FlowSat: Flow-Matching Diffusion Transformers with Metadata Conditioning for Satellite Image Generation

Digvijay Singh Parihar*
digvijaysingh.parihar@iitgn.ac.in
Rishabh Mondal*
rishabh.mondal@iitgn.ac.in
Nipun Batra
nipun.batra@iitgn.ac.in

Indian Institute of Technology
Gandhinagar
Gandhinagar, Gujarat, India

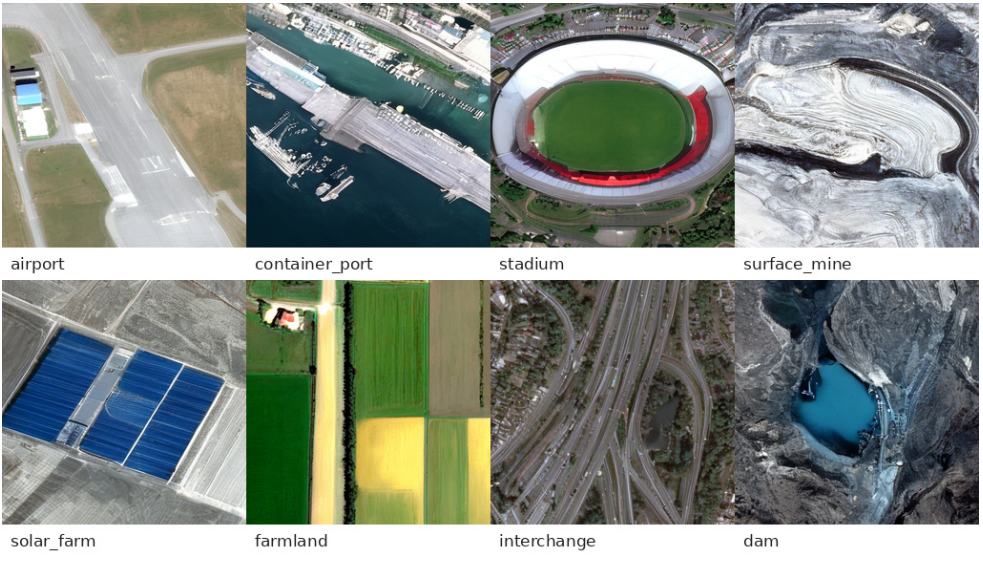


Figure 1: FlowSat samples across fMoW categories. A single model, 20 Euler steps, conditioned on a caption and held-out acquisition metadata. No per-category fine-tuning, no retrieval.

Abstract

Satellite image generation supports data augmentation, simulation, and counterfactual analysis in remote sensing, yet existing models inherit slow denoising U-Nets from natural-image diffusion and treat acquisition metadata weakly: global per-tile signals such as location, date, ground sampling distance, and cloud cover are folded into the timestep embedding without exploiting field geometry, or omitted entirely. We introduce FlowSat, a flow-matching Diffusion Transformer with structured global metadata conditioning. FlowSat fine-tunes a pretrained SANA DiT through a zero-initialised AdaLN metadata graft that preserves the backbone at initialisation, and a geometry-aware encoder representing location on the sphere, date cyclically, and sensor variables

*Equal contribution.

with multi-scale Fourier features. On fMoW-RGB, FlowSat attains the best FID and CLIP score among compared methods in only 20 Euler steps, at least 2.5× fewer than competing diffusion baselines. An encoder-controlled ablation attributes the gain to the geometry-aware encoding itself, improving FID from 46.81 to 37.96 with roughly 11× fewer conditioning parameters than a plain sinusoidal encoder. FlowSat transfers zero-shot to RSICD; ablations show zero initialisation is essential for stable fine-tuning, and controllability sweeps confirm expected seasonal and resolution changes. Code: https://github.com/sustainability-lab/flowsat-satellite-image.

1 Introduction

Generative models of satellite imagery are increasingly central to remote sensing: they augment scarce labelled data for downstream recognition, synthesise controllable scenes for simulation, and produce counterfactuals that ask how a location would appear under a different season, resolution, or sensor. What makes overhead imagery distinctive is not its appearance but its *metadata*. Every tile is tagged with a geographic coordinate, an acquisition date, a ground-sampling distance (GSD), and cloud cover; these fields are recorded for essentially all archival imagery and impose strong, physically grounded priors on what the scene should look like. A faithful generative model of satellite data must therefore be conditioned on this metadata as a first-class signal, not as an afterthought.

The dominant generative foundation model for this domain, DiffusionSat [11], established that metadata conditioning is both feasible and valuable. Yet its design predates the current state of the art in two respects. Architecturally, it fine-tunes a Stable-Diffusion U-Net trained with the denoising-diffusion (DDPM) objective and sampled with 100 DDIM steps; on natural images this recipe has since been superseded by Diffusion Transformers (DiT) [21] trained with flow matching [13, 15], which reach comparable quality in an order of magnitude fewer steps. Methodologically, metadata is embedded similarly to the diffusion timestep embedding and injected into the conditioning stream. While effective, this treatment does not explicitly account for the geometric and periodic structure of geographic, temporal, and sensor metadata. A concurrent line of work [14] modernises the backbone to a DiT but conditions only on text and spatial point queries, abandoning standard acquisition metadata entirely. To our knowledge, no existing model brings together a flow-matching DiT and principled, global metadata conditioning for the satellite domain - the combination this paper provides.

We present **FlowSat**, a flow-matching Diffusion Transformer for satellite image generation conditioned on geographic and acquisition metadata. FlowSat builds on three ideas. First, it fine-tunes a natively flow-matching SANA-0.6B backbone [33], whose deep-compression autoencoder, linear attention, and $[0, 1]$ timestep parameterisation enable high-quality generation in only 20 Euler steps. Second, metadata is injected through AdaLN-single modulation rather than spatial inputs via a *zero-initialised AdaLN graft*, preserving the pretrained backbone at initialisation and enabling metadata conditioning to emerge gradually during fine-tuning; we show that this zero initialisation is critical for stable training under a modest compute budget. Third, metadata is encoded according to its native geometry — geolocation on the sphere, temporal fields as cyclic variables, and sensor attributes across scales so that embedding distances reflect meaningful geographic and temporal relationships.

