A. Societal Impact

This research stands to make significant positive contributions, particularly in enhancing global food security and supporting sustainable agricultural practices. The developed super-resolution technology specifically addresses the challenge of monitoring agricultural land with high accuracy and timeliness, which is crucial for precision agriculture, improved yield estimation, and efficient resource management. This is especially impactful for smallholder farms, which characterize a vast portion of agricultural landscapes in the developing world and are essential to the livelihoods of billions. By enabling cost-effective and timely in-season monitoring, the technology can provide actionable insights to improve yields and mitigate losses, directly contributing to the economic well-being of these farmers and bolstering local and global food supplies.

B. Additional Experimental Results

B.1. Dataset Collection

The India Dataset is a proprietary collection of images acquired over the period 2019-2022, covering varied agricultural topographies within India. We follow established precedent from works such as[43], [51] in sourcing VHR imagery from Google Maps Satellite View. We take images sampled at 0.5 m GSD and from the following two satellite sources: Maxar Worldview, Airbus Pleiades.

Individual images represent a spatial extent of 320*320 m^2. Annotation was conducted by trained human personnel according to established protocols. An 'agricultural field' was operationally defined as a contiguous area meeting at least one of the following criteria: (1) cultivation of a single crop type, or (2) distinct visual boundaries separating it from neighboring parcels or non-agricultural features. The Vietnam Dataset was constructed using publicly available labels provided by [43]. As documented in [43], the high-resolution (HR) images at a given time t, denoted as X_t were originally collected for August 2021. However, our preliminary analysis revealed inconsistencies between these provided labels and the corresponding X_t imagery. Specifically, in some instances, reference images X_r predating X_t by six months or more exhibited greater concordance with the ground truth labels than the August 2021 images themselves. To ensure the integrity of our evaluation, the test set was subsequently filtered. This process retained only those image-label pairs where the August 2021 image (X_t) demonstrated clear and strong agreement with its associated ground truth labels, thereby excluding cases where labels appeared to reflect a significantly different temporal state. Table 5 summarises the properties of the datasets used.

B.2. Evaluation Metrics

We use standard pixel-level metrics to evaluate the performance of our binary classification task. Let TP_i, TN_i, FP_i, FN_i denote the number of true positive, true negative, false positive, false negative pixels respectively in a single test image i. For a test set with N images the semantic IoU $IoU_S = \frac{1}{N}\sum_1^N \frac{TP_i}{TP_i+FP_i+FN_i}$. At the instance level, we use the definition recommended for evaluating agricultural field instances (e.g., [12]), viz., Instance IOU $IoU_I = \frac{1}{M}\sum_1^M (P_m^i \cap P^i)/(P_m^i \cup P^i)$, where M is the number of ground truth instances in the test set, P_m^i is the total number of pixels in all the predicted i^{th} instances which have more than $t\%$ overlap with the i^{th} ground truth instance and P^i is the total number of pixels in the ground truth i^{th} instance. Accuracy, Precision, Recall and F1 score are computed from TP_i, TN_i, FP_i, FN_i and averaged over the test set.

B.3. Performance of Teacher and Human Labels

The Teacher model predictions i.e. the predictions from the HR-FM on the HR images yield an $mIoU_S$ of 74.18 % and an $mIoU_I$ of 71.17 % for the India Dataset and $mIoU_S$ of 68.47 % and an $mIoU_I$ of 63.29 % for the Vietnam subset of the AI4SmallFarms Dataset.

B.4. Ablations with Inference Timesteps

Experimental results suggested that stopping the reverse diffusion process early produced better results. Table 6 displays the results for both India dataset and Vietnam dataset. Figure 5 shows some qualitative examples from the last 100 denoising steps during inference time.

B.5. Qualitative results: Additional Examples

To provide a more nuanced understanding of our model's performance characteristics, Figure 6 presents a selection of additional qualitative examples. These examples highlight the efficacy of our proposed method under certain conditions and its failure modes.

