VectorSynth: Fine-Grained Satellite Image Synthesis with Structured Semantics

Daniel Cher*   Brian Wei*   Srikumar Sastry   Nathan Jacobs

Washington University in St. Louis

{cher, b.j.wei, s.sastry, jacobsn}@wustl.edu

*Equal contribution

Figure 1. VectorSynth logo synthesized using learned OSM-based pixel embeddings. Each letter is generated with distinct OSM tag combinations (e.g. industrial, farmland, geological features), demonstrating VectorSynth's fine-grained semantic control over satellite-image synthesis.

Abstract

We introduce VectorSynth, a diffusion-based framework for pixel-accurate satellite image synthesis conditioned on polygonal geographic annotations with semantic attributes. Unlike prior text- or layout-conditioned models, VectorSynth learns dense cross-modal correspondences that align imagery and semantic vector geometry, enabling fine-grained, spatially grounded edits. A vision language alignment module produces pixel-level embeddings from polygon semantics; these embeddings guide a conditional image generation framework to respect both spatial extents and semantic cues. VectorSynth supports interactive workflows that mix language prompts with geometry-aware conditioning, allowing rapid what-if simulations, spatial edits, and map-informed content generation. For training and evaluation, we assemble a collection of satellite scenes paired with pixel-registered polygon annotations spanning diverse urban scenes with both built and natural features. We observe strong improvements over prior methods in semantic fidelity and structural realism, and show that our trained vision language model demonstrates fine-grained spatial grounding. The code and data are available at https://github.com/mvrl/VectorSynth.
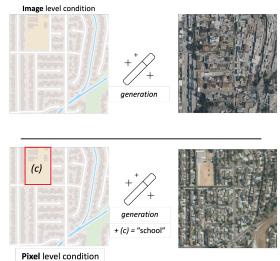
Figure 2. VectorSynth enables precise, fine-grained control over both spatial location and semantic content during satellite image synthesis. top: Image-level conditioning lacks the ability to target specific regions or object types. bottom: VectorSynth allows conditioning on individual polygons with semantic labels (e.g., c = "school"), producing coherent imagery that respects both spatial extent and semantics.

1. Introduction

Text-to-image generative models have witnessed rapid progress in recent years, driven by advances in large-scale vision-language pretraining. Models such as DALL·E 2 [34] and Stable Diffusion [35] demonstrate the remarkable ability to synthesize images from natural language prompts, ranging from surreal abstractions to photo-realistic renderings. Beyond raw generation, these models enable downstream tasks such as inpainting [30], layout-to-image translation [26], and text-guided image editing [4, 18]. Architectures like ControlNet [48] extend this further by introducing a framework to adapt these large pre-trained diffusion models for domain-specific applications.

Recently, the remote sensing community has begun exploring generative models for various Earth observation

tasks, including disaster response, environmental monitoring, poverty estimation, and urban planning, which have shown to benefit from generated satellite imagery [10, 13, 31, 32]. Diffusion-based approaches, such as DiffUCD [50] and HySCDG [3], have shown that using an appropriate conditioning technique yields rich synthetic satellite data that can support downstream remote sensing tasks.

However, existing approaches for satellite image synthesis [3, 36, 50] rely on coarse-grained semantic supervision, typically using OpenStreetMap (OSM) raster tiles and/or global text prompts. Although OSM tiles are abundant and visually interpretable, they lack semantic depth. Distinct object classes, such as residential buildings, hospitals, or schools, are often depicted similarly in OSM stylings, failing to capture these semantic properties. In addition, image-level text supervision limits fine-grained control and does not allow more specific region-level editing.

To enable more expressive and semantically grounded image synthesis, we propose shifting from coarse, image-level conditioning to fine-grained, local-level representations. This shift enables users to specify distinct textual prompts for specific regions of an image, allowing for flexible editing, richer abstraction, and precise semantic control. However, achieving this level of generation requires a model that can accurately align textual descriptions with corresponding spatial regions, such as polygons or pixel masks. As illustrated in Figure 2, traditional image-level conditioning lacks fine-grained control, producing globally plausible imagery, but failing to reflect localized semantics. In contrast, our proposed approach enables fine-grained control by conditioning the synthesis process on sub-image annotations with detailed semantics.

OpenStreetMap (OSM), with its vast and growing repository of structured geographic annotations, is an ideal source of semantic grounding for spatial reasoning tasks. While recent vision-language models (VLMs) such as RemoteSAM [45] and RemoteCLIP [27] demonstrate pixel-level grounding capabilities, their vocabularies are typically constrained to general object categories (e.g., 'car', 'road') and lack alignment with the rich, structured, and domain-specific taxonomy used in OSM. These models are not trained to handle compositional or hierarchical tag structures (e.g., 'building residential', 'shop retail') that are common in OSM, and thus fall short in tasks that require detailed semantic understanding of geographic features. To address this gap, we propose learning fine-grained alignment between satellite imagery and OSM-style textual descriptions at the polygon level. Datasets like SkyScript [43] offer global tag supervision, but do not include the vector geometries necessary for region-specific grounding.

To this end, we propose a framework that enables local-level alignment between satellite imagery and OSM-based semantic descriptions. By grounding image synthesis at the polygon level, we allow for precise spatial control over generative models, enabling composition, editing, and abstraction beyond what is possible with coarse-level supervision.

Key Contributions: We introduce a framework for pixel-level semantic control of satellite image synthesis. The contributions of our work are threefold:

1. **COSA: Contrastive OSM-Satellite Alignment Vision-Language Model.** A model trained to align OSM tag descriptions and satellite imagery through polygon level contrastive learning.
2. **VectorSynth: Text-to-Image Generation with Pixel-Level Control.** A synthesis pipeline that enables compositional, fine-grained generation from multiple textual prompts, controlling content at the pixel-level.
3. **OSM-Polygon Dataset.** A novel dataset that aligns satellite images with OSM polygon-level tags, allowing for fine-grained grounding of semantic regions.

2. Related Work

Fine-grained Contrastive Learning. Previous works have extended global vision-language models [16, 24, 33] to capture token-level alignment of image and text for improved fine-grained understanding. RegionCLIP [51] uses pre-trained CLIP to label region-text pairs, guiding contrastive learning. LOUPE [23] generates semantic regions and performs region-text alignment. MaskCLIP [52] leverages CLIP's spatial tokens for dense prediction masks. Subsequent methods [17, 22, 47] advance these approaches for tasks like open-vocabulary semantic segmentation.

