MARVEL: MASKED AUTOREGRESSIVE VISUAL UNDERSTANDING AND GENERATION FOR SPECTRAL REMOTE SENSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Spectral remote sensing observations support both scene understanding and the generation of unobserved spectral responses. Bringing understanding and generation into unified pretraining allows generative supervision to support transferable representation learning. However, more accurate generation does not necessarily yield better features for understanding. To address this challenge, we introduce **MARVEL**, a masked autoregressive pretraining framework for spectral understanding and generation. A bidirectional encoder integrates all visible observations, and full-context conditional alignment converts the resulting features into conditions for each target wavelength and spatial position. Spectral content autoregression then uses these conditions to generate responses in wavelength order. Each response is supervised against its own target and re-encoded to guide subsequent predictions. Through this reuse, feedback from subsequent predictions also contributes to shared representation learning. MARVEL achieves competitive performance on 14 downstream tasks. Without additional generation training, it also achieves strong results on spectral generation tasks using Sentinel-2 and Landsat-8 data. Controlled comparisons support the benefits of decoder organization, prediction reuse, and cross-step gradient feedback.

1 INTRODUCTION

Spectral remote sensing captures the spatial distribution of objects and their responses across wavelengths. These observations provide information for object recognition and scene understanding. However, sensors do not always capture all the spectral information needed for analysis. The observed bands provide both evidence for scene understanding and conditions for generating unobserved spectral responses. Bringing understanding and generation into unified pretraining allows generative supervision to support transferable representation learning.

Existing research has advanced both capabilities. Flexible spectral encoders support diverse band configurations and sensor inputs (Xiong et al., 2024; Baumann et al., 2026; Waldmann et al., 2025). Spectral generation models predict unobserved responses from available bands (Cai et al., 2022; zuopeng et al., 2026), and broader multimodal frameworks connect generation with Earth observation analysis (Jakubik et al., 2025). Masked image modeling provides a basis for learning these capabilities together (He et al., 2022; Hong et al., 2024). Predicting masked content trains the model to generate missing responses and provides supervision for learning representations of visible observations. The question is how to organize this learning process to benefit both understanding and generation.

A key difficulty is that improvements in generation quality do not necessarily lead to better representations. Generation depends on the encoder and decoder jointly, but understanding tasks use the encoder's features. More accurate predictions may therefore reflect better decoding without a corresponding gain in feature transferability. Our controlled comparisons illustrate this distinction (Figure 1). Replacing full-attention decoding with autoregressive-only decoding improves generation but reduces frozen-encoder transfer performance. This difference motivates us to examine how generation interacts with the learning of observation representations. Targets at different wavelengths and spatial positions require different conditions from the available observations. We jointly process observation features and target information to establish these conditions before sequential generation.

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.

Reusing supervised predictions then allows later generation to draw on earlier responses and provide further feedback for shared representation learning.

To realize this approach, we propose **MARVEL**, a masked autoregressive pretraining framework for spectral understanding and generation (Figure 2). A bidirectional encoder integrates all visible observations into shared features. To use these features for generation, we design a decoder that combines full-context conditional alignment with spectral content autoregression. Full-context conditional alignment jointly processes the encoded features and target queries to establish conditions for each requested wavelength and spatial position. Spectral content autoregression then uses these conditions to generate responses in wavelength order. We keep all visible observations accessible throughout generation, applying the autoregressive order only to dependencies among target contents. To condition subsequent generation on supervised spectral content, we re-encode each completed prediction as history. Each response receives supervision from its own reconstruction target, giving this history an explicit spectral and spatial meaning. We keep this reuse differentiable, allowing later prediction errors to propagate through earlier outputs to the encoder.

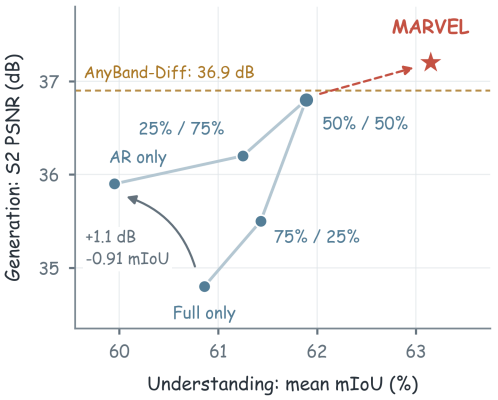

Figure 1: Understanding and generation across Full/AR allocations at a fixed depth of four decoder blocks. ★ denotes our complete method.

This generation process also needs to account for unequal spectral intervals and unreliable history early in training. We describe each band by its center wavelength and bandwidth. Since wavelength order alone does not capture spectral distances, we introduce λ-RoPE to encode relative wavelength differences in attention, building on rotary positional embeddings (Su et al., 2024). Early prediction errors can also affect subsequent generation. To address this issue, we design Spectral History Forcing (SHF), inspired by scheduled sampling (Bengio et al., 2015). SHF blends ground-truth and predicted history representations and gradually reduces ground-truth assistance during training. This schedule prepares the model to use only predicted history during inference. The predicted branch remains differentiable throughout training. For understanding, we evaluate the frozen encoder with trained task heads. For generation, we use the complete model without additional training. Our contributions are as follows: ❶ We propose MARVEL, a masked autoregressive pretraining framework that learns transferable spectral representations and conditional generation within a shared model. ❷ We design a decoder that establishes target-specific conditions from all visible observations and reuses supervised predictions as differentiable history. Each prediction is trained both to reconstruct its own target and to support subsequent generation. ❸ We demonstrate competitive performance on 14 downstream tasks and strong spectral generation without additional generation training. Controlled comparisons support the contributions of decoder organization, history content, and cross-step gradient feedback.

2 Related Work

❶ **Spectral representation learning.** SatMAE and SpectralGPT adapt masked reconstruction to spectral imagery (Cong et al., 2022; Hong et al., 2024). Wavelength conditioning, sensor metadata, and flexible tokenization support heterogeneous bands (Xiong et al., 2024; Baumann et al., 2026; Prexl & Schmitt, 2024; Li et al., 2026; Si et al., 2026), while Panopticon and Copernicus-FM extend representation learning across sensors (Waldmann et al., 2025; Wang et al., 2025). We examine how spectral recovery dependencies shape transferable encoder features. ❷ **Masked and autoregressive visual modeling.** Visual pretraining learns representations through masked reconstruction, autoregressive prediction, or their combination (He et al., 2022; El-Nouby et al., 2024; Liu & Yi, 2025). MAR extends masked autoregressive generation to continuous tokens using a diffusion loss (Li et al., 2024). HSI-VAR applies next-scale prediction to hyperspectral restoration (Wang et al., 2026). Our decoder predicts spectral responses in wavelength order, retaining all observations and reusing completed predictions as history. ❸ **Spectral generation and unified learning.** Regression and diffusion models address spectral reconstruction and restoration (Cai et al., 2022; Pang et al., 2024; Deng et al., 2025),

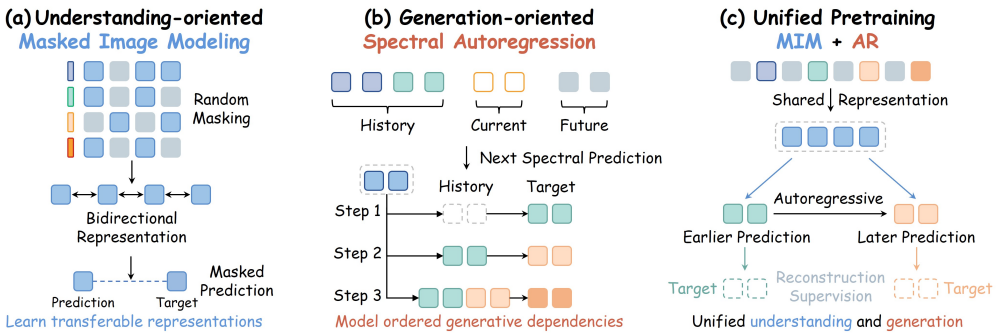

Figure 2: Learning from partial spectral observations. Understanding extracts transferable features from available measurements; conditional generation recovers responses at specified wavelengths. MARVEL connects both uses through masked pretraining with target-specific alignment and feedback from predictions generated in wavelength order.

while AnyBand-Diff supports arbitrary-band recovery (zuopeng et al., 2026). MAGE unifies representation learning and image synthesis through masked token modeling (Li et al., 2023), while TerraMind develops generative multimodal representations for Earth observation (Jakubik et al., 2025). MARVEL couples generation with representation learning through target-specific alignment and differentiable prediction reuse.

3 METHOD

MARVEL learns shared representations for understanding and conditional generation through masked spectral recovery. As shown in Figure 3, a bidirectional encoder aggregates visible observations, and a full-to-autoregressive decoder establishes conditions for each target before recovering missing content in wavelength order. Each prediction receives its own reconstruction supervision and is re-encoded to condition subsequent targets, incorporating cross-band dependencies into pretraining.

3.1 MASKED BIDIRECTIONAL SPECTRAL ENCODING

Let $\mathbf{X}_c \in \mathbb{R}^{H\times W}$ denote band c and $\mathbf{x}_{c,u}$ its patch at spatial position u. Random masking partitions spatial–spectral patch indices into visible and target sets, $\mathcal{V}$ and $\mathcal{M}$, respectively. Only visible patches enter the encoder. A band may contain both visible and masked patches, while whole-band generation masks all patches of each requested band. Both settings use the same target-query interface.

To identify bands across sensors, we describe band c by its center wavelength and bandwidth, $\mathbf{d}_c = (\lambda_c, \Delta\lambda_c)$. We construct its band embedding as $\mathbf{e}_c^{\text{band}} = f_{\text{band}}([\phi(\bar{\lambda}_c); \phi(\overline{\Delta\lambda}_c)])$, where the overbars denote normalized descriptors, ϕ is a multifrequency Fourier mapping, and f_{band} is an MLP. We combine this embedding with a patch projection P, shared across bands, and a spatial encoding, then encode the visible tokens bidirectionally:

$$\begin{aligned} \mathbf{t}_{c,u} &= P(\mathbf{x}_{c,u}) + \alpha\mathbf{e}_c^{\text{band}} + \mathbf{e}_u^{\text{sp}}, \\ \mathbf{H}_{\text{enc}} &= E_{\text{bi}}([\mathbf{t}_{\text{cls}}; \{\mathbf{t}_{c,u}\}_{(c,u)\in\mathcal{V}}]), \qquad \mathbf{M}_{\text{obs}} = \Pi(\mathbf{H}_{\text{enc},\mathcal{V}}). \end{aligned} \tag{1}$$

Here, α is a learnable gain, $\mathbf{e}_u^{\text{sp}}$ encodes spatial position, and $\mathbf{t}_{\text{cls}}$ is a global token. The projection Π maps visible-position encoder features into decoder space to form observation memory $\mathbf{M}_{\text{obs}}$. Bidirectional attention allows each visible token to integrate information across observed bands and spatial positions. All visible observations remain accessible throughout generation, including those beyond the current target wavelength. Wavelength causality applies only among generated targets.

3.2 FULL-TO-AUTOREGRESSIVE SPECTRAL DECODING

Conditional spectral generation requires both target-specific access to observations and dependencies among generated responses. Full-context conditional alignment establishes observational conditions

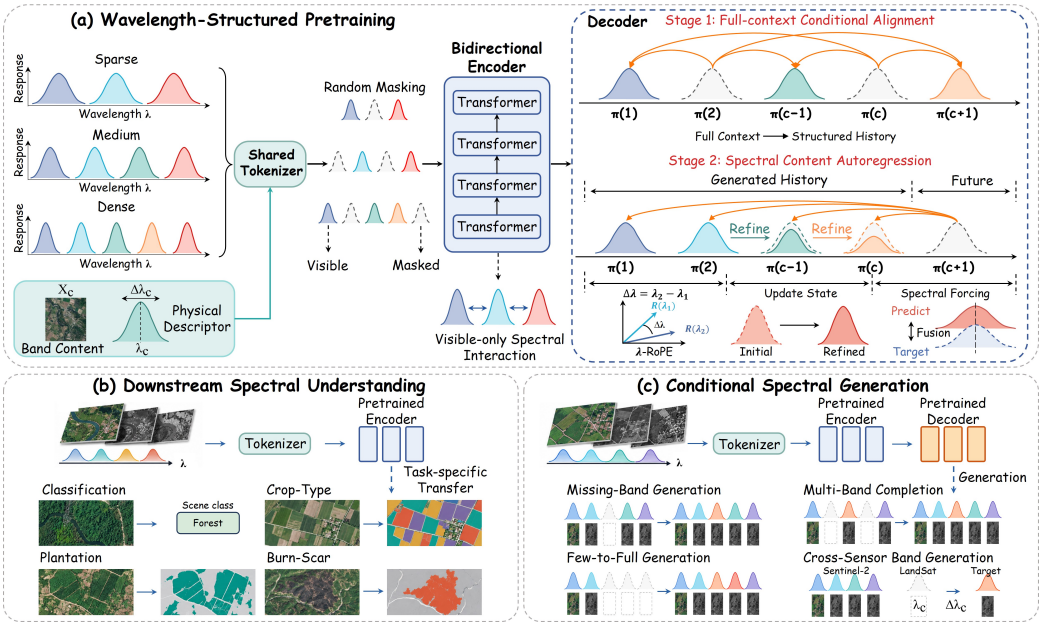

Figure 3: Overview of MARVEL. (a) A bidirectional encoder aggregates all visible patches; full-context conditional alignment establishes target-specific states. Target bands are generated in wavelength order, with supervised predictions re-encoded as differentiable history. Later reconstruction losses train the shared encoder through earlier predictions. (b) Understanding uses the frozen encoder with trained heads. (c) Conditional generation uses the complete model without further training.

for all targets. Spectral content autoregression then combines these conditions with completed predictions to generate target content in wavelength order.

