SATLEN: EVALUATING SEMANTIC-TO-SATELLITE TRANSLATION ACROSS NESTED SCALES

Anonymous authors
Paper under double-blind review

ABSTRACT

Image generators are increasingly asked to render the world from a map. We ask whether they respect a map's *physical scale*. We introduce **SatLen**, a benchmark that renders the same 100 scenes from 45 countries and territories as land-cover maps at four ground footprints, from 0.625 to 5 km per side, at a fixed input resolution. Across 21 generator configurations and 8,400 generations, we find a *coverage paradox*: as the footprint widens, every configuration becomes more similar to the real image under CLIP while matching the requested layout less closely. After removing the overlap that correct class proportions earn by chance, spatial placement falls by 56%, and three configurations fall to near-chance placement at 5 km. Pixel-level diagnostics that bypass the segmenter show where the failure lies. Generated edges still follow the map's boundaries, and do so more sharply at wide coverage, yet the land-cover identity of the enclosed regions is lost. A segmenter-free nearest-class-mean classifier confirms that this semantic loss is not a segmenter artifact. Models also treat scale as a texture setting: output textures are too smooth close up and gain fine detail far faster than real imagery as coverage grows. A prompt ablation without the extent sentence shows the same ramp, indicating that models read scale mainly from the map. Finally, a model's close-up and wide views of the same location rarely depict the same place. Current generators *keep the lines but lose the land*.

1 INTRODUCTION

Map-conditioned satellite image generation renders overhead imagery from a dense land-cover map, with applications in remote-sensing data augmentation, urban planning and disaster-scenario simulation. Recent multimodal generators and editors (OpenAI, 2026; Google, 2025b; Seedream, 2026; Black Forest Labs, 2025; Deng et al., 2025; Wu et al., 2025; Liu et al., 2025) now accept such maps directly through a prompt-plus-image interface, and map-conditioned synthesis (Espinosa & Crowley, 2023; Tang et al., 2024) is increasingly used for this purpose. Yet a central question remains open: do these models understand the *geography* they are asked to draw, or only its *appearance*? The question is most acute for physical scale. A satellite photograph of a city block and one of the whole city depict the same places at different ground sampling: buildings shrink to a few pixels, roads thin to lines, and the scene becomes a denser mosaic of the same land cover.

Evaluating this capability requires three ingredients. First, the scene must be held fixed while the scale varies, so that differences between outputs reflect scale rather than location or imagery source. Second, layout agreement must be measured beyond what correct class proportions earn by chance, because a map with the right mix of land cover in the wrong places still overlaps its target. Third, the diagnosis must not rest on a single learned segmenter but be confirmed directly from pixels.

However, existing benchmarks do not provide these ingredients. General image-editing and generation benchmarks such as ImagenHub, ImgEdit and HQ-Edit (Ku et al., 2024; Ye et al., 2025; Hui et al., 2025) focus on natural images at a single, implicit scale. Remote-sensing generators such as DiffusionSat and GeoSynth (Khanna et al., 2024; Sastry et al., 2024) are trained and tested at fixed ground sampling, and their evaluations report appearance metrics that do not test layout preservation. Whether generators follow a spatial condition across scales thus remains untested.

To address this gap, we introduce **SatLen**, a benchmark that holds each scene fixed while varying its physical scale. SatLen contains 100 scenes from 45 countries and territories, selected from $48{,}468$

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.

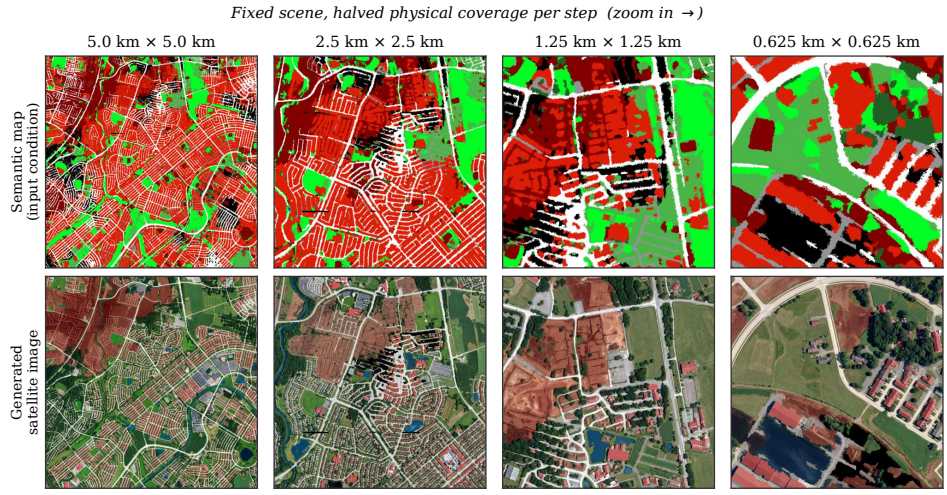


Figure 1: SatLen renders one source scene at four ground footprints of 5 to 0.625 km per side at the same input resolution. A model receives the semantic map in the top row and produces the satellite imagery below it. Because the 5 km view contains every smaller crop, each model's outputs can be checked for agreement with the map, the real image, and each other.

quality-controlled global anchors, and represents each as a four-level crop pyramid from 5 to 0.625 km per side, as shown in Figure 1. Every level is shown to the model at the same 1024-pixel resolution with a coverage-aware prompt, and the full 5 km view contains every smaller crop. Figure 2 outlines the construction and evaluation pipeline. We evaluate 21 closed- and open-weight configurations on all 400 instances, yielding 8,400 generations.

The results reveal a *coverage paradox*: from 0.625 to 5 km, every evaluated configuration becomes more similar to the real image under CLIP while matching the requested layout less closely. Beyond chance, spatial placement falls by 56%. Pixel-level diagnostics show that the generated edges still follow the map's boundaries, whereas the regions between them lose their land-cover identity. Models also render scale mainly as texture density, and their close-up and wide views of the same location rarely depict the same place. Current generators *keep the lines but lose the land*.