- **Demonstrating Robustness to New Structures** Examples a,b showcase instances where our model successfully identifies and segments newly developed structures or significant emergent changes within the scene. This demonstrates our model's capability to detect objects or alterations not present in the reference image but from the LR inputs.
- **Superiority of Reference-Based Methods in Specific Contexts** In contrast, c,f,g illustrate specific scenarios where Ref-SR methods(RefDiff[10] and SEED-SR(Ours) exhibit superior performance compared to non-reference-based approaches. These cases typically involve subtle textural changes.
- **Failure Mode:Logit Inconsistency** Example d reveals a specific failure mode observed in our model. In this in-

Table 5. Summary of Input Data Characteristics

Category	Item	Detail / Specification
LR Inputs	Satellite Sources	Sentinel-2 L1C, Landsat 8/9, and Sentinel-1 GRD
	Spatial Resolution	**10**m GSD (Resampled)
	Temporal Span	**4** weeks before the target time (t)
	Embedding Dimension from E_{LFM}	(128,128,64) $->$ (32,32,64) [After Cropping]
	Per Embedding Spatial Coverage	$320 \times 320m^2$
HR Inputs	Spatial Resolution	**0.5**m GSD
	Embedding Dimension from E_{HFM}	(120,120,3840)
	Per Embedding Spatial Coverage	$320 \times 320m^2$
	India Dataset Collection Time	2019–2022
	AI4SmallFarms Data Collection Time	August 2021 [43]

Table 6. Comparision of Performance on early stopping during Inference time for India and Vietnam Dataset. A: Accuracy, P: Precision, R: Recall

	India Dataset											
	Teacher Predictions						Human Labels					
Method	$mIoU_I$	$mIoU_S$	A	P	R	$F1$	$mIoU_I$	$mIoU_S$	A	P	R	$F1$
$T_{Inf,500}$	49.16	72.44	84.89	60.49	61.27	58.67	48.21	70.71	83.48	59.80	57.28	56.58
$T_{inf,470}$	58.44	74.96	86.33	62.11	64.68	61.51	56.61	71.06	83.55	60.94	59.38	58.34
	Vietnam Dataset (AI4SmallFarms)											
Method	$mIoU_I$	$mIoU_S$	A	P	R	$F1$	$mIoU_I$	$mIoU_S$	A	P	R	$F1$
$T_{inf,500}$	49.14	72.30	82.10	76.68	77.92	75.92	48.77	68.20	79.11	72.77	78.30	72.59
$T_{inf,470}$	50.91	71.01	81.34	76.76	74.97	74.43	49.49	67.61	78.72	72.94	75.65	71.89

stance, the introduction of noisy artifacts or erroneous detections in the output change map can be attributed to significant disagreement among the four distinct segmentation logits generated internally by our model prior to taking an average.

- **Comparative Advantage: Reduced Noise/Oversegmentation** Finally, e provides a comparative example where other existing methods exhibit common failure modes such as oversegmentation of changed regions or the introduction of spurious noisy artifacts in their predictions.

B.6. Evaluation of Segmentation Performance using SAM for Baseline Methods

We present results in Figure 8 for zero-shot performance of SAM[25] on the superresolved images generated by baseline methods and provide a qualitative comparison against segmentation results produced by HR-FM and SEED-SR's superresolved segmentation output. Visual results clearly demonstrate that the segmentation from HR-FM and SEED-SR are considerably more accurate and coherent for agricultural fields compared to zero-shot SAM, which largely struggles with delineating farm boundaries.

B.7. Foundation Model Dependency and Failure Modes

SEED-SR's performance is coupled with the underlying FMs, a limitation we discuss in the conclusions. To further investigate this dependency, we isolated a subset of the test data where the per example mIoU of the teacher predictions (HR-FM's performance on ground truth HR Images) against the human labels is poor (instance IoU $< 60\%$). This analysis confirms that SEED-SR's performance is bounded by the capabilities of the HR-FM. On the subset where the HR-FM performs poorly ($<$60% IoU), HR-FM achieves 52.02 IoU, and SEED-SR achieves 37.56 IoU, as seen in Table 7. This expected behavior validates that the model successfully learns the mapping to the target latent space, but cannot correct for errors inherent in that space. Visual inspection shows that most of these images correspond to particularly difficult regions with few fields, or extremely small fields or dry/hilly terrains. As expected, the best baseline method, RefDiff also performs poorly at $mIoU_I = 25.44$ in this set.

