Supplement Material for Foreground-Background Separation through Concept Distillation from Generative Image Foundation Models

1. Evaluation Strategy

We evaluate the accuracy of our masks by training a plain U-Net on the task of binary classification, following an approach close to the one proposed by [7]. For CUB we use the provided segmentation masks as ground-truth. For all the other datasets, we use the provided bounding boxes. We train the U-Net for 12,000 steps using a batch size of 32 and Adam optimizer with a learning rate 0.001. During training, we crop images randomly to 128×128 pixels and during inference, we employ center-cropping.

2. Finetuning

We fine-tune the diffusion models on the datasets by taking the avenue provided by [6]. Fine-tuning for foreground generation is straightforward by training the model to perform full image synthesis. For background generation, we select a random rectangular patch from the image, exclude any pixels covered by the preliminary mask, and train the model to reproduce the remaining background pixels in the background (see Fig. 1 for examples). The diffusion models are trained on foreground and background generation simultaneously, with each objective being trained in an equal proportion.

Computation of the refined masks, which requires the computation of the preliminary masks, takes 7.5 seconds for a batch of three samples on a single GPU.

3. Empirical Proof of Simplified Equation

In this section, we show empirically that it is not necessary to repeat single diffusion steps in order to achieve better results on preliminary masks. Formally we evaluate

$$\hat{M} = \sum_{t=1}^{T_0} \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z_t}|\mathbf{z_{t+1}},\mathbf{z_0})}[\sum_{l} \psi_{\mathbf{z_{t,l}}}(Q_l, K_l^T)], \quad (1)$$

and show that it can be simplified to

$$\hat{M} = \sum_{t=1}^{T_0} \sum_{l} \psi_{\mathbf{z_{t,l}}}(Q_l, K_l^T) \quad (2)$$

by computing accuracy metrics on CUB for the case of $T_0 = 1$.

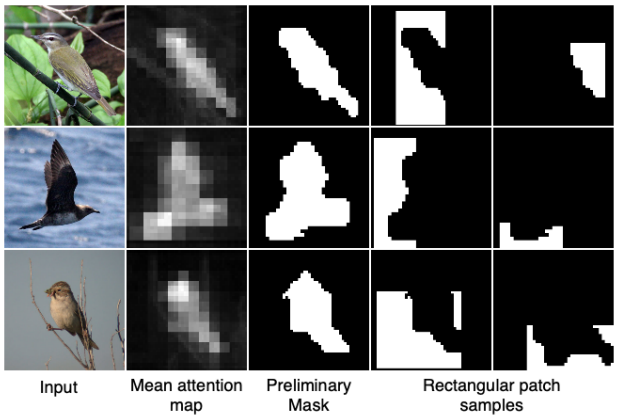

Figure 1. Illustration of how we extract rectangular patches for background inpainting during finetuning.

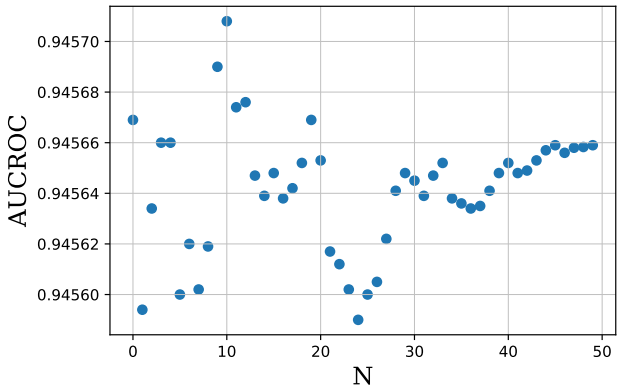

Figure 2. AUCROC of CUB over an increasing number of monte carlo samples N.

The motivation behind this is that the computation of the preliminary masks is the bottleneck of our pipeline, and we want the computation time to remain reasonable. The execution time increases linearly with the number of repetitions. We estimate the expectation in Equation (1) using monte carlo sampling and denote the number of samples as N and compute the AUCROC as a function over it. The results are shown in Fig. 2. These results suggest that increasing N also slightly increases the absolute AUCROC

value, while simultaneously decreasing the variance. However, these improvements are within a very small margin. Intuitively this means that the diffusion model is quite robust towards different latent inputs z_t. We conclude from this that it is unreasonable to compute attention masks over multiple steps and therefore perform all experiments using Eq. (2).