Our contributions are as follows:

- We formulate the evaluation of scale understanding in map-conditioned satellite generation as a within-scene, multi-scale protocol that combines chance-corrected layout scoring with segmenter-free pixel diagnostics.
- We introduce SatLen, a benchmark of 100 globally sampled scenes at four coverage levels, with paired semantic maps, prompts and reference imagery, and 8,400 scored generations from 21 model configurations.
- We identify the coverage paradox in all 21 configurations and diagnose its cause. Current generators preserve the map's geometry but lose its land-cover semantics, render scale as texture, and produce cross-scale views that are largely inconsistent.

2 Related Work

2.1 Layout-Conditioned Generation and Editing Benchmarks

Spatially conditioned synthesis builds on paired image translation (Isola et al., 2017; Wang et al., 2018), unpaired translation (Zhu et al., 2017), and semantic-image synthesis with spatially adaptive normalization (Park et al., 2019). Latent diffusion (Rombach et al., 2022), ControlNet (Zhang et al., 2023), and InstructPix2Pix (Brooks et al., 2023) extend image synthesis with spatial or instruction-based conditioning. Unified generation and editing models include GPT-Image (OpenAI, 2026), Nano Banana (Google, 2025b), Seedream (Seedream, 2026; Team Seedream et al., 2025), FLUX.2 (Black Forest Labs, 2025), Qwen-Image (Wu et al., 2025), Bagel (Deng et al., 2025), Step1X-Edit (Liu et al., 2025), OmniGen (Xiao et al., 2025), DreamOmni2 (Xia et al., 2025), and ICEdit (Zhang et al., 2025).

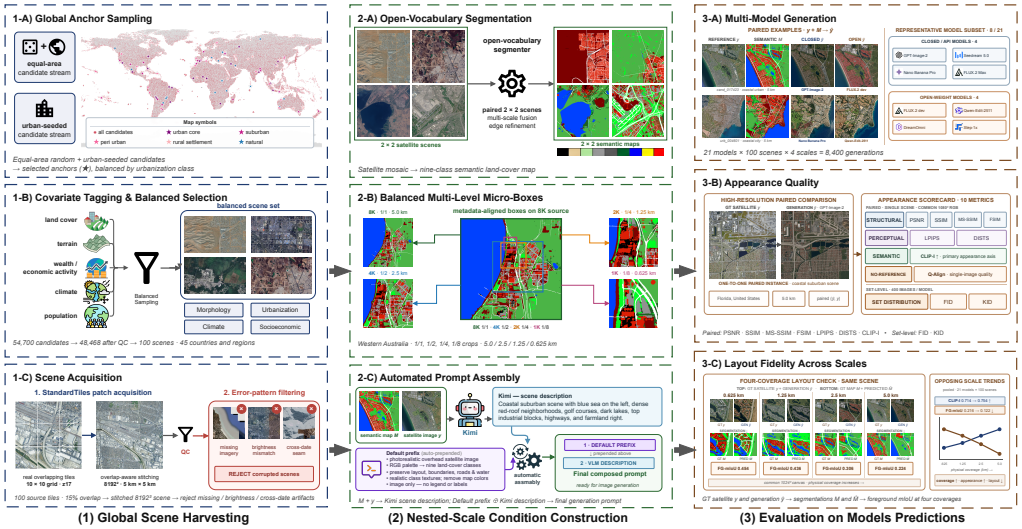


Figure 2: The SatLen task. A model receives a dense land-cover map M and a coverage-aware prompt t and generates an overhead image $\hat{y}$. Appearance is evaluated against reference imagery y, and layout fidelity is evaluated against the conditioning map M using segmentation.

Evaluation resources such as ImagenHub (Ku et al., 2024), GEdit-Bench (Liu et al., 2025), ImgEdit (Ye et al., 2025), and HQ-Edit (Hui et al., 2025) support conditional generation or instruction-based editing through benchmark protocols, datasets, and quality measures. All of them evaluate at a single, implicit scale. SatLen instead varies physical scale while holding the scene fixed, which exposes failures that single-scale evaluation cannot detect.

2.2 Remote-Sensing Generation and Evaluation

Remote-sensing generators include DiffusionSat (Khanna et al., 2024), GeoSynth (Sastry et al., 2024), map-conditioned synthesis (Espinosa & Crowley, 2023), and controllable remote-sensing diffusion (Tang et al., 2024). SatLen complements them by comparing general-purpose editors across ground coverages. Its semantic taxonomy follows the eight foreground land-cover classes of OpenEarthMap (Xia et al., 2023), with an additional background label. Maps are inferred from reference imagery with the SegEarth-OV-3 open-vocabulary segmenter (Li et al., 2025).

The evaluation combines the structural metrics PSNR, SSIM (Wang et al., 2004), MS-SSIM (Wang et al., 2003) and FSIM (Zhang et al., 2011), the learned perceptual distances LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022), the semantic similarity CLIP-I (Radford et al., 2021), the distributional distances FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018), and the no-reference quality score Q-Align (Wu et al., 2024). We show that several of these metrics improve as layout fidelity falls with coverage. We therefore complement them with a chance-corrected layout score and with pixel-level diagnostics that do not rely on any learned model.

3 The SatLen Benchmark

3.1 Principle and Motivation

SatLen isolates physical scale as the experimental variable. Each source scene supplies a four-level crop pyramid, and every level is presented at the same input resolution, so only the ground area per pixel changes between levels. The 100 scenes span all inhabited continents and a balanced mix of settlement types, as shown in Figures 3 and 4. Geography, acquisition conditions and imagery source are held fixed within each scene. The full 5 km view contains every smaller crop, so a model's outputs at different scales can be compared with one another as well as with the map and the reference. Crop placement maximizes label diversity at each level, as described in Section 3.3, and the chance-corrected measure of Section 3.4 accounts for the resulting class proportions.

