
A APPENDIX

In this appendix, we provide additional experimental results and analyses to support the findings of the main paper. The content is organized as follows:

A. Additional Ablation Studies

A.1 Effect of Unknown Word Set Size

A.2 Effect of Number of Samples per Unknown Word

A.3 Effect of Unknown Sample Size

A.4 Effect of Meta-Train / Meta-Test Split Ratio

B. Experiments with Multiple Backbones

C. Effect of Expanding Unknown Head Capacity

D. Impact of Including Target Unknown Words

E. Fairness of Comparison When Using Synthetic Unknowns

F. Effect of Using Real Images as Unknowns

G. Limitation Discussion and Broader Societal Impact

H. Use of Large Language Models (LLMs)

A ADDITIONAL ABLATION STUDIES

A.1 EFFECT OF UNKNOWN WORD SET SIZE

We compared the segmentation performance with varying numbers of unknown words used to generate unknown samples. Specifically, we randomly sampled 20, 40, 60, and 88 class names from CIFAR-100 after excluding known and target unknown classes, and used these as prompts for Stable Diffusion to synthesize diverse unknowns. As shown in Table 1, increasing the number of unknown words consistently led to performance gains, especially in the detection of Private. The H-Score improved from 33.97 with 20 unknown words to 40.65 with 88 words. This improvement can be attributed to the increased semantic diversity of the training unknown samples, which provided richer supervision for modeling the concept of “unknown.” By being exposed to a broader range of synthetic unknowns, the model was able to learn more generalized representations, resulting in improved detection of previously unseen unknown objects in target domains.

A.2 EFFECT OF NUMBER OF SAMPLES PER UNKNOWN WORD

To examine the effect of data abundance, we varied the number of synthetic samples generated per unknown word. As shown in Table 2, increasing the number of samples from 10 to 100 led to steady improvements in both common and private segmentation performance, with the H-Score rising from 37.82 to 40.65. Interestingly, when the number was further increased to 200, the performance plateaued, and the H-Score slightly dropped to 40.31. These results indicate that generating approximately 100 samples per word is sufficient to construct robust unknown representations. Beyond this threshold, additional samples provide limited benefit.

A.3 EFFECT OF UNKNOWN SAMPLE SIZE

We investigate the role of the spatial resolution of the generated unknown objects. For each unknown sample, we randomly selected a square size $H = W$ from a predefined range and used it to resize the unknown object before pasting it into the source image. As shown in Table 3, we experimented with various size ranges, including fixed ranges like [32:128] and [64:256], as well as broader intervals such as [16:128] and [128:256]. The results indicated that smaller object sizes generally led to better segmentation performance. Notably, the size range [32:128] resulted in the highest H-Score of 40.65. This experiment result suggests that using smaller unknown objects was more effective in the OSDG-SS setting, likely because such objects better aligned with the typical scale of foreground regions and minimized interference with known class regions. Conversely, larger unknown objects often occluded significant portions of the scene or disrupted the spatial context of known class regions,

Table 1: Effect of the number of unknown words used for generating synthetic samples. Evaluation is on Cityscapes after training on GTA5.

# Unknown words	GTA5 → Cityscapes		
	Common	Private	H-Score
20	57.43	24.12	31.21
40	56.74	27.45	34.78
60	57.93	28.41	36.04
88	61.84	30.28	40.65

Table 2: Effect of the number of generated samples per unknown word. Evaluation was conducted on Cityscapes after training on GTA5.

# Samples per word	GTA5 → Cityscapes		
	Common	Private	H-Score
10	60.78	24.12	37.82
50	61.42	29.54	39.89
100	61.84	30.28	40.65
200	60.54	30.21	40.31

Table 3: Effect of varying object size ranges for synthetic unknown samples. The height and width were set equal ($H = W$) and randomly sampled within the given interval. Evaluation was on Cityscapes after training on GTA5.

Object Size	GTA5 → Cityscapes		
	Common	Private	H-Score
[256:256]	44.33	18.63	26.23
[128:256]	53.41	21.21	30.36
[64:256]	58.11	24.61	34.58
[32:256]	59.40	28.61	38.61
[64:128]	58.92	27.23	37.25
[32:128]	61.84	30.28	40.65
[16:128]	60.35	28.15	38.39

Table 4: Effect of meta-train/meta-test split ratio on segmentation performance. Evaluation was conducted on Cityscapes after training on GTA5.

