GroundControl: Satellite-Conditioned Multi-View Diffusion for Reconstruction-Free Novel View Synthesis

Akshay Krishnan[1,2*] Matias Turkulainen[1,3*] Jamie Wynn[1] Zawar Qureshi[1]
Mohamed Sayed[1] Guillermo Garcia-Hernando[1] Sara Vicente[1] James Hays[2]
Michael Firman[1] Daniyar Turmukhambetov[1]
[1]Niantic Spatial [2]Georgia Institute of Technology [3]Aalto University

Abstract

Multi-view or video diffusion models have recently emerged as state-of-the-art approaches for novel-view synthesis from sparse image collections. While they can produce high quality images, they tend to hallucinate unrealistic spatial layouts when synthesizing views outside the input views (scene extrapolation). We propose to address this challenge for outdoor scenes by using satellite images as an additional information source to condition the multi-view generative model. We introduce a new satellite-conditioned multi-view diffusion model that combines conditioning from a satellite image and street-level input images when generating novel-views. Our cross-attention based conditioning allows us to retain pretrained multi-view diffusion priors to generate seamless, realistic, novel views from a handful of source views, while also adhering to the scene layout as seen from satellite images. Our model does not use 3D scene reconstruction, and only relies on posed street and satellite views. Our experiments on outdoor datasets show that our model is better than both satellite-conditioned or street-view conditioned baselines on novel-view synthesis, especially on challenging cases where target views have very little overlap with source views. Project page: https://nianticspatial.github.io/groundcontrol

1. Introduction

Imagine looking at a photo of a street, *e.g.* Fig 1. Given just that photo, could you realistically picture what is on either side of the street? Current generative models for novel-view synthesis (*e.g.* video models [8, 26–28] or camera-controlled diffusion models [21, 81]) can extrapolate scenes from sparse input views, but they hallucinate outside the input views, generating buildings and roads that may not exist. But what if you also had a satellite image? The satellite view provides a good spatial layout prior, although it gives you a limited sense of appearance, *e.g.* the satellite view may not see the facades of buildings and the appearance of smaller objects. The street image and the satellite image are both useful, but incomplete. In this work, we fuse the best of both: we introduce a generative model that uses the satellite image as a strong spatial prior to guide ground-level multi-view novel-view synthesis, allowing us to generate new, photorealistic views that adhere to the satellite layout and to the content of street-level photos.

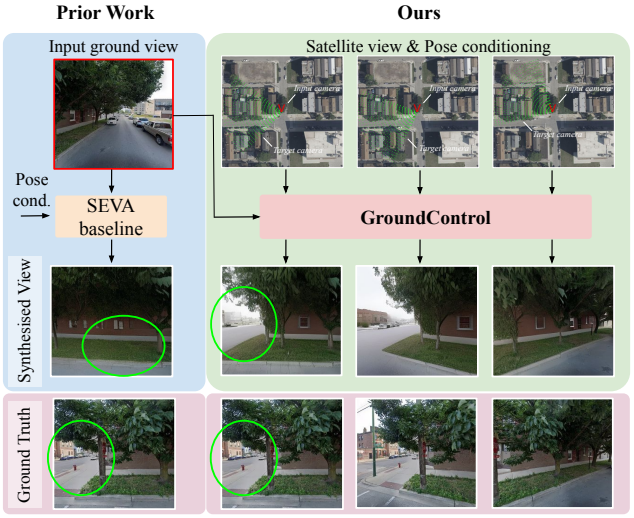


Figure 1. **GroundControl** builds on recent novel-view synthesis (NVS) diffusion models, and enables more accurate synthesis of novel views by additionally conditioning on a top-down satellite image. When compared to the NVS diffusion model SEVA [81] (left), our model has better fidelity to the scene's spatial layout.

While there is extensive work on generating ground-level images or panoramas from satellite images alone (Table 1), [37, 70, 71, 80], doing this without any ground-level information is a highly challenging task. Many such works rely on *oblique* aerial or satellite imagery, where the facade of buildings (and hence street-level textures and geometry) are captured [18, 36, 70, 71]. They struggle with the more com-

*Work done as interns at Niantic Spatial

mon orthographic satellite images which give very limited information about ground-level aesthetics that can only be inferred from street-level source views. Furthermore, in applications where sparse ground-level images are available, it is unclear how these methods can be leveraged to complete such partially observed ground-scenes. It is therefore appealing to combine satellite images for layout-guidance together with sparse street-level views to generate street-views with accurate textures and appearance.

Text-to-image diffusion models learn a rich image generative prior from large datasets which has been leveraged in many controlled-generation settings, by conditioning on various modalities [4, 46, 48, 79]. However, the combination of multi-view and satellite images poses unique challenges to these conditioning methods due to stark differences in the appearance and viewpoints between them. To this end, we introduce a novel satellite-conditioned multi-view diffusion model for outdoor scenes. By finetuning only our newly added lightweight ray-guided cross-attention layers, we retain the rich multi-view generative priors of the diffusion model. Our approach is reconstruction-free, as in does not assume access to satellite height maps or street-level 3D scene reconstructions or pixel-accurate 6DoF street view-satellite alignment. It only requires relative camera poses and coarse 3DoF (X, Y, heading) satellite alignment. We demonstrate our model's effectiveness across a range of different datasets, from domains including self-driving car data, street-level panoramas, and sparse collections of images from hand held cameras.

In summary, our key contributions are:

- We introduce a satellite-conditioned multi-view diffusion model *GroundControl* for novel view synthesis of street scenes. To the best of our knowledge, this is the first satellite-guided multi-view generation approach that does not use height-maps or street-level reconstruction.
- We adopt a ray-guided cross-attention method for conditioning state-of-the-art camera controlled diffusion models on 2D satellite image features. Our approach is efficient as it only trains the new cross-attention parameters we add, and on a relatively small dataset.
- Our experiments on diverse datasets show that GroundControl can generate novel views that adhere to the spatial layout of the satellite view, while retaining the multi-view consistency and generative prior of the diffusion model.

2. Related Work

2.1. Satellite-to-Street Generation

Street view generation from oblique satellite and aerial views. These works typically use *oblique* multi-view satellite imagery to generate a coarse scene geometry, to which visible texture (color) is applied from satellite images. This coarse reconstruction is then rendered from street-level viewpoints and subsequently refined by a dedicated network. For example, [70, 71] use a coarse textured mesh; [18] use a 3D Gaussian Splat [30]; [36] use a voxel-based representation; and AerialGo [80] use a 3D point representation similar to [18]. Skyfall-GS [34] also generates splats from multi-view satellite images, but unlike [18] they improve the Gaussian splats themselves. Unlike these approaches, our method has no restrictions on the types of satellite views it can use as conditioning signal.

	Street cond.	Ortho sat.	MV cons.	Public code	Intermediate representation	Output modality
[70]	N	N	N	N	Mesh	Panorama
[71]	N	N	Y	N	Mesh	Pose-controlled
[18]	N	N	Y	N	Splat	Pose-controlled
[36]	N	N	N	N	Voxels	Panorama
[58]	N	Y	N	N	Warp	Panorama
[49]	N	Y	N	Y	Volume	Panorama
[42, 58]	N	Y	N	Y	Heightmap	Panorama
[37]	N	Y	Y	N	Voxels	Panorama
[38]	N	Y	Y	Y	Deep point cloud	Image
[3, 52, 68, 72]	N	Y	N	Y	Direct (GAN)	Panorama
[62]	N	Y	N	Y	Polar warp	Panorama
[77]	N	Y	N	N	Ray-to-height map	Panorama
[50, 78]	N	Y	Y	N	Triplane	Panorama
[10]	Y	N*	Y	N	Heightmap	Pose-controlled
Ours	Y	Y	Y	Y†	Ray-guidance	Pose-controlled

Table 1. Compared with other related works, GroundControl enables street-level conditioning and orthorectified satellite views. We control views via pose conditioning, and outputs are multi-view consistent. Abbreviations: Street cond. = street-level conditioning; Ortho sat. = orthorectified satellite; MV cons. = multi-view consistency. * [10] accepts map input but not satellite images; † we will release code and models upon acceptance.

Street view generation from orthorectified satellite views. Methods that use orthorectified ("top-down") satellite inputs are more general, they can generate appearance of invisible building facades (without oblique imagery). Many of these methods still rely on intermediate representations: Sat2Density [49] estimates volume density, Sat2Scene [38] estimates a noisy 3D point cloud and denoises it via diffusion, Sat2Vid [37] estimates a semantically labeled point cloud, [62] uses a polar warp, [36] uses street-level voxelized geometry, and other methods estimate depth [42, 58]. More recent works Sat2Density++ [50] and SatDreamer360 [78] use ray-guided cross-attention with intermediate triplane representations. These intermediate representations lack data for large-scale supervision, and can create a quality bottleneck as they introduce artifacts downstream. Other works directly map satellite views to ground-level images using GANs [3, 19, 52, 53, 68, 72, 83]. StreetScapes [10] synthesizes street-level videos from semantic maps and height maps (without true satellite images). Other work generates fully synthetic worlds, sometimes via generated top-down views [39].

Our method is the first to directly condition a multi-view generation model on satellite images without any intermediate representations or supervision. It is closely related to [77, 78]. SatDreamer360 [78] uses an intermediate triplane

representation. It also requires panorama sequences, while ours is trained on multi-view images that are more available. [77] uses a street-to-satellite transformation similar to ours, but their "attention" model is actually a height-distribution prediction, not street-to-satellite attention. It also does not allow conditioning on street-view images (only text input).

Other tasks with satellite and street-level data. Other works combine satellite and ground-level views for related tasks. For example, [67] uses ground-level images to improve satellite segmentation, and Horizon-GS [29] combines satellite and street-level inputs for splatting. Ye *et al*. [73] address the inverse setting by synthesizing satellite views from street images. Feng *et al*. [12] use GPS as a conditioning signal for image generation, but primarily to guide high-level semantics rather than accurate scene structure.

2.2. Novel-view Synthesis and Multi-View Diffusion

Novel view synthesis (NVS) from one or more posed input images has a long history. Classical methods used image-based rendering with geometry proxies to warp input pixels [9, 23, 59]. With deep learning, DeepStereo [16] used a network to select pixels from posed input views, while [82] learned to warp input-view pixels. Another line of work directly renders explicit 3D scene representations [1, 30].

More recently, conditional diffusion models have unlocked direct generation of realistic target views [41]. Works like CAT3D [17], EscherNet [33], MultiDiff [47], PolyOculus [75] and [55, 65] *implicitly* learn multi-view consistency using ray-based positional encodings [45], which are now common in novel-view synthesis. SEVA [81] and MVGD [21] further improved temporal consistency by training on fixed-length sequences [81] and performing multi-view diffusion jointly rather than sequentially [21]. WildCAT3D [2] further extended this to in-the-wild images. A key benefit of these approaches is easier scene extrapolation, but they can hallucinate incorrect layouts and semantic artifacts, even when outputs appear plausible. An alternative line of work generates video from images without camera control [8, 26–28]; we consider camera control to be important for interactive real-world interfaces. Our work builds on the SEVA [81] backbone approach for novel view synthesis, and our innovations enable the injection of additional priors from satellite views.

3. Method

We propose an efficient approach to condition a pretrained camera-controlled multi-view diffusion model on satellite imagery. We extend the widely accepted UNet-based diffusion model architecture for satellite view conditioning by adding ray-guided cross attention parameters between the diffusion embeddings and satellite embeddings from a pretrained satellite encoder. This allows us to leverage both pretrained multi-view diffusion priors as well as pretrained satellite encoders, reducing the need for large numbers of sequence-and-satellite training data. We first provide a brief background on multi-view diffusion models in Section 3.1, then explain the satellite conditional generation approach in Section 3.2 and the ray-guided cross-attention in Section 3.3. Our training details are discussed in Section 3.4.

