BIFRÖST: MULTI-SENSOR DIFFUSION MODEL FOR CROSS-VIEW PANORAMIC SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthesizing ground-level 360° panoramas from overhead imagery typically assumes access to sub-meter aerial photography (0.3m/px). In practice, this restricts synthesis to densely surveyed metropolitan regions. Moving to globally available 10m/px Sentinel-2 multi-spectral data avoids this coverage bottleneck, but the low-resolution conditioning yields severe structural ambiguity. To overcome this, we introduce Bifröst, a diffusion framework designed to run inference strictly on open-access satellite data (10m/px Sentinel-2 and AlphaEarth embeddings) while retaining high-resolution geometric detail. Rather than feeding high-resolution imagery into the generator, we use it solely as training-time supervision via a frozen Perspective-Inversion Critic (PIC) that enforces structural consistency. This gives us the novel ability to generate street-view using only highly available low-resolution information. To handle cross-view spatial transformation, we align overhead polar coordinates directly with equirectangular spherical grids inside cross-attention layers. On a benchmark of 560k spatially disjoint locations, Bifröst eliminates the blur typical of low-resolution baselines, cutting relative depth error (AbsRel) from 0.138 to 0.106, matching models conditioned on true sub-meter aerial inputs (AbsRel 0.101) while improving FID by 77.86% over prior methods. On the public CVUSA dataset, Bifröst achieves an improvement of 68.4% FID over prior methods. These results demonstrate that the proposed approach effectively overcomes the need for high resolution information for cross-view generation.

1 INTRODUCTION

Synthesizing immersive 360° ground-level panoramas from overhead satellite imagery requires bridging an extreme viewpoint disparity. Because a single nadir satellite tile corresponds to an under-constrained distribution of ground-level scenes, deterministic mappings inevitably yield blurred, mean-seeking predictions (Zhai et al., 2017) that are insufficient for downstream applications like crop-mapping (Soler et al., 2024), autonomous driving (Gao et al., 2024) or planning. Recent works resolve this ambiguity by conditioning generation on explicit geometric surrogates, such as neural density fields (Qian et al., 2023) or height maps (Li et al., 2024), or by relying directly on high-resolution aerial photography ($\leq$ 0.3 m/px) (Ze et al., 2025; Zhang et al., 2023). While effective in dense urban corridors, these methods fail in practice: sub-meter aerial surveys are geographically sparse and non-existent across most rural or agricultural regions worldwide, due to imaging constraints and deployment costs (Gao et al., 2026).

A natural alternative is to condition exclusively on open-access, globally ubiquitous data, such as 10 m/px Sentinel-2 multi-spectral imagery. However, low-resolution data alone suffers from spectral degeneracy. The network captures regional surface reflectance but lacks the spatial frequency required to resolve roads, foliage boundaries, or building outlines. Further, additional annualized AlphaEarth (Brown et al., 2025) also are based on low-resolution observations and lack the necessary spatial frequency. The required high-resolution aerial tiles are unfortunately not available with sufficient temporal cadence and geographical coverage. Therefore, this directly limits the generation of the street-view information and this can only be generated in a highly sparse geographical and temporal cadence. The need for streetview generation is required for tasks like crop-mapping and monitoring (Soler et al., 2024). Doing so with sparse coverage would defeat the purpose for cross-view generation.

To resolve this problem, we propose ***Bifröst***, a multi-sensor diffusion framework that decouples test-time operational constraints from training-time geometric supervision. Bifröst operates strictly on

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.

open-access data at inference .i.e., requiring only 10 m/px Sentinel-2 multi-spectral bands, AlphaEarth foundation embeddings, and geo-temporal metadata. Rather than feeding high-resolution imagery into the generator, we internalize high-frequency structural priors entirely through training supervision. Specifically, we freeze a pre-trained *Perspective-Inversion Critic (PIC)* that maps generated street-level panoramas back to nadir aerial space. By penalizing structural discrepancies against unobserved high-resolution targets using an energy-based spatial covariance loss, the generator learns to infer sharp, geometrically plausible layouts from ambiguous low-resolution inputs without requiring high-resolution data at test time.

In summary, our primary contributions are:

- **Global-Scale Inference Paradigm:** To the best of our knowledge, we are the first to formulate a cross-view synthesis pipeline that eliminates inference-time dependencies on sub-meter aerial imagery, synthesizing coherent 360° panoramas exclusively from globally available 10 m/px Sentinel-2 bands, AlphaEarth embeddings, and geo-temporal metadata.
- **Cyclic Supervision:** We introduce a training-only supervision strategy using a frozen Perspective-Inversion Critic (PIC) that maps synthesized panoramas back to nadir space. We show that formulating the critic under a Gibbs distribution allows optimizing a composite spatial energy loss (boundary-weighted L_1 and MS-SSIM), internalizing sub-meter structural priors without test-time overhead(section 3.5).
- **Coordinate-Aware Cross-Attention:** We design a coordinate-augmented cross-attention mechanism that couples nadir polar coordinates directly with equirectangular spherical coordinates, establishing a smooth algebraic bridge between overhead planar orientation and panoramic horizontal azimuth (section 3.3).
- **Bridging the High-Resolution Performance Gap:** Across both a large-scale global benchmark (560k geo-temporally images) and the standard CVUSA dataset (Workman et al., 2015; Zhai et al., 2017), we show that inference using exclusively low-resolution inputs reaches parity with models conditioned on true sub-meter aerial photography, achieving competitive relative depth consistency (AbsRel $= 0.106$ vs. 0.101), improved distribution coverage (FID $= 3.87$ on our benchmark; 6.73 on CVUSA), and robust semantic layout fidelity (section 4.2).

2 Related Work

Cross-View Synthesis: Geometric Warping to Foundation-Guided Diffusion. Early overhead-to-ground image translation relied on conditional GANs, employing intermediate semantic segmentation masks to regularize the ill-posed mapping (Regmi & Borji, 2018; Tang et al., 2019; Wu et al., 2023). However, these frameworks remained fragile due to the circular challenge of predicting accurate ground-level semantics from top-down views. Subsequent methods turned to explicit 3D inductive biases, employing projective geometry, height estimation, and homography warping to transform aerial features into ground perspectives (Lu et al., 2020; Shi et al., 2022). While effective in metropolitan scenes characterized by sharp vertical facades, explicit geometry collapses across rural and agricultural landscapes where topographic relief is subtle and monocular depth cues degenerate.

Recent approaches leverage the generative scale of diffusion and flow-matching models to synthesize coherent street-level scenes. CrossViewDiff (Li et al., 2024) and CS2S (Ze et al., 2025) achieve structural consistency through spatial conditioning, while Sat2GroundScape (Xu & Qin, 2025) and SatDreamer360 (Ze et al., 2026) enforce cross-view and multi-frame panoramic coherence. Extending beyond unidirectional synthesis, Geo^2 (Zhang et al., 2026) unifies cross-view geo-localization and bidirectional synthesis by embedding aerial and ground views into a shared 3D-aware latent space guided by geometric foundation models such as VGGT (Wang et al., 2025a). Similarly, Mixed-View Diffusion (Xiong et al., 2025) synthesizes novel street views by interpolating between satellite imagery and sparse proximal ground panoramas. Nonetheless, all these generative frameworks remain tethered to sub-meter aerial photography (≤ 0.3 m/px) or pre-existing ground captures at inference time, rendering them inapplicable across the vast non-urban regions of the globe.

Implicit 3D Representations and Neural Scene Rendering. To avoid the discretization artifacts of explicit geometry, continuous 3D representations have been increasingly explored for cross-view tasks. Sat2Density (Qian et al., 2023) and its multi-view extension Sat2Density^{++} (Qian et al., 2026a) learn continuous neural radiance fields from aerial imagery, decoupling ground radiance fields from

view-dependent sky and illumination effects to enable novel view synthesis. In parallel, 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) has been adapted to cross-view domains: BevSplat (Wang et al., 2025b) leverages Gaussian primitives for weakly-supervised cross-view localization, Horizon-GS (Jiang et al., 2025) introduces multi-resolution level-of-detail anchors to bridge aerial and street-level rendering, and Skyfall-GS (Lee et al., 2026) combines satellite photogrammetry with generative diffusion refinement. To resolve facade floaters and rooftop geometry collapse in sparse-view settings, Sat3DGen (Qian et al., 2026b) introduces a geometry-first feed-forward framework with perspective supervision.

Despite their geometric fidelity, these neural representations rely on dense multi-view sequences, calibrated camera trajectories, or scene-specific per-location optimization, preventing zero-shot feed-forward generation across arbitrary global coordinates. Moreover, because their inductive biases depend heavily on visible parallax across vertical architectural planes, their performance degrades across rural environments where nadir satellite data contains no vertical facade observations.

Geospatial Foundation Models and Variational Privileged Learning. Recent advances in remote sensing demonstrate that planetary-scale foundation models, such as DiffusionSat (Khanna et al., 2024), TerraMind (Jakubik et al., 2026), and AlphaEarth (Brown et al., 2025), etc., encode rich macro-semantic and ecological land-cover priors. Rather than enforcing joint multi-modal conditioning during generation, our formulation builds on Learning Using Privileged Information (LUPI) (Vapnik & Vashist, 2009) and variational information maximization (Barber & Agakov, 2003; Poole et al., 2019). Grounding this principle in an asymmetric training setup, Bifröst decouples test-time operational constraints from training-time supervision. By employing a frozen Perspective-Inversion Critic (PIC) that projects generated panoramas back to nadir aerial space, Bifröst treats high-resolution aerial imagery as a variational structural prior. This design enables the generator to run inference exclusively on globally ubiquitous 10 m/px Sentinel-2 multi-spectral bands and foundation embeddings, achieving sub-meter structural consistency without requiring high-resolution aerial inputs at test time.

3 METHODOLOGY

Bifröst is a multi-sensor diffusion model that generates high-fidelity, north-aligned 360° street-level panoramas from open-access satellite data. Rather than conditioning on high-resolution aerial imagery during inference, Bifröst relies solely on ubiquitous low-resolution inputs: 10 m/px Sentinel-2 multi-spectral bands, AlphaEarth foundation embeddings, and geo-temporal metadata. High-resolution geometric priors are decoupled from test-time evaluation and internalized entirely during training via a frozen Perspective-Inversion Critic (PIC). We first formalize the cross-view synthesis objective (Section 3.1), outline the multi-sensor representation and coordinate-aware attention (Sections 3.2, 3.3), and present the training formulation that drives geometric internalization (Section 3.5).

3.1 PROBLEM FORMULATION

Let $\mathbf{x} \in \mathbb{R}^{H_p \times W_p \times 3}$ denote the ground-truth north-aligned equirectangular panorama. We model the cross-view conditional distribution $p(\mathbf{x} \mid \mathcal{C}_{\text{lr}})$, where the inference context $\mathcal{C}_{\text{lr}} = \{\mathbf{c}_{\text{ms}}, \mathbf{c}_{\text{ae}}, \mathbf{c}_{\text{geo}}\}$ consists exclusively of globally available inputs. Here, $\mathbf{c}_{\text{ms}} \in \mathbb{R}^{H_m \times W_m \times 13}$ represents 13-band Sentinel-2 multi-spectral image (10 m/px); $\mathbf{c}_{\text{ae}} \in \mathbb{R}^{H_a \times W_a \times 64}$ contains annualized AlphaEarth foundation embeddings (Brown et al., 2025) encoding land-cover and topological semantics; and $\mathbf{c}_{\text{geo}} = (\text{latitude}(\theta), \text{longitude}(\phi), \text{day}(d), \text{month}(m), \text{year}(y))$ denotes geographic coordinates and capture timestamps that constrain regional illumination and seasonal vegetation states.

Cross-view synthesis from low-resolution nadir imagery is fundamentally under-constrained. Because a 10 m/px Sentinel-2 pixel integrates surface reflectance over a $100\,\text{m}^2$ ground footprint, the forward mapping exhibits an extensive null space: an infinite family of high-frequency ground-level geometries (road widths, fence placements, foliage boundaries) maps to the identical low-resolution multi-spectral observation. Furthermore, while high-resolution aerial imagery $\mathbf{c}_{\text{hr}} \in \mathbb{R}^{H_h \times W_h \times 3}$ (0.3 m/px) is available in curated datasets during development, conditioning directly on $\mathbf{c}_{\text{hr}}$ at test time destroys global scalability and introduces spatial-temporal feature conflicts when concatenated with annualized embeddings $\mathbf{c}_{\text{ae}}$. Hence, our objective is to maximize the conditional likelihood $p_\Theta(\mathbf{x} \mid \mathcal{C}_{\text{lr}})$ while structuring training such that the generator internalizes the structural distribution of $\mathbf{c}_{\text{hr}}$ without requiring it as an input during sampling.

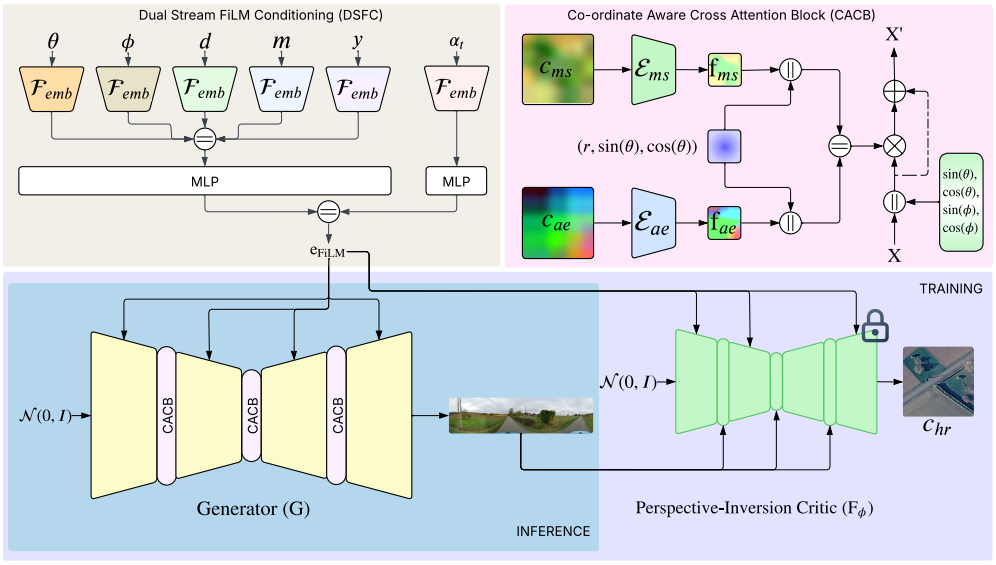

Figure 1: **The Bifröst Multi-Sensor Diffusion Architecture.** The framework decouples inference-time requirements from training-time supervision. **Inference pathway (Blue):** Globally ubiquitous multi-spectral bands ($\mathbf{c}_{\text{ms}}$) and foundation embeddings ($\mathbf{c}_{\text{ae}}$) are encoded into a unified token sequence, while geo-temporal metadata ($\mathbf{c}_{\text{geo}}$) and diffusion timesteps are injected via Dual-Stream FiLM Conditioning (DSFC). The generator backbone uses Coordinate-Aware Cross-Attention Blocks (CACB) to align overhead polar geometry (r, θ) with equirectangular panorama coordinates $(\theta_{\text{pano}}, \phi_{\text{pano}})$. **Training-only supervision (Purple):** A frozen Perspective-Inversion Critic (F_ϕ) maps the synthesized panorama back to nadir aerial space to reconstruct unobserved high-resolution targets ($\mathbf{c}_{\text{hr}}$), penalizing geometric drift via an energy-based spatial covariance loss. Both the critic and high-resolution data dependencies are discarded at test time. (Operations: $\otimes$: Cross-Attention, $\oplus$: Residual Addition, $\textcircled{||}$: Sequence Concatenation, $\textcircled{=}$: Channel Concatenation).

3.2 Multi-Sensor Input Representation and Conditioning

To process the heterogeneous inputs within the low-resolution context $\mathcal{C}_{lr}$, we employ independent convolutional encoders $\mathcal{E}_{ms}$ and $\mathcal{E}_{ae}$. The encoders utilize auto-strided convolutions to compress the modalities down to a unified spatial grid of size $H_s \times W_s$. Let $\mathbf{f}_{ms} = \mathcal{E}_{ms}(\mathbf{c}_{ms})$ and $\mathbf{f}_{ae} = \mathcal{E}_{ae}(\mathbf{c}_{ae})$, where each $\mathbf{f} \in \mathbb{R}^{H_s \times W_s \times C_s}$ and C_s is the shared latent channel dimension.

Compressing these inputs into a uniform grid yields two key benefits. First, flattening the grids produces a fixed sequence length of $H_s W_s << H_m W_m$ tokens per modality, ensuring memory-efficient downstream cross-attention. Second, the architecture dynamically accommodates the available modalities. The $\mathbf{f}_{ms}$ and $\mathbf{f}_{ae}$ tokens are concatenated along the sequence dimension to form a unified key-value sequence of length $2H_s W_s$ without requiring structural changes to the attention blocks. (More in Section 3.3).

Global Geo-Temporal FiLM Conditioning: To ground regional illumination and seasonal vegetation states without corrupting fine-grained spatial feature maps, we inject macro-environmental metadata alongside the diffusion noise level. Following (Khanna et al., 2024), the scalar geo-temporal tuple $\mathbf{c}_{\text{geo}} = (\theta, \phi, d, m, y)$ is decomposed using independent continuous sinusoidal Fourier feature embeddings $\mathcal{F}_{emb}(.)$ and projected via a lightweight MLP to yield the geo-temporal embedding $\mathbf{h}_{\text{geo}}$. Concurrently, we follow the network architecture as well as the continuous-time formulation of simple diffusion (Hoogeboom et al., 2023). The continuous signal scaling factor α_t is transformed into a global noise embedding e_{t}. We concatenate these representations along the channel dimension to construct the unified modulation context $\mathbf{e}_{\text{FiLM}} = [\mathbf{e}_t \parallel \mathbf{h}_{\text{geo}}]$. Following (Perez et al., 2018), $\mathbf{e}_{\text{FiLM}}$ modulates feature activations across all resolution stages of the U-ViT backbone via Feature-wise Linear Modulation (FiLM; complete architectural layers, MLP projections in Appendix D.1).

