Supplementary Material: Exploring Limits of Diffusion-Synthetic Training with Weakly Supervised Semantic Segmentation

1 Implementation details

1.1 Used codebases

For generating training images and pseudo-masks. we used diffusers v0.13.1 from https://github.com/huggingface/diffusers. We applied minor modifications to keep and extract cross-attention maps from generation processes, but mathematically, this did not affect the generated images. For weakly supervised segmentation training, we used BECO's official pytorch implementation from https://github.com/ShenghaiRong/BECO.

For DreamBooth-LoRA-based adaptation, we used diffusers' training script from https://github.com/huggingface/diffusers/blob/main/examples/dreambooth/train_dreambooth_lora.py. We will release our scripts to run the abovementioned codes and generated datasets upon acceptance.

1.2 Hyperparameters

The list of hyperparameters in segmentation training is summarized in Tab. 1. The list of hyperparameters in fine-tuning DreamBooth-LoRA is summarized in Tab. 2.

2 Comparisons with text-based segmentation methods

Stable-Diffusion-based segmentation training of attn2mask may be compared with text-based segmentation methods with CLIP in terms of using vision-language pretraining as their foundations. Table 3 shows comparisons of attn2mask and CLIP-based segmentation methods that support zero-shot transfer without access to the downstream-domain training images or labels. While they do not need retraining of segmentation models, they often suffer from inaccurate boundaries. Our attn2mask has an advantage in improving mask accuracy by synthetic retraining for segmentation, especially when combined with the robust co-training, at the cost of computation for the training data generation and retraining of the segmentation model.

3 Additional examples

Figure 1 visualizes examples of the attention maps corresponding to each prompt token, which are the source of our pseudo-labels. We can see that Stable Diffusion distributes attention well in response to the prompts, even in the multi-object

Table 1: Training setting and hyperparameters

Parameter name	Value
Architecture	DeepLabv3+ & ResNet50 or Swin-B
Optimizer	SGD
Learning-rate schedule	Polynominal
Initial learning rate	0.001 (ResNet50) 0.01 (Swin-B)
Learning-rate power	0.9
Weight decay	0.00001
Batch size	6 per GPU × 2 GPUs
Number of iterations	25,000

Table 2: Fine-tuning setting of DreamBooth-LoRA in Cityscapes

Parameter name	Value
Optimizer	AdamW
Learning-rate schedule	Polynomial
Initial learning rate	0.00002
Learning-rate power	1.0
Weight decay	0.01
Batch size	1 per GPU × 4 GPUs
Number of iterations	40,000

generation. This suggests that visual grounding emerges in learning large-scale text2img generation without explicit location-based supervision.

To understand the behavior of pseudo-masks extracted from the attention maps, Fig. 2 visualizes typical examples of pseudo-masks and corresponding reliability maps. Unreliable regions in the reliability maps (black) are often caused by dCRF's propagation failures in hard-to-classify regions, as well as interpolation entailed in upsampling of the attention maps around object edges. The reliability maps prevent the labeling failure regions from harming the segmentation training but may drop recovered edges by dCRF. Overall, we found that the former was more beneficial in improving the final mIoU of downstream segmentation models.

Figure 3 shows more examples generated for PASCAL VOC. We observed the successful generation of images and pseudo-labels even in the multiple-object generation cases. Figure 4 shows examples generated for Cityscapes. While images before adaptation (a) are not similar to the real examples (c), ones after adaptation (b) are more aligned with the real ones in terms of color balance and scene composition.

Table 3: Comparisons of attn2mask and text-based zero-shot-transfer segmentation methods in VOC.

Method	mIoU
Attn2mask-ResNet50 (ours)	62.2
Attn2mask-SwinB (ours)	**71.0**
MaskCLIP [5]	22.1
ALIGN [1]	29.7
GroupViT [3]	41.1
OVS [4]	44.6
CLIPpy [2]	50.8

Table 4: Abalative analyses in VOC12 with ResNet50.

Full	(a) W/o co-training	(b) W/o adaptive threshold.	(c) W/o prompt aug.
62.2	56.3	61.4	61.3

4 Additional ablation study

Although we conducted the experiment adding modules one-by-one in the main text, we additionally performed the ablation of major modules by removing one while keeping the others fixed for rigor. Table 4 shows the results, confirming the performance gains by the modules.

