Supplementary Materials for AlignDiff: Aligning Diffusion Models for General Few-Shot Segmentation

Ri-Zhao Qiu[1,2], Yu-Xiong Wang[1†], and Kris Hauser[1†]

[1] University of Illinois Urbana-Champaign
[2] University of California San Diego
{rizhaoq2, yxw, kkhauser}@illinois.edu
[†]equal advising
http://github.com/RogerQi/AlignDiff

1 Ablation Study

In addition to the ablation on the FSS-1000 [11] dataset in the main paper, in Tab. 1, we ablate different components in AlignDiff on GFSS, compared to training directly using synthetic samples from the Stable diffusion [15] model. We use GAPS [14] as the base model to apply AlignDiff and GD [12] to for the ablation study. Following previous works in semantic segmentation [2], we use the last five classes of the Pascal VOC dataset as novel classes.

To simulate training directly with pure synthetic samples without any conditioning on real samples, when 'SMask' (semi-supervised learning for pixel-annotation) is turned off, we use the method of [12] to generate masks in a zero-shot manner. Finally, when T.Inv. (normalized masked textual inversion) is turned off, we generate samples using only class names as text guidance.

AlignDiff provides high-quality pixel-level annotations. Compared to the zero-shot segmentation method proposed by [12], which requires a well-trained COCO instance segmentation model to provide annotation during base training, our semi-supervised mask generation method provides pixel-level annotations of higher quality and improve the novel IoU by approximately 10%.

Textual inversion generates more diverse samples. Compare to samples synthesized with pure text conditioning, the instance-specific embedding learned by AlignDiff introduces more diversity and further improves the novel IoU by approximately 10%.

2 CLIP Score

We use the CLIP score [10] as a proxy to evaluate the alignment of generated samples to the underlying class distribution. CLIP score [10] is known as a proxy for approximating how humans would rate the similarity between provided captions and the image.

Table 1: Ablating components of AlignDiff on 1-shot Pascal-5^3, which is a fold of Pascal-5^i. Ablation is consistent with results on FSS even on more common classes. Standard deviations over 5 runs are reported.

SMask	T.Inv.	Novel IoU
—	—	36.8 ± 1.4
✓	—	38.4 ± 1.2
✓	✓	**41.5** ± 1.2

Table 2: Comparison of CLIP scores of AlignDiff-generated samples to text-conditioned samples on rare categories of the FSS-1000 [11] dataset, which serves as a proxy to how humans would rate the similarity between texts and images. (Best is bolded)

Con. Method	CLIPScore↑ [10]
AlignDiff (ours)	**0.79**
Text-based [12, 17]	0.77

We present the results in Tab. 2. We evaluate the CLIP score on the out-of-distribution rare categories on the FSS-1000 dataset by comparing AlignDiff-generated samples and simple text-conditioned samples from Stable Diffusion, which is used in DiffuMask [17] and GD [12]. CLIP score suggests that our generated samples are preferred.

3 Dataset Description

Pascal-5^i is artificially built from the PASCAL VOC 2012 dataset [4] with augmented annotations from the SBD [8] dataset. The original VOC segmentation dataset provides segmentation annotations for 20 object categories, and the Pascal-5^i dataset manually splits the original dataset into 4 folds for cross-validation. For each fold, 5 categories are selected as novel categories, while the remaining 15 are regarded as base categories. The construction of the COCO-20^i dataset uses the 80 thing classes in COCO similarly, where the dataset is split into 4 folds with 20 categories per fold.

4 More Results on Few-Shot Segmentation for Out-of-Distribution Generation

For a more in-depth analysis, we present a class-wise IoU difference between the second row and the first row of Table. 1 (from the main paper) in Fig. 1. We can observe that the effects of synthetic samples generated by GD are mixed for individual classes. For 50 out of the 240 categories in the testing split of FSS-1000, synthetic samples generated by GD improve the final IoU. However, for the rest 190 categories, GD-generated samples negatively impact the IoU, which suggests that the generated image-mask pairs for these classes are inaccurate. We mark these 190 categories as 'out-of-distribution' categories since synthetic samples generated by GD negatively impact the final segmentation performance. The results presented in Table 3 of the main paper are computed using the top-5 classes at the end ('pidan', 'Samarra Mosque', 'Chess queen', 'American

