Appendix / Supplemental Material

A Open Source Dataset and Experiments

To advance research in real-world open-vocabulary and domain-generalized semantic segmentation, we introduce **AustinScapes**, a new large-scale street-view dataset collected from the city of Austin, Texas, USA. The dataset comprises 100,000 high-resolution street-level images, each accompanied by pixel-level semantic annotations. The annotation schema follows the Cityscapes 19-class labels, ensuring compatibility with existing benchmarks.

AustinScapes captures a wide variety of urban scenes under diverse lighting, architectural styles, and traffic conditions, making it well-suited for evaluating both fine-grained recognition and robustness across real-world complexities. We benchmark our method and previous state-of-the-art approaches on AustinScapes to assess generalization performance. Figure 8 shows the visualization results of DGSS and OVSS methods on the AustinScapes dataset. Experimental results show that our framework maintains superior segmentation quality, particularly in challenging object categories and under domain shifts, further validating its effectiveness on unseen urban domains.

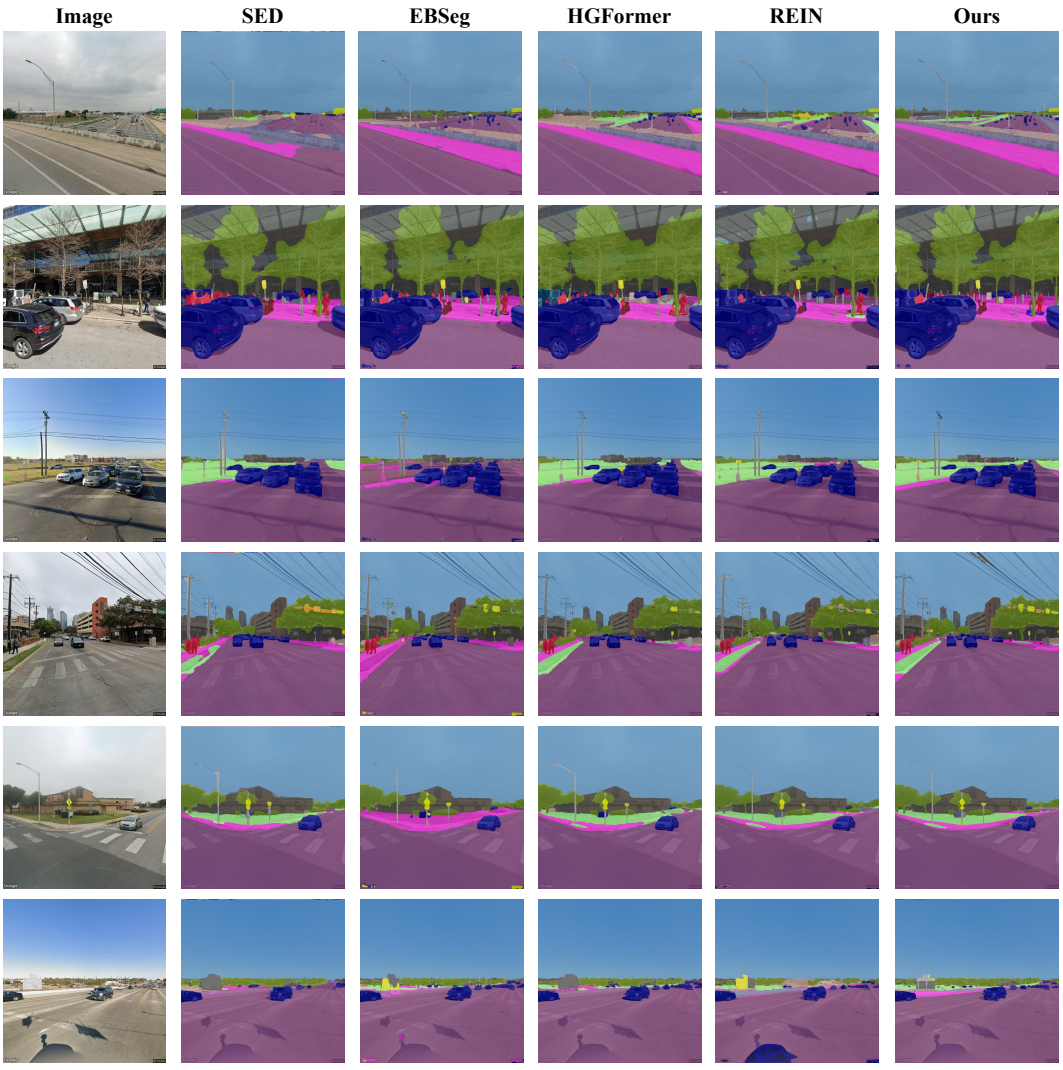

Figure 8: Key zero-shot segmentation examples from existing DGSS and OVSS methods, and our Vireo, on the AustinScapes dataset.

