Semantic Library Adaptation:
LoRA Retrieval and Fusion for Open-Vocabulary Semantic Segmentation

Supplementary Material

Figure 7. **t-SNE visualization of the LoRA Library and relative datasets in the CLIP embedding space.** Similar domains are clustered together, indicating areas with higher LoRA support and potentially stronger performance improvements. A maximum of 15,000 samples per dataset is used for the t-SNE fit and only 15% of the total datapoints are used for plotting.

Introduction

In this supplementary document, we provide additional details and analyses to support and extend the findings presented in the main paper. Here we report additional details about the implementation specifics, auxiliary experimental results, extensive ablation studies, discuss practical considerations for real-world deployment, and offer additional qualitative examples that highlight the effectiveness and robustness of our proposed method, SemLA. The rest of this document is organized as follows:

- **Section A** details the implementation aspects of our method, including training procedures, hyperparameters, model architecture modifications, and code availability for replication purposes.
- **Section B** provides an in-depth analysis of the LoRA adapter library, including visualizations of the embeddings from training samples using t-SNE, labeling of the adapters with natural language using BLIP-2 [33] for the sake of interpretability, analyses of adapter contributions dataset by dataset, and support score analysis.
- **Section C** presents extensive ablation studies and performance analyses. We explore the effectiveness of fully fine-tuned models versus LoRA adapters, conduct hyperparameter sensitivity analysis, and assess alternative domain navigators such as DINOv2 [48] versus CLIP.
- **Section D** showcases additional qualitative results across various domains, further demonstrating the adaptability and efficacy of our approach.
- **Section E** discusses pragmatic considerations for real-world deployment of SemLA, including strategies to handle computational overhead, domain navigation in specialized domains, scalability concerns, and methods to ensure efficiency and reliability in production environments.

A. Implementation Details

LoRA Training Details. We attach LoRAs to every *nn.Linear* layer in the CAT-Seg architecture, except for CLIPs token embedding layer, as this parameter was not trained in the original CAT-Seg implementation either. All LoRAs are trained with the same LoRA configuration – i.e., rank r=8 and α=16. The training hyperparameters are largely the same as CATSeg with minor modifications: compared to the original CAT-Seg implementation, we use a base learning rate of 1e-4, weight decay of 1e-5, and 1000 warm-up iterations. For ACDC and MUSES adapters we use a batch size of 2 and a warm-up factor of 0.01. For BDD and CS we increase the batch size to 4 while keeping other hyperparameters the same. For the remaining datasets, we increased the warm-up factor to 0.1 while keeping other hyperparameters the same. All the adapters were trained until convergence.

Code and Models. Full source code and documentation are available in our project page https://thegoodailab.org/semla.

B. Interpretability of the LoRA Library

B.1. t-SNE Visualization of the LoRA Library

Figure 7 presents a t-SNE [60] visualization of the LoRA adapters' centroids and their associated datasets in the CLIP embedding space. This visualization illustrates the distribution and relationships among different adapters and domains, highlighting similarities between them.

We observe that domains with similar visual characteristics are positioned closely, such as foggy conditions or nighttime scenes. This clustering validates the effectiveness of using CLIP embeddings for adapter selection.

B.2. BLIP for Labeling LoRA Adapters

To highlight the transparency and interpretability of our system, we leverage the connection between the CLIP embedding space and natural language. By processing the centroids of our LoRA adapters with BLIP-2 [33], we obtain natural language captions describing the domain encapsulated in each adapter's training set. This provides semantic insights into the content and characteristics of the training set used for each adapter.

Table 4 reports, for each dataset, the caption provided by BLIP, together with the answers to two simple questions "Where is this?" and "Describe the environment in two words", when processing domain centroids. We can appreciate how both captions and answers are strongly related to the content in each dataset, even though retrieved information remains limited and coarse. Nevertheless, the possibility of extracting natural language captions of the LoRA centroids is an interesting feature, further motivating the use of CLIP as our domain navigator.

B.3. Adapter Contributions

Figure 8 shows the adapter weight distribution for all datasets composing our benchmark involved in the leave-one-out experiments. The parameters used in this experiment are $\tau = 0.01$ and top-$K = 7$. The weights represent the relative contribution of each adapter to the fused model, highlighting their respective roles in the overall composition.

