A APPENDIX

A.1 IMPLEMENTATION DETAILS

Data Preprocessing **WorldStrat** We obtain the RGB satellite images provided by Cornebise et al. (2022). We follow the preprocessing steps given by Cornebise et al. (2022), that crop the HR images into patches of 192×192 and LR images into corresponding pathes of the HR images of size 63×63. For each HR image, there are 16 corresponding LR images of different time. We extract the RGB band of the HR images and the RGB band of every LR images. We then resize LR image into the size of 192×192. We use the same training and validation split provided by Cornebise et al. (2022) and then form a training and validation set. We then extract the timestamp from the metadata and compute the dt_i for each LR_i.

fMoW We obtain the high-resolution images from Christie et al. (2018), and the paired Sentinel-2 images from Cong et al. (2022). We first identify the area of interest on the HR images as given by Christie et al. (2018), and then crop out other areas. Then we crop the corresponding LR images following the pre-processing steps given by Cong et al. (2022). We crop the HR images into patches of 512×512, and align LR images into patches based on each HR image patch. Then we resize each LR image into the size of 512×512 in accordance with the HR image. We consider 6 categories: airport, amusement parks, car dealership, crop field, educational institution, electric substation for training and testing. For training, we filter out HR images that do not have a corresponding LR, and those do not have three channels. We take the same training and validation split from Christie et al. (2018). When training, we consider all images from Christie et al. (2018), and when testing, we pick the first 100 images from each selected category of the validation set of Christie et al. (2018). We also extract dt_i from the metadata of fmow provided by Christie et al. (2018), and the metadata of Sentinel-2 data provided by Cong et al. (2022).

Model Training We take the pretrained checkpoint (SD1.2) provided by Rombach et al. (2022), and then fine tune on the processed Christie et al. (2018) and Cornebise et al. (2022) datasets. We rescale every LR image and HR image to the scale of [0,1]. Then, we use a learning rate $1e^{-5}$, and a batch size of 4 for both datasets. We stop training when the FID of sampled images stops improving. For Cornebise et al. (2022) dataset, we only train 800 iterations (partly due to the small dataset size). For Christie et al. (2018) dataset, we train 4400 iterations. We then take the model for downstream inference tasks. For WorldStrat dataset, we use the prompt "Satellite images" for training.

Model Inferencing We use 50 DDIM steps with $\eta = 0$ for inferencing. We perform optimization every 5 steps for computational efficiency, otherwise, we just perform conditional sampling. We set $\lambda = 0.1$ and $\alpha = 0.2$ for both datasets.

CLIP zero-shot classification We use the pretrained CLIP model to compute the image embedding and text embedding. For each class of fMoW, we take the reconstructed images from each methods and compute its CLIP embedding using the ViT-B/32 model, which gives an embedding of shape (1,512). Then we use the text encoder to compute the embedding, which gives a shape of (77, 512). The prompt for each class is given by "a satellite image of {class} from an overhead view". Then we commpute the cosine similarity of the image embedding and the text embeddings, which takes the mean of the text embedding over dimension 0. Then we pick the class with the highest cosine similarity.

A.2 MORE ABLATION STUDIES

Effect of LPIPS weight α and optimization weight λ There are two hyper-parameters in our inference-time algorithm, and that is the weight for LPIPS distance α v.s. L2 distance, and the weight λ for balancing the original predicted clean image component and the one after optimization. We expect the perceptual quality of reconstructed images to improve when we increase α from 0. We also expect the reconstruction quality to improve when λ increases from 0 since the weight of fusion increases. In Fig.6, we present the LPIPS score on 100 samples on the WorldStrat dataset with varying α and λ. We find that LPIPS score improves when both α and λ increase from 0. Then, performance converges as α keeps increasing, and marginally degrades as λ continues to increase. We observe that generally, the performance of our algorithm is insensitive to hyperparameter change.

