REVERSE STABLE DIFFUSION: WHAT PROMPT WAS USED TO GENERATE THIS IMAGE? (SUPPLEMENTARY)

Anonymous authors
Paper under double-blind review

ABSTRACT

In the supplementary, we expand our work in several ways. First, we analyze the impact of the diffusion step at which we switch the original U-Net in Stable Diffusion with our model, to generate images faithful to the text prompts. Next, we present another direct application of our framework, namely image captioning. Then, we discuss related work on curriculum learning and compare our approach with other state-of-the-art curriculum learning methods. We continue by showcasing some training examples with various difficulty levels, illustrating that the proposed difficulty scores are correlated with the alignment levels between images and prompts. Then, we present details about the hyperparameters of our method, and report additional results with the DAKL method using various parameter combinations. Finally, we present results with additional neural architectures.

1 WHEN TO REPLACE U-NET IN STABLE DIFFUSION?

We hereby present a qualitative analysis of the impact of the denoising step at which we substitute the original U-Net with ours. To illustrate how this change affects the final samples, we showcase two prompt examples and several images in Figure 1. We vary the number of denoising steps performed by our U-Net. Specifically, we use it to perform the last 49, 40, and 25 denoising steps, respectively. The illustrated examples indicate that the optimal text-to-image alignment is achieved for both prompt examples, when only a single denoising step is performed with the original U-Net. However, when we make the switch in the later stages of the denoising diffusion process, the impact on the final image becomes less meaningful. In the first example, when we introduce the model after the first 10 steps, the output is still aligned with the text, but for the second example, the horse is removed when performing 10 steps with the original model. Overall, we conclude that the first part of the denoising process has the highest impact on the content of the final image. When we switch the model in the second half of the denoising process, the results are very similar to the case when we use only the original U-Net. Based on these observations, we decided to replace the original U-Net with our own right after the first sampling step in our application to image generation discussed in the main article.

2 QUALITATIVE EVALUATION OF THE NOVEL COMPONENTS ADDED TO U-NET

To study the effect of our novel components on the alignment of images generated by the fine-tune U-Net, we generated several images with the fine-tuned U-Net, before and after introducing our extra techniques, namely the multi-label classification head and the curriculum learning strategy. We present the images in Figure 2. The figure shows several cases where the enhanced version of U-Net produces images that are better aligned with the prompts. For example, in the right-hand side example from the first row, the vanilla fine-tuned U-Net does not generate the man. Similarly, in the left-hand side example from the second row, it does not generate the hat on the bird's head. The vanilla model also fails to generate the female android in the right-hand side example from the second row. In all these cases, our enhanced version of U-Net generates the objects referred in the prompts,