B Related Works

This section provides a more comprehensive discussion of the research areas that are highly relevant to our work: Open-Vocabulary Semantic Segmentation (OVSS) and Domain-Generalized Semantic Segmentation (DGSS). By examining these areas, we aim to highlight our contributions in bridging the gap between DGSS and OVSS.

B.1 Open-Vocabulary Semantic Segmentation

Existing OVSS methods [23, 12, 29, 4] primarily leverage the rich semantic knowledge embedded in pre-trained vision-language models (VLMs) such as CLIP [53]. Early approaches often adopt two-stage pipelines. For example, ZSSeg [54] classifies class-agnostic masks using CLIP, and MaskCLIP [22] adapts CLIP for instance mask proposal classification. Towards end-to-end solutions, LSeg [21] generates pixel-level embeddings aligned with text features, and OVSeg [1] further enhances this by learning discriminative pixel embeddings using a class-agnostic decoder. OpenSeg [55] presents a unified, language-driven framework for open-vocabulary segmentation.

To reduce dependency on region proposals, FreeSeg [3] introduces a grouping-based strategy to cluster pixels. Generalized decoders such as X-Decoder [56] enable language-conditioned pixel-level predictions. GroupViT [57] explores hierarchical region grouping with textual concepts. SegCLIP [13], a task-specific adaptation, enhances spatial understanding from CLIP for dense prediction tasks. More recently, OpenSeeD [58] aims to unify multiple segmentation tasks under the open-vocabulary setting via a conditioned decoder, and Cat-seg [12] improves multimodal matching by jointly modeling image and text in the cost function. Additionally, [59] addresses evaluation issues by proposing a mask-matching-based framework to handle semantic ambiguity.

Despite these advances, challenges remain in transferring coarse-grained VLM knowledge to pixel-level predictions and in resolving fine-grained linguistic ambiguities. Our work builds upon this line of research by incorporating depth cues to enhance language-guided segmentation, particularly in disambiguating objects within complex scenes.

B.2 Domain-Generalized Semantic Segmentation

Previous methods primarily focus on learning feature representations that are both discriminative for semantic segmentation and invariant to domain-specific variations. A foundational strategy in DGSS is introduced by DANN [26], which employs adversarial training with a domain classifier. Li et al. [60] investigate generative approaches using GANs to model the joint distribution of images and labels, while RobustNet [27] mitigates domain-specific biases through instance-selective whitening. SAN-SAW [15] introduces semantic-aware normalization using SAN and SAW modules, and BlindNet [61] enhances robustness by disentangling style and content representations.

More recently, pre-trained VFMs have been effectively leveraged in DGSS frameworks [10, 8, 62, 6]. These methods utilize frozen VFM features as universal backbones and optimize learnable tokens to improve cross-domain performance.

Building upon these advancements, our work posits that integrating the semantic generality of language with the geometric stability of depth provides a compelling approach to achieving robust domain-generalized open-vocabulary segmentation.

C Expend Experiments and Analysis

In this section, we present per-class method comparisons, model details, and visual results that are excluded from the main manuscript owing to page constraints.

C.1 More Data Analysis and Model Details

From Table 5, our approach achieves a mIoU of 66.73%, significantly outperforming other OVSS approaches. Compared to the second-best SED (48.16% mIoU), this represents an improvement of 18.57 percentage points, demonstrating a substantial enhancement in cross-domain generalization. Moreover, our approach attains the best performance in 18 categories.

Table 5: Performance comparison across OVSS methods on per-class IoU and mIoU (City.→BDD.). Top three results are highlighted as best , second , and third , respectively(%).

Method	road	side.	build.	wall	fence	pole	light	sign	vege	terr.	sky	pers.	rider	car	truck	bus	train	moto.	bicy.	mIoU
EBseg	89.72	42.47	81.24	25.11	31.24	45.95	40.14	44.32	80.57	17.67	93.20	50.52	24.96	83.97	35.60	62.83	0.00	39.60	40.02	40.59
FC-CLIP	93.63	57.06	79.53	18.63	38.78	48.86	24.17	48.45	83.89	45.27	83.26	69.16	55.44	88.55	44.13	75.35	0.00	66.74	**60.34**	51.41
CAT-Seg	92.08	50.94	79.41	29.08	38.85	13.88	25.88	28.88	79.26	44.20	91.11	45.51	30.07	82.74	40.65	64.85	0.00	43.23	36.18	44.28
SED	93.78	60.35	82.90	30.07	44.48	39.26	44.97	46.22	83.40	45.55	92.53	60.06	29.23	86.92	43.64	42.94	0.31	43.88	42.12	48.16
Ours	**95.40**	**68.08**	**88.29**	**34.81**	**46.11**	**60.95**	**62.31**	**59.80**	**87.26**	**51.10**	**95.27**	**72.94**	**62.63**	**90.93**	**53.82**	**83.08**	**22.35**	**74.40**	58.37	**66.73**

Table 6: Comparison of per-class IoU across OVSS methods on the DELIVER dataset. Top three results are highlighted as best , second , and third , respectively(%).