Trained on fMoW-RGB alone and fine-tuned from a public checkpoint, FlowSat attains the best FID and the best CLIP score among compared methods on the in-domain fMoW

benchmark, surpassing the strongest published baseline while using at least 2.5× fewer sampling steps, and it transfers to the out-of-domain RSICD benchmark without retraining. Controllability sweeps confirm that the trained model genuinely *uses* the metadata pathway: holding the caption and noise fixed and varying a single field reproduces the expected change in season or resolution. Our contributions are:

- **FlowSat**, the first flow-matching DiT for satellite imagery with global metadata conditioning, reaching state-of-the-art in-domain FID and CLIP at 20 sampling steps.
- A *zero-initialised AdaLN graft* that injects metadata into the modulation pathway while leaving the pretrained backbone unchanged at initialisation, with an ablation isolating the zero initialisation as the component that makes training feasible.
- A *geometry-aware metadata encoder* that respects the spherical, cyclic, and multi-scale structure of the underlying fields, isolated by an encoder-only controlled comparison against a plain sinusoidal encoder that is 11× larger (Sec. 4.3).

2 Related Work

Diffusion and flow-matching generation. Diffusion models [3, 8, 28], latent diffusion [23], and fast samplers [9, 16, 27] underpin image synthesis; DiT [21] introduced transformer backbones, PixArt-α/Σ [1] AdaLN-single conditioning. Flow matching [13], rectified flow [15], and SiT [18] enable velocity-based generation and distillation [5]; SD3's MMDiT [4] added logit-normal timestep sampling. We build on SANA [33]: linear attention [10], QK-normalisation a Gemma text encoder [31], a 32× autoencoder.

Generative models for remote sensing. DiffusionSat [11] pioneered metadata-conditioned satellite generation; CRS-Diff [30], MetaEarth [35], and GeoDiT [26] extend it, yet all keep DDPM training and omit metadata or fold it into the timestep pathway. FlowSat is the first flow-matching DiT with *structured global* metadata conditioning and a prior-preserving zero-init graft (Tab. 1).

Conditioning representations. Zero-init branches preserve pretrained behaviour [37]; location embeddings [19], spherical encoders [12, 24], and Fourier features [20, 29] inform our geometry-aware metadata encoder.

Table 1: **Positioning against prior satellite generators.** FlowSat uniquely combines a flow-matching (FM) DiT, structured geometry-aware global metadata conditioning (Struct.), and a zero-initialised pathway preserving the pretrained prior (Z-init), with the fewest sampling steps. *w/ t*: folded into the timestep embedding; *partial*: subset of standard metadata fields.

Method	Backbone	FM	Metadata	Struct.	Z-init	Steps
DiffusionSat [11]	U-Net	✗	w/ *t*	✗	✗	100
CRS-Diff [30]	U-Net	✗	✗	✗	✗	∼50
MetaEarth [35]	U-Net	✗	partial	✗	✗	∼100
GeoDiT [26]	DiT	✗	✗	✗	✗	∼50
FlowSat (ours)	DiT	✓	global	✓	✓	**20**

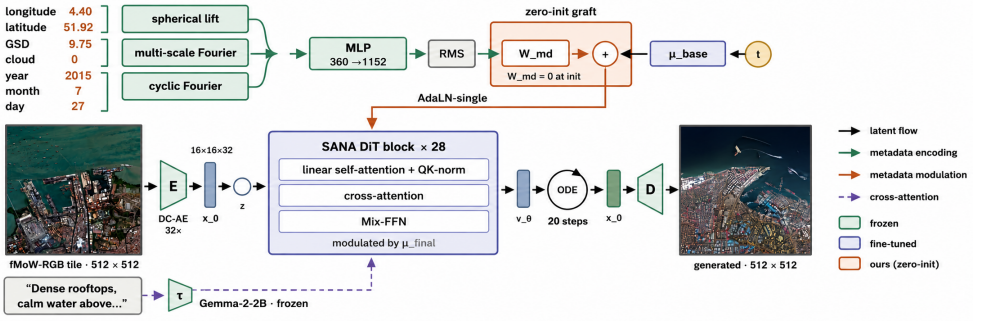


Figure 2: **FlowSat architecture.** Images are compressed into DC-AE latents and processed by a pretrained SANA-0.6B flow-matching DiT conditioned on text and metadata. Our key contribution is a zero-initialised metadata branch that injects geometry-aware metadata into the global AdaLN modulation pathway, preserving the pretrained velocity field at initialization while enabling controllable satellite image generation.

3 Method

3.1 Overview

FlowSat is a latent flow-matching DiT (Fig. 2): a frozen deep-compression autoencoder yields latents, SANA predicts the noise-to-data velocity field, text enters via cross-attention, and timestep *and metadata* via AdaLN-single.

3.2 Latent Flow Matching

We operate in the latent space of a deep-compression autoencoder (DC-AE) [33]: an image $I \in \mathbb{R}^{512\times512\times3}$ is encoded to $x_0 = \mathcal{E}(I) \in \mathbb{R}^{32\times16\times16}$, a 32× compression yielding only 256 tokens and bounding per-block cost. With $z \sim \mathcal{N}(0,\mathbf{I})$ of the same shape, we follow the rectified-flow form of flow matching (FM) [13, 15] and define the linear interpolant

$$x_t = (1-t)\,x_0 + t\,z, \qquad t \in [0,1], \tag{1}$$

with constant target velocity $v_{\text{target}} = z - x_0$, which the network v_θ regresses under conditioning c (text and metadata, Sec. 3.6):

$$\mathcal{L} = \mathbb{E}_{x_0,z,t}\left\| v_\theta(x_t,t,c) - (z-x_0)\right\|_2^2. \tag{2}$$

We draw t from a logit-normal distribution [4], concentrating supervision on the mid-trajectory regime. The backbone is initialised from the SANA checkpoint, itself flow-matching pretrained, so its $[0,1]$ timestep convention matches ours with no schedule remapping. At inference we integrate the ODE $\dot{x} = v_\theta(x,t,c)$ from t=1 (noise) to t=0 (data) with an Euler solver in N=20 steps.

3.3 Backbone

We adopt SANA-0.6B [33] as the velocity-field backbone: L=28 blocks, hidden d=1152, H=16 heads (d_h=72), ≈600M parameters. With **patch size** p=1, each position of the

16×16 latent grid is one token (N=256); PixArt-Σ [1] instead applies p=2 to an 8× VAE latent (N=1,024 at 512 px), so the 32× DC-AE compression with p=1 gives a 4× shorter sequence at no loss of spatial resolution at the transformer input. Each block applies linear self-attention [33] ($\mathcal{O}(Nd)$ vs. $\mathcal{O}(N^2d)$ for softmax), cross-attention to text, and a Mix-FFN [33] (expansion 2.5), all modulated by the shared AdaLN-single parameters of Sec. 3.4.

