Dataset Augmentation by Mixing Visual Concepts Supplementary Material

Abdullah Al Rahat, Hemanth Venkateswara
Georgia State University, USA
mkutubi1@student.gsu.edu, hvenkateswara@gsu.edu

In this supplemental material, we provide further explanations of the experiments, and more qualitative results followed by additional ablative studies, limitations, and future directions.

1. More Experimental details

Stable Diffusion. Stable Diffusion (SD) represents a state-of-the-art text-to-image model capable of generating photo-realistic images from textual prompts. Operating as a conditional diffusion model, SD gradually removes noise from a noisy sample, transforming it into a realistic image that reflects the input text. This process iteratively refines the image's latent representation to match the visual context described by the text.

The model functions within a latent space framework, where images are initially embedded using a Variational Autoencoder (VAE) [6]. A Noise Scheduler [5] introduces noise to the latent image representation, and a U-Net [9] is trained to predict and remove this noise. During inference, the Noise Scheduler and U-Net collaborate to progressively refine the latent space representation, culminating in the generation of denoised images in the pixel space.

In our approach, we leverage the SD model as a foundational text-to-image generator. We adapt the input layer of SD's U-Net to accommodate images from the source dataset and a prompt as conditioning inputs. This prompt combines a caption generated by BLIP-2 [7], an image captioning model, with a class label prompt such as "a photo of class type class label". The inclusion of the class label prompt addresses potential issues with missing or inaccurate class labels from the captioning model.

Training Details. For fine-tuning the SD (Stable Diffusion) model, we utilize an end-to-end training approach inspired by InstructPixtoPix [1] on our curated dataset. Here are the specific details: **Training Approach:** The training is conducted using conditioning by images and prompts, likely following the methodology outlined in InstructPixtoPix. **Hyperparameters:** Learning Rate(LR): 0.0001, Optimizer: Adam optimizer with default settings, Batch Size: Chosen as 50. **CIFAR-10 and CIFAR-100:** Training commence from scratch. Hyperparameters are set following methodologies akin to AutoAugment [3], FastAutoAugment [8], and RangAugment [4] specifically tailored for CIFAR-10 and CIFAR-100. **Caltech101:** LR: 0.0001, Optimizer: Adam optimizer with default settings, Batch Size: Chosen as 50. **Brain Tumor Dataset:** Different hyperparameter settings are chosen for various classification models. For RestNet50 and WideRestNet-50-2, LR:0.001, ADAM optimizer, batch size:50. For EfficientNetb0, LR:0.02, Optimizer: SGD with Multistep Scheduler and gamma=0.1. These settings are chosen to optimize the fine-tuning process for the SD model, aiming to enhance its performance in generating diverse and high-quality outputs aligned with the conditioning inputs and prompts provided during training

Dataset Preparation. We rearrange the images in the dataset by randomly selecting two images with the same class label and pairing them together. For instance, suppose we have five images $(I_0, I_1, I_2, I_3, \text{and}, I_4)$ in class A. We randomly select two images from the set and pair them using the indices as follows: $(0, 1), (0, 4), (4, 0), (3, 4), (1, 2), \ldots$. We can select as many pairs as we like. The main objective of this pairing is to use the first image of the pair as an input image (i.e., the condition image) and generate the second image using the Diffusion model. During training, the goal is to minimize the Mean Squared Error (MSE) loss between the generated sample and the second image in the pair. In our case, we create two separate random number lists, each containing several items equal to the number of images in each class in the dataset. These lists are then paired one-to-one and only the index pairs are stored rather than the image pairs. This approach keeps the storage requirements for the images from increasing. During training, a mini-batch of index pairs is created, and only the images corresponding to those indices are accessed.