3.3 Coordinate Aware Cross Attention

The Geometric Inductive Bias: Satellite imagery and street-level panoramas view the world through completely different geometries. If you rotate a satellite image, features spin around the center of the image. But if you turn around on the street, the scenery simply shifts left or right across the 360° panorama. Because standard attention treats image tokens like an unordered bag of features, it has no built-in sense of direction and struggles to map overhead positions to ground-level lines of sight.

Polar-to-Spherical Coordinate Augmentation: To provide the attention layers with a structural inductive bias, we explicitly ground tokens in their native geometric reference frames:

1. *Satellite Polar Grid:* For overhead feature maps of size $H_s \times W_s$, we construct a normalized polar grid $\mathbf{C}_{\text{polar}} = (r, \sin\theta, \cos\theta) \in \mathbb{R}^{H_s \times W_s \times 3}$ [1], where $r \in [0, 1]$ is the normalized radial distance from the image center and $\theta \in (-\pi, \pi]$ is the clockwise azimuth relative to North ($\theta = 0$). We concatenate $\mathbf{C}_{\text{polar}}$ along the channel dimension of each modality before flattening, yielding the coordinate-grounded overhead tokens $\tilde{\mathbf{F}}_{\text{ctx}} = [\text{flatten}(\mathbf{f}_{\text{ms}} \parallel \mathbf{C}_{\text{polar}}) \parallel_{\text{seq}} \text{flatten}(\mathbf{f}_{\text{ae}} \parallel \mathbf{C}_{\text{polar}})] \in \mathbb{R}^{2H_sW_s \times (C_s+3)}$.
2. *Panorama Spherical Grid:* For panorama query features $\mathbf{X}_l \in \mathbb{R}^{H_l \times W_l \times C_l}$ at network layer l, we construct an equirectangular coordinate grid $\mathbf{C}_{\text{pano}} = (\sin\theta_{\text{pano}}, \cos\theta_{\text{pano}}, \sin\phi_{\text{pano}}, \cos\phi_{\text{pano}}) \in \mathbb{R}^{H_l \times W_l \times 4}$, where $\theta_{\text{pano}} \in (-\pi, \pi]$ represents horizontal azimuth and $\phi_{\text{pano}} \in [\phi_{\min}, \phi_{\max}]$ spans the vertical elevation field of view (full coordinate derivations are provided in Appendix B). We form the augmented query tokens as $\tilde{\mathbf{X}}_l = [\mathbf{X}_l \parallel \mathbf{C}_{\text{pano}}] \in \mathbb{R}^{H_l \times W_l \times (C_l+4)}$.

Cross-Attention Routing: The shared azimuthal harmonic encoding $(\sin\theta, \cos\theta)$ provides a direct algebraic bridge between overhead rotational orientation and panoramic horizontal panning. Linearly projecting the augmented tokens yields the cross-attention queries, keys, and values:

$$\mathbf{X}'_l = \mathbf{X}_l + \mathbf{W}_{\text{out}} \cdot \text{Attn}(\tilde{\mathbf{X}}_l\mathbf{W}_Q, \tilde{\mathbf{F}}_{\text{ctx}}\mathbf{W}_K, \tilde{\mathbf{F}}_{\text{ctx}}\mathbf{W}_V), \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable projection matrices for query, key and value respectively. This inductive bias guides the attention heads to softly align multi-sensor overhead contexts with panoramic lines of sight while preserving the capacity to model complex occlusions.

3.4 The Perspective-Inversion Critic (PIC)

Decoupling Inference Constraints from Training Supervision: High-resolution satellite imagery $\mathbf{c}_{\text{hr}}$ (0.3 m/px) provides the fine structural detail needed to place roads, fences, and tree lines accurately. However, requiring $\mathbf{c}_{\text{hr}}$ at test time defeats our goal of global coverage. We resolve this by using $\mathbf{c}_{\text{hr}}$ purely during training. We introduce a *Perspective-Inversion Critic* (F_ϕ), a pre-trained diffusion network that performs the inverse task: taking a generated ground-level panorama $\hat{\mathbf{x}}$ together with the low-resolution context $\mathcal{C}_{\text{lr}}$, and predicting the corresponding overhead satellite view $\mathbf{c}_{\text{hr}}$. The generator never sees $\mathbf{c}_{\text{hr}}$ directly; it only receives feedback on whether its generated panoramas can be inverted back into plausible aerial layouts.

Frozen Critic to stabilize the training: Using a critic often invites the min-max optimization instabilities typically associated with adversarial training frameworks. We avoid this entirely by keeping the critic F_ϕ frozen during generator training. Using a standard L_2 objective, the critic diffusion model is trained to estimate the conditional expectation of clean data (Efron, 2011):

$$F_\phi(\mathbf{z}_t, \hat{\mathbf{x}}, \mathcal{C}_{\text{lr}}, t) \approx \mathbb{E}\left[\mathbf{c}_{\text{hr}} \mid \mathbf{z}_t, \hat{\mathbf{x}}, \mathcal{C}_{\text{lr}}\right]. \tag{2}$$

Because F_ϕ is pre-trained and fixed (see Appendix C.1 for pre-training details), the generator G_Θ receives stationary, non-adversarial regression gradients through $\hat{\mathbf{x}}$. The critic simply acts as a geometric score function, penalizing layouts that contradict overhead satellite geometry.

Matching Noise Levels Between Generator and Critic: Diffusion models generate coarse layouts at early steps (high noise) and fine textures at later steps (low noise). If the critic evaluated a coarse

[1] *Notational clarification:* We overload the symbol θ to denote both the azimuthal angle in our geographic (latitude) and geometric coordinate systems (see fig. 3 and fig. 4). The exact meaning can be unambiguously inferred from the immediate context.

early-stage panorama using low noise, it would demand fine details before the layout was even formed. To keep both networks synchronized, we couple the critic's timestep t_{critic} to the generator's current training step $\bar{t}_{\text{gen}}$ with a small random jitter:

$$t_{\text{critic}} = \text{clip}\left(\bar{t}_{\text{gen}} + \epsilon,\ t_{\text{min}}^{\text{critic}},\ t_{\text{max}}^{\text{critic}}\right), \quad \epsilon \sim \mathcal{U}(-j, j). \tag{3}$$

This ensures the critic evaluates macro-layouts early in the denoising process of G_Θ, and focuses on fine surface textures toward the end. Clamping avoids noise collapse at $t_{\text{max}}^{\text{critic}}$ (where excessive Gaussian noise destroys gradient signal) and avoids gradient explosion at $t_{\text{min}}^{\text{critic}}$ (where near-deterministic pixel regression causes high-frequency ringing).

3.5 Cyclic Objective and Training Dynamics

To internalize unobserved high-resolution aerial geometry without requiring high-resolution inputs at test time, the generator must encode features predictive of $\mathbf{c}_{\text{hr}}$ beyond the spectral information already present in $\mathcal{C}_{\text{lr}}$.

Gibbs Formulation & Structural Critic Loss. Assuming an isotropic Gaussian for q_ϕ yields an L_2 regression loss, which collapses under multi-modal uncertainty into mean-seeking, blurry aerial predictions. To prevent this mean-seeking blur, we parameterize the surrogate q_ϕ as a Gibbs distribution:

$$q_\phi(\mathbf{c}_{\text{hr}} \mid \hat{\mathbf{x}}, \mathcal{C}_{\text{lr}}) = \frac{1}{Z} \exp\left(-E(\mathbf{c}_{\text{hr}}, \hat{\mathbf{c}}_{\text{hr}})\right), \quad \text{where} \quad \hat{\mathbf{c}}_{\text{hr}} = F_\phi(\mathbf{z}_{t_c}, \hat{\mathbf{x}}, \mathcal{C}_{\text{lr}}, t_c). \tag{4}$$

Here, F_ϕ is the frozen Perspective-Inversion Critic evaluated at critic timestep $t_c \approx \bar{t}_{\text{gen}}$, dynamically matched to the generator's noise level to synchronize spatial frequency scales. Following established practices in neural image synthesis (Zhao et al., 2017), we instantiate the composite spatial energy function $E(\mathbf{c}_{\text{hr}}, \hat{\mathbf{c}}_{\text{hr}})$ using boundary-weighted L_1 and multi scale MS-SSIM (Wang et al., 2003) to capture anisotropic boundary sharpness and multi-scale covariance without requiring intractable partition function estimation:

$$\mathcal{L}_{F_\phi} = E(\mathbf{c}_{\text{hr}}, \hat{\mathbf{c}}_{\text{hr}}) = w_{\ell_1} \left\|(\hat{\mathbf{c}}_{\text{hr}} - \mathbf{c}_{\text{hr}}) \odot \mathbf{\Lambda}\right\|_1 + w_{\text{ssim}} \left(1 - \text{MS-SSIM}(\hat{\mathbf{c}}_{\text{hr}}, \mathbf{c}_{\text{hr}})\right), \tag{5}$$

where $\mathbf{\Lambda} = 1 + (1 - \lambda_{\text{edge}})\mathbf{M}_{\text{sobel}}$ is an edge-weighting mask computed from the target aerial image.

Observation 1 Approximate Partition Function Invariance: Under the assumption of local contrast preservation, the energy function $E(\mathbf{c}_{\text{hr}}, \hat{\mathbf{c}}_{\text{hr}})$ acts as a strictly spatial residual mapping. Consequently, the normalizing partition function $Z = \int \exp(-E(\mathbf{c}_{\text{hr}}, \hat{\mathbf{c}}_{\text{hr}}))\, d\mathbf{c}_{\text{hr}} \approx Z_0$ is approximately invariant to the generator parameters θ.

By Section 3.5 (proof in Appendix A), $\nabla_\Theta \log Z \approx \mathbf{0}$.

Total Training Objective. The primary generator G_Θ is trained end-to-end under the combined objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_G + \lambda_{\text{critic}}(t) \cdot \mathcal{L}_{F_\phi}, \tag{6}$$

where $\mathcal{L}_G = \mathbb{E}\left[\lambda_G(t) \|\mathbf{\Omega} \odot (\hat{\mathbf{x}} - \mathbf{x})\|_2^2\right]$ is the continuous diffusion loss modulated by the spatial horizon weighting matrix $\mathbf{\Omega}$ (Appendix D.3), and $\lambda_{\text{critic}}(t)$ follows schedule as described in Appendix D.4. Crucially, because F_ϕ is frozen, the critic acts as an analytical energy oracle rather than an adversarial discriminator, completely avoiding min-max training instability. At inference, synthesis operates purely via variance-corrected DDIM conditioned on $\mathcal{C}_{\text{lr}}$, bypassing F_ϕ and $\mathbf{c}_{\text{hr}}$ entirely.

4 Experiments

4.1 Experimental Setup

Datasets and Leak-Free Spatial Partitioning. To evaluate cross-view synthesis under realistic global remote sensing conditions, we curate a large-scale, multi-sensor dataset spanning diverse biomes, rural topologies, and road networks across North America and Europe (More details in Appendix F:

Table 1: Quantitative comparison on the cross-view panorama generation task on CVUSA dataset. Bifröst relies only on low-resolution inputs at inference to achieve performance at par with the state-of-the-art models which use high resolution satellite images to generate the panoramas. Best overall results are in **bold**; best LR-only results are underlined. S: Sentinel-2 (10m), H: High-Res Satellite (0.3m), A: AlphaEarth.

Configuration	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑
Prior State-of-the-Art Baselines (Input - High-Res satellite image)				
ControlNet (Zhang et al., 2023)	22.55	0.51	12.83	0.32
CS2S (Ze et al., 2025)	21.30	**0.44**	**13.67**	**0.38**
CrossViewDiff (Li et al., 2024)	23.67	**0.44**	12.00	0.37
Bifröst (Ours: LR Inference + Cyclic PIC Training)				
S w/ cyclic	7.15	0.62	11.69	0.32
A w/ cyclic	**6.63**	0.59	12.14	0.33
S + A w/ cyclic	6.73	0.59	12.17	0.33

1. **Global Multi-Sensor Dataset:** The primary benchmark contains 556k paired instances comprising Google Street View (GSV) panoramas ($H_p \times W_p = 256 \times 1024$), 13-band Sentinel-2 L2A multispectral imagery (10 m/px), annual AlphaEarth 64-dimensional foundation embeddings (Brown et al., 2025), and auxiliary high-resolution aerial imagery (0.3 m/px, used strictly during training). To ensure valid real-world correspondence, we enforce a strict $\leq$ 7-day temporal window between Sentinel-2 captures and GSV panorama timestamps. We partition the data into 456k training pairs, 50k validation pairs, and 50k held-out test pairs. Crucially, to prevent geographic data leakage, train/val/test splits are partitioned across disjoint Google S2 Level-8 cells ($\approx 1300\text{km}^2$) (Larkin-York et al., 2015), ensuring that test locations share no spatial overlap or immediate border tiles with training regions.
2. **CVUSA**(Workman et al., 2015): To benchmark against conventional cross-view models that operate strictly on RGB inputs, we also evaluate on the standard CVUSA dataset, adhering to the official CVPR split (Zhai et al., 2017) ($\approx$ 35.5k training and 8.8k test image-satellite pairs) along with corresponding Sentinel-2 images and AlphaEarth embeddings fetched seperately.

Baselines: We benchmark *Bifröst* against four representative classes of cross-view synthesis models: (i) *Spatial layout conditioned diffusion*: ControlNet (Zhang et al., 2023) adapted to overhead cross-view conditioning; (ii) *Dedicated cross-view diffusion architectures*: CS2S (Ze et al., 2025) and CrossViewDiff (Li et al., 2024); (iii) *High-Resolution Upper Bounds*: Models conditioned on 0.3 m/px Google Maps aerial tiles (H), both standalone and fused with Sentinel-2 ($S + H$) and AlphaEarth ($S + H + A$), representing the empirical ceiling when fine-grained overhead imagery is available; (iv) *Low-Resolution Baselines*: Standard conditional diffusion models trained on low-resolution Sentinel-2 and AlphaEarth (S, A, $S + A$) without cyclic supervision.

Evaluation Protocol. We evaluate synthesized panoramas across four complementary axes: (1) *Perceptual Realism:* FID (Heusel et al., 2017) and LPIPS (Zhang et al., 2018); (2) *Photometric & Texture Fidelity:* PSNR, SSIM (Wang et al., 2004), and Haralick GLCM MAE (Haralick et al., 1973) across Contrast, Entropy, and Correlation; (3) *3D Geometry:* Zero-shot depth consistency (AbsRel, $\delta_1 < 1.25$) via DepthAnythingV2 (Yang et al., 2024); and (4) *Semantic Alignment:* we obtain mIoU between zero shot semantic segmentation masks of real and generated panoramas, using pretrained Mask2Former (Cheng et al., 2021) (details in Appendix E).

We conduct extensive component and modality ablations isolating the impact of the Perspective-Inversion Critic, coordinate grids, geo-temporal metadata, and explicit depth vs. multi-spectral cues; detailed results and discussion are provided in Appendix I (Table 6).

4.2 Quantitative Evaluations

We evaluate ***Bifröst*** against state-of-the-art cross-view synthesis frameworks and internal modality configurations to assess perceptual realism, surface texture fidelity, and 3D structural alignment. To isolate the impact of our Perspective-Inversion Critic (PIC), we group models by their inference-time input requirements in table 2.

Comparison with Prior State-of-the-Art. We first evaluate against established cross-view baselines: ControlNet (Zhang et al., 2023), CS2S (Ze et al., 2025), and CrossViewDiff (Li et al., 2024). For a rigorous comparison, all prior methods were retrained from a Stable Diffusion checkpoint on our identical geographic dataset splits. As shown in the top block of table 2, while ControlNet achieves a high PSNR (15.64 dB) characteristic of mean-seeking pixel regression, it struggles perceptually (FID $= 24.95$) and yields poor geometric layout reconstruction (AbsRel $= 0.226$). Similarly, existing cross-view models struggle on unconstrained rural and road topologies, exhibiting elevated GLCM texture errors and geometric depth distortion (AbsRel ≥ 0.221).

Table 2: Quantitative comparison on the cross-view panorama generation task on our Global Multi-Sensor Dataset. Models are grouped by their inference-time data requirements. Bifröst relies only on low-resolution inputs at inference, yet bridges the structural gap (AbsRel) to high-resolution baselines and achieves state-of-the-art FID. Best overall results are in **bold**; best LR-only results are underlined. S: Sentinel-2 (10m), H: High-Res Satellite (0.3m), A: AlphaEarth.

Method / Active Modalities	Perceptual & Photometric				GLCM Texture Error (MAE)			Depth Geometry	
	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	Contrast ↓	Entropy ↓	Correl. ↓	AbsRel ↓	δ_1 ↑
Prior State-of-the-Art Baselines (Retrained on our multi-sensor benchmark; Input - High-Res satellite image)									
ControlNet (Zhang et al., 2023)	24.9512	0.5957	**15.6396**	0.3602	0.4872	1.7140	0.0165	0.2260	0.7687
CS2S (Ze et al., 2025)	18.8600	0.6879	13.5698	0.3015	0.3736	0.3445	0.0342	0.3181	0.6929
CrossViewDiff (Li et al., 2024)	17.4773	0.6411	14.1714	0.3215	0.2785	0.2927	0.0257	0.2212	0.7506
HR Dependent Baselines (Require 0.3m/px HR data at Inference)									
HR Satellite Only (H)	4.3554	0.4916	13.9490	0.4569	0.1886	0.2638	0.0097	0.1010	0.9077
A + H	4.2418	0.4917	13.9416	0.4567	0.1907	0.2682	0.0097	0.1021	0.9065
S + H	3.9274	0.4912	13.9865	0.4569	**0.1880**	**0.2634**	0.0096	**0.1007**	0.9079
S + H + A	3.9933	**0.4892**	14.0333	**0.4589**	0.1889	0.2663	**0.0094**	0.1010	0.9075
Low-Resolution Baselines (Ubiquitous 10m/px Inference)									
Sentinel-2 Only (S)	4.3079	0.5338	13.0272	0.4271	0.2665	0.3490	0.0114	0.1356	0.8841
AlphaEarth Only (A)	4.6529	0.5363	13.0987	0.4333	0.2541	0.3352	0.0112	0.1390	0.8837
S + A	4.4003	0.5362	13.0873	0.4330	0.2514	0.3397	0.0111	0.1385	0.8830
Bifröst (Ours: LR Inference + Cyclic PIC Training)									
S w/ cyclic	3.8324	0.5183	12.8512	0.4385	0.2503	0.3280	0.0123	0.1174	0.9012
A w/ cyclic	**3.7572**	0.5167	12.9110	0.4384	0.2284	0.3065	0.0119	0.1180	0.8983
S + A w/ cyclic	3.8707	0.5038	13.2401	0.4478	0.2069	0.2855	0.0113	0.1057	**0.9086**

Benchmarking on Standard CVUSA. As shown in table 1, the structural internalization enabled by our cyclic PIC objective transfers robustly to the standard CVUSA benchmark (Workman et al., 2015; Zhai et al., 2017). Baselines conditioned directly on sub-meter aerial imagery (0.3 m/px) hold a slight edge on pairwise reference metrics like PSNR (13.67 dB vs. 12.17 dB), SSIM (0.38 vs. 0.33) and LPIPS (0.44 vs. 0.59). This gap is expected: in addition to the extreme test-time resolution disparity (10 m/px Sentinel-2 vs. 0.3 m/px aerial), CVUSA lacks per-image acquisition dates, preventing exact temporal co-registration between our retrofitted satellite inputs and the historical street captures (see Appendix F.2. Despite these constraints, *Bifröst* maintains comparable photometric fidelity while substantially outperforming high-resolution baselines, improving FID from 21.30 to $\mathbf{6.73}$.