5 Details of prompt augmentation

The lists of words used for the synonym-and-hyponym replacement are shown in Table 5. The lists were collected by querying ChatGPT "`please raise fifteen examples of synonyms or hyponyms for ${class_names}`," and later manually curated. In the curation process, apparently improper words, for example, "bus" as a hyponym for "car", were removed. We also modified ambiguous words; for example "persian" as a hyponym of "cat" may cause a generation of carpets and we changed it into more specific "persian cat". Using these lists, the base names that appeared in the original caption corpus were randomly replaced with their corresponding synonyms and hyponyms with the probability of 50%.

References

1. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: ICML. pp. 4904–4916. PMLR (2021) 3
2. Ranasinghe, K., McKinzie, B., Ravi, S., Yang, Y., Toshev, A., Shlens, J.: Perceptual grouping in contrastive vision-language models. In: ICCV (2023) 3

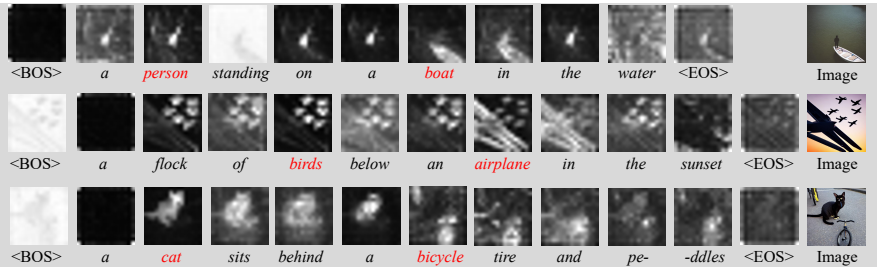

Fig. 1: Visualized attention maps corresponding to each of prompt tokens.

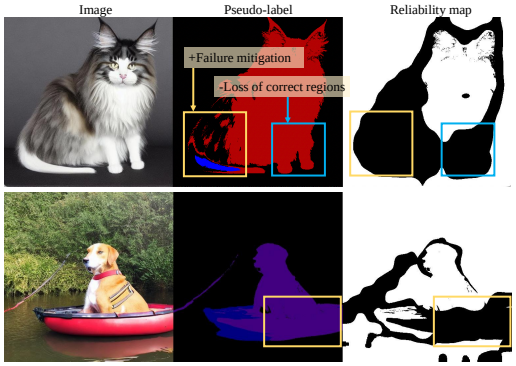

Fig. 2: Examples of pseudo-masks and corresponding reliability maps.

3. Xu, J., De Mello, S., Liu, S., Byeon, W., Breuel, T., Kautz, J., Wang, X.: GroupViT: Semantic segmentation emerges from text supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18134–18144 (2022) 3
4. Xu, J., Hou, J., Zhang, Y., Feng, R., Wang, Y., Qiao, Y., Xie, W.: Learning open-vocabulary semantic segmentation models from natural language supervision. In: CVPR. pp. 2935–2944 (2023) 3
5. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from CLIP. In: ECCV (2022) 3