However, these CLIP-based image encoders have inherently low latent resolutions that require upsampling for dense prediction tasks, limiting their capability in capturing fine-grained details [29]. In complex environments like urban remote sensing imagery, detailed information is lost when using these low-resolution latent models [21, 37]. FeatUp [8] addresses this with a learnable feature upsampler. Applied to MaskCLIP, the model can capture more fine-grained text-image alignment.

While these approaches have better fine-grained text-image alignment, they keep the text encoder frozen to retain the benefits of CLIP pretraining. This poses limitations in the remote sensing domain, where textual semantics are often highly correlated. For example, 'building height 5m' and 'building height 30m' are linguistically similar, but may refer to visually distinct structures like a small house and a high-rise apartment, respectively. RemoteCLIP [27] attempts to address this by training both the image and text encoders for vision-language alignment in the remote sensing domain. Other approaches such as Sat2Cap [7] align satellite imagery with ground-level imagery to improve fine-grained understanding. However, all such models still have low-resolution latent image features.

In this work, we propose a contrastive learning framework that jointly trains a high-resolution image encoder and a text encoder useful for fine-grained remote sensing synthesis.

Satellite Image Synthesis. Recent advances in satellite image synthesis [19, 46] have shown promise across a range of remote sensing applications, including change detection [3], cloud removal [41], and synthetic data generation for discriminative tasks [38]. These approaches often rely on either modality-to-modality translation (e.g., SAR-to-optical [1]) or style-conditioned generation using simplified semantic inputs. GeoSynth [36] is a notable work using semantic information for satellite image generation. It conditions a ControlNet [48] based diffusion model on Open-StreetMap (OSM) tile images that serve as a proxy for objects and land use structure. However, OSM stylings, while visually intuitive, are limited in both semantic depth and compositional control. They compress diverse geographic information into a fixed set of hand-crafted visual representations, which cannot easily capture multi-label, hierarchical, or region-specific semantics. Our work seeks to move beyond fixed visual stylings by leveraging the rich, structured tag data available in OSM. Rather than treating the OSM input as a 2D styled image, we encode raw OSM tag sets directly into the *text space*, using vision-language models (VLMs) to establish semantically grounded and compositional controls. Our method opens up a new avenue for semantic satellite image synthesis by embedding structured geographic knowledge into a generative language-driven pipeline.

3. Dataset

To achieve fine-grained conditioning, as seen in Figure 2, we construct a dataset coupling polygon-level OSM vector data with high-resolution satellite imagery. This enables fine-grained local alignment and compositional semantics for precise control over generated content.

To capture a wide range of urban layouts, we focus our data collection on five major cities. During training, we sample from Los Angeles, New York City, Paris, and Berlin. Each training city is split into spatial blocks and divided 60/20/20 into train/validation/test dataset splits. To assess the model's ability to generalize beyond these known contexts, we designate Chicago as a holdout city to evaluate out-of-distribution performance. High-resolution satellite imagery is sourced from the Mapbox Static Tiles API[1] at zoom level 16, corresponding to a spatial resolution of approximately 0.6 meters per pixel. Each tile is 512×512 pixels, covering an area of roughly 300×300 meters, consistent with prior work [36] to support direct comparisons.

Vector data is obtained from GeoFabrik's OSM extracts [9]. To retain only semantically relevant features vis-

---

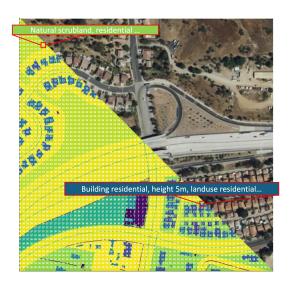[1] https://docs.mapbox.com/api/maps/static-tiles/

Figure 3. Illustration of pixel-level tag assignments. Each pixel inherits tags from overlapping polygons, resulting in compositional tag lists used for downstream learning tasks.

ible in overhead imagery, we filter out point geometries, remove rare tags ($< 0.2\%$ of tiles), and keep only tiles with $\geq 70\%$ vector feature coverage. To further enhance structural diversity, we incorporate building height data from GlobalFootprintsLM [39], enriching the representation of vertical variation across scenes.

To represent semantic content at the pixel level, as seen in Figure 3, we render the filtered vector features by collecting tags from overlapping annotations per pixel. Each pixel gets a multi-tag composition, and nearby pixels with identical compositions form polygon instances. For example, a single pixel might inherit tags such as ['building residential', 'place island', 'height 6m'], reflecting multiple overlapping semantic layers.

We illustrate the richness and granularity of the tag annotations by visualizing multi-tag composition overlays on a sample tile shown in Figure 3. The overlay reveals dense and semantically consistent tagging across spatial structures, such as roads, residential blocks, and natural features, highlighting the high quality and compositional expressiveness of our annotations.

We also render all tiles using custom Mapbox styles to generate stylized OSM maps to conduct consistent evaluation against previous work [36]. Finally, we caption each satellite image using LLaVA [28], a multimodal vision language model. The prompt used for captioning is: 'Describe the contents of the image'. The final OSM-Satellite dataset includes approximately 1,000 unique OSM tags and over 400,000 unique tag combinations associated with individual pixels. We generate around 20,000 image tiles, each paired with satellite imagery, polygon-level vector annotations, pixel-level semantic masks, global satellite descrip-
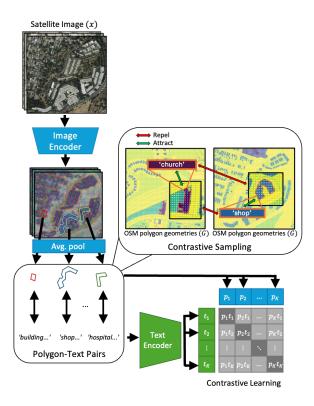
Figure 4. Architecture overview showing dual encoders for satellite imagery and OSM tag descriptions, with polygon-guided average pooling to extract region-specific embeddings. We align polygon embeddings with grounded OSM tags, enabling fine-grained spatial conditioning for satellite image synthesis.

tions and corresponding OSM stylizations. This dataset serves as the foundation for pixel-level contrastive learning and image synthesis tasks.

4. Methodology

In this section, we describe our proposed approach to build a semantically aligned pixel-level vision-language model and a fine-grained satellite image synthesis framework.