The Low-Resolution Ceiling and the Temporal Inconsistency. Grouping our ablations (table 2) by sensor constraints exposes two fundamental empirical phenomena:

- **The Low-Resolution Ceiling:** Models conditioned solely on ubiquitous low-resolution data (S, A, and $S + A$) hit a rigid structural plateau. Across all three LR combinations, relative depth error stagnates at AbsRel ≈ 0.136–0.139, accompanied by degraded GLCM texture contrast (0.251–0.267). Because 10 m/px Sentinel-2 pixels blend narrow roadways and fence lines into homogeneous spectral averages, standard feed-forward conditioning lacks the high-frequency cues required to resolve layout boundaries.
- **The Temporal Inconsistency:** Conversely, high-resolution models (0.3 m/px) set the empirical ceiling for structural alignment (AbsRel ≈ 0.101). Furthermore, annualized AlphaEarth embeddings aggregate multi-temporal features, introducing temporal friction when forced to reconcile with instantaneous, fine-grained aerial snapshots in early network layers.

Supervising the generator with our cyclic PIC objective decisively breaks the low-resolution plateau. As detailed in the bottom block of table 2, cyclic training systematically improves across multiple perceptual and geometric metrics for every low-resolution permutation (S, A, and $S + A$).

Crucially, our full model, *Bifröst* ($S + A$ w/ cyclic), relies strictly on 10 m/px inputs at test time, yet drives relative depth error down to $\mathbf{0.106}$ and achieves a threshold accuracy of $\boldsymbol{\delta_1 = 0.909}$, matching



Figure 2: **Qualitative comparison of Bifröst's cross-view synthesis.** The model successfully internalizes structural layouts without high-resolution geometry at inference. From left to right: the RGB patch from Sentinel-2, the ground truth panorama, and the Bifröst (S+A) generated panorama. (a) Rural road with dense foliage. (b) Open field. (c) and (d) Highways. (e) Harvested terrain.

the 0.3 m/px HR-conditioned upper bound ($\delta_1 = 0.908$). Furthermore, PIC supervision significantly refines local surface synthesis, lowering GLCM texture contrast to 0.207 and achieving an FID of $\mathbf{3.87}$. These results confirm that *Bifröst* learns detailed geometric priors to enable high-fidelity synthesis without high-resolution inference imagery.

Evaluation of Critic Model To confirm F_ϕ converges to a reliable geometric estimator prior to freezing, we benchmark its performance on the isolated inversion task (Street View + C_{lr} $\rightarrow$ HR Satellite) on the 50k held-out test split.

Table 3: **Quantitative evaluation of the pre-trained Perspective-Inversion Critic (F_ϕ).** The task evaluates inverse nadir synthesis (0.3 m/px) conditioned on ground-level Street View (SV) and low-resolution satellite inputs (S: 10 m/px Sentinel-2; A: AlphaEarth).

Conditioning Inputs	FID ↓	LPIPS ↓	PSNR (dB) ↑	SSIM ↑
SV + Sentinel-2 (S)	11.0124	0.5447	15.8843	0.2823
SV + AlphaEarth (A)	**10.5327**	0.5436	15.5763	**0.2838**
SV + Sentinel-2 + AlphaEarth (S+A)	10.6130	**0.5427**	**15.8943**	0.2814

5 Conclusion

In this work, we introduced *Bifröst*, a cross-view generative framework that achieves global-scale 360° street-view synthesis without requiring high-resolution aerial imagery at inference. By shifting the paradigm from explicit high-resolution conditioning to *Variational Cyclic Internalization*, *Bifröst* leverages a frozen Perspective-Inversion Critic during training to internalize fine-grained structural priors into a Coordinate-Aware Cross-Attention backbone, relying solely on globally accessible 10 m/px Sentinel-2 multi-spectral observations and AlphaEarth embeddings at test time. Our approach closes the resolution gap, reducing relative depth error down to $\mathbf{0.105}$ (nearing the 0.101 high-resolution upper bound) and establishing a state-of-the-art FID of $\mathbf{3.87}$.

By proving that high-frequency terrestrial geometry can be recovered from ubiquitous low-resolution satellite inputs, *Bifröst* democratizes planetary-scale immersive view synthesis beyond proprietary aerial monopolies. We discuss nadir blindspots, sub-pixel foreground limits, temporal dynamics, and failure modes in section H.

REPRODUCIBILITY STATEMENT

We have provided detailed explanations on the architecture (Sec. 3), loss functions (Appendix A) and implementation details (Appendix D) in order to enable a reader to reproduce the work. We are limited by organizational regulations and are not at liberty to release the 456k dataset and code for the work. However, we have included comparisons against publicly reported results on state of the art methods, on the publicly available CVUSA dataset (Table 1). On this public dataset, we only include publicly available Sentinel-2 and AlphaEarth embeddings (as described in Appendix F.2) for the training and inference, to enable readers to reproduce results on this public benchmark. We have also provided details on the evaluation metrics used in Appendix E.

AI USE STATEMENT

In this work, we used generative AI tools to formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, refine hypotheses, and write code/implement methods. Additionally, we used these tools to suggest experimental parameters, create artifacts (For e.g., fig. 3 and fig. 4), draft parts of the research paper, summarize and analyze existing literature, brainstorm, search for information, edit the manuscript to improve readability, identify relevant literature, format references, suggest manuscript structure, and propose titles and keywords.

We have not used generative AI tools to help develop the models or conceptual frameworks, to generate visualizations to support qualitative and thematic data analysis, or to design and provide feedback on research methodology and experiments. Tasks involving the generation of synthetic datasets, data cleaning and reformatting, and translation were not applicable to this work.

We have thoroughly reviewed all AI-assisted work. All manuscript changes and mathematical claims proposed by LLMs were manually verified by two authors. Similarly, all AI-generated code was tested, verified, and utilized by multiple authors. We take full responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

David Barber and Felix Agakov. Information maximization in noisy channels : A variational approach. In S. Thrun, L. Saul, and B. Schölkopf (eds.), *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2003. URL https://proceedings.neurips.cc/paper_files/paper/2003/file/a6ea8471c120fe8cc35a2954c9b9c595-Paper.pdf.

Christopher Brown, Steven Brumby, Brookie Guzder-Williams, Tanya Birch, Samantha Hyde, Joseph Mazzariello, Wanda Czerwinski, Valerie Pasquarella, Robert Haertel, Simon Ilyushchenko, Kurt Schwehr, Mikaela Weisse, Fred Stolle, Craig Hanson, Oliver Guinan, Rebecca Moore, and Alexander Tait. Dynamic world, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9:251, 06 2022. doi: 10.1038/s41597-022-01307-4.

Christopher F Brown, Michal Kazmierski, et al. Alphaearth foundations: An embedding field model for accurate and efficient global mapping from sparse label data. *arXiv preprint arXiv:2507.22291*, 2025. URL https://arxiv.org/abs/2507.22291.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CoRR*, abs/2112.01527, 2021. URL https://arxiv.org/abs/2112.01527.

Fabian Deuser, Wejdene Mansour, Hao Li, Konrad Habel, Martin Werner, and Norbert Oswald. Temporal resilience in geo-localization: Adapting to the continuous evolution of urban and rural environments. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pp. 442–451, 2025. doi: 10.1109/WACVW65960.2025.00055.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181. URL https://doi.org/10.1198/jasa.2011.tm11181. PMID: 22505788.

Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. MagicDrive: Street view generation with diverse 3d geometry control. In *International Conference on Learning Representations*, 2024.

Tianxiong Gao, Shuyan Zhang, Wutao Yao, Erping Shang, Jin Yang, Yong Ma, and Yan Ma. Srct: Structure-preserving method for sub-meter remote sensing image super-resolution. *Sensors*, 26(2): 733, 2026.

Robert M. Haralick, K. Shanmugam, and Its'Hak Dinstein. Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-3(6):610–621, 1973. doi: 10.1109/TSMC.1973.4309314.

Jonathan Heek, Emiel Hoogeboom, and Tim Salimans. Multistep consistency models, 2024. URL https://arxiv.org/abs/2403.06807.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. URL http://arxiv.org/abs/1706.08500.

Emiel Hoogeboom, Jonathan Heek, and Tim Salimans. Simple diffusion: end-to-end diffusion for high resolution images. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

Johannes Jakubik, Felix Yang, Benedikt Blumenstiel, Erik Scheurer, Rocco Sedona, Stefano Maurogiovanni, Jente Bosmans, Nikolaos Dionelis, Valerio Marsocci, Niklas Kopp, Rahul Ramachandran, Paolo Fraccaro, Thomas Brunschwiler, Gabriele Cavallaro, Juan Bernabe-Moreno, and Nicolas Longépé. Terramind: Large-scale generative multimodality for earth observation, 2026. URL https://arxiv.org/abs/2504.11171.

Lihan Jiang, Kerui Ren, Mulin Yu, Linning Xu, Junting Dong, Tao Lu, Feng Zhao, Dahua Lin, and Bo Dai. Horizon-gs: Unified 3d gaussian splatting for large-scale aerial-to-ground scenes. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26789–26799, 2025.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. URL https://arxiv.org/abs/2308.04079.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 5586–5604, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/16c3c941409d0581286eff49b180930f-Paper-Conference.pdf.

Diederik P. Kingma and Ruiqi Gao. Understanding diffusion objectives as the elbo with simple data augmentation, 2023. URL https://arxiv.org/abs/2303.00848.

Diederik P Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. On density estimation with diffusion models. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=2LdBqxc1Yv.

Dan Larkin-York, Google Inc., Koordinates Limited, Mike Playle, and Tiago Brito. S2 geometry library. https://github.com/google/s2geometry, 2015. URL https://github.com/google/s2geometry.

Jie-Ying Lee, Yi-Ruei Liu, Shr-Ruei Tsai, Wei-Cheng Chang, Chung-Ho Wu, Jiewen Chan, Zhenjun Zhao, Chieh Hubert Lin, and Yu-Lun Liu. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery, 2026. URL https://arxiv.org/abs/2510.15869.

Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhimeng Zheng, Zilong Huang, Dahua Lin, and Conghui He. Crossviewdiff: A cross-view diffusion model for satellite-to-street view synthesis, 2024. URL https://arxiv.org/abs/2408.14765.

Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R. Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 5171–5180. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/poole19a.html.

Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3683–3692, October 2023.

Ming Qian, Bin Tan, Qiuyu Wang, Xianwei Zheng, Hanjiang Xiong, Gui-Song Xia, Yujun Shen, and Nan Xue. Seeing through satellite images at street views, 2026a. URL https://arxiv.org/abs/2505.17001.

Ming Qian, Zimin Xia, Changkun Liu, Shuailei Ma, Wen Wang, Zeran Ke, Bin Tan, Hang Zhang, and Gui-Song Xia. Sat3DGen: Comprehensive street-level 3d scene generation from single satellite image. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=E7JzkZCofa.

Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Y. Shi, D. Campbell, X. Yu, and H. Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 44 (12):10009–10022, 2022.

Jordi Laguarta Soler, Thomas Friedel, and Sherrie Wang. Combining deep learning and street view imagery to map smallholder crop types. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 22202–22212, 2024.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *CoRR*, abs/2006.10739, 2020. URL https://arxiv.org/abs/2006.10739.

Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J. Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

Vladimir Naumovich Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural networks : the official journal of the International Neural Network Society*, 22 5-6:544–57, 2009. URL https://api.semanticscholar.org/CorpusID:12342641.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer, 2025a. URL https://arxiv.org/abs/2503.11651.

Qiwei Wang, Shaoxun Wu, and Yujiao Shi. Bevsplat: Resolving height ambiguity via feature-based gaussian primitives for weakly-supervised cross-view localization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Z. Wang, E.P. Simoncelli, and A.C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems Computers, 2003*, volume 2, pp. 1398–1402 Vol.2, 2003. doi: 10.1109/ACSSC.2003.1292216.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. doi: 10.1109/TIP.2003.819861.

Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 1–9, October 2015. doi: 10.1109/ICCV.2015.451. Acceptance rate: 30.3%.

Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 25:3546–3559, 2023. doi: 10.1109/TMM.2022.3162474.

Zhexiao Xiong, Xin Xing, Scott Workman, Subash Khanal, and Nathan Jacobs. Mixed-view panorama synthesis using geospatially guided diffusion. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=ylUVRikhTL.

Ningli Xu and Rongjun Qin. Satellite to groundscape – large-scale consistent ground view generation from satellite views, 2025. URL https://arxiv.org/abs/2504.15786.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2, 2024. URL https://arxiv.org/abs/2406.09414.

Xianghui Ze, Zhenbo Song, Qiwei Wang, Jianfeng Lu, and Yujiao Shi. Controllable satellite-to-street-view synthesis with precise pose alignment and zero-shot environmental control. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=f92M45YRfh.

Xianghui Ze, Beiyi Zhu, Zhenbo Song, Jianfeng Lu, and Yujiao Shi. SatDreamer360: Multiview-consistent generation of ground-level scenes from satellite imagery. In *International Conference on Learning Representations*, 2026.

Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

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 (ICCV)*, pp. 3836–3847, October 2023.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018. URL http://arxiv.org/abs/1801.03924.

Yancheng Zhang, Xiaohan Zhang, Guangyu Sun, Zonglin Lyu, Safwan Wshah, and Chen Chen. Geo^2: Geometry-guided cross-view geo-localization and image synthesis, 2026. URL https://arxiv.org/abs/2603.25819.

Hang Zhao, Orazio Gallo, Iuri Frosio, and Jan Kautz. Loss functions for image restoration with neural networks. *IEEE Transactions on Computational Imaging*, 3(1):47–57, 2017. doi: 10.1109/TCI.2016.2644865.

A GIBBS FORMULATION OF THE COMPOSITE OBJECTIVE

In Section 3.4, we introduced the cyclic objective regularizing the generator G_Θ using a frozen Perspective-Inversion Critic (PIC) F_ϕ. Here, we formally prove that our composite loss corresponds to the negative log-likelihood of a shift-invariant Gibbs distribution, bypassing the blurriness of isotropic Gaussian assumptions.

While F_ϕ estimates the posterior mean, supervising the generator with a secondary Gaussian L_2 loss forces mean-seeking blur and ignores spatial covariance. We therefore parameterize the variational predictive density q_ϕ as a Gibbs distribution:

$$q_\phi(\mathbf{c}_{hr} \mid \hat{\mathbf{x}}, \mathbf{c}_{lr}) = \frac{1}{Z(\hat{\mathbf{y}})} \exp\left(-E(\mathbf{c}_{hr}, \hat{\mathbf{y}})\right), \quad \text{where } \hat{\mathbf{y}} = F_\phi(\hat{\mathbf{x}}, \mathbf{c}_{lr}), \tag{7}$$

with the composite energy function:

$$E(\mathbf{y}, \hat{\mathbf{y}}) = w_{l1} \left\|(\mathbf{y} - \hat{\mathbf{y}}) \odot \mathbf{\Lambda}\right\|_1 + w_{\text{ssim}} \left(1 - \text{MS-SSIM}(\mathbf{y}, \hat{\mathbf{y}})\right). \tag{8}$$

Here, $\mathbf{\Lambda}$ is a static boundary-weighting edge mask derived deterministically from the target $\mathbf{c}_{hr}$.

Observation 1 (Approximate Partition Function Invariance). *Under the assumption of local contrast preservation, the partition function $Z(\hat{\mathbf{y}}) = \int \exp(-E(\mathbf{y}, \hat{\mathbf{y}}))d\mathbf{y}$ is approximately independent of the generator parameters θ. Consequently, minimizing the cyclic loss $\mathcal{L}_{F_\phi}$ acts as a tractable surrogate for maximizing the variational lower bound on $I(\mathbf{c}_{hr}; \hat{\mathbf{x}})$.*

Proof. Under the translation $\mathbf{u} = \mathbf{y} - \hat{\mathbf{y}}$, the Jacobian determinant is unity ($|\det \mathbf{I}| = 1$). While the boundary-weighted L_1 norm depends strictly on the spatial residual $\mathbf{u}$, the MS-SSIM (Wang et al., 2003) term introduces a local variance dependency ($\sigma_{\mathbf{y}}^2 + \sigma_{\hat{\mathbf{y}}}^2$) that is not strictly shift-invariant.