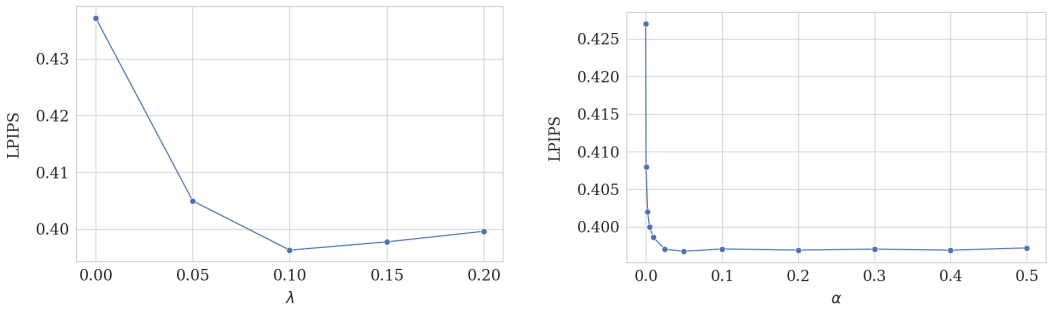

Figure 6: (Left) Ablation study on optimization weight λ. (Right) Ablation study on LPIPS weight α.

A.3 Implementation Details of Baselines

WorldStrat We follow the original codebase of Cornebise et al. (2022), where we train the HighResNet model on both Christie et al. (2018) and Cornebise et al. (2022) datasets. We tune the hyperparameters of the loss function based on our validation set. We stop training when the validation performance converges. On Cornebise et al. (2022) we train 125000 iterations with a batch size of 32, and on Christie et al. (2018) we take 100000 iterations with a batch size of 32.

MSRResNet We follow the original codebase of Wang et al. (2018b), where we train the MSRResNet model on both Christie et al. (2018) and Cornebise et al. (2022) datasets. During training, we randomly pick a LR image and its paired HR image. We tune the hyperparameters of the loss function based on validation set performance. We train for 160000 iterations for both Cornebise et al. (2022) and Christie et al. (2018) datasets with a batch size of 16.

DBPN We follow the original codebase of Haris et al. (2018), where we train the MSRResNet model on both Christie et al. (2018) and Cornebise et al. (2022) datasets. During training, we randomly pick a LR image and its paired HR image. We tune the hyperparameters of the loss function based on validation set performance. We train for 100 epochs for both Cornebise et al. (2022) and Christie et al. (2018) datasets with a batch size of 16.

Pix2Pix We follow the original codebase of Isola et al. (2017), where we train the MSRResNet model on both Christie et al. (2018) and Cornebise et al. (2022) datasets. During training, we randomly pick a LR image and its paired HR image. We tune the hyperparameters of the loss function based on validation set performance. We train for 30 epochs for Christie et al. (2018) and 100 epochs for Cornebise et al. (2022) with a batch size of 16.

ControlNet We follow the original codebase of Zhang et al. (2023). During training, we randomly pick a LR image and its paired HR image. We tune the hyperparameters of the loss function based on validation set performance. We train for 12500 iterations with a batch size of 16 for Cornebise et al. (2022), and 21500 iterations with a batch size of 16 for Christie et al. (2018).

DiffusionSat We implemented the 3D ControlNet architecture as mentioned in Khanna et al. (2024). Then, we take the RGB band and the SWIR, NIR band from LR image for training 3D ControlNet. We tune the hyperparameters of the loss function based on validation set performance. We train for 12500 iterations with a batch size of 16 for Cornebise et al. (2022), and 20000 iterations with a batch size of 16 for Christie et al. (2018). We observe that further training worsened FID scores on both datasets.

A.4 More Results

We present additional results on reconstruction for paired LR and HR (taken at the same time) in Fig. 7. Notice that we can reconstruct accurate details in this setting. We report additional results on