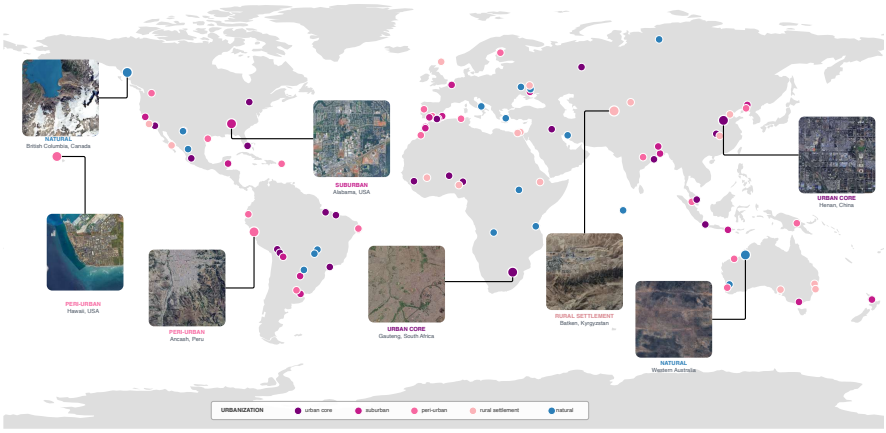


Figure 3: Geographic distribution of the 100 evaluation scenes. Anchors are drawn from an equal-area global sample augmented with an urban-seeded stream, then category-balanced. The set spans 45 countries and territories on all inhabited continents.

3.2 Task Definition

Figure 2 illustrates an instance, a triple (M, t, y). The condition $M \in \{0, \ldots, 8\}^{H \times W}$ is a semantic land-cover map over nine coarse classes, namely background, bareland, grass, pavement, road, tree, water, cropland and building, rendered as a flat color image. The prompt t instructs the model to convert M into a photorealistic overhead satellite image, spells out the color–class legend, states the ground extent, for example "this overhead scene covers approximately $5\,\text{km} \times 5\,\text{km}$", and appends a one-sentence scene description. The reference y is the real satellite tile from which the map was derived. A model under test produces $\hat{y} = G(M, t)$, which we compare with y and M.

3.3 Dataset Construction & Composition

Scene acquisition and semantic maps. Each scene originates as an 8192×8192 pixel satellite mosaic covering a $5\,\text{km} \times 5\,\text{km}$ ground footprint centered on a sampled anchor. Mosaics are assembled from static-map tiles at the zoom level whose native resolution first meets the target ground-sampling distance, stitched with 15% overlap and resampled to a common 8192-pixel canvas. Scenes with more than 20% blank tiles are dropped. Semantic maps are inferred from each mosaic using the open-vocabulary SegEarth-OV-3 pipeline (Li et al., 2025) and mapped into the nine-class taxonomy. The supplement documents text aliases, multi-scale inference, and mask refinement. The same segmenter labels the generated images, and Section 3.4 adds segmenter-free diagnostics.

The within-scene crop pyramid. From each 8192-pixel scene we extract crops at four scale fractions, $1/1$, $1/2$, $1/4$ and $1/8$, and resample every crop to a common 1024-pixel input. The resulting square footprints have side lengths of 5.0, 2.5, 1.25, and 0.625 km, respectively. Throughout the paper, coverage denotes footprint side length, not area, and the effective ground-sampling distance halves at each successive level. The $1/1$ crop is the full scene itself. For the three sub-crops we employ a *label-balanced placement* strategy: candidate positions are scanned on a block-aligned grid and the position maximizing semantic diversity is selected. Specifically, let $\mathcal{Q}_s$ denote the set of block-aligned candidate positions for a crop of side s pixels, and let $n_i(q)$ be the exact pixel count of foreground class i inside candidate q. We select the crop position that maximizes normalized Shannon entropy over the thresholded class distribution:

$$q^\star = \arg\max_{q \in \mathcal{Q}_s} \left[-\frac{1}{\log 8} \sum_{i=1}^{8} p_i(q) \, \log p_i(q) \right], \qquad (1) \quad p_i(q) = \frac{n_i(q)\, \mathbf{1}\big[n_i(q) \geq \lceil \rho\, s^2 \rceil\big]}{\sum_{j=1}^{8} n_j(q)\, \mathbf{1}[n_j(q) \geq \lceil \rho\, s^2 \rceil]}, \qquad (2)$$

with $\rho = 0.001$ the minimum-area threshold that suppresses tiny fragments and the convention $0 \log 0 = 0$ in Equation 1. The denominator $\log 8$ normalizes to $[0, 1]$ over eight foreground classes. Ties favour more scored classes, then more foreground pixels, then the scene center.

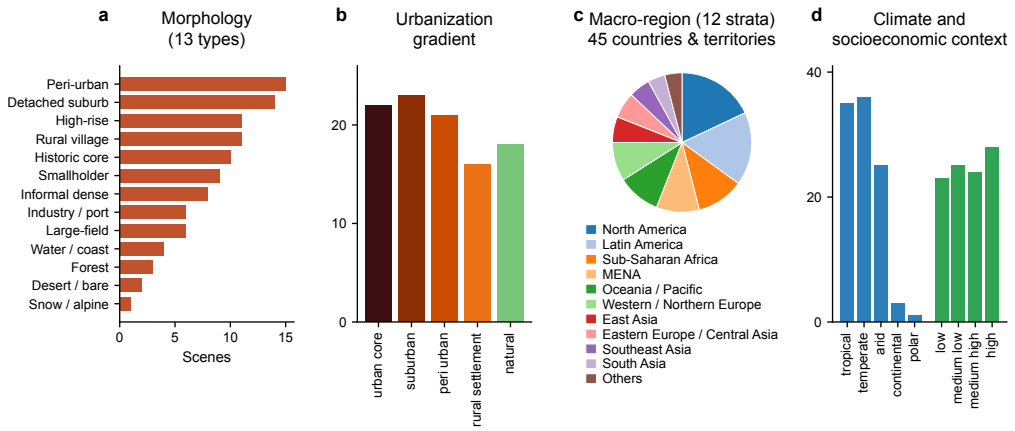


Figure 4: Composition of the 100 evaluation scenes by morphology, urbanization, world region, climate group and socioeconomic-context quartile, the five dimensions used in scene selection.

Crop positions are selected independently at each size, and every crop lies inside the full 5 km scene. Each crop has its own semantic map, coverage-aware prompt, scene description, and real reference.

