Cross-View Meets Diffusion:

Aerial Image Synthesis with Geometry and Text Guidance

Ahmad Arrabi[1,*], Xiaohan Zhang[1*], Waqas Sultani[3], Chen Chen[4], Safwan Wshah[1,2†]

[1]Department of Computer Science, University of Vermont, Burlington, USA
[2]Vermont Complex Systems Center, University of Vermont, Burlington, USA
[3]Intelligent Machine Lab, Information Technology University, Pakistan
[4]Center for Research in Computer Vision, University of Central Florida

WACV 2025 TUCSON, ARIZONA • FEB 28 - MAR 4

Scan me for code and weights

Motivations

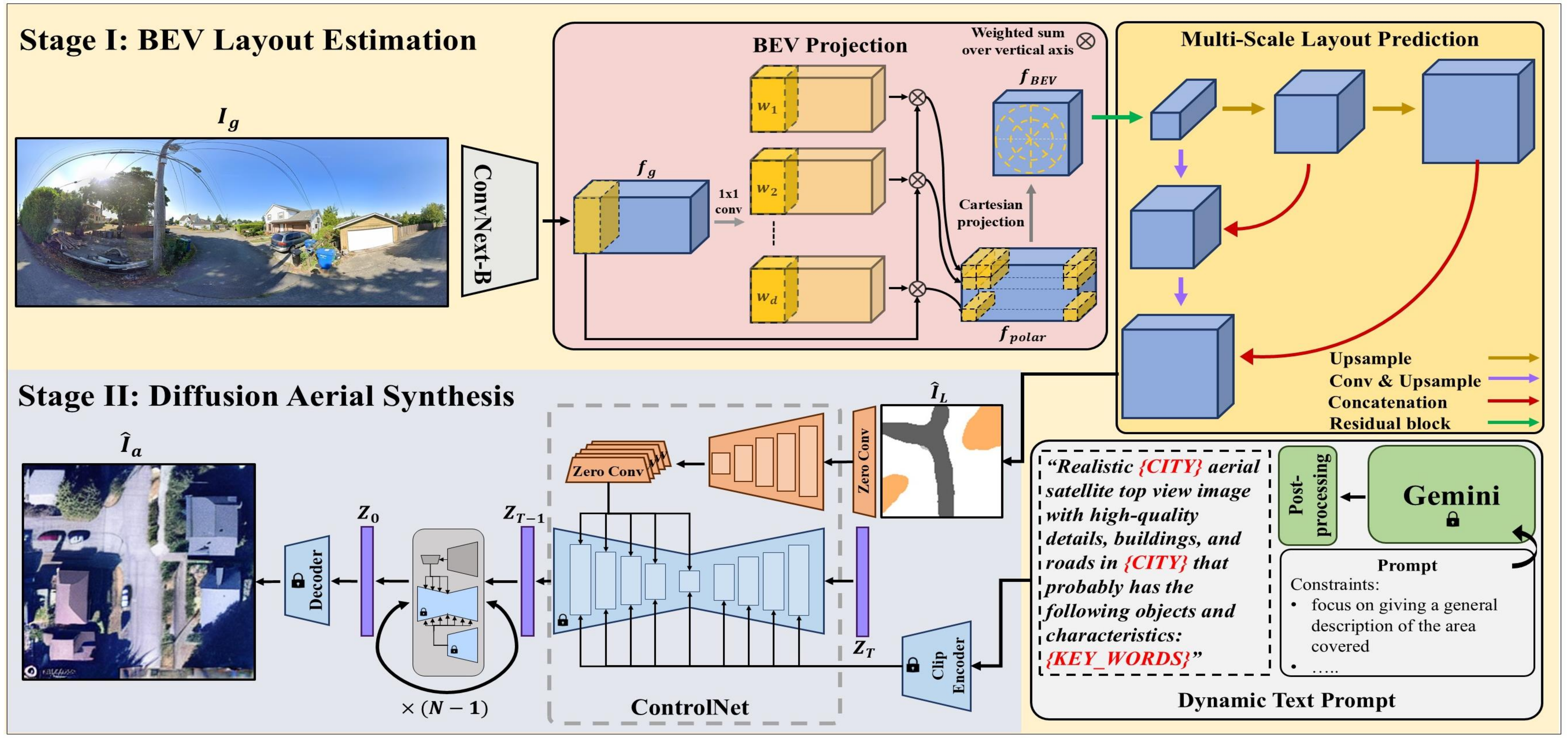
- Aerial images provide high-resolution, detailed views but are costly and effort-intensive to capture, often relying on UAVs or drones
- In contrast, ground images are abundant, cost-effective, and readily available through autonomous vehicles and crowdsourcing platforms
- Ground-to-aerial (G2A) image synthesis offers a promising, cost-effective solution by generating aerial images from corresponding ground views

"The image shows an urban street intersection with several buildings on each side. The buildings are mostly commercial and residential with a few trees and cars along the street. There is a German car repair shop on the right side of the intersection. The street is paved and has a bike lane. There are a few people walking on the street."

