Metadata, Wavelet, and Time Aware Diffusion Models for Satellite Image Super Resolution

Luigi Sigillo, Renato Giamba, Danilo Comminiello
Sapienza University of Rome

ICLR

Motivation

Real-world Impact. Environmental monitoring, disaster response, and agricultural management require detailed imagery unavailable at scale.

Access Limitation. High costs and sensor constraints limit frequent HR imaging. Widely available alternatives like Sentinel-2 (10–60m resolution) lack the detail needed for tasks like crop mapping or urban infrastructure analysis.

Modeling Challenges. Satellite SR is challenged by heterogeneous spatial and temporal resolutions, metadata-rich inputs, and strong environmental variability necessitating context-aware generative approaches.

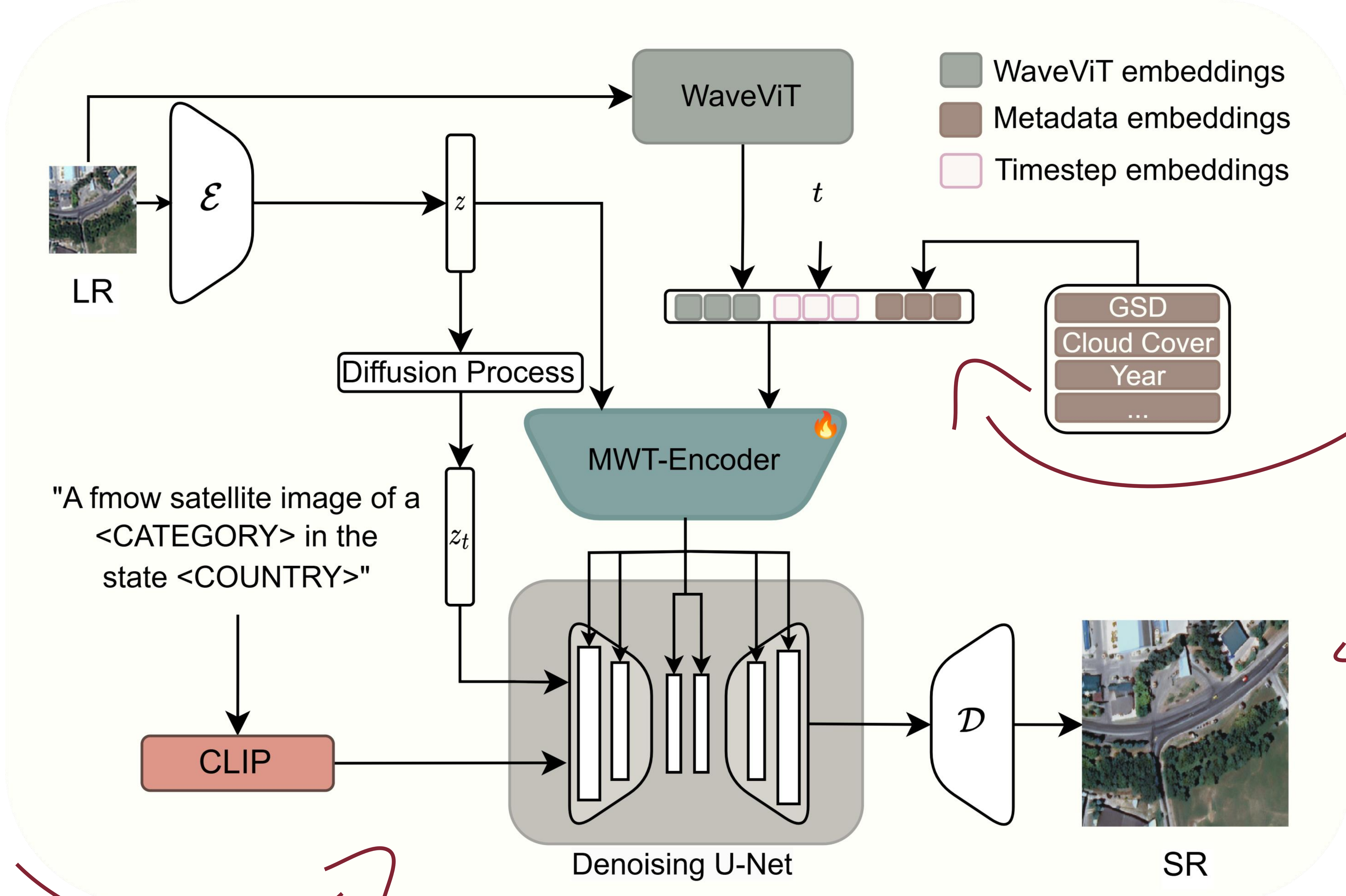
MWT-Diff at a Glance

What is MWT-Diff? Our framework is a novel approach combining **latent diffusion** models with **wavelet** and **metadata** integration to generate super-resolved satellite imagery.

Why it works? Our model **preserves critical spatial characteristics** while demonstrating significant improvements. Leverages the fusion of metadata, wavelet features, and temporal information of the **MWT-Encoder** at multiple scales.

MWT-Diff

1) Input. A low-resolution satellite image is first encoded by a pretrained VAE encoder into a latent representation z.