3.1. Multi-view Diffusion Models

Text-conditioned image diffusion models generate an image $\mathbf{x}$ from text conditions $\mathcal{T}$ by iteratively denoising an initial noise sample ϵ in T timesteps. They are probabilistic models that sample from the distribution of images conditioned on text $\mathcal{P}(\mathbf{x}|\mathcal{T})$. Camera-controlled multi-view diffusion models [21, 81] extend this generative prior to N views, with added conditioning on camera poses and intrinsic parameters. Given a set of input images $\mathcal{X}^{\text{inp}}$ and their camera parameters $\mathcal{C}^{\text{inp}}$, they generate a set of target views $\mathcal{X}^{\text{tgt}}$ for the queried target cameras $\mathcal{C}^{\text{tgt}}$. Probabilistically, they sample from the distribution $\mathcal{P}(\mathcal{X}^{\text{tgt}}|\mathcal{X}^{\text{inp}}, \mathcal{C}^{\text{inp}}, \mathcal{C}^{\text{tgt}})$.

Diffusion models [25] are trained to denoise time-dependent noisy data samples x_t. In a variance-exploding diffusion process, the noisy samples x_t are obtained from clean samples x_0 by $x_t = x_0 + \sigma_t \epsilon$ where timestep t ranges from [0, T], $\epsilon \sim \mathcal{N}(0, I)$ is noise sampled from a unit Gaussian distribution, and σ_t depends on the choice of noise schedule. The model can be trained with different prediction targets; our work uses ϵ-prediction [25].

Following standard practice, we use DDIM sampling [61] with classifier-free-guidance (CFG) during inference [24]. Starting from a pure Gaussian noise sample $x_T \sim \mathcal{N}(0, \sigma_T^2 + 1)$ at $t = T$, the latent diffusion model (LDM) predicts an ϵ_t at each timestep t, from which we compute the predicted $\tilde{x}_0$ and $\tilde{x}_{t-1} = \tilde{x}_0 + \sigma_{t-1}\epsilon_t$ using Tweedie's formula. CFG allows us to condition the ϵ_t prediction on guidance signals (*e.g.* text), so each $\epsilon_t = \epsilon_{t(u)} + \lambda_{\text{cfg}}(\epsilon_{t(c)} - \epsilon_{t(u)})$ where $\epsilon_{t(c)}$ and $\epsilon_{t(u)}$ are conditional and unconditional predictions respectively. A higher λ_{cfg} value biases the generation closer to the condition.

In the case of LDMs [54], clean samples $\mathbf{x_0}$ are lower-resolution latents encoded from the image using a pretrained VAE's encoder [31]. For multi-view LDMs, a clean sample comprises multiple target images $\mathcal{X}^{\text{tgt}}$. The model is conditioned on additional input views $\mathcal{X}^{\text{inp}}$, along with camera parameters of both input and target views $\{\mathcal{C}^{\text{inp}}, \mathcal{C}^{\text{tgt}}\}$.

3.2. GroundControl: Satellite-conditioned Multi-view Diffusion

We introduce a satellite-conditioned multi-view diffusion model GroundControl, that samples from the conditional distribution $\mathcal{P}(\mathcal{X}^{\text{tgt}} \mid \mathcal{X}^{\text{inp}}, \mathcal{C}^{\text{inp}}, \mathcal{C}^{\text{tgt}}, S)$. Here, S is the satellite image captured at roughly the same location as $\{\mathcal{C}^{\text{inp}}, \mathcal{C}^{\text{tgt}}\}$. Effectively Conditioning on satellite images S is non-trivial. The design choices in GroundControl follow

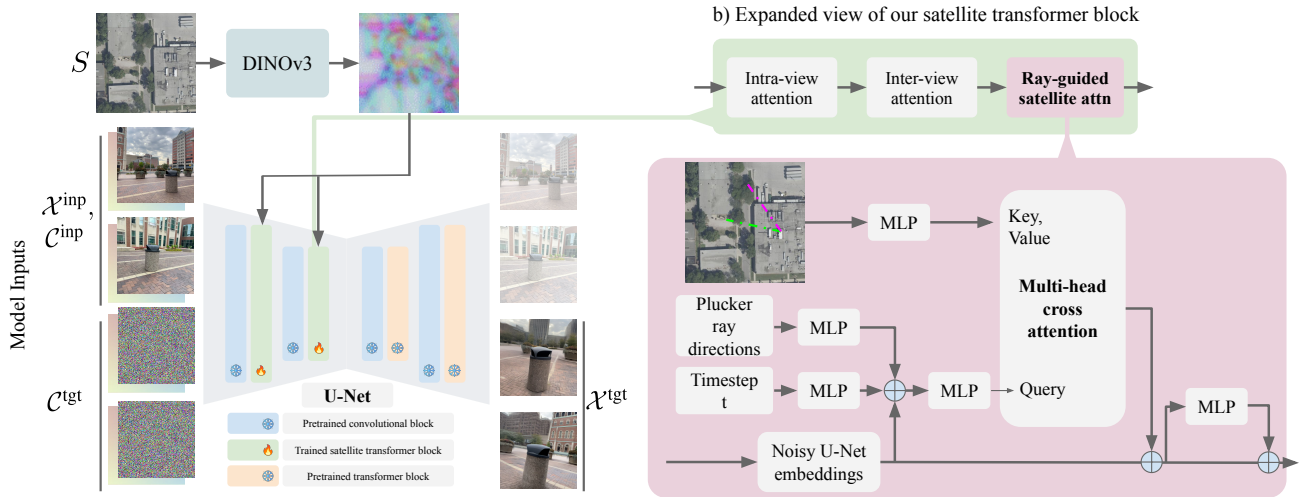


Figure 2. **GroundControl** incorporates satellite images as an efficient conditioning signal for pose-controllable multi-view diffusion models. (a) Our inputs include a DINOv3 encoded satellite image, known ground-view image(s) $\mathcal{X}^{\text{inp}}$ with associated camera parameters $\mathcal{C}^{\text{inp}}$, and target camera parameters $\mathcal{C}^{\text{tgt}}$ corresponding to the views to be synthesized. GroundControl then generates output images $\mathcal{X}^{\text{tgt}}$ for each target camera. (b) A key contribution is the integration of satellite information into the SEVA architecture [81]. (c) Our ray-guided attention sub-block fuses satellite features with noisy latent embeddings via multi-head cross-attention.

two main principles. First, we retain the multi-view generative prior $\mathcal{P}(\mathcal{X}^{\text{tgt}} \mid \mathcal{X}^{\text{inp}}, \mathcal{C}^{\text{inp}}, \mathcal{C}^{\text{tgt}})$ of the pretrained model while increasing the likelihood of samples that adhere to the satellite image condition S. Second, we use specialized parameters for satellite images because their semantics differ substantially from those of street-level images.

Hence, GroundControl is finetuned from a multi-view diffusion model [81], which provides the rich generative multi-view prior from large-scale pretraining on posed images. Our key contribution is to add residual cross-attention layers to incorporate the satellite conditioning signal into the diffusion model. In particular, the residual cross-attention layers contain a zero-initialized linear output layer which ensures that the multi-view diffusion prior is retained at the start of training. Motivated by our second design principle, we encode satellite images with DINOv3 [60]. Our street-level multi-view input images are encoded by the diffusion model itself, alongside the noisy target images. This allows us to condition the multi-view diffusion model on generalizable satellite embeddings, much like the CLIP text embeddings used in text-to-image diffusion models [51].

Conditioning on a satellite image assumes that the multi-view images are geolocalized. In particular, we assume known metric relative poses between views and approximate latitude, longitude and planar orientation of one of the views. This allows us to localize each camera with a 2D translation and 1D orientation in the satellite image.

3.3. Ray-guided Cross Attention

An effective way to condition a latent diffusion model is to add additional embeddings (or residuals) from the conditioning signal to specific blocks of the model, directly in the perspective of the generated image. When conditioning embeddings are not image-level signals (*e.g.* text, object poses [69], etc.), they are incorporated using cross-attention layers where the image embeddings cross-attend to the conditioning embeddings. Since satellite images are very different from the street-level image perspective, a viable conditioning method would be to introduce layers for the image tokens to cross-attend to the full satellite image. For a multi-view diffusion model, each image could cross-attend to a transformed satellite image which is centered at and oriented along the corresponding camera's axis.

We find that the multi-view diffusion model struggles to learn the conditioning signal from the full satellite image, likely as it is challenging for noisy image tokens to attend to within-view components of the satellite image. We address this issue by introducing ray-guided cross-attention layers, inspired by similar works [77, 78]. Our ray-guided cross-attention (RGCA) layers restricts the cross-attention to be along the 1D ray-direction in the satellite view, allowing it to learn a conditioning signal more efficiently (Figure 3). The queries in RGCA are the intermediate UNet features, along with an added embedding computed from Plucker Ray coordinates and the diffusion timestep. We find that the added Plucker and timestep embeddings are important, since the UNet features are noisy in the early diffusion stages. The architecture is shown in Figure 2 (c). The final MLP layer that produces the RGCA embeddings is zero-initialized so as to retain the pretrained LDM's latent prior. We give full details of RGCA in the supplement.



Figure 3. Our **ray-guided cross-attention** attends to tokens of the satellite image that are along the 3D ray of the street-view token.

3.4. Training

Loss. Our model is trained with an L2-loss on the predicted ϵ noise. We apply this loss only to target views.

Datasets. We use multiple training datasets that contain geotagged multi-view / panoramic images. These include self-driving datasets with GPS measurements, and geotagged street-view panoramas. In particular, we use nuScenes [5], Argoverse-v2 [66], Metropolis [43], and VIGOR [84]. For each training sample, we sample crops from different views of multi-camera rigs in [5, 43, 66] and from the same panorama in VIGOR. Additionally, we identify pairs of panoramas that are within 20m in VIGOR, and sample views from both panoramas, introducing a camera baseline to reduce scene scale ambiguity. To retain the ability of the multi-view diffusion model to be conditioned on non-driving camera trajectories, we additionally train on training set of DL3DV-10K [40] dataset (excluding Benchmark scenes) without any satellite conditioning.

Satellite images. For our training datasets, we compile satellite imagery from various sources (details are provided in the supplementary material). For each multi-view image sample, we crop the satellite images with the center as the center of all cameras, and oriented along north. We used crops of 128m$\times$128m in size, with 4 pixels per meter.

Implementation details. We add our satellite-conditioning blocks to the SEVA model and train only the satellite-conditioning blocks, keeping the SEVA multi-view diffusion weights frozen. We train on multi-view examples with 21 views, with number of input views sampled randomly from 1 to 8. We train with the Adam optimizer on 16$\times$A100 GPUs. We use a learning rate of 3e-5 and a batch size of 192. For our ray-guided cross-attention, we sample 20 points along each ray starting from a minimum distance of 0.5m up to a maximum distance of 62m in the satellite image. We use 1000 diffusion timesteps during training and 50 DDIM steps [61] at inference.

3.5. Inference

During inference, the model is conditioned on both input views and satellite images.

Dual classifier-free guidance. Our model allows the strength of conditioning on the input image and the satellite view to be controlled independently by applying separate classifier-free guidance to each of them:

$$\hat{\epsilon}_t(i,s) = \epsilon_t(u) + \lambda_i(\epsilon_{t(i)} - \epsilon_{t(u)}) + \lambda_s(\epsilon_{t(i,s)} - \epsilon_{t(i)}). \quad (1)$$

Here $\epsilon_{t(i,s)}$ is the model's output conditioned on both input and satellite views; $\epsilon_{t(i)}$ is conditioned only on input images; and $\epsilon_t(u)$ is the model's unconditional output. To enable dual CFG, we drop the input view conditioning at training time with probability 0.3, and drop both the satellite and input view conditioning with probability 0.15.