These pie charts provide insights into how different adapters contribute to the final model for specific target domains, demonstrating the effective combination of knowledge from relevant adapters.

B.4. LoRA Support Score Analysis

We analyze the relationship between the LoRA support score and mIoU performance. We define LoRA support score for a test image $\mathbf{x}_t$:

$$\text{Support Score}(\mathbf{x}_t) = \sum_{i \in \mathcal{K}} \frac{w_i}{\|\mathbf{e}_t - \mathbf{c}_i\|_2}, \tag{15}$$

where w_i is the weight assigned to adapter i, $\mathbf{e}_t$ is the CLIP embedding of the test image, $\mathbf{c}_i$ is the centroid of adapter i, and $\mathcal{K}$ is the set of top-K selected adapters.

We compute the support score for a sample of images and plot it against their corresponding mIoU scores. Figure 9 shows that images with higher support scores tend to have higher mIoU, confirming that the LoRA support score is a good predictor of segmentation performance. We also notice that for low values of support

Dataset	Caption:	Question: Where is this? Answer:	Question: Describe The environment in two words? Answer:
bdd	the view from the driver's seat of a car on a street in san francisco, ca, june 2018	the city of los angeles, california	The environment in two words is the environment in which it is located.
idd	road in kolkata, india, photo by person	a street in bangalore, India	city, road, traffic jam
nyu	a view of the kitchen in the house i'm renting in san francisco, ca, in summer 2008	the house i grew up in, in san francisco, california, usa	blue and white
acdc-rain	the rain is coming down hard, but the streets are dry, and the cars are moving along the road	berlin, germany, street view, rain	rain
acdc-fog	a view from the driver's seat of a car on a highway on a foggy morning in kiev, ukraine	the highway in bordeaux, france, on a foggy day in october 2018	foggy, rainy, cloudy, misty
muses-snow-day	the view from the driver's seat of a car on a city street with buildings in sight	the city of berlin, germany, on a rainy winter day	Rainy day in vienna, austria, with trees and buildings
coconutL	person	a small town in the middle of nowhere, nyc, usa	The environment is where the person lives, works, and plays.
acdc-night	street at night in kiev, ukraine, with traffic lights and a car on the road	austria	dark and light, city, traffic
a150	the blue house	a house in the middle of the woods	The environment is where the person lives, works, or plays.
Cityscapes	street view of berlin, germany	berlin, germany, in the year 2014	city, street, road, traffic light
pc59	person	the house of the person in the picture	blue sky, green grass
muses-fog-day	a view from the driver's seat of a car on a rainy day in bordeaux, france	a foggy day in bordeaux, france, driving on the autoroute	foggy, rainy, misty
muses-clear-day	the car driving on the street in the city	berlin, germany, in the year 2040, a virtual reality simulation	city, road, street
acdc-snow	street in krasnodar, russia, april 2019	austria	snow, winter
muses-fog-night	the road at night, with car lights visible	the road in the dark, in the middle of nowhere, at night	dark and light
muses-clear-night	the car on the road at night, with city lights in the background	austria	night, city, traffic, street lights
mv	street view of kuala lumpur, malaysia, with the city's main road visible	australia	urban, city, cityscape
muses-snow-night	a view of the city from a car's windshield at night, with city lights and snow visible	a city in the uk, in wintertime, with a car driving on the road	snow, rain, night, city

Table 4. **Text generation results using BLIP-2 [33].** For the image embedding, the average embedding across all images from each dataset was computed. Then different prompts were given to the model, as presented at the top of the table.