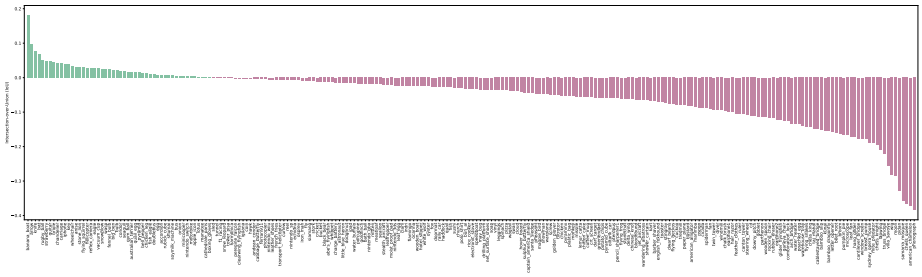

Fig. 1: GD [12] may synthesize samples that are harmful to the overall IoU. Class-wise IoU difference of HSNet [13] on the FSS-1000 dataset [11] under the 1-shot setting with support set of 1) only 1 real sample and 2) 1 real sample and 20 text-conditioned synthetic samples generated by GD [12]. Green bars denote classes whose text-conditioned samples improve the final IoU, whereas red bars denote classes whose text-conditioned samples negatively influence the final IoU. We mark classes with red bars as out-of-distribution categories. 190 out of 240 total categories are considered out-of-distribution because synthetic samples generated by GD are harmful to the segmentation performance.

Chamelon', and 'Phonograph'). Analysis of different failure patterns for out-of-distribution categories is given in Sec. 6.

Finally, AlignDiff, which conditions both the image generation process and the mask generation process using a few provided real images, addresses both aforementioned issues. This results in improved IoU compared with all other settings using a single real sample and using text-conditioned synthetic samples. The class-wise detailed analysis of the IoU difference between AlignDiff-augmented few-shot segmentation and 1-shot real-sample-only few-shot segmentation is given in Fig. 2. Synthetic samples generated by AlignDiff improve the IoU of 178 out of the 240 testing categories, which results in better overall IoU and the OOD IoU of classes where GD-synthesized samples negatively impact the segmentation performance.

5 Implementation Details

5.1 Text-to-Image Synthesis Model

We use the pre-trained checkpoints and the codebase of Stable Diffusion [15], which is available at[3]. The weights of the Stable Diffusion are frozen.

5.2 Training Details for Generalized Few-shot Semantic Segmentation

We follow the standard sequential learning procedure that two recently published works, PIFS [1] and GAPS [14], use in their papers. Specifically, these two works

[3] https://github.com/CompVis/stable-diffusion

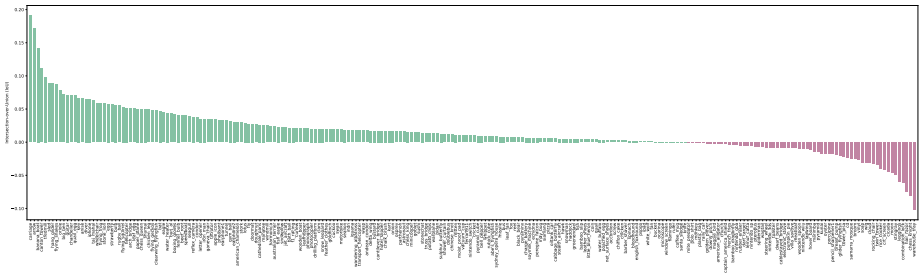

Fig. 2: AlignDiff handles out-of-distribution generation and improves the overall IoU of 1-shot segmentation on the FSS-1000 dataset [11]. Class-wise IoU difference of HSNet [13] on the FSS-1000 dataset [11] under the 1-shot setting with support set of 1) only 1 real sample and 2) 1 real sample and 20 synthetic samples generated by AlignDiff (ours). Green bars denote classes whose synthetic samples improve the final IoU, whereas red bars denote classes whose synthetic samples negatively influence the final IoU. AlignDiff improves IoU for a total of 178 out of 240 classes and improves the overall IoU.