Global sampling and transparent tagging. Candidate anchors come from an equal-area random sample over land, with a minimum spatial separation, and from an urban-seeded stream drawn from the 2020 GHSL built-surface mask. For every anchor we reduce a battery of public geospatial products over concentric footprints: ESA WorldCover land cover (Zanaga et al., 2022), GHSL built surface and population (Schiavina et al., 2023), SRTM terrain (Farr et al., 2007), VIIRS nighttime lights (Román et al., 2018), WorldClim climate (Hijmans et al., 2005), gridded HDI (Kummu et al., 2018), and administrative boundaries. A targeted subset of dense urban anchors has its morphology refined with Overture building footprints (Overture Maps Foundation, 2024). From these covariates we derive by transparent ordered rules the five selection dimensions summarized in Figure 4, namely a *morphology class* with 14 physical-form types, 13 of them in the evaluation split, a five-level *urbanization class*, a *custom coarse climate group*, and an *area-level socioeconomic-context quartile*. Quality control retains 48,468 candidates, from which a greedy constrained-diversity selector first selects 1,000 anchors. A second stage chooses the 100 evaluation scenes from 293 acquired and segmented scenes that pass imagery review, as detailed in the supplement.

The resulting set covers every urbanization level and socioeconomic quartile in roughly equal numbers, and 13 morphology classes ranging from dense historic cores and high-rise superblocks to large-field agriculture and desert bare sparse land.

3.4 Evaluation Protocol

Appearance quality. We report ten metrics in four families. Classical structure uses PSNR, SSIM (Wang et al., 2004), MS-SSIM (Wang et al., 2003) and FSIM (Zhang et al., 2011). Learned perceptual distance uses LPIPS (Zhang et al., 2018) and DISTS (Ding et al., 2022). Semantic similarity uses the CLIP image–image cosine (Radford et al., 2021), denoted CLIP-I. Distribution and no-reference quality use FID (Heusel et al., 2017) and KID (Bińkowski et al., 2018) per model, and Q-Align (Wu et al., 2024). Paired metrics use a common 1080×1080 RGB normalization, and CLIP and Inception use their canonical preprocessing.

Layout fidelity. SegEarth-OV-3 labels each generated image $\hat{y}$ into the nine-class taxonomy, and we compute foreground mean intersection-over-union, FG-mIoU, against M. CLIP-I and FG-mIoU are the two primary axes, since a map constrains *where* categories lie but not their texture.

Chance-corrected placement. FG-mIoU rewards correct class proportions even when regions are misplaced. For an output with predicted class histogram $\hat{n}_i$ and target histogram n_i over N pixels, independent placement gives expected intersection $n_i \hat{n}_i / N$. This defines a chance score mIoU_0 under the same classwise averaging as FG-mIoU. The best score achievable with the same histograms,

Table 1: Main leaderboard. Metrics are averaged over all 100 scenes and four coverage levels, and FID and KID are computed per model. Arrows indicate the preferred direction. Bold and underlining mark the best and second-best unrounded values in each column, and displayed values can tie after rounding. GPT-Image-2 leads on CLIP-I, while Seedream 5.0 leads on FG-mIoU, FID, and KID. Rows are ordered by category-balanced appearance rank within each group. IDs identify the model points in Figure 5a.

ID	Model	PSNR↑	SSIM↑	MS-SSIM↑	FSIM↑	LPIPS↓	DISTS↓	CLIP-I↑	FID↓	KID↓	Q-AI↑	FG-mIoU↑
Closed / API models												
1	GPT-Image-2 (OpenAI, 2026)	**13.05**	**0.203**	**0.284**	0.649	**0.571**	0.276	**0.833**	127.5	0.044	4.09	0.310
2	Seedream 4.5 (Seedream, 2025)	12.57	0.196	0.263	**0.650**	0.608	**0.272**	0.813	131.5	0.048	4.26	0.281
3	Nano Banana 2 (Google, 2026)	12.28	0.168	0.243	0.631	0.615	0.283	0.807	128.4	0.047	4.37	0.283
4	Seedream 5.0 (Seedream, 2026)	12.00	0.173	0.245	0.633	0.621	0.281	0.802	**116.0**	**0.040**	4.22	**0.327**
5	Nano Banana Pro (Google, 2025b)	12.10	0.161	0.233	0.625	0.626	0.287	0.799	136.3	0.054	4.34	0.270
6	Seedream 4.0 (Team Seedream et al., 2025)	11.09	0.143	0.208	0.628	0.652	0.296	0.774	127.3	0.049	4.36	0.243
7	Nano Banana (Google, 2025a)	12.47	0.172	0.238	0.624	0.642	0.295	0.794	135.3	0.050	3.70	0.265
8	GPT-Image-1.5 (OpenAI, 2025)	12.46	0.112	0.142	0.627	0.646	0.311	0.806	129.8	0.056	4.25	0.159
9	FLUX.2 Max (Black Forest Labs, 2025)	11.48	0.121	0.124	0.569	0.732	0.320	0.763	164.6	0.080	4.43	0.074
10	FLUX.2 Pro (Black Forest Labs, 2025)	11.39	0.119	0.124	0.565	0.738	0.322	0.760	163.9	0.083	**4.43**	0.072
Open-weight models												
11	FLUX.2 dev (Black Forest Labs, 2025)	12.99	0.178	0.261	0.645	0.621	0.293	0.814	134.9	0.046	4.00	0.243
12	Qwen-Edit-2511 (Wu et al., 2025)	11.82	0.168	0.215	0.617	0.643	0.336	0.772	151.2	0.072	4.43	0.207
13	DreamOmni2 (Xia et al., 2025)	11.51	0.154	0.210	0.534	0.835	0.382	0.661	153.4	0.054	3.74	0.125
14	Step-1x (Liu et al., 2025)	11.45	0.190	0.182	0.534	0.784	0.375	0.682	185.2	0.082	3.35	0.100
15	Step-1x (think+reflect) (Liu et al., 2025)	11.12	0.191	0.189	0.529	0.787	0.401	0.662	168.6	0.061	3.13	0.100
16	Bagel (Deng et al., 2025)	11.18	0.145	0.178	0.547	0.865	0.403	0.679	174.0	0.083	4.02	0.109
17	Step-1x (think) (Liu et al., 2025)	11.29	0.203	0.203	0.526	0.777	0.408	0.671	179.3	0.073	2.84	0.111
18	Bagel (think) (Deng et al., 2025)	11.09	0.169	0.205	0.551	0.860	0.423	0.670	174.1	0.074	3.54	0.131
19	ICEdit (MoE) (Zhang et al., 2025)	11.38	0.142	0.192	0.557	0.767	0.433	0.661	198.6	0.093	2.76	0.098
20	ICEdit (Zhang et al., 2025)	11.68	0.165	0.197	0.533	0.802	0.476	0.638	206.7	0.098	2.61	0.074
21	OmniGen (Xiao et al., 2025)	8.74	0.202	0.222	0.363	0.936	0.605	0.588	291.1	0.192	1.57	0.024