4. Experiments

Our experiments evaluate the quality of our satellite-conditioned novel view generations on outdoor datasets.

Baselines. While several works address satellite-conditioned street-view generation or multi-view/video generation in isolation, our work is unique in combining both these areas. We compare to baselines from both families for the multi-view novel-view synthesis (NVS) task. For multi-view diffusion models, we compare to the state-of-the-art SEVA [81]. As our model adopts their backbone, we find SEVA to be the closest point of comparison, as pretraining determines NVS quality to a great extent. Among models that only use satellite-conditioning, we compare to Sat2Density [49], which is one of the very few publicly-available models, see Table 1.

Datasets. We evaluate our model on validation / test splits of Metropolis [43] (36 scenes) and nuScenes [5] (150 scenes). nuScenes is a self-driving dataset with a 6-camera rig: front, front-left, front-right, back, back-left, back-right. We use 7 front camera views as input, collected in a sequence at equal time intervals. The 35 views from the other five cameras are used as targets. Metropolis contains a 3-camera rig facing forward, left, and right. We use the forward-facing camera as input and left and right facing cameras as target. This configuration allows us to evaluate target views with very little overlap with input views. These datasets are in-distribution with our training data.

As an out-of-distribution generalization experiment, we evaluate on outdoor scenes from the well-established DL3DV benchmark [40]. The original benchmark is not geo-registered; we use the geo-alignment of [63]. We use the same input-target splits from SEVA [81]: target frames are equally spaced along the trajectory (every 8th frame) and input frames are chosen by applying K-medoids clustering to the camera poses of the remaining frames. Despite this, the input-target overlap in DL3DV is substantially higher than that of our in-distribution evaluation datasets, leaving less room for improvement using satellite images.

Metrics. Following standard practice in novel-view synthesis [81] we evaluate **generation quality for the target views** (PSNR, SSIM, LPIPS). Since we focus on satellite-guided controllability in unseen regions, we evaluate the adherence of the generated images to the satellite layout. We compute the semantic masks for roads and buildings using [6], and report their **semantic IoU**. This serves as a good

	nuScenes [5]							Metropolis [43]						
	PSNR ↑	SSIM ↑	LPIPS ↓	DINO ↓	SEG ↑	IoU_{road} ↑	IoU_{bdg} ↑	PSNR ↑	SSIM ↑	LPIPS ↓	DINO ↓	SEG ↑	IoU_{road} ↑	IoU_{bdg} ↑
Street-View only														
SEVA	11.334	0.308	**0.616**	**6.140**	0.492	0.425	0.182	10.884	0.315	0.586	5.654	0.551	0.581	0.291
Satellite only														
Sat2Density [49]	10.679	0.301	0.714	6.657	0.507	–	–	9.074	0.346	0.743	6.775	0.537	–	–
Street-view + Satellite														
Ours	**11.880**	**0.331**	0.619	6.145	**0.498**	**0.437**	**0.200**	**12.809**	**0.401**	**0.547**	**5.247**	**0.605**	**0.630**	**0.372**

Table 2. Comparison with baseline methods on test datasets that **match our training domain**, *i.e.* driving datasets. Our model improves over the baseline systems. Semantic IoU includes vehicles in both the road and building masks.



Figure 4. **Qualitative results on in-distribution dataset: Metropolis** Top 4 rows show satellite images along with corresponding input and target views. Final two rows show synthesis of target views using GroundControl and SEVA. GroundControl is significantly better than SEVA at adhering to the layout of the scene.

metric for layout alignment, since obtaining vectorized layout annotations for our data or generations is non-trivial. We also report the **semantic similarity** of predictions using features from foundational semantic models as an additional layout adherence metric. In particular, we use L2 error of DINO [7] features and cosine similarity of Segment Anything [32] (SEG in tables) encoder features as proposed in [77]. We evaluate the **multi-view consistency** of our generated images using SIFT correspondences [57] on overlapping views. We report TSED (Thresholded Symmetric Epipolar Distance) [74, 81] for correspondences in these views, % of consistently generated views (with at least 10 correspondences under 8 px^2 error, and the median number of correspondences. We report the metrics on a single



Figure 5. **Qualitative results on in-distribution datasets:** Metropolis and nuScenes. Each row shows, from left: The satellite view S; an example ground-level 'seed' input view $\mathcal{X}^{\text{inp}}$; a ground truth target view $\mathcal{X}^{\text{tgt}}$; then predictions from GroundControl, [81] and [49]. Satellite images have input camera frustums plotted in red, and the target frustum in green. † We crop Sat2Density results to make pinhole images from their panoramic outputs; see supplementary for their full panoramic output.

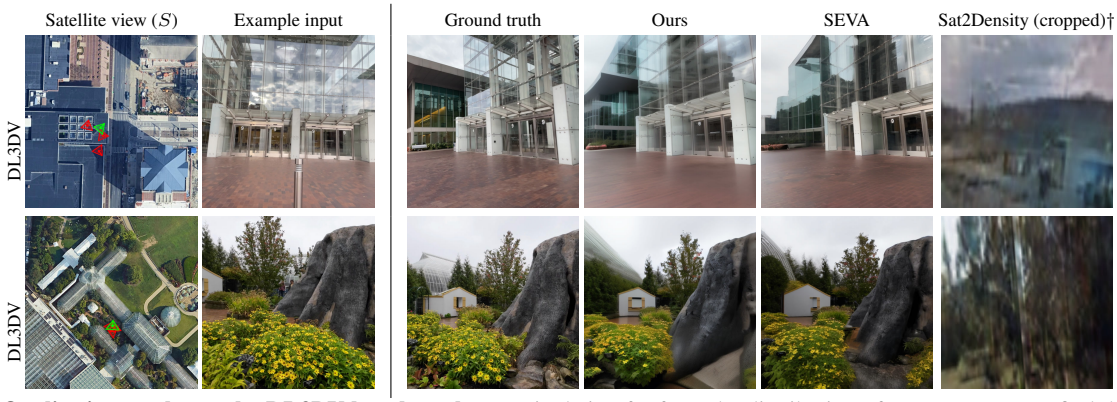


Figure 6. **Qualitative results on the DL3DV benchmark:** Despite being far from the distribution of street scenes, we find that GroundControl learns to leverage the satellite image when generating regions without sufficient overlap from sparse input views (*e.g.* buildings in the background). † We resample Sat2Density panoramas to make pinhole images; see supplementary for their full panoramic output. Red triangles in satellite view are inputs, green is the shown target view.

randomly sampled generation for each example. For a particular example, we use the same random noise sample for SEVA and our model.

In-distribution results. Table 2 shows that Sat2Density struggles to recover street-level images from satellite imagery and an appearance code. The street-view-only SEVA baseline produces perceptually plausible target views from its learned multi-view prior, but has weaker semantic IoU, indicating a weaker semantic layout-adherence. GroundControl improves image quality and layout-adherence on both datasets. On nuScenes, our gains are primarily in PSNR, SSIM, and semantic IoU, while LPIPS and DINO remain comparable to SEVA, thus slightly improving layout adherence while preserving perceptual quality. The gains

	DL3DV [40]							
	PSNR ↑	SSIM ↑	LPIPS ↓	DINO ↓	SEG ↑	TSED ↓	%Consistent ↑	#Corrs. ↑
Sat2Density [49]	10.204	0.232	0.768	7.116	0.451	N/A	0.0	0
ViewCrafter* [76]	13.350	0.260	0.547	5.230	0.541	3845.57	2.02	**332.62**
SEVA [81]	14.912	0.360	**0.407**	**3.534**	**0.674**	22.75	62.99	287
Ours	**15.170**	**0.372**	0.414	3.653	**0.674**	**18.35**	**65.56**	**290**

Table 3. Comparison with baseline methods on DL3DV, a dataset which is **out of our training data distribution**, as it contains general outdoor (not driving) scenes. We use the geolocalized dataset split from [63]. We use a total of 39 sequences with 3 input views and 37 test views each. *ViewCrafter is incompatible with our input-target split as it only generates continuous trajectories, it is evaluated on a different continuous input-target split.

	PSNR ↑	DINO ↓	SEG ↑
Satellite view as input view	11.312	5.862	0.558
W/o ray guided cross-attention (RGCA)	11.675	5.583	0.573
W/o Ray and timestep embed. in RGCA query	11.372	5.309	0.579
UNet decoder conditioning	11.291	5.400	0.572
W/o separate CFG	12.706	**5.144**	**0.608**
Ours	**12.809**	5.247	0.605

Table 4. Ablation study on the Metropolis dataset.

are substantially larger on Metropolis, where SEVA is much worse than on nuScenes. GroundControl improves across all metrics (*e.g.*, 10.88→12.81 PSNR, 0.291→0.372 building IoU). This suggests that satellite condition is more useful when the street-view prior alone provides weaker constraints on unseen scene content.

Out-of-distribution results. On DL3DV (Table 3), GroundControl generalizes well despite being trained only on driving datasets. It improves PSNR and SSIM, while remaining comparable to SEVA in LPIPS, SEG, and DINO. Satellite conditioning also improves multi-view consistency, reducing TSED and increasing the fraction of consistent views. Since DL3DV has much higher overlap between input and target views than Metropolis or nuScenes, there is less room for satellite conditioning to help. Sat2Density however performs substantially worse, highlighting the difficulty of generating these views from satellite imagery alone. Overall, these results suggest that satellite conditioning can improve generation quality and multi-view consistency even outside our training distribution.

4.1. Ablations

We ablate some of the key design choices of our method on the Metropolis dataset in Table 4. We provide more qualitative comparisons and ablation details in the supplement.

Without ray-guided cross-attention (RGCA). An alternative to using ray-guidance during cross-attention is for each street-image token to attend to all satellite tokens. We find that given our dataset size, this baseline struggles to leverage the satellite image for the novel-view synthesis, as evidenced by the lower performance in Table 4.

UNet decoder conditioning. We find that adding the satellite cross-attention layers in the UNet encoder performs better than in the decoder.

RGCA query structure: We add a raymap and diffusion timestep embedding to the queries in all RGCA layers (Fig. 2). We find that this is important, as queries are otherwise noisy, especially early in the diffusion process.

Satellite view as input view. We consider a conditioning design that leverages the same multi-view diffusion model for the satellite images. We provide the satellite image as another input view, with a learned embedding for the plucker ray input that SEVA requires. This is comparable to methods in related work [64]. It re-uses the pretrained diffusion model to encode satellite images, and the information between satellite and ground views transfers through the cross-view attention layers in SEVA. We finetune the self- and cross-view layers. As reported in Table 4, we find that this does not condition the generation as expected. This can be explained by the lack of specialized learned modules to process satellite images, as they are significantly different from ground view images captured by consumer cameras.

Without dual CFG. We replace our dual CFG scheme with a more standard single-parameter CFG, using $\epsilon_t(i, s) = \epsilon_t(u) + \lambda(\epsilon_{t(i,s)} - \epsilon_{t(u)})$ instead of eq. 1, using $\lambda = 4$. We note that using different classifier-free guidance strengths for input views versus satellite images can slightly improve generation quality (PSNR).