Full-context conditional alignment. Encoder features represent visible positions, whereas generation requests specify target wavelengths and spatial locations. For each target position $(c, u) \in \mathcal{M}$, we construct a query $\mathbf{q}_{c,u}$ from a learned mask embedding $\mathbf{m}$, its physical descriptor $\mathbf{d}_c$, and spatial encoding $\mathbf{e}^{\mathrm{sp}}_u$. Collecting these queries as $\mathbf{Q}_{\mathcal{M}} = [\mathbf{q}_{c,u}]_{(c,u)\in\mathcal{M}}$, we jointly process them with observation memory through a full-attention decoder:

$$\begin{aligned} \mathbf{q}_{c,u} &= f_{\mathrm{query}}(\mathbf{m}, \mathbf{d}_c, \mathbf{e}^{\mathrm{sp}}_u), \\ [\mathbf{M}^{(0)}_{\mathrm{obs}}; \mathbf{Z}^{(0)}_{\mathcal{M}}] &= D_{\mathrm{full}}([\mathbf{M}_{\mathrm{obs}}; \mathbf{Q}_{\mathcal{M}}]). \end{aligned} \tag{2}$$

The outputs are updated observation memory $\mathbf{M}^{(0)}_{\mathrm{obs}}$ and target-specific latent conditions $\mathbf{Z}^{(0)}_{\mathcal{M}}$. Joint attention lets every target gather evidence from all visible measurements and interact with other target queries. These queries contain known descriptors and spatial coordinates, without target pixel values. Generated content enters the conditioning through the subsequent autoregressive history. Alignment is learned through the shared pretraining objective, without an additional alignment loss.

Spectral content autoregression. Starting from the aligned states, we generate target responses sequentially and re-encode completed predictions as history. Let π order the K target bands by increasing center wavelength. For band $\pi(k)$, $\mathcal{U}_k = \{u : (\pi(k), u) \in \mathcal{M}\}$ contains its target positions, and $\mathbf{Y}_k = \{\mathbf{x}_{\pi(k),u} : u \in \mathcal{U}_k\}$ denotes their content, with aligned states $\mathbf{Z}^{(0)}_k$. Thus, $\mathbf{Y}_k$ contains the masked patches of that band during pretraining and all its patches during whole-band generation. At step k, we use aligned states, observation memory, and preceding predictions:

$$\begin{aligned} \mathbf{S}_{<k} &= \mathrm{Concat}_{j<k}\, \rho_\theta(\widehat{\mathbf{Y}}_j, \mathbf{d}_{\pi(j)}, \mathcal{U}_j), \\ \widehat{\mathbf{Y}}_k &= F_{\mathrm{AR},\theta}(\mathbf{Z}^{(0)}_k, \mathbf{M}^{(0)}_{\mathrm{obs}}, \mathbf{S}_{<k}), \\ \rho_\theta(\mathbf{x}_{c,u}, \mathbf{d}_c, u) &= P_{\mathrm{hist}}(\mathbf{x}_{c,u}) + \mathbf{e}^{\mathrm{dec}}_{c,u}. \end{aligned} \tag{3}$$

Here, the history encoder ρ_θ operates patchwise: P_{hist} projects patch values into decoder space, and $\mathbf{e}^{\mathrm{dec}}_{c,u}$ combines their band and spatial embeddings. History therefore consists of response estimates

that have passed through the reconstruction head and receive supervision at their own target positions. Each prediction serves both as a supervised output and as a condition for later targets. This reuse remains differentiable, allowing loss gradients to propagate from later predictions through earlier outputs to the encoder. Autoregression requires K sequential steps across target bands, while target patches within each band are predicted in parallel.

Each AR layer accesses observations and history through two cross-attention branches. Let $\mathcal{A}_\lambda^{(\ell)}(\mathbf{A}, \mathbf{B})$ denote wavelength-conditioned attention with queries from $\mathbf{A}$ and keys and values from $\mathbf{B}$. For $\ell = 0, \ldots, L-1$, the layer update and output prediction are

$$\begin{aligned} \mathbf{U}_k^{(\ell)} &= \mathbf{Z}_k^{(\ell)} + \mathcal{A}_\lambda^{(\ell)}(\mathbf{Z}_k^{(\ell)}, \mathbf{M}_{\text{obs}}^{(0)}) + \mathcal{A}_\lambda^{(\ell)}(\mathbf{Z}_k^{(\ell)}, \mathbf{S}_{<k}), \\ \mathbf{Z}_k^{(\ell+1)} &= \mathbf{U}_k^{(\ell)} + \text{MLP}^{(\ell)}(\mathbf{U}_k^{(\ell)}), \qquad \widehat{\mathbf{Y}}_k = f_{\text{out}}(\mathbf{Z}_k^{(L)}). \end{aligned} \tag{4}$$

The observation branch accesses all visible measurements, while the history branch accesses only preceding predictions. At the first step, $\mathbf{S}_{<1} = \varnothing$ and the history-attention contribution is zero. After L layers, the output head produces the current response, which is re-encoded and appended to history for the next target. AR layers and the output head are shared across steps; observation memory and initial target states are computed once before autoregression. The AR blocks use no separate within-band spatial self-attention. Spatial interactions arise through joint full-context alignment and attention to spatially distributed observation and history tokens.

The conditional attention must also account for spectral distance: consecutive bands in the generation order can have different wavelength intervals. We therefore apply λ-RoPE in both attention branches, rotating projected queries $\mathbf{q}_i$ and keys $\mathbf{k}_j$ according to their normalized center wavelengths. Following rotary encoding (Su et al., 2024),

$$[R(\bar{\lambda}_i)\mathbf{q}_i]^\mathsf{T}[R(\bar{\lambda}_j)\mathbf{k}_j] = \mathbf{q}_i^\mathsf{T} R(\bar{\lambda}_j - \bar{\lambda}_i)\mathbf{k}_j. \tag{5}$$

The rotation R thus introduces relative wavelength differences into the attention score. Band embeddings identify individual responses through center wavelength and bandwidth, while λ-RoPE represents spectral distances between targets and their conditions.

3.3 History scheduling and pretraining objective

Spectral History Forcing. Early prediction errors can make history unreliable, while using only ground-truth history would introduce a mismatch with inference. Inspired by scheduled sampling (Bengio et al., 2015), Spectral History Forcing (SHF) blends ground-truth and predicted content in the history encoding space, gradually removing the ground-truth contribution. For a preceding target position $(\pi(j), u)$, we use the same history encoder to obtain $\rho_{j,u}^{\text{gt}} = \rho_\theta(\mathbf{x}_{\pi(j),u}, \mathbf{d}_{\pi(j)}, u)$ and $\widehat{\rho}_{j,u} = \rho_\theta(\widehat{\mathbf{x}}_{\pi(j),u}, \mathbf{d}_{\pi(j)}, u)$. At epoch e, the history used by target k is

$$\begin{aligned} \mathbf{m}_{j,u}(e) &= r(e)\,\text{stopgrad}[\rho_{j,u}^{\text{gt}} + \boldsymbol{\epsilon}_{j,u}] + [1 - r(e)]\widehat{\rho}_{j,u}, \\ \mathbf{S}_{<k}(e) &= \text{Concat}_{j<k}[\mathbf{m}_{j,u}(e)]_{u \in \mathcal{U}_j}. \end{aligned} \tag{6}$$

Here, $r(e) = 0.3\max(1 - e/50, 0)$ controls the ground-truth contribution, and $\boldsymbol{\epsilon}_{j,u}$ is small feature noise applied to the ground-truth branch. This branch is detached, while the predicted branch remains differentiable. During pretraining, $\mathbf{S}_{<k}(e)$ replaces $\mathbf{S}_{<k}$ in Equation 3. Predicted content contributes from the start of training and supplies all history from epoch 50 onward.

Ground-truth assistance is restricted to preceding targets, $j < k$. It does not enter observation memory or full-context alignment, and the ground-truth values of current and future targets are never supplied as conditioning inputs. Predictions are still generated sequentially in wavelength order.

Pretraining objective. Using the scheduled history, we jointly optimize the encoder and decoder through pointwise reconstruction and spectral curvature matching. The reconstruction loss supervises predictions at masked positions:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_k} \left\| \widehat{\mathbf{x}}_{\pi(k),u} - \mathbf{x}_{\pi(k),u} \right\|_2^2. \tag{7}$$

To additionally supervise spectral shape, we match wavelength-aware second-order differences to those of the target. For sample b, let λ_b denote its ordered band wavelengths. At spatial patch u,

$\mathbf{R}_{b,u}$ combines predictions at masked positions with measured values at visible positions, and $\mathbf{Y}_{b,u}$ contains the corresponding ground truth. With $\mathcal{A}_b$ denoting spatial patches masked in at least one band, the curvature matching loss is

$$\mathcal{L}_{\text{curv}} = \frac{1}{B}\sum_{b=1}^{B} \frac{\sum_{u\in\mathcal{A}_b} \left\| \mathcal{D}^{(2)}_{\lambda_b}\mathbf{R}_{b,u} - \mathcal{D}^{(2)}_{\lambda_b}\mathbf{Y}_{b,u} \right\|_1}{\max(|\mathcal{A}_b|, 1)}. \tag{8}$$

Here, B is the batch size, and $\mathcal{D}^{(2)}_{\lambda_b}$ compares neighboring spectral slopes while accounting for unequal wavelength spacing. The entrywise ℓ_1 norm sums over band triplets and pixels within each patch. Matching target curvature penalizes spectral shape errors without forcing curvature toward zero. The overall objective is $\mathcal{L} = \mathcal{L}_{\text{rec}} + \beta\mathcal{L}_{\text{curv}}$, with $\beta = 10^{-6}$. The reconstruction term directly supervises each response, while curvature matching constrains spectral shape across bands. Both losses train the shared encoder through aligned target states and observation memory. Differentiable prediction reuse provides an additional path: gradients reaching later predictions can also propagate through history encoding and earlier outputs to the encoder features that produced them. Each prediction is therefore optimized both to fit its own target and to support subsequent generation. These coupled paths incorporate dependencies among generated responses into shared representation learning, without downstream understanding labels. After pretraining, understanding tasks use the frozen shared encoder with trained task heads. Conditional generation uses the complete model without further training, with $r = 0$ and no history noise. For a single target band ($K = 1$), generation uses its aligned states and observation memory. For multiple target bands ($K > 1$), completed predictions additionally provide history in wavelength order.

Table 1: GEO-Bench segmentation mIoU (%). Avg. is the arithmetic mean over six tasks. Bold and underlined values mark the best and second-best scores per column. Blue shading marks MARVEL.