However, (Zhao et al., 2017) demonstrate that optimizing a joint L_1 and MS-SSIM objective decouples local luminance from structure, actively penalizing contrast mismatches such that the generated local variance converges to the target ($\sigma_{\hat{\mathbf{y}}} \approx \sigma_{\mathbf{y}}$). Furthermore, because the pre-trained Perspective-Inversion Critic (F_ϕ) is optimized via diffusion score matching, its structural predictions inherently recover the true data variance. Under this practical generative condition of matched contrast, the variance terms in the MS-SSIM denominator act as scaling constants relative to the spatial residual, allowing the energy function to be approximated as:

$$E(\mathbf{y}, \hat{\mathbf{y}}) \approx w_{l1} \|\mathbf{u} \odot \mathbf{\Lambda}\|_1 + w_{\text{ssim}}\, g(\mathbf{u}) = \tilde{E}(\mathbf{u}). \tag{9}$$

Changing variables in the partition function under this approximation gives:

$$Z(\hat{\mathbf{y}}) \approx \int_{\mathbb{R}^D} \exp\left(-\tilde{E}(\mathbf{u})\right) d\mathbf{u} = Z_0 = \text{const}. \tag{10}$$

Because Z_0 and $\mathbf{\Lambda}$ (derived from fixed $\mathbf{c}_{hr}$) are approximately invariant to $\hat{\mathbf{x}}$, taking the negative log-likelihood yields:

$$-\log q_\phi(\mathbf{c}_{hr} \mid \hat{\mathbf{x}}, \mathbf{c}_{lr}) \approx w_{l1} \left\|(\mathbf{c}_{hr} - F_\phi(\hat{\mathbf{x}}, \mathbf{c}_{lr})) \odot \mathbf{\Lambda}\right\|_1 + w_{\text{ssim}} \left(1 - \text{MS-SSIM}(\mathbf{c}_{hr}, F_\phi(\hat{\mathbf{x}}, \mathbf{c}_{lr}))\right) + \text{const}. \tag{11}$$

Thus, $\nabla_\Theta \log Z \approx 0$, and $\min_\Theta \mathcal{L}_{F_\phi}$ acts as a tractable surrogate to maximize $\mathbb{E}[\log q_\phi]$, which monotonically tightens the lower bound on $I(\mathbf{c}_{hr}; \hat{\mathbf{x}})$. □

Remark on Training Dynamics. It is important to note that the assumption of local contrast preservation ($\sigma_{\hat{y}} \approx \sigma_y$) is an asymptotic property that holds tightly as the generator converges. During the earliest stages of training, the generator's outputs lack physical contrast, meaning $Z(\hat{y})$ strictly depends on Θ. However, in practice, our dynamic critic gating mechanism (see Appendix D.4) shields the generator from critic gradients during high-noise, unformed layout phases. By the time the critic loss is actively applied and the network reaches the low-noise regime, the generator has established sufficient base contrast such that $\nabla_\Theta \log Z \approx 0$ serves as a stable, tractable approximation for optimization.

B GEOMETRIC INDUCTIVE BIAS & ATTENTION EQUIVARIANCE

Cross-view synthesis between overhead satellite imagery and ground-level street-view panoramas is fundamentally constrained by a drastic geometric perspective change: a planar, top-down orthographic observation must be mapped onto an outward-looking spherical manifold. In this section, we formalize the coordinate systems introduced in Section 3.2 and rigorously prove that our Coordinate-Aware Cross-Attention block natively spans rotational-to-translational equivariance without requiring expensive group-equivariant convolutions or constrained architectural operators.

B.1 DERIVATION OF POLAR COORDINATE PROJECTIONS FROM NADIR GRIDS

Let the satellite feature tensor be denoted by $\mathbf{F}_{\text{sat}} \in \mathbb{R}^{H_s \times W_s \times C_s}$, where $(i, j) \in \{0, \dots, H_s - 1\} \times \{0, \dots, W_s - 1\}$ index discrete spatial grid locations on the nadir plane. Under the standard geospatial orientation convention (North-up, East-right), let the metric center of the observation patch be:

$$(c_y, c_x) = \left(\frac{H_s - 1}{2}, \frac{W_s - 1}{2} \right). \tag{12}$$

The discrete spatial offsets relative to the camera origin (ground-level projection center) are:

$$\Delta y_{ij} = -(i - c_y), \quad \Delta x_{ij} = j - c_x, \tag{13}$$

where Δy increases toward geographic North and Δx increases toward geographic fig. 3.

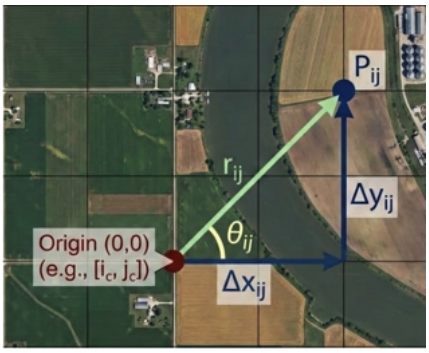

Figure 3: Illustration of satellite to polar coordinate transformation.

To map these Euclidean grid coordinates to an azimuth-range reference frame centered at the camera viewpoint, we define the normalized radial distance $r_{ij} \in [0, 1]$ and the continuous azimuthal angle $\theta^s_{ij} \in (-\pi, \pi]$:

$$r_{ij} = \frac{\sqrt{(\Delta x_{ij})^2 + (\Delta y_{ij})^2}}{\max_{u,v} \sqrt{(\Delta x_{uv})^2 + (\Delta y_{uv})^2}}, \tag{14}$$

$$\theta^s_{ij} = \arctan2(\Delta x_{ij}, \Delta y_{ij}). \tag{15}$$

By defining $\theta^s_{ij} = \arctan2(\Delta x_{ij}, \Delta y_{ij})$, geographic North ($\Delta x = 0, \Delta y > 0$) maps to $\theta = 0$, East maps to $\theta = \pi/2$, South maps to $\pm\pi$, and West maps to $-\pi/2$, establishing a standard clockwise azimuthal reference frame.

To eliminate the coordinate boundary discontinuity at $\pm\pi$ and provide smooth, continuous positional features for learned linear layers, we map θ^s_{ij} to the unit circle $\mathbb{S}^1$ via first-order harmonic embeddings:

$$\mathbf{C}_{\text{polar}}(i, j) = \left[r_{ij},\ \sin\theta^s_{ij},\ \cos\theta^s_{ij} \right]^\top \in \mathbb{R}^3. \tag{16}$$

B.2 DERIVATION OF SPHERICAL PROJECTION FOR EQUIRECTANGULAR PANORAMAS

Let the intermediate diffusion query feature map be $\mathbf{X} \in \mathbb{R}^{H_l \times W_l \times C_l}$, representing a 360° cylindrical panorama under an equirectangular projection. Let $(u, v) \in \{0, \dots, H_l - 1\} \times \{0, \dots, W_l - 1\}$ index the vertical (elevation) and horizontal (azimuthal) coordinates of the query tokens.

In equirectangular geometry, the horizontal pixel axis corresponds linearly to longitude (azimuth), and the vertical pixel axis corresponds linearly to latitude (elevation). Assuming the panorama is centered such that the optical axis of the central column ($v = \lfloor W_l/2 \rfloor$) aligns with geographic North ($\theta = 0$) fig. 4:

$$\theta_v^p = 2\pi \left(\frac{v - W_l/2}{W_l} \right) \in [-\pi, \pi), \tag{17}$$

which preserves identical clockwise orientation: moving rightward (v increasing) corresponds to looking toward the East ($\theta^p > 0$).

The vertical pixel index u parameterizes the optical elevation angle ϕ_u^p, mapped linearly across the camera sensor's vertical field of view (FOV) $[\phi_{\min}, \phi_{\max}]$:

$$\phi_u^p = \phi_{\max} - u \left(\frac{\phi_{\max} - \phi_{\min}}{H_l - 1} \right) \in \left[-\frac{\pi}{2}, \frac{\pi}{2} \right], \tag{18}$$

where $u = 0$ corresponds to the zenith ($\phi_{\max}$) and $u = H_l - 1$ corresponds to the nadir ground plane ($\phi_{\min}$).

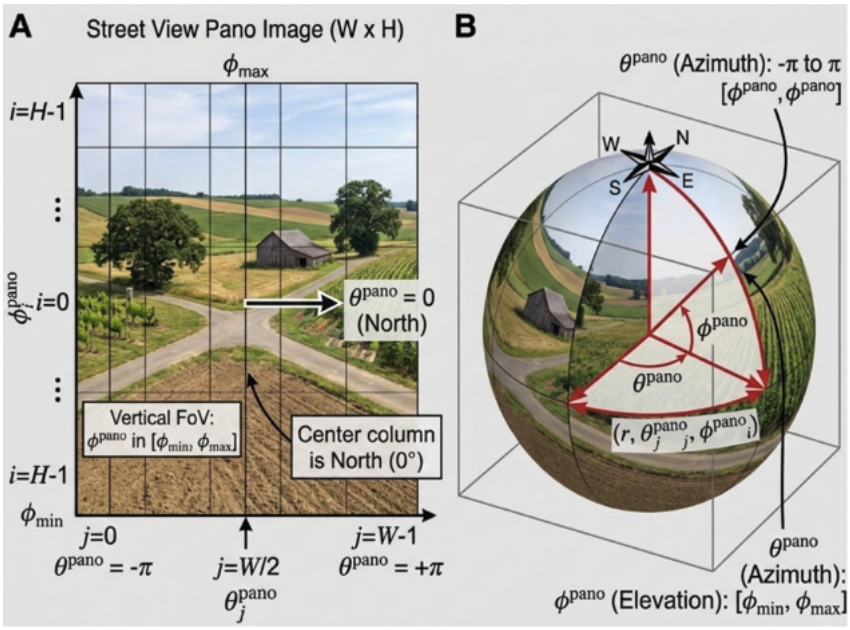

Figure 4: Illustration of street view panorama to polar coordinate transformation

The continuous spherical embedding for panoramic query tokens is therefore constructed as:

$$\mathbf{C}_{\text{spherical}}(u, v) = [\sin\theta_v^p, \cos\theta_v^p, \sin\phi_u^p, \cos\phi_u^p]^\top \in \mathbb{R}^4. \tag{19}$$

B.3 Proof of Rotational-to-Translational Equivariance

We analyze how our coordinate-aware attention mechanism maps 2D rotations of the overhead satellite imagery to horizontal translations of the equirectangular panorama.

Proposition 1 (Representational Span for Equivariance). *Let $\mathbf{F}_{sat}(r, \theta_s)$ be the polar-indexed satellite features and $\mathbf{X}(u, \theta_p)$ be the panorama query features. There exists a parameterization of the linear projection matrices $(\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V)$ such that a clockwise planar rotation of the satellite input by α induces an exact horizontal translation of the output panorama by $\Delta\theta = \alpha$.*

Proof. **1. Decoupled Spatial Subspace:** Let the linear projection matrices partition feature and coordinate channels independently:

$$\mathbf{W}_Q = \begin{bmatrix} \mathbf{W}_Q^{\text{feat}} & \mathbf{W}_Q^{\text{coord}} \end{bmatrix}, \quad \mathbf{W}_K = \begin{bmatrix} \mathbf{W}_K^{\text{feat}} & \mathbf{W}_K^{\text{coord}} \end{bmatrix}. \tag{20}$$

By setting cross-terms between coordinates and features to zero, the unnormalized attention logit splits additively into a semantic score and a geometric score: $S = S_{\text{feat}} + S_{\text{geom}}$.

2. Relative Phase Extraction: Let $\mathbf{W}_Q^{\text{coord}}$ and $\mathbf{W}_K^{\text{coord}}$ project the azimuthal harmonics $(\sin\theta, \cos\theta)$ with scale $\gamma > 0$ and zero out orthogonal coordinate dimensions:

$$S_{\text{geom}}(\theta_p, \theta_s) = \frac{\gamma^2}{\sqrt{d}} \left(\sin\theta_p \sin\theta_s + \cos\theta_p \cos\theta_s\right) = \frac{\gamma^2}{\sqrt{d}} \cos(\theta_p - \theta_s). \tag{21}$$

Applying a rotation $\theta_s \mapsto \theta_s - \alpha$ to the satellite inputs and a matching shift $\theta_p \mapsto \theta_p - \alpha$ to the queries yields:

$$S_{\text{geom}}(\theta_p - \alpha,\, \theta_s - \alpha) = \frac{\gamma^2}{\sqrt{d}} \cos\left((\theta_p - \alpha) - (\theta_s - \alpha)\right) = S_{\text{geom}}(\theta_p, \theta_s). \tag{22}$$

3. Softmax and Value Invariance: Because the 2D rotation of the satellite keys represents a permutation of tokens over a closed circular grid, the denominator sum in the Softmax operator is invariant under rotation:

$$\sum_k \exp\left(S(\theta_p - \alpha,\, \theta_{s,k} - \alpha)\right) = \sum_k \exp\left(S(\theta_p,\, \theta_{s,k})\right). \tag{23}$$

Hence, the normalized attention weights satisfy $A(\theta_p - \alpha, \theta_s - \alpha) = A(\theta_p, \theta_s)$. Finally, by setting the coordinate slice of the value projection to zero ($\mathbf{W}_V^{\text{coord}} = \mathbf{0}$), the output features depend purely on rotated satellite visual features, ensuring the aggregated representation shifts cyclically by $\Delta\theta = \alpha$. □

C Perspective-Inversion Critic (PIC) Pre-training

This section details the architecture, training dynamics, and empirical convergence of the Perspective-Inversion Critic (F_ϕ). During Bifröst's cyclic training phase, F_ϕ is held frozen to supervise the generator G_Θ without adversarial instability by mapping synthesized equirectangular panoramas $\hat{\mathbf{x}} \in \mathbb{R}^{H_p \times W_p \times 3}$ and ubiquitous context $C_{lr} = \{\mathbf{c}_{ms}, \mathbf{c}_{ae}, \mathbf{c}_{geo}\}$ back to nadir aerial space $\hat{\mathbf{c}}_{hr} \in \mathbb{R}^{H_h \times W_h \times 3}$ (0.3 m/px).

C.1 Network Architecture and Training Objective

Architectural Symmetry and Coordinate Inversion. To maintain structural symmetry with G_Θ, the critic F_ϕ uses an identical U-ViT continuous-time diffusion backbone and dual-stream FiLM conditioning. Because F_ϕ predicts an orthographic nadir aerial patch rather than a spherical projection, the coordinate conditioning frames are inverted:

- **Queries:** Diffusion queries $\mathbf{Z}_t \in \mathbb{R}^{H_s \times W_s \times C}$ reside on an overhead Cartesian grid and are augmented with polar coordinates $\mathbf{C}_{polar} \in \mathbb{R}^3$ centered at the camera nadir point (derivation in Appendix B.1).
- **Conditioning:** Ground-level panorama features $\mathbf{S}_{sv} \in \mathbb{R}^{H_l \times W_l \times C}$ are embedded with equirectangular spherical coordinates $\mathbf{C}_{spherical} \in \mathbb{R}^4$ (Appendix B.2). The conditioning from low resolution satellite data follows the same protocol as the Queries.

Decoupled Sequential Attention. To avoid the $\mathcal{O}((N_{sv} + N_{lr})^2)$ cost and gradient interference of joint concatenation, F_ϕ routes conditioning features through sequential cross-attention layers:

$$\mathbf{X}' = \mathbf{X} + \mathbf{W}_{out}^{sv} \cdot \text{Attn}(\mathbf{X}\mathbf{W}_Q^{sv},\, \mathbf{S}_{sv}\mathbf{W}_K^{sv},\, \mathbf{S}_{sv}\mathbf{W}_V^{sv}), \tag{24}$$

$$\mathbf{X}'' = \mathbf{X}' + \mathbf{W}_{out}^{lr} \cdot \text{Attn}(\mathbf{X}'\mathbf{W}_Q^{lr},\, \mathbf{S}_{lr}\mathbf{W}_K^{lr},\, \mathbf{S}_{lr}\mathbf{W}_V^{lr}), \tag{25}$$

where $\mathbf{X} = [\mathbf{Z}_t \parallel_{chan} \mathbf{C}_{polar}]$. This guarantees high-frequency geometry from $\mathbf{S}_{sv}$ is locked before macroscopic spectral context from $\mathbf{S}_{lr} = [\mathbf{f}_{ms} \parallel_{seq} \mathbf{f}_{ae}]$ is integrated.

Continuous-Time Objective & Stationary Gradients. Overhead nadir imagery lacks an elevation-dependent horizon, so the spatial horizon weight is set to uniform ($\Omega = \mathbf{I}$). F_ϕ is optimized via score-matching:

$$\mathcal{L}_{PIC}(\phi) = \mathbb{E}_{t,\mathbf{c}_{hr},\boldsymbol{\epsilon}}\left[w(t)\left\|\mathbf{c}_{hr} - F_\phi(\mathbf{z}_t, \mathbf{x}_{sv}, C_{lr}, t)\right\|_2^2\right]. \tag{26}$$

By Tweedie's formula, the converged network directly estimates the conditional posterior expectation $\mathbb{E}[\mathbf{c}_{hr} \mid \mathbf{z}_t, \mathbf{x}_{sv}, C_{lr}]$. When frozen during generator training, evaluating synthesized panoramas $\hat{\mathbf{x}}$ yields stationary, non-adversarial regression gradients:

$$\nabla_{\hat{\mathbf{x}}}\mathcal{L}_{F_\phi} = \nabla_{\hat{\mathbf{x}}}\left(w_{\ell_1}\left\|F_\phi(\mathbf{z}_{t_{critic}}, \hat{\mathbf{x}}, C_{lr}) - \mathbf{c}_{hr}\right\|_1 \odot \mathbf{\Lambda} + w_{ssim}(1 - \text{MS-SSIM})\right), \tag{27}$$

completely sidestepping the discriminator drift and mode collapse of adversarial cyclic training.

Training Configuration. F_ϕ is pre-trained for 300,000 steps across 16 NVIDIA A100 (80GB) accelerators with batch size 512 using AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$, weight decay 10^{-4}). The learning rate follows a 2,000-step linear warmup to a peak of 10^{-4}, annealed via cosine decay. Targets are 256×256 aerial patches (0.3 m/px).

D Implementation and Training Details

D.1 Dual-Stream FiLM Conditioning and Fourier Frequency Encodings

Global temporal diffusion signals and geographical context are modulated into the U-ViT feature maps using Feature-wise Linear Modulation (FiLM) (Perez et al., 2018). To prevent cross-modal interference, the conditioning pipeline operates as two independent streams before joint projection.

Sinusoidal Fourier Embedding. Scalar inputs $s \in \mathbb{R}$ are projected into a d_f-dimensional continuous Fourier feature space (Tancik et al., 2020):

$$\mathcal{F}_{emb}(s) = \left[\sin(2^0\pi s), \cos(2^0\pi s), \ldots, \sin(2^{K-1}\pi s), \cos(2^{K-1}\pi s)\right]^\top \in \mathbb{R}^{d_f}, \tag{28}$$

where $d_f = 2K$ with K octaves.