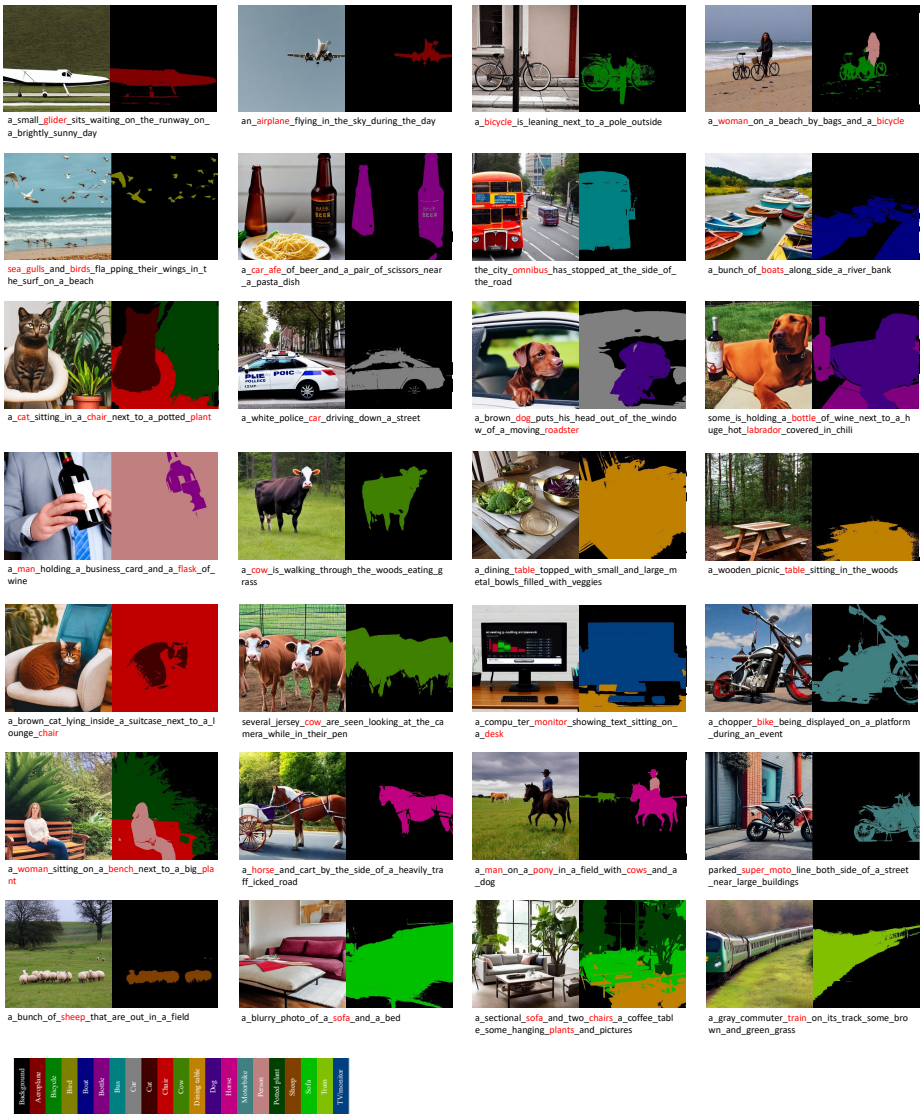

Fig. 3: Examples of the generated dataset for VOC with attn2mask. Prompts are separated by "_" that represents the sub-word tokenization. The class words are highlighted in red.

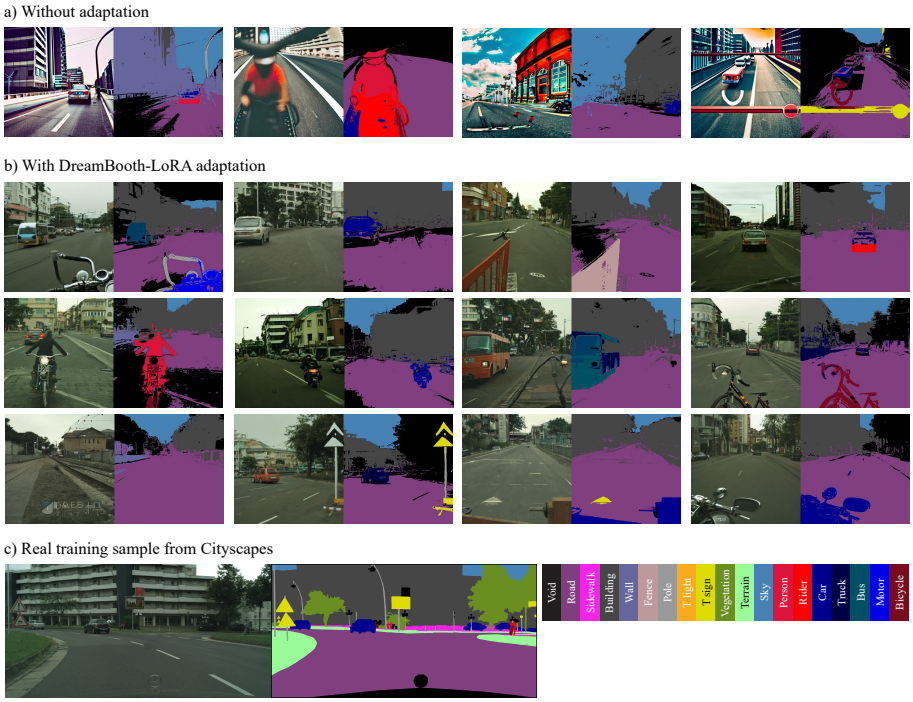

Fig. 4: Examples of the generated dataset for Cityscapes with and without adaptation.