separate the learning process into two stages. In the first base learning stage, they use the standard training procedure to train a segmentation model on the base datasets. To avoid information leaks from unseen classes, the base dataset excludes all images with at least one pixel of the novel classes. The architecture of the segmentation model is DeepLab-V3 [3], which uses the ResNet-101 backbone and replaces the last per-pixel classification layer with a cosine-similarity-based per-pixel classification layer. Following GAPS [14], on both the COCO and the PASCAL VOC datasets, the base training uses a batch size of 32, a polynomial learning rate schedule with an initial learning rate of 0.01, and training for 20 epochs. The SGD optimizer is used with 0.9 momentum and 0.0001 weight decay. Following existing work [1,14], standard augmentation techniques such as random scaling, random cropping, and random horizontal flipping are used.

During the novel learning stage, we again resort to the training settings proposed by GAPS [14]. We fine-tune the segmentation models every time the model is presented with a new class (*e.g.,* a total of five fine-tunings for 5 classes on each split of the Pascal-5^i dataset). For GAPS-based methods, on the PASCAL VOC dataset, we train the segmentation model for 200 iterations using a batch size of 16, a backbone learning rate of 0.001, and a classifier learning rate of 0.01. For the GAPS-based method, we use the memory-replay buffer construction strategy proposed in GAPS [14] to construct a small subset of 500 base examples, $\hat{\mathcal{D}}^B$, for copy-pasting the novel samples onto. For finetune-based methods, the training settings such as batch size are similar to GAPS. However, the subset of base examples to paste on, $\hat{\mathcal{D}}^B$, is constructed by randomly selecting 500 examples from the base dataset.

In addition to the learning settings, we also list details of how AlignDiff uses the synthetic samples. When AlignDiff is combined with plain fine-tuning, we use the copy-paste strategy [7] to create samples with realistic scene layouts.

When it is combined with GAPS [14], which is a method based on copy-paste, we follow the procedure described in [14] and directly add synthetic samples with their masks as candidates for copy-paste [7] in GAPS. In both cases, 50% of the synthetic samples are selected from the few-shot samples or samples conditioned via normalized masked textual inversion; whereas the images of the rest of the 50% samples are purely synthetic using text conditioning.

5.3 Training Details for Few-Shot Semantic Segmentation

We follow the standard training procedure in few-shot semantic segmentation using support-query episodes [5, 13, 16]. In particular, we use HSNet [13], which is a commonly used method in few-shot segmentation, to investigate the performance of AlignDiff to handle out-of-distribution generation on the FSS-1000 dataset. We use the same training settings in HSNet, where support and query images are resized to 400 by 400, and parameters are optimized via an Adam optimizer with a learning rate of 1e-3.

Unlike PIFS [1] and GAPS [14] from generalized few-shot segmentation, HSNet does not require fine-tuning on novel data. Instead, given a support set and a query image, HSNet refines the predictions of the mask of the query image using the samples given in the support set. Therefore, we focus on augmenting the support set with additional synthetic samples to improve the performance. In order to avoid including degenerate samples for out-of-distribution categories, we perform a simple estimation step. For every text-conditioned sample, we use it as a 1-shot support set and treat the given real support image as a query image to perform 1-shot segmentation. If the IoU of the predicted mask and the given query mask exceeds a pre-defined threshold β (we set $\beta = 0.5$), then this particular text-conditioned sample is added to the final support set.

6 Analysis of Failure Patterns of Plain Text Conditioning

As illustrated in the text-conditioned sample column of five out-of-distribution categories from FSS-1000 [11] in the main paper, a few types of failure patterns can be observed for GD [12] and Stable Diffusion [15] in general when the generative process is guided using plain texts. These factors necessitate methods for conditioning the generative process using a few input image-mask pairs.