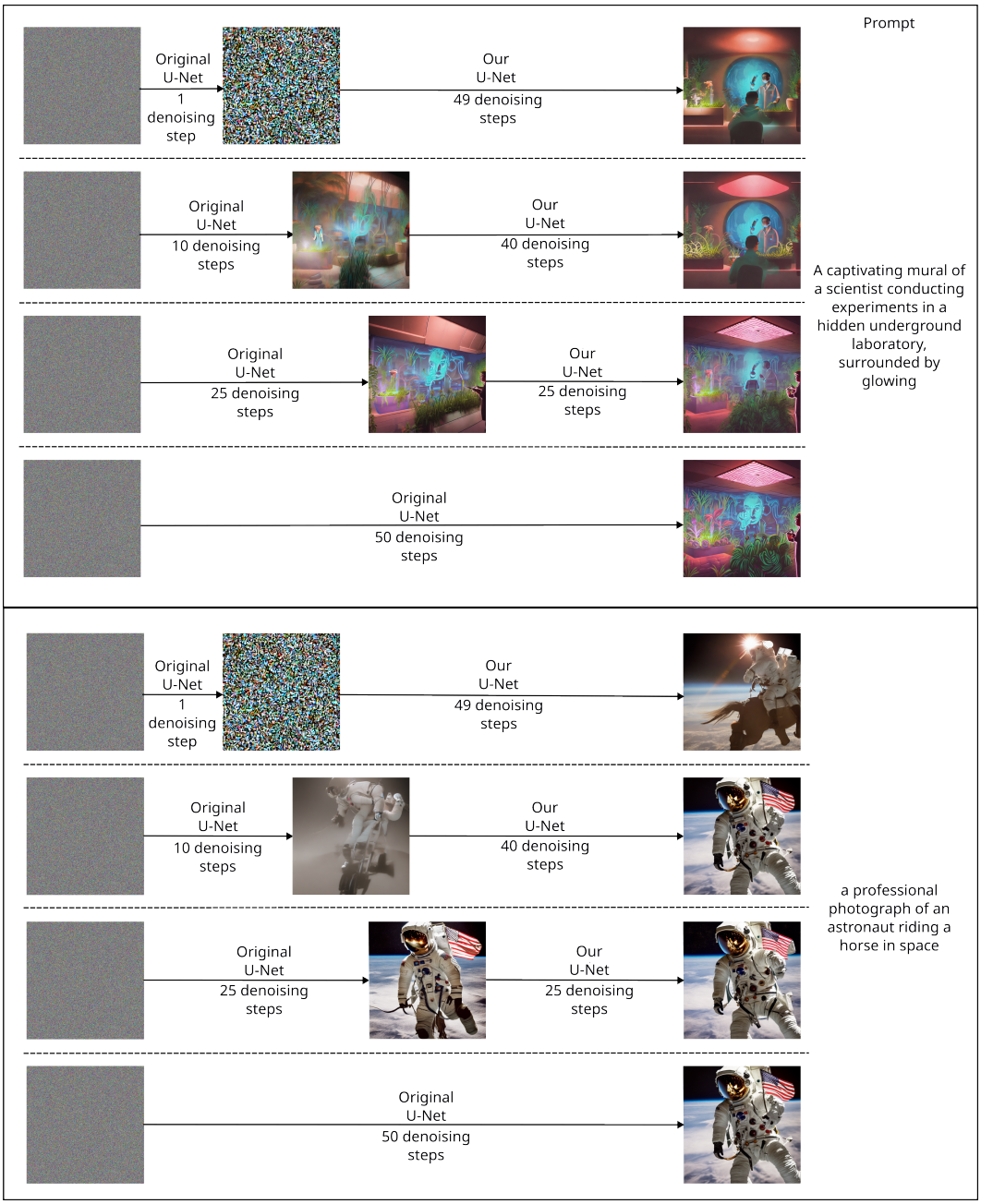

Figure 1: Examples that demonstrate the impact of replacing the original U-Net in Stable Diffusion with our U-Net on text prompt generation, at various steps of the denoising diffusion process.

confirming the benefits of adding our novel components. Note that, in the main paper, we showed quantitative results demonstrating that the enhanced U-Net achieves higher cosine similarity scores. Hence, the qualitative results shown in Figure 2 are consistent with the quantitative results from the main paper.

3 Application to Generated Image Captioning

In the main paper, we studied the task of prompt embedding prediction, which is a well-posed reverse engineering task. Predicting the actual prompt is also possible, but we