QK-normalisation. Queries and keys in self- and cross-attention are RMS-normalised [6, 36] before the scaled dot product, $\hat{Q} = Q/(\|Q\|_{\text{rms}} + \varepsilon)$ and likewise K. This keeps attention logits finite in `bf16`: without it, $QK^\top/\sqrt{d_h}$ can overflow during flow-matching fine-tuning and yield NaN forward passes. Fine-tuning the architecturally similar PixArt-Σ-XL/2, which lacks QK-norm, under an identical recipe diverged repeatedly (gradient norms $> 10^7$); SANA converges stably.

Text conditioning. A frozen Gemma-2-2B [31] decoder-only LLM encodes captions: one forward pass, last-layer hidden states $E_{\text{cap}} \in \mathbb{R}^{S\times 2304}$. Each block cross-attends with queries from its hidden state and keys/values from E_{cap}. Against encoder-style T5-XXL, Gemma-2-2B needs ≈5 GB in `bf16` rather than ≈11, freeing ≈6 GB per GPU that we reallocate to batch size.

Output head. The final projection maps hidden states to C_z=32 channels, matching the DC-AE latent. Since the checkpoint was pretrained with velocity prediction rather than DDPM ε-prediction and the channel count already matches, *no output-head surgery is required*: the pretrained weights directly predict $v_\theta \in \mathbb{R}^{32\times 16\times 16}$.

3.4 AdaLN-Single Modulation

Every transformer block is conditioned through adaptive layer normalisation. SANA adopts the *single* variant (AdaLN-single) of PixArt-Σ [1], in which one shared affine projection of the timestep embedding supplies the modulation for all L blocks, in place of the L independent projections of the original diffusion transformer (DiT) [21]. Let $t_{\text{emb}} = \phi_t(t) \in \mathbb{R}^d$ denote the sinusoidal timestep embedding. The shared projection produces a global modulation vector

$$\mu = W_t\,\sigma(t_{\text{emb}}) + b_t, \qquad \mu \in \mathbb{R}^{6d}, \tag{3}$$

where $\sigma(\cdot)$ is the SiLU activation and $W_t \in \mathbb{R}^{6d\times d}$, $b_t \in \mathbb{R}^{6d}$ are learned. The vector μ is split into six components $(\gamma_1, \beta_1, \alpha_1, \gamma_2, \beta_2, \alpha_2)$, each in $\mathbb{R}^d$: within a block, (γ_1, β_1) scale and shift the input to the self-attention sub-layer and (γ_2, β_2) that of the feed-forward sub-layer, while α_1 and α_2 gate the respective residual contributions. A small per-block learned offset is added to the shared μ so that blocks remain individually expressive despite sharing one projection. Sharing this projection removes the redundant per-block modulation parameters of the original DiT while preserving per-block control through the gating terms, and it is the point at which we inject metadata in Sec. 3.6.

3.5 Geometry-Aware Metadata Encoding

Metadata is what separates a satellite tile from a natural image, yet prior work feeds it to the network with little regard for its structure. DiffusionSat [11] sinusoidally embeds each field and sums a per-field sinusoidal-MLP embedding with the timestep embedding inside the UNet's conditioning pathway. We instead observe that each field has a *geometry*: longitude and latitude lie on the sphere $\mathbb{S}^2$, month and day are cyclic, and sensor parameters vary across scales. Encoding a field without its geometry creates avoidable failures: a longitude of 179°

and one of -179° are adjacent on Earth but maximally distant as scalars, and December and January are one month apart in the world yet eleven units apart on a number line. We encode each field according to its true geometry, so that proximity in the embedding reflects proximity on the globe and on the calendar, and feed the result to the metadata graft of Sec. 3.6. Sec. 4.3 tests this claim directly by swapping only the encoder for a plain per-field sinusoidal alternative.

Fields and normalisation. Every FMoW tile carries the seven-field vector

$$m = \big(\text{lon}, \text{lat}, \text{GSD}, \text{cloud\%}, \text{year}, \text{month}, \text{day}\big) \in \mathbb{R}^7. \tag{4}$$

For reproducible comparison with DiffusionSat [11], each field is first linearly mapped to $[0, 1000]$:

$$\begin{aligned} m_{\text{lon}} &= \tfrac{\text{lon}+180}{360} \cdot 1000, & m_{\text{lat}} &= \tfrac{\text{lat}+90}{180} \cdot 1000, \\ m_{\text{GSD}} &= \tfrac{\text{GSD}}{\text{GSD}_{\max}} \cdot 1000, & m_{\text{cc}} &= \tfrac{\text{cloud\%}}{100} \cdot 1000, \\ m_{\text{yr}} &= \tfrac{\text{year}-1980}{120} \cdot 1000, & m_{\text{mo}} &= \tfrac{\text{month}}{12} \cdot 1000, \quad m_{\text{day}} = \tfrac{\text{day}}{31} \cdot 1000. \end{aligned} \tag{5}$$

Three geometries, three encoders. A raw normalised scalar is an impoverished input: it carries no periodicity (months and days are cyclic), no spherical geometry (longitude and latitude lie on $\mathbb{S}^2$, not the plane), and no multi-scale structure. We therefore route the seven fields through three specialist sub-encoders, each matched to the geometry of its field group; a spherical lift for geolocation, a periodic Fourier expansion for acquisition time, and a multi-scale Fourier expansion for sensor scalars.

(i) Geolocation: spherical lift with a Fourier basis. Longitude λ and latitude ϕ are first lifted to the unit sphere through the standard bijection

$$\mathbf{p}(\lambda, \phi) = \big(\cos\phi\cos\lambda,\ \cos\phi\sin\lambda,\ \sin\phi\big) \in \mathbb{R}^3, \tag{6}$$

and the resulting 3-vector is expanded at F_{geo}=32 dyadic frequencies $\omega_k = 2^k\pi$, $k = 0, \ldots, F_{\text{geo}}-1$:

$$\phi_{\text{geo}}(\lambda, \phi) = \big[\mathbf{p},\ \{\sin(\omega_k\mathbf{p})\}_k,\ \{\cos(\omega_k\mathbf{p})\}_k\big] \in \mathbb{R}^{3+6F_{\text{geo}}}, \tag{7}$$

giving a $3+6\times32 = 195$-dimensional feature. The spherical lift makes the encoding continuous across the antimeridian ($\lambda = \pm180^\circ$) and at the poles, which a direct (λ, ϕ) encoding cannot guarantee: the two coordinates above map to neighbouring points $\mathbf{p}$ and hence to nearby embeddings.