4.2. Qualitative Results

Figure 5 shows how our model compares with baselines [81] and [49] on driving datasets [5] and [43]. Each row shows a different scene from the two datasets. The first image shows the satellite view S; the second is an example input view. The final four columns show the same target viewpoint: with the ground truth target $\mathcal{X}^{\text{tgt}}$, our generation, SEVA [81], and the Sat2Density [49] baselines respectively. We crop Sat2Density outputs from full panoramas to match the FOV of our target cameras. We observe that our method captures the structure of unseen parts of the scene better than the baselines, giving more accurate placements of trees, buildings and street furniture. In the supplementary material we show visualize the attention maps for the satellite images.

5. Conclusion

We have introduced GroundControl, which enables effective satellite conditioning of multi-view diffusion models. Our architectural innovations ensure the satellite image is incorporated efficiently and effectively. Our extensive evaluations on public benchmark datasets validate the utility of our contributions and set a new state-of-the-art for free-viewpoint satellite-conditioned novel-view synthesis.

Acknowledgements We thank Filippo Aleotti, Gabriel Brostow, and Alan Paul for insightful discussions and support.

References

[1] Kara-Ali Aliev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *ECCV*, 2020. 3

[2] Morris Alper, David Novotny, Filippos Kokkinos, Hadar Averbuch-Elor, and Tom Monnier. Wildcat3d: Appearance-aware multi-view diffusion in the wild. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2025. 3

[3] Khawlah Bajbaa, Abbas Anwar, Muhammad Saqib, Hafeez Anwar, Nabin Sharma, and Muhammad Usman. From satellite to street: A hybrid framework integrating stable diffusion and panogan for consistent cross-view synthesis. *arXiv preprint arXiv:2509.24369*, 2025. 2

[4] Edurne Bernal-Berdun, Ana Serrano, Belen Masia, Matheus Gadelha, Yannick Hold-Geoffroy, Xin Sun, and Diego Gutierrez. Precisecam: Precise camera control for text-to-image generation. In *CVPR*, 2025. 2

[5] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 5, 6, 8, 14, 15

[6] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts, 2025. 5

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 6

[8] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024. 1, 3

[9] Shenchang Eric Chen and Lance Williams. View interpolation for image synthesis. In *SIGGRAPH*, 1993. 3

[10] Boyang Deng, Richard Tucker, Zhengqi Li, Leonidas Guibas, Noah Snavely, and Gordon Wetzstein. Streetscapes: Large-scale consistent street view generation using autoregressive video diffusion. In *SIGGRAPH*, 2024. 2

[11] Esri. Esri World Imagery. https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9. Accessed: 2025-10-01. 14

[12] Chao Feng, Ziyang Chen, Aleksander Holynski, Alexei A Efros, and Andrew Owens. GPS as a control signal for image generation. In *CVPR*, 2025. 3

[13] Florian Fervers. Tiled Web Maps. https://github.com/fferflo/tiledwebmaps. Accessed: 2025-10-01. 14

[14] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Uncertainty-aware vision-based metric cross-view geolocalization. In *CVPR*, 2023.

[15] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. Statewide visual geolocalization in the wild. In *ECCV*, 2024. 14

[16] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *CVPR*, 2016. 3

[17] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul P. Srinivasan, Jonathan T. Barron, and Ben Poole. CAT3D: Create anything in 3D with multi-view diffusion models. *NeurIPS*, 2024. 3

[18] Zhiyuan Gao, Wenbin Teng, Gonglin Chen, Jinsen Wu, Ningli Xu, Rongjun Qin, Andrew Feng, and Yajie Zhao. Skyeyes: Ground roaming using aerial view images. In *WACV*, 2025. 1, 2

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[20] Google. Google Maps Platform Documentation. https://developers.google.com/maps/documentation. Accessed: 2025-10-01. 14

[21] Vitor Guizilini, Muhammad Zubair Irshad, Dian Chen, Greg Shakhnarovich, and Rares Ambrus. Zero-shot novel view and depth synthesis with multi-view geometric diffusion. In *CVPR*, 2025. 1, 3

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 17

[23] Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel J. Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics (TOG)*, 2018. 3

[24] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 3

[26] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022. 1, 3

[27] Zhihao Hu and Dong Xu. VideoControlNet: A motion-guided video-to-video translation framework by using diffusion model with controlnet. *arXiv preprint arXiv:2307.14073*, 2023.

[28] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *arXiv*, 2022. 1, 3

[29] 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 *CVPR*, 2025. 3

[30] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *TOG*, 2023. 2, 3

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 3

[32] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023. 6

[33] Xin Kong, Shikun Liu, Xiaoyang Lyu, Marwan Taher, Xiaojuan Qi, and Andrew J Davison. Eschernet: A generative model for scalable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9503–9513, 2024. 3

[34] 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. *arXiv preprint arXiv:2510.15869*, 2025. 2

[35] Teng Li, Guangcong Zheng, Rui Jiang, Shuigen Zhan, Tao Wu, Yehao Lu, Yining Lin, and Xi Li. Realcam-i2v: Real-world image-to-video generation with interactive complex camera control. *arXiv preprint arXiv:2502.10059*, 2025. 12

[36] 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. *arXiv preprint arXiv:2408.14765*, 2024. 1, 2

[37] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Rongjun Qin, Marc Pollefeys, and Martin R. Oswald. Sat2Vid: Street-view panoramic video synthesis from a single satellite image. In *ICCV*, 2021. 1, 2

[38] Zuoyue Li, Zhenqiang Li, Zhaopeng Cui, Marc Pollefeys, and Martin R. Oswald. Sat2Scene: 3D urban scene generation from satellite images with diffusion. In *CVPR*, 2024. 2

[39] Chieh Hubert Lin, Hsin-Ying Lee, Willi Menapace, Menglei Chai, Aliaksandr Siarohin, Ming-Hsuan Yang, and Sergey Tulyakov. InfiniCity: Infinite-scale city synthesis. In *ICCV*, 2023. 2

[40] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10k: A large-scale scene dataset for deep learning-based 3D vision. In *CVPR*, 2024. 5, 8, 15

[41] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3D object. In *ICCV*, 2023. 3

[42] 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 *CVPR*, 2020. 2

[43] Mapillary. Mapillary metropolis dataset. https://www.mapillary.com/dataset/metropolis. Accessed: 5 November 2025. 5, 6, 8, 14, 15

[44] Microsoft. Azure Maps. https://azure.microsoft.com/en-us/products/azure-maps. Accessed: 2025-10-01. 14

[45] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 3

[46] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, 2024. 2

[47] Norman Müller, Katja Schwarz, Barbara Rössle, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Multidiff: Consistent novel view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10258–10268, 2024. 3

[48] Dmitry Petrov, Pradyumn Goyal, Divyansh Shivashok, Yuanming Tao, Melinos Averkiou, and Evangelos Kalogerakis. Shapewords: Guiding text-to-image synthesis with 3D shape-aware prompts. In *CVPR*, 2025. 2

[49] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *ICCV*, 2023. 2, 5, 6, 7, 8

[50] 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. *arXiv preprint arXiv:2505.17001*, 2025. 2

[51] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 4

[52] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *CVPR*, 2018. 2

[53] Bin Ren, Hao Tang, and Nicu Sebe. Cascaded cross mlp-mixer GANs for cross-view image translation. *arXiv preprint arXiv:2110.10183*, 2021. 2

[54] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3

[55] Mehdi SM Sajjadi, Henning Meyer, Etienne Pot, Urs Bergmann, Klaus Greff, Noha Radwan, Suhani Vora, Mario Lučić, Daniel Duckworth, Alexey Dosovitskiy, et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *CVPR*, 2022. 3

[56] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulo, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. OrienterNet: Visual Localization in 2D Public Maps with Neural Matching. In *CVPR*, 2023. 12

[57] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[58] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *PAMI*, 2022. 2

[59] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3D photography using context-aware layered depth inpainting. In *CVPR*, 2020. 3

[60] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 4

[61] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3, 5

[62] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *CVPR*, 2021. 2

[63] Matias Turkulainen, Akshay Krishnan, Filippo Aleotti, Mohamed Sayed, Guillermo Garcia-Hernando, Juho Kannala, Arno Solin, Gabriel Brostow, and Daniyar Turmukhambetov. Cross-View Splatter: Feed-Forward View Synthesis with Georeferenced Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2026. 5, 8

[64] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *CVPR*, 2025. 8

[65] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 3

[66] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *NeurIPS Datasets and Benchmarks*, 2021. 5, 14, 15

[67] Scott Workman, M Usman Rafique, Hunter Blanton, and Nathan Jacobs. Revisiting near/remote sensing with geospatial attention. In *CVPR*, 2022. 3

[68] 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*, 2022. 2

[69] Ziyi Wu, Yulia Rubanova, Rishabh Kabra, Drew A. Hudson, Igor Gilitschenski, Yusuf Aytar, Sjoerd van Steenkiste, Kelsey Allen, and Thomas Kipf. Neural Assets: 3d-aware multi-object scene synthesis with image diffusion models. In *Advances in Neural Information Processing Systems*, 2024. 4

[70] Ningli Xu and Rongjun Qin. Geospecific view generation: Geometry-context aware high-resolution ground view inference from satellite views. In *ECCV*, 2024. 1, 2

[71] Ningli Xu and Rongjun Qin. Satellite to GroundScape: Large-scale consistent ground view generation from satellite views. In *CVPR*, 2025. 1, 2

[72] Hongji Yang, Yiru Li, and Yingying Zhu. Retrieval-guided cross-view image synthesis. *arXiv preprint arXiv:2411.19510*, 2024. 2

[73] Junyan Ye, Jun He, Weijia Li, Zhutao Lv, Yi Lin, Jinhua Yu, Haote Yang, and Conghui He. Leveraging BEV paradigm for ground-to-aerial image synthesis. In *ICCV*, 2025. 3

[74] Jason J. Yu, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023. 6

[75] Jason J. Yu, Tristan Aumentado-Armstrong, Fereshteh Forghani, Konstantinos G. Derpanis, and Marcus A. Brubaker. Polyoculus: Simultaneous multi-view image-based novel view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 3

[76] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 8, 12

[77] 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. *arXiv preprint arXiv:2502.03498*, 2025. 2, 3, 4, 6

[78] Xianghui Ze, Beiyi Zhu, Zhenbo Song, Jianfeng Lu, and Yujiao Shi. SatDreamer360: Multiview-consistent generation of ground-level scenes from satellite imagery. *arXiv preprint arXiv:2506.00600*, 2025. 2, 4

[79] Lvmin Zhang, Maneesh Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *CVPR*, 2023. 2

[80] Fuqiang Zhao, Yijing Guo, Siyuan Yang, Xi Chen, Luo Wang, Lan Xu, Yingliang Zhang, Yujiao Shi, and Jingyi Yu. AerialGo: Walking-through city view generation from aerial perspectives. *arXiv preprint arXiv:2412.00157*, 2024. 1, 2

[81] Jensen (Jinghao) Zhou, Hang Gao, Vikram Voleti, Aaryaman Vasishta, Chun-Han Yao, Mark Boss, Philip Torr, Christian Rupprecht, and Varun Jampani. Stable virtual camera: Generative view synthesis with diffusion models. In *ICCV*, 2025. 1, 3, 4, 5, 6, 7, 8, 12, 19

[82] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *ECCV*, 2016. 3

[83] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[84] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, 2021. 5, 14

[85] Chen Ziwen, Hao Tan, Kai Zhang, Sai Bi, Fujun Luan, Yicong Hong, Li Fuxin, and Zexiang Xu. Long-lrm: Long-sequence large reconstruction model for wide-coverage gaussian splats. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. 13

GroundControl: Satellite-Conditioned Multi-View Diffusion for Reconstruction-Free Novel View Synthesis

Supplementary Material

This supplementary material provides additional experiments and analysis, more qualitative results, along with more details on datasets, the model, and more baselines. In Section A, we provide additional experiments that evaluate our robustness to satellite alignment, quantify performance across viewpoint overlaps, and provide additional baselines. We provide additional details on our training data in Section C and the ray-guided cross-attention architecture in Section D. We provide more qualitative results in B, highlighting our multi-view consistency, challenging cases, comparisons on our ablations, and a discussion on the impact of dual CFG. We also include a discussion of our limitations and directions for future work.