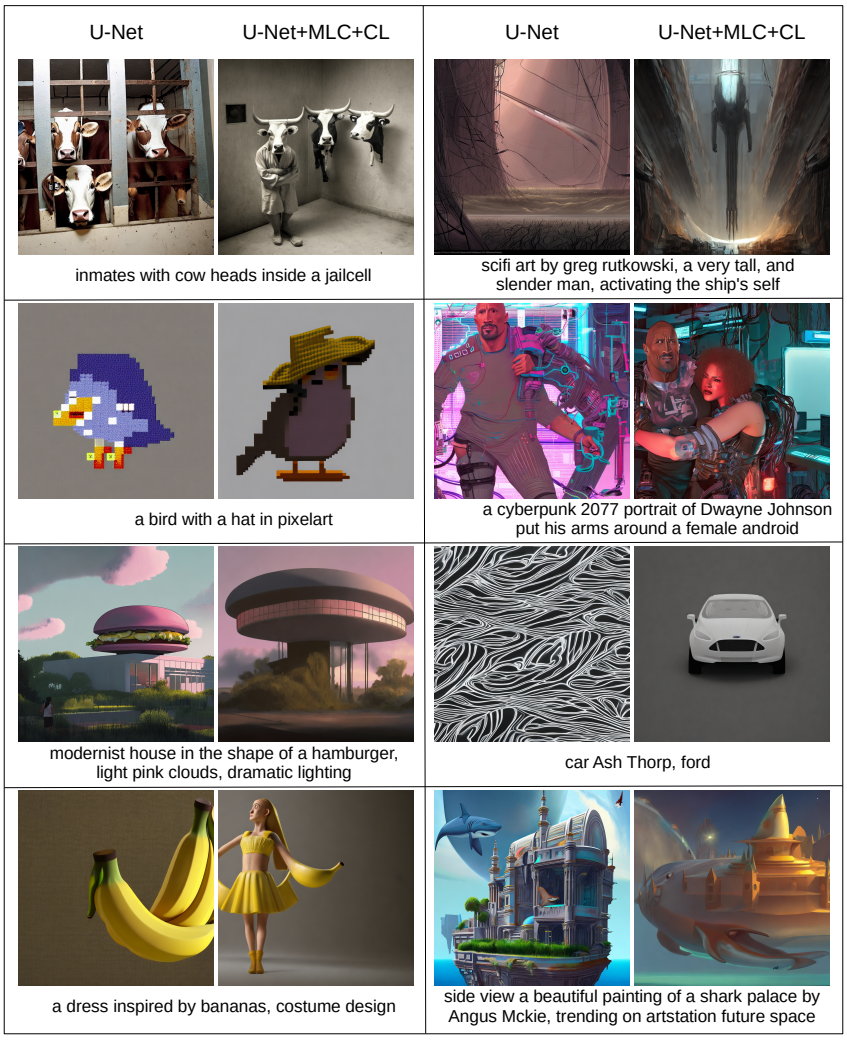

Figure 2: Images generated by the fine-tuned U-Net, before and after adding our novel components, namely the multi-label classification (MLC) head and the curriculum learning (CL) strategy. Best viewed in color.

consider this task ill-posed. This is because we have only one prompt for each generated image, while image captioning benchmarks typically have several alternative ground-truth captions for each image (Chen et al., 2015), and models are evaluated against the best matching ground-truth caption. Hence, predicting the exact prompt is significantly more difficult than predicting the prompt embedding. In the embedding space, we essentially allow the models to predict semantically related prompts without being needlessly penalized.

Nevertheless, we next study two approaches to apply our framework on generated image captioning. Given a generated image, one naive approach is to assign a caption via a 1-nearest neighbor (1-NN) model applied on the prompt embedding space learned by our joint framework. We obtain a prompt embedding for the query image and compare the resulting embedding with every embedding in the training set, keeping the most similar one in terms of the cosine similarity. This approach does not involve any fine-tuning on prompt generation. A more sophisticated approach is to employ our framework based on curriculum learning (CL) and an extra multi-label classification (MLC) head to fine-tune the BLIP model on the image captioning task.

We qualitatively compare our approaches with a fine-tuned vanilla BLIP. We display a set of representative samples in Figure 3. Our 1-NN method showcases comparable results

Ground truth. joe biden riding an elephant
1-NN. an elephant
BLIP. elephants are walking down the road with a man riding on the back of a cart, a digital rendering, magic realism, joe biden riding a dragon, emperor biden.
BLIP MLC+CL. joe biden riding an elephant.

Ground truth. Field of skeletons, award winning horror photography.
1-NN. barrel of skeletons, horror scene, detailed, photorealistic
BLIP. arafed skeleton standing in a field of bones with a man in the middle, digital art, surrealism, with a skeleton army, skeletal omens, all skeletons
BLIP MLC+CL. a man standing in front of a bunch of skeletons, behance contest winner, gothic art.

Ground truth. Sky full of travel balloon and stars.
1-NN. hot air balloon flying through the sky, double rainbow, illustration by dr seuss, oh! the places you'll go, watercolor
BLIP. there are two hot air balloons flying in the sky with stars, a storybook illustration, naive art, hot air balloons, floating among stars
BLIP MLC+CL. A hot air balloon floating in the sky, a storybook illustration by chiho aoshima, behance contest winner, space art, storybook illustration.

Ground truth. Harry Maguire as a Brighton soccer player
1-NN. football player in the style of Michaelangelo
BLIP. there are two men playing soccer on a field with a crowd watching, a picture, renaissance, tyrell wellick, max prentis, tom burke
BLIP MLC+CL. Harry Kane playing for Reading FC.

Ground truth. tall, security guard checks the bags of a worried looking woman, art, satire.
1-NN. criminal sketch of a clown.
BLIP. cartoon of a man and woman standing in a doorway with a bag, a cartoon, net art, an illustration, hotel room, film still from an cartoon
BLIP MLC+CL. a man in a suit and tie standing in front of a woman in a dress, a comic book panel by david cronenberg, tumblr contest winner, gothic art.