Method	PV4Ger	NZ-Cattle	NeonTree	Cashew	SA-Crop	Chesapeake	Avg.
MAE	86.0	75.6	52.1	52.4	29.7	50.4	57.70
SpectralGPT	90.6	72.0	53.2	54.2	31.1	53.6	59.12
DOFA	91.4	78.6	53.5	57.2	31.7	55.1	61.25
AOM	91.7	78.9	53.6	58.0	32.1	55.4	61.62
Panopticon	92.0	79.1	53.6	58.0	33.2	56.1	62.00
Copernicus-FM	91.6	<u>79.8</u>	53.8	58.6	33.4	56.0	62.20
TerraMind	<u>92.3</u>	**80.0**	**54.6**	<u>59.4</u>	**34.0**	<u>57.1</u>	<u>62.90</u>
MARVEL	**92.6**	79.5	<u>54.5</u>	**59.7**	<u>33.9</u>	**58.7**	**63.15**

4 EXPERIMENTS

4.1 EXPERIMENTAL PROTOCOL

We pretrain on Sentinel-2 (S2), Landsat-8 (L8), and RGB imagery for 100 epochs at 128×128 resolution, with two Full and two AR decoder blocks. Internal controls match encoder architecture, data, masking, training duration, and decoder depth; equal depth does not imply equal computation. Understanding evaluation covers six GEO-Bench segmentation tasks using mean mIoU (Lacoste et al., 2023), plus eight classification and dense-prediction benchmarks, including three optical Copernicus-Bench tasks (Marsocci et al., 2024; Wang et al., 2025). Generation uses held-out scenes with ten S2 and six L8 bands. Missing withholds one band, Multi three fixed bands, and Few generates non-RGB bands from RGB. All methods receive identical observations and target requests, with all target pixels excluded from inputs and PSNR computed only on targets. Target ground truth is used only for metric computation, except in the privileged history oracle. The appendix provides details on the data sources and evaluation protocols.

4.2 UNDERSTANDING TRANSFER AND OBSERVATION ROBUSTNESS

MARVEL achieves the highest six-task mean mIoU of 63.15, exceeding AOM by 1.53 points and TerraMind by 0.25 points (Table 1). It leads on three tasks and outperforms AOM on all six. Across

eight additional benchmarks, MARVEL ranks first on six, ties for second on MADOS, and ranks second on LC100Cls-S3 (Table 2). These gains span classification and dense prediction with a frozen encoder, showing that the pretrained features support multiple task types without backbone adaptation. We also evaluate transfer with fewer labels and observed bands. With fewer labels, MARVEL exceeds ChannelMAE by 1.5–3.6 mIoU points (Figure 5(a)). With the fewest input bands, it exceeds AOM by 2.7 mean mIoU points on Cashew and SA-Crop (Figure 5(b)).

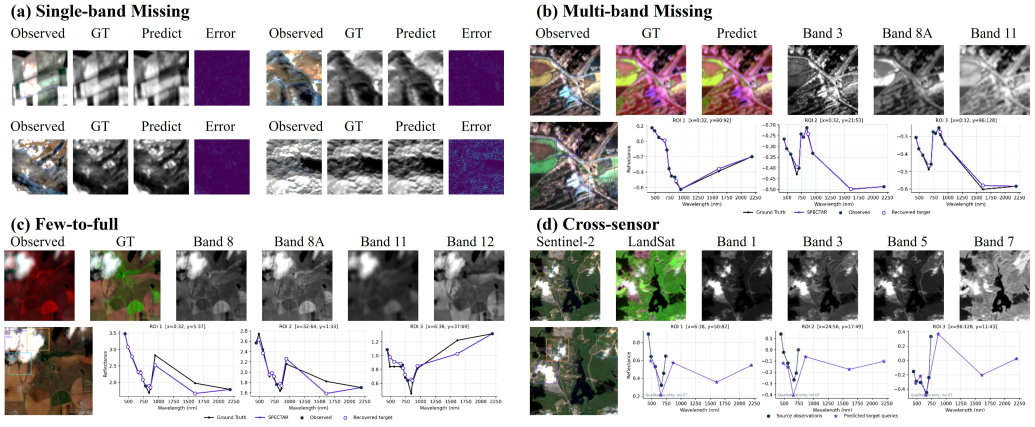

Figure 4: Qualitative recovery: (a) Missing, (b) Multi, (c) Few, and (d) cross-sensor queries. Ground truth and errors are shown where available. The unpaired cross-sensor examples in (d) are not included in the quantitative evaluation.

Table 2: Transfer to classification and dense-prediction tasks. BEN reports mAP; EuroSAT, So2Sat, and LC100Cls-S3 report accuracy; the remaining tasks report mIoU. All values are percentages.

	Additional transfer tasks					Copernicus-Bench subset		
Method	**BEN mAP ↑**	**EuroSAT Acc. ↑**	**So2Sat Acc. ↑**	**HLS Burn mIoU ↑**	**MADOS mIoU ↑**	**Cloud-S2 mIoU ↑**	**DFC2020-S2 mIoU ↑**	**LC100Cls-S3 Acc. ↑**
MAE	76.4	94.6	63.6	74.8	49.6	73.8	51.6	68.4
SpectralGPT	79.0	95.2	65.1	76.1	51.3	76.4	54.7	70.8
DOFA	80.2	96.1	67.1	78.5	53.7	79.2	58.3	74.5
AOM	81.0	96.4	68.0	79.1	54.5	80.5	60.1	75.6
Panopticon	82.4	96.7	68.5	<u>80.7</u>	<u>55.9</u>	82.1	<u>63.2</u>	77.2
Copernicus-FM	82.1	95.8	67.7	79.4	54.8	<u>82.8</u>	62.6	**80.1**
TerraMind	<u>82.6</u>	<u>96.9</u>	<u>68.9</u>	80.6	**56.2**	81.6	61.9	78.0
MARVEL	**82.8**	**97.3**	**69.3**	**81.3**	<u>55.9</u>	**83.5**	**64.1**	<u>79.5</u>

Table 3: Conditional generation PSNR (dB) on withheld bands. Missing: one target; Multi: three fixed targets; Few: RGB-to-remaining-band generation. Dashes denote unevaluated settings.

	Sentinel-2			Landsat		
Method	**Missing**	**Multi**	**Few**	**Missing**	**Multi**	**Few**
Spectral interpolation	–	29.8	–	–	29.4	–
SpectralMAE	35.5	33.0	30.6	34.2	32.3	29.8
Full-reconstruction MAE	33.9	32.1	29.9	32.8	31.6	29.2
Autoregressive only	35.9	33.5	31.0	34.6	32.8	30.2
MST++	–	–	31.8	–	–	31.1
AnyBand-Diff	<u>36.9</u>	**35.1**	**32.8**	**35.9**	<u>33.7</u>	**32.0**
MARVEL	**37.2**	<u>34.9</u>	<u>32.6</u>	<u>35.8</u>	**34.1**	<u>31.8</u>

4.3 CONDITIONAL SPECTRAL GENERATION

MARVEL achieves competitive generation performance without additional training (Table 3). It outperforms SpectralMAE in all six settings, remains within 0.4 dB of AnyBand-Diff, and exceeds MST++ by 0.8/0.7 dB on S2/L8 Few. Gains over AR-only decoding reach 1.4/1.3 dB on S2/L8 Multi and persist with fewer observed bands (Figure 5(c)). Missing uses no generated history at inference. Multi and Few additionally test the use of generated history, with each completed response re-encoded to provide conditions for the targets that follow. Figures 4 and 6 illustrate generation across settings and scenes. A separate paired cross-sensor evaluation yields 33.6/33.2 dB for S2-to-L8/L8-to-S2 generation (Table A7); both sensors are represented in pretraining.

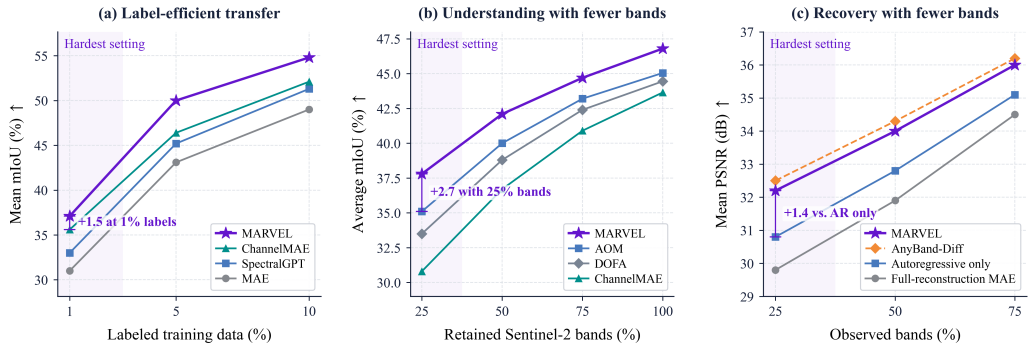

Figure 5: Robustness to fewer labels and observed bands. (a) Six-task GEO-Bench mean mIoU. (b) Mean mIoU on Cashew and SA-Crop. (c) Recovery PSNR averaged over S2 and L8. Shading marks the sparsest setting; band fractions in (c) differ from the fixed RGB inputs in Few.

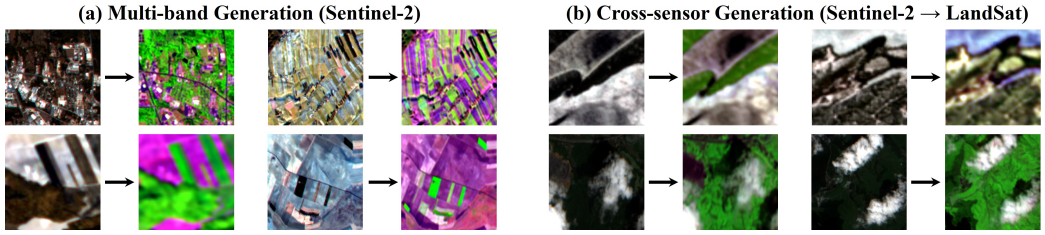

Figure 6: Qualitative examples of conditional spectral generation across diverse scenes. Each pair shows the observed input (left) and the generated output (right).

4.4 STRUCTURAL ANALYSIS

❶ **History content.** Predicted-content history outperforms query-state, shuffled-history, and observation-only conditioning in multi-target generation (Figure 7). The shuffled control preserves the history token count and band and spatial encodings, but breaks the correspondence between predicted content and spatial position. Its weaker results support the importance of spatially aligned history. Per-target gains remain positive across later steps in the six-target sequence, showing that the model can use its own imperfect predictions to support subsequent generation. ❷ **History reuse and gradients.** Both detached and differentiable history improve understanding and generation over observation-only conditioning (Table 4(b)). The complete history configuration gains 1.25 points in frozen-encoder mean mIoU and 1.2 dB in S2 Multi PSNR over observation-only conditioning. For fixed weights, inputs, and random draws, detaching earlier predicted patch values leaves forward values unchanged. This control blocks gradients from later losses through earlier predictions while preserving their own reconstruction supervision. With history construction and SHF unchanged, retaining cross-step gradients adds 0.31 points and 0.4 dB over detached history, supporting a contribution beyond forward history reuse. ❸ **Decoder and training components.** At fixed decoder depth, two-Full/two-AR decoding improves understanding and generation over Full-only and AR-only decoding (Figure 1). Component ablations further support λ-RoPE, bandwidth information,

curvature matching, and SHF (Table 4(a)). Without curvature matching, mean mIoU drops by 0.98 points, while S2 Missing PSNR decreases by 0.1 dB. This illustrates why the pretraining design is evaluated through both generation and frozen-encoder transfer. ❹ **Observation visibility.** Removing observations above the target wavelength at inference reduces Missing PSNR on both S2 and L8, with model weights fixed (Table 4(c)). This supports retaining all observations while applying causal ordering only to dependencies among generated targets.

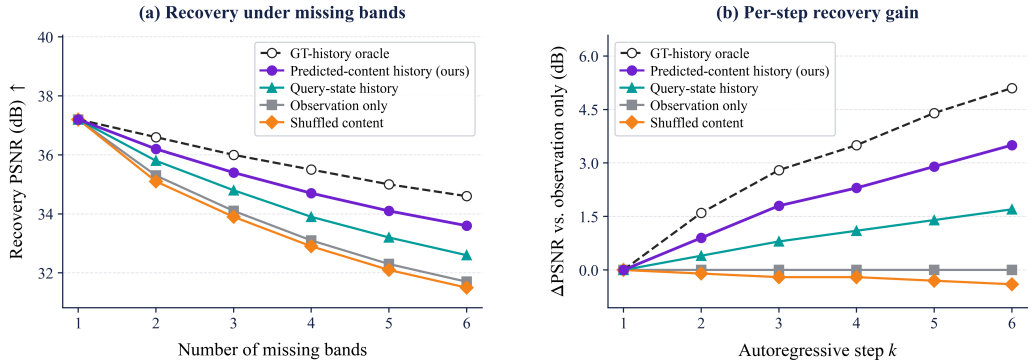

Figure 7: History-content controls, each trained separately. (a) PSNR versus missing-band count. (b) Per-target PSNR gains over observation-only conditioning in the six-target setting, not cumulative gains. The GT-history oracle accesses preceding ground-truth targets.