Additional images were generated using our MVC method for dataset augmentation. Figures 1, 2, and 3 compare real images with images showing coarse-grained

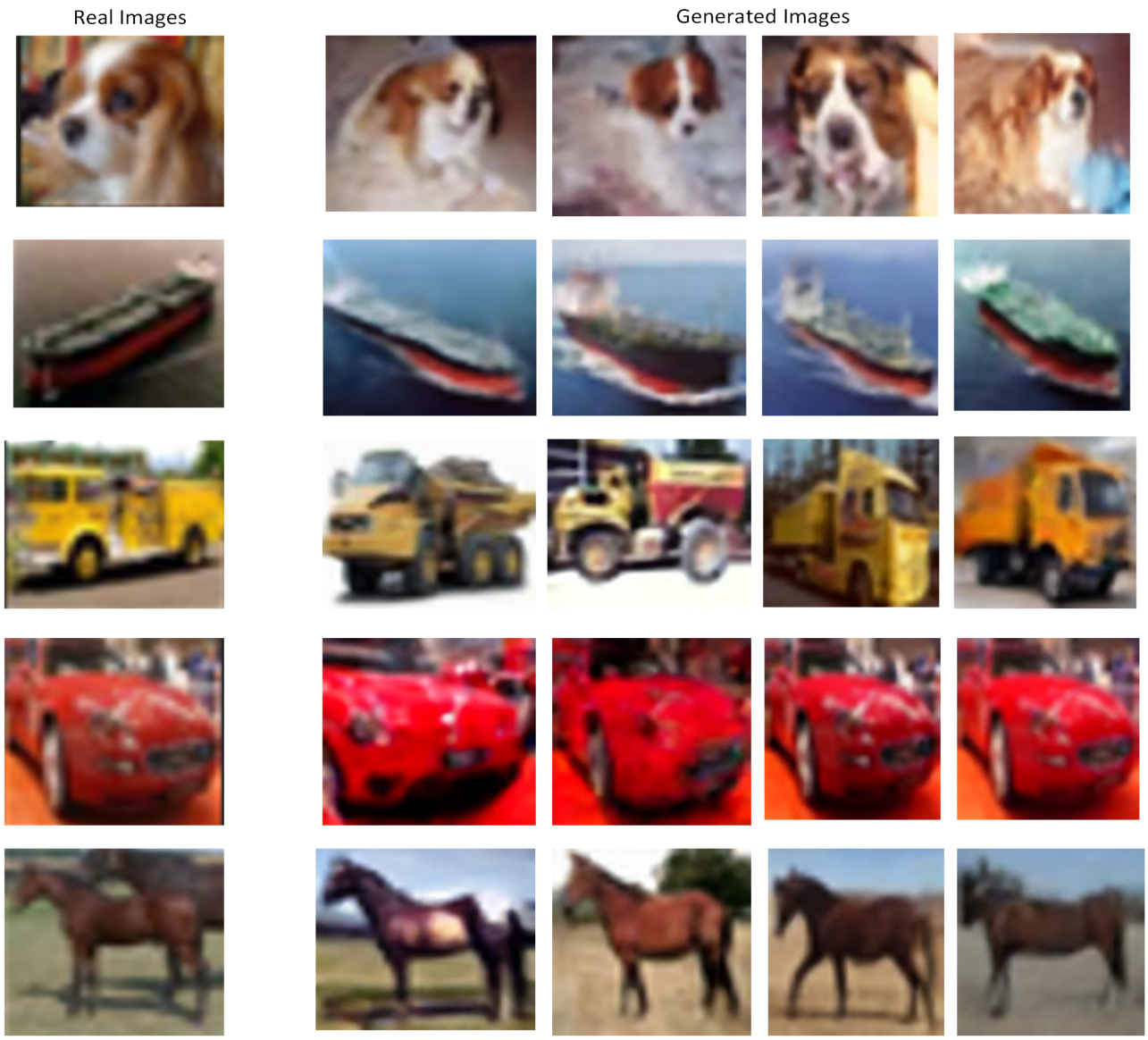

Figure 1. Column 1 depicts real images from CIFAR-10. In the remaining columns, the first two columns show mostly coarse-grained change and the next two columns mostly show fine-grained change.

and fine-grained changes from CIFAR-10, CIFAR-100, and Brain Tumor Datasets, respectively.