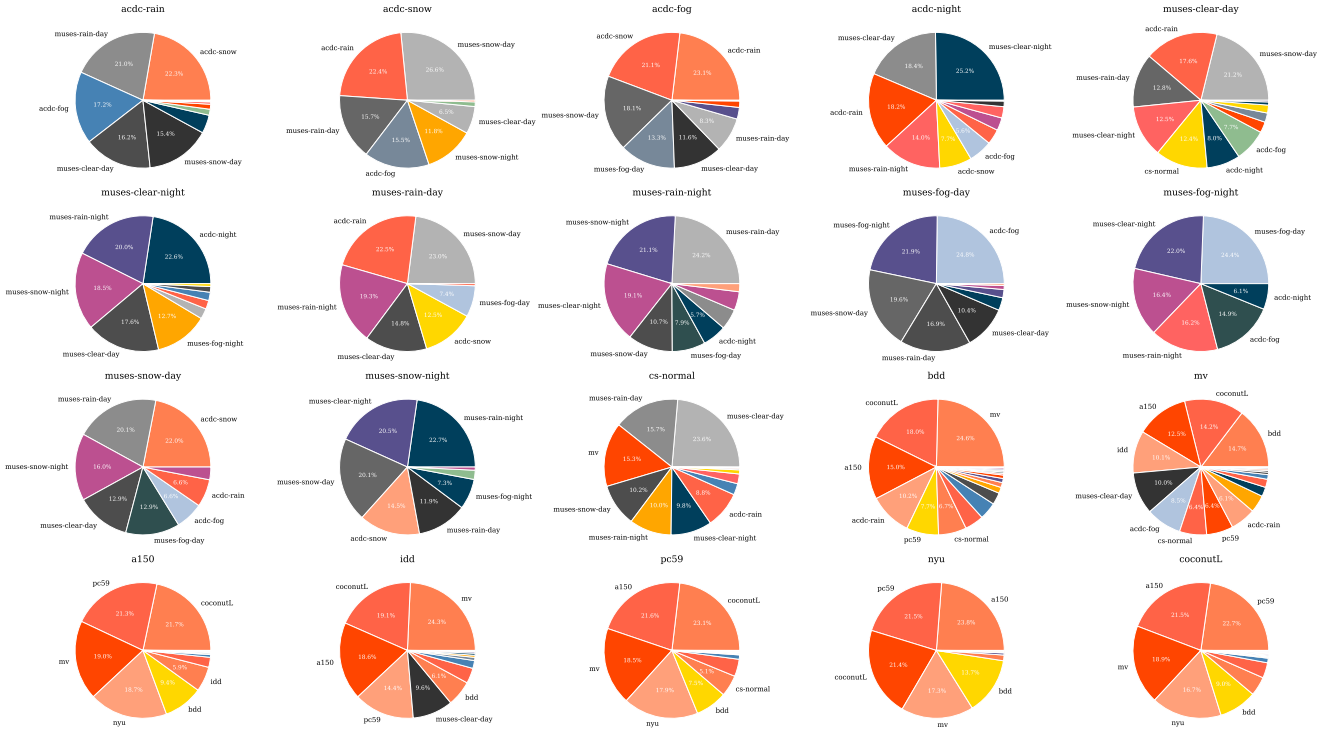


Figure 8. **Adapters weight distribution for each benchmark dataset.** Each pie chart is divided into sections proportional to the average contribution provided by each adapter based on CAT-Seg leave-one-out adaptation settings.

score (*e.g.* below 0.09), an improvement in support score does not strictly imply a stronger improvement in mIoU, showing that the underlying relation is likely not linear.

Overall, this analysis validates our assumption that proximity in the CLIP embedding space, combined with the weighting mechanism, is an effective heuristic for adapter selection.

C. Ablations and Analysis

C.1. Hyper-parameters Study

We conduct ablations over τ and K, the two hyper-parameters controlling our system at test time.

- **Number of Adapters** (K): Increasing K includes more adapters in the fusion, potentially providing more context but risking the introduction of unrelated knowledge while increasing the LoRA merging computational overhead.
- **Temperature** (τ): Regulates the weighting of adapters based on their distances. Lower τ emphasizes closer adapters; higher τ promotes a more uniform weighting.

Figure 10 shows a heatmap of overall performance across different values of K and τ. Performance peaks at $K = 7$ and $\tau = 0.01$, balancing relevance and diversity in adapter selection.