- **Inaccurate attention.** Stable diffusion may attend to only a single word in a multi-word object phrase, which is shown in the 'samarra mosque' sample in the first row. Stable Diffusion incorrectly attends to only the word 'mosque' and generates images of a common mosque, rather than the Samarra mosque. The chess queen samples in the fifth row fail in a similar pattern, where Stable Diffusion incorrectly attends mainly to the word 'queen'.
- **Uncomprehensive description.** The second row shows samples for the 'phonograph' class from FSS-1000. In this case, images of phonograph in

FSS-1000 are all instances of vintage phonographs with copper horns. However, when given the prompt 'a photo of a phonograph,' Stable Diffusion generates images of common phonographs with no horns and a turnable player, as shown in the samples.

– **Rare concept.** Pidan (third row) and American Chameleon (fourth row), as a type of uncommon Asian traditional food, are also classes from the FSS-1000 dataset. However, Stable Diffusion fails completely on those novel concepts when given only text prompts and generates irrelevant images.

7 More Qualitative Results

We provide additional representative synthetic samples generated by our AlignDiff and compare them with GD [12] in this section.

Fig. 3 gives results of text-conditioned image synthesis for the last five classes of the PASCAL VOC dataset with masks generated by GD [12]. We notice that the PASCAL VOC dataset contains only commonly seen classes. Thus, Stable Diffusion can generate diverse instances for PASCAL VOC using only text conditioning. However, GD fails to generate accurate masks, which hinders the training of segmentation models.

Fig. 4 shows results of text-conditioned image synthesis with masks generated by the semi-supervised mask generator that we proposed. We can notice how AlignDiff is able to generate masks with crispy boundaries.

Fig. 5 shows interesting qualitative results on the PASCAL-5-3 split using the proposed normalized masked textual inversion method for conditioning. The first row contains the 1-shot image-mask pairs provided to AlignDiff, and the rest of the image-mask pairs are generated by AlignDiff. The figure illustrates two intriguing findings: 1) AlignDiff is able to generate relevant images despite the objects of interest may only occupy a small region in the original image, which is not possible with plain textual inversion [6] and 2) Compared to text-conditioned samples in Fig. 3 and Fig. 4, samples synthesized by AlignDiff increas the diversity of synthesized samples (*e.g.,* AlignDiff generates sofas with more realistic camera viewpoint and CRT monitors with variations). This helps better capture intra-class variation, which is also quantitatively validated in the ablation study in the main paper.

References

1. Cermelli, F., Mancini, M., Xian, Y., Akata, Z., Caputo, B.: Prototype-based incremental few-shot semantic segmentation. In: BMVC (2021) 3, 4, 5
2. Cha, S., Yoo, Y., Moon, T., et al.: Ssul: Semantic segmentation with unknown label for exemplar-based class-incremental learning. NeurIPS (2021) 1
3. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 4
4. Everingham, M., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. IJCV (2010) 2