Ground truth. photograph of a ginger male, middle aged balding superhero.
1-NN. portrait photo of a yellow - skinned bald wrinkly man with stubble and big eyes, he wears blue overalls no shirt, he looks like a human minion, moody cinematic lighting, realistic facial features, hyper detailed, crisp image.
BLIP. there is a man with a beard and a white shirt, a portrait, renaissance, red beard, reddish beard, acting headshot
BLIP MLC+CL. a balding man with a red beard, wearing a white shirt, studio lighting.

Ground truth. handsome Danny DeVito in an Iron Man suit, western, D&D, fantasy, intricate, elegant, highly detailed
1-NN. Iron Man Giant Armor, pencil sketch, concept art.
BLIP. a close up of a cartoon iron man in a suit, vector art, process art, iron man, superior iron man, ironman
BLIP MLC+CL. iron man, trending on artstation.

Ground truth. A Landscape by Charles Addams and salvador dali
1-NN. geometric painting of tomorrowland by hieronymus bosch
BLIP. painting of a landscape with a tree, a yellow apple, and a dog, a surrealist painting, surrealism, surrealist landscape, surrealist landscape painting, painting of a dreamscape
BLIP MLC+CL. a landscape with trees, a storybook illustration by charles addams, behance contest winner, american scene painting, storybook illustration.

Figure 3: Examples of captions for generated images. We compare the prompts returned by a fine-tuned vanilla BLIP with those of a 1-NN model applied on pre-trained prompt embeddings, and those of an enhanced version of BLIP based on multi-label classification (MLC) and curriculum learning (CL). Best viewed in color.

with those of BLIP (Li et al., 2022). For the less complex images, a matching prompt is usually found by our nearest neighbor approach. The predicted prompts for the harder examples are still representative and depict certain nuances of the text, but they often fail to precisely describe all aspects of the input images. A similar behavior is often observed for the vanilla BLIP. Our version of BLIP (based on MLC+CL) produces improved prompts in a number of cases, *e.g.* the first image on the first column, the second image on the second column, or the last image on the second column. Although the prompts of our best model are representative, they are still far from the ground-truth prompts, suggesting that the generated image captioning task is indeed ill-posed. A representative ill-posed case is the third image on the second column, depicting Iron Man, where it is impossible to predict the prompt, as Danny DeVito is hidden by the Iron Man suit. Indeed, there is no visual clue to indicate the presence of Danny DeVito in the source prompt.

We further perform a quantitative analysis on the entire test set to compare our approaches with vanilla BLIP. As evaluation metric, we use the reference-augmented version of the recently proposed CLIPScore (Hessel et al., 2021), namely the RefCLIPScore. The corresponding results are presented in Table 1. The BLIP model is a state-of-the-art captioning model that obtains a RefCLIPScore of 30.93. Although our 1-NN approach is not particularly tuned for image captioning, it produces competitive captioning results, reaching a RefCLIPScore of 25.53. The proposed version of BLIP based on multi-label classification and curriculum learning yields superior results, increasing the RefCLIPScore of BLIP from 30.93 to 31.88. These results show that our framework can improve image captioning results, thus extending its applicability from prompt embedding generation to generated image captioning.

Table 1: Comparison between a fine-tuned vanilla BLIP and the proposed image captioning approaches. Our first approach is based on a 1-NN applied on pre-trained prompt embeddings. Our second approach is an enhanced version of BLIP, which integrates our multi-label classification head and curriculum learning strategy. The best score is highlighted in bold.

Model	Multi-Label Classification	Curriculum Learning	RefCLIPScore
1-NN (ours)	✓	✓	25.53
BLIP	-	-	30.93
BLIP (ours)	✓	-	31.77
BLIP (ours)	✓	✓	**31.88**

Table 2: Comparison between our curriculum learning method and two state-of-the-art curriculum learning approaches, CBS (Sinha et al., 2020) and LeRaC (Croitoru et al., 2022). The best score is highlighted in bold.