2. Details on Ablation Study

Challenges with Pretrained Diffusion Models. Pretrained diffusion models often produce samples from the captions of training images that do not align well with the domain of the train-test dataset. This issue is particularly noticeable in fine-grained datasets such as Caltech101 and medical imaging datasets. The problem is likely due to the training dataset of the diffusion model not adequately representing fine-grained and medical images. Consequently, when used in the medical domain, the pretrained diffusion models produce less effective and lower-quality samples. This limitation is a result of the model's training biases towards non-medical imagery, which hinders its ability to create accurate representations of medical images. Figure 1 in main paper illustrates this phenomenon for the brain tumor dataset. However, the pretrained models may generate highly diversified, high-quality samples for commonly available categories

Effect of different ways of conditioning on generation: We used various conditioning techniques to diversify images within a category. This included conditioning the dif-

Figure 2. Column 1 depicts real images from CIFAR-100. In the remaining columns, the first two columns show the mostly coarse-grained change and the next two columns mostly show fine-grained change.

fusion model using only image embedding, only image caption embedding, and "class label with image number" embedding (e.g., "A Photo 1/n of type 'class label'"). Conditioning with caption embedding provided stable and diverse generation. We also attempted to obtain the difference between two image embeddings to derive an editing direction ($E = E(I1) - E(I2)$) for conditioning the diffusion model. However, this resulted in unstable generation, often producing images of other classes and distorted images. As a result, we decided to condition the model with caption embedding along with image latent conditioning in the fine-tuned model. Figure 4 illustrates the effect of different conditioning choices on generation. We observed that generation from image embedding conditioning resulted in poor and less realistic images. We attribute this to image embedding sometimes not capturing proper information from images, especially when the image resolution is poor compared to caption embedding. The last row of the figure depicts the generation when conditioning is done with both caption embedding and shape structure, but these also produced less realistic images.

Effect number of tokens Mixing: We have conducted tests involving different numbers of token shuffling, where each token is represented by an embedding vector. After shuf-