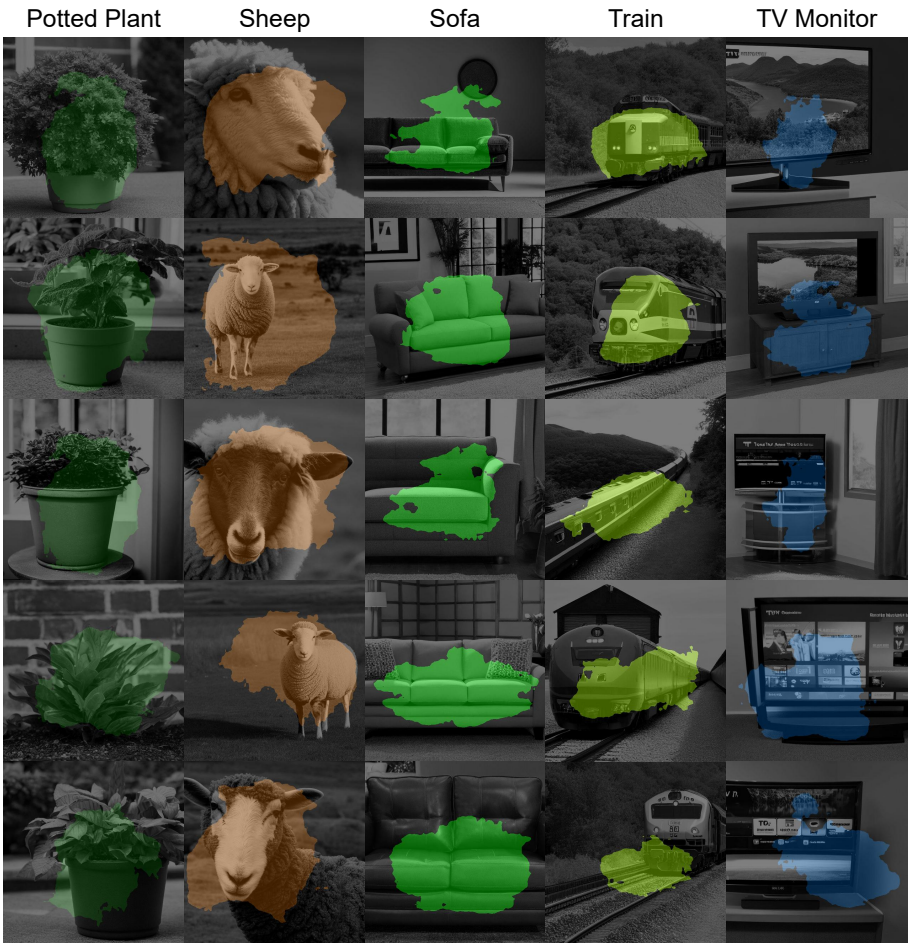

Fig. 3: Qualitative results on the PASCAL-5-3 split using plain text conditioning. Masks are generated by GD [12]. Best viewed in color.

5. Fan, Q., Pei, W., Tai, Y.W., Tang, C.K.: Self-support few-shot semantic segmentation. In: ECCV (2022) 5
6. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 6, 9
7. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: CVPR (2021) 4, 5
8. Hariharan, B., Arbelaez, P., Bourdev, L., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011) 2
9. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022) 10