Class	**Cloud**					**Fog**					**Night**					**Rain**					**Sun**				
	Ours	SED	CAT-Seg	FC-CLIP	EBseg	Ours	SED	CAT-Seg	FC-CLIP	EBseg	Ours	SED	CAT-Seg	FC-CLIP	EBseg	Ours	SED	CAT-Seg	FC-CLIP	EBseg	Ours	SED	CAT-Seg	FC-CLIP	EBseg
Building	73.71	81.43	81.63	53.96	81.64	52.58	64.43	59.10	50.79	62.79	56.24	77.20	74.99	58.13	75.90	57.34	72.25	79.73	60.24	71.22	85.55	83.17	81.47	49.24	84.39
Fence	39.24	18.22	25.89	28.31	29.54	37.52	14.10	19.50	25.53	21.56	21.20	7.07	10.65	12.99	8.94	38.06	11.77	16.88	24.36	20.86	37.10	10.67	23.77	22.65	28.26
Other	0.00	0.00	0.00	0.11	0.04	0.00	0.00	0.00	0.02	0.06	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	1.88	0.03
Pedestrian	68.86	39.43	21.84	4.21	16.64	73.61	36.72	14.02	2.96	36.21	56.44	31.99	9.00	8.78	5.93	66.21	33.65	19.85	3.28	17.14	64.61	30.83	13.25	3.37	23.59
Pole	32.17	19.50	15.62	6.74	31.35	31.91	21.96	17.90	5.82	35.65	24.20	20.46	14.19	6.51	33.20	32.85	26.89	20.69	6.69	34.53	37.35	22.33	19.39	4.20	34.08
RoadLine	0.00	5.87	2.30	4.46	2.95	0.00	5.44	2.28	3.94	1.30	0.00	5.09	1.94	3.59	2.68	0.00	5.63	2.11	4.01	2.38	0.00	4.80	2.09	3.96	3.02
Road	91.70	29.72	89.55	0.00	65.80	91.72	24.25	89.15	0.01	80.96	90.78	26.22	86.28	0.00	64.23	91.65	32.11	88.48	0.01	67.85	91.94	22.69	88.78	0.00	63.06
SideWalk	71.50	57.22	57.36	59.27	59.04	66.91	53.90	55.25	47.48	43.27	69.18	56.95	58.65	54.13	38.89	69.50	56.40	56.95	52.37	45.44	71.16	48.96	55.14	44.70	27.36
Vegetation	69.36	61.11	62.12	29.30	39.67	68.58	58.86	58.96	33.79	35.20	38.46	40.39	31.46	24.22	20.33	66.69	62.29	62.37	28.17	31.89	68.85	58.12	59.30	54.17	68.62
Cars	63.84	60.20	57.84	66.49	65.73	68.99	60.90	58.78	72.55	68.84	77.26	66.19	60.66	74.09	67.51	64.80	59.94	53.38	62.80	60.64	70.38	65.32	59.65	66.64	64.05
Wall	3.46	0.08	2.05	2.67	4.14	0.30	0.00	0.00	0.08	0.43	2.11	1.46	0.00	1.47	1.53	0.47	0.41	0.01	0.49	0.23	4.07	5.95	0.00	0.29	3.80
TrafficSign	22.00	26.99	25.02	30.59	23.55	32.28	36.64	24.98	10.11	25.88	20.83	35.63	33.87	12.18	26.76	13.01	28.84	26.61	14.26	20.98	32.60	41.36	42.29	29.98	32.87
Sky	86.90	92.22	94.09	26.28	59.90	67.51	82.47	82.87	18.12	41.63	43.22	79.51	56.92	21.98	7.81	65.98	77.80	92.65	11.00	13.61	96.37	95.20	95.22	23.70	97.22
Ground	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Bridge	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.00	0.04	0.13	0.00	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.29	0.00	0.00
RailTrack	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.21	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00
GroundRail	0.00	12.80	0.00	15.07	0.00	0.00	8.32	0.03	11.33	0.00	0.00	11.20	0.00	6.31	0.00	0.00	6.93	0.00	8.85	0.00	0.00	6.72	0.00	11.54	0.00
TrafficLight	57.73	25.71	23.39	24.94	13.05	67.39	41.82	29.52	10.24	21.85	51.48	30.92	21.47	2.01	4.45	58.03	23.86	22.72	4.60	10.50	61.49	39.72	25.74	3.17	17.78
Static	0.00	0.07	0.00	0.12	0.00	0.00	0.01	0.01	0.00	0.22	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.04	0.02	0.00
Dynamic	0.00	0.00	0.00	0.01	0.36	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.53	0.00	0.00	0.00	0.01	0.18	0.00	0.00	0.00	0.00	0.78
Water	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.68	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Terrain	48.31	32.42	38.57	15.22	4.35	45.35	27.11	35.20	17.08	2.42	37.71	13.06	28.41	4.41	0.45	47.53	25.24	33.96	17.23	9.46	43.44	25.18	31.03	15.49	5.16
TwoWheeler	51.66	37.99	16.31	20.55	12.10	60.03	41.90	22.45	29.29	11.42	48.04	31.61	14.29	25.50	13.74	68.66	52.24	41.47	34.45	41.11	60.42	47.01	38.37	31.28	40.16
Bus	0.00	0.01	0.00	3.76	0.00	0.00	6.11	0.00	4.45	0.00	0.00	0.93	0.00	0.00	0.00	0.00	0.00	0.00	23.64	12.30	0.00	0.00	0.00	1.73	0.05
Truck	28.00	9.22	41.70	45.42	55.62	53.19	43.36	49.80	68.80	60.14	50.18	33.77	10.99	50.29	14.40	54.24	45.11	45.65	52.43	47.98	67.95	64.92	69.59	55.30	66.04
Mean	**33.89**	24.40	26.22	17.50	22.62	**35.80**	25.25	24.80	17.26	22.00	**29.89**	22.79	20.56	14.93	15.50	**33.46**	25.18	26.53	16.59	20.35	**38.49**	27.14	28.21	16.93	26.41