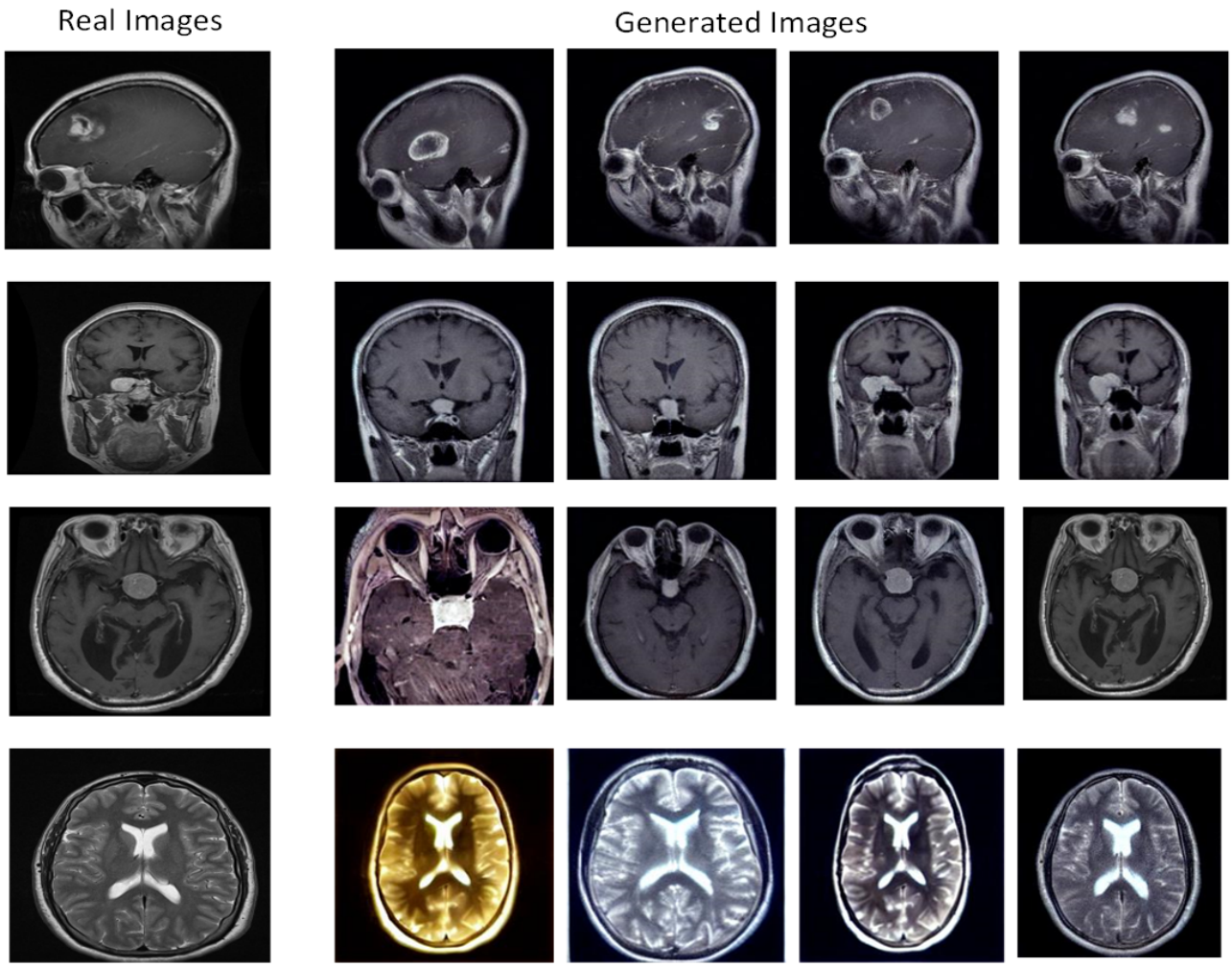

Figure 3. Column 1 depicts real images from Brain Tumor Dataset [2]. In the remaining columns, the first two columns show mostly coarse-grained change and the next two columns mostly show fine-grained change.

fling in the embedding space, we observed that increasing the number of tokens shuffled does not significantly affect diversity. In text embedding, the initial tokens, equivalent to the number of words in the caption, are the most important, with additional tokens coming from null text ("") embedding among 77 tokens. As a result, shuffling after the main caption's tokens is less impactful. Hence, it is essential to set the shuffling to encompass some of the tokens from the caption. In Figure 5, the lower sub-figure illustrates token shuffling, with Label 1 provided the best diversity by covering caption tokens, while gradually including more tokens from null text or padding results in less diverse input caption embedding, thus generating less diverse samples. Another observation is that by shuffling a small part of each embedding vector, fine-grained modifications can be made. We noted that increasing the size of the shuffling part leads to increasingly distorted images being generated. In Figure 5, the upper sub-figure represents the effect of fine-grained changes, where a portion of randomly selected vectors is replaced. From the figure, it was observed that Label 1 produced the most realistic results. However, as the shuffling parts are increased, the input embedding deviates increasingly from the original, resulting in distorted images. Therefore, for coarse-grained changes, it is acceptable to shuffle any number of tokens, including the first few tokens representing the captions. For fine-grained changes, a small part (less than 10) among the 768 is replaced.

Two-Phase Training. In the second approach, we follow Two-Phase Training. In the initial phase,we trained the classification models using a combined dataset. This combined dataset consisted of the real dataset along with a significantly larger portion of synthetic data (three to four times the size of the real dataset).During this phase, the model focuses on learning general patterns and features present in