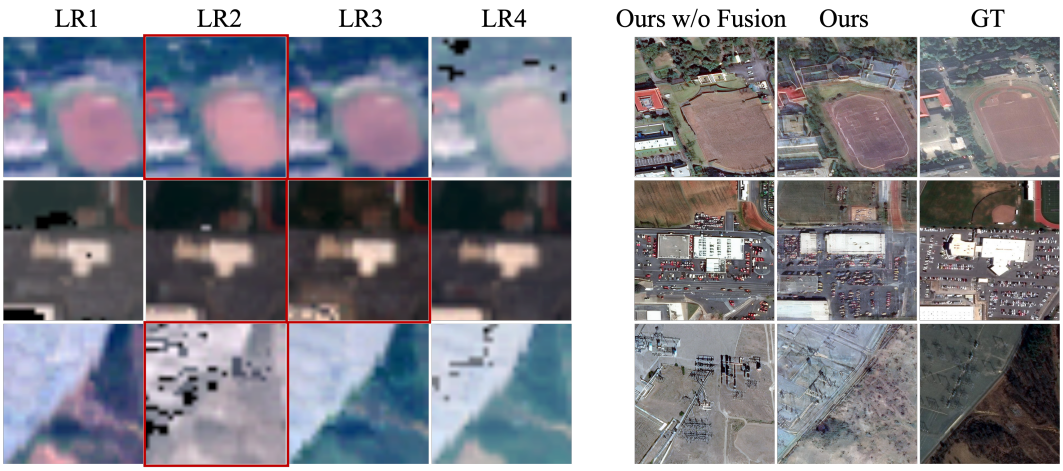

Figure 7: Super-resolution results of SatDiffMoE with paired low-resolution image as the input.

unconditional generation and conditioning on dt_i as demonstrated in Fig.8, and Fig.9. We observe that we can generate realistic satellite images. Conditioning on dt_i makes semantic changes in the image and can be applied to tasks such as cloud removal. We also report the PSNR and SSIM metrics in Table 7. The error bars are presented in Table 8.

Method	WorldStrat		fMoW	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
WorldStrat Cornebise et al. (2022)	17.98	0.396	**13.42**	**0.443**
MSRResNet Wang et al. (2018b)	**19.81**	**0.512**	<u>13.01</u>	<u>0.290</u>
DBPN Haris et al. (2018)	19.17	<u>0.471</u>	11.90	0.268
Pix2Pix Isola et al. (2017)	<u>19.76</u>	0.448	12.21	0.180
ControlNet Zhang et al. (2023)	11.89	0.113	10.82	0.117
DiffusionSat Khanna et al. (2024)	12.34	0.133	10.63	0.109
SatDiffMoE (Ours)	17.40	0.396	11.96	0.172

Table 7: Comparison of PSNR and SSIM metrics for super-resolution on WorldStrat dataset and fMoW. Best results are in bold. Second best results are underlined.

Method	WorldStrat	MSRResNet	DBPN	Pix2Pix	ControlNet	DiffusionSat	SatDiffMoE(Ours)
WorldStrat	0.081	0.077	0.079	0.069	0.079	0.091	0.076
fMoW	0.092	0.081	0.052	0.045	0.034	0.034	0.044

Table 8: Standard deviation of the quantitative metric LPIPS presented in Table 2 for super-resolution on fMoW and WorldStrat dataset.