4.1. Polygon-Level Contrastive Learning

We learn a textual embedding space for OSM tags aligned with satellite imagery to enable fine-grained conditioning during synthesis. As described in Section 3, polygons define spatial units associated with multi-tag compositions in the image. We apply contrastive learning to pull closer the embeddings of aligned polygon–tag composition pairs (e.g., a group of satellite pixels and its corresponding multi-tag composition), while pushing apart the embeddings of dissimilar pairs (e.g., polygons associated with different multi-tag compositions), as seen in Figure 4.

Previous satellite image-text contrastive learning frameworks [27] aim to enhance image encoders, so their representations more closely align with a pretrained text embedding space, such as CLIP. These models are often optimized for image retrieval or segmentation, but not text-guided generation. In contrast, we focus on improving the text encoder to align better with the image embedding space for the purpose of fine-grained generation.

To support this, we opt for dense pixel-level representations. Previous dense contrastive frameworks form contrastive pairs at the patch level [52] or use self-supervised spatial cues [5, 44], while our approach leverages vector polygon annotations with explicit tag labels from OSM. These polygons are directly aligned with image regions, which enables polygon-guided contrastive learning.

Architecture. Our model, **COSA**, is shown in Figure 4. The architecture consists of a learnable image encoder f_{img}, a learnable text encoder f_{text}, and a polygon-guided contrastive loss objective that aligns OSM tag compositions with corresponding polygon image features.

Let $x \in \mathbb{R}^{3 \times H \times W}$ denote a satellite image with height H and width W. Let $\mathcal{C} = \{c_1, \ldots, c_K\}$ be OSM multi-tag compositions corresponding to polygon geometries $G = \{g_1, \ldots, g_K\}$ in the image, where each multi-tag composition c_i is a sentence. The text and image encoders process c_i and x respectively to produce a corresponding text embedding e and dense image embeddings $z_{\text{img}} \in \mathbb{R}^{D \times H' \times W'}$, where D is the embedding dimension and (H', W') is the spatial resolution of the image feature map.

Polygon-Guided Contrastive Loss. We aim to generate polygon-text contrastive pairs for training. For each polygon geometry g_i in the satellite image x, let $M_{g_i} \in \{0,1\}^{H' \times W'}$ be a binary mask indicating the spatial extent of the polygon on the image feature map. We obtain this binary mask by interpolating and thresholding from the resolution of the image features. The polygon embedding $p_i \in \mathbb{R}^D$ is computed by average pooling z_{img} over the masked region:

$$p_i = \frac{1}{\sum M_{g_i}(h, w)} \sum_{h=1}^{H'} \sum_{w=1}^{W'} M_{g_i}(h, w) \cdot z_{\text{img}}[:, h, w] \quad (1)$$

Given K polygon-text pairs $\{(p_i, e_i)\}_{i=1}^{K}$, we use a symmetric InfoNCE loss which is defined as follows:

$$\mathcal{L}_{p,e} = -\frac{1}{2K} \sum_{i=1}^{K} \left[\log \frac{\exp\left(\text{sim}(p_i, e_i)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}(p_i, e_j)/\tau\right)} \right.$$
$$\left. + \log \frac{\exp\left(\text{sim}(e_i, p_i)/\tau\right)}{\sum_{j=1}^{K} \exp\left(\text{sim}(e_i, p_j)/\tau\right)} \right] \quad (2)$$

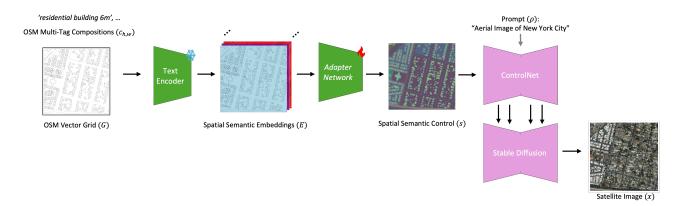where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a learnable temperature parameter.

Figure 5. This figure presents our semantic-guided image synthesis pipeline which employs a pretrained text encoder to generate dense pixel-level control from input vector geometry.

4.2. Image Synthesis

Architecture. We train conditional generative models to synthesize satellite images x given an image-level natural language text description ρ and spatial semantic control s derived from OpenStreetMap (OSM). Specifically, we optimize a latent diffusion model to approximate the conditional distribution $p(x \mid \rho, s)$.

We build upon the ControlNet [48] architecture, which extends pre-trained diffusion models by incorporating additional control inputs. ControlNet consists of a trainable copy of the encoding layers of the base diffusion model, connected via zero-initialized convolution layers. This design preserves the original model's capabilities while enabling additional control. The control branch processes the conditioning information at multiple scales and feeds it into the main U-Net through residual connections.

Figure 5 illustrates our complete pipeline. The generation is guided by two controls: a global text prompt ρ and a spatial semantic control s.

We derive s from OSM vector geometries. First, we render the OSM data into a grid $G \in \mathcal{C}^{H \times W}$, where each pixel (h, w) contains a multi-tag composition $c_{h,w} \in \mathcal{C}$. Next, we encode each composition with a text encoder T, producing spatial semantic embeddings $E \in \mathbb{R}^{D \times H \times W}$:

$$E[h, w] = T(c_{h,w}) \quad (3)$$

To align E with ControlNet, we pass it through a lightweight adapter $\mathcal{A}$, yielding a 3-channel raster:

$$s = \mathcal{A}(E) \in \mathbb{R}^{3 \times H \times W} \quad (4)$$

The spatial semantic control s provides both layout guidance and fine-grained OSM tag information, while the text prompt ρ provides global context.

Finally, ControlNet conditions the diffusion process on both s and ρ, training the denoising network ϵ_θ according

to the objective:

$$\mathcal{L} = \mathbb{E}_{z_0, s, \rho, \epsilon} \left[\|\epsilon - \epsilon_\theta(z_t, s, \rho)\|_2^2 \right] \quad (5)$$

where z_t is the noisy latent representation at diffusion timestep t.

During inference, users render OSM multi-tag compositions into a pixel grid, which is encoded into the same control representation s. ControlNet then synthesizes images consistent with these pixel-wise semantics.

4.3. Implementation

For the COSA VLM, we use a SatlasNet [2] backbone with a learnable MLP adapter network as an image encoder. We use CLIP as our default learnable text encoder, but also experiment with BERT [6] and E5 [40], which are strong sentence level embedding models. For the image synthesis framework, we use Stable Diffusion v2.1 [35].