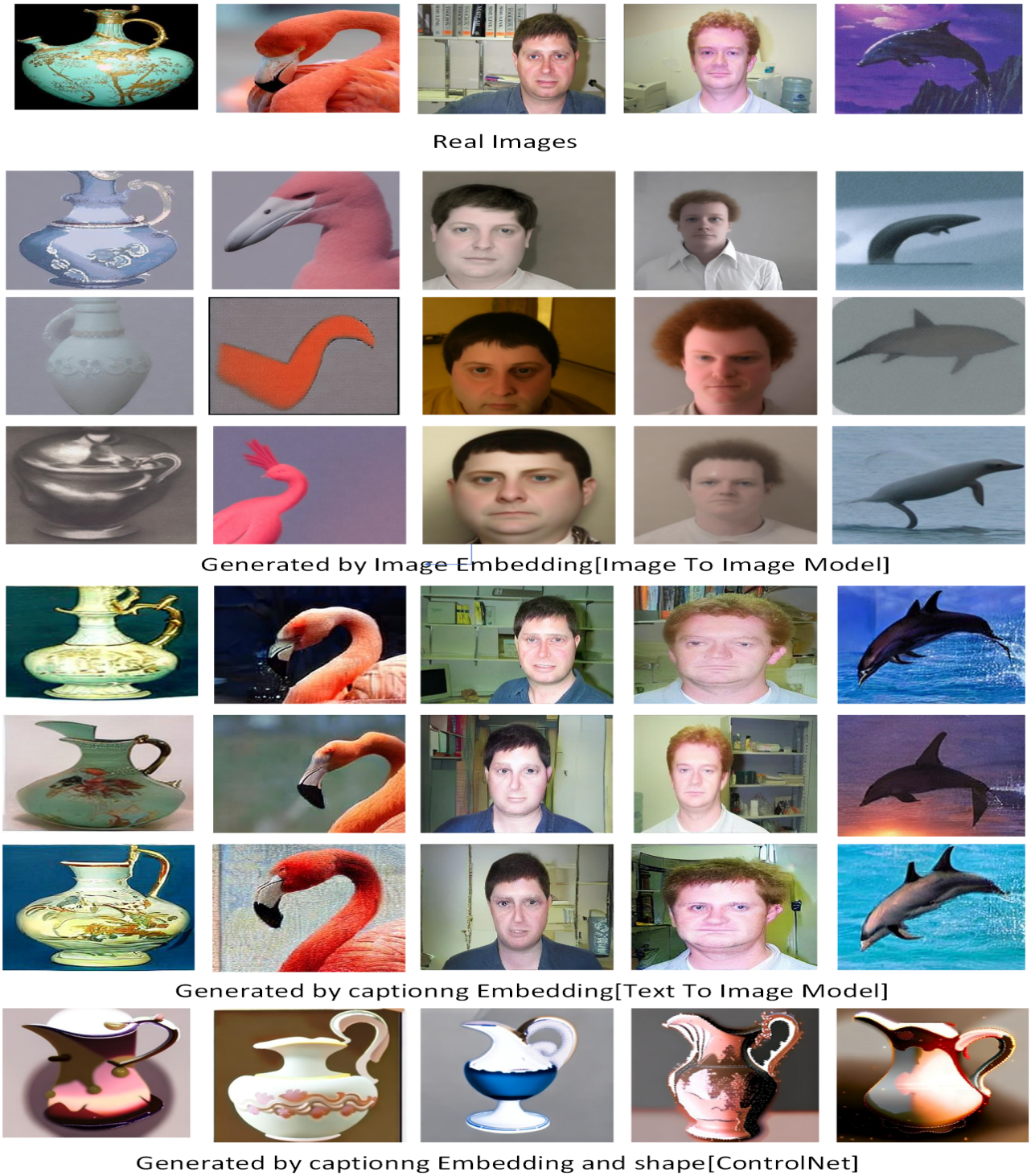

Figure 4. First row represents the original images. The next three rows depict images generated by Image Embedding, the subsequent three rows represent the images generated by caption embedding, and the last row represents the images generated by ControlNet [10].

both real and synthetic data. Exposure to a large volume of synthetic data helps in understanding the broader distribution of possible inputs.After the initial training phase, we fine-tuned the model using only the real dataset with reduced training steps and smaller learning rate. This phase aims to refine the model's understanding specifically on