For small targets prone to missed detections under extreme conditions (such as traffic lights and pedestrians), our method achieves 62.31% and 72.94% (17.34% and 12.88% higher than SED), respectively. The results reflect the robustness of our method in complex scenarios.

As shown in Table 6, on the DELIVER dataset, our method achieves the highest mIoU across all five extreme weather conditions, outperforming the runner-up by an average of 8–11 percentage points. Under Sun and Cloud conditions, large classes achieve over 80% IoU. In more adverse conditions, our method still maintains IoUs above 30%. Even in the most challenging Night scenario, we outperform others by a margin of 7.10%. For fine-grained categories such as TwoWheeler, Pedestrian, TrafficSign, and TrafficLight, our method consistently outperforms competitors, demonstrating strong recognition of low-contrast and small-scale objects. Although rare categories exhibit lower absolute IoUs, we still rank among the top in most of these cases. Overall, our approach demonstrates superior performance across both frequent and challenging classes under diverse weather conditions.

Table 7: Efficiency and Performance Comparison

Method	**Iter.**	**Train (4 BS)**	**Mem.**	**FPS**	**Infer.**	**Perform.**
SED	40k	6h59min	20.8G	1.73	33min	57.0%
CAT-Seg	40k	3h3min	9.7G	1.66	35min	48.9%
FC-CLIP	40k	5h34min	18.7G	1.45	40min	58.7%
EBSeg	40k	6h34min	22.1G	0.62	94min	49.7%
REIN	40k	12h6min	13.8G	1.88	31min	69.4%
Vireo	40k	15h2min	18.4G	1.36	43min	73.0%

Table 7 compares training time, memory usage, FPS, inference time, and final mIoU (averaged across ACDC, BDD100K, and Mapillary) for various OVSS methods and the baseline, all evaluated under a unified 40k-iteration schedule with a batch size of 8. While Vireo's training time (15 hours 2 minutes) and inference time (43 minutes) are not the shortest, it achieves the highest mIoU of 73.0%, utilizing 18.4 GB of memory at 1.36 FPS, substantially outperforming all other approaches.

Importantly, we follow the previous configuration of intermediate DINOv2 feature layers ($\{\hat{f}_{l_i}^V\}_{i=1}^4, l_i \in \{8, 12, 16, 24\}$), which has been demonstrated in prior work to offer an optimal

trade-off between accuracy and efficiency. This ensures a fair and consistent comparison across all methods in our experiments.

C.2 More Visualization Results

Figure 9 presents a visual comparison of segmentation results between DGSS methods and ours on the *Cityscapes* $\rightarrow$ *ACDC* target domain. The first row shows the original outputs, while the second and third rows display zoomed-in crops for closer inspection. In the original images, other models clearly miss regions on the left sidewalk. In the second-row crops, our method yields much finer and more complete boundaries around small objects such as traffic lights and signs, detecting nearly every instance. In the third-row crops, even distant vehicles and signals are accurately segmented. Moreover, Vireo maintains continuous, clean segmentation of roads and sidewalks in low-contrast conditions (nighttime, rain, snow), without breaks or noise artifacts. It preserves highly consistent segmentation across large background areas like sky, buildings, and vegetation, suppressing "speckle" noise. It precisely captures very fine structures such as curbs and road markings. And it produces coherent, complete contours for dynamic objects (pedestrians, cyclists, vehicles). Whether in clear daytime or foggy, rainy nighttime scenarios, Vireo demonstrates superior cross-scenario robustness.