(ii) Acquisition time: multi-scale Fourier encoding. The normalised year, month, and day are expanded with F_{tmp}=16 dyadic frequencies and stacked with their raw values:

$$\phi_{\text{tmp}}(y, \mu, \delta) = \big[(y, \mu, \delta),\ \{\sin(\omega_k(y, \mu, \delta))\}_k,\ \{\cos(\omega_k(y, \mu, \delta))\}_k\big] \in \mathbb{R}^{3+6F_{\text{tmp}}}, \tag{8}$$

a $3+6\times16 = 99$-dimensional feature. The Fourier basis captures both the annual periodicity of month and day and the secular trend across years.

(iii) Sensor scalars: per-field Fourier encoding. GSD and cloud cover are each encoded independently with F_{sc}=16 frequencies,

$$\phi_{\text{sc}}(s) = \big[s,\ \{\sin(\omega_k s)\}_k,\ \{\cos(\omega_k s)\}_k\big]_{k=0}^{F_{\text{sc}}-1} \in \mathbb{R}^{1+2F_{\text{sc}}}, \tag{9}$$

giving $1+2\times16 = 33$ dimensions per field.

Projection. The four encoded components are concatenated into

$$\phi(m) = \left[\phi_{\text{geo}},\ \phi_{\text{tmp}},\ \phi_{\text{sc}}(\text{GSD}),\ \phi_{\text{sc}}(\text{cloud})\right] \in \mathbb{R}^{195+99+33+33} = \mathbb{R}^{360}, \tag{10}$$

and projected to the backbone hidden dimension by a three-layer multi-layer perceptron (MLP) with GELU activations:

$$m_{\text{emb}} = \text{MLP}_{360\to512\to512\to1152}\big(\phi(m)\big) \in \mathbb{R}^{d}, \qquad d = 1152. \tag{11}$$

With Xavier-uniform initialisation, the projection adds ≈1.04M trainable parameters ($360\times512 + 512\times512 + 512\times1152 = 1{,}036{,}288$, excluding biases) - under 0.2% of the backbone. For comparison, a plain per-field sinusoidal encoder over the same seven fields requires 11,370,240 parameters at d=1152, an order of magnitude more (Sec. 4.3). The embedding m_{emb} is consumed by the metadata graft of Sec. 3.6.

Relation to SatCLIP. Unlike SatCLIP [12], which encodes only (λ, ϕ) for contrastive retrieval, we additionally model temporal and sensor metadata, replace spherical harmonics with a spherical-lifted Fourier encoding (Eq. 7), and train the representation end-to-end through the flow-matching loss (Eq. 2) for AdaLN-based generation.

3.6 Zero-Initialised AdaLN Metadata Graft

We inject the metadata embedding m_{emb} (Sec. 3.5) through AdaLN-single modulation. A zero-initialised graft preserves the pretrained backbone at initialization while gradually enabling metadata conditioning during fine-tuning.

Why AdaLN instead of cross-attention. AdaLN-single is the canonical global-conditioning pathway in DiT, PixArt-Σ, and MMDiT [1, 2, 21]. Because metadata fields (location, GSD, cloud cover, date) are image-global, we inject them through AdaLN rather than cross-attention. This avoids competition with caption tokens for attention capacity, an issue indirectly reflected in DiffusionSat, where adding numerical metadata to captions reduces CLIP score (0.1720 vs. 0.1862) [11]. **Formulation.** We extend the AdaLN-single projection of Eq. 3 with a parallel, zero-initialised term:

$$\mu_{\text{final}} = \underbrace{W_t\,\sigma(t_{\text{emb}}) + b_t}_{\text{timestep (pretrained)}} + \underbrace{W_{\text{md}}\,\tilde{m}_{\text{emb}}}_{\text{metadata graft}}, \qquad W_{\text{md}}\big|_{\text{init}} = \mathbf{0}, \tag{12}$$

with $W_{\text{md}} \in \mathbb{R}^{6d\times d}$ and $\tilde{m}_{\text{emb}}$ the RMS-normalised metadata embedding,

$$\tilde{m}_{\text{emb}} = \frac{\sqrt{d}\,m_{\text{emb}}}{\|m_{\text{emb}}\|_2 + \varepsilon}, \qquad \|\tilde{m}_{\text{emb}}\|_2 \approx \sqrt{d}. \tag{13}$$

Because $W_{\text{md}} = \mathbf{0}$ at initialisation, $\mu_{\text{final}} = \mu_{\text{base}}$ exactly, and the grafted model reproduces the pretrained SANA output. As training proceeds, W_{md} accumulates gradient from the velocity loss (Eq. 2) and the metadata contribution grows smoothly from zero.

Relation to zero-init residual conditioning. The zero-initialised parallel branch follows the pattern introduced by ControlNet [57] and IP-Adapter [54], which attach branches with zero-initialised output projections so that a pretrained generator is recovered exactly at step zero. Those methods inject into the spatial feature maps of a UNet; we inject into the *modulation scalars* of a DiT. Applying the zero-init pattern to AdaLN conditioning rather than to feature maps is, to our knowledge, unexplored, and it preserves the pretrained velocity field

at initialisation - the period in which an un-warmed conditioning branch would otherwise perturb the image priors most.

Decoupling encoder scale from conditioning magnitude. RMS normalisation (Eq. 13) fixes $\|\tilde{m}_{\text{emb}}\|_2 \approx \sqrt{d}$, making the metadata contribution depend only on $\|W_{\text{md}}\|_F$. As W_{md} starts at zero and grows smoothly, the graft remains independent of the metadata encoder's output scale, enabling encoder replacement without retuning. This property is what makes the encoder-only ablation of Sec. 4.3 a clean comparison: the two encoders can be substituted without retuning the graft or the schedule.