Stream 1: Signal-Level Context ($\mathbf{e}_t$). Rather than embedding the raw scalar timestep $t \in [0, 1]$, the network embeds the exact continuous signal amplitude $\alpha_t \in (0, 1)$ extracted from the shifted log-SNR schedule (Section D.2):

$$\mathbf{e}_t = \text{MLP}_{noise}\left(\mathcal{F}_{emb}(\alpha_t)\right) \tag{29}$$

Stream 2: Geo-Temporal Prior ($\mathbf{h}_{geo}$). The five-tuple scalar metadata $\mathbf{c}_{geo} = (\theta, \phi, d, m, y)$ is embedded using independent Fourier projections, concatenated, and mapped through a dedicated MLP (Khanna et al., 2024):

$$\mathbf{h}_{geo} = \text{MLP}_{geo}\left([\mathcal{F}_{emb}(\theta) \parallel \mathcal{F}_{emb}(\phi) \parallel \mathcal{F}_{emb}(d) \parallel \mathcal{F}_{emb}(m) \parallel \mathcal{F}_{emb}(y)]\right). \tag{30}$$

Crucially, the final linear projection of MLP_{geo} is initialized with Gaussian weights of near-zero variance ($\sigma = 10^{-3}$) and zero bias. This guarantees that at the start of training, $\mathbf{h}_{geo} \approx \mathbf{0}$, preventing stylistic or geographic priors from destabilizing early spatial layout learning.

Joint FiLM Modulation. The representations are concatenated to form a joint FiLM embedding:

$$\mathbf{e}_{FiLM} = [\mathbf{e}_t \parallel \mathbf{h}_{geo}]. \tag{31}$$

D.2 Continuous-Time Diffusion Schedule and Resolution Log-SNR Shift

We employ a continuous-time diffusion framework, uniformly sampling $t \sim \mathcal{U}(0, 1)$ mapped through a continuous cosine noise schedule parameterized in log-SNR space Kingma et al. (2021)

Analytical Derivation of the Resolution Log-SNR Shift ($\Delta = -2.77$). When a diffusion backbone pre-trained on square images of reference resolution $d \times d$ ($d = 256$) is fine-tuned to synthesize equirectangular panoramas of dimensions $H_p \times W_p = 256 \times 1024$, the effective spatial frequency spectrum shifts towards higher token redundancy. Following (Hoogeboom et al., 2023), preserving the expected signal-to-noise ratio per unit area under spatial pooling/downsampling requires shifting the log-SNR curve by:

$$\Delta = -2\log\left(\frac{\max(H_p, W_p)}{d}\right) = -2\log\left(\frac{\max(256, 1024)}{256}\right) = -2\log(4) = -4\log(2) \approx -2.7726. \tag{32}$$

Continuous-Time Shifted Schedule Formulation. Base boundary limits are defined across $\log \mathrm{SNR}_{\min} = -15.0$ and $\log \mathrm{SNR}_{\max} = 15.0$. Applying the resolution shift $\Delta = -2.77$ yields corrected boundary endpoints:

$$\log \mathrm{SNR}'_{\min} = \log \mathrm{SNR}_{\min} - \Delta = -12.23, \quad \log \mathrm{SNR}'_{\max} = \log \mathrm{SNR}_{\max} - \Delta = 17.77. \tag{33}$$

The continuous integration bounds for the cosine schedule are computed via:

$$a = \arctan\left(\exp\left(-\frac{\log \mathrm{SNR}'_{\min}}{2}\right)\right), \quad b = \arctan\left(\exp\left(-\frac{\log \mathrm{SNR}'_{\max}}{2}\right)\right). \tag{34}$$

For any continuous time step $t \in [0, 1]$, the shifted log-SNR is analytically evaluated as:

$$\log \mathrm{SNR}(t) = -2\log\left(\tan\left((a - b)t + b\right)\right) + \Delta. \tag{35}$$

The forward marginal coefficients α_t, σ_t satisfying $\alpha_t^2 + \sigma_t^2 = 1$ are:

$$\alpha_t = \left(1 + e^{-\log \mathrm{SNR}(t)}\right)^{-1/2}, \quad \sigma_t = \left(1 + e^{\log \mathrm{SNR}(t)}\right)^{-1/2}. \tag{36}$$

Therefore,

$$z_t = \alpha_t x + \sigma_t \epsilon \quad \epsilon \sim \mathcal{N}(0, I) \tag{37}$$

Sigmoid-Weighted ELBO Objective. To maximize the continuous variational lower bound without gradient variance explosions at the extreme boundaries $t \to 0$ and $t \to 1$, we employ the sigmoid-weighted loss (Kingma & Gao, 2023):

$$w(t) = -\frac{1}{2}\frac{d \log \mathrm{SNR}(t)}{dt} \cdot \sigma\left(\log \mathrm{SNR}(t) - n\right), \tag{38}$$

where $\sigma(\cdot)$ is the standard logistic sigmoid and $n = 0.0$ specifies the central log-SNR bias.

D.3 Spatial Horizon Weighting Schedule and Annealing Dynamics

Equirectangular street-view panoramas exhibit an inherent vertical structural asymmetry: the upper rows ($i \to 0$) depict largely featureless, uninformative sky regions, whereas the lower rows ($i \to H_p$) contain intricate ground geometry, roads, curb lines, and structural boundaries.

Vertical Weighting Profile. To bias model capacity toward ground-level geometry, we apply an anisotropic vertical weight $\omega(i)$ to row index $i \in \{0, \ldots, H_p - 1\}$:

$$\omega(i) = \omega_{sky} + (\omega_{ground} - \omega_{sky}) \cdot \sigma\left(\frac{i - \tau H_p}{\gamma H_p}\right), \tag{39}$$

with the following empirically validated parameters:

- $\omega_{sky} = 0.4$: Prevents the generator from expending capacity fitting amorphous sky gradients.
- $\omega_{ground} = 1.0$: Retains full gradient magnitude over road surfaces, buildings, and vegetation.
- $\tau = 0.4$: Horizon center boundary, corresponding to the empirical camera pitch horizon line in the dataset.
- $\gamma = 0.05$: Transition smoothness factor governing the transition width across the horizon.

Linear Warmup Annealing. Applying equation 39 from initialization causes gradient instability because the network has not yet learned basic spatial canvas coordinates. We introduce a linear annealing schedule over a $K_{warm} = 10,000$-step warmup phase:

$$\Omega(k,i) = \left(1 - \min\left(1, \frac{k}{K_{warm}}\right)\right) \cdot 1.0 + \min\left(1, \frac{k}{K_{warm}}\right) \cdot \omega(i), \tag{40}$$

where k denotes current optimization step. The spatially weighted generator loss is given by:

$$\mathcal{L}_G = \mathbb{E}_{t,\mathbf{x},\boldsymbol{\epsilon}}\left[w(t)\left\|\sqrt{\boldsymbol{\Omega}} \odot (\mathbf{x} - \hat{\mathbf{x}})\right\|_2^2\right]. \tag{41}$$

D.4 Dynamic Critic Gating, Timestep Coupling, and Boundary Parameters

During training, the Perspective-Inversion Critic (F_ϕ) provides non-adversarial, structural gradient feedback by projecting the generator's denoised estimate back into nadir aerial space. However, two critical edge-case instabilities arise if this loss is applied naively across the full diffusion trajectory:

High-Noise Gradient Degradation. At extreme forward diffusion noise levels ($t \to 1$ or low $\log \mathrm{SNR}$), the generator's instantaneous clean prediction $\hat{\mathbf{x}}_0 = (\mathbf{z}_t - \sigma_t \hat{\boldsymbol{\epsilon}})/\alpha_t$ is dominated by unstructured Gaussian noise. Passing this high-variance pseudo-sample through the non-linear critic F_ϕ produces erratic gradients that degrade early global canvas formation.

Dynamic Gating Mechanism. To ensure that structural supervision is injected exclusively when the generator's spatial layout has coalesced, we enforce an indicator gating threshold τ:

$$\lambda_{critic}(t) = \begin{cases} \lambda_{\text{target}}, & \text{if } \bar{t}_{\text{gen}} \leq \tau \\ 0, & \text{otherwise,} \end{cases} \tag{42}$$

where $\bar{t}_{\text{gen}} \in [0,1]$ represents the normalized continuous diffusion timestep of the generator batch, $\tau = 0.60$ is the empirically validated gating ceiling, and $\lambda_{\text{target}} = 0.10$ denotes the active critic loss weight. When $\bar{t}_{\text{gen}} > \tau$, the critic pathway is short-circuited ($\lambda_{F_\phi} = 0$), preserving computational throughput and shielding the backbone from high-noise gradient shock.

Spatial Loss Hyperparameters and Edge Masking. Once projected into the aerial footprint $\hat{\mathbf{c}}^{\text{hr}} = F_\phi(\hat{\mathbf{x}}_0, C_{\text{lr}}, t_{\text{critic}})$, the composite spatial objective is evaluated against ground truth high-resolution aerial imagery $\mathbf{c}^{\text{hr}}$:

$$\mathcal{L}_{F_\phi} = w_{\ell_1}\left\|\left(\hat{\mathbf{c}}^{\text{hr}} - \mathbf{c}^{\text{hr}}\right) \odot \boldsymbol{\Lambda}\right\|_1 + w_{\text{ssim}}\left(1 - \text{MS-SSIM}\left(\hat{\mathbf{c}}^{\text{hr}}, \mathbf{c}^{\text{hr}}\right)\right), \tag{43}$$

with loss weights set to $w_{\ell_1} = 1.0$ and $w_{\text{ssim}} = 0.5$. The diagonal edge penalty matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{256\times256}$ assigns a multiplier of $\lambda_{\text{edge}} = 4.0$ to pixels within 8 px of the aerial tile perimeter, penalizing boundary seam artifacts caused by polar-to-planar projection transforms.

D.5 Variance-Adjusted DDIM Sampling Algorithm

At inference, the continuous interval $t \in [1,0]$ is discretized into $N = 256$ uniform steps $\{t_k\}_{k=1}^N$. Standard deterministic DDIM yields over-smoothed building textures, whereas fully stochastic DDPM adds high-frequency noise that degrades linear road boundaries. We deploy a variance-adjusted DDIM sampler with dynamic clipping and variance-corrected stochasticity (Algorithm 1).

Algorithm 1 Variance-Adjusted DDIM Sampling with Dynamic Clipping

Require: Conditioning inputs $C_{lr} = \{\mathbf{c}_{ms}, \mathbf{c}_{ae}, \mathbf{c}_{geo}\}$, trained generator G_Θ, total steps $N = 256$, stochasticity multiplier $\eta = 0.15$, clipping ceiling $\tau_{\max} = 0.3$.

Discretize continuous time: $1 = t_N > t_{N-1} > \cdots > t_1 > t_0 = 0$.

Sample initial Gaussian noise: $\mathbf{z}_{t_N} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

for $k = N$ **downto** 1 **do**

  Compute schedule parameters $\alpha_{t_k}, \sigma_{t_k}$ and next-step parameters $\alpha_{t_{k-1}}, \sigma_{t_{k-1}}$ via Section D.2.

  Evaluate model prediction: $\hat{\mathbf{x}}_0 = G_\Theta(\mathbf{z}_{t_k}, C_{lr}, t_k)$.

  Compute normalized residual noise: $\hat{\boldsymbol{\epsilon}}_{pred} = \frac{\mathbf{z}_{t_k} - \alpha_{t_k}\hat{\mathbf{x}}_0}{\sigma_{t_k}}$.

  Estimate data-space empirical error variance (Heek et al., 2024):

$$v_x = \max\left(0,\ 1 - \frac{1}{D}\sum_{d=1}^{D}(\hat{\boldsymbol{\epsilon}}_{pred}^{(d)})^2\right) \cdot e^{-\log \mathrm{SNR}(t_k)}.$$

  Apply strict threshold clipping: $\bar{v}_x = \min(v_x,\ \tau_{\max})$.

  Sample exploration noise $\boldsymbol{\epsilon}_k \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and form variance-corrected clean estimate:

$$\hat{\mathbf{x}}_0' = \hat{\mathbf{x}}_0 + \sqrt{\eta \cdot \bar{v}_x} \cdot \boldsymbol{\epsilon}_k.$$

  Recalculate effective residual noise: $\hat{\boldsymbol{\epsilon}}' = \frac{\mathbf{z}_{t_k} - \alpha_{t_k}\hat{\mathbf{x}}_0'}{\sigma_{t_k}}$.

  Compute deterministic DDIM transition:

$$\mathbf{z}_{t_{k-1}} = \alpha_{t_{k-1}}\hat{\mathbf{x}}_0' + \sigma_{t_{k-1}}\hat{\boldsymbol{\epsilon}}'.$$

end for

return Final synthesized panorama $\hat{\mathbf{x}} = \mathbf{z}_0$.

D.6 Hardware Infrastructure, Compute Profiling, and Training Budget

All experiments are conducted across NVIDIA A100s. Detailed infrastructure specifications and profiling metrics are documented in Table 4.

Table 4: **Hardware infrastructure, computational footprint, and training configurations.**

Hyperparameter / System Specification	**Generator** (G_Θ)	**Critic Pre-training** (F_ϕ)
Hardware Accelerator Pod	16 × A100	16 × A100
High-Bandwidth Memory (HBM per chip)	80 GB	80 GB
Total Accelerator Memory	1.28 TB	1.28 TB
Numerical Precision	`bfloat16`	`bfloat16`
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999$)
Weight Decay	10^{-4}	10^{-4}
Peak Learning Rate	10^{-4}	10^{-4}
Learning Rate Schedule	Cosine decay to 10^{-6}	Cosine decay to 10^{-6}
Warmup Steps	2,000 steps (linear)	2,000 steps (linear)
Global Batch Size	512	512
Total Optimization Steps	300,000 steps	300,000 steps
Spatial Resolution	$256 \times 1024 \times 3$	$256 \times 256 \times 3$

Cyclic Critic Timestep Coupling Parameters. During joint generator training, the critic's operating noise level is tethered via Equation 3 with uniform jitter $j = 0.05$, noise bounds $t_{\min} = 0.02$, $t_{\max} = 0.98$, and gating threshold $\tau = 0.60$. The edge penalty multiplier is fixed to $\lambda_{edge} = 4.0$, and spatial loss scalars are set to $w_{\ell_1} = 1.0$ and $w_{ssim} = 0.5$.

E EVALUATION METRICS FORMULATION & PROTOCOL

Because overhead-to-ground synthesis is fundamentally ill-posed, high-frequency ground geometry and vertical structural facades are unobservable from nadir satellite crops, standard deterministic pixel metrics alone penalize plausible multi-modal completions of the unobserved null space. We establish a four-tier evaluation protocol spanning distributional realism, photometric and physical surface micro-textures, 3D geometric layout, and semantic structural concordance.

E.1 STANDARD VISUAL, TEXTURAL, AND GEOMETRIC METRICS

- **Distributional & Perceptual Realism:** We compute Fréchet Inception Distance (FID) (Heusel et al., 2017) using the penultimate layer activations of a pre-trained Inception-v3 to quantify global visual manifold coverage. To assess local pairwise perceptual fidelity robust to spatial phase stochasticity, we compute Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) using a pre-trained VGG backbone (for the 456k dataset) and AlexNet backbone for CVUSA (to make it comparable with results reported in prior methods (Ze et al., 2025)).
- **Photometric Correspondence:** We report Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) (Wang et al., 2004) against target ground-truth panoramas to benchmark baseline spatial adherence and luminance consistency.
- **Surface Micro-Texture Fidelity:** To evaluate whether synthesized surfaces reflect physical material properties rather than mode-averaged blur, we compute Haralick Gray-Level Co-occurrence Matrix (GLCM) (Haralick et al., 1973) statistics. Co-occurrence frequencies are calculated across offsets $d \in \{1, 3\}$ pixels along four canonical orientations $\{0°, 45°, 90°, 135°\}$. We report the Mean Absolute Error (MAE) between ground truth and synthesized panoramas across three physical dimensions:
  1. *Contrast MAE*: Quantifies spatial intensity variations and micro-roughness (e.g., differentiating gravel verges from smooth asphalt).
  2. *Entropy MAE*: Measures spatial randomness and disorder (e.g., unmanaged forest canopies versus monolithic concrete pads).
  3. *Correlation MAE*: Assesses directional linear dependencies, capturing structural continuity in tractor furrows, crop rows, and lane markings.
- **3D Geometric Layout Alignment (Depth Proxy):** In the absence of planetary-scale ground-truth LiDAR, we evaluate 3D spatial coherence using the frozen zero-shot foundation model DepthAnythingV2 (Yang et al., 2024) as a pseudo-ground-truth oracle. Following the same evaluation protocols as DepthAnythingV2, We report Absolute Relative Error (AbsRel) and the threshold accuracy metric ($\delta_1 < 1.25$).

E.2 SEMANTIC ALIGNMENT VIA SEGMENTATION

We measure the semantic alignment between generated and ground truth street view images, to evaluate if the generation is semantically meaningful. Direct evaluation of semantics is difficult due to the lack of ground truth labels for objects like roads, trees etc. To this end, we used a pretrained Mask2Former (Cheng et al., 2021) segmentation model, trained on ADE20K (with 150 classes), to run zero shot semantic segmentation on pairs of ground truth and generated street view panoramas. To make the pretrained model compatible with the classes in our data, the 150 classes were collapsed into four foreground classes (sky, road, crops, foliage) and a background class. The segmentation outputs of the real and generated street view images are shown in Fig. 5. Table 5 show the results on our dataset. mIoU is first calculated for each class and then averaged to get the per-image mIoU. This was then averaged over a subset of 500 validation images, to obtain the reported mIoU and its 95% CI. We also report per-class mIoU, calculated across all pixels of the class.

Baseline Metric Sources. The baselines for CVUSA dataset in Table 1, FID, PSNR, and SSIM values are taken directly from the official evaluations in Ze et al. (2025) and Li et al. (2024).

Table 5: mIoU (best in **bold**) between semantic segmentation outputs from ground truth and Bifröst generated street view images, using a pretrained Mask2Former model in a zero shot setting.