Table 4: Core ablations. GB: six-task mean mIoU (%); generation: PSNR (dB). Panels (a,c) use Missing; (b) uses Multi. History gradients pass through earlier predictions; – means not applicable. Panel (c) varies observations at inference with fixed pretrained weights. Bold/underlined scores indicate best/second-best results per panel; blue shading marks the full/default configuration.

(a) Spectral and training components

Configuration	GB mIoU ↑	S2 Missing ↑
Without λ-RoPE	61.74	36.6
Prediction-only history	61.71	36.5
Fixed teacher forcing	61.44	36.8
Center wavelength only	61.76	36.7
Without curvature regularizer	<u>62.17</u>	<u>37.1</u>
Full model	**63.15**	**37.2**

(b) Predicted-history reuse

Predicted history	History gradient	GB mIoU ↑	S2 Multi ↑
×	–	61.90	33.7
✓	×	<u>62.84</u>	<u>34.5</u>
✓	✓	**63.15**	**34.9**

(c) Observation visibility

Observations	S2 Missing	L8 Missing
All available	**37.2**	**35.8**
Wavelength-truncated	<u>35.8</u>	<u>34.4</u>

5 Conclusion

In this paper, we presented MARVEL, a masked autoregressive pretraining framework for spectral understanding and generation. Generation quality depends on both the encoder and decoder, so better predictions do not necessarily imply stronger encoder representations. MARVEL establishes target-specific conditions from all visible observations and re-encodes supervised predictions as differentiable history. Each prediction receives its own reconstruction supervision and feedback from later predictions, incorporating cross-band dependencies into shared representation learning. Experiments show competitive performance on 14 downstream tasks and strong conditional generation on Sentinel-2 and Landsat-8 without additional generation training. Ablations support contributions from both content reuse and gradients through earlier predictions. Future work will explore efficient spectral decoding, predictive uncertainty, and fuller sensor-response descriptions, and evaluate transfer to unseen sensors and spectral fidelity on a physical radiometric scale.

AI USE STATEMENT

We used ChatGPT to assist with language polishing, grammar and wording corrections, and the preparation of selected figures and tables. We reviewed the AI-assisted content and take responsibility for the final manuscript and reported results.

ETHICS STATEMENT

We use publicly available remote sensing datasets that permit research use and comply with their respective license terms.

REPRODUCIBILITY STATEMENT

We describe the model and pretraining objectives in Section 3 and Appendix B. We summarize the data sources, training settings, and evaluation protocols in Appendices A.2 and C, and provide generation target sets and metric definitions in Appendices C.1 and A.5, respectively. We detail the controlled experiments assessing the history-gradient pathway in Appendix C.4.

REFERENCES

Alexander Baumann, Leonardo Ayala, Silvia Seidlitz, Jan Sellner, Alexander Studier-Fischer, Berkin Özdemir, Lena Maier-Hein, and Slobodan Ilic. Carl: Camera-agnostic representation learning for spectral image analysis. In *International Conference on Learning Representations*, volume 2026, pp. 23335–23357, 2026.

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. *Advances in neural information processing systems*, 28, 2015.

Yuanhao Cai, Jing Lin, Zudi Lin, Haoqian Wang, Yulun Zhang, Hanspeter Pfister, Radu Timofte, and Luc Van Gool. Mst++: Multi-stage spectral-wise transformer for efficient spectral reconstruction. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 744–754. Ieee, 2022.

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. *Advances in Neural Information Processing Systems*, 35:197–211, 2022.

Keli Deng, Jie Nie, and Yuntao Qian. Rgb pre-training enhanced unobservable feature latent diffusion model for spectral reconstruction. *arXiv preprint arXiv:2507.12967*, 2025.

Alaaeldin El-Nouby, Michal Klein, Shuangfei Zhai, Miguel Ángel Bautista, Vaishaal Shankar, Alexander T Toshev, Joshua M. Susskind, and Armand Joulin. Scalable pre-training of large autoregressive image models. In *Forty-first International Conference on Machine Learning*, 2024.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Danfeng Hong, Bing Zhang, Xuyang Li, Yuxuan Li, Chenyu Li, Jing Yao, Naoto Yokoya, Hao Li, Pedram Ghamisi, Xiuping Jia, et al. Spectralgpt: Spectral remote sensing foundation model. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5227–5244, 2024.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, et al. Terramind: Large-scale generative multimodality for earth observation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7383–7394. IEEE, 2025.

Alexandre Lacoste, Nils Lehmann, Pau Rodriguez, Evan Sherwin, Hannah Kerner, Björn Lütjens, Jeremy Irvin, David Dao, Hamed Alemohammad, Alexandre Drouin, et al. Geo-bench: Toward foundation models for earth monitoring. *Advances in Neural Information Processing Systems*, 36: 51080–51093, 2023.

Tianhong Li, Huiwen Chang, Shlok Mishra, Han Zhang, Dina Katabi, and Dilip Krishnan. Mage: Masked generative encoder to unify representation learning and image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2142–2152, 2023.

Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vector quantization. *Advances in Neural Information Processing Systems*, 37: 56424–56445, 2024.

Xuyang Li, Chenyu Li, and Danfeng Hong. Any-optical-model: A universal foundation model for optical remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 6539–6547, 2026.

Yunze Liu and Li Yi. Map: Unleashing hybrid mamba-transformer vision backbone's potential with masked autoregressive pretraining. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. i–x. IEEE, 2025.

Valerio Marsocci, Yuru Jia, Georges Le Bellier, David Kerekes, Liang Zeng, Sebastian Hafner, Sebastian Gerard, Eric Brune, Ritu Yadav, Ali Shibli, et al. Pangaea: A global and inclusive benchmark for geospatial foundation models. *arXiv preprint arXiv:2412.04204*, 2024.

Li Pang, Xiangyu Rui, Long Cui, Hongzhong Wang, Deyu Meng, and Xiangyong Cao. Hir-diff: Unsupervised hyperspectral image restoration via improved diffusion models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3005–3014. IEEE, 2024.

Jonathan Prexl and Michael Schmitt. Senpa-mae: Sensor parameter aware masked autoencoder for multi-satellite self-supervised pretraining. In *DAGM German Conference on Pattern Recognition*, pp. 317–331. Springer, 2024.

Haozhe Si, Yuxuan Wan, Yuqing Wang, Minh Do, and Han Zhao. Lessvit: Robust hyperspectral representation learning under spectral configuration shift. *arXiv preprint arXiv:2605.18541*, 2026.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Leonard Waldmann, Ando Shah, Yi Wang, Nils Lehmann, Adam 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*, pp. 2229–2239, 2025.

Xiangming Wang, Benteng Sun, Yungeng Liu, Haijin Zeng, Yongyong Chen, Jingyong Su, and Jie Liu. Hsi-var: Rethinking hyperspectral restoration through spatial-spectral visual autoregression. *arXiv preprint arXiv:2602.00749*, 2026.

Yi Wang, Zhitong Xiong, Chenying Liu, Adam J. Stewart, Thomas Dujardin, Nikolaos Ioannis Bountos, Angelos Zavras, Franziska Gerken, Ioannis Papoutsis, Laura Leal-Taixé, and Xiao Xiang Zhu. Towards a unified copernicus foundation model for earth vision, 2025.

Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *European conference on computer vision*, pp. 432–448. Springer, 2018.

Zhitong Xiong, Yi Wang, Fahong Zhang, Adam J Stewart, Joëlle Hanna, Damian Borth, Ioannis Papoutsis, Bertrand Le Saux, Gustau Camps-Valls, and Xiao Xiang Zhu. Neural plasticity-inspired multimodal foundation model for earth observation. *arXiv preprint arXiv:2403.15356*, 2024.

Zhao zuopeng, Ying Liu, xiaoyu Li, Su Luo, Lu Li, and Wenwen Liu. Anyband-diff: A unified remote sensing image generation and band repair framework with spectral priors. In *Forty-third International Conference on Machine Learning*, 2026.

A Supplementary Experiments

A.1 Decoder allocation and joint performance

Table A1 gives the numerical results for Figure 1. At a fixed depth of four blocks, AR-only decoding improves generation over Full-only decoding but weakens understanding. Allocating two blocks to each component improves both; increasing the AR share reduces both scores. Single-band inference uses no generated history, so these results assess generation after pretraining. The multi-target experiments in Section 4.4 additionally test history reuse during inference. PSNR follows the definitions in Appendix A.5.

Table A1: Decoder allocation at a fixed depth of four parameterized blocks. Full denotes full attention and AR denotes autoregressive decoding. GB is average GEO-Bench mIoU; generation is Sentinel-2 single-band recovery. Configurations correspond to the controls in Table A2. Bold and underlined values indicate the best and second-best results; blue shading marks the default configuration.

Decoder allocation	GB mIoU ↑	S2 PSNR ↑
Full $\times 4$	60.86	34.8
Full $\times 3$ + AR $\times 1$	<u>61.43</u>	35.5
Full $\times 2$ + AR $\times 2$	**61.89**	**36.8**
Full $\times 1$ + AR $\times 3$	61.25	<u>36.2</u>
AR $\times 4$	59.95	35.9

Table A2 extends the comparison to Landsat-8 and few-to-full generation. The combined decoder improves on either component alone across the reported settings. Adding the spectral conditioning and training components yields further gains in both understanding and generation.

Table A2: Understanding transfer and conditional generation across pretraining designs. GB: six-task GEO-Bench mean mIoU. S2/L8 Missing: single-band PSNR (dB); Few-to-full: mean of the two sensor-level Few PSNRs.

Pretraining design	GB mIoU ↑	S2 Missing ↑	L8 Missing ↑	Few-to-full ↑
Full-reconstruction MAE	60.58	33.9	32.8	29.6
Full attention only	60.86	34.8	33.7	29.6
Autoregressive only	59.95	35.9	34.6	30.6
Full initialization + AR	<u>61.89</u>	<u>36.8</u>	<u>35.4</u>	<u>31.7</u>
Full model	**63.15**	**37.2**	**35.8**	**32.2**

A.2 Pretraining data and input representation

The pretraining corpus includes Sentinel-2 imagery from SSL4EO-S12, Landsat-8 imagery from Activefire, and RGB imagery from GeoPile, fMoW, and OpenEarthMap (Table A3). Bands used together within a sample are co-registered on a common spatial grid. Center wavelength and bandwidth are normalized using fixed statistics from the training data. The shared spatial projection can be implemented by a three-dimensional convolution with kernel size and stride $(1, p, p)$, applying the same projection independently to each band.

Table A3: Pretraining data sources, comprising 1,560,000 samples in total. RGB collections are grouped together.

Source	Imagery	Resolution	Samples
SSL4EO-S12	Sentinel-2	10–60 m	1,004,000
Activefire	Landsat-8	30–100 m	146,000
GeoPile, fMoW, OpenEarthMap	RGB	0.1–30 m	410,000

Image inputs are standardized independently for each image and band using z-score normalization. In standard generation evaluation, target ground truth is used only for scoring; its values and statistics

are not used to normalize the observed inputs or to rescale predictions. Appendix A.5 defines the standardization and evaluation metrics. Ground-truth history used by SHF during pretraining is distinct from this evaluation protocol.

A.3 LABEL EFFICIENCY

For label fraction q, label efficiency is evaluated with the frozen pretrained backbone and task heads trained using that fraction of each task's training labels. The aggregate in Figure 5(a) is the arithmetic mean over the six GEO-Bench segmentation tasks:

$$\overline{\mathrm{mIoU}}(q) = \frac{1}{6}\sum_{d=1}^{6} \mathrm{mIoU}_d(q). \tag{9}$$

Table A4: Transfer with different label fractions. Values are six-task GEO-Bench mean mIoU (%).

Method	1% labels	5% labels	10% labels
MAE	31.0	43.1	49.0
SpectralGPT	33.0	45.2	51.3
ChannelMAE	<u>35.6</u>	<u>46.4</u>	<u>52.1</u>
MARVEL	**37.1**	**50.0**	**54.8**

A.4 UNDERSTANDING WITH REDUCED SPECTRAL OBSERVATIONS

Table A5 reports the mean mIoU on m-Cashew-Plantation and m-SA-Crop-Type at different observation fractions. At full observations, MARVEL scores 46.8, the mean of its 59.7 and 33.9 results in Table 1. At the sparsest reported setting, its 37.8 mIoU exceeds AOM's 35.1 by 2.7 points.