A. Additional Experimental Details

A.1. Performance across different input-target view overlap levels

In this section we compare the interpolation vs extrapolation capabilities of our model. So, we evaluate how our model's performance varies across different levels of overlap between input and target views. We use 20 scenes from Metropolis, and sample 20 different input-target view configurations for each scene at random. For each configuration, we compute the maximum number of SIFT correspondences for each target view across all input views. This indicates the maximum amount of visual overlap for that target view. We plot the PSNR for the target views bucketed by this indicator of visual overlap in Figure I. We find that our model provides improvements across all levels of visual overlap. The improvement in the low-overlap region ($<$ 200 correspondences) is noticeably consistent.

A.2. Robustness to satellite pose misalignment

Our method assumes that the transformation from the reference frame used in the diffusion model to the satellite image is known. In practice, this geo-alignment can be achieved using methods like OrienterNet [56]. To benchmark on this more practical setting, we introduce random pose offsets in our satellite alignment for nuScenes and DL3DV in Table I. We add random translation and rotation offsets to the alignment (upto 10m translation, 10° rotation offset). OrienterNet falls within this range (99.5% recall $<$ 5m, 5°). GroundControl performance only suffers very little degradation in this range, and remains competitive with SEVA with improvements in PSNR/SSIM.

A.3. Additional baselines

In the main paper, we use Sat2Density as our satellite conditioned street view generation baseline. Although it mainly relies on the satellite image, it also uses the input street-level views to compute a global appearance embedding that contains lighting information, which the generation model is conditioned on. For our novel-view synthesis diffusion baseline, we primarily compare to SEVA [81].

	nuScenes				DL3DV			
offset (m,°)	PSNR↑	SSIM↑	DINO↓	Seg↑	PSNR↑	SSIM↑	DINO↓	Seg↑
SEVA	11.334	0.308	**6.140**	0.492	14.912	0.360	**3.534**	**0.674**
0,0	**11.880**	**0.331**	6.145	**0.498**	**15.170**	**0.372**	3.653	**0.674**
5,5	11.518	0.311	6.156	0.490	15.131	**0.372**	3.679	0.670
10,10	11.509	0.311	6.156	0.490	15.124	0.371	3.682	0.669

Table I. **Robustness to pose misalignment:** Performance degradation is mild with alignment offsets.

Since GroundControl is built on top of SEVA, it is comparable in terms of number of parameters and pretraining data. Here we discuss comparisons to more novel-view synthesis baselines. We also discuss comparisons to video generation models.

A.3.1. Video generation models

Vanilla image-to-video (I2V) diffusion models are ill-suited for novel-view synthesis (NVS) as they cannot generate specific target viewpoints. Even when finetuned for camera conditioning, the target frames are required to form a continuous trajectory, which is incompatible with our task of predicting arbitrary viewpoints. They are also trained with only 1 or 2 input-view conditioning. Our task involves multiple input views. We make a best-effort comparison of our model to some video generation models as described below.

ViewCrafter on DL3DV We compared against the ViewCrafter [76] baseline in Table 3 of our paper. It uses a video diffusion model for novel-view synthesis by conditioning it on partial scene point clouds, obtained from a sparse set of input views. It also assumes that the frames to be generated sequentially follow the first frame (input). We therefore provide it an additional view, in the form of the first target view. This makes the comparison unfair (in ViewCrafter's favor). We use the 4 views to construct the point cloud. We then render the point cloud for 25 test views and use the diffusion model to generate the corresponding frames conditioned on these renders. As the model can only generate 25 frames at a time, and our splits have about 40 target views, we repeat this process for subsequent frames. When a set of frames has fewer than 25 test views, we duplicate the last test frame as needed so that we always generate a full set of 25 frames. Despite this, both SEVA and our model significantly outperform ViewCrafter in Table 3 (main paper).

Comparison on continuous trajectories To compare against the more recent video models, *e.g.* RealCamI2V [35] and Veo 3.1 (proprietary), we created a 1-view (first frame as input) and 2-view (first and last frames as input) sequential input-target split on Metropolis in Tab. II. This is a specialization of the more general novel-view synthesis task that restricts the target views to a

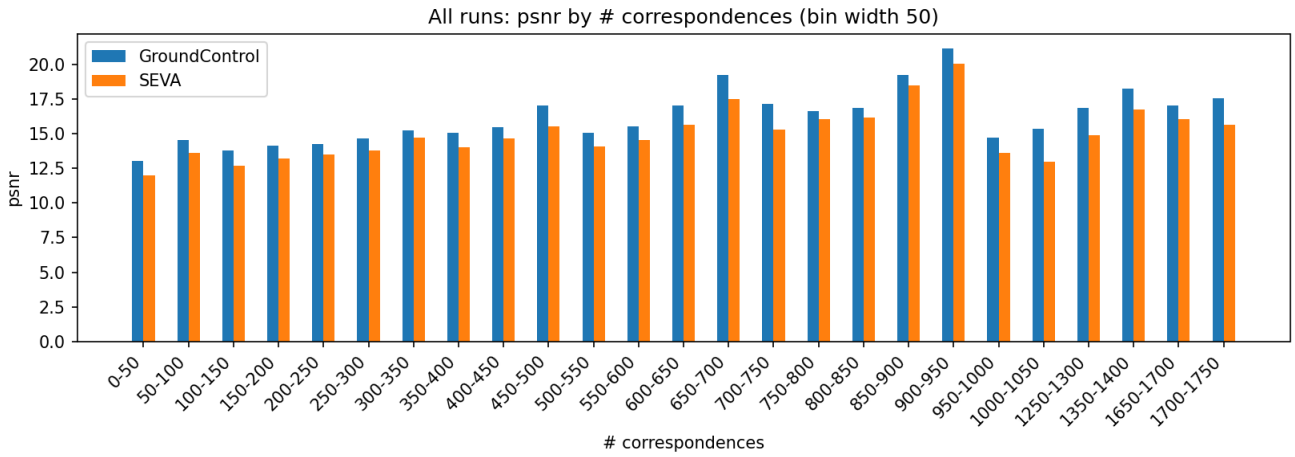


Figure I. **PSNR performance vs. image overlap measured by the number of correspondences on Metropolis dataset**. GroundControl obtains better performance compared to the SEVA baseline across all overlap regimes.

Views	Model	Straight				Turns			
		PSNR↑	LPIPS↓	DINO↓	Seg.↑	PSNR↑	LPIPS↓	DINO↓	Seg.↑
1	RealCamI2V	11.57	0.508	**4.482**	**0.627**	10.28	**0.599**	**5.672**	**0.557**
	SEVA	12.35	**0.505**	4.718	0.611	10.48	0.616	5.968	0.523
	Ours	**12.73**	0.518	4.687	0.623	**11.12**	0.612	5.787	0.548
2	*Veo 3.1*	***14.81***	***0.395***	***3.245***	***0.707***	***11.99***	***0.538***	*5.068*	*0.574*
	SEVA	12.99	0.472	4.427	0.633	11.29	0.545	5.196	0.571
	Ours	13.69	0.474	4.313	0.654	11.82	0.545	**5.057**	**0.593**

Table II. Comparison to image-to-video diffusion models on the Metropolis dataset for a video generation task (continuous trajectories). We divide our test set into straight car trajectories vs turns. Our model is slightly better than RealCamI2V in both cases, while being comparable to Veo on turns, with better semantic-feature metrics.

continuous trajectory. We further divide this into straight and turning trajectories to analyze model performances separately in the two cases. RealCamI2V is a camera-controlled video model that can be conditioned on a single input frame only. Even on this video generation task, we are slightly better than RealCamI2V (noticeably better on PSNR, while being slightly worse on DINO and segment-anything feature similarity). We also compare against the proprietary Veo model on a subset of Metropolis with 2 input views (Veo supports a start and end frame conditioning). It is a proprietary model likely trained with more data and model parameters. It is better than our model on the straight trajectories, while being comparable on turns.

A.3.2. Feedforward Regression Models

In our evaluation we exclude baselines which are trained for the task of reconstruction, rather than generation. This is because our task focuses on synthesizing accurate content in regions which are not visible in the input views using information from the satellite view. Comparing to models which do not even attempt to generate such content is unfair to those models and uninformative to the performance on our desired task. In Fig. X, we illustrate this with an example output from LongLRM [85] on a scene from our 3-view DL3DV test split. We repeat our 3 input views so that LongLRM gets 32 input views total, and run inference at its native resolution, both of which favor its performance. LongLRM does not perform well because much of the target view is not visible in the input views at all (resulting in much of the target view being blank), and the highly extrapolative nature of the task means that visible areas of the scene also do not look good. For this reason we do not consider feedforward reconstruction models such as LongLRM to be suitable baselines for our task.

Dataset	p100	p75	p50	p25	p10	p0
nuScenes	2.714	0.850	0.527	0.221	-0.095	-0.817
Metropolis	3.444	2.396	1.962	1.449	0.941	0.258

Table III. Distribution (percentile values) of per-scene $(\text{PSNR}_{\text{ours}} - \text{PSNR}_{\text{SEVA}})$. These show that satellite conditioning improves generations in most cases on nuScenes and Metropolis.

A.4. Effectiveness of satellite imagery

To better quantify our model's performance compared to SEVA (no satellite conditioning), we further analyze the metrics in Table 2 (main paper). For each scene, we compute $\delta = (\text{PSNR}_{\text{ours}} - \text{PSNR}_{\text{SEVA}})$, the improvement in PSNR for our generation vs SEVA. We tabulate the percentile distribution for this metric in Table III. Satellite helps in all cases on Metropolis (lowest value of δ is positive). On nuScenes, our model is noticeably worse ($>$ 0.5 dB) for only 3.7 % scenes. We found *2 common challenges* in these cases on nuScenes: a) A large vehicle (truck / tram) occludes most of the input FOV, or b) the camera is stationary (*e.g.* while waiting at intersections). Both these cases give GroundControl very little ground-level context, which it needs to fuse information from satellite view.

We also performed another experiment on misaligned satellite images. Instead of the correct satellite image, we provided different images with 2 types of misalignments (i) by flipping the satellite image horizontally, and (ii) by using a mismatched satellite image from another random scene. We performed this analysis on

	PSNR ↑	DINO ↓	SEG ↑
Flipped satellite	12.017	5.782	0.513
Mismatched satellite	12.052	5.591	0.530
Ours	**12.809**	**5.247**	**0.605**

Table IV. Impact of corrupted satellite imagery: using flipped satellite images or mismatched satellite images degrades performance.

Metropolis and report the results in Table IV. These corruptions indeed degrade generation quality. However, the degradation is not too significant. This is likely because many scene layouts in driving datasets are similar (even under the flip transformation). Moreover, the input image condition provides a strong prior that the model does not violate even with a misaligned satellite image (it prefers the street image signals, under a conflicting satellite signal).

B. Additional Qualitative Results

Figures II and III show additional qualitative results on the Metropolis dataset. Figure II particularly highlights the multi-view consistency of our model, and the satellite adherence in comparison to SEVA. Figure IV shows our results on some challenging scenes from nuScenes where the input images have reduced visibility due to weather conditions. For the Sat2Density crops used in the main paper, we include the full generated panoramas in Figure VIII. We also include a qualitative comparison to state-of-the-art proprietary video model Veo in Figure XI.