2) Wavelet Decomposition & Embedding. Pretrained WaveViT applies Discrete Wavelet Transform (DWT) to extract multi-frequency features (textures/edges).

3) Conditioning Fusion. MWT-Encoder combines:

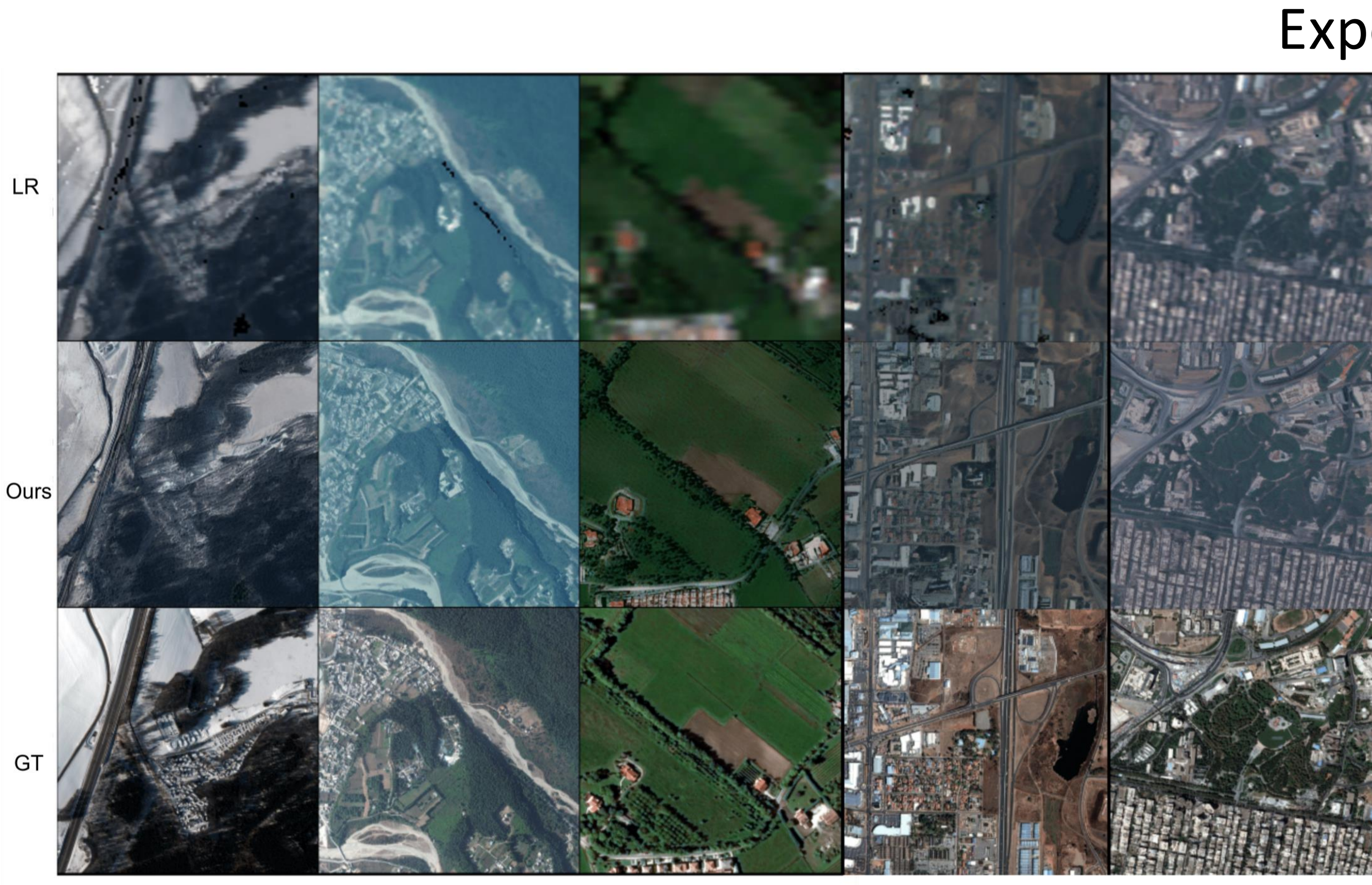
- Wavelet embeddings
- Metadata (sinusoidal encoding)
- Timestep data

→ Outputs a 3072 guidance vector.

4) Denoising Process. The denoising U-Net iteratively refines the latent space using the conditional guidance, simulating the reverse diffusion process.

5) High-Res (HR) Output. 512×512 photorealistic image, preserving geospatial details (e.g., crop boundaries, urban layouts).

Experimental Results



Comparison of the LR Sentinel-2 input, the output of the model, and the corresponding HR fMoW.