Table 5: List of class-name aliases used for prompt augmentation in VOC.

Base names	Synonyms and hyponyms
“airplane”, “airplanes”, “plane”, “planes”	“aircraft”, “jet”, “airliner”, “glider”, “drone”, “seaplane”, “ultralight aircraft”
“bicycle”, “bicycles”	“pushbike”, “road bike”, “mountain bike”, “racing bike”, “bmx”, “tandem bike’, “recumbent bicycle”, “folding bike”, “electric bike”
“bird”, “birds”	“avian”, “songbird”, “sparrow”, “falcon”, “hawk”, “eagle”, “owl”, “crow”, “seagull”, “parrot”, “swallow”, “woodpecker”, “pigeon”
“boat”, “boats”, “ship”, “ships”, “yacht”, “yachts”	“vessel”, “canoe”, “kayak”, “rowboat”, “sailboat”, “catamaran”, “dinghy”, “speedboat”, “barge”, “fishing boat”, “gondola”
“bottle”, “bottles”	“flask”, “vial”, “decanter”, “carafe”, “squeeze bottle”, “perfume bottle”
“bus”, “buses”, “busses”	“omnibus”, “motorcoach”, “school bus”, “transit bus”, “double-decker bus”, “trolleybus”, “shuttle bus”, “intercity bus”, “articulated bus”
“car”, “cars”	“automobile”, “motorcar”, “sedan”, “coupe”, “convertible”, “suv”, “pickup truck”, “minivan”, “hatchback”, “station wagon”, “roadster”
“cat”, “cats”	“feline”, “kitten”, “tabby cat”, “siamese cat”, “persian cat”, “maine coon”, “bengal cat”, “ragdoll cat”, “bobcat”
“chair”, “chairs”	“armchair”, “recliner”, “rocking chair”, “lounge chair”, “dining chair”, “barstool”, “office chair”, “folding chair”, “bench”, “high chair”
“cow”, “cows”	“bovine”, “heifer”, “steer”, “bull”, “dairy cow”, “beef cow”, “holstein”, “angus”, “jersey cow”
“table”, “tables”	“desk”, “dining table”, “coffee table”, “side table”, “end table”, “kitchen table”, “writing desk”, “console table”, “card table”
“dog”, “dogs”	“canine”, “pooch”, “hound”, “puppy”, “mutt dog”, “terrier”, “bulldog”, “poodle”, “labrador”, “shepherd”, “retriever”
“horse”, “horses”	“equine”, “stallion”, “mare”, “gelding”, “pony”, “foal”, “thoroughbred”, “clydesdale”, “palomino horse”, “appaloosa”
“motor”, “motors”, “motorcycle”, “motorcycles”	“chopper bike”, “sportbike”, “dirt bike”, “scooter”, “cafe racer”, “bobber bike”, “moped”, “supermoto”
“person”, “people”, “man”, “woman”, “men”, “women”, “player”, “players”	“human”, “homo sapiens”, “citizen”
“plant”, “plants”	“flora”, “greenery”, “vegetation”, “herb”, “tree”, “flower”, “cactus”, “succulent”
“sheep”	“ewe”, “woolly animal”, “merino sheep”, “suffolk sheep”, “dorper sheep”, “shetland sheep”, “border leicester”, “jacob sheep”
“sofa”, “sofas”	“couch”, “settee”, “loveseat”, “chesterfield”, “chaise lounge”, “sectional sofa”, “daybed”, “sleeper sofa”
“train”, “trains”	“locomotive”, “subway train”, “metro train”, “commuter train”, “bullet train”
“monitor”, “monitors”, “tv”, “tvs”, “television”, “televisions”	“display”, “screen”, “vdu”, “computer monitor”, “flat panel monitor”, “crt monitor”, “led monitor”, “touchscreen”, “dual monitor”