Backbone	Curriculum learning method	Cosine similarity
ViT	No curriculum	0.6526
	LeRaC (Croitoru et al., 2022)	0.6521
	CBS (Sinha et al., 2020)	0.6254
	Ours	**0.6544**

4 Related Work on Curriculum Learning

Since we employ a novel curriculum learning regime to boost the performance of the studied models, we can also consider work on curriculum learning as related. The research community has extensively utilized this learning paradigm across a range of domains, including both computer vision (Bengio et al., 2009; Croitoru et al., 2022; Ionescu et al., 2016; Shi & Ferrari, 2016; Soviany et al., 2021; Chen & Gupta, 2015; Sinha et al., 2020; Zhang et al., 2021a) and natural language processing (Croitoru et al., 2022; Liu et al., 2018; Platanios et al., 2019). However, given the unique nature of each application, distinct data organization approaches have been developed to ensure optimal results. For example, in vision, the number of objects in the image is one criterion (Soviany et al., 2021; Shi & Ferrari, 2016), while, in natural language processing, both word frequency (Liu et al., 2018) and sequence length (Kocmi & Bojar, 2017; Tay et al., 2019; Zhang et al., 2021b) are utilized. Other contributions tried to avoid estimating sample difficulty by implementing curriculum learning on the model itself (Karras et al., 2018; Sinha et al., 2020; Croitoru et al., 2022), or by selecting the samples dynamically, based on the performance of the model (Kumar et al., 2010; Jiang et al., 2015). Different from related approaches based on ordering data samples according to their difficulty (Bengio et al., 2009; Soviany et al., 2021; Shi & Ferrari, 2016), we propose to employ a novel approach to assess the difficulty level. More specifically, we utilize the mean cosine similarity between the prompt embedding produced by the model and the ground-truth embedding vector, measured at various stages of the standard training process.

5 Comparison with Other Curriculum Learning Methods

To assess the performance of our proposed curriculum learning technique, we compare it with other curriculum learning methods from the recent literature. We choose two state-of-the-art methods, namely Curriculum by Smoothing (CBS) (Sinha et al., 2020) and Learning Rate Curriculum (LeRaC) (Croitoru et al., 2022). Each of these methods have additional hyperparameters, which we tried to carefully tune via grid search. Unfortunately, we did not manage to make them surpass the performance of the vanilla training regime. As shown in Table 2, the results demonstrate the net superiority of our curriculum learning method in image-to-prompt generation with the ViT backbone. The success of our technique relies on harnessing the misalignment level between the image and the prompt in each training pair to create the curriculum schedule. In contrast, CBS and LeRaC do not take the misalignment into account. We believe this explains why our results are better.

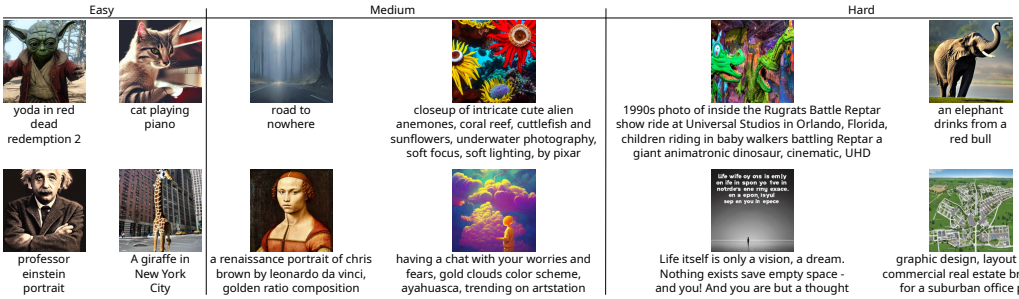

Figure 4: Examples labeled as easy, medium and hard by our difficulty estimation procedure based on monitoring the cosine similarity of samples during conventional training. Best viewed in color.

Table 3: Varying the number of k-means clusters r and the parameter γ of the RBF kernel used in DAKL.

DAKL	r	γ	Cosine similarity
-	-	-	0.6900
✓	1,000	0.001	0.6899
✓	5,000	0.001	0.6905
✓	10,000	0.0001	0.6898
✓	10,000	0.001	0.6917
✓	10,000	0.01	0.6909
✓	10,000	0.1	0.6878

6 Qualitative Analysis of Proposed Difficulty Score

In Figure 4, we illustrate some examples showing how our method categorizes image-prompt pairs based on their difficulty (easy, medium, or hard). The easy samples contain short straight-forward descriptions, which are very well aligned with the generated images. The medium examples involve descriptions that produce abstract images, or descriptions that require rich creativity. Finally, the hard samples consist of prompts that cannot be associated with a visual representation, *e.g.* quotes, or very complex prompts with multiple, especially unreal, elements. These examples indicate that our difficulty scores are easy to interpret visually, suggesting a reasonable organization of the easy, medium and hard example batches, which is correlated with the alignment between images and prompts.