C.2. Distance Metrics Comparison

We compare Euclidean distance (used in SemLA) against alternative distance measures, specifically cosine similarity and Mahalanobis distance. As shown in Table 5, cosine similarity – which would be a natural choice for CLIP embeddings – yields aligned performance with Euclidean distance, which is expected given CLIP embeddings exhibit almost uniform norms, making cosine similarity essentially a monotonic function of Euclidean distance. Conversely, Mahalanobis distance performs worse since covari-

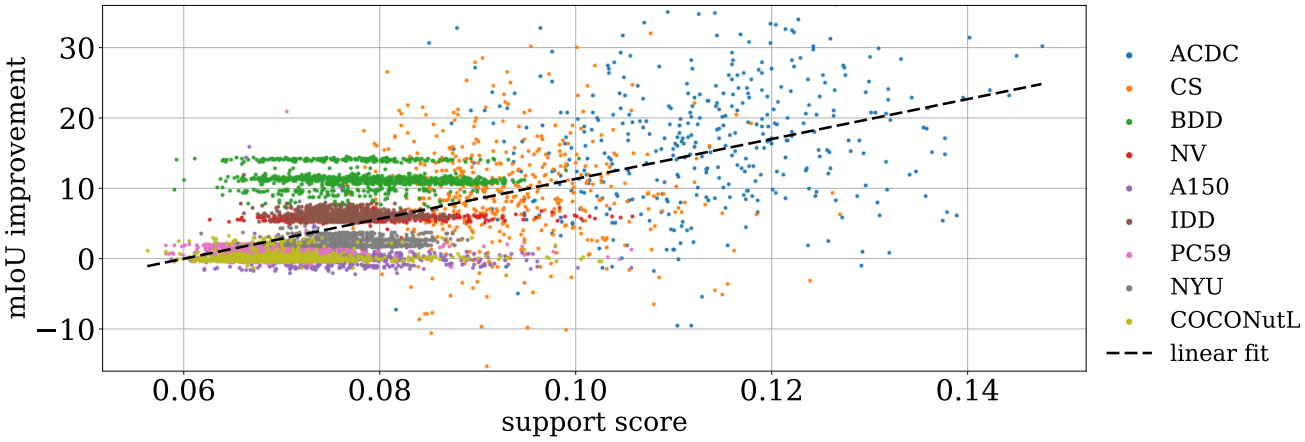


Figure 9. **Correlation between LoRA support score and mIoU.** Higher support scores correlate with better segmentation performance.

Method	ACDC				MUSES								CS	BDD	MV	A150	IDD	PC59	NYU	COCONutL*	h-mean
	rain	snow	fog	night	clear (d)	clear (n)	rain (d)	rain (n)	fog (d)	fog (n)	snow (d)	snow (n)									
Uniform [35]	67.40	66.35	69.71	49.98	58.28	55.78	54.70	45.09	73.75	45.16	61.02	49.08	62.18	58.19	31.51	37.25	38.83	63.06	48.93	(67.62)	51.89
SemLA with **Euclidean** (ours)	67.71	68.95	71.92	51.73	61.09	60.06	57.60	47.35	72.97	52.38	67.28	55.92	63.91	57.30	31.12	38.18	40.16	64.75	51.35	(67.26)	54.16
SemLA with **Cosine**	67.76	68.67	72.52	51.24	61.55	60.22	57.76	47.03	73.03	48.10	66.82	56.67	63.53	57.52	30.24	38.10	39.80	64.65	50.93	(67.31)	53.70
SemLA with **Mahalanobis** †	59.94	63.18	67.70	45.90	56.26	50.54	48.96	36.71	75.74	34.93	56.84	35.89	57.30	56.54	30.03	37.85	39.78	64.43	50.55	(67.69)	47.87

Table 5. **Ablation study – Distance Metrics Comparison (CAT-Seg [10]).** Comparing SemLA with alternative distances. On MUSES, (d) and (n) stand for day and night. ( ) means excluded from h-mean. † For Mahalanobis, source domains where the covariance cannot be computed are excluded from the library. The parameters for Cosine, Mahalanobis, and Late Fusions (τ and K) are tuned independently to achieve the best results with each variant.

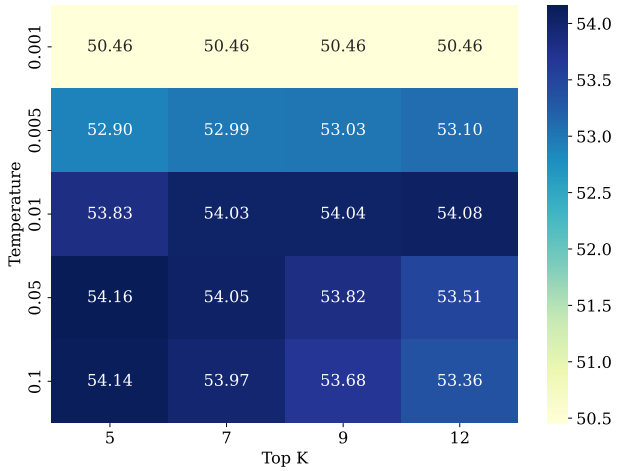


Figure 10. **Hyper-parameters Study.** Impact of K (number of adapters) and τ (temperature) on overall performance (mIoU).

ance estimation becomes numerically unstable for domains with limited samples (fewer than 500 samples), necessitating the exclusion of some adapters and thus degrading performance. Overall, Euclidean distance emerges as the simplest, most robust choice for our method.