We precompute text embeddings of all OSM taglists in our dataset for training- and inference-time efficiency. We train six model variants, experimenting with different text encoders and adapter network architectures. For the adapter network, we experimented with deeper convolutional stacks and residual connections, but found that a 2D convolution followed by a sigmoid activation consistently yielded the best results. Following ControlNet [48], we apply random prompt masking. This encourages the model to leverage spatial semantics from the control image even when text is absent, improving robustness and generalization. Each model was trained on a single NVIDIA H100 GPU (80GB) for a total of 24 hours, using the Adam optimizer with a learning rate of $1e{-}5$ and a batch size of 8.

5. Results and Discussion

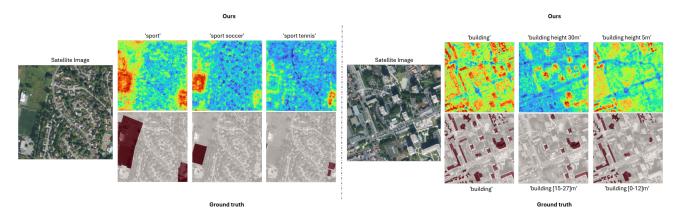We conduct extensive evaluations of VectorSynth, with ablations to assess contributions.

Figure 6. Similarity heatmaps given different text queries highlighting the fine-grained understanding of the proposed contrastive training. The bottom row shows the respective ground truth polygons. The alignment shows the ability of our approach to disentangle correlated semantic structures in OSM tag text.

VLM	B@1 ↑	Sem@20 ↑	R@5 ↑	R@10 ↑
CLIP	88.68	88.30	25.48	46.79
RemoteCLIP	87.83	87.39	26.58	51.73
Ours (E5)	89.86	90.37	21.14	48.14
Ours (BERT-base)	<u>90.11</u>	<u>91.64</u>	<u>29.49</u>	<u>53.10</u>
Ours (CLIP)	**91.61**	**92.47**	**32.40**	**54.06**

Table 1. Polygon-to-text retrieval results. We compare baseline VLMs with our COSA variants that use different text encoders. **Note:** B@1 is BERTScore@1, Sem@20 is Semantic-nDCG@20, R@K is Recall@K. Best results are **bold** and second best are <u>underlined</u>.

VLM	Parent Acc. ↑	Child Acc. ↑	Mixed F1 ↑
CLIP	43.16	29.41	0.171
RemoteCLIP	44.34	31.58	0.172
Ours	**82.84**	**77.09**	**0.272**

Table 2. OSM tag prediction results. 'Ours' denotes our COSA VLM with a CLIP text encoder. **Note:** Parent/Child accuracy measure correctness at broad vs. fine-grained tag levels, and Mixed F1 is their averaged F1. Best results are **bold**.

5.1. Cross-Modal Evaluation

We evaluate COSA, our contrastively trained vision-language model, on its ability to align polygon-level satellite imagery with OSM multi-tag composition. Performance is assessed through cross-modal retrieval and polygon-level tag prediction, testing fine-grained semantic grounding and generalization.

Cross-Modal Retrieval. We report BERTScore@1 [49] for top-1 semantic similarity, semantic nDCG@20 [15] for ranked semantic relevance, and Recall@5/10 for retrieval

accuracy. Table 1 reports polygon-to-text retrieval results on the test set. The VLM baselines, CLIP and RemoteCLIP, achieve reasonable performance, however our COSA model variants show consistent improvements across all metrics, with the CLIP-based text encoder achieving the highest BERTScore@1 (91.61), semantic nDCG@20 (92.47), Recall@5 (32.40), and Recall@10 (54.06). These gains highlight the importance of the choice of the text encoder. Furthermore, our contrastive training provides better alignment for fine-grained semantic retrieval compared to baseline VLMs.

Polygon-Level Tag Prediction. Beyond retrieval, we also evaluate our model's ability to directly predict OSM tags for polygons. We evaluate this task using three metrics: Parent accuracy, which measures correctness for parent tags (i.e. 'building') to capture broader semantic categories; Child accuracy, which measures correctness for child tags within a parent (i.e. 'apartments') to capture more fine-grained categories; Mixed F1 score, the average of parent- and child-level F1. Compared to CLIP and RemoteCLIP, COSA substantially improves across parent- and child-level accuracy, along with Mixed F1 score. This demonstrates that contrastive training not only enhances retrieval but also enables stronger tag prediction, opening the door to effective pseudo-labeling for sparsely annotated OSM regions.

We visualize the normalized cosine similarity matrix of tag embeddings in Figure 6 to illustrate the model's fine-grained semantic understanding. After contrastive training, embeddings capture more structured and meaningful relationships, particularly among closely related categories (e.g., 'sport tennis' vs. 'sport soccer'). These patterns reflect improved sensitivity to subtle distinctions in OSM tags, such as different types of sports facilities and buildings.

Model	Finetuned Text Encoder	In-Distribution Test			Out-of-Distribution Test		
		FID ↓	SSIM ↑	PSNR ↑	FID ↓	SSIM ↑	PSNR ↑
GeoSynth-OSM [36]	✗	95.30	0.16	9.92	108.33	0.15	9.89
Ours (VectorSynth)							
BERT	✗	52.16	0.15	12.59	72.15	0.10	10.91
E5	✗	46.71	0.15	12.92	63.66	0.12	11.24
CLIP	✗	33.07	0.20	14.04	45.13	0.16	12.03
RemoteCLIP	✗	48.75	0.14	12.89	61.54	0.11	11.58
Ours (VectorSynth + COSA)							
COSA-BERT	✓	40.61	0.15	12.66	67.72	0.11	11.07
COSA-E5	✓	57.06	0.17	13.27	68.22	0.14	11.69
COSA-CLIP	✓	**29.20**	**0.21**	**14.15**	**41.12**	**0.17**	**12.10**
Gain (%)		+69.36	+31.25	+42.64	+62.04	+13.33	+22.35

Table 3. Quantitative evaluation of different text encoders used in control generation. We show VectorSynth with and without contrastive OSM-Satellite alignment on in-distribution and out-of-distribution test sets used for tag conditions at the sub-image level. **Note**: All models have the same global text encoder for the SD2.1 base model.

Figure 7. Comparison of fine-grained semantic edits. Each set shows the local caption used for editing across models: (a) GeoSynth, (b) GeoSynth w/ Inpainting and (c) VectorSynth

5.2. Semantically Grounded Image Synthesis

We evaluate the quality and controllability of satellite image synthesis under various input representations and conditioning types. Specifically, we assess: (1) the impact of

different embedding sources for pixel-level control, and (2) fine-grained semantic editing capabilities.