Split Ratio (Meta-train:Meta-test)	GTA5 → Cityscapes		
	Common	Private	H-Score
1:1	58.9	27.42	37.42
2:1	59.39	29.26	39.20
3:1	61.84	30.28	40.65
4:1	61.15	29.77	40.04
5:1	61.64	29.41	39.82

which adversely affected learning. These findings underscore the importance of appropriately sizing synthetic unknowns to provide effective supervision for unknown-aware segmentation.

A.4 EFFECT OF SPLIT RATIO IN META-LEARNING

We evaluated how the ratio between meta-train and meta-test subsets influences generalization in our meta-learning strategy. As shown in Table 4, we varied the ratio from 1:1 to 5:1 while keeping the total number of unknown words and source images fixed. Both the unknown word set and the source domain were partitioned using semantic class identity and RGB-based subdomain splits, respectively. The results in Table 4 indicated that generalization performance improved as the ratio increased from 1:1 to 3:1, where the H-Score reached its highest value of 40.65. This suggested that allocating a larger portion of unknown samples to the meta-train phase allowed the model to better acquire generalized representations of the unknown space. However, increasing the ratio beyond 3:1 led to marginal or slightly decreased performance. We hypothesized that overly reducing the meta-test set would limit the diversity of held-out unknowns, thereby weakening the effect of entropy-based regularization. Overall, the 3:1 configuration offered the best trade-off between supervised learning and generalization in the meta-learning framework.

B EXPERIMENTS WITH MULTIPLE BACKBONES

To assess the generalizability of our framework, we evaluated performance across three backbone architectures: SAM (Kirillov et al., 2023), EVA02 (Fang et al., 2023), and DINO-V2 (Oquab et al., 2023). As shown in Table 5, our method consistently outperformed both Rein (Wei et al., 2024) and FADA (Bi et al., 2024) under all backbone settings. Notably, with DINO-V2, our method achieved the best H-Score of 40.65 on Cityscapes and an average H-Score of 33.85 across three target domains, indicating strong robustness to both semantic and visual domain shifts. To evaluate the impact of our framework on ResNet-based architectures, we conducted an additional ablation using IBN-Net (Pan et al., 2018), which achieved the best performance among ResNet-50-based DG-SS baselines. Table 6 presents the results. Incorporating our method into the IBN-Net baseline yielded a substantial improvement in average H-Score, from 14.78 to 20.34. Furthermore, our method

Table 5: Backbone ablation with ResNet-50. Our method provides substantial performance gains even without using the Vision Foundation Model (VFM) backbones, and improves upon the strong ResNet-based IBN-Net baseline.

Backbone	Method	Trained on GTA5			
		→ Cityscapes	→ BDD100k	→ Mapillary	AVG
		H-Score			
SAM Kirillov et al. (2023)	Rein	17.43	15.42	10.13	14.33
	FADA	24.86	17.39	18.45	20.23
	Ours	34.41	28.91	24.23	26.57
EVA02 Fang et al. (2023)	Rein	17.64	13.57	10.84	14.02
	FADA	24.41	15.64	21.84	20.63
	Ours	34.42	26.50	26.44	29.12
DINO-V2 Oquab et al. (2023)	Rein	20.14	13.04	15.77	16.32
	FADA	27.21	14.96	24.43	22.20
	Ours	40.65	27.48	33.41	33.84

Table 6: Backbone ablation with ResNet-50. Our method improves upon the strong IBN-Net baseline and maintains robustness even without VFM backbones.

Method	Backbone	Trained on GTA5									Average		
		→ Cityscapes			→ BDD 100k			→ Mapillary					
		Common	Private	H-Score	Common	Private	H-Score	Common	Private	H-Score	Common	Private	H-Score
IBN-Net	ResNet-50	22.64	18.76	20.51	19.34	9.94	13.13	16.92	7.83	10.70	19.63	12.17	14.78
Ours (IBN-Net base)		33.46	22.56	26.95	30.78	11.56	16.80	25.13	13.16	17.27	29.79	15.76	20.34
Ours (FADA base)	DINO-V2	61.84	30.28	40.65	53.48	17.34	27.48	53.64	24.48	33.41	47.05	24.98	33.84

maintained strong performance even when using ResNet-50 as the backbone instead of DINO-V2, achieving an average H-Score of 33.84.