Parameter count. The graft adds $W_{\text{md}} \in \mathbb{R}^{6d\times d}$, i.e. $6\times 1152\times 1152 = 7{,}962{,}624 \approx 7.96$M parameters; with the ≈ 1.04M-parameter metadata encoder (Sec. 3.5) this is ≈ 9.0M new parameters, roughly 1.5% of the ≈ 592M backbone. The backbone itself is fully fine-tuned; only the DC-AE and text encoder remain frozen.

3.7 Conditioning Dropout and Caption Curriculum

Two training-time choices govern how the model learns to use, and to do without, its conditioning: independent dropout of the caption and metadata signals, which enables classifier-free guidance (CFG) at inference, and a mixed caption curriculum that makes the model robust to caption length.

Independent conditioning dropout. CFG requires the model to have seen an unconditional signal during training. We drop the two conditioning streams independently so that the model learns the joint, both marginal, and the unconditional distributions. With probability p_{drop}=0.10, the encoded caption is replaced by the zero embedding per sample,

$$E_{\text{cap}} \leftarrow \mathbf{1}[\xi_i > p_{\text{drop}}]\cdot E_{\text{cap}}, \qquad \xi_i \sim \text{Uniform}(0,1), \tag{14}$$

while the cross-attention *padding mask is left unchanged*: zeroing the mask rather than the embedding would make every cross-attention row a uniform average over padded positions, yielding degenerate outputs. With the same probability, each of the seven metadata fields is masked independently *before* the encoder,

$$m_j \leftarrow m_j \cdot \mathbf{1}[\xi_{ij} > p_{\text{drop}}], \qquad j = 1,\ldots,7. \tag{15}$$

Per-field masking, rather than all-or-nothing dropout, exposes the model to partially specified metadata during training; we find this improves robustness at inference when some fields are missing (Sec. 4.3). The unconditional branch used by CFG sets both $E_{\text{cap}} = \mathbf{0}$ and $m = \mathbf{0}$.

Caption curriculum. Because FMoW lacks captions, we generate a 100–200 word VLM description for each tile. Training only on these rich captions creates a distribution shift, yielding poor performance on the short prompts common at inference. We therefore sample captions as 70% rich, 20% short, and 10% empty:

$$\text{caption} = \begin{cases} \text{rich} & \zeta < 0.70, \\ \text{short} & 0.70 \le \zeta < 0.90, \\ \varnothing & \zeta \ge 0.90, \end{cases} \qquad \zeta \sim \text{Uniform}(0,1). \tag{16}$$

Short captions are constructed from the category, reverse-geocoded city/country, and acquisition season. Missing VLM captions default to the short-caption mode.

Interaction of the two mechanisms. The empty-string mode in Eq. 16 and the embedding-level dropout in Eq. 14 both contribute unconditional caption signal. We use $p_{\text{caption\_drop}} = 0$

so the embedding level dropout does not affect us and our split falls back to Eq. 16 which is our default setting. The curriculum's benefit comes at no annotation cost: every short caption is derived from fields already present in FMoW, so length robustness is obtained for free.

3.8 Datasets

Training data: fMoW-RGB. We train on fMoW-RGB [2]: ≈257K images, 62 categories, up to 2048×2048, each with longitude, latitude, GSD, cloud cover, and timestamp, encoded as in Sec. 3.5. Images are resized and centre-cropped to 512×512; malformed TIFFs and samples with missing metadata (≈0.05%) are excluded.
Captions. fMoW lacks captions, so each image receives a 100–200 word VLM description and a short metadata-derived caption, sampled in training as 70% rich, 20% short, 10% empty (for classifier-free guidance).
Evaluation. Following DiffusionSat [11], we evaluate on the official fMoW-RGB test split (10,000 images) with same-pipeline VLM captions, plus short metadata-derived captions to separate caption quality from caption *style* (Sec. 4).
Out-of-domain evaluation. We evaluate zero-shot on RSICD [17] (10,921 images, 30 scene categories), whose human-written captions and different sensors and regions shift both text and image distributions; performance thus reflects robustness rather than memorisation.

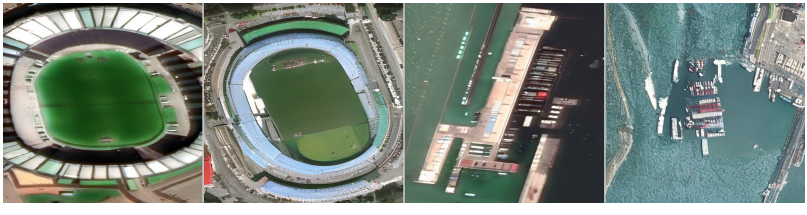


Figure 3: **Qualitative comparison on fMoW-RGB.** Columns 1 and 3: DiffusionSat (100 DDIM steps); columns 2 and 4: FlowSat (20 Euler steps). Left prompt: *"a large oval sports stadium with green field, tiered seating, and parking lots"*; right: *"a commercial port with piers, container ships, dockside cranes, and stacked shipping containers"*. FlowSat gives comparable or sharper structure at 5× fewer sampling steps.

3.9 Implementation details

Model. FlowSat fine-tunes Sana-0.6B [33][1] (28 blocks, hidden 1152), flow-matching pre-trained with QK-norm. A frozen DC-AE (32×) maps 512×512×3 to 16×16×32 latents; a frozen Gemma-2-2B-IT decoder (hidden 2304, ≤ 256 tokens, bf16; ∼5 GB vs. T5-XXL's ∼11) encodes text. Metadata passes through our geometry-aware encoder (Sec. 3.5) into the zero-init AdaLN graft (Sec. 3.6). Backbone (≈592M), encoder, and graft (≈9M) train end-to-end (≈601M total); DC-AE and text encoder stay frozen.
Optimisation. Training: 512×512, 125K steps from the public checkpoint; AdamW (β=(0.9, 0.9
wd 10^{-2}, ε=10^{-8}), peak LR 10^{-5} (cosine, 1000 warm-up), gradient clip $\|g\|_2 \leq 1.0$, bf16, gradient checkpointing, effective batch 24 (8×3 GPUs). Flow matching: logit-normal time sampling (μ=0, σ=1) on $x_t = (1-t)x_0 + tz$ with target $v = z - x_0$; zero-initialised W_{md} makes step-0 output exactly match pretrained Sana, adapting thereafter.