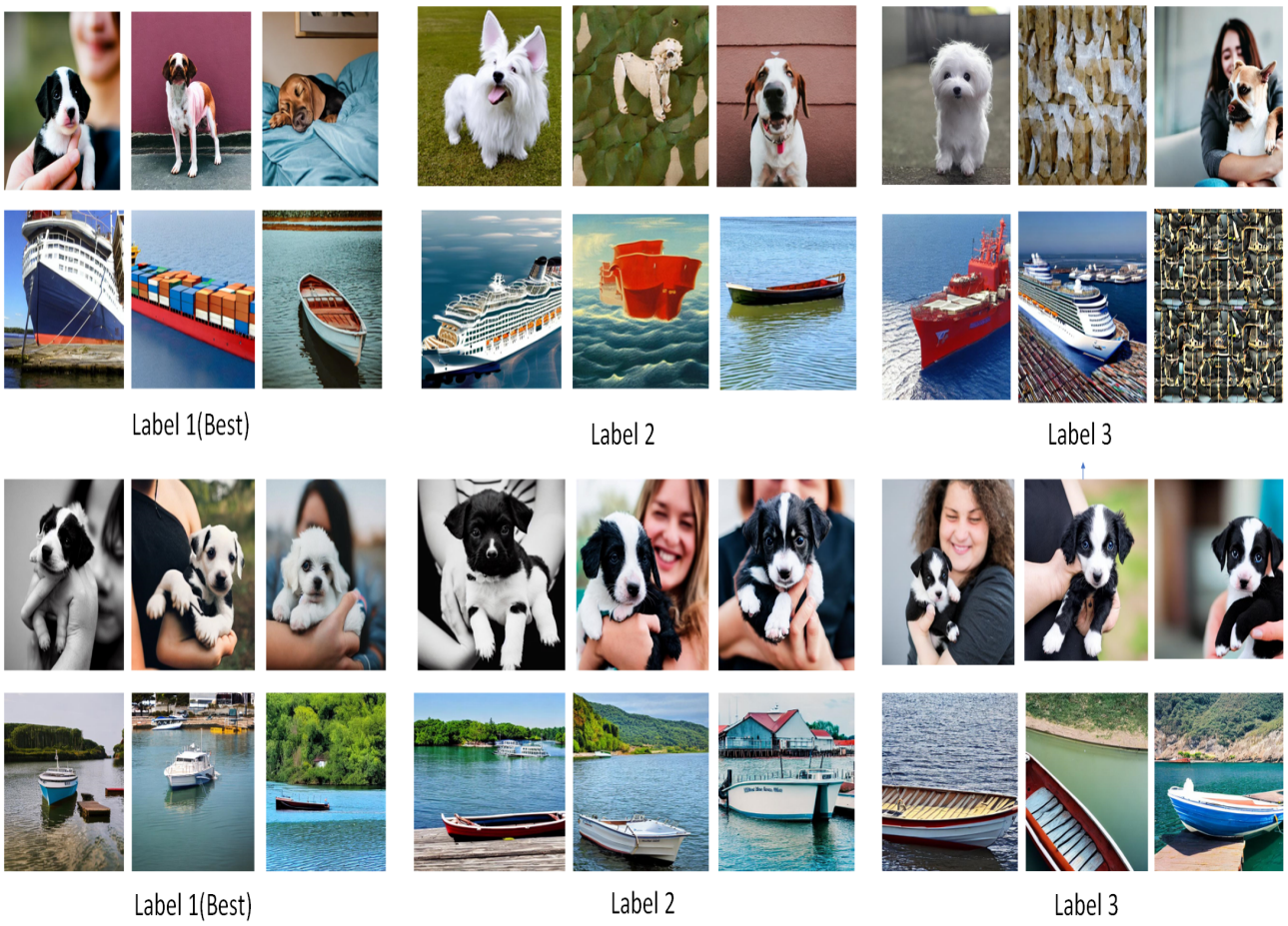

Figure 5. First two rows depict the effect of randomly replacing parts of embedding vectors increasingly. The lower two rows illustrate the effect of tokens shuffling increasingly from label 1 to label 3.

real-world examples. By limiting the number of training iterations during fine-tuning, we prevent the model from overfitting to the real dataset. This approach encourages the model to retain the broader insights gained from the synthetic data while fine-tuning on real examples. Applying a very small learning rate during fine-tuning ensures that the model updates its parameters slowly. This cautious adjustment helps in consolidating the knowledge acquired from the initial phase and minimizes the risk of losing beneficial insights gained from the synthetic data.This two-phase approach effectively addresses domain shift issues by balancing exposure to synthetic and real data. It allows the model to leverage the advantages of synthetic data in the initial phase while ensuring that final performance metrics are primarily influenced by real-world data.