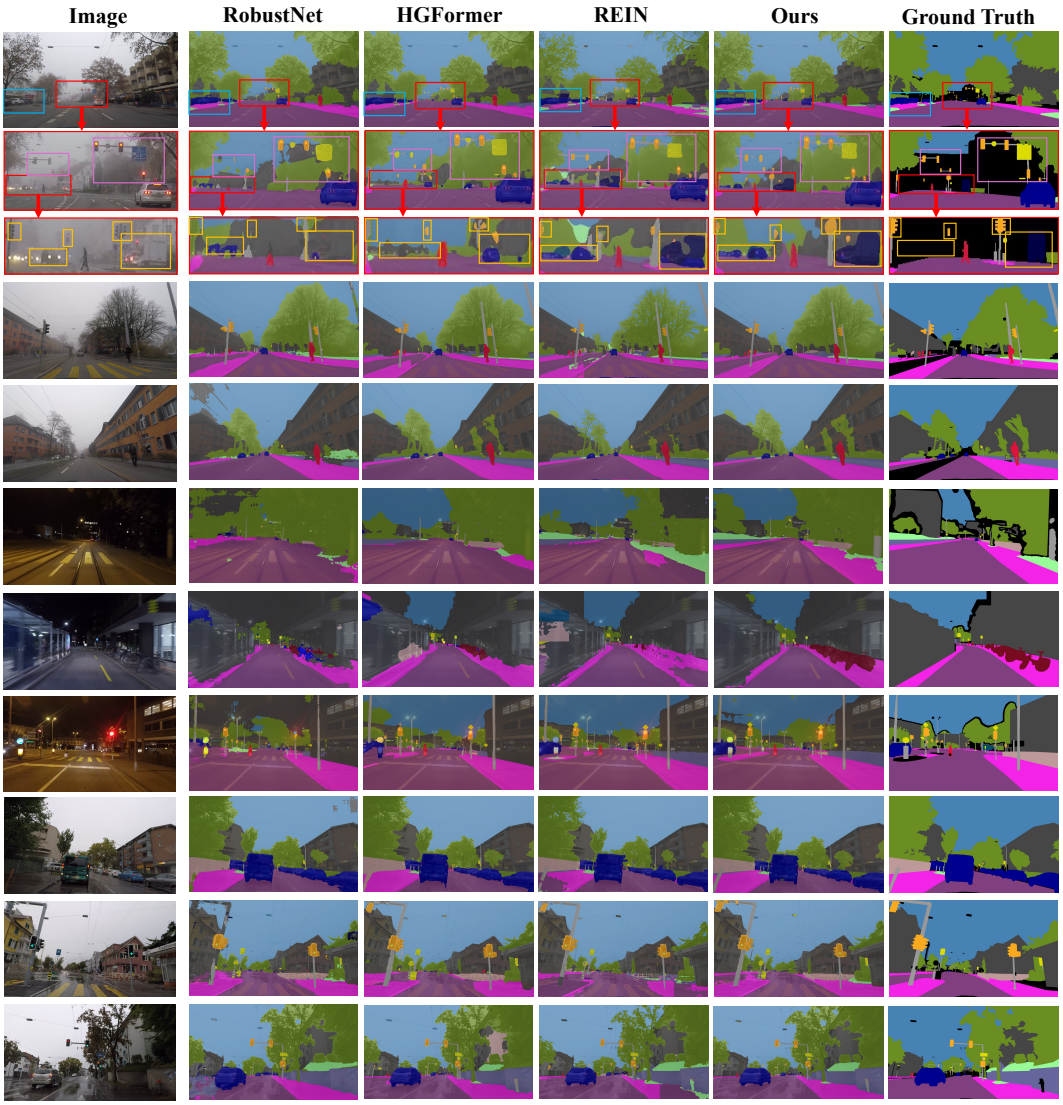

Figure 9: Key segmentation examples of existing DGSS methods and Vireo under the Citys. $\rightarrow$ ACDC unseen target domains under Night, Snow, Rain, and Fog conditions.

Figure 10 presents a visual comparison of segmentation results between DGSS methods and ours on the *GTA5* $\rightarrow$ *Cityscapes* + *BDD* + *Mapillary* target domains. Across these varying scene distributions, RobustNet, HGFormer, and REIN often misclassify or omit road markings, crosswalks, clusters of pedestrians, and parked vehicles. In contrast, our Vireo consistently preserves the complete contours of dynamic objects and slender structures (e.g., lampposts, traffic signals). It segments curbs and lane lines continuously without breaks and delineates building–sky boundaries more sharply. It renders vegetation and other background regions with uniform, noise-free color blocks. Even for distant groups of pedestrians or in highly cluttered street scenes, Vireo accurately separates adjacent classes. This further demonstrates its superior generalization and robustness to unseen domains.