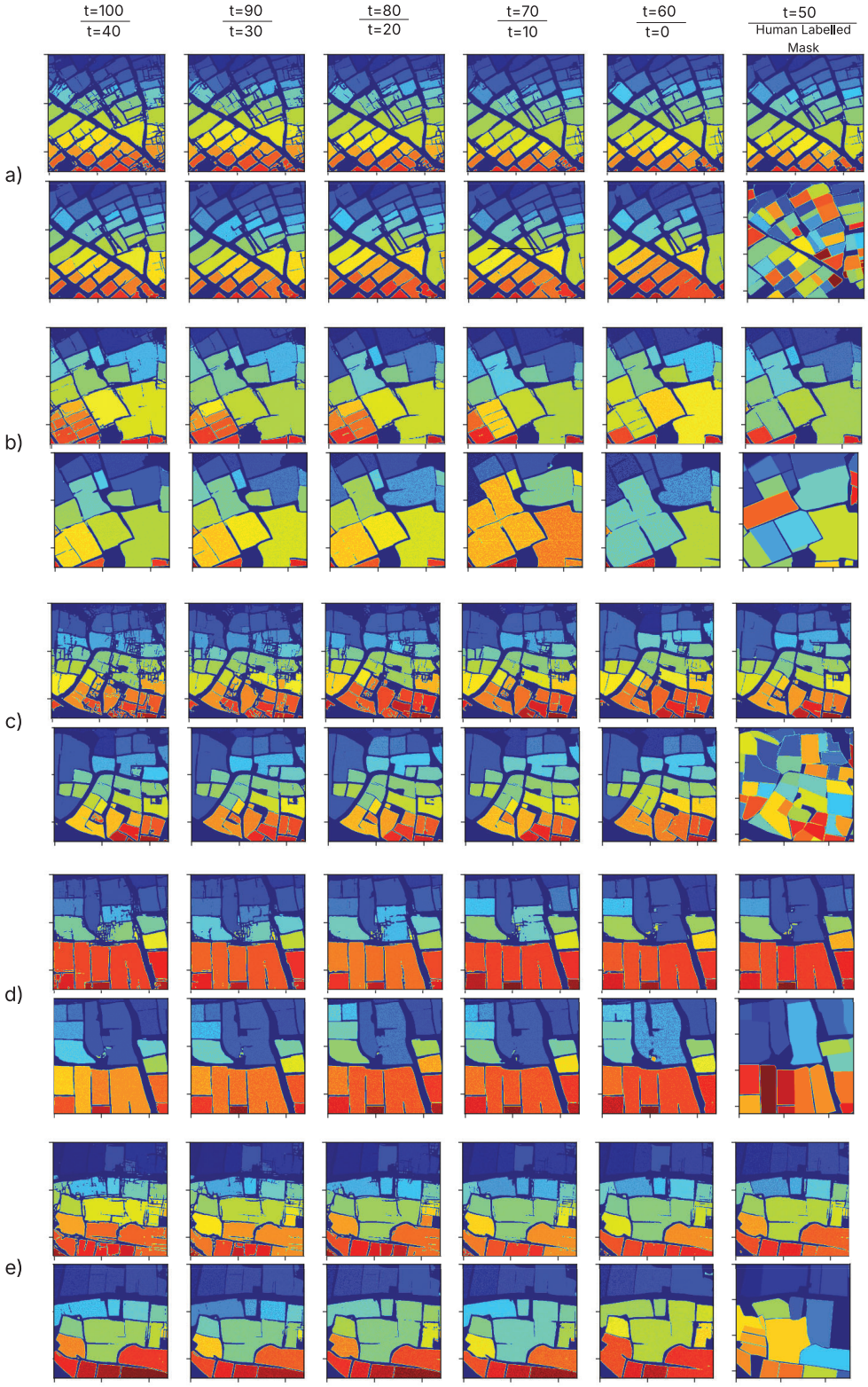

Figure 5. Qualitative results of last 100 denoising steps during inference for India Dataset.