B.1. Multiple Noise Samples

We provide qualitative comparisons of multiple generations from SEVA and our model in Figure V. This allows us to compare the diversity of samples from both models. We find that the satellite conditioning indeed reduces the variance of layouts generated by our model in comparison to SEVA. Our model's generations still retain diversity in the appearance or height of buildings as this cannot be observed from the satellite view.

B.2. Effect of Dual CFG

As discussed in the main paper, we use a dual-CFG scheme which allows us to separately control the strength of the conditioning on the input views and on the satellite image. In Fig. VI we show the impact of varying the satellite CFG value while keeping the image CFG value fixed (something made possible by our dual-CFG approach). As we increase the satellite CFG, the model correctly extends the monorail and the lane divider, and removes the building in the background, in accordance with the satellite image. However at very high values, the quality of the generated image begins to decrease. Our dual-CFG approach gives control over this trade-off. (Note that at satellite CFG 0 the output is independent of the content of the satellite image, because the satellite guidance term from eq. 1 of the main submission gets multiplied by zero.)

B.3. Ablations

Figure VII shows our ablations where we compare our model with variants which don't have satellite imagery as input (**No satellite input**) and a model which uses satellite as input but doesn't include ray-guided cross-attention (**No ray CA**). We see that our model has better fidelity to the overall structure of the scene when compared with the ground truth, vs. both the alternative formulations.

Dataset	MV	Pano	#Scenes	Contrib	Satellite source
Argoverse 2 [66]	Yes	No	700	10%	Google + ArcGIS
nuScenes [5]	Yes	No	700	13%	ArcGIS
Metropolis [43]	Yes	No	75	7%	Original dataset
VIGOR [84]	Yes*	Yes	52105†	70%	Azure Maps

Table V. Training datasets. * We pair nearby panoramas to create a multiview version of the dataset. † Unlike other datasets, the number of scenes in VIGOR corresponds to the number of panoramas. MV = is multi-view? Pano = is panorama? Contrib = training contribution. We train jointly on all datasets.

B.4. Visualizations of Attention Maps for Satellite Rays

See Figure IX for visualization of satellite image attention of one of our map conditioning blocks.

C. Data processing

C.1. Evaluation datasets

Metropolis Each scene contains 7 input views from the front camera on the car, and 14 target views (7 each from the left and right cameras). The front, left, and right cameras each have a 90° FOV and separated by 90° in their view directions, leaving little-to-no overlap between input and target views at the same timestep. However, there remains some overlap due to the car's motion over the 7 timesteps.

nuScenes We used the front camera on the 6-camera rig as the input view and the remaining 5 cameras as the target views, for 7 timesteps (7 input views, 35 target views total). The images are center-cropped to square, with an FOV of approximately 40°. This setup is similar to that of Metropolis, with very little overlap between target and input views at the same timestep. Any overlap comes from the camera motion across timesteps.

C.2. Training datasets

For our training datasets, we augment the existing dataset with satellite imagery from various sources. Given GPS latitude and longitude information from the original datasets, we query tiled web map providers such as Google Maps [20], Azure Maps [44], and Esri World Imagery [11].

The queries are done using the TiledWebMaps package [13–15] at a sampling density of 4 pixels per meter, at a resolution of 128m $\times$ 128m.

We will provide the exact GPS coordinates, code, and query parameters used to query satellite images for each scene so that our experiments can be reproduced using any compatible tile provider (Esri, Google, Azure/Microsoft).

VIGOR[84] The original VIGOR dataset contains panorama images with non-centered satellite images with large zoom levels. We regenerate the dataset satellite images such that they are centered at the panorama latitude and longitude location.

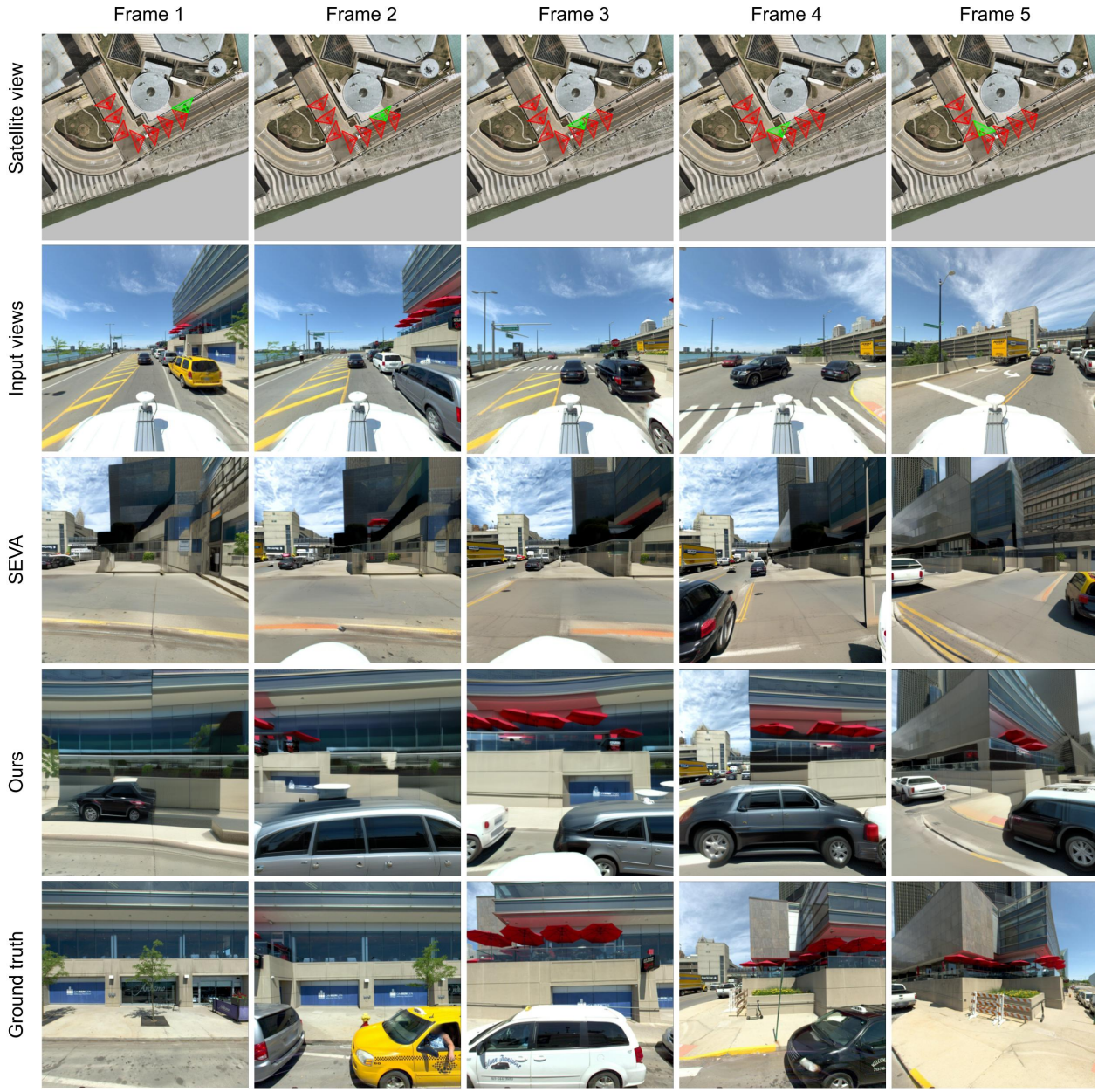


Figure II. **Multi-view generations** on the Metropolis dataset: These are multi-frame generations from a single diffusion process with 7 input views (shown in red on the satellite images). We show target views from 5 consecutive timesteps (green in the satellite images). Our model's generations are multi-view consistent, and adheres to the spatial layout information in the satellite image.

Metropolis[43] The Mapillary Metropolis dataset provides high-resolution satellite imagery from which we use a crop that is centered at the centroid of the input and target camera views.

Argoverse 2[66] We use GPS locations to query Esri World Imagery (ArcGIS) for the satellite images, except Pittsburgh where we use Google Maps imagery because we found that the ArcGIS satellite image quality was poor for this region.

nuScenes[5] We use GPS locations to query Esri World Imagery (ArcGIS) for the satellite images.

We found that training our satellite conditioning parameters on relatively small datasets (*e.g.* Argoverse or VIGOR alone) can cause them to overfit to the camera trajectory distributions on these datasets. Most of our datasets use a multi-view inside-out facing camera rig with straight-line motion (or no motion in case of panoramas). The model on these alone does not generalize to other scenes with handheld cameras. For this reason, we add data that is closer to the training distribution of SEVA to retain its performance beyond our training dataset. We include scenes from DL3DV [40] with the benchmark scenes removed during training. For these scenes, we do not use any satellite encoder features

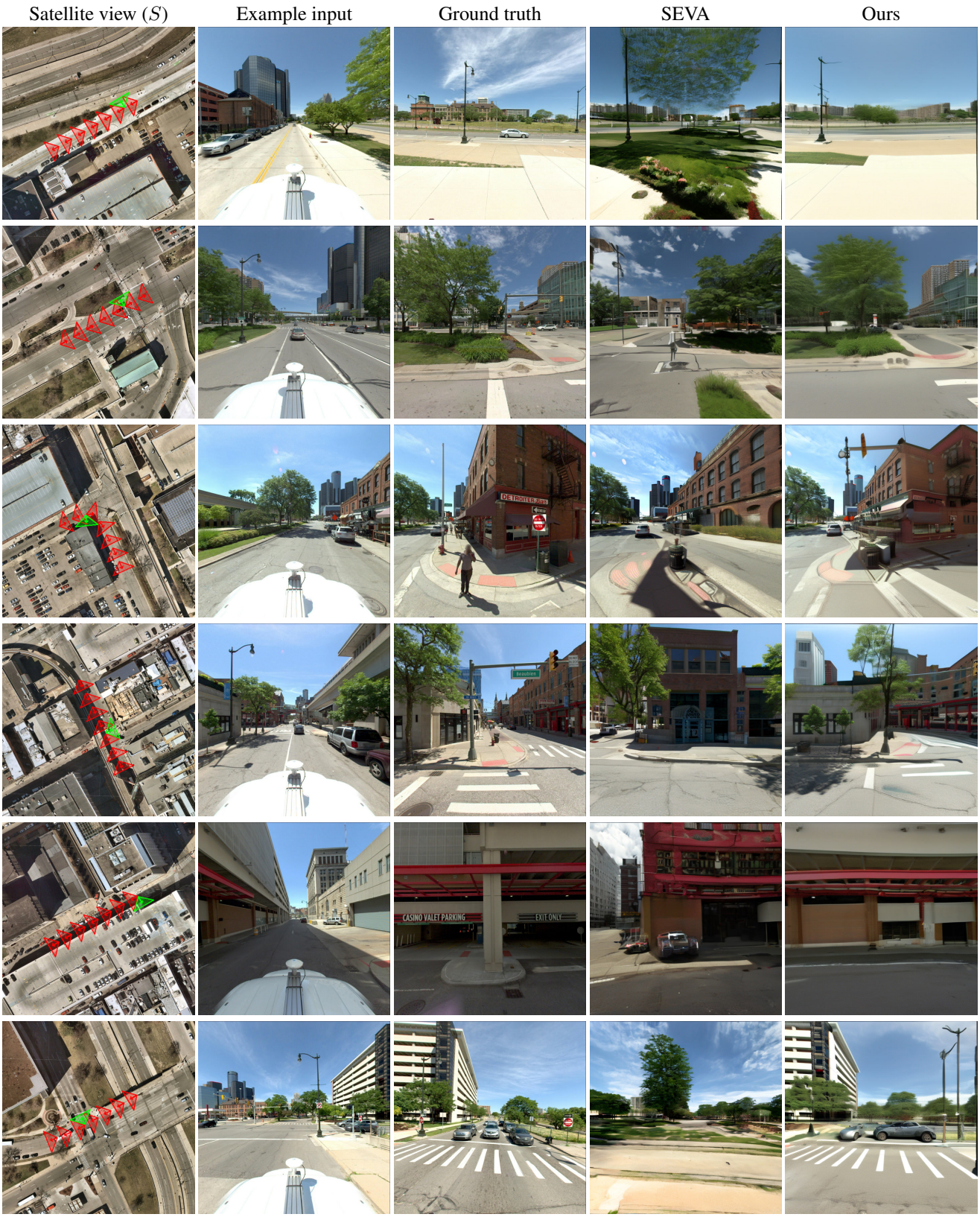


Figure III. **Qualitative results on the Metropolis dataset.** We show results of our satellite-conditioned model on sequences of length 21 frames. This includes 7 input frames (from the front camera) and 14 target frames (from the left and right cameras at the corresponding timestamps). The SEVA model without satellite conditioning underperforms in this sparse input-view regime which has little overlap with the target views.

