	0.9391	0.7719	0.0928	18.7208
Generative Models							
SD 3.5 Esser et al. (2024)	0.0175	0.0005	32.98	0.8398	0.8711	0.4528	0.1533
MCVD Voleti et al. (2022)	0.0314	0.0031	31.28	0.8637	0.7665	0.1890	0.1956
Ours (Panopticon) Waldmann et al. (2025)	0.0179	0.0005	32.89	0.8398	0.8750	0.4475	0.1475
Ours	0.0158	0.0004	33.81	0.8672	0.8984	0.4449	0.1475

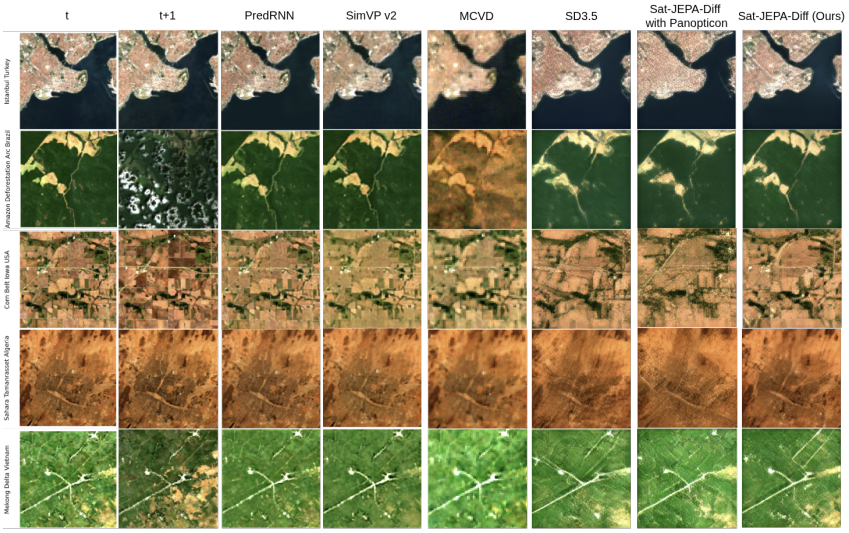

Figure 2: Qualitative comparison of next-frame predictions ($t \to t+1$). While deterministic baselines (PredRNN, SimVP) suffer from spectral blurring, Sat-JEPA-Diff preserves high-frequency details and geospatial boundaries.

Qualitative Analysis. Figure 2 presents a visual comparison of next-frame predictions ($t \to t+1$). Deterministic baselines like PredRNN and SimVP suffer from spectral blurring, masking Istanbul's urban density and smearing Amazonian agricultural borders. Sat-JEPA-Diff overcomes this by generating crisp, realistic textures. Our model preserves intricate street networks and distinct forest edges, validated by our superior GSSIM scores. While diffusion-based generation introduces minor stochastic variations, the semantic grounding provided by the IJEPA module ensures these variations remain geosemantically plausible. We further validate long-horizon temporal consistency through autoregressive rollout experiments in Appendix C.

5 Conclusion

We presented Sat-JEPA-Diff, a novel framework integrating the semantic reasoning of IJEPA with the generative power of Latent Diffusion Models. By shifting forecasting from pixel space to a semantic latent space, we successfully mitigate the blurring artifacts of deterministic baselines. Our results demonstrate superior performance in structural integrity (GSSIM) and perceptual quality (FID). Future work will explore replacing learned embeddings with vision-language scene descriptions as temporally forecastable conditioning signals for generation.

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. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15619–15629, 2023.

Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6228–6237, 2018.

Christopher F Brown, Michal R Kazmierski, Valerie J Pasquarella, William J Rucklidge, Masha Samsikova, Chenhui Zhang, Evan Shelhamer, Estefania Lahera, Olivia Wiles, Simon Ilyushchenko, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. arXiv preprint arXiv:2507.22291, 2025.

Yezhen Cong, Samar Khanna, Chenlin Meng, Patrick Liu, Erik Rozi, Yutong He, Marshall Burke, David Lobell, and Stefano Ermon. Satmae: Pre-training transformers for temporal and multi-spectral satellite imagery. In Advances in Neural Information Processing Systems, volume 35, pp. 197–211. Curran Associates, Inc., 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In Forty-first international conference on machine learning, 2024.

Zhangyang Gao, Cheng Tan, Lirong Wu, and Stan Z Li. Simvp: Simpler is better for video prediction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3950–3959, 2022.

Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. Advances in neural information processing systems, 27, 2014.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 16000–16009, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabé-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation. IEEE/CVF International Conference on Computer Vision (ICCV), 2025.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022.

Michael Mathieu, Camille Couprie, and Yann LeCun. Deep multi-scale video prediction beyond mean square error. arXiv preprint arXiv:1511.05440, 2015.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10684–10695, 2022.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(4):4713–4726, 2022.

Xingjian Shi, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. Advances in neural information processing systems, 28, 2015.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. Advances in neural information processing systems, 35:23371–23385, 2022.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam J. Stewart, Zhitong Xiong, Xiao Xiang Zhu, Stefan Bauer, and John Chuang. Panopticon: Advancing any-sensor foundation models for earth observation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2025.