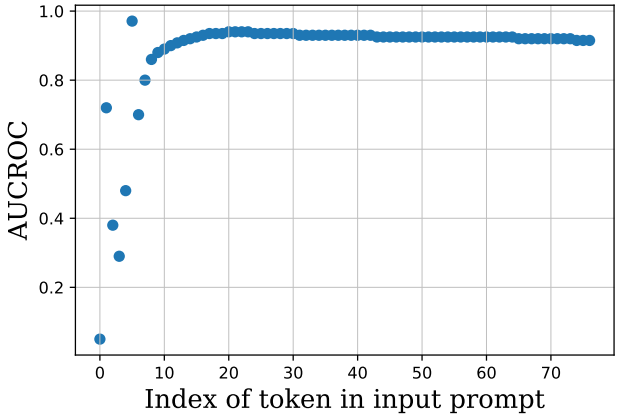

Figure 3. AUCROC values of the preliminary masks extracted for every token using the prompt "a photo of a bird". The language model is inherited from [6] and uses a BERT-tokenizer [3]. Therefore a "startofstring" token is added to the beginning of the input and the length is padded to a length of 77. The peak is at the "bird" token.

Initially, we also experimented using only the object token or using a list of objects as an input prompt but both approaches result in more noisy preliminary masks and inferior AUCROC values.

4. Outlier Cases

Since our method is self-supervised it can be prone to some errors, as the model has never seen an example of a real segmentation. In Fig 4 we show examples of rare selected failure cases that we have observed during testing.

The first is that branches on which birds are sitting are often segmented as part of the bird. This is a result of the preliminary masks which sometimes include these branches into objects. Consequently, the U-Net is uncertain about these areas. We observe that this only happens to birds that are clinging to branches, as can be seen in the left image in Fig. 4. In rare occasions, very low contrast also leads the model to accidentally predict parts of the image as background (such as in the right two bird examples). This is likely because the pixel intensity distribution of the bird is too close to that of the background, causing the image difference when computing the refined masks to be too small, resulting in it being misclassified as background.

In the dog dataset, the method often struggled if humans were holding the dog (see left-most dog example). In these cases, the final segmentation only excludes parts of the human in the background but not all of it. One possible cause of this is that the model struggles to reconstruct the human when performing background inpainting, due to feature complexity, causing it to have a high-intensity difference when computing the refined masks. We observed that this did not happen if the humans were positioned further back in the background, as can be seen in the third example in Figure 4. Finally, the method also struggled with dogs that are only black and white. We believe that this is because of a bimodal pixel distribution assumption. If we perform inpainting for mask refinement for these dogs the white and black parts have very different contrasts compared to the background and are consequently assigned different modes of the bimodal Gaussian mixture model.

In the case of the cars our method seemed to have limited performance if the input image had a plain white background. We believe this is because during the refinement stage the model does not expect the image to be entirely empty, and therefore always tries to inpaint something in the image center. However, the problem of segmenting cars in front of white background can be solved using trivial methods. The predictions for Human3.6m were consistent throughout the whole dataset, with the exception of occasional under-segmentation of the legs, as explained in the main paper.

5. Inpainting Ablation

To verify that our inpainting strategy does not make our proposed pipeline unnecessarily complicated, we try to refine the masks using a simpler approach that crops regions of the background and uses them to inpaint. We do this by extracting the largest background region according to the preliminary masks and then flipping it into the region of the foreground object.

6. Prompt Engineering

To further justify the choice of our text-conditioning y we compute the AUCROC for every token of the prompt "a photo of a bird". Internally this prompt is preceded by a fixed "startofstring" token. The results are shown in Fig. 3. We can clearly observe that the highest response happens if we compute $\hat{M}$ for the token "bird". After that, the AUCROC remains high, albeit not at the same level as before. We decide against incorporating the preliminary maps of different tokens into the pipeline because they slightly decrease the AUCROC value for the preliminary masks while simultaneously reducing the interpretability of our approach. The same holds true for the "startofstring" token, which has very low activation on all the bird pixels. By inverting the attention scores we could therefore also locate objects. However, this observation is a direct consequence