$\mathrm{mIoU}^{\star}$, uses intersection $\min(n_i, \hat{n}_i)$. We report the *placement gain* $\Delta = \mathrm{mIoU} - \mathrm{mIoU}_0$, the layout agreement that class proportions alone cannot explain.

Segmenter-free diagnostics. Four further measures use only pixels on a common grid, with no learned model. *Boundary-edge AUC* asks whether image edges follow the map. It ranks pixels by Sobel gradient magnitude and scores how well they separate map boundaries from region interiors, where 0.5 means no alignment. *Texture energy* is the share of spectral energy above 64 cycles per image, expressed as a $\log_2$ ratio to the real crop. *Cross-zoom consistency* correlates the 16×16 luminance layout of a model's crop with the matching window of its own 5 km output. *NCM mIoU* checks whether regions look like their assigned land cover. It applies the minimum-distance-to-means classifier (Richards & Jia, 2006), also known as nearest-class-mean (Mensink et al., 2013), with class means of CIELAB colour and local contrast fitted on the real crop. Generated pixels take the nearest class, and the score is reported as a fraction of the same classifier's FG-mIoU on the real crop.

4 Experiments

4.1 Experimental Setup

We evaluate 21 model configurations, including inference-mode variants. The ten closed-source configurations are GPT-Image-1.5 and GPT-Image-2 (OpenAI, 2025; 2026), Nano Banana, Nano Banana 2 and Pro (Google, 2025a; 2026; 2025b), Seedream 4.0, 4.5 and 5.0 (Team Seedream et al., 2025; Seedream, 2025; 2026), and FLUX.2 Max and Pro (Black Forest Labs, 2025). The eleven open-weight configurations are FLUX.2 dev (Black Forest Labs, 2025), Qwen-Edit-2511 (Wu et al., 2025), DreamOmni2 (Xia et al., 2025), Step-1x with its think and think+reflect variants (Liu et al., 2025), Bagel with its think variant (Deng et al., 2025), ICEdit and ICEdit-MoE (Zhang et al., 2025), and OmniGen (Xiao et al., 2025). Closed-source models are accessed through official APIs, and open-weight models run locally on NVIDIA H200 GPUs with default inference settings. For each scene–scale instance, all models receive the same map and frozen prompt, giving $100 \times 4 \times 21 = 8{,}400$ generations. Intervals are 95% scene-bootstrap percentiles over $20{,}000$ resamples.

4.2 Main Results

Table 1 reports all configurations. Layout fidelity spans an order of magnitude, from 0.327 FG-mIoU for Seedream 5.0 to 0.024 for OmniGen, and the best open-weight model, FLUX.2 dev at 0.243, matches Seedream 4.0. Appearance and layout rankings disagree. GPT-Image-2 leads on five appearance metrics, including CLIP-I at 0.833, but not on mIoU. FLUX.2 Max and Pro reach

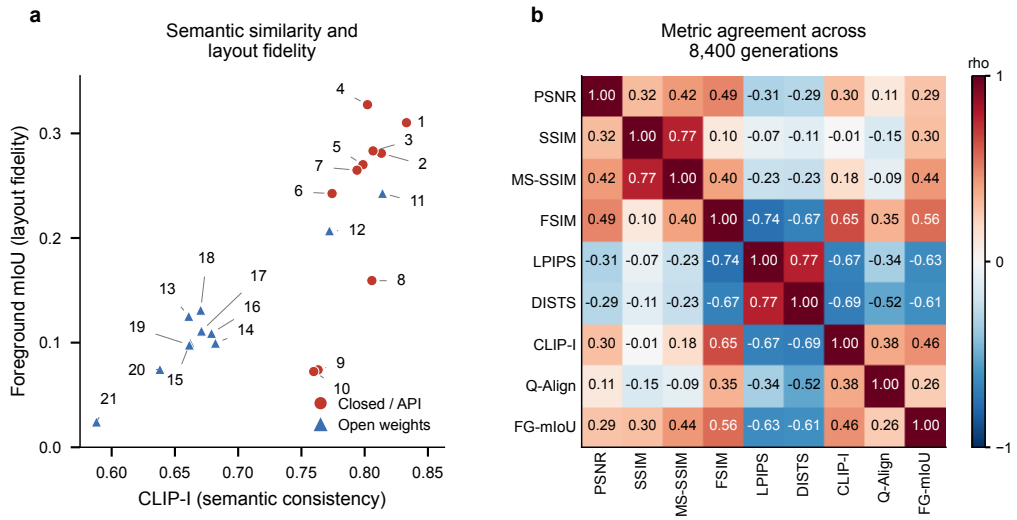


Figure 5: **(a)** Semantic similarity (CLIP-I) and segmentation-based layout fidelity (foreground mIoU), with one mean per model configuration. Similar CLIP-I scores can accompany different mIoU scores. **(b)** Per-sample Spearman correlations among nine metrics over 8,400 generations. FID and KID are per-model measures and are omitted. LPIPS and DISTS are lower-is-better. Model IDs in (a) refer to Table 1.