Method	mIoU	mIoU 95% CI	Sky	Road	Crops	Foliage
Bifröst (S+A)	**0.480**	0.010	**0.814**	**0.535**	**0.538**	0.288
Bifröst (A)	0.458	0.010	0.802	0.486	0.525	0.274
Bifröst (S)	0.445	0.010	0.793	0.491	0.487	0.252
ControlNet (Zhang et al., 2023)	0.453	0.013	0.808	0.502	0.503	**0.309**
CrossViewDiff (Li et al., 2024)	0.360	0.012	0.702	0.340	0.427	0.150

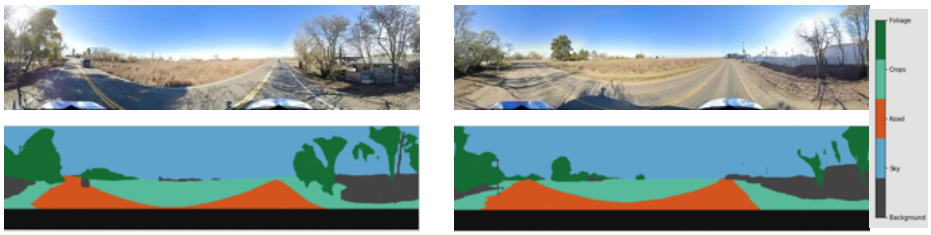

Figure 5: Figure shows the ground truth street view panorama (top left), the corresponding generated street view from Bifröst (top right), Mask2Former generated segmentation masks for ground truth image (bottom left), generated image (bottom right).

F Dataset Details

F.1 Global Multi-Sensor Dataset

We utilize Google Street View (GSV) as our source for ground-level panoramas. To ensure our dataset strictly focuses on relevant agricultural and rural environments, we filter the globe using S2 geometry (Larkin-York et al., 2015) at Level 13, isolating only cells that contain greater than 80% crop content as classified by the Dynamic World V1 dataset (Brown et al., 2022). Within these agricultural cells, we intersect the location and timestamp metadata across four distinct data sources: high-resolution (HR) satellite imagery, low-resolution (LR) Sentinel-2 multispectral imagery, AlphaEarth Brown et al. (2025) embeddings and the GSV panoramas.

To guarantee accurate physical and seasonal correspondence between the ground and overhead views, we enforce a strict temporal constraint: the three modalities, HR, LR, GSV must have been captured within a maximum 7-day window of one another, while the annualized AlphaEarth embeddings are used for the same year.

Spatial Constraints and Modality Specifications For each street-view capture location, we extract the corresponding overhead views centered precisely on those coordinates:

- *HR Satellite Imagery:* Extracted as 256×256 pixel tiles at a spatial resolution of 0.3 m/px.
- *LR Multispectral Imagery:* Extracted as 8×8 pixel patches from Sentinel-2, representing the exact corresponding geographic footprint.
- *AlphaEarth Embeddings*: The AlphaEarth embeddings are also generated at a resolution of 10m x 10m/px. Therefore, we fetch the embeddings for 8x8 pixel patch aroung the GSV capture location.
- *Ground-Level Panoramas:* To ensure a unified geometric coordinate space between the satellite conditioning and the ground view, we utilize the heading direction provided in the GSV metadata to apply a cyclic horizontal translation to the equirectangular panorama, strictly aligning the center of the image to True North.

To guarantee spatial independence and prevent identical overhead features from appearing in multiple examples, we enforce a strict minimum distance threshold between any two street-view capture locations. This separation is defined by the physical diagonal length of the HR satellite tile ($256 \times$

$\sqrt{2} \times 0.3)$ meters, equating to approximately 108.6 meters. This geometric constraint mathematically guarantees zero pixel overlap between the HR conditioning patches of any two dataset entries.

Geographic Scope and Strict Splitting Our collection process pulls data from across multiple continents to capture maximum global diversity in agricultural topologies, atmospheric conditions, and regional foliage. To construct our training, validation, and test sets, we partition the dataset geographically using larger Level 8 S2 cells. Entire Level 8 cells are distinctly assigned to a single split. This macro-level spatial segregation explicitly prevents geographic data leakage, ensuring the model's evaluation metrics reflect genuine zero-shot spatial generalization rather than memorized regional features.

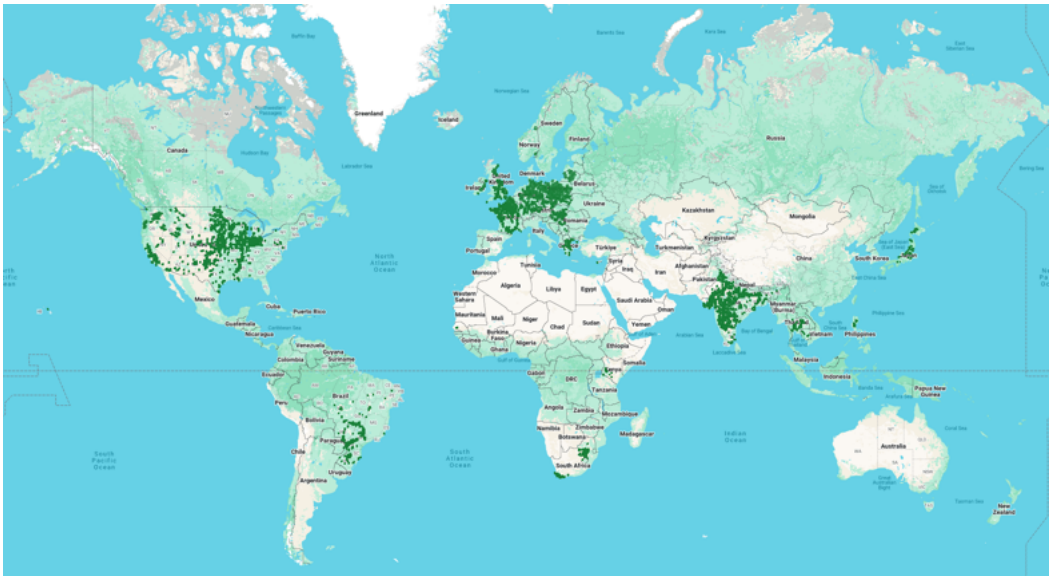

Figure 6: Geographic distribution of the dataset. Highlighted regions indicate the sampled Level 8 S2 cells used to source the co-registered satellite and street-view triplets.

F.2 CVUSA Dataset

To evaluate cross-view synthesis against established baselines on standard RGB benchmarks, we evaluate on the CVUSA dataset (Workman et al., 2015), adhering strictly to the canonical split protocol established by (Zhai et al., 2017).

- **Canonical Split and Modalities:** The standard CVUSA benchmark consists of about 1.5 million geotagged pairs of ground and aerial images from across the United States. The split used in (Zhai et al., 2017) is a subset with $35{,}532$ image pairs for training and $8{,}884$ held-out pairs for evaluation. Ground-level observations consist of Google Street View panoramas warped using camera extrinsics to achieve orientation alignment with overhead imagery and cropped vertically to suppress non-informative sky and ground pixels. Corresponding overhead views are sourced from Microsoft Bing Maps aerial imagery at zoom level 19 (≈ 0.3 m/px).
- **Multispectral and Low-Resolution Embeddings:** Because CVUSA natively provides only RGB aerial photography, we retrospectively curate the corresponding low-resolution conditioning modalities for all the geographic coordinates using Sentinel-2 multispectral images and annualized AlphaEarth embeddings.
- **Temporal Grounding and Timestamp Alignment:** Due to unavailability of per-image timestamp acquision in public space, we ground our sourcing strategy in the empirical temporal analysis of (Deuser et al., 2025). Specifically, CVUSA was crawled in 2015, establishing an empirical inflection point where approximately 56.81% of Street View panoramas post-date 2015 and 43.19% pre-date 2015.



Figure 7: Qualitative analysis of Bifröst's cross-view synthesis to check for memorization, on CVUSA dataset. From left to right, the columns display: the conditioning Sentinel-2 RGB patch, the ground truth panorama, and the generated panorama utilizing low-resolution inputs (S+A). (a) shows that the model does not generate the bridge but instead creates a building. (b) and (c) show that the model does in fact generate the fields but at a seemingly different time period, based on the crop stage. This shows that the model is using the conditioning information instead of regenerating what it might have memorized.

Sentinel-2 achieved operational status in late 2015 and AlphaEarth produces annual aggregations from 2017, we retrieve Sentinel-2 tiles from mid 2015 to mid 2016 and pair them with the annualized AlphaEarth foundation representations for 2017. This aligns our low-resolution conditioning with the median capture epoch of the benchmark as closely as possible to ensure comparable evaluation against RGB-only generative models.

G Qualitative Results

In this section, we examine the generated results qualitatively and identify the specific regions where the model generates structures accurately (marked in green boxes in fig. 8) and the regions where the model needs to improve further (marked in red boxes in fig. 8). As can be seen in the figure, the model accurately recovers complex macro-geometries such as major road intersections Fig.8(a), secondary unpaved paths joining main roads Fig.8(b), and distant field divisions Fig.8(c, d). Conversely, relying solely on low-resolution data introduces fundamental ambiguities, leading to hallucinations of unobservable vertical features like dense tree lines Fig.8(a, d), loss of high-frequency foreground textures on immediate road shoulders Fig.8(b), and seasonal crop-state mismatches due to annualized semantic priors Fig.8(c). Fig. 9 and 10 show a similar analysis for the baselines trained with high and low resolution data, and on the CVUSA dataset respectively.

Checking for Memorization We additionally try to verify qualitatively if our model is memorizing the data. We examine a small subset of validation set images from CVUSA, as shown in Fig. 7. The model can generate structures like roads but does not generate structures like bridges or underpasses (Fig. 7 (a)). Fig. 7(b) and (c) show that the model generates fields but at a seemingly different time period, based on the crop stage. In the CVUSA data, due to lack of temporal information, we had to rely on low resolution images which may be from a different time, resulting in different crop seasons (Appendix F.2). This difference in the generated indicates that the model does generate images based on the condition.

H Extended Discussion, Limitations, and Future Work

Synthesizing immersive, photorealistic 360° street-level panoramas from 10 m/px satellite observations represents a fundamentally ill-posed inverse problem. Because a single Sentinel-2 ground sampling unit encompasses an entire 100 m^2 footprint on the Earth's surface, the mapping from nadir spectral averages to horizontal lines of sight inherently contains an under-constrained null space. While *Bifröst* demonstrates that internalizing high-frequency geometry via cyclic inverted supervision substantially recovers road corridors, tree lines, and regional topology, the physical formulation

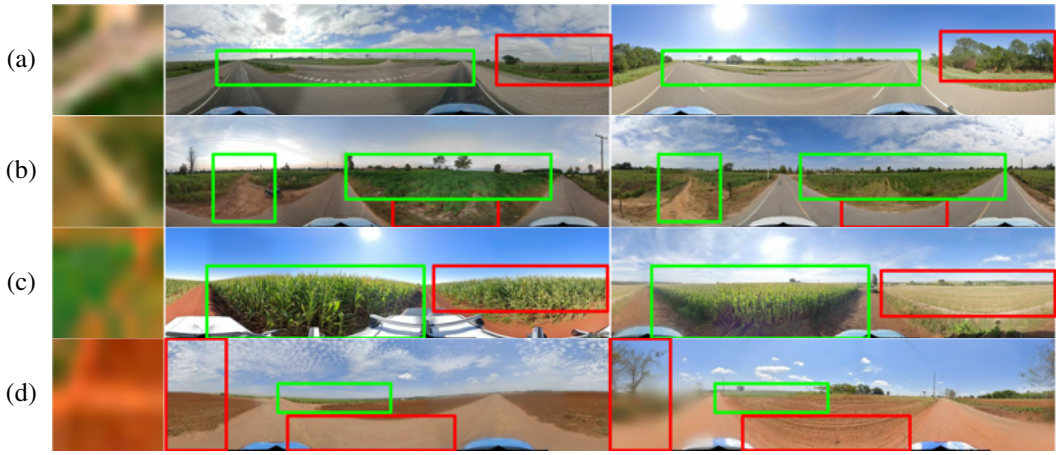

Figure 8: Qualitative analysis of Bifröst's cross-view synthesis highlighting macro-structural successes (green boxes) and unobservable or temporal failure modes (red boxes). From left to right, the columns display: the conditioning Sentinel-2 RGB patch, the ground truth panorama, and the generated panorama utilizing low-resolution inputs (S+A). The model accurately recovers complex macro-geometries such as major road intersections (a), secondary unpaved paths joining main roads (b), and distant field divisions (c, d). Conversely, relying solely on low-resolution data introduces fundamental ambiguities, leading to hallucinations of unobservable vertical features like dense tree lines (a, d), loss of high-frequency foreground textures on immediate road shoulders (b), and seasonal crop-state mismatches due to annualized semantic priors (c).

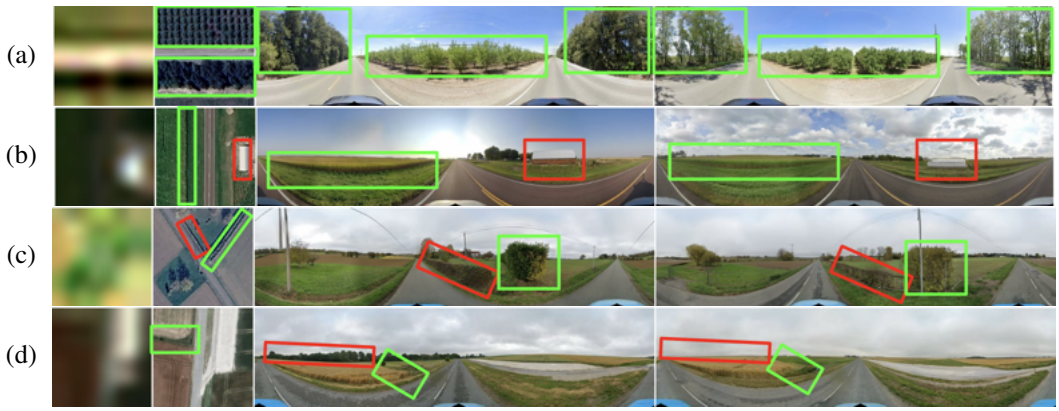

Figure 9: Qualitative analysis of Baselines trained and inferred with H + S, where the green boxes indicating accurate generation and red boxes indicating areas for improvement (left to right, the columns show, RGB bands of sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image). For instance, our model can accurately generate row-wise cultivation of shrubs and trees (a), granular variations in greenery (b), field boundaries (c, d) and shrub-walls (c). The model does hallucinate the type of trees (a) and fences (c), cannot accurately generate houses/barns since the side view is unavailable (b) as well as structures not seen in the overhead imagery (trees, in (d)).



Figure 10: Qualitative analysis of Bifröst's cross-view synthesis highlighting macro-structural success (green box) and failure modes (red box) on the CVUSA dataset, where the green boxes indicating accurate generation and red boxes indicating areas for improvement (left to right, the columns show, RGB bands of sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image). For instance, our model can identify regions with dense foliage and junctions (row 2, 3 and 6). The model cannot accurately generate markings on the road, hallucinates the type of trees and cannot generate vehicles (row 1, 4, 5 and 6). This is likely due to the coarse resolution and lack of instantaneous satellite images for this dataset.

exhibits intrinsic boundaries. Below, we provide an exhaustive discussion of these failure modes and structural trade-offs.

H.1 Nadir Blindspots and Vertical Facade Hallucination

A fundamental geometric constraint of overhead satellite remote sensing is the absence of vertical surface visibility. At strictly nadir or near-nadir viewing angles (off-nadir angle $< 10^\circ$ for Sentinel-2), an overhead sensor records building roofs, upper foliage canopies, and horizontal terrain, but receives zero photon radiance reflected from vertical architectural planes (e.g., building walls, door frames, window embrasures, or fence posts).

H.2 Temporal Phase Lags in Annualized Foundation Features

To provide rich regional semantics, *Bifröst* conditions on 64-dimensional AlphaEarth foundation embeddings (Brown et al., 2025). However, a structural characteristic of AlphaEarth is its multi-temporal, annualized aggregation strategy: embeddings are computed by pooling satellite observations across an entire calendar year to produce cloud-free, noise-resilient representations.

Dynamic Biome Sensitivity. In dynamic agricultural environments, rapid land-use transitions occur over short intervals (e.g., harvesting of wheat fields within 48 hours, sudden spring tilling, or rapid seasonal crop rotations). While our dataset enforces a strict $\leq$ 7-day temporal window between the raw Sentinel-2 multi-spectral pass and the ground panorama, the annualized foundation feature vector c_{ae} reflects the time-averaged semantic state of that parcel.

When a panorama is captured immediately post-harvest (dry stubble/bare soil) but the annualized embedding is dominated by the dense vegetative growth phase:

1. The Dual-Stream FiLM mechanism receives contradictory cues: dense vegetation in c_{ae} versus bare soil in the instantaneous Sentinel-2 L2A bands (c_{ms}).
2. In rare cases, this modality tension results in minor vegetation phase artifacts, such as generating semi-green roadside ditches adjacent to bare, freshly plowed agricultural parcels.

Temporal Decoupling and Phenological Drift in CVUSA. A notable empirical failure mode manifests when evaluating on the legacy CVUSA benchmark (Workman et al., 2015). Because the original CVUSA release lacks per-instance acquisition timestamps, exact temporal co-registration is unattainable. As documented by (Deuser et al., 2025), the ground-level street panoramas span years prior to and following the 2015 crawl epoch, whereas our retrofitted Sentinel-2 imagery and AlphaEarth embeddings are anchored to late-2015/2016 and 2017 respectively. When an evaluated ground panorama was captured several years prior to or after this baseline window (e.g., a 2011 street capture paired with a 2015 satellite tile), severe multi-year environmental shifts can occur.

In these out-of-sync instances, *Bifröst* faithfully synthesizes the land-cover state dictated by the 2015 satellite conditioning (e.g., a newly cleared agricultural parcel, updated roadside fence lines, or deforested tree stands), which directly contradicts the historical ground truth panorama depicting dense mature orchards or outdated unpaved road surfaces. While this discrepancy penalizes reference-based pixel metrics (PSNR, SSIM, and LPIPS) against the historical ground truth, qualitative inspection confirms that the failure stems from true physical landscape evolution across the unrecorded temporal gap, rather than geometric breakdown in the coordinate-aware cross-attention mechanism.

H.3 Evaluation Dependency on Foundation Model Proxies

Because acquiring planetary-scale, pixel-aligned ground-truth LiDAR and dense semantic segmentation for unconstrained rural panoramas is physically intractable, our evaluation protocol relies on off-the-shelf foundation models as pseudo-ground-truth oracles. Specifically, we utilize DepthAnythingV2 (Yang et al., 2024) for 3D geometric layout evaluation and Mask2Former (Cheng et al., 2021) for semantic structural alignment (mIoU).