Table A5: Understanding with different fractions of Sentinel-2 bands retained. Values are average mIoU on m-Cashew-Plantation and m-SA-Crop-Type.

Method	100% bands	75% bands	50% bands	25% bands
ChannelMAE	43.65	40.9	36.7	30.8
DOFA	44.45	42.4	38.8	33.5
AOM	<u>45.05</u>	<u>43.2</u>	<u>40.0</u>	<u>35.1</u>
MARVEL	**46.8**	**44.7**	**42.1**	**37.8**

A.5 ADDITIONAL GENERATION METRICS AND TARGET QUERIES

Evaluation scale. We apply z-score standardization independently to each image and band. For image n, band c, and spatial position u, let $I_{n,c,u}$ denote the original pixel value. The standardized value is

$$X_{n,c,u} = \frac{I_{n,c,u} - \mu_{n,c}}{\sigma_{n,c}}, \tag{10}$$

where $\mu_{n,c}$ and $\sigma_{n,c}$ are the spatial mean and standard deviation computed separately for band c of image n. Each observed band uses its own statistics. Predictions are compared directly with ground-truth references on the standardized scale. The same valid regions are used for predictions and references, and predicted values are not clipped before evaluation. PSNR uses a fixed data range of $L = 1$; this convention does not constrain standardized values to $[0, 1]$. Ground-truth target bands are standardized only to form the metric references; their statistics are not used to normalize observed inputs or rescale predictions.

PSNR and SSIM. For image n and requested target set $\mathcal{T}$, let $\mathcal{I}_{n,\mathcal{T}}$ contain the valid band–pixel indices across the requested target bands. With $\widehat{X}_{n,c,u}$ and $X_{n,c,u}$ denoting the standardized prediction

and reference, respectively, the image-level error and PSNR are

$$\mathrm{MSE}_{n,\mathcal{T}} = \frac{1}{|\mathcal{I}_{n,\mathcal{T}}|} \sum_{(c,u)\in\mathcal{I}_{n,\mathcal{T}}} \left(\widehat{X}_{n,c,u} - X_{n,c,u}\right)^2,$$
$$\mathrm{PSNR}_{n,\mathcal{T}} = 10\log_{10} \frac{1}{\mathrm{MSE}_{n,\mathcal{T}}}. \tag{11}$$

For multi-target requests, squared errors are pooled across the requested bands before conversion to PSNR, rather than averaging independently computed per-band PSNRs. Missing evaluates one target band per request. Reported PSNR scores are arithmetic means of image-level PSNR values over the test set, rather than PSNR computed from a dataset-level pooled MSE. SSIM is computed separately for target-band images and averaged over successful evaluations. Sensor-averaged results give equal weight to the two reported sensor scores.

Spectral fidelity. PSNR, SSIM, and correlation characterize complementary aspects of reconstruction. Physical spectral fidelity additionally requires evaluation on an appropriate radiometric scale, including spectral-angle and application-index errors. Independent band standardization changes spectral angles, so the reported image-recovery metrics alone do not establish that form of fidelity.

Table A6: Additional metrics for single-band generation under the evaluation protocol in Appendix A.5. Higher PSNR, SSIM, and correlation (Corr.) indicate better performance.

	Sentinel-2			Landsat		
Method	**PSNR**	**SSIM**	**Corr.**	**PSNR**	**SSIM**	**Corr.**
Full-reconstruction MAE	33.9	0.905	0.948	32.8	0.893	0.941
Autoregressive only	35.9	0.923	0.959	34.6	0.912	0.952
AnyBand-Diff	<u>36.9</u>	**0.947**	**0.971**	**35.9**	**0.934**	**0.964**
MARVEL	**37.2**	<u>0.944</u>	<u>0.969</u>	<u>35.8</u>	<u>0.931</u>	<u>0.962</u>

Table A7: Descriptor-conditioned target-band generation. Target pixels are withheld from model inputs; center wavelength and bandwidth specify each request. Target ground truth is used only as the metric reference. Cross-sensor targets use harmonized measured bands.

Query setting	**PSNR ↑**	**SSIM ↑**
S2 input, held-out S2 target	37.2	0.944
L8 input, held-out L8 target	35.8	0.931
S2 input, Landsat target	33.6	0.901
Landsat input, S2 target	33.2	0.895

A.6 Spectral generation across observation fractions

Table A8: Few-to-full generation across observed band fractions. Values are PSNR (dB), averaged over Sentinel-2 and Landsat. This fraction sweep is evaluated separately from the fixed Few setting in Table 3.

Method	**75% observed**	**50% observed**	**25% observed**
Full-reconstruction MAE	34.5	31.9	29.8
Autoregressive only	35.1	32.8	30.8
AnyBand-Diff	**36.2**	**34.3**	**32.5**
MARVEL	<u>36.0</u>	<u>34.0</u>	<u>32.2</u>

A.7 History scheduling, band descriptors, and curvature matching

The curvature matching loss has weight $\beta = 10^{-6}$ and uses the wavelength-aware operator and patch-wise reduction in Appendix B.5. In SHF, the mixing weight of ground-truth history decreases linearly from 0.3 to zero over the first 50 training epochs. Subsequent training and inference use representations of the model's own predicted content.

Table A9: Training and descriptor choices. GB is average GEO-Bench mIoU; generation is Sentinel-2 single-band PSNR (dB). Rankings are computed within each block.

Variant	GB mIoU ↑	S2 Missing ↑
Prediction-only history	61.71	36.5
Fixed teacher forcing	61.44	36.8
SHF	**63.15**	**37.2**
Center wavelength only	61.76	36.7
Wavelength + bandwidth	**63.15**	**37.2**
+ sensor-ID embedding	61.98	37.0
Without curvature regularizer	62.17	37.1
With curvature regularizer	**63.15**	**37.2**

A.8 Spectral order and observation visibility

Table A10 expands the λ-RoPE comparison in Table 4(a) with alternative band orders and Landsat-8 results. Wavelength-based orders outperform random and channel-index orders, and λ-RoPE further improves both understanding and recovery.

Table A10: Spectral order and λ-RoPE at fixed decoder depth. GB: mean mIoU; S2/L8 Missing: single-band PSNR (dB).

Variant	GB mIoU ↑	S2 Missing ↑	L8 Missing ↑
Random band order	60.94	35.6	34.2
Channel-index order	61.28	36.2	34.7
Reverse wavelength order	61.93	36.9	35.5
Ascending wavelength, no λ-RoPE	61.74	36.6	35.2
Ascending wavelength, λ-RoPE	**63.15**	**37.2**	**35.8**

Table A11 evaluates observation visibility using the same pretrained checkpoint. Restricting observations to wavelengths below each target reduces Missing PSNR by 1.4 dB on both sensors, supporting access to all measured bands. Few is unchanged because every RGB observation precedes its targets. The inference protocol is described in Appendix C.3.

Table A11: Observation visibility with the same pretrained checkpoint. Truncation removes observed wavelengths above the current target before encoding. Values are PSNR (dB); Few-to-full averages both sensors.

Condition memory	S2 Missing	L8 Missing	Few-to-full
All available observed bands	**37.2**	**35.8**	**32.2**
Wavelength-truncated	35.8	34.4	**32.2**

A.9 QUALITATIVE MULTI-BAND GENERATION

Figures 8–11 show 32 examples of simultaneous multi-band generation by MARVEL, complementing the quantitative Multi results in Table 3.

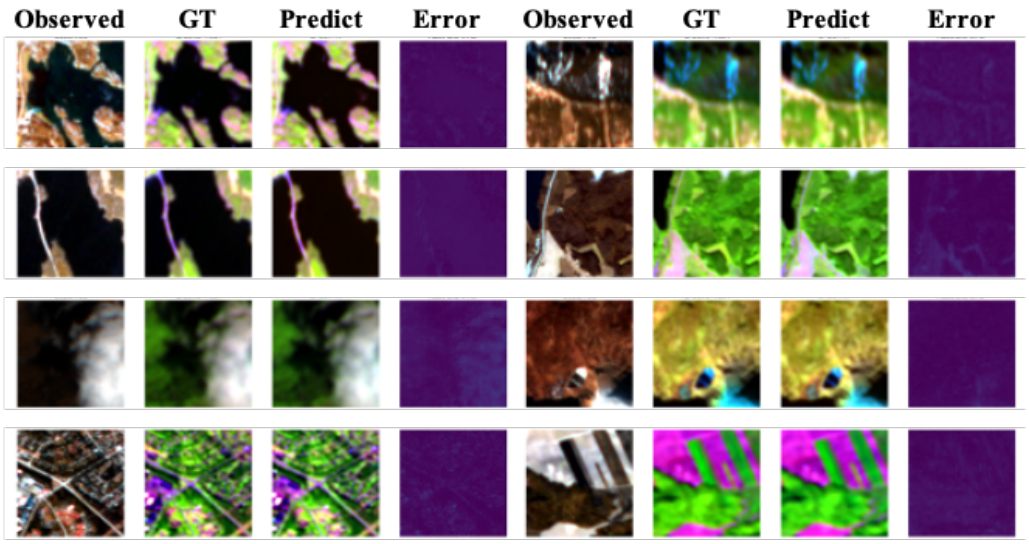

Figure 8: Multi-band generation by MARVEL (examples 1–8). From left to right: observed input, ground-truth composite, generated target-band composite, and absolute error map.

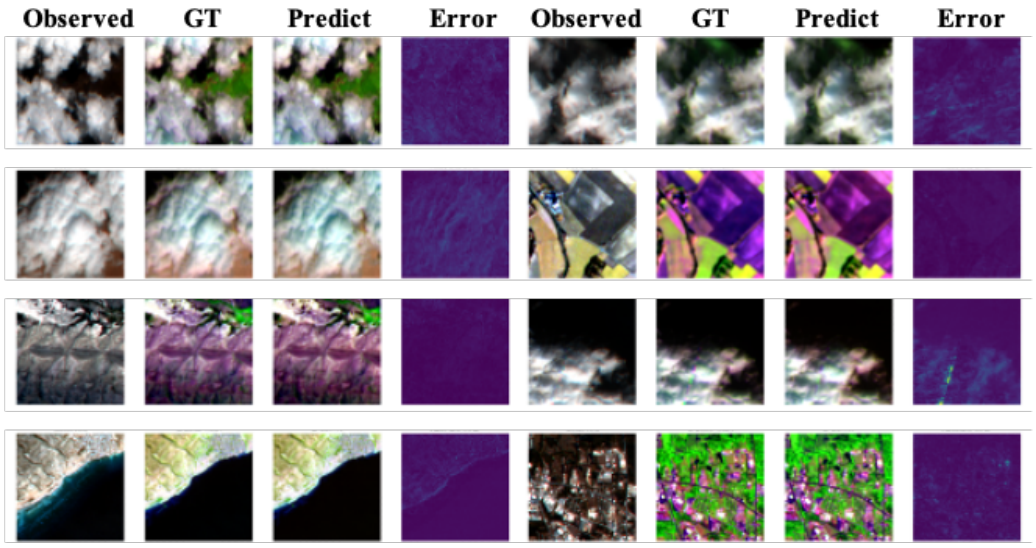

Figure 9: Multi-band generation by MARVEL (examples 9–16). From left to right: observed input, ground-truth composite, generated target-band composite, and absolute error map.

B EXPANDED METHOD FORMULATION

This appendix details the shared representation, conditional decoder, history construction, and curvature loss used in Section 3.