Table 2: Metrics by physical coverage, pooled over 21 configurations and 100 scenes, with 2,100 outputs per row. As the footprint widens, CLIP-I, PSNR, MS-SSIM and FSIM rise while FG-mIoU falls by 43%.

Coverage	PSNR↑	SSIM↑	MS-SSIM↑	FSIM↑	LPIPS↓	DISTS↓	CLIP-I↑	Q-Align↑	FG-mIoU↑
0.625×0.625 km	11.22	0.168	0.191	0.556	0.723	0.355	0.714	3.75	0.216
1.25×1.25 km	11.40	0.158	0.199	0.571	0.719	0.357	0.726	3.68	0.193
2.5×2.5 km	11.81	0.161	0.212	0.587	0.711	0.356	0.749	3.73	0.156
5.0×5.0 km	12.27	0.176	0.229	0.598	0.727	0.357	0.754	3.78	0.122
Spearman ρ	+0.23	+0.00	+0.12	+0.21	+0.01	+0.01	+0.18	-0.04	-0.23

CLIP-I near 0.76 and the highest Q-Align, yet score below 0.08 mIoU. Figure 5a makes the gap explicit. GPT-Image-1.5 and Nano Banana 2 are tied on CLIP-I at 0.806 and 0.807 but differ by $1.8\times$ on mIoU. Explicit reasoning does not reliably help. The think variant of Bagel gains 0.022 mIoU, whereas Step-1x reflection leaves it unchanged at 0.100.

4.3 The Coverage Paradox

Widening the footprint moves image similarity and layout fidelity in opposite directions, as shown in Table 2 and Figure 6. From 0.625 to 5 km, pooled CLIP-I rises from 0.714 to 0.754, a change of $+0.041$ with CI $[0.033, 0.048]$. Over the same range FG-mIoU falls from 0.216 to 0.122, a change of -0.094 with CI $[-0.105, -0.083]$. **All 21 configurations show both changes**, and all 21 have a positive CLIP-I slope and a negative mIoU slope against $\log_2$ coverage. The paradox also affects the strongest model. GPT-Image-2 gains 0.060 CLIP-I as its mIoU halves from 0.394 to 0.196. PSNR, MS-SSIM and FSIM follow CLIP-I, while SSIM, LPIPS and DISTS stay flat with $|\rho| \leq 0.01$. The most widely used similarity measures therefore rate wide-coverage outputs higher even as their layout agreement falls, a pattern visible for every configuration in Figure 6c,d.

The paradox is not a pooling artifact. Pooled over all generations, CLIP-I and mIoU are positively correlated, as the metric correlations in Figure 5b show. Within each coverage level they correlate at $\rho = 0.52$–0.55, and across the 21 model means at $\rho = 0.80$. Better models improve both axes, yet within every model, coverage drives the two axes apart. The following analyses diagnose which part of the condition models stop following, using the measures in Figure 7.

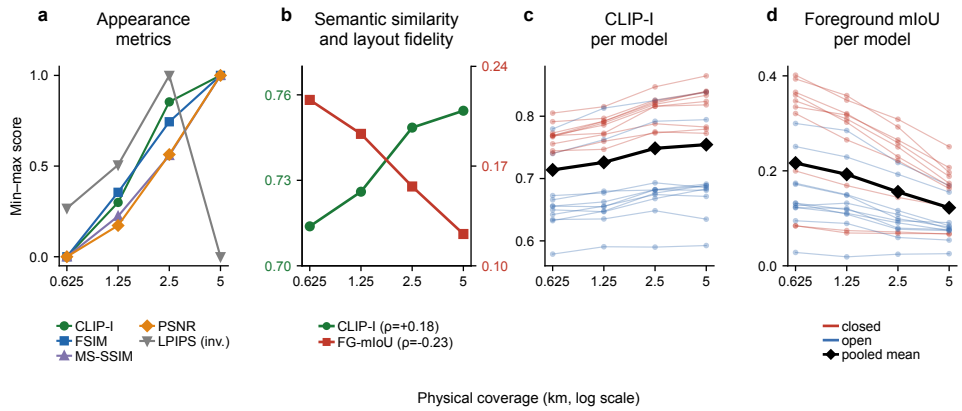


Figure 6: The coverage paradox. **(a)** Selected appearance metrics, min–max normalized across coverage (LPIPS inverted so higher is better). CLIP-I, FSIM, MS-SSIM and PSNR rise with coverage, while LPIPS does not. **(b)** Pooled CLIP-I rises while FG-mIoU falls. **(c,d)** The same opposing trends hold for all 21 configurations (thin lines) around the pooled mean (black).

4.4 Placement Collapses toward Chance

Part of FG-mIoU is earned by class proportions alone. The chance score mIoU_0 stays nearly constant, moving from 0.056 to 0.051, so the decline occurs almost entirely in the placement gain shown in Figure 7a. The gain falls from 0.160 to 0.071, a 56% loss with CI $[-0.099, -0.079]$ on the change, compared with 43% for raw mIoU. It declines in 20 of 21 configurations. At 5 km, FLUX.2 Max, FLUX.2 Pro and OmniGen fall to within 0.02 of chance, and 16.7% of all outputs score at or below chance, up from 12.0%. Composition overlap also falls from 0.60 to 0.50, but the placement gain falls faster than the composition-limited ceiling $\mathrm{mIoU}^\star$, which moves from 0.40 to 0.28.

4.5 Geometry Survives, Semantics Do Not

A falling mIoU could mean that the generated geometry has dissolved, or that the segmenter misreads it. Boundary-edge AUC separates these cases without the segmenter. Generated edges align with the map's boundaries *better* at wide coverage. As Figure 7b shows, AUC rises from 0.622 to 0.656 with CI $[0.024, 0.044]$ in 17 of 21 configurations, closing on the real image at the same coverage, which moves from 0.682 to 0.709. The measure is consistent with layout fidelity. It correlates positively with mIoU in 98% of model–coverage cells and at $\rho = 0.68$ across model means.