Embedding Sources for Control. We evaluate three different approaches to synthesize satellite imagery conditioned on different representations of OpenStreetMap (OSM) control signals. The first framework, **GeoSynth-OSM**, follows the Geosynth-OSM baseline [36] and uses rasterized OSM tiles as direct pixel-level input. The second framework, **VectorSynth**, incorporates OSM tags as textual input by embedding them with a pretrained text encoder (e.g., CLIP), then projecting the resulting embeddings to a pixel-level control map. The third framework, **VectorSynth-COSA**, builds on this approach but replaces the off-the-shelf text encoder with our aforementioned COSA model's text encoder. Across both text-based variants, we experiment with different text encoders and study the effect of contrastive learning on grounding semantic control in image synthesis. All results reported use a 2D convolutional adapter network.

Table 3 highlights the quantitative gains of our approach, showing strong and consistent improvements across all standard metrics. We evaluate the quality of generated images using three standard metrics: Fréchet Inception Distance (FID) [14] to measure distributional similarity to real images, Structural Similarity Index (SSIM) [42] to assess perceptual quality, and Peak Signal-to-Noise Ratio (PSNR) to quantify pixel-level reconstruction accuracy. Our method significantly outperforms GeoSynth-OSM when using text-based control inputs. We see further gains when using COSA-aligned encoders over their vanilla (non-aligned) counterparts. Notably, we also compare our method with RemoteCLIP's text encoder specifically tuned for remote sensing tasks and see that our COSA-aligned text encoders

Figure 8. Examples of semantic edits across urban planning, and landuse generation applications.

Category	GeoSynth + Inpaint	VectorSynth
building	18.98	13.76
natural	29.50	28.40
place	26.32	21.30
landuse	17.21	12.95
highway	23.28	16.29
Combined	16.19	**11.34**

Table 4. FID scores (↓) for semantic edits across categories. GeoSynth + Inpaint uses Stable Diffusion inpainting [35] with GeoSynth weights, while VectorSynth uses our standard pipeline. **Note:** For fairness, non-edited regions are preserved from the original image, lowering FID relative to full image synthesis.

consistently perform better, demonstrating stronger semantic grounding.

Fine-Grained Semantic Control. To evaluate semantic editing capabilities, we introduce targeted edits and examine their localized impact to the output image. Figure 7 shows diverse examples across semantic categories and spatial contexts, demonstrating the model's ability to produce precise, semantically meaningful edits. We compare VectorSynth to GeoSynth [36], and Stable Diffusion Inpainting [35] using GeoSynth weights. VectorSynth produces more realistic and coherent edits: buildings exhibit consistent structure and alignment, while features such as soccer fields appear more regular and well-formed. Quantitatively, we evaluate semantic edits in Table 4. VectorSynth editing achieves lower FID scores across various semantic categories, indicating high fidelity and better alignment with the original distribution. When all categories are combined, VectorSynth also outperforms. Note that in the combined

FID, all edits are aggregated into a larger, lower-variance set, which yields lower scores than per-category metrics. Additional capabilities are also shown in Figure 8, lending to potential applications of this model such as in urban planning and landuse generation.

6. Conclusion

We introduced VectorSynth, a novel approach for satellite image synthesis that provides fine-grained pixel-level semantic control, moving beyond the coarse-grained conditioning of prior work, which relied on broad categories, such as buildings, parks, and roads. By allowing users to define more detailed semantics for different regions in the image, VectorSynth enables a broad range of applications, from data generation for machine learning models to citizen-driven urban design. Our approach aligns the representation space using polygon-level contrastive learning, outperforming strong, off-the-shelf embedding networks. The current design is well-suited to editing scenarios, where precise, localized control is essential. Future work includes enhancing the model's ability to learn from sparse or incomplete annotations, thereby increasing its applicability in data-limited settings. Additionally, enabling the network to hallucinate uncontrolled map regions more effectively would allow users to specify fewer semantic regions while still generating coherent, high-quality scenes.

Acknowledgments

References

[1] Xinyu Bai, Xinyang Pu, and Feng Xu. Conditional diffusion for sar to optical image translation. *IEEE Geoscience and Remote Sensing Letters*, pages 1–1, 2023. 3

[2] Favyen Bastani, Piper Wolters, Ritwik Gupta, Joe Ferdinando, and Aniruddha Kembhavi. Satlaspretrain: A large-scale dataset for remote sensing image understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16772–16782, 2023. 5, 12

[3] Yanis Benidir, Nicolas Gonthier, and Clément Mallet. The change you want to detect: Semantic change detection in earth observation with hybrid data generationf. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2204–2214, 2025. 2, 3

[4] 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*, pages 18392–18402, 2023. 1

[5] Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: from patch and token embeddings to finite discrete tokens. In *CVPR*, pages 15095–15104, 2023. 4

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 5

[7] Aayush Dhakal, Adeel Ahmad, Subash Khanal, Srikumar Sastry, Hannah Kerner, and Nathan Jacobs. Sat2cap: Mapping fine-grained textual descriptions from satellite images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 533–542, 2024. 2

[8] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 2

[9] Geofabrik GmbH. Geofabrik download server. https://download.geofabrik.de/, 2024. Accessed: 2025-07-17. 3

[10] Muhammed Goktepe, Amir hossein Shamseddin, Erencan Uysal, Javier Muinelo Monteagudo, Lukas Drees, Aysim Toker, Senthold Asseng, and Malte von Bloh. Ecomapper: Generative modeling for climate-aware satellite imagery. In *Forty-second International Conference on Machine Learning*, 2025. 2

[11] Mordechai Haklay and Patrick Weber. Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, 7(4):12–18, 2008. 17

[12] Tengda Han, Dilara Gokay, Joseph Heyward, Chuhan Zhang, Daniel Zoran, Viorica Patraucean, Joao Carreira, Dima Damen, and Andrew Zisserman. Learning from streaming video with orthogonal gradients. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 13651–13660, 2025. 13, 14

[13] Yutong He, Dingjie Wang, Nicholas Lai, William Zhang, Chenlin Meng, Marshall Burke, David Lobell, and Stefano Ermon. Spatial-temporal super-resolution of satellite imagery via conditional pixel synthesis. *Advances in Neural Information Processing Systems*, 34:27903–27915, 2021. 2