Yunbo Wang, Haixu Wu, Jianjin Zhang, Zhifeng Gao, Jianmin Wang, Philip S Yu, and Mingsheng Long. Predrnn: A recurrent neural network for spatiotemporal predictive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2):2208–2225, 2022.

A Dataset Details

To train and evaluate the Sat-JEPA-Diff architecture, we curated a large-scale, multi-modal dataset spanning diverse geographical landscapes.

A.1 Data Sources

- Embeddings: Alpha Earth Foundation Model Embeddings. We use pre-computed 64-dimensional feature vectors per pixel, which contain semantic information resistant to noise from the atmosphere.
- Optical Imagery: Sentinel-2 Surface Reflectance (RGB bands), harmonized to 10m GSD. Images are normalized to the range [0,1].

A.2 Spatiotemporal Distribution

The data set covers a period from 2017 to 2024, and there are 100 unique Regions of Interest (RoIs).

Figure 3: Geographical distribution of the 100 selected Regions of Interest (RoIs).

B Implementation Details

Our architectural configurations, hyperparameters, and training infrastructure are summarized below.

B.1 Network Architecture

IJEPA Module. We use a Vision Transformer (ViT) backbone as the encoder E_θ. The encoder processes 128×128 RGB input images with patch size 8, producing $N = 256$ patch tokens. The embedding dimension is $D = 768$ for the base configuration. The predictor P_ϕ is a lightweight transformer with depth configurable via hyperparameter (default: 6 layers). The target encoder E_ξ is an exponential moving average (EMA) copy of E_θ, updated with momentum $\tau \in [0.999, 1.0]$ following a cosine schedule.

Projection Head. A linear projection layer maps the predictor's raw output from 768 dimensions to the target embedding dimension of 64, matching the Alpha Earth Foundation Model embedding space used as supervision signal.

Conditioning Adapter. The adapter A_ψ transforms IJEPA embeddings and coarse RGB signals into SD3.5-compatible conditioning. It consists of:

- IJEPA Token Projection: A three-layer MLP ($768 \to 1024 \to 1024 \to 4096$) with LayerNorm and GELU activations, projecting semantic tokens to the cross-attention dimension.
- Coarse RGB Projection: The 32×32 reference image is patchified into $8 \times 8 = 64$ tokens (patch size 4×4), each with dimension 48 (flattened $4 \times 4 \times 3$ RGB values). A similar MLP ($48 \to 1024 \to 1024 \to 4096$) projects these to cross-attention space.
- Pooled Projection: Mean-pooled IJEPA tokens are projected via a two-layer MLP ($768 \to 1024 \to 2048$) for global conditioning.
- Fusion Gate: A learned sigmoid gate balances IJEPA semantic signals and coarse RGB structural signals: $h = \alpha \cdot h_{\text{semantic}} + (1 - \alpha) \cdot h_{\text{coarse}}$.

Learnable positional embeddings (up to 1024 tokens for IJEPA, 64 tokens for coarse RGB) are added before projection. Total adapter parameters: $\sim$25M.

Stable Diffusion Backbone. We use Stable Diffusion 3.5 Medium Esser et al. (2024) as the generative backbone. The core transformer is kept frozen, with only Low-Rank Adaptation (LoRA) Hu et al. (2022) modules trained on the attention projections. LoRA configuration: rank $r = 8$, alpha $\alpha = 16$. The VAE encoder/decoder operates at $8\times$ spatial compression.

B.2 Training Configuration

Optimization. We use AdamW optimizer with the following schedule:

- Base learning rate: 1×10^{-4}
- Start learning rate: 2×10^{-5} (linear warmup)
- Final learning rate: 1×10^{-6} (cosine decay)
- Warmup epochs: 1
- Weight decay: $0.04 \to 0.4$ (cosine schedule)
- Total epochs: 100

Mixed Precision. Training uses bfloat16 automatic mixed precision for memory efficiency. The VAE is kept in float32 for numerical stability during encoding/decoding operations.

Batch Size. We trained the model with a batch size of 8 per GPU. For multi-GPU training, we use PyTorch DistributedDataParallel with synchronized batch normalization.

EMA Schedule. Target encoder momentum follows: $\tau_t = \tau_{\text{base}} + t \cdot (\tau_{\text{final}} - \tau_{\text{base}})/T$, where $\tau_{\text{base}} = 0.999$, $\tau_{\text{final}} = 1.0$, and T is the total training iterations.

B.3 Loss Function Weights

IJEPA Hybrid Loss. We employ a multi-component loss to prevent spatial collapse:

- L1 reconstruction weight (λ_1): 20.0
- Cosine similarity weight (λ_2): 2.0
- Spatial variance weight (λ_3): 2.0
- Contrastive (InfoNCE) weight (λ_4): 0.5
- Feature regression weight: 5.0
- Contrastive temperature: $\tau = 0.1$

Diffusion Loss. Flow matching MSE loss with SSIM regularization (weight 0.1). Total loss combines IJEPA and diffusion objectives with SD loss weight $\lambda = 1.0$.