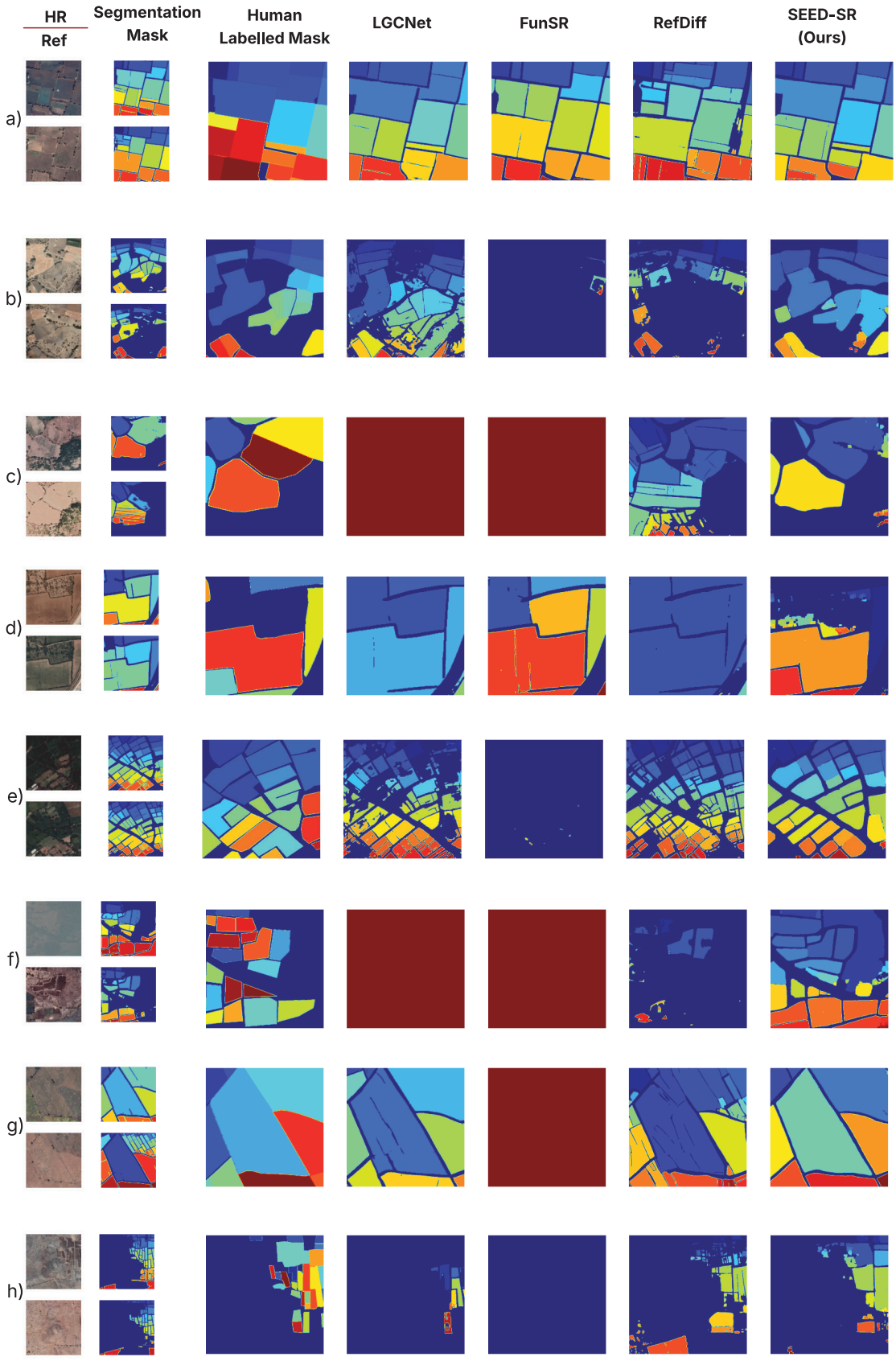

Figure 6. Additional Qualitative Examples from the India Dataset for Visual Comparison