[1] `Efficient-Large-Model/Sana_600M_512px_diffusers`

Hardware. Training: 3× A100 80 GB (∼1.7 s/step, ∼60 h to 125K steps). Inference: 20-step Euler at CFG 2.5, native 512×512 on one GPU.

DiffusionSat reproduction. We re-train DiffusionSat [11] (SD U-Net + DDPM, authors' code) on the same fMoW-RGB split and VLM captions, matching effective batch, precision, optimiser, and EMA; only backbone (U-Net vs. DiT), objective (DDPM vs. flow matching), metadata injection (timestep-summed scalar vs. AdaLN graft), and sampler (100-step DDIM vs. 20-step Euler) differ.

3.10 Evaluation

We report FID [7], LPIPS (AlexNet) [58], SSIM [52], and CLIP score [22]. To compare sampling cost, the NFE column counts model forward passes per image, equal to the sampler step count for one-pass solvers (DDIM, Euler). All models are evaluated with the same FID implementation, the same reference statistics, the same preprocessing, and the same 10,000-image sample set; CLIP scores are computed with a single CLIP variant throughout, and are compared only within a fixed caption distribution.

4 Results

4.1 In-domain (FMoW) and out-of-domain (RSICD)

Table 2: **Quantitative comparison** on fMoW-RGB (in-domain) and RSICD (zero-shot); baselines from GeoDiT [26]. Ours: 20 Euler steps vs. ∼50–100. **Bold**=best, underline=second; †text+location. FlowSat: single seed, 125K checkpoint; across three seeds FID 31.53±0.32, CLIP 0.3018±0.0007, mean still best. Only DiffusionSat is protocol-controlled (our split, captions, budget, evaluation; Sec. 3.9); others as published, unmatched on data, preprocessing, or FID statistics.

		FMoW-RGB (In-Domain)				RSICD (Zero-Shot)			
Method	Steps	FID↓	LPIPS↓	SSIM↑	CLIP↑	FID↓	LPIPS↓	SSIM↑	CLIP↑
DiffusionSat [11]	100	35.27	0.5749	0.1299	0.2854	48.71	0.5125	0.2023	0.2831
GeoSynth [25]	∼50	70.55	0.5498	0.1948	0.2832	47.84	0.4989	0.2442	0.2699
CRS-Diff [30]	∼50	45.06	0.5566	0.1074	0.2769	32.99	0.5071	0.1882	0.2863
GeoDiT-α [26]	∼50	33.32	0.5400	0.2013	0.2956	**31.76**	**0.4992**	**0.2680**	0.2764
GeoDiT-2Σ [26]	∼50	32.11†	**0.5233**†	**0.2212**†	0.2967†	–	–	–	–
FlowSAT (ours)	**20**	**31.10**	0.6853	0.1923	**0.3016**	48.32	0.6555	0.2029	**0.2871**

In-domain (FMoW). FlowSat attains the best FID (31.10) and best CLIP score (0.3016) among all compared methods, surpassing the strongest published baseline (GeoDiT-2Σ, which additionally conditions on point locations) on both distribution-level fidelity and caption alignment. This is achieved with 20 Euler steps against ≥ 50 for the competing diffusion models ($\geq 2.5\times$ fewer forward passes) and without a larger or different training corpus. To confirm that this is not a favourable-seed artifact, we repeat training and evaluation under three seeds: FID $= 31.53 \pm 0.32$ and CLIP $= 0.3018 \pm 0.0007$ over 10,000 samples per run. The mean retains the lead over the strongest published baseline (GeoDiT-2Σ, FID 32.11), though at 0.58 FID the margin is within roughly two standard deviations and we do not claim a statistically separated gap on FID alone; the CLIP margin (0.3018 vs. 0.2967) is

comparatively larger relative to its seed spread. SSIM (0.1923) trails the best GeoDiT variant (0.2212), and LPIPS is higher than the baselines; we analyse both below, as they reflect the same underlying trade-off.

Zero-shot (RSICD). FlowSat outperforms DiffusionSat on both FID (48.32 vs. 48.71) and CLIP (0.2871 vs. 0.2831), achieving the best CLIP score overall. Although GeoDiT-α and CRS-Diff obtain stronger absolute FID, all methods degrade under the RSICD distribution shift. FlowSat's preserved CLIP lead indicates stronger transfer of text alignment than image fidelity.

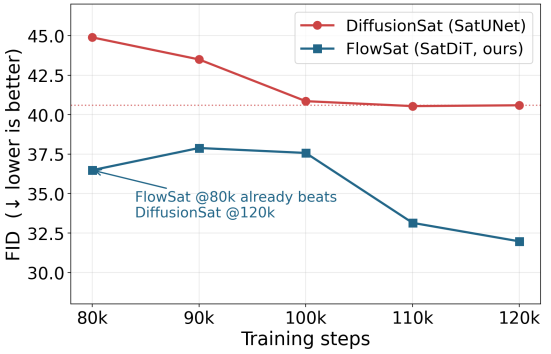
CLIP under an independent caption distribution. Tab. 2 computes CLIP with the training VLM captions, which risks rewarding style familiarity. We re-evaluate with short metadata-derived captions built from the fMoW category, reverse-geocoded city and country, and acquisition season, independent of the VLM pipeline. FlowSat scores 0.2983 (n=10,000), 0.003 below its rich-caption 0.3016; retrained DiffusionSat scores 0.2852 on identical prompts, images, and CLIP variant. The advantage persists, ruling out caption-style leakage. The control is conservative: short captions sit closer to DiffusionSat's native distribution than ours, and neither model is evaluated on its own training caption distribution.

Reference-based vs. distributional metrics. FlowSat achieves the best FID and CLIP but weaker LPIPS and SSIM. We bound the gap. First, the reference is one of many valid acquisitions: real-vs-real LPIPS at the same location under temporal variability is 0.4251 (95% CI $\pm$0.0063, n=1500), roughly 62% of FlowSat's 0.6853 before any modelling error. Second, the autoencoder adds a floor: DC-AE reconstructions give LPIPS 0.1405/SSIM 0.6108 (n=2000) vs. 0.1171/0.7070 for the 8$\times$ SD-VAE of DiffusionSat and GeoDiT; the 32$\times$ latent is a less faithful target, so the gap partly inherits the compression the 256-token, 20-step pipeline requires. Third, FlowSat's SSIM (0.1923) approaches the real-vs-real 0.2138. The gap thus reflects temporal ambiguity and stronger compression, not weaker distributional generation (FID 31.10 vs. 35.27). A lower-compression autoencoder would improve these metrics at the token- and step-budget cost; a trade-off, unsettled without a matched-backbone experiment.