Random Selection with Probability. Instead of straightforwardly combining all synthetic data with the real dataset, we introduced a probabilistic approach. This involved randomly selecting synthetic images with a specified probability (e.g., 80%) from a pool of synthetic datasets and combining them with batches from the real dataset during training. This dynamic integration aimed to balance synthetic and real data effectively. Our findings showed comparable performance to a two-phase strategy. This method offers flexibility by incrementally adding synthetic data, optimizing computational resources by focusing on a subset per batch rather than overwhelming the model with extensive synthetic datasets.

For a qualitative and quantitative comparison, please take a look at the main paper.

3. Limitation and Future Direction

Our extensive experiments have shown that synthetic data can enhance classifier learning and achieve new state-of-the-art performance with a comparable margin. Furthermore, our results indicate that current synthetic data have strong potential for model pre-training, which helps to bridge domain gaps in small datasets. However, we

have noticed some limitations of pretrained diffusion models. These models often generate samples that do not align well with the dataset's domain. While they produce high-quality samples for common categories, they are less effective for fine-grained and medical image domains, likely due to training biases.

To address this limitation, we have explored the critical role of fine-tuning and domain-specific augmentation strategies in mitigating out-of-domain sample generation issues when using pretrained diffusion models. Our proposed fine-tuning strategies and augmentation prompting methods are essential for improving the quality and applicability of generated samples in specialized domains, ultimately enhancing the overall effectiveness of machine learning models in real-world applications. We have also investigated the use of synthetic and real images to improve performance and found that the second and third approaches mentioned in section 4.2 in main paper are more effective for training the classification model.

However, we have identified limitations and challenges in using synthetic data for image augmentation. One limitation is that due to limited computational resources, we were unable to further scale up for the larger datasets such as Imagenet, which would require months to train a diffusion model and generating snthetic dataset of millions of scale. Additionally, we were unable to explore larger model sizes and advanced architectures in our current investigation, which is an area worth exploring in future research.

In our recent studies, we've encountered significant challenges when addressing domain shifts in generated images while attempting to increase generation diversity. Despite visually similar appearances upon inspection, augmented images often exhibit differences in pixel-level distributions that can impact the learning process of classification models. Our extensive experiments aimed at achieving a substantial increase in classification accuracy to 99% through augmented images have not yielded success due to these domain diversity issues. The generated samples do not adequately cover the entire distribution seen in the test data.

Looking ahead, our future research direction will focus on developing novel training strategies for classification models. Specifically, we aim to explore methods where the visual distribution of images influences the pixel-level numerical distribution. This approach holds promise for advancing zero and few-shot learning capabilities. By learning from visual distributions, akin to how humans naturally learn, we may potentially reduce the reliance on large datasets and generalize more effectively across diverse image datasets.

Human learning is inherently rooted in visual patterns rather than underlying numerical distributions. Thus, future breakthroughs in image augmentation for any-shot learning will likely hinge on our ability to effectively capture and leverage visual distributions in the training process.

In summary, our future research will concentrate on pioneering methodologies that bridge the gap between visual and numerical distributions in image data. This pursuit aims to unlock new avenues for enhancing the robustness and efficiency of machine learning models, particularly in scenarios requiring adaptation to varied and challenging image datasets.

References

[1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 1

[2] Jyotismita Chaki and Marcin Wozniak. Brain tumor mri dataset, 2023. 4

[3] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019. 1

[4] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020. 1

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 1

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1

[8] Sungbin Lim, Ildoo Kim, Taesup Kim, Chiheon Kim, and Sungwoong Kim. Fast autoaugment. *Advances in neural information processing systems*, 32, 2019. 1

[9] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 1

[10] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5