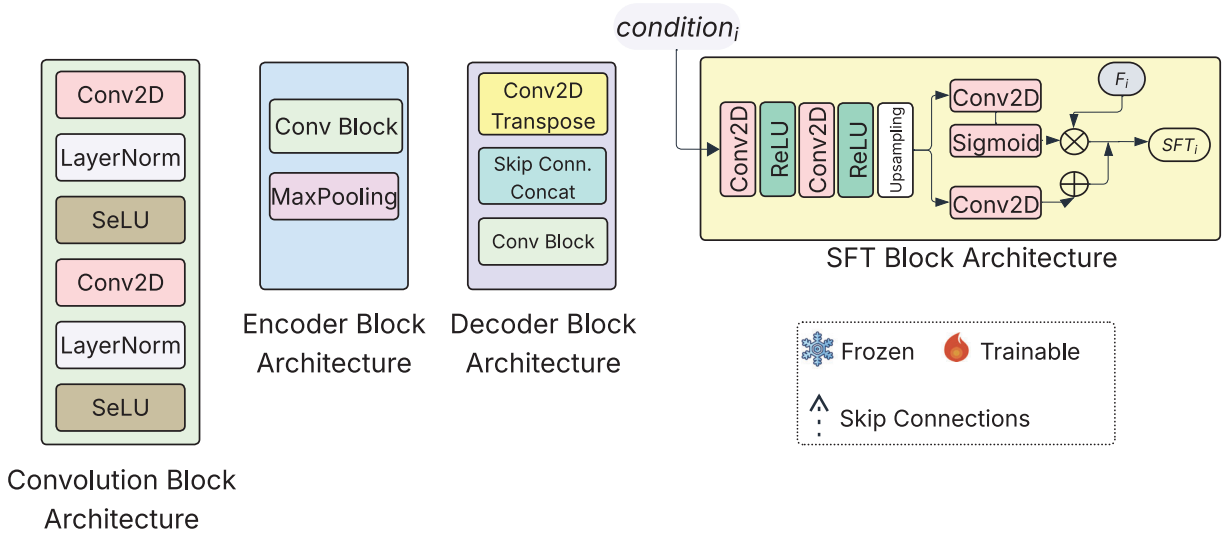

Figure 7. Architecture of ConvBlock, Encoder Block, Decoder Block and SFT as used in Figure 2

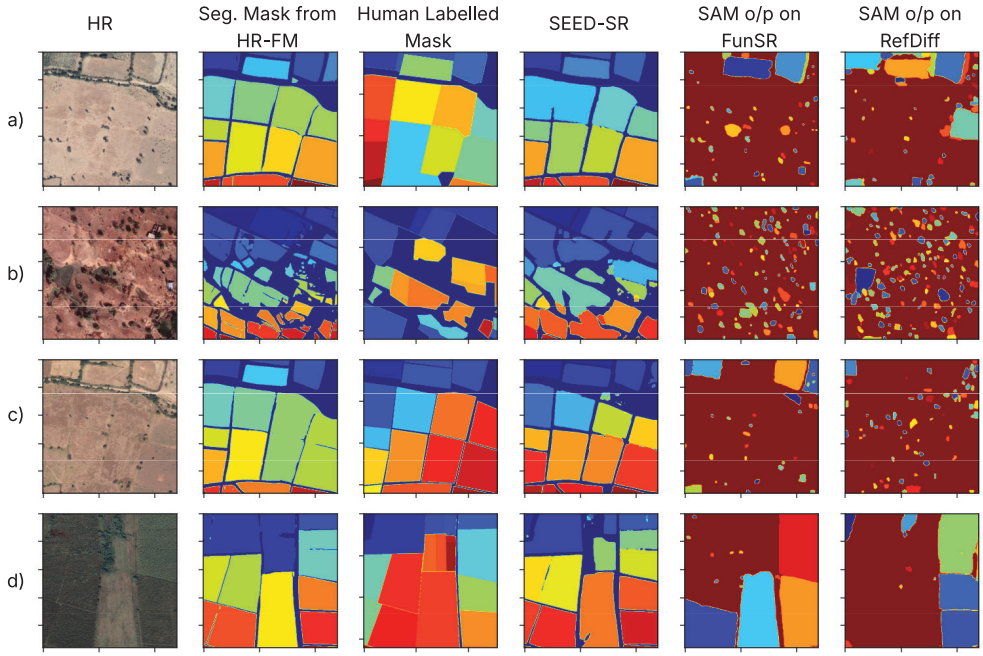

Figure 8. Zero Shot Instance Segmentation Outputs by SAM on Image Superresolution Baselines (RefDiff[10] and FunSR[7]) on the India Dataset

B.8. Computational Cost

Our model runs for 200,000 training steps. Each training step for a batch size of 128 takes approximately 1 second over 8 Google TPUv5(s). The complete training framework took 58.3 hours. The complete inference framework along with metrics computation took approximately 3 hours. The inference step takes 2.64 seconds per denoising step for a batch size of 320.