7 Hyperparameter Tuning

We establish the hyperparameters during preliminary experiments on the validation set. Due to the large scale of the training set, we train all models for three epochs. The image resolution is either 224×224 or 256×256, depending on the model, while the mini-batch size is set to 64. The models are optimized with Adam. We set the learning rate to 10^{-4} for ViT / Swin models, and $5 \cdot 10^{-6}$ for CLIP-based models. With the introduction of the classification head, there are two extra hyperparameters. These are the weight λ of the additional classification loss and the size of the vocabulary m. We set $\lambda = 0.1$ and $m = 1000$. For DAKL, we set the number of k-means clusters to 10K and $\gamma = 0.001$. To foster future research and allow others to fully reproduce our results, we release our code as open source.

To provide a more comprehensive overview of the hyperparameters of our DAKL method, we present additional results by varying the parameter σ of the RBF transformation. Additionally, we explore various choices for the number of centroids r used by the k-means algorithm. We present the corresponding empirical results in Table 3. There are multiple hyperparameter combinations surpassing the baseline, but the best results are obtained for $r = 10,000$ clusters and $\gamma = 0.001$.

Table 4: Comparison between several neural architectures, which are divided into two categories: black box (■) and white box (□). Black-box models do not have access to the weights of Stable Diffusion. In contrast, the white-box model starts the fine-tuning process from the weights of Stable Diffusion.

Image Encoder	Type	Cosine similarity
CLIP-Huge + k-NN	■	0.6189
BLIP	■	0.5129
CLIP-Huge	■	0.6725
Swin-L	■	0.6624
ViT	■	0.6526
U-Net$_{enc}$	□	0.6130

Table 5: Cosine similarity scores between predicted and ground-truth prompt embeddings, while employing different combinations of neural architectures. Individual models are compared with combinations of two, three and four models. The top score is highlighted in bold.

#Models	CLIP-Huge	U-Net$_{enc}$	Swin-L	ViT	Cosine similarity
1	✓	-	-	-	0.6750
	-	✓	-	-	0.6497
	-	-	✓	-	0.6671
	-	-	-	✓	0.6550
2	✓	✓	-	-	0.6785
	✓	-	✓	-	0.6887
	✓	-	-	✓	0.6854
	-	✓	✓	-	0.6787
	-	✓	-	✓	0.6732
	-	-	✓	✓	0.6792
3	✓	✓	✓	-	0.6900
	✓	✓	-	✓	0.6901
	✓	-	✓	✓	0.6901
	-	✓	✓	✓	0.6820
4	✓	✓	✓	✓	**0.6917**

8 Results with Other Tested Models

Additional baselines. In the main paper, we reported results with three models having no knowledge about the internals of Stable Diffusion, treating the diffusion model as a black box. We also employed the U-Net encoder from Stable Diffusion, which comes with the pre-trained weights of the diffusion model. Hence, we consider the approach based on U-Net as a white-box method. As underlying models, we initially considered two more black-box architectures. The first one is a k-nearest neighbors (k-NN) algorithm, applied in a regression setting. Leveraging the power of a fine-tuned CLIP to match the image and text representations, the embedding of the image is compared to all the embeddings obtained from reference training prompts. Then, based on the distance to the closest neighbors, the output embedding (in the sentence transformer space) is regressed as a weighted mean. The second baseline is represented by a BLIP model (Li et al., 2022), a recent approach with state-of-the-art results in image captioning, which is fine-tuned on our task.

Results. In Table 4, we compare the four models included in the main paper with the two additional models. For a fair comparison, all models are trained with the vanilla training procedure. We emphasize that the three black-box models (CLIP-Huge, Swin-L, and ViT) chosen as underlying architectures for our novel training framework are the most competitive ones. Hence, increasing the performance levels of these models by employing our learning framework is more challenging. This is why our learning framework exhibits the highest performance boost for the U-Net encoder, which starts from a lower average cosine similarity compared to the top three black-box models.