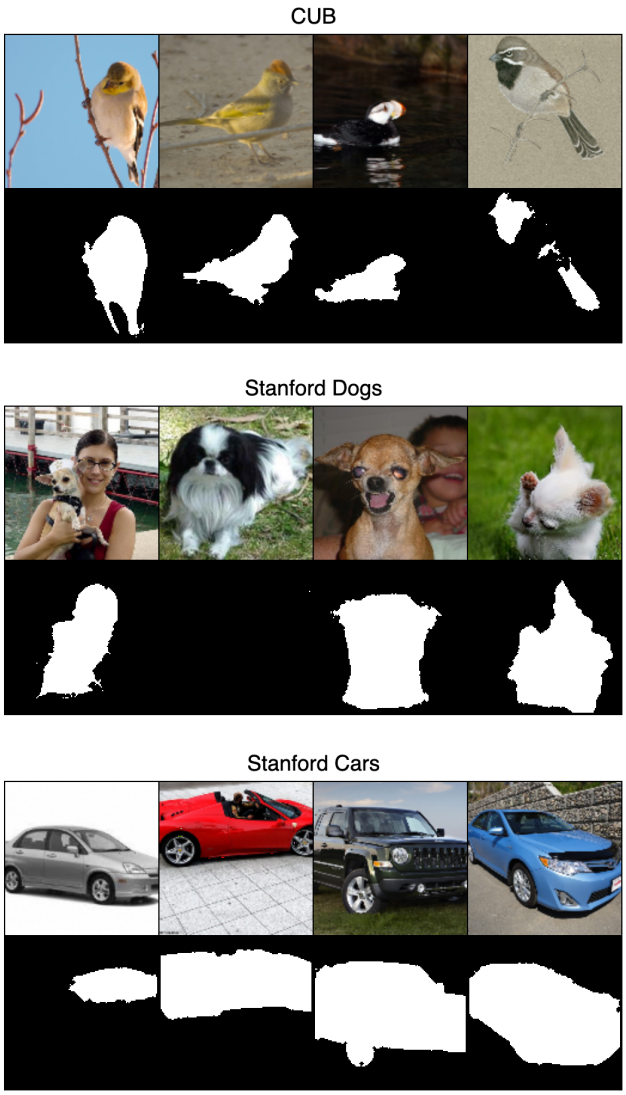

Figure 4. Examples of failure cases of the U-Net model.

of the computation of the attention probabilities. Attention is computed for every pixel as a probability of belonging to a token using softmax normalization on the attention scores. Since this probability has to sum up to one, the activations of non-object pixels have to be high for some tokens. Furthermore, we analyzed the stability of the extracted preliminary masks in terms of minor changes to the input prompt. In Fig. 5, we show the difference of the segmentation masks if we integrate more prior knowledge by describing the image composition in the prompt. From these images, we can see that there are only minor changes to the silhouette of the human and, consequently, that the results are mostly independent of the prompt. We made the same observation when prompting on "person" instead.

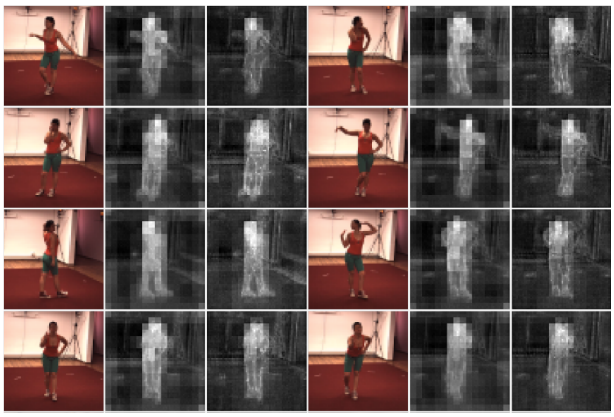

Figure 5. Preliminary masks computed on the prompt "A photo of a human standing in a room" (left) and on the prompt "A photo of a human" (right).

7. Classifier-free Guidance

We use classifier-free guidance as proposed by [4]:

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, y_f) = w\epsilon_\theta(\mathbf{z}_t, y_f) - (w-1)\epsilon_\theta(\mathbf{z}_t, y), \quad (3)$$

where w denotes the classifier-free guidance scale, and ϵ the update term of the diffusion process. In our case, we assume that the latent space representation of the images $\mathbf{z}_t$ conditioned on the prompts is reduced to the background and the foreground clusters. Consequently, we can replace the unconditional prompt with the background prompt from the equation, which changes it to

$$\tilde{\epsilon}_\theta(\mathbf{z}_t, y_f) = w\epsilon_\theta(\mathbf{z}_t, y_f) - (w-1)\epsilon_\theta(\mathbf{z}_t, y_b). \quad (4)$$

Finally, we can also perform classifier-free guidance for background generation by setting the scale to $w = -1$ which is equivalent to switching the prompts and setting $w = 2$.

To verify this, Fig. 6 shows the influence of w in more detail. Images with high guidance towards the background (*i.e.*, low w) do not show any signs of the object. By increasing this value, we can see a bird growing from a part of the image. To further illustrate this process we added a few video samples of this to the supplements. Judging from these images we concluded that our assumption of the clustering is correct and that the model has indeed learned what background information is.

8. Medical Image Analysis

To analyze whether LDMs are interpretable after being adapted to domain-specific tasks, we evaluate our proposed extraction method on an LDM fine-tuned on MIMIC [5] following an approach similar to the one suggested by [2].