Table 8 shows the average inference time per example at maximum batch size for SEED-SR and three base-

Method	$mIoU_I$ (vs. Human Labels)
HR-FM (Teacher prediction)	52.02
SEED-SR Prediction	37.56
RefDiff Prediction	25.44

Table 7. Performance on subset of the test data where the per example $mIoU_I$ of the teacher predictions(HR-FM's performance on ground truth) against the human labels is poor ($mIoU_I < 60\%$)

lines. While diffusion methods (SEED-SR, RefDiff) are slower than non-generative methods (LGCNet, FunSR), the substantial performance gain justifies this trade-off for high-accuracy applications. Notably, SEED-SR (3.968s) is slightly faster than the comparable SOTA diffusion baseline, RefDiff (4.12s). The average inference time per example at maximum batch size was calculated at maximum throughput to take advantage of the hardware's parallelization capabilities across models.

Table 8. Model Inference Time Comparison

Method	Average Inference time per Example
SEED-SR(Ours)	3.968 s
RefDiff [10]	4.12 s
FunSR [7]	0.7 s
LGCNet [27]	0.64 s

Table 9. Sensitivity to number of masks

No. of Masks	Teacher Predictions ($mIoU_I$)	Human Labels ($mIoU_I$)
1	49.9693	49.7525
2	55.2753	53.8581
3	55.5124	54.6811
4	**57.7448**	**56.0503**
5	57.7419	56.0247
6	57.6382	56.4666
7	57.8608	56.5775
8	58.3667	56.5203

to downsample the feature map spatially (to 7x7) before applying cross-attention at the bottleneck (L207-217); at extremely high dimensions. This innovation reduced memory usage from 14 GB to 2.75 MB and FLOPs from 423 GFLOPs to 5 MFLOPs (See Table), making diffusion feasible in this feature space.

Table 10. Computational Cost Comparison justifying the design choice of DDPM architecture

Metric	Case A (7×7)	Case B (120×120)
Tokens per sample	49	14,400
Attention memory	∼0.15 MB	$\sim 13.2 \times 10^3$ MB (float32)
Total memory	∼2.75 MB	$\sim 14 \times 10^3$ MB
FLOPs	∼5 MFLOPs	$\sim 423 \times 10^3$ MFLOPs

B.9. Choice of different number of masks for the production of the final segmentation map

Diffusion models are heavily reliant on the random seed for the variability in examples and often a mean of predictions is taken to reduce stochastic variance. The masks were computed for the validation set of the India dataset(291 images). Table 9 shows the performance for different number of masks. The masks were randomly chosen for each example. The results show clear gains up to N=4, with diminishing returns thereafter. We selected N=4 as the optimal balance between performance ($\approx 2.3\ mIoU_I$ gain over N=2) and inference cost.

C. Additional Model Details

C.1. Novelty in DDPM Architecture

Our contribution is the novel integration of two distinct geospatial FMs via a latent diffusion model for 20× task-specific SR. This is non-trivial due to the high-dimensional embeddings (120×120×3840) . The novelty in our DDPM architecture Sec. 4.3 includes the integration strategy for multi-modal conditioning (LR and Ref) using self-attention on concatenated inputs and introducing a convolution block

C.2. Experimental Details

SEED-SR was trained using the AdamW optimizer [33] with a learning rate 5×10^{-4}, a batch size of 128, for a total of $200,000$ steps. A total of $T = 500$ timesteps were employed for the forward process and $T = 470$ steps were employed for the reverse diffusion processes. The noise variance schedule for the diffusion process was linearly interpolated from $\beta_1 = 10^{-4}$ to $\beta_T = 0.02$. The training was done on Google's TPUv5.