Another interesting observation is that the white-box U-Net is not necessarily the best model. Indeed, the privilege of having access to the weights of the Stable Diffusion model

seems to fade out in front of very deep architectures, such as Swin-L and CLIP-Huge, that benefit from large-scale pre-training.

9 Ablating the Ensemble

Since our main goal is to assess how well the text embeddings can be recovered from generated images, we motivate our use of an ensemble of multiple models via the focus on minimizing the possibility of reporting low cosine similarity scores due to a poor model choice for the reverse task. In Table 1 from the main paper, there is a noticeable gap between the individual models and the ensemble. We thus believe the scores reported for the ensemble better reflect the misalignment of the original Stable Diffusion model. To better motivate the proposed combination of models, we conduct additional experiments with various ablated combinations of models. The corresponding results are shown in Table 5. We observe that combining every two models leads to better results than using the individual counterparts. Further performance gains are obtained by combining every three models. Still, the top cosine similarity is reached when we combine all four models. In conclusion, our results clearly indicate that all individual models contribute to improving the proposed ensemble.

Furthermore, we observe that the gains saturate each time we increase the number of models in the ensemble. This suggests that adding even more models would generate marginal gains. Thus, we limit ourselves to using the proposed ensemble based on four models.

References

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of ICML*, pp. 41–48, 2009. 5

Xinlei Chen and Abhinav Gupta. Webly Supervised Learning of Convolutional Networks. In *Proceedings of ICCV*, pp. 1431–1439, 2015. 5

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 3

Florinel-Alin Croitoru, Nicolae-Catalin Ristea, Radu Tudor Ionescu, and Nicu Sebe. LeRaC: Learning Rate Curriculum. *arXiv preprint arXiv:2205.09180*, 2022. 5

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In *Proceedings of EMNLP*, pp. 7514–7528, 2021. 4

Radu Tudor Ionescu, Bogdan Alexe, Marius Leordeanu, Marius Popescu, Dim Papadopoulos, and Vittorio Ferrari. How hard can it be? Estimating the difficulty of visual search in an image. In *Proceedings of CVPR*, pp. 2157–2166, 2016. 5

Lu Jiang, Deyu Meng, Qian Zhao, Shiguang Shan, and Alexander G. Hauptmann. Self-Paced Curriculum Learning. In *Proceedings of AAAI*, pp. 2694–2700, 2015. 5

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of ICLR*, 2018. 5

Tom Kocmi and Ondřej Bojar. Curriculum Learning and Minibatch Bucketing in Neural Machine Translation. In *Proceedings of RANLP*, pp. 379–386, 2017. 5

M. Kumar, Benjamin Packer, and Daphne Koller. Self-Paced Learning for Latent Variable Models. In *Proceedings of NIPS*, volume 23, pp. 1189–1197, 2010. 5

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of ICML*, 2022. 4, 7

Cao Liu, Shizhu He, Kang Liu, and Jun Zhao. Curriculum Learning for Natural Answer Generation. In *Proceedings of IJCAI*, pp. 4223–4229, 2018. 5

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. Competence-based curriculum learning for neural machine translation. In *Proceedings of NAACL*, pp. 1162–1172, 2019. 5

Miaojing Shi and Vittorio Ferrari. Weakly Supervised Object Localization Using Size Estimates. In *Proceedings of ECCV*, pp. 105–121, 2016. 5

Samarth Sinha, Animesh Garg, and Hugo Larochelle. Curriculum by smoothing. In *Proceedings of NeurIPS*, pp. 21653–21664, 2020. 5

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum self-paced learning for cross-domain object detection. *Computer Vision and Image Understanding*, 204:103–166, 2021. 5

Yi Tay, Shuohang Wang, Anh Tuan Luu, Jie Fu, Minh C. Phan, Xingdi Yuan, Jinfeng Rao, Siu Cheung Hui, and Aston Zhang. Simple and Effective Curriculum Pointer-Generator Networks for Reading Comprehension over Long Narratives. In *Proceedings of ACL*, pp. 4922–4931, 2019. 5

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. FlexMatch: Boosting Semi-Supervised Learning with Curriculum Pseudo Labeling. In *Proceedings of NeurIPS*, pp. 18408–18419, 2021a. 5

Wei Zhang, Wei Wei, Wen Wang, Lingling Jin, and Zheng Cao. Reducing BERT Computation by Padding Removal and Curriculum Learning. In *Proceedings of ISPASS*, pp. 90–92, 2021b. 5