While these models provide robust relative benchmarks, they introduce intrinsic prior biases into the evaluation loop. DepthAnything V2, functioning as a monocular depth estimator, inherently favors

Table 6: Ablation of architectural components and input modalities against the High-Resolution (S + H) baseline. Removing explicit spatial priors, namely, coordinate augmentation (coords) and geo-temporal metadata (c_{geo}), severely degrades both structural alignment (AbsRel) and textural fidelity. Furthermore, substituting low-resolution multispectral data (S) with an explicit depth map improves mean-seeking pixel metrics (PSNR, SSIM) but significantly harms perceptual realism (FID). This demonstrates that raw multi-band spectral features provide richer high-frequency environmental context than strictly geometric depth maps.

Configuration	Standard Metrics				GLCM Texture (MAE)			Depth Structure	
	FID ↓	LPIPS ↓	PSNR ↑	SSIM ↑	Contrast ↓	Entropy ↓	Correl. ↓	AbsRel ↓	δ_1 ↑
Baseline Reference									
S + H (Full Architecture)	**3.9274**	0.4912	13.9865	0.4569	**0.1880**	**0.2634**	0.0096	**0.1007**	**0.9079**
Architectural Ablations									
S + H w/o $\mathbf{c}_{\text{geo}}$	<u>3.9861</u>	0.5076	13.8700	0.4460	0.2054	0.2916	0.0110	0.1148	0.9027
S + H w/o coords	4.0493	0.5059	13.8012	0.4409	0.2179	0.2784	0.0106	0.1126	0.9007
Input Representation									
H + Depth Map	4.4741	**0.4860**	<u>14.4370</u>	<u>0.4580</u>	**0.1880**	0.2650	**0.0089**	0.1076	0.9001

spatially smooth, statistically probable topologies, which may penalize physically accurate but highly irregular agricultural layouts or high-frequency vegetation boundaries. Similarly, Mask2Former's segmentation priors are heavily anchored by its pre-training domain (e.g., standard autonomous driving or urban datasets); this can induce semantic shift when processing equirectangular projections of unpaved rural terrain or ambiguous crop boundaries. Consequently, improvements in these proxy metrics demonstrate strong alignment with generalized structural and semantic priors, but they should be interpreted as robust relative proxies rather than absolute physical reconstruction accuracy.

H.4 Future Work

While *Bifröst* establishes a robust foundation for globally scalable cross-view synthesis, several avenues remain for extending this framework. First, our Variational Cyclic Internalization effectively utilizes unobserved high-resolution targets while bypassing the min-max instability of adversarial training. However, explicitly isolating the structural advantages of the diffusion-based Perspective-Inversion Critic (PIC) against alternative high-resolution supervisory signals, such as direct contrastive feature-matching or standard layout-aware discriminators, remains an important direction for future empirical analysis.

Second, to resolve the temporal friction introduced by annualized foundation embeddings, future architectures could integrate dynamic, time-series conditioning directly from temporally dense satellite constellations, allowing the network to explicitly model phenological shifts rather than relying on yearly aggregates. Finally, addressing the fundamental nadir blindspot regarding vertical facades could involve bridging *Bifröst's* global priors with sparse, crowdsourced terrestrial imagery or emerging continuous 3D representations, paving the way for globally available, multi-view consistent rendering.

I Additional Ablation Studies and Structural Analysis

To isolate the contributions of our architectural components, training objectives, and input representations, we conduct systematic ablation experiments across both geometric depth proxies and semantic layout consistency.

Architectural and Objective Ablations. Table 6 evaluates our core design choices against the high-resolution baseline ($S + H$):

- **Impact of the Perspective-Inversion Critic (PIC):** Disabling the cyclic objective ($\lambda_{\text{critic}} = 0$) causes the most severe structural degradation across the entire system. Without the frozen critic enforcing nadir geometric consistency via Variational Internalization, the model collapses to the low-resolution baseline plateau, with depth error surging from AbsRel $= 0.101$ to 0.138 and FID worsening to 4.40. This confirms that cyclic inverted supervision is the primary driver bridging the resolution gap.



Figure 11: Visualization of ControlNet (Zhang et al., 2023) street-view generation when trained on high resolution Google Map images(left to right, the columns show, high res Google Map image and the ground truth image followed by the generated street view image)

- **Coordinate-Aware Cross-Attention (CACB):** Removing explicit coordinate grids ($\mathbf{C}_{\text{polar}}$ and $\mathbf{C}_{\text{pano}}$) impairs the model's ability to map rotational overhead orientations to horizontal panoramic viewing rays. Depth error increases from 0.101 to 0.113, accompanied by a rise in GLCM texture contrast ($0.188 \rightarrow 0.218$), demonstrating that coordinate grounding provides a crucial structural inductive bias.
- **Geo-Temporal FiLM Conditioning ($\mathbf{c}_{\text{geo}}$):** Omitting scalar geo-temporal metadata yields a modest increase in FID ($3.93 \rightarrow 3.99$) but causes noticeable degradation in LPIPS ($0.491 \rightarrow 0.508$) and depth alignment (0.115). While the diffusion backbone can hallucinate plausible outdoor scenery without metadata, geo-temporal FiLM is required to match regional sun angles, seasonal vegetation states, and geographic terrain profiles.
- **Explicit Depth vs. Multi-Spectral Cues:** Prior cross-view methods often argue for explicit depth map conditioning. Substituting Sentinel-2 multispectral bands with an off-the-shelf depth map (H + Depth) achieves high PSNR (14.44 dB) and low LPIPS (0.486). However, because pixel-level depth maps favor smooth, mean-seeking predictions, the depth-conditioned model significantly degrades distributional realism (FID $= 4.47$ vs. 3.93 for $S + H$). Raw multi-band spectral data provides richer ecological and textural context than purely geometric depth representations.

J Visualizing the different generations

We have provided the detailed ablation analysis that compares the different sensor modality combinations in table 2. In order to understand the corresponding visual differences, we provide a series of qualitative examples for each sensor modality combination that we have analyzed in fig. 17, fig. 18, fig. 19, fig. 20, fig. 21, fig. 22, fig. 23.

K Societal Impact and Environmental Sustainability

The generation of immersive, ground-level views from ubiquitous satellite observations presents substantial societal opportunities alongside distinct ethical and technical risks.



Figure 12: Visualization of CS2S (Ze et al., 2025) street-view generation when trained on high resolution Google Map images(left to right, the columns show, high res Google Map image and the ground truth image followed by the generated street view image)

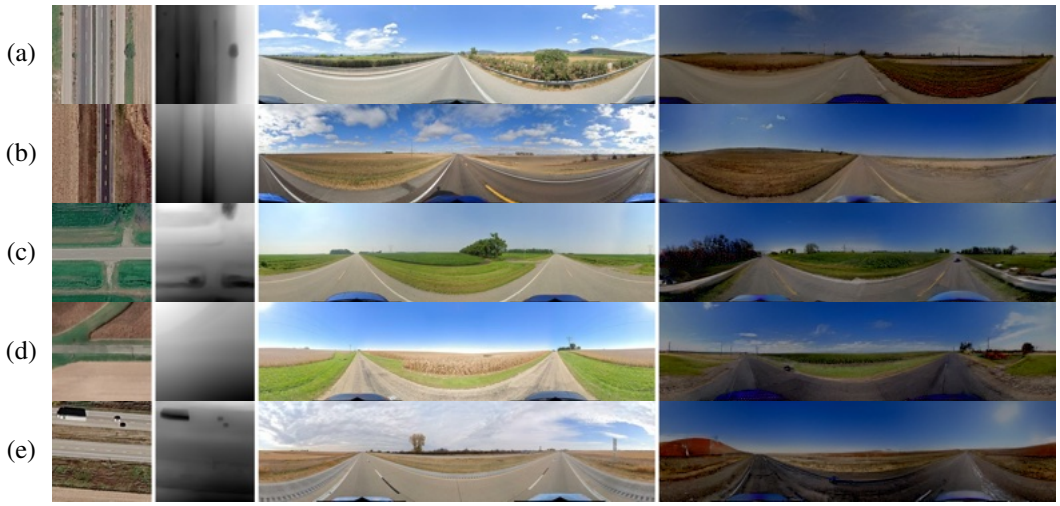

Figure 13: Visualization of CrossViewDiff (Li et al., 2024) street-view generation when trained on high resolution Google Map images(left to right, the columns show, high res Google Map image, the depth map generated by DepthAnythingV2 (Yang et al., 2024) and the ground truth image followed by the generated street view image)

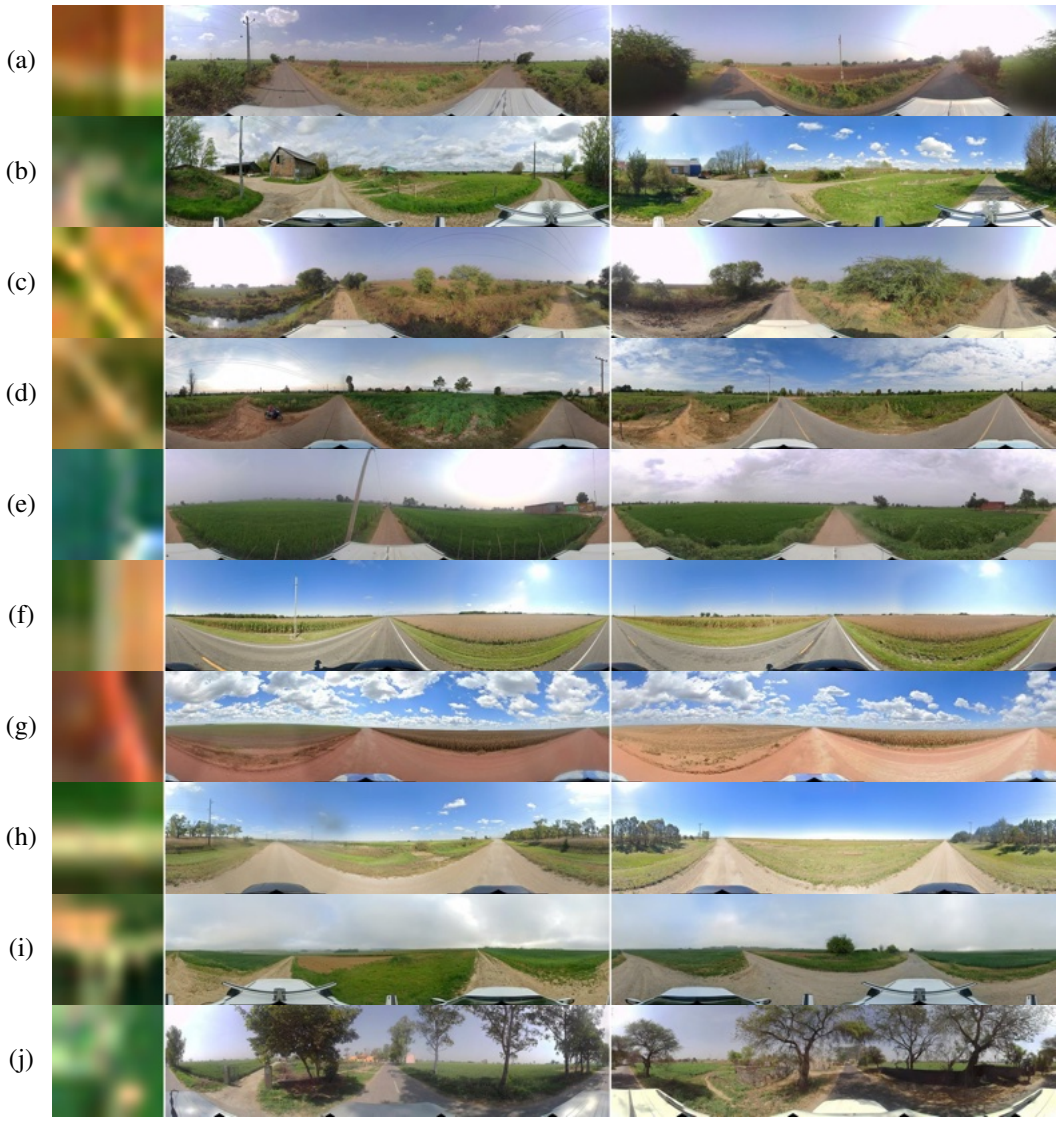

Figure 14: **S + A w/ cyclic:** Visualization of Bifröst street-view generation conditioned on low-resolution Sentinel-2 and AlphaEarth embeddings. From left to right, the columns display: the Sentinel-2 RGB patch, the ground truth street-view panorama, and the generated street-view panorama.



Figure 15: **A w/ cyclic:** Visualization of Bifröst street-view generation conditioned on low-resolution AlphaEarth embeddings Only. From left to right, the columns display: the Sentinel-2 RGB patch (Only for Visualization), the ground truth street-view panorama, and the generated street-view panorama.



Figure 16: **S w/ cyclic:** Visualization of Bifröst street-view generation conditioned on low-resolution Sentinel-2 Only. From left to right, the columns display: the Sentinel-2 RGB patch, the ground truth street-view panorama, and the generated street-view panorama.

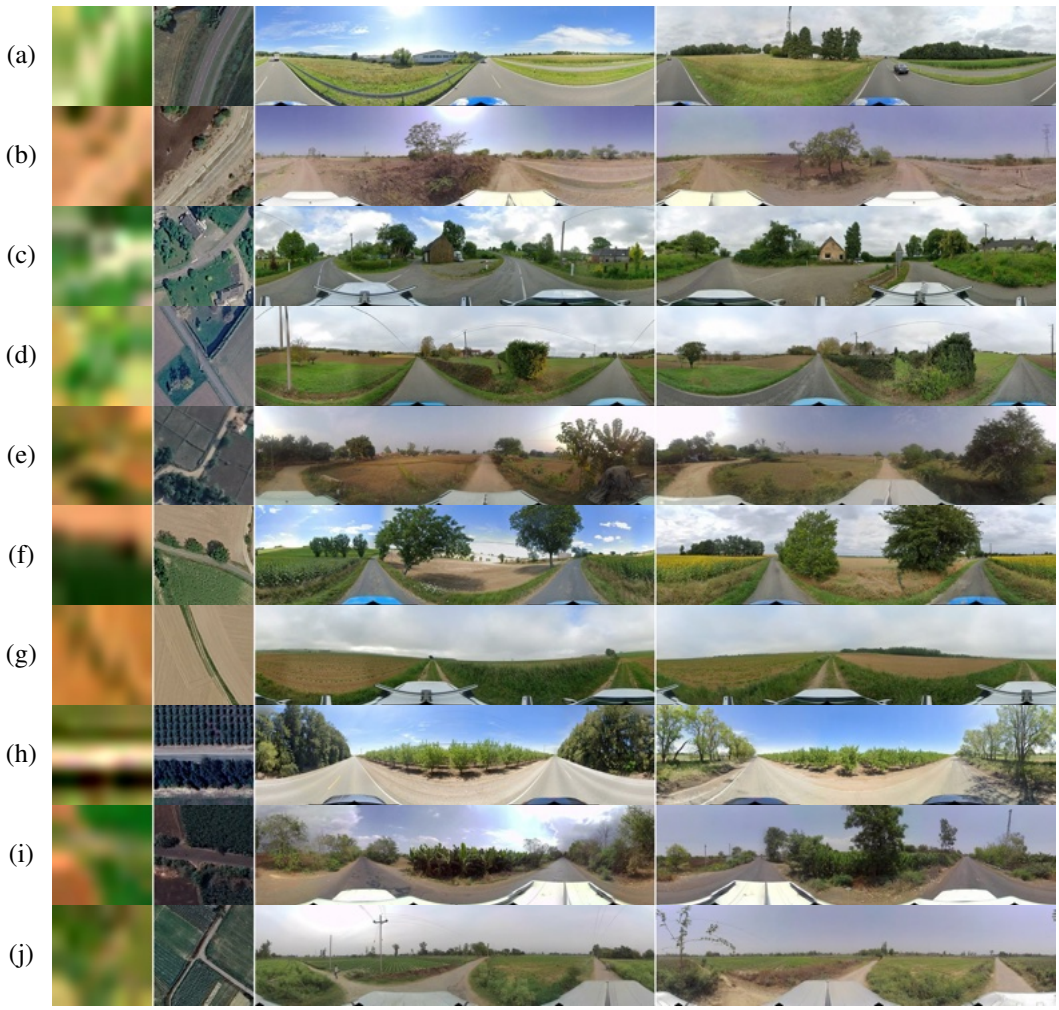

Figure 17: S+M+A: Visualization of Bifröst street-view generation when trained on low resolution Sentinel2, high resolution Google Map images and AlphaEarth embeddings. (left to right, the columns show, RGB bands of sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image)