C.3. Diffusion Process

C.3.1. Forward Processs

The forward process gradually adds Gaussian noise to the HR embedding e_h over T discrete timesteps. Given $e_{h,0}$ the noised embedding $e_{h,\tau}$ at timestep $\tau \epsilon [0, T]$ is obtained by sampling from the distribution:

$$q(e_{h,\tau}|e_{h,0}) = \mathcal{N}(e_{h,\tau}; \sqrt{\bar{\alpha}_{h,0}, (1-\bar{\alpha}_\tau)\mathbf{I}})$$

where

$$e_{h,\tau} = \sqrt{\bar{\alpha}_\tau} e_{h,0} + \sqrt{1-\bar{\alpha}_\tau}\varepsilon, for \varepsilon \approx N(0, I)$$

The variables $\alpha_\tau = 1 - \beta_\tau$ and $\bar{\alpha}_\tau = \prod_{i=1}^{\tau} \alpha_i$, where ${\beta_\tau}^T{}_{\tau=1}$ is a predefined noise variance schedule.

C.3.2. Reverse Denoising Process

The reverse process learns to recover the original clean embedding $e_{h,0}$ from its noised version $e_{h,\tau}$ conditioned on e_l, e_r and the timestep τ. This is achieved by training a neural network $G_\theta(e_{h,\tau}, \tau, c)$ to predict an estimate of the clean embedding, $\hat{e}_{h,0} = G_\theta(e_{h,\tau}, \tau, c)$, where c=condition$(e_l, e_r)$ represents the combined conditioning information. The model is trained by minimizing the mean squared error between the true clean embedding and its prediction:

$$L_{\text{diffusion}} = \mathbb{E}_{\tau, e_{h,0}, \varepsilon} \left[||e_{h,0} - G_\theta(e_{h,\tau}, \tau, c)||_2^2 \right]$$

The sampling process proceeds as below:

We begin by sampling $e_{h,\tau}$ from a standard isotropic Gaussian distribution,

$$e_{h,\tau} \sim \mathcal{N}(0, \mathcal{I})$$

For each timestep τ from T down to 1, to obtain $e_{h,\tau-1}$ from $e_{h,\tau}$ we estimate $\hat{e}_{h,0}$ from the model G_θ,

$$\hat{e}_{h,0} = G_\theta(e_{h,\tau}, \tau, c)$$

Using the predicted $\hat{e}_{h,0}$ and the current $e_{h,\tau}$, the mean of the posterior distribution $q(e_{h,\tau-1}|e_{h,\tau}, \hat{e}_{h,0})$ denoted by $\tilde{\mu}_t(e_{h,\tau}, \hat{e}_{h,0})$ is calculated as:

$$\tilde{\mu}_t(e_{h,\tau}, e_{h,0}) = \frac{\sqrt{\bar{\alpha}_{\tau-1}}\beta_\tau}{1-\bar{\alpha}_\tau} \hat{e}_{h,0} + \frac{\sqrt{\alpha_\tau}(1-\bar{\alpha}_{\tau-1})}{1-\bar{\alpha}_\tau} e_{h,\tau}$$

The sample for the previous timestep is then drawn by adding scaled Gaussian noise to the posterior mean:

$$e_{h,\tau-1} = \tilde{\mu}_t(e_{h,\tau}, e_{h,0}) + \sigma_t z$$

where $z \sim \mathcal{N}(0, \mathcal{I})$ if $\tau > 1$ and z=0 if $\tau = 1$. The choice of noise variance $\sigma_\tau^2 = \beta_\tau$ follows from [19]. So the sampling step becomes,

$$e_{h,\tau-1} = \tilde{\mu}_t(e_{h,\tau}, e_{h,0}) + \sqrt{\beta_\tau} z$$

Thus the $\hat{e}_{h,0}^{final}$ is the generated high-resolution embedding, which is then passed to the frozen decoder D_{HSM} to produce the final segmentation map.

C.3.3. Spatially Feature Transform (SFT) blocks

To specifically introduce and leverage the distinct characteristics of the low-resolution input (e_l) and the reference input (e_r), we employ Spatially Feature Transform (SFT) blocks at different stages within the decoder. Figure 7 shows a detailed architecture of the SFT Block, Convolution Block, Encoder Block and Decoder Block as shown in Figure 2. Architecturally, each SFT block first processes its conditioning input through a pair of 3x3 convolutional layers with ReLU activations (transforming channels to 128, then 64).