C.3. Full Fine-Tuning (FFT)

We explore whether our library could be constructed using fully fine-tuned models instead of LoRA adapters. Table 6 reports the results achieved either by deploying and fusing fully fine-tuned models or LoRA adapters in our library. While aggregating fully fine-tuned models is a known practice to merge different knowledge – as explored in [35] – the results indicate no benefits over our LoRA-based approach. Moreover, storing and merging full models is significantly more computationally expensive than operating with adapters, introducing a sizable overhead at inference time. Full fine-tuning is more prone to overfitting, especially on smaller datasets, whereas LoRA adapters are lightweight and can be trained effectively with limited data. This reinforces our choice of using LoRA adapters, which are modular, efficient, and easily combinable.

C.4. Domain Navigators: DINO vs. CLIP

SemLA uses CLIP [51] to navigate into the LoRA Library and pick the most relevant adapters to combine. However, different visual encoders could serve the same purpose. In Table 7, we test the use of an alternative domain navigator – DINO v2 [48] – and compare the performance achieved by SemLA variants using this latter or CLIP.

On average, the two perform comparably, with CLIP embeddings slightly outperforming DINO ones in guiding adapter selection on average, likely due to their joint text-image embedding space capturing semantic information more effectively. Nonetheless, this experiment proves that SemLA is not bound to use CLIP as the domain navigator, although this latter provides nice properties in terms of explainability – as showcased in Section B.2.

Method	ACDC				MUSES								CS	BDD	MV	A150	IDD	PC59	NYU	COCONutL*	h-mean
	rain	snow	fog	night	clear (d)	clear (n)	rain (d)	rain (n)	fog (d)	fog (n)	snow (d)	snow (n)									
Zero-shot [10]	46.53	48.04	47.09	37.93	44.43	39.29	38.95	27.78	53.73	25.35	43.56	33.29	47.11	47.95	25.69	37.83	35.39	63.33	49.38	(68.26)	39.39
Oracles (LoRA)	70.94	69.22	69.98	51.55	69.36	57.09	54.28	52.11	75.85	61.26	66.25	54.35	67.47	60.06	49.57	53.99	64.34	68.68	61.90	(70.44)	61.05
Oracles (FFT)	70.97	72.02	73.78	53.09	70.49	58.20	55.57	53.18	74.90	62.64	65.83	58.67	70.35	61.03	50.56	52.84	66.38	68.43	64.36	(68.36)	62.38
Uniform [35] (FFT)	69.01	67.91	73.28	51.71	61.44	57.28	54.99	43.13	74.14	36.91	57.81	52.99	62.29	58.05	30.62	36.34	39.73	62.60	48.09	(65.29)	51.43
SemLA (FFT)	69.54	72.07	73.20	52.87	62.78	59.49	57.44	45.59	74.24	53.22	64.54	56.75	65.52	58.28	28.44	31.26	41.33	62.01	46.02	(63.64)	52.79
SemLA (LoRA)	67.71	68.95	71.92	51.73	61.09	60.06	57.60	47.35	72.97	52.38	67.28	55.92	63.91	57.30	31.12	38.18	40.16	64.75	51.35	(67.26)	54.16

Table 6. **Ablation study – Full Fine-Tuning vs LoRA Adaptation (CAT-Seg [10]).** We use full fine-tuned models instead of LoRA adapters and measure the impact on performance over our 20-domain benchmark in leave-one-out setting. On MUSES, (d) and (n) stand for day and night. ( ) means excluded from h-mean. SemLA (LoRA) with $\tau = 0.05$, and top-$K = 5$; SemLA (FFT) with $\tau = 0.01$, and top-$K = 9$.