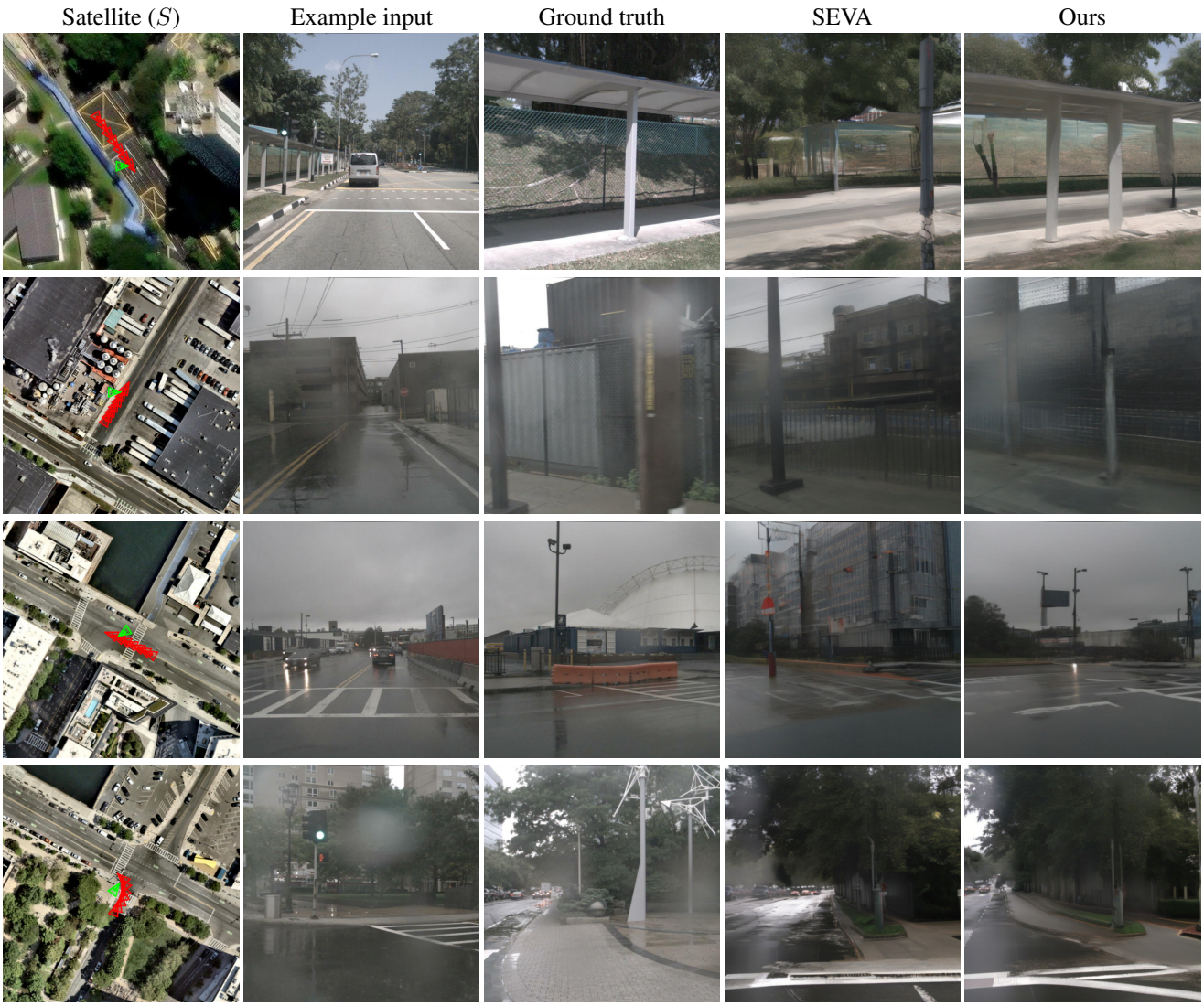


Figure IV. **Qualitative results on nuScenes dataset** These show cases with very little overlap between input views (red in satellite view) and the target views (green in satellite view). Our model's generations adhere more closely to the layout of the scene, since these are conditioned on the satellite view.

(the conditioning layers get a zero-input). Our training data mix is summarized in Table V.

D. Additional Model Details

We provide here more details about our model architecture, in particular - details about the ray-guided cross-attention block that we propose.

D.1. Ray-guided Cross Attention

We insert our **ray-guided satellite cross attention block** after the intra-view self-attention and inter-view self-attention blocks in SEVA. The satellite-conditioned cross attention block primarily contains a satellite cross-attention layer and an MLP, both of which are applied in a residual manner [22] to the diffusion embeddings that are input to the block.

Each such block receives four inputs:

- **multi-view diffusion embeddings** I: These are the intermediate SEVA features at that block. The number of channels and resolution varies throughout the model.
- **Ray embeddings** R: The Plucker coordinates for the 3D ray direction, which does not include the yaw component *i.e.* is computed using the rotation matrix with the yaw removed. These ray coordinates are passed to an MLP to encode to a fixed dimension.
- **The diffusion timestep embedding** t: This diffusion timestep t is passed to an MLP to encode to a fixed dimension.
- **Satellite feature samples** S: 20 samples along each street image ray, with 1D position embedding added along each ray. These samples are evenly spaced between the camera origin and

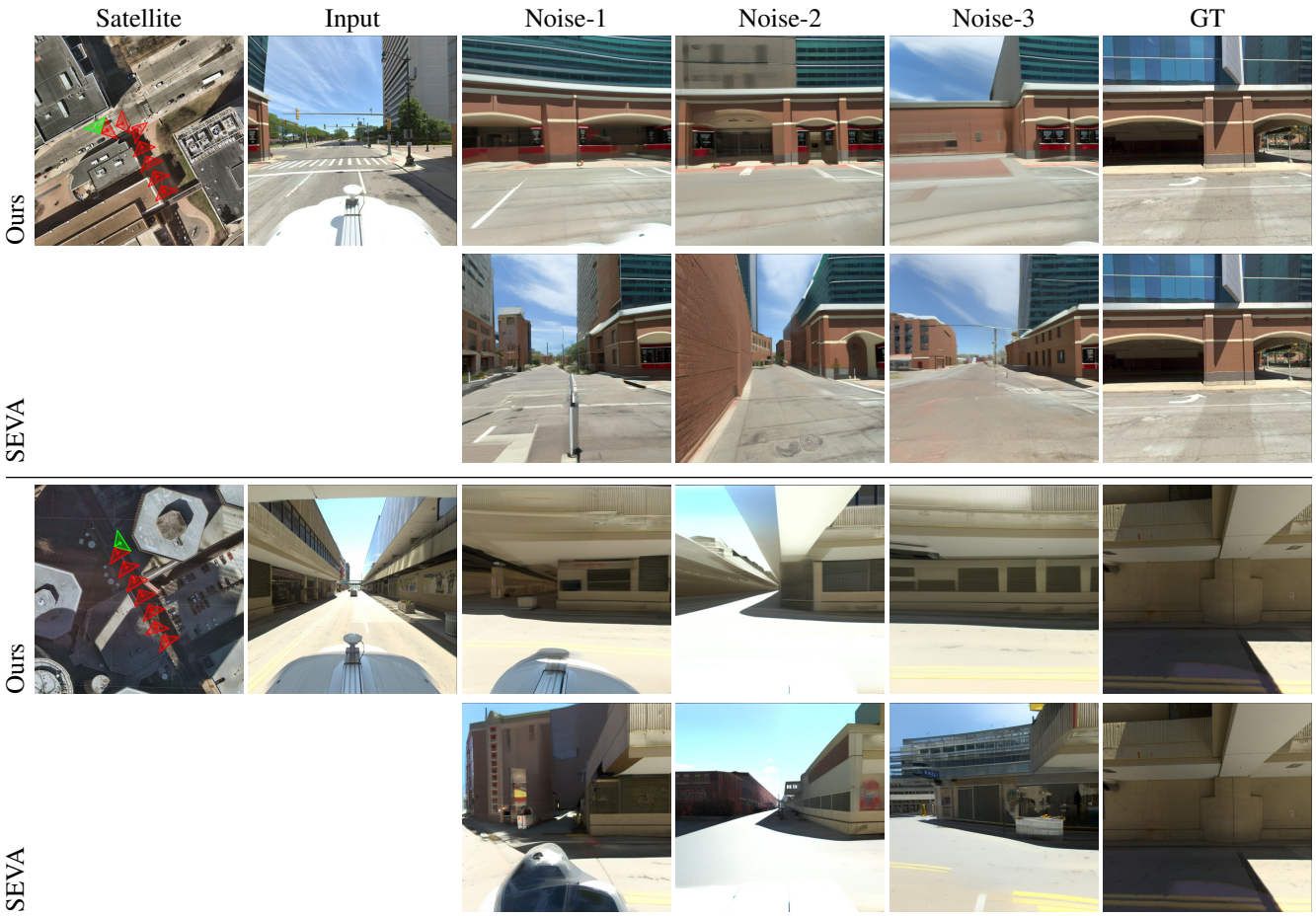


Figure V. **Qualitative results with multiple noise samples**: We compare variations in the generations from SEVA to that of our model, when starting the denoising process with different noise samples. In regions that are non-overlapping with the input view, SEVA's generations vary more, and are inconsistent with the scene layout. Since our generations are conditioned on satellite imagery, they show less variance in the generated layout, while still generating diverse samples.

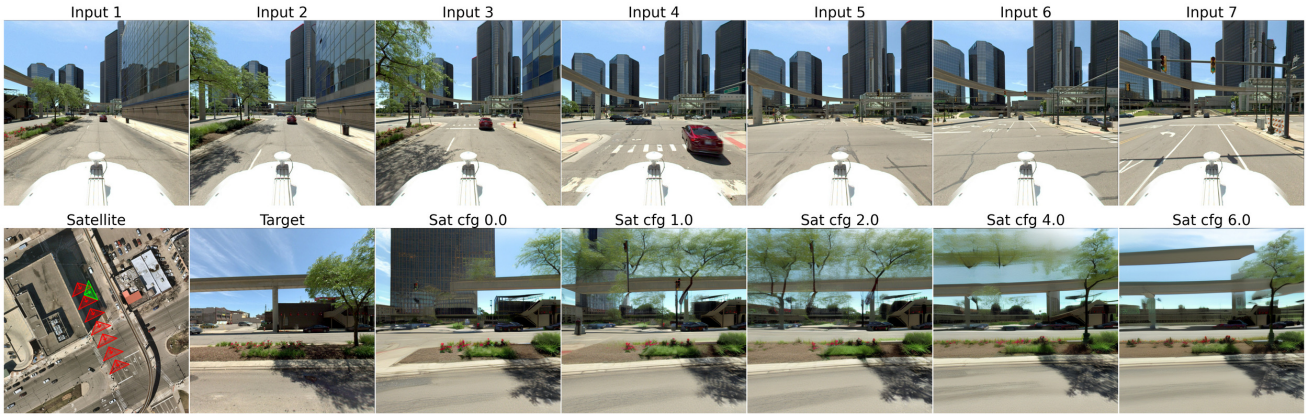


Figure VI. **Varying satellite CFG on a scene from Metropolis** Increasing the satellite CFG strengthens the adherence to the satellite image, but higher values trade off against image quality. The input views on the top row correspond to the red frustums in the satellite image, and the predicted views shown in the second row correspond to the green frustum.