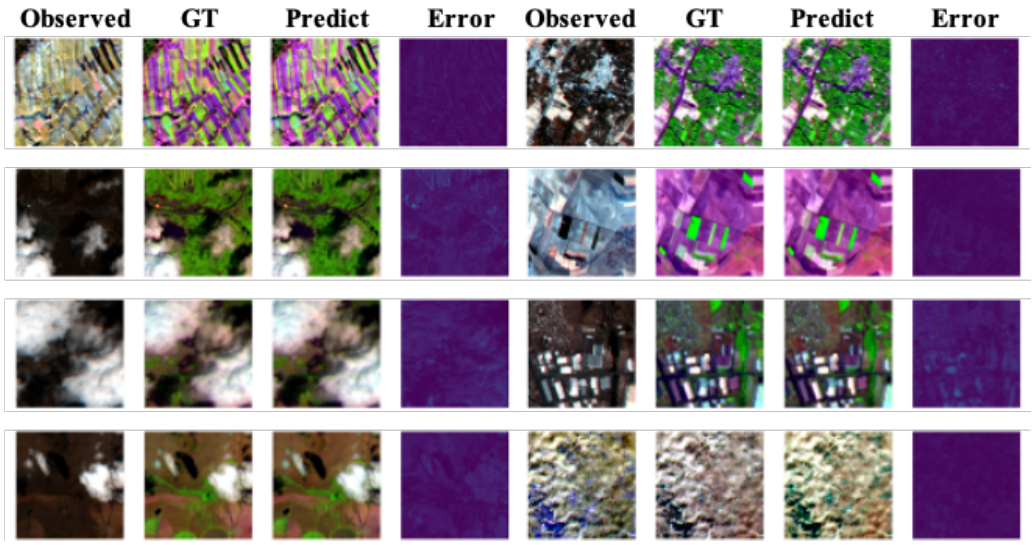

Figure 10: Multi-band generation by MARVEL (examples 17–24). From left to right: observed input, ground-truth composite, generated target-band composite, and absolute error map.

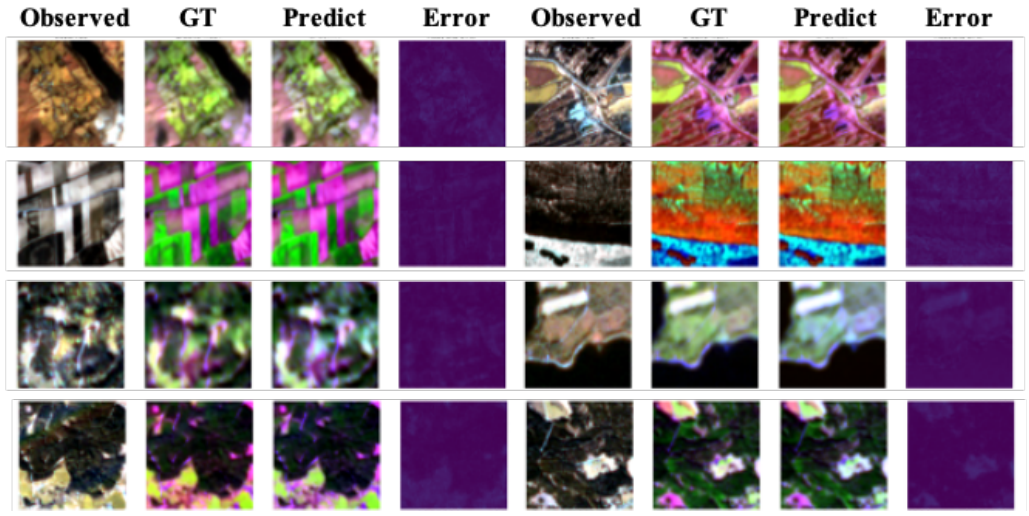

Figure 11: Multi-band generation by MARVEL (examples 25–32). From left to right: observed input, ground-truth composite, generated target-band composite, and absolute error map.

B.1 Problem formulation

A spectral image is represented as a set of bands with physical descriptors:

$$\mathcal{X} = \{(\mathbf{X}_c, \mathbf{d}_c)\}_{c=1}^{C}, \qquad \mathbf{d}_c = (\lambda_c, \Delta\lambda_c), \tag{12}$$

where $\mathbf{X}_c \in \mathbb{R}^{H\times W}$ is a band image, and λ_c and $\Delta\lambda_c$ are its center wavelength and bandwidth. Physical descriptors identify bands across configurations without requiring a fixed channel count.

Given observed bands $\mathcal{O}$, understanding uses encoder features for downstream prediction. Generation additionally receives descriptors $\mathbf{D}_\mathcal{T}$ for target bands $\mathcal{T}$. Both uses share an encoder:

$$\begin{aligned} \mathbf{F}_\mathcal{O} &= E_\theta(\mathcal{X}_\mathcal{O}), \\ \widehat{\mathbf{y}} &= H_\psi(\mathbf{F}_\mathcal{O}), \\ \widehat{\mathbf{X}}_\mathcal{T} &= G_\theta(\mathbf{F}_\mathcal{O}, \mathbf{D}_\mathcal{O}, \mathbf{D}_\mathcal{T}), \end{aligned} \tag{13}$$

where H_ψ is a downstream task head, G_θ is the conditional decoder, and $\mathbf{D}_\mathcal{O}$ is supplied with the observations. For whole-band recovery, $\mathcal{O} \cap \mathcal{T} = \varnothing$, and the target descriptors specify the spectral positions to generate.

Pretraining operates on patches indexed in space and wavelength. Let $i = (c, u)$ denote spatial patch position u in band c, and let $\mathcal{V}$ and $\mathcal{M}$ denote visible and target positions, respectively. Random masking samples $\mathcal{M}$ from all patches, so a band can contain both visible and masked positions. Whole-band recovery instead corresponds to

$$\begin{aligned} \mathcal{M} &= \{(t, u) : t \in \mathcal{T},\ u \in \mathcal{U}\}, \\ \mathcal{V} &= \{(o, u) : o \in \mathcal{O},\ u \in \mathcal{U}\}, \end{aligned} \tag{14}$$

where $\mathcal{U}$ contains all spatial patch positions. Wavelength order applies to target predictions, while all visible content remains available, including observations at wavelengths longer than the current target.

B.2 Shared spectral representation

Channel indices can denote different wavelengths across sensors. Physical descriptors place observed bands and target requests in a common spectral coordinate system, alongside their content and spatial positions.

Let $\bar{\lambda}_c$ and $\overline{\Delta\lambda}_c$ denote normalized center wavelength and bandwidth. A multifrequency Fourier mapping constructs the band embedding:

$$\begin{aligned} \phi(v) &= [\sin(\omega_1 v), \cos(\omega_1 v), \ldots, \sin(\omega_J v), \cos(\omega_J v)], \\ \mathbf{e}_c^{\text{band}} &= f_{\text{band}}([\phi(\bar{\lambda}_c); \phi(\overline{\Delta\lambda}_c)]). \end{aligned} \tag{15}$$

Here, J is the number of frequencies, and f_{band} maps continuous descriptors into feature space. A spatial projector P, shared across bands, extracts local content and combines it with band and spatial embeddings:

$$\mathbf{t}_{c,u} = P(\mathbf{x}_{c,u}) + \alpha \mathbf{e}_c^{\text{band}} + \mathbf{e}_u^{\text{sp}}, \tag{16}$$

where $\mathbf{x}_{c,u}$ is patch content, $\mathbf{e}_u^{\text{sp}}$ is a two-dimensional spatial embedding, and α is a learnable gain. The shared projector has a parameter count independent of the number of bands; descriptor embeddings distinguish their spectral responses.

We aggregate observations bidirectionally so that each visible token can incorporate context from other measured bands and spatial positions:

$$\begin{aligned} \mathbf{H}_{\text{enc}} &= E_{\text{bi}}\left([\mathbf{t}_{\text{cls}}; \{\mathbf{t}_i\}_{i \in \mathcal{V}}]\right), \\ \mathbf{M}_{\text{obs}} &= \Pi(\mathbf{H}_{\text{enc},\mathcal{V}}), \end{aligned} \tag{17}$$

where $\mathbf{t}_{\text{cls}}$ is a global token, and Π projects visible-position features into the decoder dimension to form observation memory $\mathbf{M}_{\text{obs}}$. Each memory position retains its band and spatial coordinates while integrating content from other observations.

Understanding uses pooled features for global prediction or retains spatial features for dense prediction. Generation accesses the observation memory through target descriptors.

B.3 Conditional spectral autoregression

The decoder uses all observations to initialize target-specific states, then generates target content in wavelength order. Completed predictions form history for subsequent targets, while observation memory remains available throughout.

Full-context alignment gives each target access to observations before its prediction enters history. For target band t and spatial position u, we construct the query

$$\mathbf{q}_{t,u} = f_{\text{query}}(\mathbf{m}, \mathbf{d}_t, \mathbf{e}_u^{\text{sp}}). \tag{18}$$

where $\mathbf{d}_t = (\lambda_t, \Delta\lambda_t)$ and f_{query} combines the learnable mask embedding, Fourier features of normalized band descriptors, and spatial encoding in decoder space, as in Equation 2. This query

construction is shared by random-mask pretraining and whole-band generation. All queries and observation memory pass jointly through full attention:

$$[\mathbf{M}_{\text{obs}}^{(0)}; \mathbf{Z}_{\mathcal{M}}^{(0)}] = D_{\text{full}}([\mathbf{M}_{\text{obs}}; \mathbf{Q}_{\mathcal{M}}]). \tag{19}$$

The outputs are target-specific conditional states $\mathbf{Z}_{\mathcal{M}}^{(0)}$ and updated observation memory $\mathbf{M}_{\text{obs}}^{(0)}$. Autoregressive decoding uses these states, the updated memory, and preceding predictions.

Suppose masked positions span K bands. Let π order them by increasing center wavelength and $\mathcal{U}_k = \{u : (\pi(k), u) \in \mathcal{M}\}$ contain the target positions in band $\pi(k)$. Then

$$\begin{aligned} &\lambda_{\pi(1)} \le \lambda_{\pi(2)} \le \cdots \le \lambda_{\pi(K)}, \\ &\mathbf{Y}_k = \{\mathbf{x}_{\pi(k),u} : u \in \mathcal{U}_k\}. \end{aligned} \tag{20}$$

Spatial positions within each target band are predicted together. $\mathbf{Y}_k$ contains masked patches during pretraining and all spatial positions during whole-band generation.

Completed predictions are re-encoded as history, with each prediction receiving reconstruction supervision at its own target positions. With ρ_θ denoting the history encoder, the recurrence is

$$\begin{aligned} \mathbf{S}_{<k} &= \text{Concat}_{j<k}\, \rho_\theta(\widehat{\mathbf{Y}}_j, \mathbf{d}_{\pi(j)}, \mathcal{U}_j), \\ \widehat{\mathbf{Y}}_k &= F_{\text{AR},\theta}\big(\mathbf{Z}_k^{(0)}, \mathbf{M}_{\text{obs}}^{(0)}, \mathbf{S}_{<k}\big), \end{aligned} \tag{21}$$

where $\mathbf{Z}_k^{(0)}$ is the current target's conditional state. The history encoder uses a content projection shared across bands and adds band and spatial encodings:

$$\rho_\theta(\mathbf{x}_{c,u}, \mathbf{d}_c, u) = P_{\text{hist}}(\mathbf{x}_{c,u}) + \mathbf{e}_{c,u}^{\text{dec}}, \tag{22}$$

Here, P_{hist} projects patch content, and $\mathbf{e}_{c,u}^{\text{dec}}$ combines band and spatial encodings in decoder space.

To account for irregular spectral spacing, λ-RoPE uses normalized center wavelengths to rotate queries and keys in both observation and history attention (Su et al., 2024; Si et al., 2026). For linearly projected queries $\mathbf{q}_i$ and keys $\mathbf{k}_j$,

$$\begin{aligned} \widetilde{\mathbf{q}}_i = R(\bar{\lambda}_i)\mathbf{q}_i, \qquad \widetilde{\mathbf{k}}_j = R(\bar{\lambda}_j)\mathbf{k}_j, \\ \widetilde{\mathbf{q}}_i^{\mathsf{T}}\widetilde{\mathbf{k}}_j = \mathbf{q}_i^{\mathsf{T}} R(\bar{\lambda}_j - \bar{\lambda}_i)\mathbf{k}_j. \end{aligned} \tag{23}$$

where R acts on feature pairs, with phases given by fixed frequencies multiplied by normalized center wavelength. Band embeddings describe each measurement's center wavelength and bandwidth, while λ-RoPE introduces relative wavelength differences into attention.

Starting from $\mathbf{Z}_k^{(0)}$, each autoregressive layer attends separately to observations and history:

$$\begin{aligned} \mathbf{U}_k^{(\ell)} &= \mathbf{Z}_k^{(\ell)} + \mathcal{A}_\lambda^{(\ell)}\big(\mathbf{Z}_k^{(\ell)}, \mathbf{M}_{\text{obs}}^{(0)}\big) \\ &\quad + \mathcal{A}_\lambda^{(\ell)}\big(\mathbf{Z}_k^{(\ell)}, \mathbf{S}_{<k}\big), \\ \mathbf{Z}_k^{(\ell+1)} &= \mathbf{U}_k^{(\ell)} + \text{MLP}^{(\ell)}(\mathbf{U}_k^{(\ell)}). \end{aligned} \tag{24}$$

where $\mathcal{A}_\lambda^{(\ell)}(\mathbf{A}, \mathbf{B})$ uses $\mathbf{A}$ as queries and $\mathbf{B}$ as keys and values, with linear projections and λ-RoPE. Observation attention uses all measurements; history attention uses only preceding predictions.