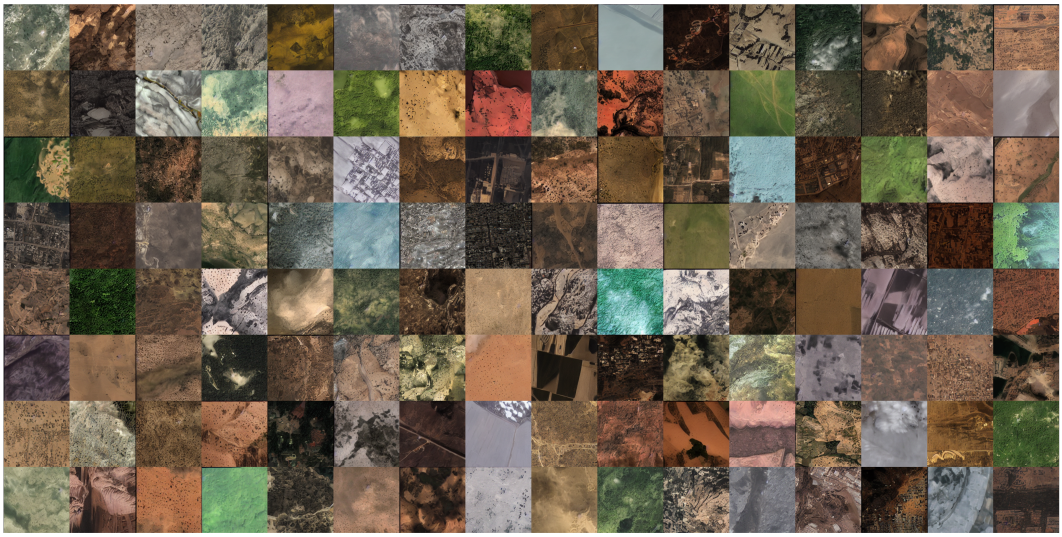

Figure 8: Generated high resolution samples from our unconditional model.

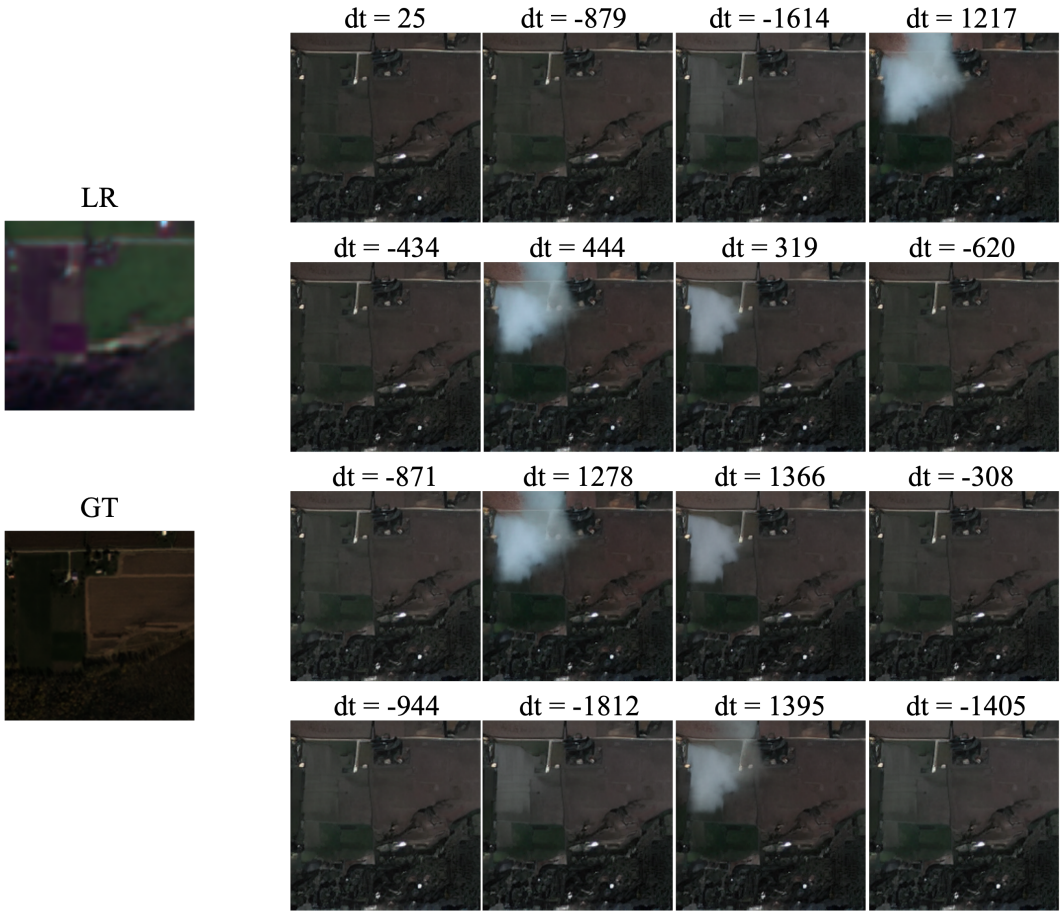

Figure 9: Here we consider the same LR image, and vary the relative time difference to generate different conditional samples.