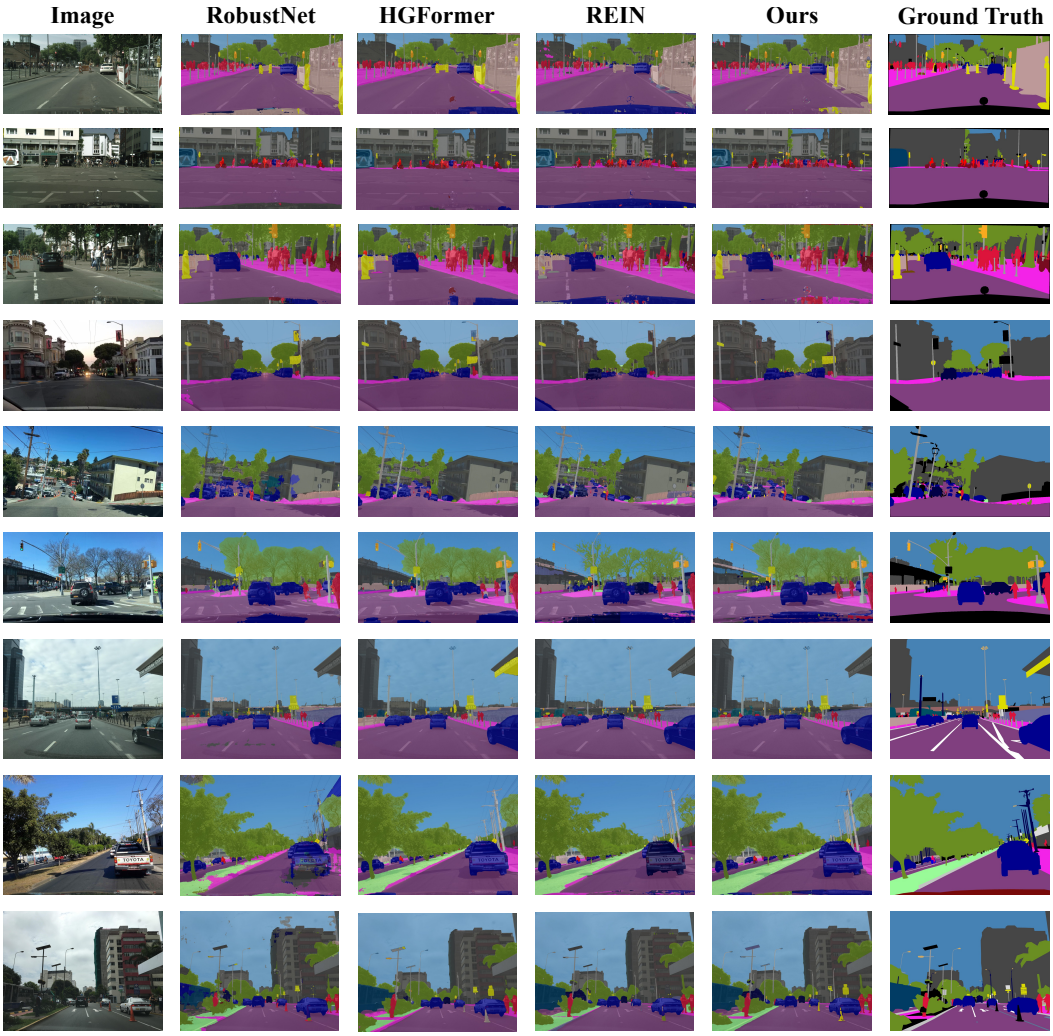

Figure 10: Key segmentation examples of existing DGSS methods and Vireo under the GTA5→City., BDD. and Map. unseen target domains.

Figure 11 presents a comparison of segmentation results between various OVSS methods and Vireo on the *Cityscapes* $\rightarrow$ *ACDC* target domain. The first column shows the original input images, followed by CAT-Seg, EBSeg, FC-CLIP, SED, Vireo, and finally the ground truth. As observed, CAT-Seg and EBSeg often yield fragmented road and sidewalk segmentations with background noise under extreme conditions such as rain, snow, and nighttime. FC-CLIP partially recovers large objects but still fails to detect small targets like traffic lights and thin poles. SED shows some improvement in road continuity but struggles with distant small objects and background consistency. In contrast, Vireo consistently produces continuous, precise segmentations of roads, lane markings, pedestrians, vehicles, and traffic signals across all weather and lighting conditions. It maintains low-noise, highly uniform background regions. Its outputs most closely match the ground truth, further demonstrating Vireo's superior performance and robustness in open-vocabulary, cross-domain scenarios.