Method	ACDC				MUSES								CS	BDD	MV	A150	IDD	PC59	NYU	COCONutL*	h-mean
	rain	snow	fog	night	clear (d)	clear (n)	rain (d)	rain (n)	fog (d)	fog (n)	snow (d)	snow (n)									
Zero-shot [10]	46.53	48.04	47.09	37.93	44.43	39.29	38.95	27.78	53.73	25.35	43.56	33.29	47.11	47.95	25.69	37.83	35.39	63.33	49.38	(68.26)	39.39
Oracles	70.94	69.22	69.98	51.55	69.36	57.09	54.28	52.11	75.85	61.26	66.25	54.35	67.47	60.06	49.57	53.99	64.34	68.68	61.90	(70.44)	61.05
Uniform [35]	67.40	66.35	69.71	49.98	58.28	55.78	54.70	45.09	73.75	45.16	61.02	49.08	62.18	58.19	31.51	37.25	38.83	63.06	48.93	(67.62)	51.89
SemLA (DINOv2)	68.40	68.26	73.57	51.18	61.94	59.58	56.06	48.43	73.81	52.60	67.42	56.33	64.04	58.23	31.02	37.45	40.12	64.40	50.52	(67.63)	54.14
SemLA (CLIP)	67.71	68.95	71.92	51.73	61.09	60.06	57.60	47.35	72.97	52.38	67.28	55.92	63.91	57.30	31.12	38.18	40.16	64.75	51.35	(67.26)	54.16

Table 7. **Ablation study – CLIP [51] vs DINOv2 [48] for domain navigation (CAT-Seg [10]).** – We generate the weights for merging the LoRAs based on features extracted from DINOv2 or CLIP, and evaluate the impact on performance over our 20-domain benchmark in a leave-one-out setting. On MUSES, (d) and (n) stand for day and night. ( ) means excluded from h-mean.

D. Additional Qualitative Results

Figure 11 presents additional qualitative segmentation results comparing the zero-shot baseline, uniform merging, and SemLA across different domains. These examples further confirm the effectiveness of our method in adapting to diverse and challenging domains without any additional training being conducted.

E. Discussion: Real-World Deployment

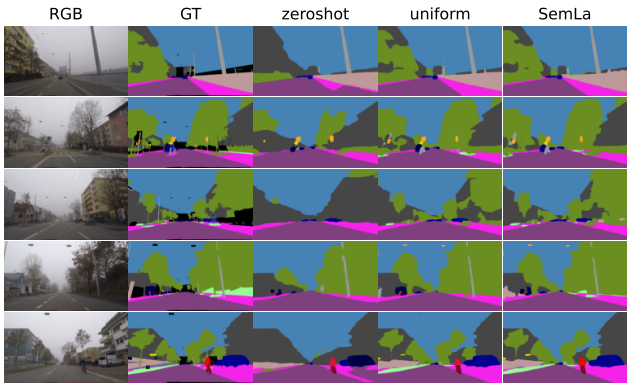
While SemLA demonstrates strong performance in controlled experimental settings, deploying it in real-world applications introduces additional challenges and considerations. In this section, we discuss practical aspects related to the use of CLIP as a domain navigator and propose strategies to address potential limitations.

CLIP as a Domain Navigator for Specific Domains. Although CLIP has shown remarkable generalization capabilities across diverse domains – as evidenced by our extensive 20-domain benchmark – it may struggle in niche or highly specialized domains [67]. When bringing SemLA into production for such specific use cases, it is important to account for this potential limitation. If the target domain is well-scoped, using a fine-tuned domain navigator, with better semantic understanding, might provide better performance.

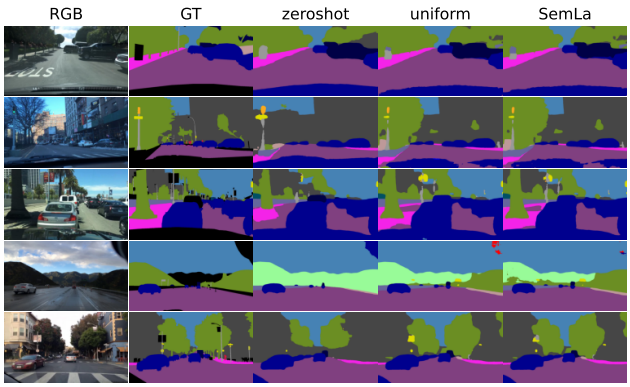
Alternatively, a hierarchical approach could be explored: a general CLIP model can provide a coarse understanding of the domain, identifying then a domain-specific CLIP expert. The expert is then tasked with computing the LoRA distances more precisely.