[14] 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. 7

[15] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446, 2002. 6

[16] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2

[17] Dong Jing, Xiaolong He, Yutian Luo, Nanyi Fei, Wei Wei, Huiwen Zhao, Zhiwu Lu, et al. Fineclip: Self-distilled region-based clip for better fine-grained understanding. *Advances in Neural Information Processing Systems*, 37:27896–27918, 2024. 2

[18] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023. 1

[19] 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. *arXiv preprint arXiv:2312.03606*, 2023. 3

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 16

[21] Kartik Kuckreja, Muhammad Sohail Danish, Muzammal Naseer, Abhijit Das, Salman Khan, and Fahad Shahbaz Khan. Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition, pages 27831–27840, 2024. 2

[22] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pages 143–160. Springer, 2024. 2

[23] Juncheng Li, Xin He, Longhui Wei, Long Qian, Linchao Zhu, Lingxi Xie, Yueting Zhuang, Qi Tian, and Siliang Tang. Fine-grained semantically aligned vision-language pre-training. *Advances in neural information processing systems*, 35:7290–7303, 2022. 2

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. pages 12888–12900. PMLR, 2022. 2

[25] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10545–10556, 2025. 12

[26] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023. 1

[27] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024. 2, 4, 15

[28] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 3

[29] Timo Lüddecke and Alexander Ecker. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7086–7096, 2022. 2

[30] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11461–11471, 2022. 1

[31] Arpan Mahara and Naphtali Rishe. Multispectral band-aware generation of satellite images across domains using generative adversarial networks and contrastive learning. *Remote Sensing*, 16(7):1154, 2024. 2

[32] Boyu Pang, Siwei Zhao, and Yinnian Liu. The use of a stable super-resolution generative adversarial network (ssrgan) on remote sensing images. *Remote Sensing*, 15(20):5064, 2023. 2

[33] 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. 2

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[35] 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. 1, 5, 8

[36] 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. 2, 3, 7, 8, 9, 12

[37] Akashah Shabbir, Mohammed Zumri, Mohammed Bennamoun, Fahad Shahbaz Khan, and Salman Khan. Geopixel: Pixel grounding large multimodal model in remote sensing. In *Forty-second International Conference on Machine Learning*, 2025. 2

[38] Aysim Toker, Marvin Eisenberger, Daniel Cremers, and Laura Leal-Taixé. Satsynth: Augmenting image-mask pairs through diffusion models for aerial semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27695–27705, 2024. 3

[39] VIDA. Google-microsoft open buildings - combined by vida. https://beta.source.coop/repositories/vida/google-microsoft-open-buildings. Accessed: 2025-05-18. 3

[40] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*, 2022. 5

[41] Yuxi Wang, Bing Zhang, Wenjuan Zhang, Danfeng Hong, Bin Zhao, and Zhen Li. Cloud removal with sar-optical data fusion using a unified spatial–spectral residual network. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–20, 2023. 3

[42] 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. 7

[43] Zhecheng Wang, Rajanie Prabha, Tianyuan Huang, Jiajun Wu, and Ram Rajagopal. Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5805–5813, 2024. 2

[44] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021. 4

[45] Liang Yao, Fan Liu, Delong Chen, Chuanyi Zhang, Yijun Wang, Ziyun Chen, Wei Xu, Shimin Di, and Yuhui Zheng. Remotesam: Towards segment anything for earth observation, 2025. 2

[46] 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*, 2024. 3

[47] Quan-Sheng Zeng, Yunheng Li, Daquan Zhou, Guanbin Li, Qibin Hou, and Ming-Ming Cheng. Maskclip++: A mask-based clip fine-tuning framework for open-vocabulary image segmentation. 2024. 2

[48] 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. 1, 3, 5

[49] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019. 6

[50] Xiangrong Zhang, Shunli Tian, Guanchun Wang, Huiyu Zhou, and Licheng Jiao. Diffucd: Unsupervised hyperspectral image change detection with semantic correlation diffusion model. *arXiv preprint arXiv:2305.12410*, 2023. 2

[51] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16793–16803, 2022. 2

[52] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 2, 4

VectorSynth: Fine-Grained Satellite Image Synthesis with Structured Semantics

Supplementary Material

A. Applications

We evaluate the utility of our generated imagery by running SegEarth-OV [25], a state-of-the-art open-vocabulary remote sensing segmentation model, on synthetic images created from structured OSM tags using our VectorSynth-COSA model. We also run the segmentation model on the grounded satellite imagery, and GeoSynth-OSM [36] for comparison. We define a subset of categories in our data that represent different land uses, buildings, and road types. We use segmentation accuracy to measure how well the generated image matches the given polygon labels, with higher accuracy meaning the generated image shows strong pixel-level fidelity to the class.

We consistently outperform GeoSynth-OSM across all categories, with particularly strong gains in fine-grained classes such as road types and distinct building uses. As shown in Table 5, our method achieves competitive results compared to real satellite imagery, and in several categories, such as land use residential, and natural regions, performance even surpasses that of real images. This indicates that our model generates semantically faithful scenes that align well with downstream open-vocabulary segmentation tasks. While challenging categories like industrial areas remain difficult due to visual ambiguity, our results demonstrate that our pretraining alignment and generation pipeline yields more spatially and semantically precise synthetic outputs.

We further perform some qualitative evaluations in Figure 9 on out of OSM distribution text prompts. We see that generated outputs adhere to spatial and semantic constraints.

B. Data

As seen in Figure 10, we densely sampled Los Angeles, New York City, Paris, and Berlin. These cities were chosen as they represent different urban planning styles: Los Angeles exemplifies low-density horizontal sprawl, New York is defined by verticality and a rigid grid system, Paris features radial layouts and dense historical cores, and Berlin reflects a blend of post-war reconstruction and structured zoning. Chicago is used as an out-of-space test city. In addition, we conduct one additional experiment on generating OSM annotations using Sydney, Australia.

We visualize the tag distribution in our training data as seen in Figure 11. Through the overlay visualization in Figure 12, we see that there is strong spatial grounding in the dataset.

Class	GeoSynth	VectorSynth	Original
place	79.54	81.26	81.81
natural region	25.26	26.04	25.55
building	19.52	32.03	32.43
industrial	4.15	18.23	36.19
apartments	5.00	14.95	22.92
school	0.50	11.39	20.69
landuse	44.62	55.90	55.06
residential	44.62	55.90	55.06
farmland	2.35	16.80	55.17
forest	12.39	28.12	36.47
sport	4.86	15.51	26.65
railway	2.93	13.51	42.04

Table 5. Segmentation accuracy (%) for parent and child classes in OSM tags using SegEarth-OV. Child classes are indented under their parent. We compare generated images from GeoSynth and VectorSynth, along with the original satellite image.