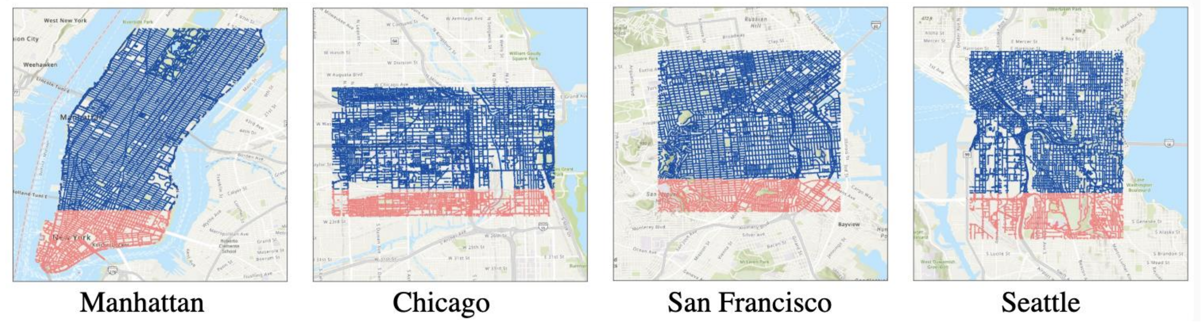
One sample from our VIGORv2 dataset. Top left is the aerial image, top middle is the street-view image, top right is the layout map, and bottom is the text description

VIGORv2

- **Expanded Modality**: VIGORv2 extends the original VIGOR dataset by adding 105,214 **center-aligned aerial-ground** image pairs, **BEV layout maps**, and **text descriptions** of ground images
- **Geographical Splits**: We Introduce geographically non-overlapping training and testing splits to prevent data leakage



Proposed GPG2A architecture



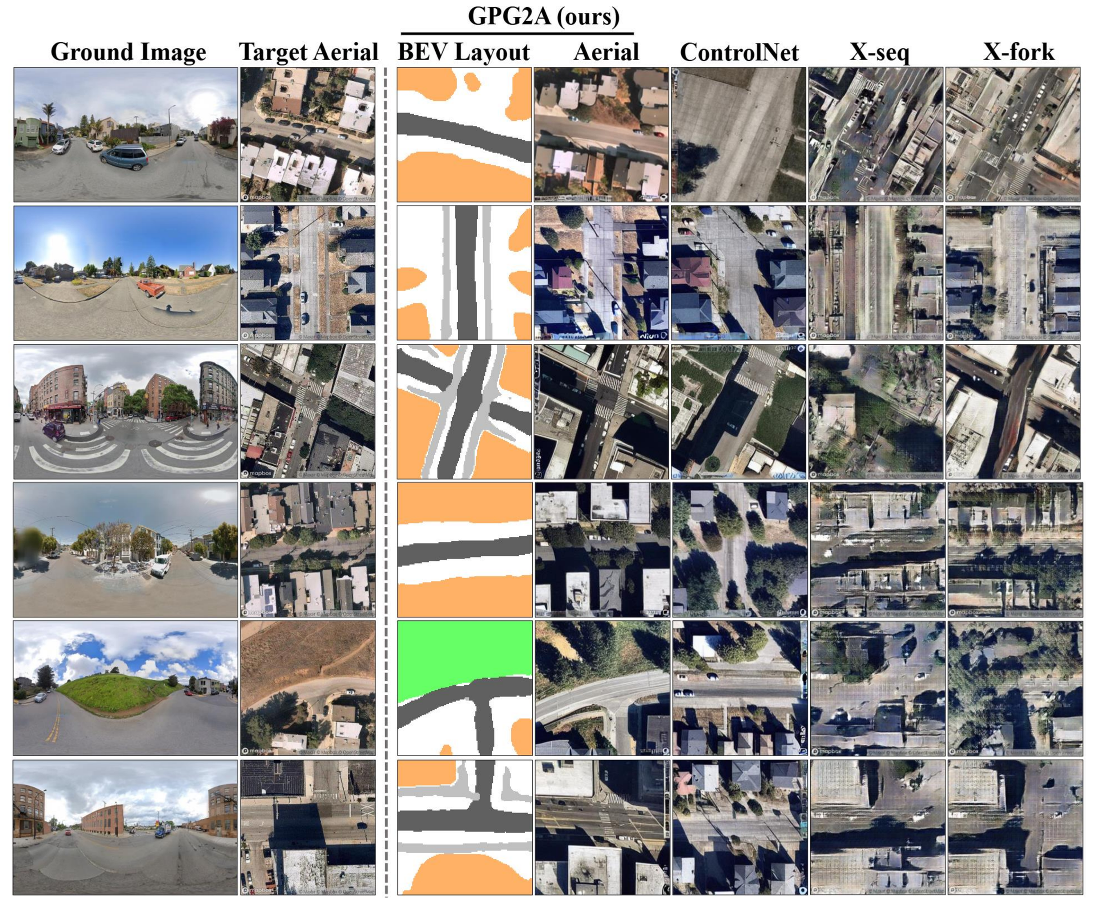
Geographically split of the training (blue) and testing (red) data of VIGORv2