Figure 11: Key segmentation examples of existing OVSS methods and Vireo under the Citys. $\rightarrow$ ACDC unseen target domains under Night, Snow, Rain, and Fog conditions.

Figure 12 compares OVSS methods and our Vireo on the *Cityscapes* $\rightarrow$ *DELIVER*transfer setting, which includes sunny, rainy, cloudy, foggy, and nighttime conditions. FC-CLIP preserves mask edges but suffers from widespread class errors. CAT-Seg and EBSeg yield fragmented road and sidewalk masks with spurious background artifacts under adverse conditions. SED improves overall continuity but still missegments slender structures like poles and signs in low-contrast scenarios. In contrast, Vireo consistently produces clear, coherent masks for zebra crossings, lane markings, sidewalks, pedestrians, vehicles, and poles. It effectively suppresses noise in sky, building, and vegetation regions. The outputs most closely match the ground truth, demonstrating superior robustness and generalization in extreme, unseen domains.

As shown in Figure 13, adding CMPE and GTP transforms the attention from a diffuse pattern into a concentrated focus on critical regions such as roads, vehicles, and pedestrians. In the Affinity Map, bright clusters reveal the model's ability to capture long-range dependencies within the same object, while dark boundaries emphasize clear separation between different semantic regions. Together, these results confirm our method's robust handling of semantic coherence and pixel-level affinities across diverse weather and lighting conditions.

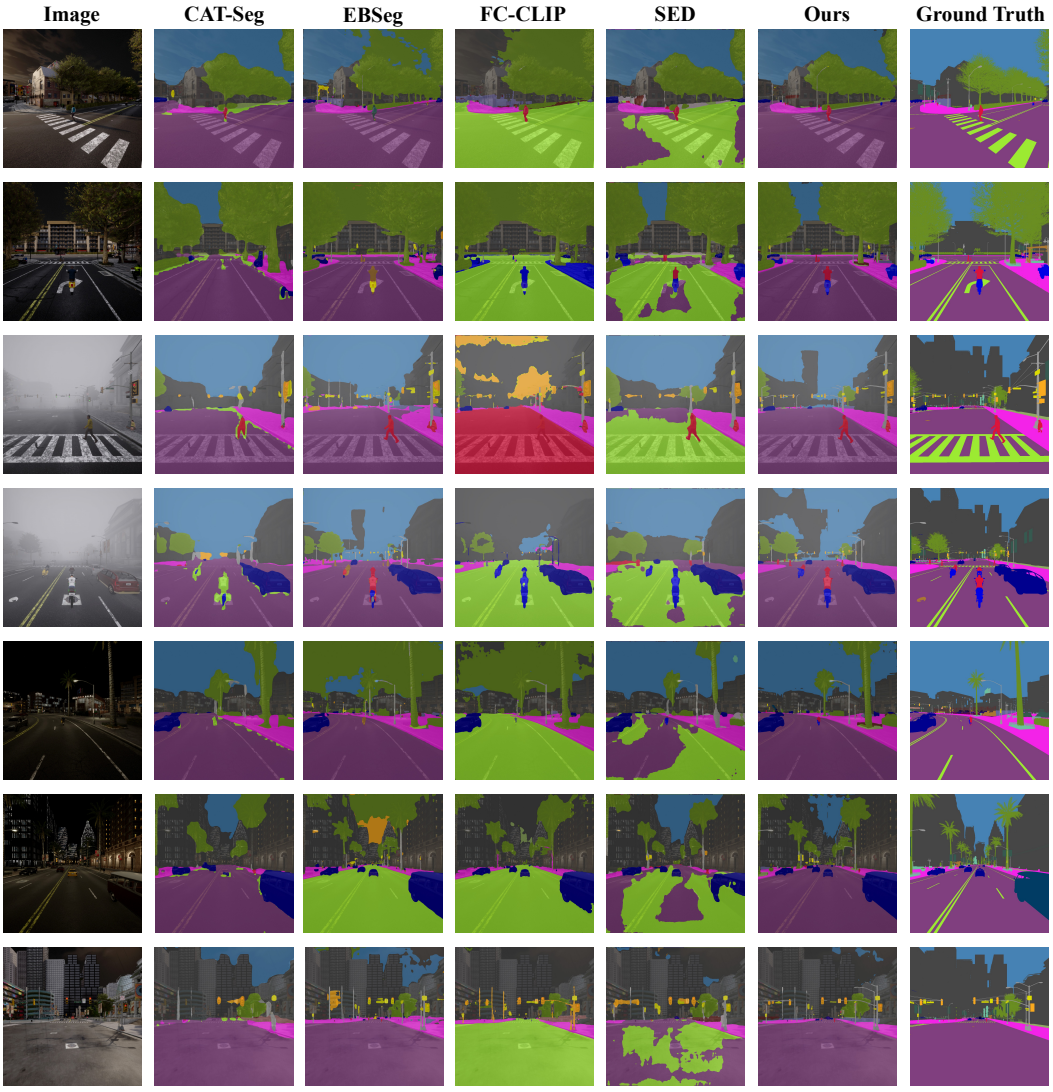

Figure 12: Key segmentation examples of existing OVSS methods and Vireo under the Citys. $\rightarrow$ DELIVER unseen target domains under Cloud, Fog, Night, Rain, and Sun conditions.