Qualitative prompt adherence. Fig. 3 illustrates FlowSat's tendency to render not only the primary subject but its surrounding context. For the prompt of a stadium, both samples reproduce the core elements - the oval structure, tiered seating, and green field but they differ in contextual completeness: one frames the stadium in isolation, while the other additionally renders the parking areas, access roads, and surrounding vegetation implied by a true overhead view. This reflects the in-domain fidelity captured quantitatively by FID (Sec. 4): plausible scenes include the infrastructure that co-occurs with the named subject, not the subject alone. We show a single case here for illustration; the aggregate trend is what the FID and CLIP results measure.

4.2 Sampling and Training Efficiency

Inference efficiency. FlowSat matches DiffusionSat's published 100-step DDIM FID at $\approx$20 Euler steps, a 5$\times$ reduction in function evaluations: a linear-interpolant flow with an Euler ODE solver needs far fewer discretisation steps than a stochastic DDIM trajectory [13, 15]. This advantage is a property of flow matching and the SANA/DC-AE backbone, not of our metadata conditioning; any flow-matching DiT inherits it. Our contribution, the geometry-aware encoder and zero-init AdaLN graft, is isolated in Tab. 3; the efficiency figures characterise the system, not the conditioning claim.



Steps	DiffusionSat FID ↓	FlowSat FID ↓
80k	44.89	**36.48**
90k	43.49	**37.88**
100k	40.85	**37.57**
110k	40.54	**33.14**
120k	40.59	**31.97**

Figure 4: **FID vs. training steps** (10,000 samples, 20 inference steps, CFG 2.5). FlowSat leads at every checkpoint: at 80k it reaches FID 36.48, already below DiffusionSat's best (40.54), and improves through 120k (31.97, ∼21% relative reduction) while DiffusionSat plateaus near 40 after 100k. Values are at the 120K checkpoint; Tab. 2's 31.10 is the 125K checkpoint, same sampler and budget.

Training efficiency. FlowSat reaches its reported numbers after 125K fine-tuning steps from the public Sana initialisation on fMoW-RGB alone; GeoDiT-2Σ trains ∼900K steps from scratch, and DiffusionSat fine-tunes ∼100K–150K steps from Stable Diffusion with additional remote-sensing corpora (SpaceNet, Satlas). Initialisation, data, and scale differ, so these counts only situate FlowSat's budget; Fig. 4 gives the matched-condition comparison against our re-trained DiffusionSat.

4.3 Ablation Studies

Metadata injection: zero-init vs. random-init. Zero-initialising the metadata projection W_{md} into AdaLN preserves the pretrained Sana prior at step 0. Against Kaiming-normal initialisation in an otherwise identical setup, zero-init yields coherent imagery by 2K steps and better samples through 10K, while Kaiming-init remains degraded (Fig. 5); the gain stems from zero initialisation itself, not the graft layer alone.

Geometry-aware vs. plain metadata encoding. To isolate the encoder from the graft, we replace *only* the encoder of Sec. 3.5 with a per-field sinusoidal–MLP encoder in the style of DiffusionSat [□], holding backbone, zero-init graft, data, caption curriculum, batch, LR schedule, and seed fixed, and train both for 50K steps. The geometry-aware encoder improves FID 46.81 → 37.96 and CLIP 0.3006 → 0.3032 (Tab. 3); with no other difference between runs, the gain is attributable to the encoding itself. It is also *smaller*: the plain encoder spends an independent $256 \to 1152 \to 1152$ MLP per field (7 fields, 11.37M parameters), while native-geometry encoding with one joint $360 \to 512 \to 512 \to 1152$ MLP needs 1.04M, i.e. better FID/CLIP with ∼11× fewer conditioning parameters, so the benefit is a structure-matched representation, not capacity. These numbers are not comparable to Tab. 2 (31.10 at 125K): both arms stop at 50K, compute-matched but unconverged.

Metadata controllability. Fixing caption and latent noise, we sweep one field at a time (Fig. 6). Month at ≈66°N (no seasonal language in the caption) gives a summer-to-winter transition, July vegetation to February snow; GSD from 3.0 to 0.3 m on a port scene gives a coarse-to-fine progression, containers and vessel detail emerging only at low GSD. Seed fixed per row, the swept field is the only changing input.

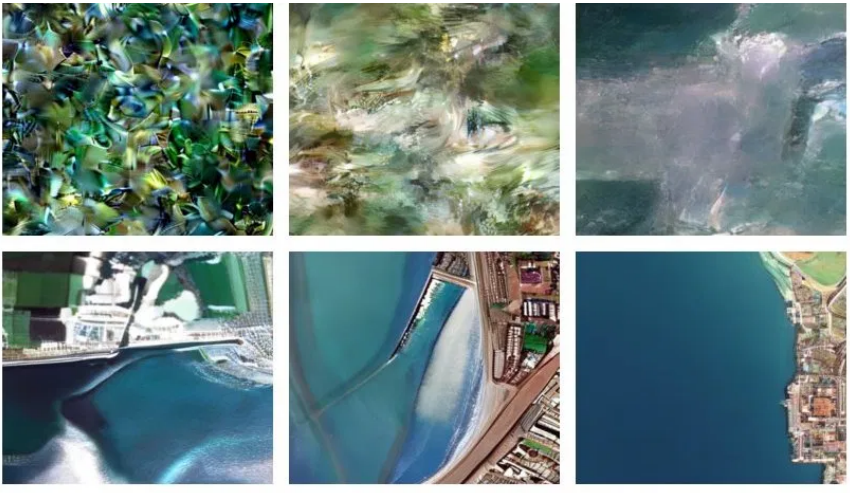


Figure 5: **Zero- vs. Kaiming-init metadata graft** (matched caption, metadata, seed). Zero-init (bottom) preserves Sana's pretrained AdaLN modulation at step 0 and inherits the image prior immediately; Kaiming-init (top) perturbs it from the first step, forcing the model to relearn the prior before using metadata.

Table 3: **Metadata encoder ablation.** Only the encoder differs; all else (backbone, graft, data, curriculum, batch, schedule, seed) is identical. Both arms: 50K steps, same fMoW-RGB test split and FID reference. Params exclude the shared W_{md} graft (≈7.96M). FID exceeds Tab. 2 (50K vs. 125K steps).