C EFFECT OF EXPANDING UNKNOWN HEAD CAPACITY

We investigated how expanding the segmentation head to include multiple unknown class nodes (i.e., $K=66$) affected performance compared to using a single unknown node ($K=1$). As shown in Table 7, expanding the head yielded consistent improvements across all model configurations. Specifically, Ours achieved an H-Score of 40.65 with $K=66$, outperforming its $K=1$ counterpart (36.04) by a margin of 4.61. Similar trends were observed in intermediate configurations. For example, when only synthetic unknown sample was applied (Config B), the H-Score improved from 30.78 ($K=1$) to 29.54 ($K=66$), and with MUG added (Config C), the H-Score increased from 34.92 to 36.60. These results indicated that modeling unknowns with multiple dedicated head nodes allowed the model to better represent the semantic diversity of unknown categories. In contrast, using a single unknown node limited the model’s capacity to express uncertainty across diverse unknown regions and likely forced overly coarse decisions. Furthermore, multi-node heads encouraged more fine-grained discrimination in the feature space, resulting in improved private-class segmentation. Overall, the head expansion strategy proved essential for achieving strong performance in open-set domain generalization.

D IMPACT OF INCLUDING TARGET UNKNOWN WORDS IN THE SYNTHETIC UNKNOWN SET

To assess the effect of including target-domain unknown categories in the synthetic unknown word pool, we compared three configurations with varying numbers of “shared unknowns” — that is, synthetic unknown words that also appear in the target-domain unknown classes. In all settings, we sampled additional object words from CIFAR-100 to create synthetic unknowns, while maintaining the meta-train/meta-test split at a 2:1 ratio across all unknowns, including shared ones.

As shown in Table 8, when no shared unknowns were included (i.e., a strict separation between training unknowns and evaluation-time unknowns), our method achieved an H-Score of 40.65. However, introducing 3 shared unknowns (out of 91 total) increased the H-Score to 43.25, and including 6 shared unknowns further boosted it to 45.81. This trend suggests that performance benefits from partial overlap between synthetic and actual unknown classes, even when the overlap is

Table 7: Comparison of segmentation performance when expanding the unknown head to $K = 66$ versus using a single unknown node ($K=1$). Evaluation was conducted on Cityscapes after training on GTA5.

Config	Method	GAT5 \rightarrow Cityscapes					
		$K = 66$			$K=1$		
		Common	Private	H-Score	Common	Private	H-Score
A	FADA with Confidence-based threshold	57.43	18.59	27.21	57.43	18.59	27.21
B	A+MUG (w/o meta-learning)	52.91	20.49	29.54	56.74	21.12	30.78
C	B+MUG	59.42	26.44	36.60	57.93	24.51	34.92
Ours	C+ODB	61.84	30.28	40.65	60.84	25.61	36.04

Table 8: Effect of including target-domain unknown classes in the synthetic unknown word set. Shared unknowns were randomly split into meta-train and meta-test subsets at a 2:1 ratio. Evaluation was conducted on Cityscapes after training on GTA5.

Shared unknown/ Total unknown word set	Meta-train # Unknown	Meta-test # Unknown	Additional # head node	GAT5 \rightarrow Cityscapes		
				Common	Private	H-Score
0 / 88	66	22	+ 66	61.84	30.28	40.65
3 / 91	68	23	+ 68	64.97	23.22	43.25
6 / 94	70	24	+ 70	62.59	36.12	45.81

limited. Notably, the gains in H-Score were mostly attributed to improvements in the private-class IoU, which increased from 30.28 to 36.12 as more shared unknowns were included. These results highlight the importance of designing synthetic unknowns that are semantically relevant to the actual deployment scenarios.

E FAIRNESS OF COMPARISON WHEN USING SYNTHETIC UNKNOWNNS

For the OSDG-CLS method, while it is designed to predict unknowns, the framework is not built to incorporate additional samples of synthetic unknowns. Indiscriminate injection of such data can conflict with the original method.

Table 9: Performance comparison of CSDG-SS methods trained on GTA5. The evaluation metric is H-Score.

Method	Cityscapes	BDD100k	Mapillary	AVG
Rein	17.64	13.57	10.84	14.01
Rein + MUG (w/o meta-learning)	24.41	15.64	21.84	20.63
Rein + Ours	34.42	26.50	26.44	29.12
FADA	27.21	14.97	24.43	24.08
FADA + MUG (w/o meta-learning)	29.54	21.84	27.12	26.16
FADA + Ours	40.65	27.48	29.70	33.84

For the CSDG-SS method, which does not handle unknowns, we can observe the impact of providing synthetic unknowns by comparing the baseline in Table 9. As shown, Rein+MUG (w/o meta-learning) and FADA+MUG (w/o meta-learning) both outperform their respective baselines, confirming that providing unknown samples can improve performance even without structural changes. However, comparing Rein+Ours and FADA+Ours highlights a much larger improvement in meta-learning. This suggests that our approach contributes more than just increasing the synthetic unknown sample datasets.