After L updates, the reconstruction head predicts the current band, and its re-encoded content is appended to history:

$$\begin{aligned} \widehat{\mathbf{x}}_{\pi(k),u} &= f_{\text{out}}(\mathbf{Z}_{k,u}^{(L)}), \qquad u \in \mathcal{U}_k, \\ \mathbf{S}_{<k+1} &= [\mathbf{S}_{<k}; \rho_\theta(\widehat{\mathbf{Y}}_k, \mathbf{d}_{\pi(k)}, \mathcal{U}_k)]. \end{aligned} \tag{25}$$

The first step uses empty history and a zero history-attention contribution. Later steps reuse completed predictions while retaining access to all observations.

B.4 TRAINING OBJECTIVE AND HISTORY SCHEDULING

Each masked target receives reconstruction supervision and can condition later targets through history. Let $\mathbf{S}_{<k}(e)$ denote the history at epoch e, constructed using the schedule below. The reconstruction objective is

$$\mathcal{L}_{\mathrm{rec}} = \frac{1}{|\mathcal{M}|} \sum_{k=1}^{K} \sum_{u \in \mathcal{U}_k} \left\| \left[F_{\mathrm{AR},\theta}\big(\mathbf{Z}_k^{(0)}, \mathbf{M}_{\mathrm{obs}}^{(0)}, \mathbf{S}_{<k}(e)\big) \right]_u - \mathbf{x}_{\pi(k),u} \right\|_2^2. \tag{26}$$

The current target's loss updates the shared encoder through its conditional state and observation memory. Differentiable history also carries later losses through earlier predictions to the encoder features that produced them. Each prediction is trained to fit its target and support later generation.

Early prediction errors can make history unreliable, while exclusive ground-truth conditioning creates a mismatch with inference. Spectral History Forcing (SHF), motivated by scheduled sampling (Bengio et al., 2015), blends ground-truth and predicted representations while gradually removing the ground-truth contribution. For an earlier position $(\pi(j), u)$, the history token is

$$\begin{aligned} \mathbf{m}_{j,u}(e) = r(e)\,\mathrm{stopgrad}\left[\rho_\theta(\mathbf{x}^{\mathrm{gt}}_{\pi(j),u}, \mathbf{d}_{\pi(j)}, u) + \boldsymbol{\epsilon}_{j,u}\right] \\ + [1 - r(e)]\rho_\theta(\widehat{\mathbf{x}}_{\pi(j),u}, \mathbf{d}_{\pi(j)}, u). \end{aligned} \tag{27}$$

Both branches use the same history encoder ρ_θ. The coefficient $r(e)$ decreases with epoch, and $\boldsymbol{\epsilon}_{j,u}$ is a small perturbation of the ground-truth representation. Ground-truth history is detached; predicted history retains gradients through its encoding and earlier outputs. Concatenating tokens for $j < k$ forms $\mathbf{S}_{<k}(e)$. Ground-truth assistance is restricted to preceding targets and does not enter observation memory or full-context alignment.

Predictions remain sequential throughout training. Once $r(e) = 0$, training and inference use the same predicted-history construction.

We combine reconstruction with the spectral curvature matching loss defined below:

$$\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \beta \mathcal{L}_{\mathrm{curv}}, \qquad \beta = 10^{-6}. \tag{28}$$

B.5 WAVELENGTH-AWARE SPECTRAL CURVATURE MATCHING

The auxiliary loss compares the local spectral shape of the completed reconstruction with that of the training target. The formulation below follows the implemented wavelength sorting, finite differences, mask selection, and reduction.

Completed reconstruction. For a batch of B samples, let C be the number of bands, T the number of spatial patches per band, and $Q = p^2$ the number of pixel values in each single-band patch. The prediction and target tensors are reshaped from $B \times (CT) \times Q$ to $B \times C \times T \times Q$. We write their entries as $\widehat{y}_{b,c,u,q}$ and $y_{b,c,u,q}$, respectively, and let $m_{b,c,u} \in \{0, 1\}$ indicate a masked patch. All band-indexed quantities are ordered by center wavelength separately for each sample, so that $\lambda_{b,1} \le \cdots \le \lambda_{b,C}$. The completed reconstruction is

$$r_{b,c,u,q} = \begin{cases} \widehat{y}_{b,c,u,q}, & m_{b,c,u} = 1, \\ y_{b,c,u,q}, & m_{b,c,u} = 0. \end{cases} \tag{29}$$

Thus, visible measurements supply the unmasked entries, while predictions supply the masked ones. The reference tensor y is used as training supervision. The matrices in Equation 8 are $\mathbf{R}_{b,u} = [r_{b,c,u,q}]_{c,q}$ and $\mathbf{Y}_{b,u} = [y_{b,c,u,q}]_{c,q}$.

Differences on the wavelength grid. Let $\varepsilon > 0$ be the denominator floor, with default value 10^{-6} in the implementation. For any band-indexed tensor z, define the first divided difference

$$\big(\mathcal{D}^{(1)}_{\lambda_b} z\big)_{i,u,q} = \frac{z_{b,i+1,u,q} - z_{b,i,u,q}}{\max(\lambda_{b,i+1} - \lambda_{b,i}, \varepsilon)}, \quad i = 1, \ldots, C-1. \tag{30}$$

The second-order operator compares neighboring slopes and divides by the wavelength span of the three-band stencil:

$$\big(\mathcal{D}^{(2)}_{\lambda_b} z\big)_{i,u,q} = \frac{\big(\mathcal{D}^{(1)}_{\lambda_b} z\big)_{i,u,q} - \big(\mathcal{D}^{(1)}_{\lambda_b} z\big)_{i-1,u,q}}{\max(\lambda_{b,i+1} - \lambda_{b,i-1}, \varepsilon)}, \quad i = 2, \ldots, C-1. \tag{31}$$

Both operators use the numeric center wavelengths supplied to the loss. The spacing factors account for an irregular wavelength grid, and the implementation introduces no additional factor of two. When the denominator floor is inactive, Equation 31 is the second divided difference, equal to one half of the usual three-point second-derivative estimate. On a uniform grid of spacing h, it reduces to $(z_{b,i+1,u,q} - 2z_{b,i,u,q} + z_{b,i-1,u,q})/(2h^2)$.

Mask weighting and reduction. A spatial patch contributes if at least one of its bands is masked:

$$w_{b,u} = \mathbb{I}\left[\sum_{c=1}^{C} m_{b,c,u} > 0\right], \qquad \mathcal{A}_b = \{u : w_{b,u} = 1\}. \tag{32}$$

For $C \geq 3$, the exact loss is

$$\ell_b^{\text{curv}} = \frac{\sum_{u=1}^{T} w_{b,u} \sum_{i=2}^{C-1} \sum_{q=1}^{Q} \left| \left(\mathcal{D}^{(2)}_{\lambda_b} r\right)_{i,u,q} - \left(\mathcal{D}^{(2)}_{\lambda_b} y\right)_{i,u,q} \right|}{\max\left(\sum_{u=1}^{T} w_{b,u}, 1\right)},$$
$$\mathcal{L}_{\text{curv}} = \frac{1}{B} \sum_{b=1}^{B} \ell_b^{\text{curv}}. \tag{33}$$

The weight $w_{b,u}$ is shared across all spectral triplets and pixels within a spatial patch. A fully observed triplet has identical entries in r and y and therefore contributes zero. Each sample is normalized only by its number of contributing spatial patches: the implementation sums, rather than averages, over the $C - 2$ triplets and Q within-patch pixels. Samples with no masked patches contribute zero, and the loss is defined as zero when $C < 3$.

Interpretation. This term penalizes differences between reconstructed and target changes in spectral slope, complementing the pointwise reconstruction objective. Because target curvature is retained as the reference, nonzero curvature in the measured spectrum is not penalized merely for being nonzero. The constraint operates on the tensor scale supplied during training and does not impose a complete physical sensor-response model. Its magnitude depends on the wavelength scale, band configuration, and patch size through the stated operator and reduction; the weight β is tied to those conventions.

C Additional experimental details

Pretraining uses Sentinel-2, Landsat-8, and RGB imagery; Appendix A.2 summarizes the data composition. We use a ViT-Base encoder with 8×8 patches and randomly mask 75% of spatial–spectral patch positions. Training runs for 100 epochs with 10 warmup epochs, a per-process batch size of 88, four gradient accumulation steps, a base learning rate of 10^{-4}, weight decay of 0.05, and 128×128 inputs. The default decoder contains two full-context blocks and two autoregressive blocks. In SHF, the weight assigned to ground-truth history decreases linearly from 0.3 to zero over the first 50 epochs. Band descriptors contain center wavelength and bandwidth, and autoregressive decoding uses λ-RoPE.

Understanding and generation use the same pretrained checkpoint. For understanding, we freeze the encoder and train a linear head for classification or a UPerNet head Xiao et al. (2018) for semantic segmentation. We use the original public train/validation/test splits of each downstream dataset. For each task, we search learning rates in $\{10^{-3}, 5 \times 10^{-4}, 10^{-4}, 5 \times 10^{-5}, 10^{-5}, 5 \times 10^{-6}\}$, select the learning rate using validation performance, and report the corresponding test-set result. Generation uses the complete pretrained model directly, with no additional optimization.

The main understanding benchmark comprises six GEO-Bench semantic segmentation tasks spanning RGB, RGBN, and Sentinel-2 imagery, evaluated by average mIoU (Lacoste et al., 2023). Eight additional transfer tasks include m-BigEarthNet, m-EuroSAT, and m-So2Sat classification; HLS Burn Scars and MADOS dense prediction under the PANGAEA protocol (Marsocci et al., 2024); and Cloud-S2, DFC2020-S2, and LC100Cls-S3 from Copernicus-Bench (Wang et al., 2025). BEN uses mAP; EuroSAT, So2Sat, and LC100Cls-S3 use accuracy; the remaining tasks use mIoU. We also evaluate label efficiency and robustness to changes in available spectral observations by reducing downstream labels and input bands, respectively.

Understanding baselines. SpectralGPT, DOFA, Panopticon, Copernicus-FM, and TerraMind are evaluated using their official released pretrained checkpoints. MAE and ChannelMAE are pretrained by us, and AOM is reproduced using our training implementation. All understanding baselines use Base-sized encoders and follow the same frozen-encoder evaluation protocol as MARVEL. Models initialized from official checkpoints retain their original pretraining data and optimization histories.

Generation baselines. SpectralMAE is pretrained by us. AnyBand-Diff is reproduced by us and trained separately for Sentinel-2 and Landsat-8 using the same sensor-specific training data as MARVEL. Inference uses 50 sampling steps, following the original paper (zuopeng et al., 2026), and retains the physics-guided sampling module. MST++ is trained separately for Sentinel-2 and Landsat-8 and evaluated in the fixed RGB-to-remaining-band setting. Within each evaluated setting, the learned generation models receive identical observed bands and target requests and are evaluated on the same scenes using the metric definitions in Appendix A.5.

Interpolation baseline. The nonparametric baseline uses linear spectral interpolation. For a target center wavelength λ_t, we select the nearest observed band on each side, with center wavelengths $\lambda_l < \lambda_t < \lambda_r$. At each aligned spatial position u, the target is estimated by

$$\widehat{X}_t(u) = (1 - a_t)X_l(u) + a_t X_r(u), \qquad a_t = \frac{\lambda_t - \lambda_l}{\lambda_r - \lambda_l}. \tag{34}$$

The weights use actual center wavelengths, accounting for unequal spectral spacing. Each target is interpolated independently from observed bands. We do not extrapolate beyond the observed wavelength range. We report results only for Multi, where all target wavelengths lie within the observed range. Missing includes boundary-band targets outside the range of the remaining observations, while Few requests non-RGB bands beyond the RGB wavelength range. We therefore do not report aggregate results for Missing or Few and mark these entries with dashes in Table 3.