Metadata encoder	Params	FID ↓	CLIP ↑
Per-field sinusoidal [□]	11.37M	46.81	0.3006
Geometry-aware (ours)	1.04M	**37.96**	**0.3032**

5 Limitations

FlowSat attains the best FID/CLIP but lower LPIPS/SSIM; Sec. 4 bounds this empirically, and fully separating the autoencoder's share needs a matched-backbone, lower-compression experiment. The 50K-step encoder ablation (Tab. 3) fixes sign, not converged magnitude; sampling efficiency (Sec. 4.2) is inherited from flow matching and SANA; controllability (Fig. 6) is qualitative (direction, not effect size). Baselines are reproduced, with possible minor pipeline differences despite matched splits; RSICD FID is not the best absolute.

6 Future Work

The decoupled conditioning and zero-init graft extend naturally to **(i)** *temporal sequence generation*, via per-frame metadata and a lightweight temporal-attention module; **(ii)** *change-guided inpainting*, via a zero-initialised reference-frame projection alongside the caption route; and **(iii)** *GSD-controlled super-resolution*, building on the coarse-to-fine GSD response of Fig. 6, for which the metadata controllability shown here is a prerequisite.

Figure 6: **Metadata controllability.** Caption fixed, one field swept. Top: month at 66°N, July → September → December → February. Bottom: GSD 3.0 → 1.5 → 0.6 → 0.3 m.

7 Broader Applications

Beyond the controllability demonstrated above, FlowSat's captioned, metadata-conditioned generation enables a range of downstream uses in Earth observation (EO). We group them by the property each relies on.

Generating labelled or unlabelled training data. FlowSat can augment scarce EO datasets by generating metadata-matched examples for under-represented classes and acquisition conditions, while also enriching self-supervised pretraining corpora with greater geographic and seasonal diversity.

Sharing restricted imagery. Metadata-conditioned synthetic imagery can provide shareable alternatives to real data when licensing or security constraints restrict access [11, 30].

Planning and counterfactuals. FlowSat enables counterfactual visualisation of proposed land-use changes and converts EO data into intuitive RGB imagery for planning, communication, and education.

Stress-testing EO pipelines. FlowSat enables controlled evaluation of EO models by generating test data with specific seasons, cloud cover, and land-cover conditions.

These are application directions enabled by FlowSat's controllable generation, not capabilities evaluated in this paper; quantifying their downstream benefit is future work.

8 Conclusion

We introduced FlowSat, a flow-matching DiT for metadata-conditioned satellite image generation. By combining a zero-initialised AdaLN graft with a geometry-aware metadata encoder, FlowSat achieves the best in-domain FID and CLIP at only 20 sampling steps, transfers zero-shot to RSICD, and demonstrates stable and controllable metadata conditioning. An encoder-controlled ablation attributes the gain to the geometry-aware encoding itself rather than to the backbone or the graft, and a short-caption evaluation confirms that the CLIP advantage is not an artifact of caption style.

References

[1] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *European Conference on Computer Vision*, pages 74–91. Springer, 2024.

[2] Gordon Christie, Neil Fendley, James Wilson, and Ryan Mukherjee. Functional map of the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6172–6180, 2018.

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

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

[5] Zhengyang Geng, Ashwini Pokle, and J Zico Kolter. One-step diffusion distillation via deep equilibrium models. *Advances in Neural Information Processing Systems*, 36: 41914–41931, 2023.

[6] Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4246–4253, 2020.

[7] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[8] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[9] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[10] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention. In *International conference on machine learning*, pages 5156–5165. PMLR, 2020.

[11] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery. In *International Conference on Learning Representations*, volume 2024, pages 5586–5604, 2024.

[12] Konstantin Klemmer, Esther Rolf, Caleb Robinson, Lester Mackey, and Marc Rußwurm. Satclip: Global, general-purpose location embeddings with satellite imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4347–4355, 2025.

[13] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

[14] Jiaqi Liu, Ronghao Fu, Haoran Liu, Lang Sun, and Bo Yang. Geodit: A diffusion-based vision-language model for geospatial understanding. *arXiv preprint arXiv:2512.02505*, 2025.

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

[16] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *Advances in neural information processing systems*, 35:5775–5787, 2022.

[17] Xiaoqiang Lu, Binqiang Wang, Xiangtao Zheng, and Xuelong Li. Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4):2183–2195, 2017.

[18] Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *European Conference on Computer Vision*, pages 23–40. Springer, 2024.

[19] Oisin Mac Aodha, Elijah Cole, and Pietro Perona. Presence-only geographical priors for fine-grained image classification. In *Proceedings of the IEEE/cvf international conference on computer vision*, pages 9596–9606, 2019.

[20] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[21] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023.

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

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

[24] Marc Rußwurm, Konstantin Klemmer, Esther Rolf, Robin Zbinden, and Devis Tuia. Geographic location encoding with spherical harmonics and sinusoidal representation networks. In *International Conference on Learning Representations*, volume 2024, pages 1746–1759, 2024.

[25] Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: Contextually-aware high-resolution satellite image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 460–470, 2024.

[26] Srikumar Sastry, Dan Cher, Brian Wei, Aayush Dhakal, Subash Khanal, Dev Gupta, and Nathan Jacobs. Geodit: Point-conditioned diffusion transformer for satellite image synthesis. *arXiv preprint arXiv:2603.02172*, 2026.

[27] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[28] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[29] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020.

[30] Datao Tang, Xiangyong Cao, Xingsong Hou, Zhongyuan Jiang, Junmin Liu, and Deyu Meng. Crs-diff: Controllable remote sensing image generation with diffusion model. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–14, 2024.

[31] Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[32] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[33] Enze Xie, Junsong Chen, Junyu Chen, Han Cai, Haotian Tang, Yujun Lin, Zhekai Zhang, Muyang Li, Ligeng Zhu, Yao Lu, and Song Han. SANA: Efficient high-resolution text-to-image synthesis with linear diffusion transformers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=N8Oj1XhtYZ.

[34] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[35] Zhiping Yu, Chenyang Liu, Liqin Liu, Zhenwei Shi, and Zhengxia Zou. Metaearth: A generative foundation model for global-scale remote sensing image generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(3):1764–1781, 2024.

[36] Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

[38] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.