F EFFECT OF USING REAL IMAGES AS UNKNOWNNS

To further validate the necessity of using synthetic unknown samples, we conducted an additional experiment by directly using real images from CIFAR-100 as unknown samples, instead of Stable

Table 10: Comparison between synthetic and real unknown samples. Synthetic denotes Stable Diffusion-generated samples, while Real denotes CIFAR-100 images. The evaluation metric is H-Score.

Trained on GTA5	Cityscapes	BDD100k	Mapillary	AVG
Ours (Synthetic)	40.65	27.48	33.41	33.84
Ours (Real)	36.41	28.12	35.12	33.21

Trained on SYNTHIA	Cityscapes	BDD100k	Mapillary	AVG
Ours (Synthetic)	30.07	24.69	28.52	27.76
Ours (Real)	28.12	26.84	31.12	28.69

Diffusion-generated images. The results are summarized in Table 10. We observe that replacing synthetic unknowns with real images yields comparable performance: when trained on GTA5, the average H-Score decreases only marginally from 33.84 (synthetic) to 33.21 (real), and when trained on SYNTHIA, the average H-Score is 27.76 (synthetic) to 28.69 (real). These results suggest that the overall effectiveness of our framework does not rely critically on the choice between real or synthetic unknowns. Nevertheless, the use of a text-to-image generation model provides the advantage of generating object-shaped unknowns that are semantically diverse and tailored to our task, making it a more flexible and scalable solution for OSDG-SS.

G LIMITATION DISCUSSION AND BROADER SOCIETAL IMPACT

Limitation Discussion Our framework for OSDG-SS demonstrates strong performance but also presents several limitations that merit discussion. First, the use of a fixed external vocabulary (e.g., CIFAR-100) and synthetic generation through Stable Diffusion introduces a dependency on curated prompt sets that may not fully capture the complexity of real-world unknowns. Although we carefully removed overlaps with known and target categories, the resulting synthetic unknowns may still differ semantically or visually from actual deployment-time unknowns. Furthermore, the segmentation head expansion by the number of unknown prompts, while effective, introduces scalability concerns when dealing with large or evolving unknown spaces. These design decisions—such as the number of unknown samples per word, prompt diversity, and split ratios in meta-learning—may require tuning across domains, which could limit general applicability. Despite these constraints, our results suggest that careful modeling of unknown diversity and meta-train/meta-test structure can lead to robust generalization.

Broader Societal Impact In terms of societal impact, our approach has the potential to improve the reliability of perception systems in open-world settings such as autonomous driving or medical diagnostics by enabling explicit handling of unfamiliar inputs. However, the use of generative models also raises concerns about representational fairness and bias; if the prompt vocabulary lacks inclusivity or reflects dataset biases, it may lead to blind spots in unknown recognition. These limitations highlight the importance of responsible dataset design and continued investigation into adaptive, scalable, and fair approaches for open-set segmentation.

H USE OF LARGE LANGUAGE MODELS (LLMs)

We clarify that large language models (LLMs) were used solely to aid or polish the writing of this paper, such as improving readability and refining phrasing. They were not involved in the design of the methodology, experiments, or analysis, and therefore did not contribute at the level of author.

REFERENCES

Qi Bi, Jingjun Yi, Hao Zheng, Haolan Zhan, Yawen Huang, Wei Ji, Yuexiang Li, and Yefeng Zheng. Learning frequency-adapted vision foundation model for domain generalized semantic segmentation. *Advances in Neural Information Processing Systems*, 37:94047–94072, 2024.

270 Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and
271 Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the*
272 *IEEE/CVF conference on computer vision and pattern recognition*, pages 19358–19369, 2023.

273 Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer
274 Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*,
275 2023.

276 Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez,
277 Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without
278 supervision. *arXiv preprint arXiv:2304.07193*, 2023.

279 Xingang Pan, Ping Luo, Jianping Shi, and Xiaoou Tang. Two at once: Enhancing learning and generalization
280 capacities via ibn-net. In *Proceedings of the european conference on computer vision (ECCV)*, pages 464–479,
281 2018.

282 Zhixiang Wei, Lin Chen, Yi Jin, Xiaoxiao Ma, Tianle Liu, Pengyang Ling, Ben Wang, Huaian Chen, and Jinjin
283 Zheng. Stronger fewer & superior: Harnessing vision foundation models for domain generalized semantic
284 segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages
285 28619–28630, 2024.

286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323