As coverage grows, the lines are kept while the regions between them lose their assigned category. For target pixels present at both endpoints, building IoU falls from 0.286 to 0.103 and road IoU from 0.316 to 0.120, while cropland IoU *rises* from 0.163 to 0.223. For GPT-Image-2, the share of target-building pixels decoded as cropland rises from 7.8% to 54.4%. Qwen-Edit rises from 21.7% to 46.3% and Seedream 5.0 from 7.6% to 29.2%. In one Argentine city at 5 km, 98.7% of building pixels are decoded as cropland despite a visible street grid. NCM mIoU confirms this without a learned model. Measured against the real image at the same scale, generated land cover is recognisable at only 41% of the real level at 0.625 km, falling to 34% at 5 km. The fraction declines in all 16 configurations with stored outputs for this measure, and it ranks models almost exactly as the segmenter does, with $\rho = 0.96$ across model means. Across scales, generators therefore preserve the *geometry* of the condition but not the land-cover *identity* of its regions.

4.6 Scale Is Rendered as Texture

Real imagery gains fine detail slowly as coverage grows. Its high-frequency energy increases by a factor of 1.9 from 0.625 to 5 km and levels off beyond 1.25 km. Generated imagery behaves differently, as Figure 7c shows. All 21 configurations are too smooth close up, at a pooled $2^{-1.45} \approx 0.37\times$ the real level, and all 21 add detail faster than real imagery, with a median factor of 3.6 over the same range. Models thus respond to coverage with texture density rather than with more, correctly placed structures. A prompt-ablation subset of 8 scenes repeats the generation for 15 configurations without the ground-extent sentence, and the ramp persists. On the four scenes with stored reference

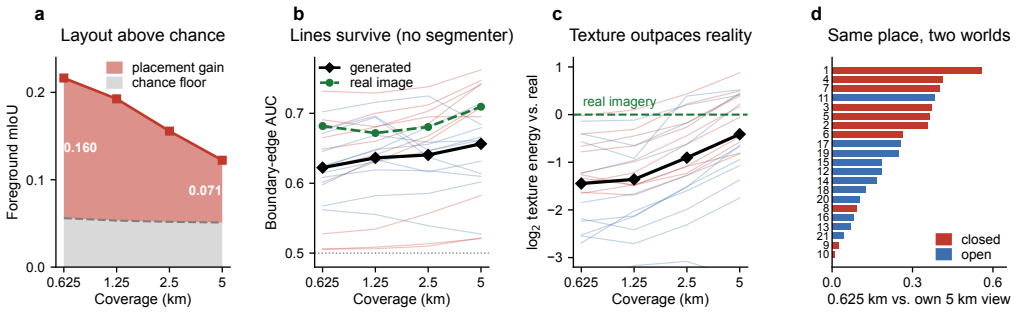


Figure 7: Diagnosing the coverage paradox. **(a)** FG-mIoU split into the chance floor and the placement gain above it. **(b)** Boundary-edge AUC from pixels alone, with the real image in green. **(c)** Fine-texture energy relative to the real crop. **(d)** Correlation of each configuration's 0.625 km output with the matching window of its own 5 km output, with IDs from Table 1. Thin lines show individual configurations, red for closed and blue for open.

imagery, the mean $\log_2$ texture slope per coverage doubling is 0.42 without the sentence, against 0.28 for the real crops. Within scenes, the ramp follows the boundary density of the input map as closely as the coverage level itself. Scale is therefore read mainly from how busy the map looks, and the stated extent is not needed to produce the ramp.

4.7 Cross-Scale Consistency

Because the 5 km view contains every crop, we can ask whether a model draws the same place twice. The mean luminance correlation between a 0.625 km output and the matching window of the model's own 5 km output is 0.22, as Figure 7d shows. It reaches 0.56 for GPT-Image-2 and falls to at most 0.04 for FLUX.2 Max, FLUX.2 Pro and OmniGen. The prompt-ablation subset also varies the scene description at a fixed map. Across its 8 scenes and 15 configurations, images of the same crop correlate at 0.44 under three differently worded prompts, but only at 0.29 across zoom levels. The generators treat each footprint as a new picture, not as a new view of one location.

Together, the chance correction and the four segmenter-free diagnostics explain the paradox. Wide-coverage outputs gain CLIP-I and PSNR because their textures and global palette look like wide-area imagery, while their fine land-cover semantics and placement degrade. Appearance metrics, including human-aligned ones such as Q-Align with coverage $\rho = -0.04$, therefore cannot detect the failure. It becomes visible only when scale is varied within a scene, the chance level is removed, and the diagnosis is confirmed without the segmenter. SatLen provides this protocol.

5 Conclusion

We present **SatLen**, a benchmark that evaluates whether image generators understand physical scale in map-conditioned satellite generation. It stems from the core observation that a satellite image carries a physical scale: the same place must remain the same place whether it is seen from 0.625 or 5 km, so proper evaluation must extend beyond image realism at a single scale. By rendering 100 globally sampled scenes at four coverage levels and scoring $8{,}400$ generations from 21 model configurations with chance-corrected layout scores and segmenter-free pixel probes, SatLen measures scale understanding rather than appearance alone. The evaluation reveals a coverage paradox in every evaluated model: as the footprint widens, outputs look more like real imagery yet follow the requested layout less closely, with spatial placement falling by 56% beyond chance. Current generators keep the map's geometry but lose its land-cover semantics, and render scale as texture rather than as geography. Ultimately, we hope SatLen guides future research toward generators that are geographically faithful across scales, rather than merely realistic at any single one.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic benchmark images and for generating the one-sentence scene descriptions used in the benchmark prompts. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in writing proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, translation, cleaning or reformatting datasets, qualitative or thematic data analysis, or interpreting results; proof-related tasks are not applicable to this work. Additionally, we used generative AI tools for editing and polishing the paper to improve readability. We have reviewed all AI-assisted work. We checked the generated scene descriptions and benchmark images for consistency with the specified maps, prompts, and experiment protocol, and checked the AI-assisted edits for accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. In *International Conference on Learning Representations*, 2018.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiao-Ping Nie, Ziang Song, Shi Guang, and Haoqi Fan. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Keyan Ding, Kede Ma, Shiqi Wang, and Eero P Simoncelli. Image quality assessment: Unifying structure and texture similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(5):2567–2581, 2022. doi: 10.1109/TPAMI.2020.3045810.