Efficiency in Production Environments. In production-intensive applications, dynamically loading and unloading dedicated LoRA adapters for each individual input image may be impractical due to computational overhead. While this overhead is significantly lower than the one introduced by retraining the model at test time – as required by most traditional test-time adaptation methods – it is still non-negligible. For applications that do not require real-time processing, such as batch processing of large volumes of images (e.g., processing data accumulated over 24 hours), a practical approach involves pre-computing the CLIP embeddings for all images. The images can then be clustered based on their embeddings, and a batch centroid can guide the fusion of relevant LoRA adapters for the entire batch. This reduces the frequency of adapter loading and improves efficiency by applying the same fused model to similar images. In contrast, real-time applications in the field of robotics and autonomous driving cannot rely on batch processing due to their immediate response requirements. In these cases, we propose implementing a debouncing mechanism that triggers adapter swapping only when there is a significant change in the domain. Specifically, the system can monitor the CLIP embeddings of incoming images—or use an exponential moving average (EMA) of these embeddings—and compare them to the embeddings associated with the currently active adapters. If the embedding distance exceeds a predetermined threshold, indicating that a new domain has been encountered, the system triggers the retrieval and fusion of new adapters. This approach ensures that the model adapts only when necessary, minimizing computational overhead while maintaining adaptability. This strategy is analogous to concepts proposed in domain-adaptive systems like HAMLET [3], where adaptation occurs only upon detecting domain shifts. Furthermore, in a real-world deployment, the prediction process can be presented with an average LoRA distance metric. As shown in our analysis, this metric provides an additional source of confidence estimation by indicating how well the selected adapters align with the target domain. Such a heuristic contributes to the study of model calibration and can be valuable for downstream tasks—effectively informing whether to trust the model's predictions in critical applications.

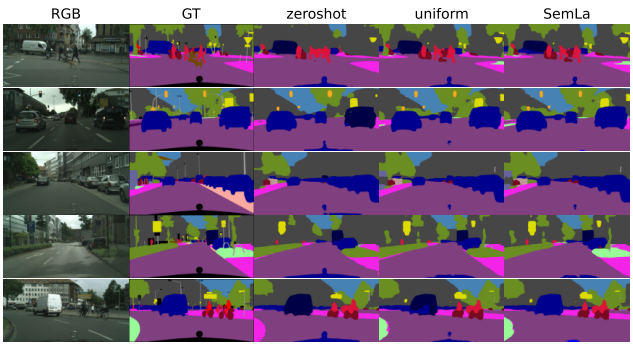
Scalability and Model Calibration. Scaling SemLA to handle a vast number of adapters introduces challenges in identifying and addressing library weaknesses. Automated strategies for recognizing gaps in the library – such as monitoring frequent



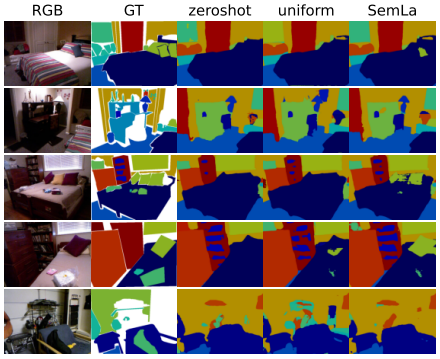
(a) ACDC Fog



(b) BDD



(c) Cityscapes (CS)



(d) NYU

Figure 11. **Additional qualitative results.** The datasets displayed are ACDC Fog, BDD, Cityscapes (CS), and NYU. For each dataset, images are shown in order: Input Image, Zero-Shot, Uniform Merging, SemLA (Ours), Ground Truth. Our method produces more accurate and detailed segmentations across various domains.

occurrences of high embedding distances – can prompt the training of new adapters to fill these gaps. Integrating a LoRA support score into the system allows for continuous monitoring of the model's performance relative to the domain coverage of the adapter library. This not only enhances scalability but also improves the system's robustness and reliability in dynamic environments.