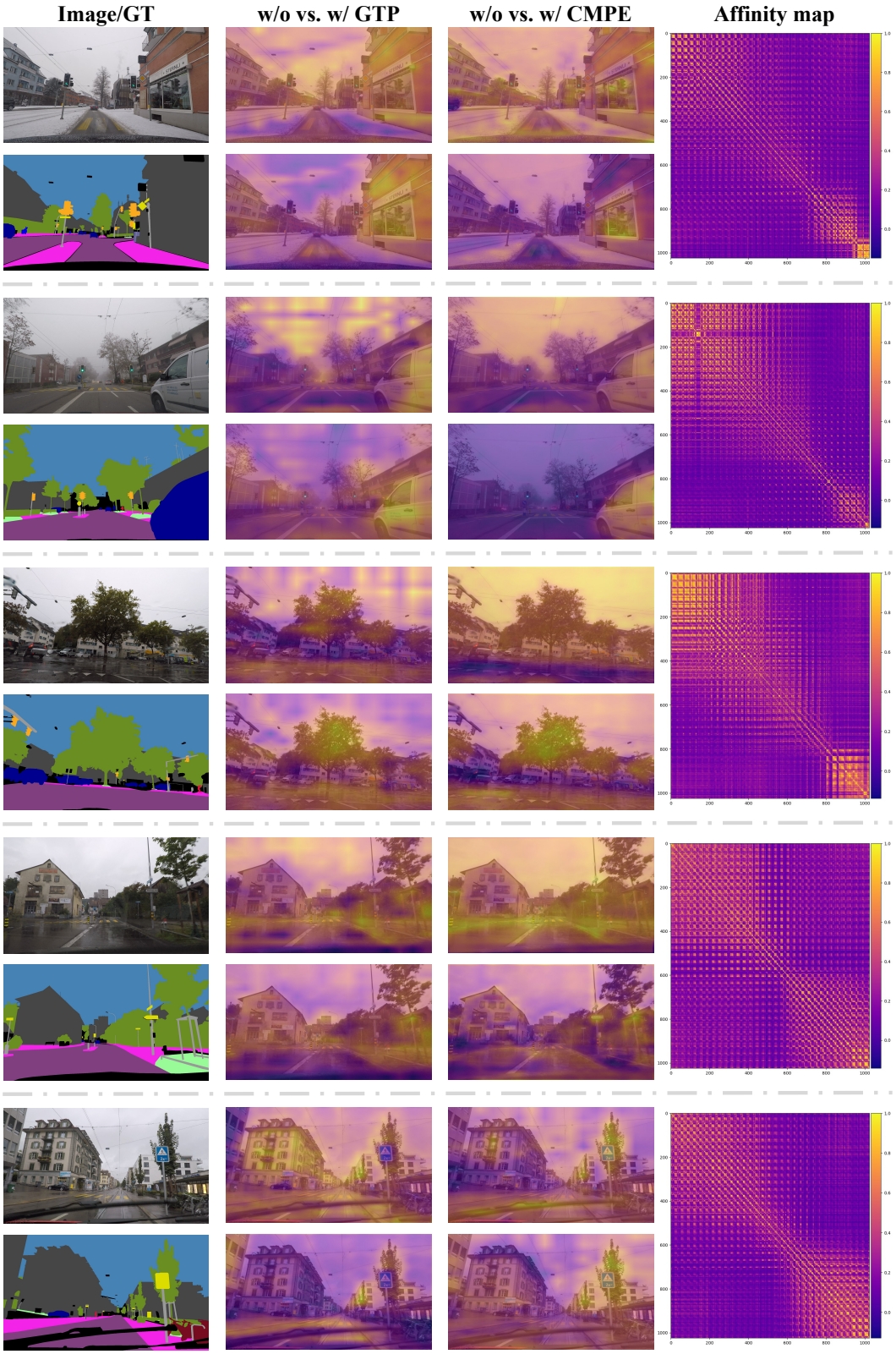

Figure 13: Visualizations of attention maps and affinity maps under different scenes, where CMPE denotes the Coarse Mask Prior Embedding and GTP represents the GeoText Prompts.