Let c_{in} be the input conditioning in the SFT block, the convolutional block followed by ReLU produces,

$$C_1 = \text{ReLU}(\text{Conv}_{128,3\times3}(c_{in}))$$

Progressively we apply another convolutional block to reduce the channels to 64,

$$C_2 = \text{ReLU}(\text{Conv}_{64,3\times3}(C_1))$$

The output C_2 is then bilinearly upsampled to match the spatial dimensions (H,W) of the main feature map F that it will modulate, resulting in C_{up}. Subsequently, from the upsampled conditioning features C_{up}, two separate 3x3 convolutional branches, are used to predict spatial-specific modulation parameters: a scale factor (γ) and a shift factor (β). Both γ and β will have the same number of channels (n_F) as the main feature map F.

$$\gamma = \sigma(\text{Conv}_{n_F,3\times3}(C_{up}))$$

$$\beta = \text{Conv}_{C_F,3\times3}(C_{up})$$

Finally, the learned scale (γ) and shift (β) parameters are applied to the main feature map F using an element-wise affine transformation. The feature map F is first scaled by γ and then the shift β is added such that,

$$F_{out} = \gamma \odot F + \beta$$

C.4. Layerwise Dimensions in Our Model

Table 5 displays the layerwise dimensions of SEED-SR's denoising diffusion model.

C.5. Changes in Ablation studies

Table 6 specifies the layerwise changes in model dimensions because of the different inputs of ablation studies.

Table 11. Layerwise Dimensions of our Model

Layer	Dimension
Input Noised HR	(120,120,3840)
Input Ref	(120,120,3840)
Input LR	(32,32,64)
Input Timestep	(1,)
Usampled LR	(120,120,64)
Concatenated Encoder Input	(120,120,7744)
Encoder Block 1	(60,60,320)
Encoder Block 2	(30,30,480)
Encoder Block 3	(15,15,560)
Encoder Block 3 with Timestep Embedding	(15,15,560)
Encoder Block 4	(7,7,640)
Attention Block	(7,7,640)
Bottleneck Conv Block	(7,7,784)
Decoder Block 1	(15,15,640)
SFT Block 1	(15,15,640)
Decoder Block 2	(30,30,560)
Decoder Block 3	(60,60,480)
SFT Block 2	(60,60,480)
Decoder Block 3 with SFT and Timestep Embedding	(60,60,480)
Decoder Block 4	(120,120,320)
Attention Block	(120,120,320)
Convolution Block	(120,120,3840)

Table 12. Layerwise Input Dimension Changes for Ablation Studies

(a) Config A: S2 + Ref, no embeddings

Layer	Dimension
Input Noised HR	(480,480,3)
Input Ref	(480,480,3)
Input LR	(32,80,80,12)
Input Timestep	(1,)
Reshaped and Upsampled LR	(480,480,384)
Concatenated Encoder Input	(480,480,390)
⋮	⋮
Convolution Block	(480,480,3)

(b) Config B: Only S2 images, no embeddings

Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	(32,80,80,12)
Input Ref	NA
Input Timestep	(1,)
Upsampled LR	(120,120,384)
Concatenated Encoder Input	(120,120,4224)
⋮	⋮
Convolution Block	(120,120,3840)

(a) Config C: Only LR Embeddings

Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	(32,32,64)
Input Ref	NA
Input Timestep	(1,)
Upsampled LR	(120,120,64)
Concatenated Encoder Input	(120,120,3904)
⋮	⋮
Convolution Block	(120,120,3840)

(b) Config D: Only Ref Embeddings

Layer	Dimension
Input Noised HR	(120,120,3840)
Input LR	NA
Input Ref	(120,120,3840)
Input Timestep	(1,)
Upsampled LR	NA
Concatenated Encoder Input	(120,120,7680)
⋮	⋮
Convolution Block	(120,120,3840)

(a) Config E: Ref Embeddings with S2 Images

Layer	Dimension
Input Noised HR	(120,120,3840)
Input Ref	(120,120,3840)
Input LR	(32,80,80,12)
Input Timestep	(1,)
Reshaped and Upsampled LR	(120,120,384)
Concatenated Encoder Input	(120,120,8064)
⋮	⋮
Convolution Block	(120,120,3840)