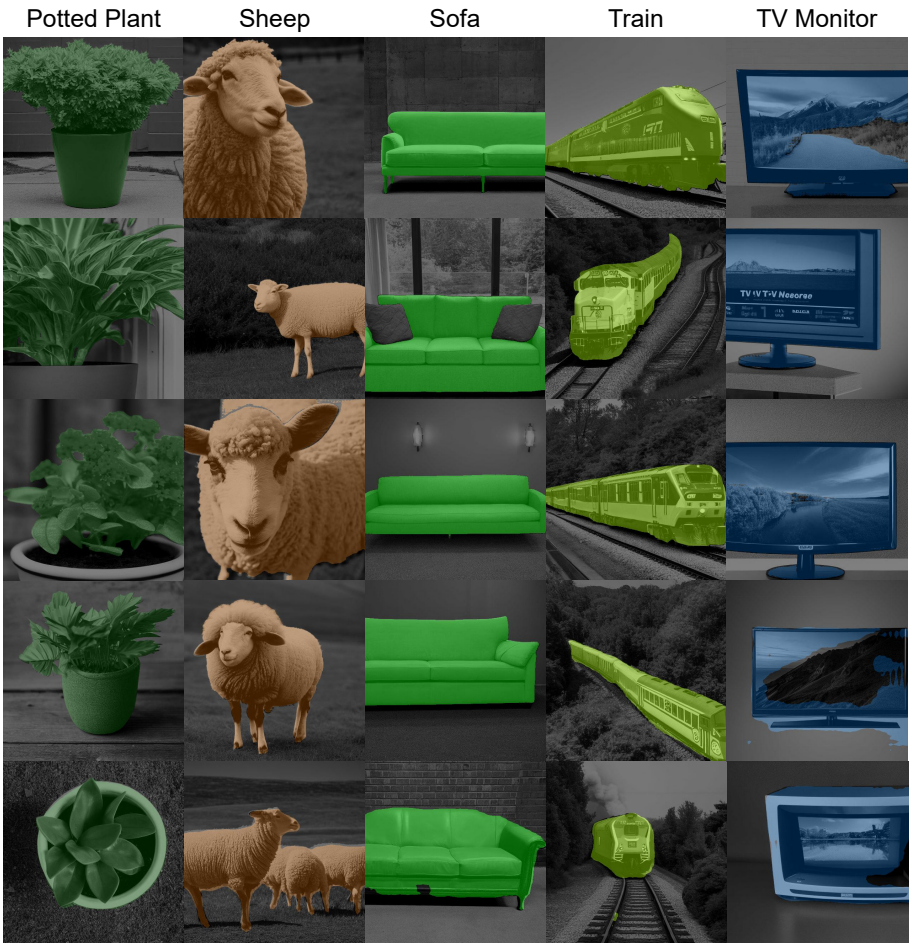

Fig. 4: Qualitative results on the PASCAL-5-3 split using plain text conditioning. Masks are generated by our AlignDiff. Best viewed in color.

10. Hessel, J., Holtzman, A., Forbes, M., Bras, R.L., Choi, Y.: Clipscore: A reference-free evaluation metric for image captioning. arXiv preprint arXiv:2104.08718 (2021) 1, 2
11. Li, X., Wei, T., Chen, Y.P., Tai, Y.W., Tang, C.K.: FSS-1000: A 1000-class dataset for few-shot segmentation. In: CVPR (2020) 1, 2, 3, 4, 5
12. Li, Z., Zhou, Q., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Guiding text-to-image diffusion model towards grounded generation. arXiv preprint arXiv:2301.05221 (2023) 1, 2, 3, 5, 6, 7
13. Min, J., Kang, D., Cho, M.: Hypercorrelation squeeze for few-shot segmentation. In: ICCV (2021) 3, 4, 5
14. Qiu, R.Z., Chen, P., Sun, W., Wang, Y.X., Hauser, K.: GAPS: Few-shot incremental semantic segmentation via guided copy-paste synthesis. In: CVPRW (2023) 1, 3, 4, 5