Reference Dropout. During training, the coarse RGB reference is dropped with probability $p = 0.15$ to encourage the model to rely on IJEPA semantic predictions alone, improving generalization.

B.4 Masking Strategy

Following IJEPA Assran et al. (2023), we use multi-block masking:

- Encoder mask scale: $[0.7, 1.0]$ (proportion of patches visible)
- Predictor mask scale: $[0.2, 0.5]$ (proportion of patches to predict)
- Aspect ratio range: $[0.75, 1.5]$
- Number of encoder masks: 1
- Number of predictor masks: 1
- Minimum patches to keep: 6
- Overlap allowed: True

B.5 Inference Configuration

Diffusion Sampling. At inference, we use the Flow Matching Euler Discrete scheduler with:

- Default sampling steps: 20
- Noise strength for latent initialization: $\sigma = 0.35$
- Single-step speed estimation to increase efficiency.

Conditioning Fusion. The 32×32 RGB image I_t, with a reduced sampling rate, captures large spatial patterns, while the IJEPA embedded vectors z_{t+1} provide detailed semantic cues. During training, the merging gate stabilizes around $\alpha \approx 0.5$, indicating that the network treats both inputs with equal weight.

B.6 Computational Resources

Hardware. All experiments were performed on a single NVIDIA RTX 5090 24GB GPU.

B.7 Text Conditioning

We use a fixed prompt for all Sentinel-2 imagery:

> "High-resolution Sentinel-2 satellite image, multispectral earth observation, natural colors RGB composite, 10m ground resolution, clear atmospheric conditions, detailed land surface features"

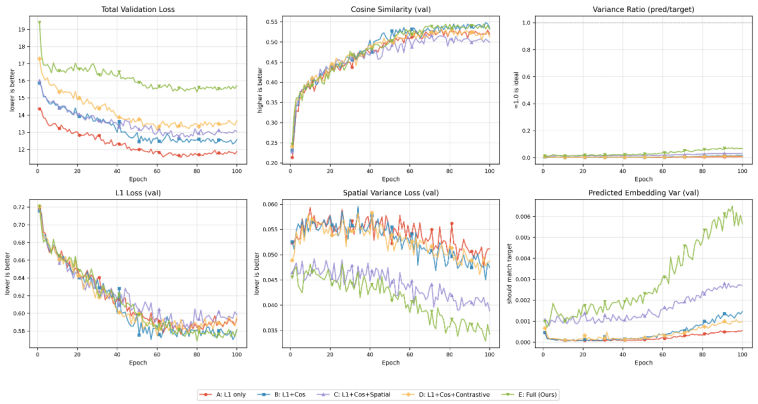

Figure 4: Systematic IJEPA Loss Ablation. Validation metrics over 100 epochs demonstrate that our full objective function (E curve) uniquely avoids representation collapse and maintains high embedding variance compared to near-zero variance in the reduced underlying models (A-D curves). Despite higher total loss, the full model maintains high cosine similarity and achieves superior spatial variance.

Figure 5: Long-horizon autoregressive rollout comparison (2018 → 2024) on the Rio de Janeiro Coast. Top Row: Ground Truth. Rows 2-3: Deterministic baselines rapidly degrade into spectral blurring (spatial collapse) after 2-3 steps. Bottom Row: Sat-JEPA-Diff maintains high contrast and structural sharpness throughout the 7-year horizon.

B.8 Ablation on IJEPA Loss Components (Training Dynamics)

To address the necessity of a multi-component loss formulation under our computational constraints, we performed an early trajectory ablation study (20 epochs) focusing on the IJEPA encoder and predictor operating on SOTA base model embedded vectors.

As shown in Figure 4, while the constrained 20-epoch window naturally allows the curves to converge to a similar limited region, early learning dynamics reveal critical differences. Removing certain adjustment terms (e.g., spatial variance or contrast loss) significantly delays the initial alignment with the target embedded vectors and destabilizes the early learning phase. This early trajectory analysis supports our reliance on full hybrid formulation to provide rapid and robust semantic orientation. A comprehensive full-scale ablation of 100 epochs with a diffusion backbone remains a key direction for future extended studies.

B.9 Reproducibility

Random Seeds. To enable CUDA's determinism, we lock the NumPy and PyTorch seed values to 0 and set torch.backends.cudnn.benchmark = True.

Data Split. The dataset is randomly shuffled before being divided into 80% training and 20% validation subsets.

Code Availability. We built this pipeline on PyTorch 2.0+, using PEFT for Diffusers for SD3.5 and LoRA integration. The YAML configuration files track all experimental settings.

C Autoregressive Stability Analysis

We assessed long-term stability using an autoregressive rolling method. Starting with the actual image from 2018, we generated predictions up to 2024 by feeding previous outputs back into the network. Figure 5 shows how deterministic methods (PredRNN, SimVP) rapidly blur high-frequency details to minimize MSE. Sat-JEPA-Diff successfully avoids this spatial collapse, preserving clear textures and sharp boundaries across all time steps.