	Model	FID ↓	LPIPS ↓
fMoW	Low Resolution	114.38	0.756 +/- 0.004
	StableSR	53.85	0.345 +/- 0.002
	MWT-Diff	**53.07** (-1.44%)	**0.336** +/- 0.002 (-2.61%)
Sentinel2-fMoW	WorldStrat Cornebise et al. (2022)	426.7	0.736 ± 0.092
	MSRResNet Wang et al. (2018b)	286.5	0.783 ± 0.081
	DBPN Haris et al. (2018)	278.2	0.750 ± 0.052
	Pix2Pix Isola et al. (2017)	196.3	0.643 ± 0.045
	SatDiffMoE Luo et al. (2024)	115.6	0.606 ± 0.044
	DiffusionSat Khanna et al. (2024)	102.9	0.638 ± 0.034
	ControlNet Zhang et al. (2023)	102.3	0.644 ± 0.034
	MWT-Diff	**98.1** (-4.11%)	**0.555** ± 0.003 (-8.41%)

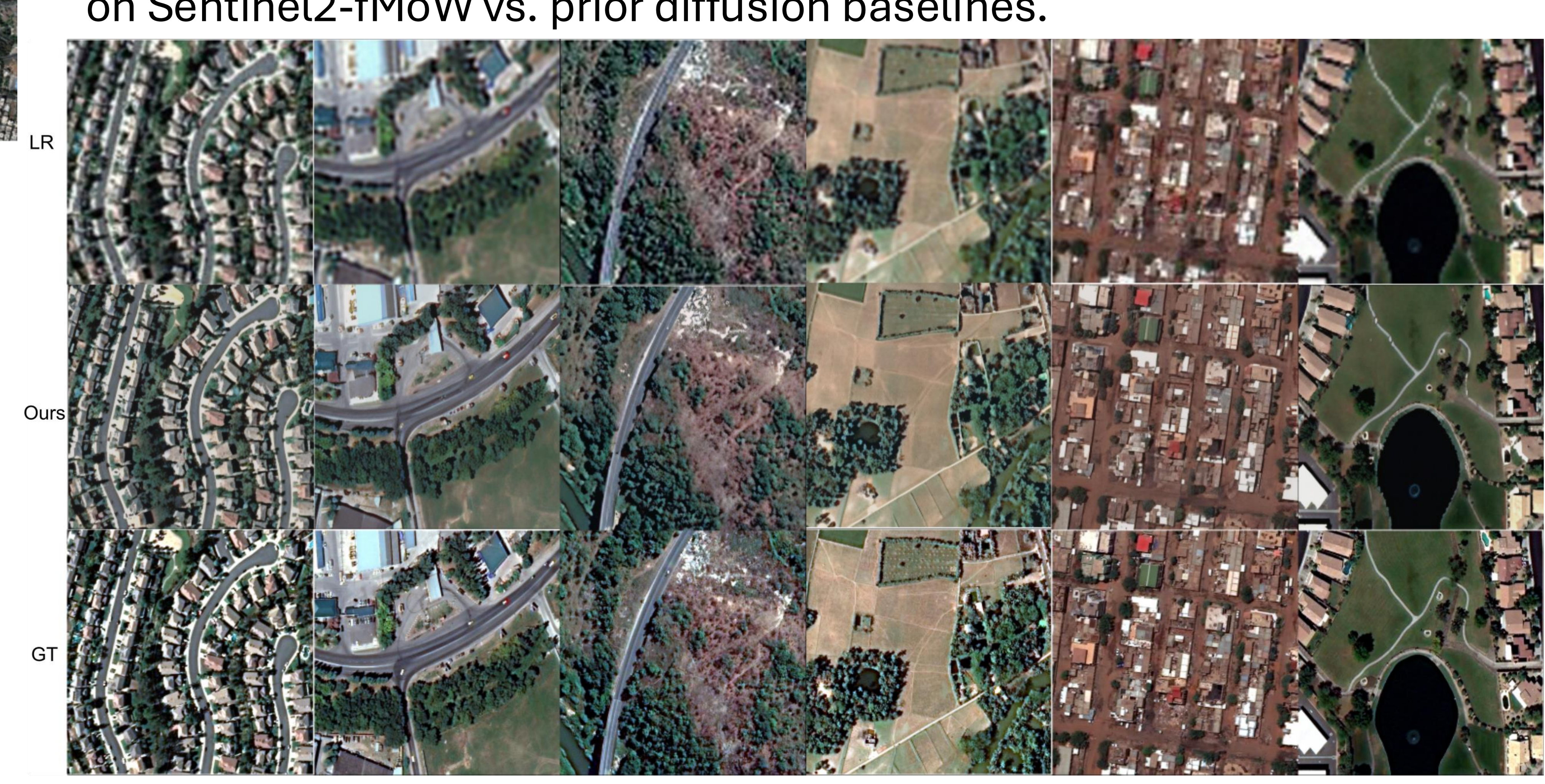
MWT-Diff advantages

Metadata-guided diffusion → Improves reconstruction fidelity by incorporating physical sensor parameters and location data from satellite observations.

Wavelet transforms (WaveViT) → Preserves textures, boundaries and high-frequency features.

Computationally efficient → Only trains the MWT-Encoder while keeping diffusion backbone frozen.

Benchmark-Leading Quality → Achieves FID ↓4.11% and LPIPS ↓8.41% on Sentinel2-fMoW vs. prior diffusion baselines.



Qualitative results on 128x128 → 512x512 with fMoW.