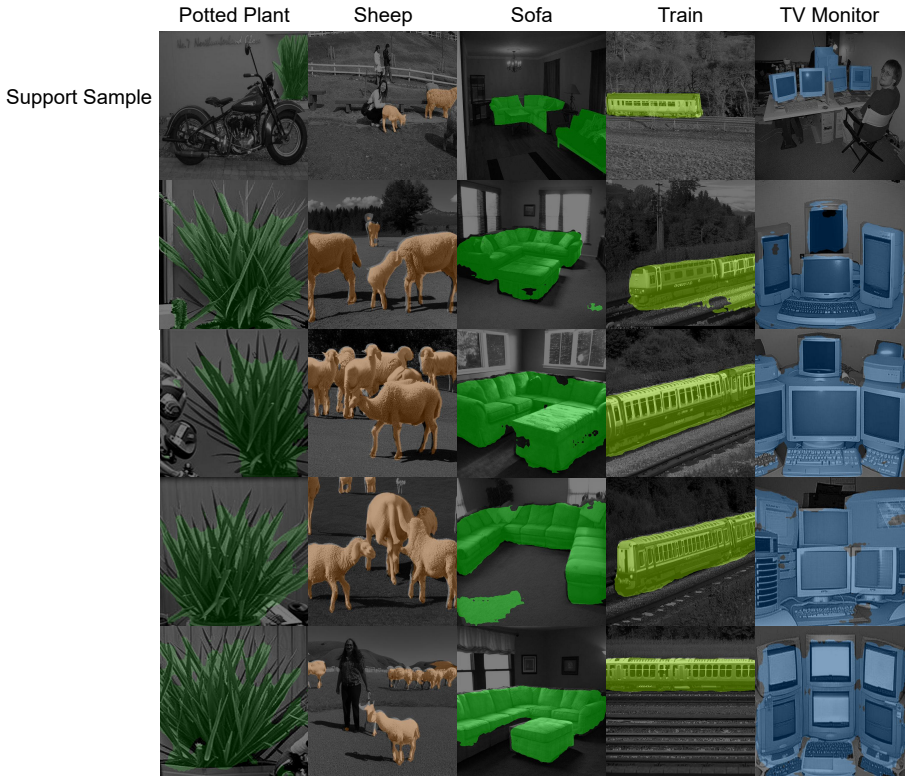

Fig. 5: Qualitative results on the PASCAL-5-3 split using the proposed normalized masked textual inversion method for conditioning. The first row contains the 1-shot image-mask pairs provided to AlignDiff, and the rest of the image-mask pairs are generated by AlignDiff. The figure illustrates two interesting findings: 1) AlignDiff is able to generate relevant images despite the objects of interest may only occupy a small region in the original image, which is not possible with plain textual inversion [6] and 2) Compared to text-conditioned samples in Fig. 3 and Fig. 4, synthesized samples here increase the diversity of training samples (*e.g.,* AlignDiff generates sofas with more realistic viewport and CRT monitors with variations). This helps better capture in-class variation, which is also quantitatively validated in the ablation study in the main paper. Best viewed in color.

15. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 1, 3, 5
16. Wang, K., Liew, J.H., Zou, Y., Zhou, D., Feng, J.: Panet: Few-shot image semantic segmentation with prototype alignment. In: ICCV (2019) 5
17. Wu, W., Zhao, Y., Shou, M.Z., Zhou, H., Shen, C.: Diffumask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models. In: ICCV (2023) 2, 10

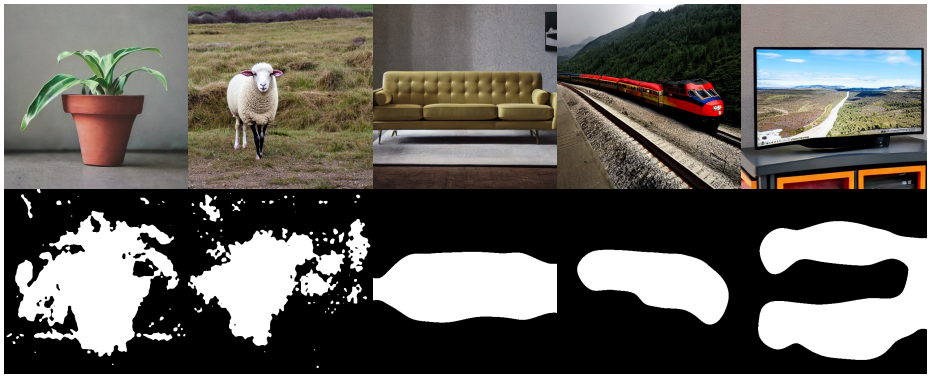

Fig. 6: Coarse masks generated by exploiting cross-attention of Stable Diffusion [9,17].

Fig. 7: Qualitative samples of optimizing adjective embedding v.s. noun cembedding. When the object of interest occupies only a small region in the input image, noun serves as an additional regularizer that retains the semantic concept.

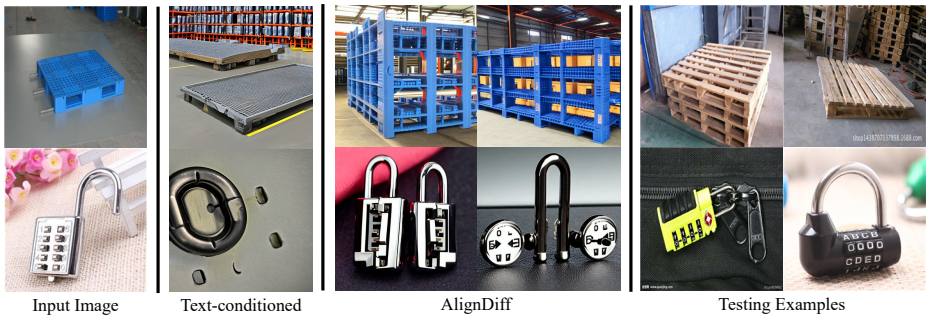

Fig. 8: Qualitative samples of categories where AlignDiff augmentation fail (correspond to red categories in Fig. 2). The augmentation fails if both text-conditioned samples and AlignDiff-generated samples have gaps from the testing distribution. For instance, the first row demonstrates 'warehouse tray', where AlignDiff is given a blue tray and the testing distribution is made of wooden trays. Similarly, the texture and style of combination locks in the second row varies significantly on a sample-by-sample basis.