C. COSA

C.1. Architecture Details

Image Encoder. Our image encoder is built on top of SatlasNet [2]. SatlasNet is pretrained on high-resolution aerial imagery, consistent with our dataset, using a Swin-V2 backbone followed by a feature pyramid network (FPN) resulting in multi-scale feature maps of varying resolution. Following the FPN, our image encoder interpolates and concatenates the multi-scale feature maps, then passes the result through a learnable MLP adapter network to align the embedding dimension of the text encoder. Our adapter network consists of sequential 1×1 convolutional layers with ReLU activations and Batch Normalization, following a Conv2d–ReLU–BN–Conv2d–ReLU–BN structure. To encourage high resolution vision-language alignment, we freeze the Swin-V2 backbone and let the feature pyramid network, and adapter network be learnable.

C.2. Training and Implementation Details

Polygon Sampling. As the number of polygon-text pairs varies within a minibatch, contrastive sampling size can also fluctuate. To address this variability, we use a combination of minibatch size B and number of sampled polygon-text pairs K such that, with at least 95% confidence, the sampled K pairs is reached. In cases where K is not reached, we sample all polygon-text pairs within the batch. If the same multi-tag composition is in multiple images, the polygon pair is randomly chosen from one of the corresponding images. Our setup naturally introduces both *intra-*
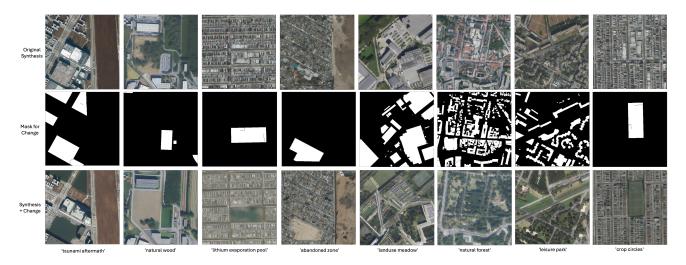
Figure 9. Qualitative evaluations on out-of-distribution text incorporated conditions, and purposeful edits. Each set shows: (top) original synthesized image, (middle) the injected control mask corresponding to the semantic change, and (bottom) the synthesized output conditioned on (below) the natural language description used for conditioning.

Figure 10. Geographic coverage of the dataset across five major cities: Los Angeles, New York City, Paris, Berlin, and Chicago. Each city includes training, validation, and held-out tiles, except for Chicago, which is fully held out and used only for testing. Training and validation tiles are shown in blue; Chicago test tiles are shown in red.

Figure 11. Word cloud of the most frequent OSM tags in the dataset. Font size reflects frequency across all tile-level tag lists. Common tags include urban structure types (e.g., `building residential`, `highway primary`), land use (`land use commercial`, `park`), and 3D features such as `height`.

image and *inter-image* negatives, encouraging distinction between semantically similar polygons-text pairs within the

same image and across different images. In addition, for each contrastive pair, we sub-sample tag words in the multi-tag compositions during training to provide better generalization to varying text queries.

Optimization With Orthogonal Gradients. Due to the spatial nature of satellite images, polygon-text pairs often exhibit strong feature correlation, both within the same image and across images in the batch. This is especially true in urban areas, where OSM tag distributions and architectural layouts likely follow recurring spatial patterns (e.g., residential blocks, road grids, building clusters). Such correlations may limit learning efficiency and lead to gradient directions with poor diversity. For this reason, we implement orthogonal gradients [12], an optimization technique designed to promote diversity by projecting updates onto the gradients orthogonal component. This approach has shown effectiveness in data domains such as sequential

Figure 12. Overlay of OSM-derived tags on top of satellite imagery. Each region is annotated with semantically meaningful labels (e.g., `building residential`, `land use park`), showcasing the compositional richness and spatial precision of the dataset.

video frames, where the data is highly correlated. Specifically, we implement the *Orthogonal AdamW* variant implemented based on [12]. Orthogonal AdamW introduces an additional term controlled by a hyperparameter β_{ort}, set to 0.9 in our experiments.

Training Details. We train our model using the AdamW optimizer with a learning rate of $1e-4$, β values of $(0.9, 0.98)$, $\epsilon = 1e-6$, and a weight decay of 0.01. We use a cosine annealing warm restart learning rate scheduler with an initial cycle length of $T_0 = 20$ epochs. We train until early stopping with a minibatch size of $B = 6$ satellite images per GPU. Each batch includes $K = 128$ sampled polygon-tag pairs, drawn across the minibatch. Training typically ended around 80 epochs. We initialize the logit scale temperature parameter as $\log(1/0.07)$ and learn it during training. To avoid numerical instability, we clamp the logit scale to a maximum of $\log(100)$. To ensure reproducibility, we set all random seeds to 42 and disabled CuDNN benchmarking. All experiments are run on two GeForce RTX 4090 GPUs (24GB).

D. VectorSynth Controls

A qualitative comparison of the different control signals is provided in Figure 15. Visually, we observe that OSM tiles provide a high-level structural prior but lack semantic richness. While text-based pixel-level control maps introduce more diverse semantic information, our COSA control

maps exhibit sharper transitions between objects, reflecting stronger inter-tag contrast and improved spatial grounding. This is especially evident in the fine-grained delineation of urban features. For example, residential and commercial buildings, as well as differences in heights of buildings, appear more homogeneous in CLIP maps, but are more distinctly separated in COSA. These improvements result from aligning OSM tag semantics with satellite imagery during pretraining, leading to control signals that are both semantically expressive and spatially localized.

E. Dealing with Sparsity

Geographic annotation datasets often suffer from inherent sparsity, where comprehensive polygon coverage is unavailable across all spatial regions. This sparsity presents significant challenges during both training and inference, as models must generate plausible geographic content even when provided with incomplete or limited control signals. We address this fundamental limitation through two complementary approaches: progressive masking during training and automated annotation enhancement using vision-language models.

E.1. Progressive Masking for Sparse Control Adaptation

To enable robust performance under sparse annotation conditions, we use a progressive masking training strategy that gradually reduces polygon coverage throughout the training process. This approach trains the model to effectively hallucinate plausible geographic features when given increasingly sparse control inputs.