Figure VII. **Qualitative ablation results on NuScenes.** Here we compare three alternatives: a model with no satellite input (No sat); a model with satellite input but no ray-based cross-attention (no ray CA), and finally our model. Columns on the left show the satellite view (S) with the input views in red and target view in green, a representative input view (input) and the ground truth (GT) corresponding to this target view. We note that our model tends to have better fidelity to the ground truth layout.

a maximum depth of 62m.

Note that the ray embedding MLP and diffusion timestep MLP are shared across different satellite conditioning blocks. These are added to the UNet's intermediate features to obtain the cross-attention queries. In our ablations, we find that it is important to use these embeddings, as the UNet intermediate features are very noisy in the early diffusion process. Since the satellite image is top-down, all patches along a particular column of the street-image attend to the same ray along the satellite image (assuming images are approximately gravity-aligned). The raymap embeddings in the attention query can therefore, help tokens from the same column attend to different satellite image tokens.

Given these inputs, the block updates the diffusion embeddings I as follows:

query $\leftarrow$ MLP(MLP(I) + R + t)
query $\leftarrow$ layernorm(query)
key $\leftarrow$ S
value $\leftarrow$ S
residual $\leftarrow$ MHCA(query, key, value)
$I \leftarrow I$ + layernorm(residual)
$I \leftarrow I$ + MLP(layernorm(I))

The output is passed onto the next diffusion model block.

The ray-guided cross-attention is different from a vanilla cross attention in that each image token selectively attends to the satellite tokens along its 3D ray, as opposed to attending to the whole satellite image. This efficient mechanism enables us to condition a large multi-view diffusion on satellite information by training the added satellite-conditioning layers on the relatively small easily accessible paired image-satellite datasets.

E. Limitations

Overall, our approach for improving NVS is only suitable for use in GPS-localized outside locations. We build our satellite-conditioned model on top of an existing multi-view diffusion model (SEVA [81]). SEVA uses an entangled camera pose and appearance representation (by concatenating raymaps and images) which means that any modifications to its embeddings can affect either appearance or poses. When used on datasets with camera trajectories that are completely out-of-distribution, we notice that

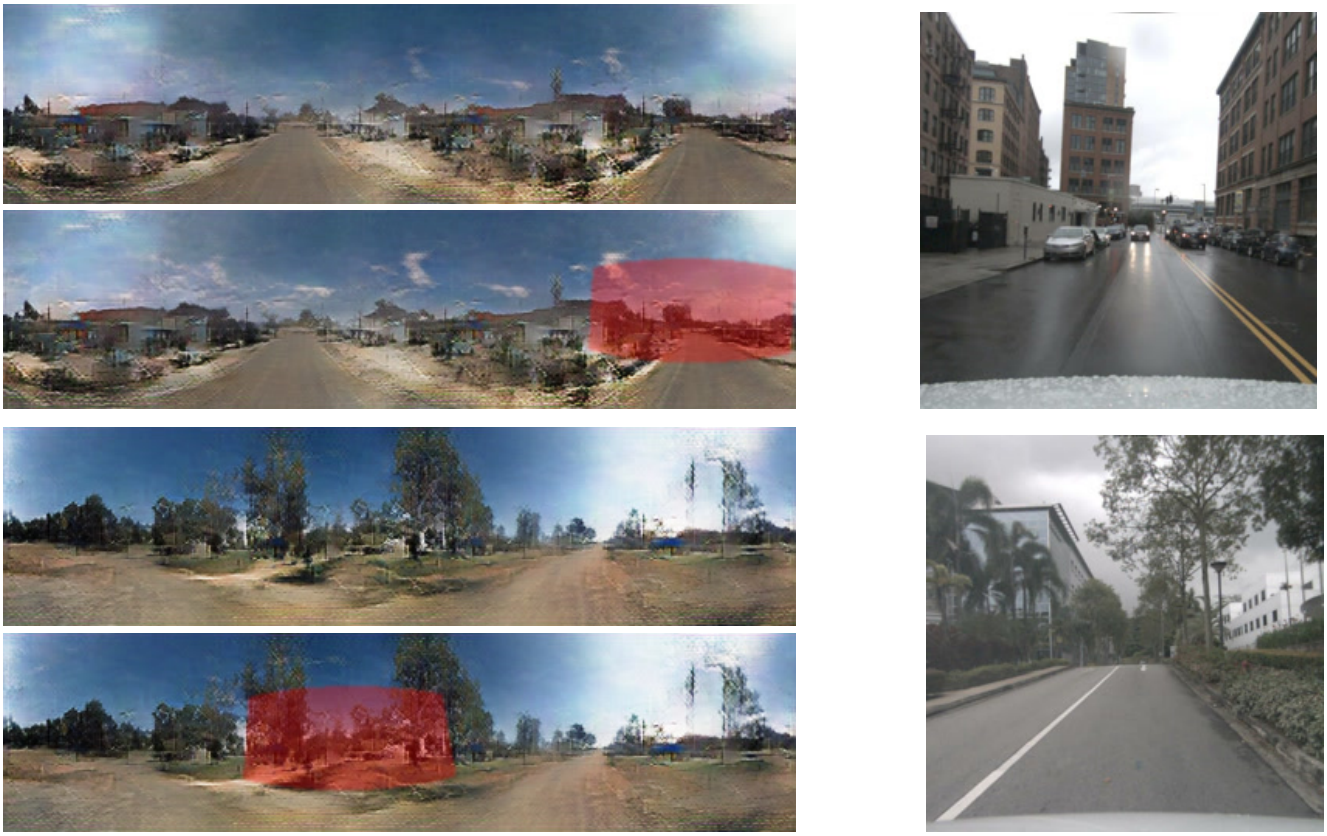


Figure VIII. Visualization of Sat2Density results on nuScenes, complementing Fig. 5. The predicted panorama is 128×512. Each row shows (left) the panorama prediction and below it a highlight showing the region used to generate the cropped image, and (right) the ground-truth target image.

satellite conditioning can affect the multi-view consistency (pose conditioning) or appearance (saturation or image quality) of the generated images. Our training minimizes this by balancing various different datasets. We also notice that in many cases the satellite conditioning can induce an incorrect generation by being sensitive to shadows of buildings that are visible in the satellite image. Unsurprisingly, our model can also struggle when the images are captured in covered areas, *e.g*. under a bridge, under tree canopies, parking lots, etc. If there have been large changes in the scene since the satellite view was captured, *e.g*. buildings have been built or demolished, then performance may degrade.

F. Future work

Our work can be revisited with emerging DiT-based diffusion models, which are more flexible than UNet-based models. Currently, our model does not use any viewpoint metadata available for satellite images. Incorporating overhead viewpoint information this could allow joint conditioning on satellite and oblique aerial images. Incorporating vectorized semantic map conditioning is also an exciting direction for future work.

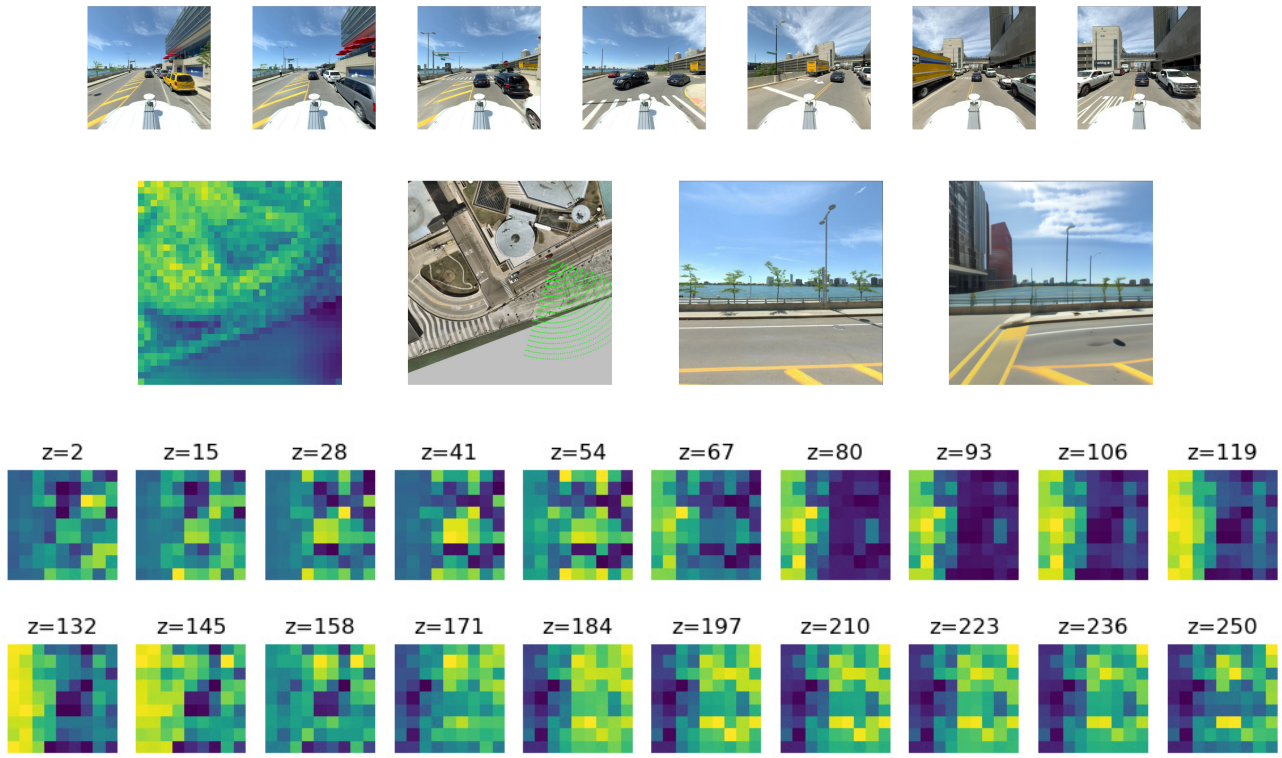


Figure IX. Visualization of Satellite attention. Top row: input images; second row: satellite features, satellite ray samples aligned with target image, target image and generated image; third row: attention maps according to the Z value along the ray.



Figure X. Top row: some input views from our 3-view DL3DV split. Bottom row: a test view, showing the ground truth, ours, and LongLRM.



Figure XI. **Veo 3.1 qualitative results** on the Metropolis dataset with 3 input frames. We generate 2 separate videos with Veo to use the 3 conditioning frames, but GroundControl generates the target views in a single pass. Veo produces highly photorealistic videos, since its model size and training data are larger by orders of magnitude. Unlike our model, it can also handle dynamic objects (including the hood of the car the camera is mounted on). However, its camera trajectory may not necessarily follow the required target poses. It can also frequently hallucinate scene content in incorrect locations (as highlighted in the image).