Miguel Espinosa and Elliot J Crowley. Generate your own scotland: Satellite image generation conditioned on maps. *arXiv preprint arXiv:2308.16648*, 2023.

Tom G Farr, Paul A Rosen, Edward Caro, Robert Crippen, Riley Duren, Scott Hensley, Michael Kobrick, Mimi Paller, Ernesto Rodriguez, Ladislav Roth, David Seal, Scott Shaffer, Joanne Shimada, Jeffrey Umland, Marian Werner, Michael Oskin, Douglas Burbank, and Douglas Alsdorf. The shuttle radar topography mission. *Reviews of Geophysics*, 45(2), 2007. doi: 10.1029/2005RG000183.

Google. Introducing gemini 2.5 flash image (nano banana). https://developers.googleblog.com/en/introducing-gemini-2-5-flash-image/, 2025a.

Google. Introducing nano banana pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025b.

Google. Nano banana 2: Combining pro capabilities with lightning-fast speed. https://blog.google/innovation-and-ai/technology/ai/nano-banana-2/, 2026.

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. In *Advances in Neural Information Processing Systems*, 2017.

Robert J Hijmans, Susan E Cameron, Juan L Parra, Peter G Jones, and Andy Jarvis. Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, 25(15):1965–1978, 2005.

Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Cihang Xie, and Yuyin Zhou. Hq-edit: A high-quality dataset for instruction-based image editing. In *International Conference on Learning Representations*, 2025.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1125–1134, 2017.

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*, 2024.

Max Ku, Tianle Li, Kai Zhang, Yujie Lu, Xingyu Fu, Wenwen Zhuang, and Wenhu Chen. Imagenhub: Standardizing the evaluation of conditional image generation models. In *International Conference on Learning Representations*, 2024.

Matti Kummu, Maija Taka, and Joseph HA Guillaume. Gridded global datasets for gross domestic product and human development index over 1990–2015. *Scientific Data*, 5:180004, 2018. doi: 10.1038/sdata.2018.4.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. SegEarth-OV3: Exploring SAM 3 for open-vocabulary semantic segmentation in remote sensing images. *arXiv preprint arXiv:2512.08730*, 2025. URL https://arxiv.org/abs/2512.08730.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, Yibo Zhu, Binxing Jiao, Xiangyu Zhang, Gang Yu, and Daxin Jiang. Step1x-edit: A practical framework for general image editing. *arXiv preprint arXiv:2504.17761*, 2025.

Thomas Mensink, Jakob Verbeek, Florent Perronnin, and Gabriela Csurka. Distance-based image classification: Generalizing to new classes at near-zero cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2624–2637, 2013. doi: 10.1109/TPAMI.2013.83.

OpenAI. GPT-Image-1.5. https://openai.com/index/new-chatgpt-images-is-here/, 2025.

OpenAI. GPT-Image-2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026.

Overture Maps Foundation. Overture maps foundation data. https://overturemaps.org, 2024.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2337–2346, 2019.

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*, pp. 8748–8763, 2021.

John A. Richards and Xiuping Jia. *Remote Sensing Digital Image Analysis: An Introduction*. Springer, Berlin, Heidelberg, 4 edition, 2006. doi: 10.1007/3-540-29711-1.

Miguel O Román, Zhuosen Wang, Qingsong Sun, Virginia Kalb, Steven D Miller, et al. NASA's black marble nighttime lights product suite. *Remote Sensing of Environment*, 210:113–143, 2018.

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*, pp. 10684–10695, 2022.

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 Workshops*, pp. 460–470, 2024.

Marcello Schiavina, Michele Melchiorri, Martino Pesaresi, Panagiotis Politis, Sergio Manuel Carneiro Freire, Luca Maffenini, Pietro Florio, Daniele Ehrlich, Katarzyna Goch, Alessandra Carioli, Johannes Uhl, Pierpaolo Tommasi, and Thomas Kemper. GHSL data package 2023. Technical Report JRC133256, Publications Office of the European Union, 2023.

Seedream. Seedream 4.5. `https://seed.bytedance.com/en/seedream4_5/`, 2025.

Seedream. Seedream 5.0 Lite. `https://seed.bytedance.com/en/seedream5_0_lite/`, 2026.

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. doi: 10.1109/TGRS.2024.3453414.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8798–8807, 2018.

Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multiscale structural similarity for image quality assessment. In *The Thirty-Seventh Asilomar Conference on Signals, Systems & Computers*, volume 2, pp. 1398–1402, 2003. doi: 10.1109/ACSSC.2003.1292216.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. In *International Conference on Machine Learning*, 2024.

Bin Xia, Bohao Peng, Yuechen Zhang, Junjia Huang, Jiyang Liu, Jingyao Li, Haoru Tan, Sitong Wu, Chengyao Wang, Yitong Wang, Xinglong Wu, Bei Yu, and Jiaya Jia. DreamOmni2: Multimodal instruction-based editing and generation. *arXiv preprint arXiv:2510.06679*, 2025. URL `https://arxiv.org/abs/2510.06679`.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023.

Shitao Xiao, Yueze Wang, Junjie Zhou, Huaying Yuan, Xingrun Xing, Ruiran Yan, Shuting Wang, Tiejun Huang, and Zheng Liu. Omnigen: Unified image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13294–13304, 2025.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. Imgedit: A unified image editing dataset and benchmark. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025.

Daniele Zanaga, Ruben Van De Kerchove, Dirk Daems, Wanda De Keersmaecker, Carsten Brockmann, et al. Esa worldcover 10 m 2021 v200. *Zenodo*, 2022. doi: 10.5281/zenodo.7254221.

Lin Zhang, Lei Zhang, Xuanqin Mou, and David Zhang. Fsim: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386, 2011.

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*, pp. 3836–3847, 2023.

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*, pp. 586–595, 2018.

Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, and Yi Yang. In-context edit: Enabling instructional image editing with in-context generation in large scale diffusion transformer. In *Advances in Neural Information Processing Systems*, 2025.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.