Our progressive masking scheme linearly increases the proportion of masked polygons over training iterations (100% to 30%). This curriculum learning approach allows the model to first establish strong associations between dense annotations and corresponding geographic features, then gradually adapt to scenarios with limited supervisory information.

The progressive masking strategy demonstrates clear benefits for sparse control scenarios. As illustrated in Figure 16, models trained with this approach exhibit improved robustness when polygon coverage falls below 60% of the image area. However, we observe a trade-off in performance: while the progressively masked model excels with sparse controls, it slightly underperforms compared to the baseline model when provided with very dense annotation coverage. This behavior aligns with our training objective, as the model learns to rely less heavily on comprehensive annotation signals.
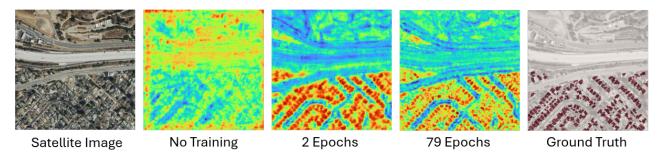
Figure 13. Similarity heatmaps for a satellite image given the text query 'house' inferred from COSA with no training, 2 epochs of training, and 79 epochs of training.
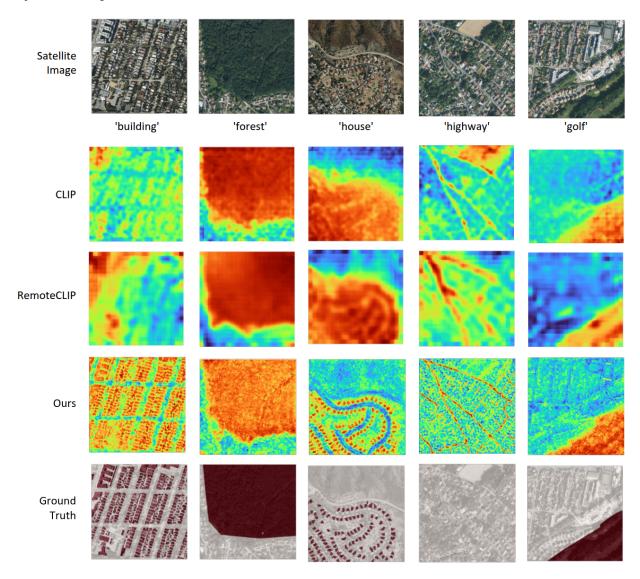

Figure 14. Similarity heatmaps given text queries comparing CLIP, RemoteCLIP, and our approach—COSA. Taking inspiration from [27], we use a sliding window inference approach to show high-resolution similarity heatmaps for CLIP and RemoteCLIP.
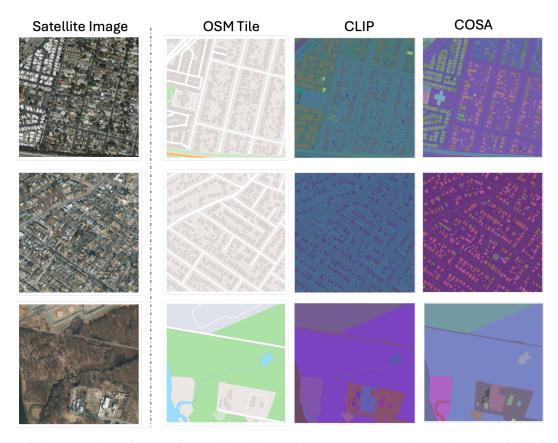
| Satellite Image | OSM Tile | CLIP | COSA |

Figure 15. Qualitative comparison of control signals. OSM tiles provide coarse structural priors but lack semantic detail. Text-based maps offer richer semantics, while COSA maps show sharper object boundaries and better spatial grounding—especially in distinguishing urban features like residential and commercial buildings. These improvements stem from aligning OSM tags with satellite imagery during pretraining.

Model	FID ↓	SSIM ↑	PSNR ↑
GeoSynth	170.25	0.18	12.16
VectorSynth	177.17	0.17	11.99
VectorSynth (w/ generated tags)	154.13	0.18	12.11

Table 6. Comparison of FID, SSIM, and PSNR of satellite imagery across Sydney, Australia

E.2. Automated Annotation Enhancement via COSA VLM

To further address annotation sparsity, we leverage our COSA vision-language model (VLM) to automatically generate additional semantic annotations for sparse regions. This approach combines the Segment Anything Model (SAM) [20] for mask generation with our specialized COSA VLM for polygon-text retrieval, creating a pipeline that densifies sparse annotations with contextually appropriate semantic labels.

The annotation enhancement pipeline operates in three stages. First, we apply SAM [20] to the input satellite imagery to generate comprehensive segmentation masks covering all visible geographic features. Next, we utilize our COSA VLM to perform polygon-retrieval, generating semantically grounded text descriptions for each SAM-generated mask. These automatically generated text annotations are then integrated with existing sparse annotations to provide richer control signals during generation.

We evaluate this approach on an out-of-distribution dataset featuring high-resolution imagery of Sydney, Australia. Sydney's harbor-centric development and organic street patterns contrast with our training data from NYC, LA, Berlin, and Paris, which feature more geometric grids and radial planning structures. We compare three generation approaches: VectorSynth using only available Open-
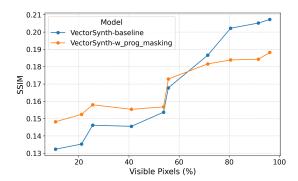
Figure 16. SSIM performance versus polygon coverage. Progressive masking (orange) outperforms baseline (blue) below 60% coverage but underperforms at dense coverage above 80%.

StreetMap (OSM) [11] tags without filtering, the baseline GeoSynth model, and VectorSynth enhanced with SAM + COSA VLM annotations. We note that we do not filter the coverage of the dataset; therefore, the OSM tags are very sparse, and many images do not contain any OSM tag information.

In Table 6, we see that using our text generation pipeline improves upon strictly using the OSM tags, and outperforms other baselines. Our experimental results demonstrate that the automated annotation enhancement pipeline can be an effective way to mitigates sparsity limitations and generate data that is useful for our vectorsynth generation.

The combination of progressive masking training and automated annotation enhancement provides a comprehensive solution to the sparsity challenge in geographic image synthesis. While progressive masking enables the model to perform well with inherently sparse controls, the COSA VLM pipeline allows us to artificially densify annotations when computational resources permit, achieving the best of both sparse and dense control paradigms.