Internal controls. Full-reconstruction MAE uses the same pretraining data and training budget as MARVEL. The Full-only, AR-only, and combined-decoder controls use the same encoder architecture, pretraining data, masking strategy, and training duration, with four parameterized decoder blocks. The Full-only decoder uses target descriptors and full attention in every block, whereas the AR-only decoder allocates all four blocks to autoregressive prediction. The combined decoder uses two full-context blocks followed by two autoregressive blocks. The complete model additionally incorporates λ-RoPE, bandwidth descriptors, SHF, and spectral curvature matching. Full-only and AR-only controls are trained in our implementation. These decoder-allocation comparisons hold parameterized block depth fixed, which does not imply equal parameter counts or training FLOPs.

C.1 Generation data and target sets

The same-sensor evaluation uses independent Sentinel-2 Level-2A and Landsat-8 surface-reflectance scenes, separated from pretraining at the scene or geographic-region level before extracting evaluation patches. Bands within a sensor are aligned to a common spatial grid. The Sentinel-2 evaluation set is $\mathcal{B}_{\mathrm{S2}} = \{\mathrm{B2}, \mathrm{B3}, \mathrm{B4}, \mathrm{B5}, \mathrm{B6}, \mathrm{B7}, \mathrm{B8}, \mathrm{B8A}, \mathrm{B11}, \mathrm{B12}\}$, and the Landsat-8 set is $\mathcal{B}_{\mathrm{L8}} = \{\mathrm{B2}, \mathrm{B3}, \mathrm{B4}, \mathrm{B5}, \mathrm{B6}, \mathrm{B7}\}$. These sets cover visible, near-infrared, and shortwave-infrared responses; thermal bands are not included.

For each sensor, the evaluation set comprises approximately 100 geographically distributed scenes and approximately 6,000 valid patches retained after invalid-data filtering. Scenes are selected to cover urban areas, croplands, forests, water bodies, and bare land. Patch sampling aims to minimize spatial overlap and avoid concentrating samples in a small number of scenes. Missing, Multi, and Few reuse the same evaluation patches, changing only the observed and target band sets.

Missing iterates over each band $c \in \mathcal{B}$, using $\mathcal{O} = \mathcal{B} \setminus \{c\}$ and $\mathcal{T} = \{c\}$. Multi uses the fixed target sets in Table A12 with the remaining bands observed. Few supplies RGB bands and recovers every other band in the sensor's evaluation set. All methods use identical observation sets, target sets, and evaluation scenes. Only target-band predictions contribute to recovery metrics. Cross-sensor evaluation uses separate harmonized pairs and is not part of these same-sensor holdouts. Qualitative examples of simultaneous multi-band generation appear in Figures 8–11.

Table A12: Fixed target sets for conditional recovery. For Multi, observations are the complement of the target set within $\mathcal{B}$. For Few, observations are RGB bands B2, B3, and B4.

Setting	Sentinel-2 targets	Landsat-8 targets
Multi	B3, B8A, B11	B3, B5, B6
Few	B5, B6, B7, B8, B8A, B11, B12	B5, B6, B7

C.2 History-content controls

History representations. The predicted-content, query-state, spatially shuffled, and observation-only configurations are trained separately. Predicted-content history re-encodes completed pixel predictions using the history encoder. Query-state history instead reuses the final decoder states $\mathbf{Z}_j^{(L)}$ of preceding targets, before the reconstruction head, as history for subsequent predictions. These states also receive reconstruction supervision through the output head. This control therefore compares direct decoder-state reuse with history constructed by re-encoding explicit pixel predictions. Observation-only conditioning omits generated history. The GT-history oracle uses preceding ground-truth targets as history and serves as a privileged reference.

Target-count sweep. Figure 7 uses the ten-band Sentinel-2 evaluation set $\mathcal{B}_{\mathrm{S2}}$. We construct nested target sets from the ordered sequence

$$(b_1, \ldots, b_6) = (\mathrm{B5}, \mathrm{B6}, \mathrm{B7}, \mathrm{B8A}, \mathrm{B11}, \mathrm{B12}).$$

For a missing-band count $m \in \{1, \ldots, 6\}$, the target and observation sets are

$$\mathcal{T}_m = \{b_1, \ldots, b_m\}, \qquad \mathcal{O}_m = \mathcal{B}_{\mathrm{S2}} \setminus \mathcal{T}_m.$$

Thus, each successive setting adds one target band and retains the remaining bands as observations. At $m = 6$, the observed bands are $\{\mathrm{B2}, \mathrm{B3}, \mathrm{B4}, \mathrm{B8}\}$. All configurations use the same evaluation images, observation sets, and target sets. Targets are generated in increasing wavelength order. Panel (a) reports PSNR over the requested target bands using the aggregation defined in Appendix A.5. The one-target point evaluates B5 alone, and the three-target point evaluates B5, B6, and B7. This sweep is distinct from the aggregate Missing and fixed-target Multi protocols in Table 3.

Spatially shuffled history. The shuffled control retains the number of history tokens and their spectral and spatial encodings but permutes predicted patch content among spatial positions within each preceding band. For a spatial permutation σ, the history token is

$$\widetilde{\mathbf{s}}_{j,u} = P_{\mathrm{hist}}(\widehat{\mathbf{x}}_{\pi(j),\sigma(u)}) + \mathbf{e}^{\mathrm{dec}}_{\pi(j),u}. \tag{35}$$

The permutation changes content-to-position correspondence, not merely token storage order. It uses only completed predictions and is applied in both training and evaluation. Where bands share the same history-position set, the same spatial permutation is used across those bands. Spatial sets that differ under random masking require permutations within their available positions; no ungenerated patch is introduced. Nontrivial permutations avoid fixed points where possible, and singleton sets remain unchanged. Evaluation permutations are reproducible with fixed random seeds.

Per-target generation gains. Panel (b) fixes the six-target setting $(\mathcal{O}_6, \mathcal{T}_6)$ throughout evaluation. Step k corresponds to target band b_k in the ordered sequence above. Let $P_{v,k}$ denote the test-set mean PSNR for configuration v on that target. The plotted gain is

$$\Delta_{v,k} = P_{v,k} - P_{\text{obs-only},k}. \tag{36}$$

Each point compares the same target band and evaluation images against observation-only conditioning. The gains are computed separately for individual targets and are not accumulated across steps. The initial observation set remains $\mathcal{O}_6$, while available history grows as preceding targets are generated. Different target wavelengths can have different generation difficulty. Because configurations are trained separately, their first-step scores need not agree even when no preceding history is available.

C.3 Observation visibility control

For target $\pi(k)$, the truncated setting retains only original observations with wavelengths at or below the target:

$$\mathcal{O}_k = \{c \in \mathcal{O} : \lambda_c \leq \lambda_{\pi(k)}\}. \tag{37}$$

Higher-wavelength observations are removed before the bidirectional encoder and cannot contribute to its features or Full-context states. These states are rebuilt using $\mathcal{O}_k$ for each step, while history contains predictions completed at preceding steps. This control is applied at inference to the same pretrained checkpoint. This differs from the default model, which encodes all observations and computes Full states once. The comparison tests access to observed evidence and does not impose equal inference computation.

C.4 Gradient-path control for predicted history

Table A13 complements the history-content comparisons in Figure 7 by separating forward content reuse from gradients through earlier predictions. We compare observation-only, detached-history, and differentiable-history configurations under matched training settings.

Configurations. All three configurations use the same encoder architecture and a two-Full/two-AR decoder. The observation-only control retains observation attention and target-specific initialization but receives neither predicted nor ground-truth history. The other two configurations differ only in whether gradients pass through the predicted patch values entering history. For a preceding target position $(\pi(j), u)$, define

$$\begin{aligned} \widehat{\rho}_{j,u}^{\text{full}} &= \rho_\theta(\widehat{\mathbf{x}}_{\pi(j),u}, \mathbf{d}_{\pi(j)}, u), \\ \widehat{\rho}_{j,u}^{\text{detach}} &= \rho_\theta(\text{stopgrad}[\widehat{\mathbf{x}}_{\pi(j),u}], \mathbf{d}_{\pi(j)}, u). \end{aligned} \tag{38}$$

Detachment applies only to predicted content entering the history encoder. The history projection and descriptor embeddings remain trainable, and each undetached prediction receives its own reconstruction loss. Gradients through observation memory, Full-context states, and other loss paths are retained. Both predicted-history configurations use Equation 6 with the same ground-truth mixing schedule, noise, and detached ground-truth branch.

Matched training and evaluation. Each configuration is pretrained once, with matched data, initialization seeds, masking, target order, descriptors, loss weights, optimizer, and update budget. For fixed weights, inputs, and random draws, detached and differentiable history yield identical forward values and differ only in their training gradients.

For understanding, each encoder is frozen and identical task heads use matched settings and random seeds. Learning rates are selected on validation data; GB mIoU averages the corresponding test results over six GEO-Bench tasks.

Generation uses each pretrained model without ground-truth history or further optimization. Multi and Few follow Table 3, with PSNR averaged over test images and Few averaged over sensors. These averages do not quantify variation across independent pretraining runs.

Table A13: Control for gradients through predicted history. GB is six-task GEO-Bench mean mIoU (%); recovery metrics are PSNR (dB). Few averages the S2 and L8 Few scores. History gradient denotes the route from later losses through earlier predicted content.

Configuration	History gradient	GB mIoU ↑	S2 Multi ↑	L8 Multi ↑	Few ↑
Observation-only	–	61.90	33.7	33.0	31.0
Predicted history, detached	Stopped	<u>62.84</u>	<u>34.5</u>	<u>33.8</u>	<u>31.8</u>
Predicted history, differentiable	Retained	**63.15**	**34.9**	**34.1**	**32.2**

Results and interpretation. Detached predicted history improves understanding and all reported generation settings over observation-only conditioning. Retaining history gradients yields further gains, including 0.31 mIoU points and 0.4 dB on S2 Multi. The matched history construction supports a contribution from gradient flow beyond forward content reuse in this setting.

D FURTHER DISCUSSION

What is shared. Both tasks share an encoder pretrained without downstream understanding labels. Understanding uses a trained task head; generation uses the pretrained decoder. The history-gradient control supports differentiable prediction reuse for both capabilities.

Radiometry and sensor generalization. Recovery metrics and spectral-curvature regularization use standardized tensors. Standard evaluation uses target ground truth only for scoring, without prediction calibration. Spectral-angle and application-index evaluations on an appropriate radiometric scale could further assess physical fidelity. Band descriptors encode center wavelength and bandwidth; both evaluated sensors, S2 and L8, occur in pretraining. Full spectral-response modeling and transfer to unseen sensors remain future work.

Empirical scope. Evaluation covers six GEO-Bench segmentation tasks, eight additional transfer benchmarks (including three optical Copernicus-Bench tasks), and S2/L8 generation. Broader Copernicus-Bench coverage and additional modalities remain to be explored. Random spatial–spectral masking and whole-band removal yield different observation patterns; mixed masking objectives could help align pretraining with generation.

Practical use and broader impacts. Generated bands should remain distinguishable from measurements and undergo independent radiometric and task-level validation for agricultural monitoring or disaster assessment. The deterministic model returns a single estimate; uncertainty estimation could characterize ambiguity under sparse observations.

D.1 COMPUTATION AND SCALING

For dense attention (Section 3.2), let N_v, K, and P count observation tokens, requested bands, and target patches per band, respectively. Let d be the decoder width and L_F, L_A the Full and AR block counts. Equal patch counts describe whole-band recovery; random masking may yield different counts. Excluding encoder, projection, and feed-forward costs, attention over $N_v + KP$ tokens gives

$$C_{\text{full}} = O\big(L_F(N_v + KP)^2 d\big)\,. \tag{39}$$

At step k, P queries attend to N_v observation tokens and $(k-1)P$ history tokens, giving

$$\begin{aligned} C_{\text{AR}} &= O\left(L_A d \sum_{k=1}^{K} P[N_v + (k-1)P]\right) \\ &= O\left(L_A d \left[KPN_v + \frac{K(K-1)}{2}P^2\right]\right). \end{aligned} \tag{40}$$

Observation features and Full states are computed once; projection caching could reduce repeated work. History requires $O(KPd)$ feature storage, and training without gradient truncation retains activations across steps. Attention workspace depends on the kernel and checkpointing strategy.

Generation requires K dependent steps, with parallel computation across spatial positions within each band. Equal block depth does not imply equal computation. Deployment comparisons require latency, throughput, and peak-memory measurements across resolutions and band counts. Grouped generation or bounded history could reduce cost but would change the dependency structure.