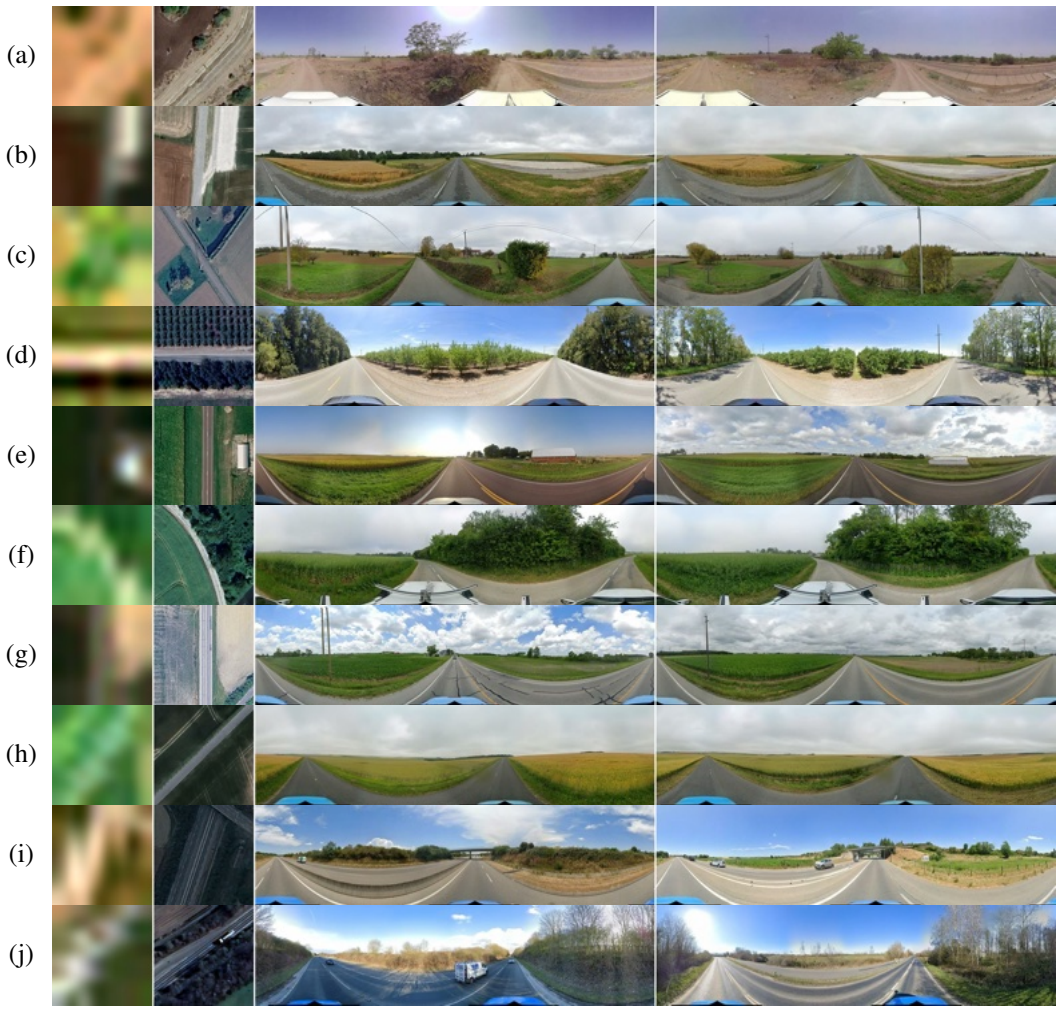

Figure 18: S+M: Visualization of Bifröst street-view generation when trained on low resolution Sentinel2 and high resolution Google Map images. (left to right, the columns show, RGB bands of sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image)



Figure 19: S: Visualization of Bifröst street-view generation when trained on low resolution Sentinel2. (left to right, the columns show, RGB bands of sentinel 2 and the ground truth image followed by the generated street view image)

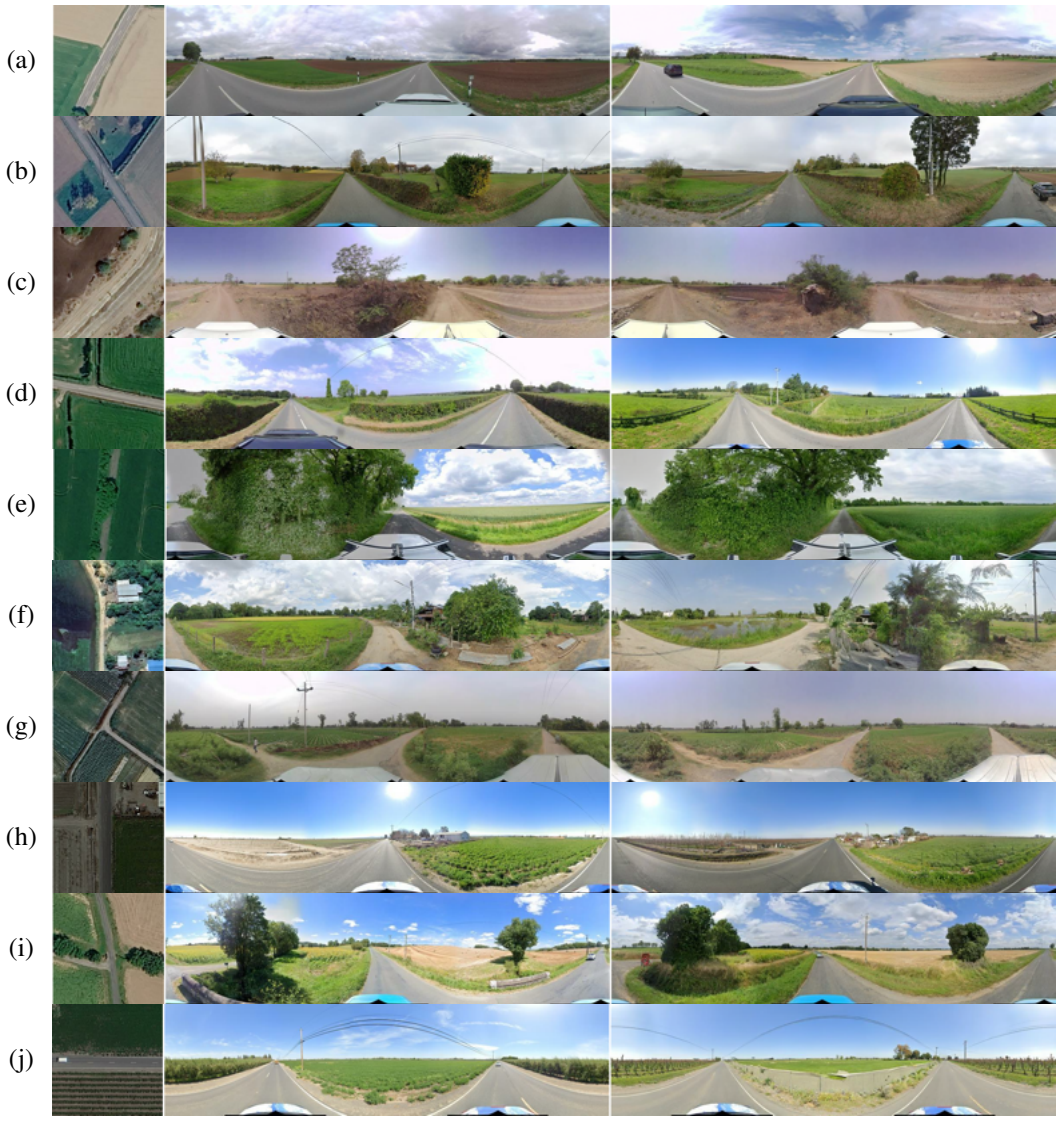

Figure 20: M: Visualization of Bifröst street-view generation when trained on high resolution Google Map images. (left to right, the columns show high res Google Map image and the ground truth image followed by the generated street view image)



Figure 21: A: Visualization of Bifröst street-view generation when trained on AlphaEarth embeddings. (left to right, the columns show, RGB bands of sentinel 2 and the ground truth image followed by the generated street view image)

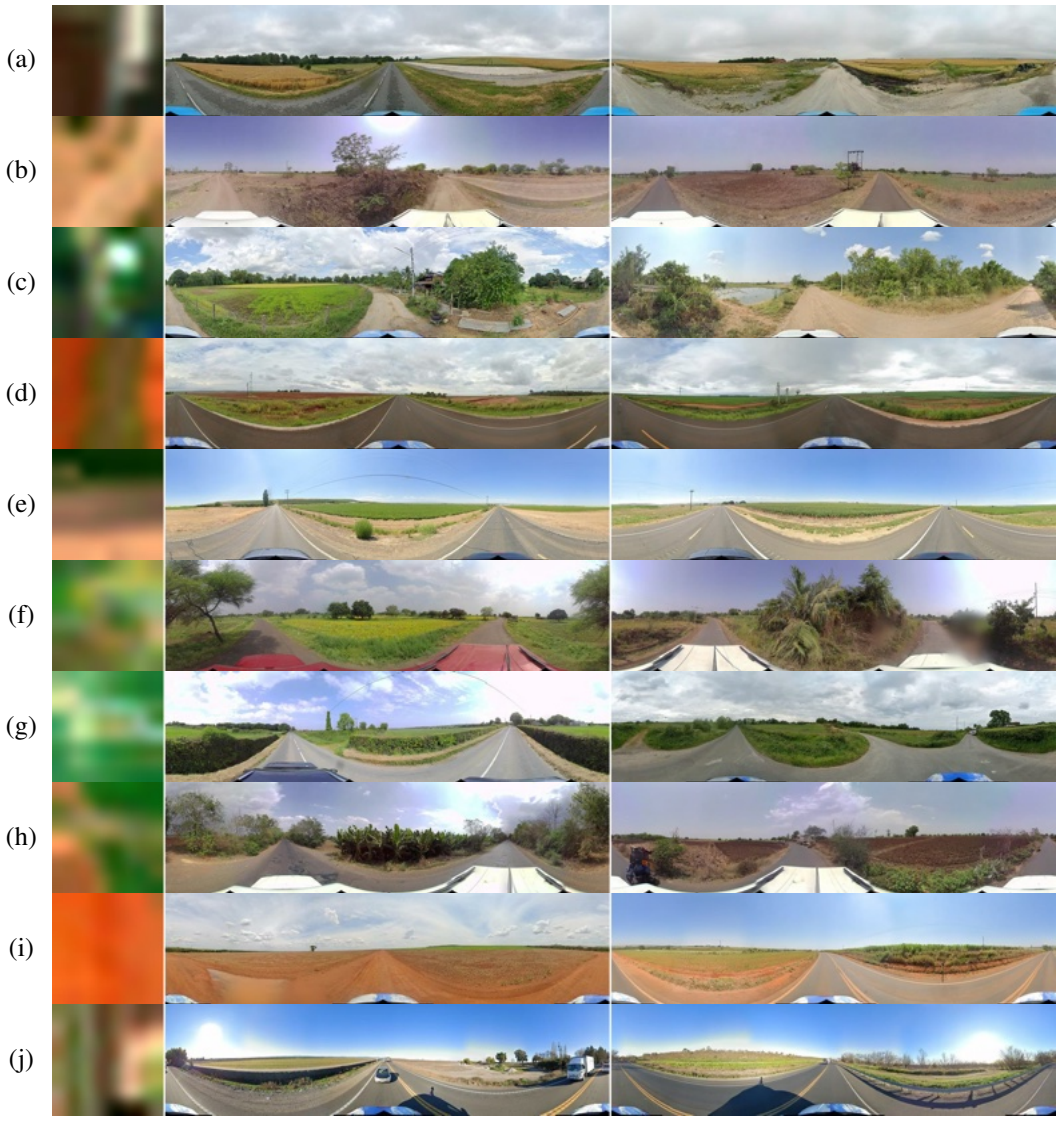

Figure 22: A+S: Visualization of Bifröst street-view generation when trained on low resolution Sentinel2 and AlphaEarth embeddings. (left to right, the columns show, RGB bands of sentinel 2 and the ground truth image followed by the generated street view image)

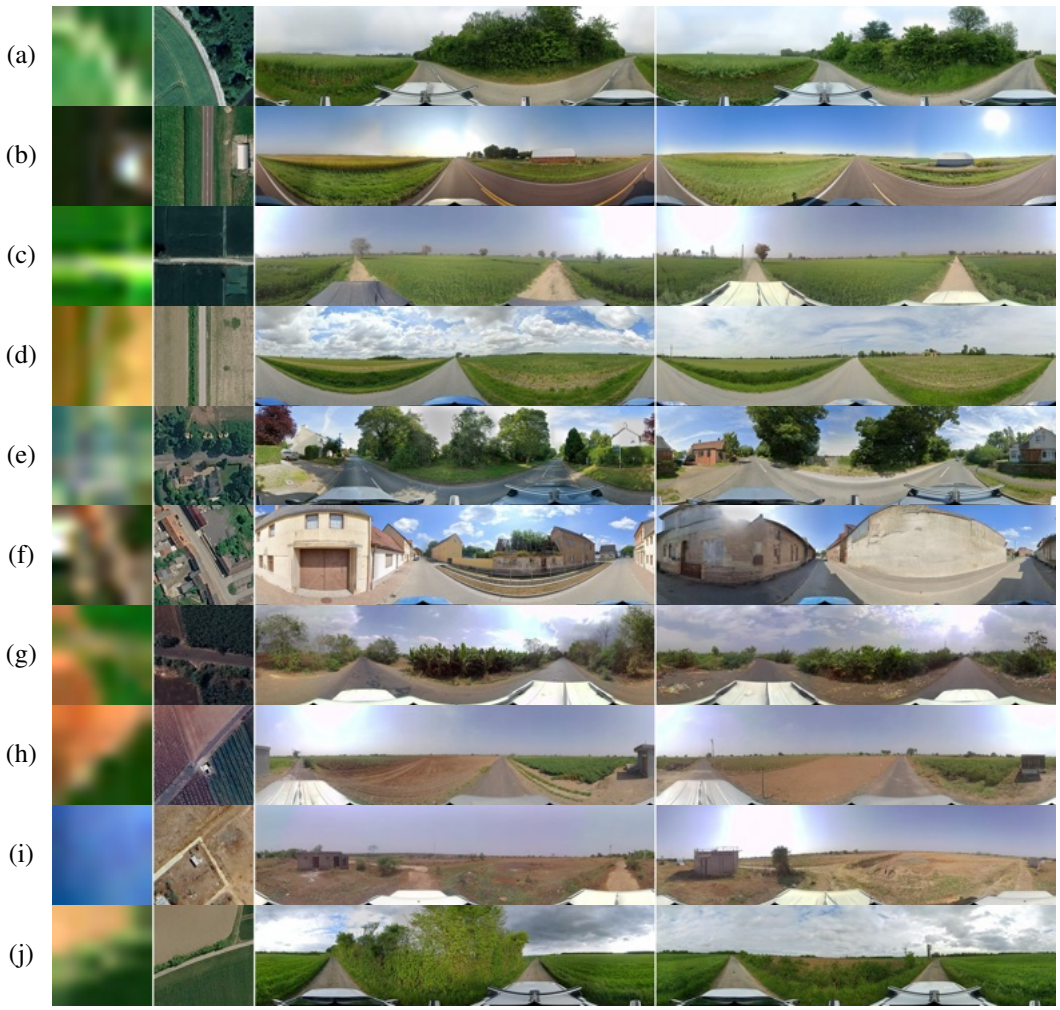

Figure 23: M+A: Visualization of Bifröst street-view generation when trained on high resolution Google Map images and AlphaEarth embeddings. (left to right, the columns show, RGB bands of sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image)



Figure 24: CVUSA S w/ cyclic: Visualization of Bifröst street-view generation when generator is trained on low resolution Sentinel2 images only. (left to right, the columns show, RGB bands of Sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image). High res image is only shown for reference, and is not used in the inference.



Figure 25: CVUSA A w/ cyclic: Visualization of Bifröst street-view generation when generator is trained on AlphaEarth embeddings only. (left to right, the columns show, RGB bands of Sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image). Sentinel2 and high res images are only shown for reference, and are not used in the inference.



Figure 26: CVUSA S+A w/ cyclic: Visualization of Bifröst street-view generation when generator is trained on low resolution Sentinel2 images and AlphaEarth embeddings. (left to right, the columns show, RGB bands of Sentinel 2, high res Google Map image and the ground truth image followed by the generated street view image). High res image is only shown for reference, and is not used in the inference.

K.1 Downstream Positive Applications

Conventional cross-view synthesis frameworks are tethered to proprietary, high-resolution aerial photography (≤ 0.3 m/px). Such data is overwhelmingly concentrated over high-density metropolitan zones in affluent nations, creating a structural geographic bias. By demonstrating that high-frequency ground geometry can be synthesized strictly from freely accessible, globally ubiquitous Sentinel-2 (10 m/px) and open foundation embeddings at inference time, democratizes immersive spatial awareness for historically underserved domains.

Precision Agriculture and Rural Stewardship. Over 80% of our benchmark encompasses agricultural and rural ecosystems. Current monitoring workflows rely on top-down vegetation indices (e.g., NDVI, NDRE), which inherently compress vertical structural dynamics. Synthesizing ground-level panoramas directly from multi-spectral passes enables farmers and agricultural researchers to inspect crop height variations, roadside drainage conditions, field boundary degradation, and localized canopy stress without deploying continuous, capital-intensive ground sensor networks or autonomous rovers.

Rapid Disaster Assessment and First-Response Logistics. In the wake of natural catastrophes (e.g., flash flooding, wildfires, severe wind events), terrestrial transit corridors are frequently disrupted, and commercial ground-view mapping vehicles (such as GSV fleets) cannot access affected sectors. While commercial aerial surveying requires days to charter and clear airspace, Sentinel-2 offers a predictable 5-day global revisit cycle. facilitates immediate, zero-shot ground-perspective assessment of road navigability, collapsed utility corridors, and standing water boundaries, empowering logistical planners with street-level orientation before dispatching emergency response teams.

Decentralized Environmental Monitoring. Deforestation, illegal rural waste dumping, and soil erosion typically initiate at rural margins concealed from nadir view by dense forest canopies. Converting multispectral spectral signatures into oblique, ground-facing perspectives provides conservation authorities with intuitive visual proxies to identify encroaching logging paths or unauthorized ground excavation in protected biomes.

K.2 Ethical Considerations, Dual-Use, and Safety Safeguards

Like all generative visual models conditioned on spatial and geographic primitives, poses societal risks regarding privacy, verification, and deceptive misuse.

Privacy and Identity Obfuscation. Our ground-truth training source utilizes public Google Street View data, which applies automated blurring algorithms over human faces and vehicle license plates. Furthermore, because our model synthesizes ground views from 10 m/px satellite pixels, where an entire human or passenger car occupies a fraction of a single pixel, our model does not and cannot reconstruct real-time personal identities, human presence, or vehicle registrations. It generates plausible regional geometric topology (e.g., road width, tree line, field crops) rather than verifiable surveillance footage.

Hallucination Hazards in Critical Decision-Making. A critical vulnerability of diffusion-based synthesis is plausible hallucination. In mission-critical workflows (e.g., evaluating bridge structural integrity or precise floodwater depth), generative output must **not** be treated as ground truth forensic evidence. As emphasized in Section 3.1, the model produces samples from the conditional posterior distribution $p(x \mid C_{\text{lr}})$; occluded architectural facades and sub-pixel road obstructions are stochastically inferred. Downstream decision-support tools must display uncertainty overlays and explicitly inform human operators that synthesized images are predictive geometric hypotheses.

Digital Forensics and Provenance. To prevent malicious dissemination (e.g., generating falsified imagery of damaged municipal infrastructure or fabricated rural disaster zones), all assets synthesized via should incorporate cryptographic C2PA provenance metadata and invisible spatial watermarking in the latent equirectangular frequency domain, ensuring verifiable attribution and preventing generative visual disinformation.