$$Sim_s = \frac{1}{N}\sum_{i=1}^{N}\frac{2-2\times(f^a\cdot\hat{f}^a)}{4},$$

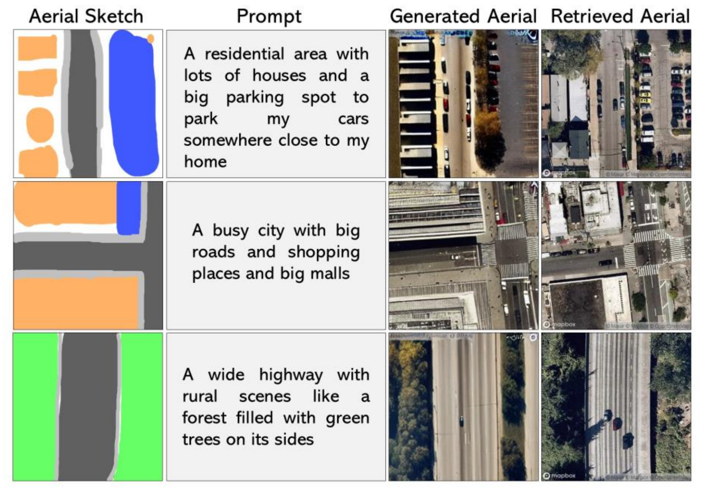
$$\mathrm{FID}_{\mathrm{SAFA}} = ||\mu^a - \hat{\mu}^a|| + Tr(\Sigma^a + \hat{\Sigma}^a - 2(\Sigma^a\hat{\Sigma}^a)^{\frac{1}{2}})$$

Method	Same-area			Cross-area		
	$Sim_s \downarrow$	$Sim_c \downarrow$	$\mathrm{FID}_{\mathrm{SAFA}} \downarrow$	$Sim_s \downarrow$	$Sim_c \downarrow$	$\mathrm{FID}_{\mathrm{SAFA}} \downarrow$
X-seq	0.392	0.438	0.411	0.392	0.454	0.570
X-fork	0.341	0.423	0.151	0.372	0.445	0.357
ControlNet†	0.435	0.415	0.154	0.446	0.405	0.386
ControlNet‡	0.369	0.412	0.110	0.409	0.420	0.220
GPG2A (ours)	0.295	0.402	0.079	0.333	0.392	0.197

Benchmarking with vanilla ControlNet, X-fork, and X-seq.



This is a qualitative comparison of same-area (top 3 rows) and cross-area (bottom 3 rows) images



Sketch-based region search application results

$$\hat{x} = \begin{cases} \lambda\, x_{fake} + (1-\lambda)\, x_{real} & p \le p_\circ \\ x_{real} & p > p_\circ, \end{cases}$$

	$p_\circ$	R@1↑	R@5↑	R@10↑	R@1%↑
same-area	0	83.58%	93.72%	96.08%	99.27%
	0.4	86.33%	94.76%	96.27%	**99.38%**
	0.6	**86.84%**	**94.98%**	96.30%	99.34%
	0.8	85.77%	94.71%	**96.44%**	99.30%
cross-area	0	50.03%	70.17%	77.70%	94.03%
	0.4	51.55%	71.96%	78.18%	94.10%
	0.6	**52.84%**	**72.32%**	**78.45%**	**94.19%**
	0.8	50.11%	70.98%	77.60%	93.99%

Data augmentation on SAFA by using our GPG2A synthesized images

Model Overview

Why Two stages?

- The problem is simplified! reducing the domain gap between aerial and ground views
- The BEV layout map explicitly preserves geometry correspondence between the views
- Leverage strong pre-trained diffusion foundation models (stage II).

Why add text?

To further improve the synthesis quality and fuse surrounding information not fully represented in the BEV layout map.

FOV	BEV Accuracy		Synthesis Quality		
	Avg F1	mIoU	$Sim_s \downarrow$	$Sim_c \downarrow$	$\mathrm{FID}_{\mathrm{SAFA}} \downarrow$
90°	0.259	0.149	0.413	0.414	0.290
180°	0.411	0.258	0.385	0.406	0.181
270°	0.458	0.297	0.369	0.404	0.143
360°	0.565	0.394	0.295	0.402	0.079

Ablation study on input ground image with variant field-of-view (FOV).

Applications

1. **Sketch-based Region Search**

Sketch: Draw what you have in mind

Describe: Add text about the area

Discover: GPG2A synthesizes an image, and we find the closest match from a database

2. **Data Augmentation for Cross-view Geo-localization**: Leveraging the synthesized aerial images from GPG2A to augment cross-view geo-localization training

* These authors contribute equally   † Senior and corresponding author