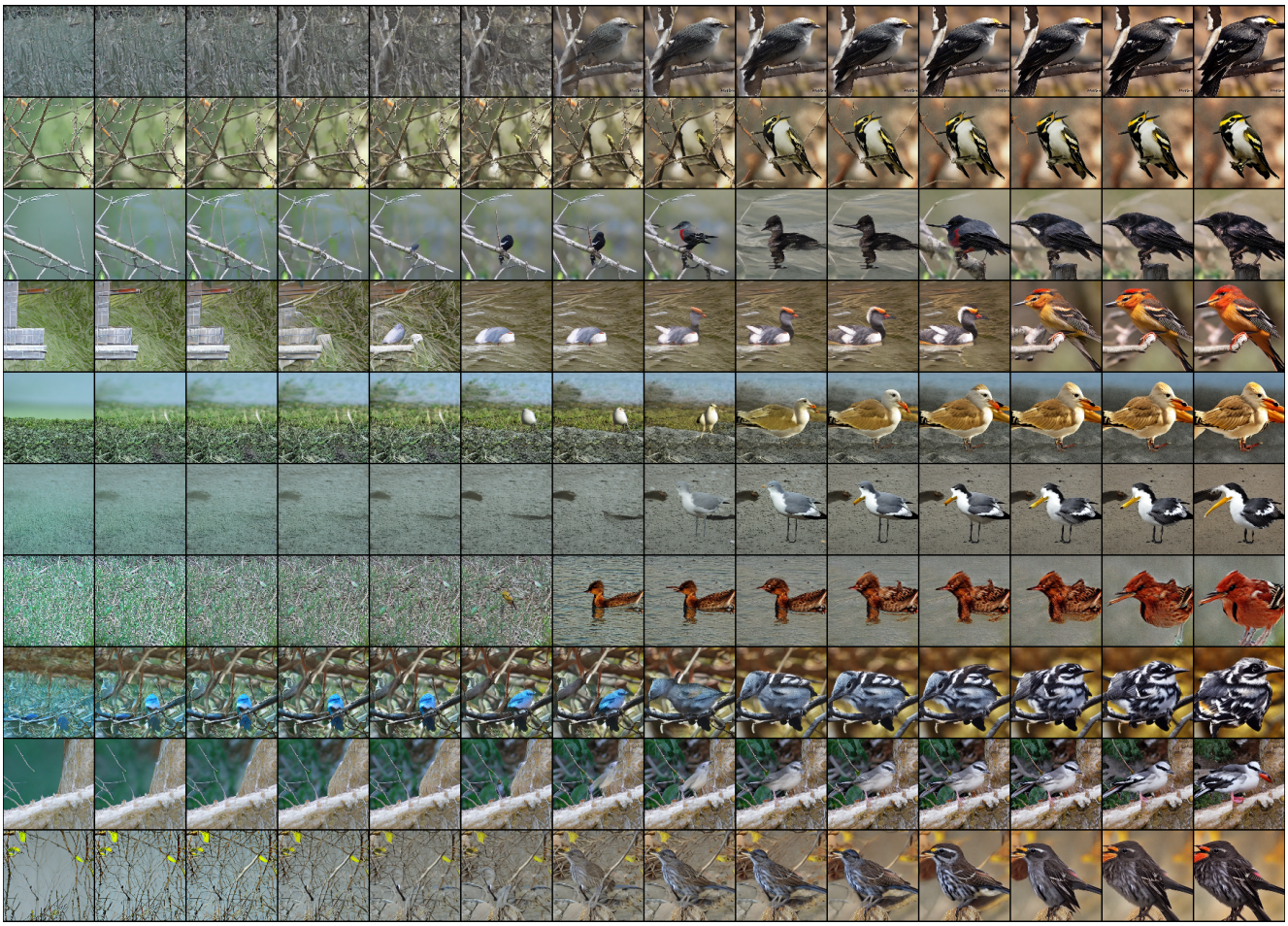

Figure 6. Synthesis results starting from the same seeds while increasing the scale of classifier-free guidance. The guidance scale w ranges from -6.5 to 6.5 and is increased in steps of 1.

Fine-tuning is done for 60k steps over $\sim$160000 images and the *impression* section of the radiology reports corresponding to the images. The learning rate is set to 5×10^{-5}, and the language encoder is kept frozen. We set the batch size during fine-tuning to 256, spread over 16 80GB A100 GPUs during roughly 470 hours of computation. To evaluate the localization accuracy, we take the impressions of the MS-CXR subset [1], which we left as a hold-out set during training. Then, we use the impressions from [1] and compute $\hat{M}$ and M_{pre} on the tokens corresponding to the eight different diseases of the dataset, and compare the predicted region with the ground-truth bounding boxes. Because some words are unknown to the language encoder, they were split into different tokens. In this case, we compute the sum over the attention maps of all tokens.

References

[1] Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, et al. Making the most of text semantics to improve biomedical vision–language processing. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVI*, pages 1–21. Springer, 2022. 4

[2] Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacin, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 2

[4] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[5] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019. 3

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2

[7] Yu Yang, Hakan Bilen, Qiran Zou, Wing Yin Cheung, and Xiangyang Ji. Learning foreground-background segmentation from improved layered gans. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2524–2533, 2022. 1