

Diffusion-based Dataset Augmentation for Downstream Crop Segmentation

Alex Senden¹, Masoomeh Gomroki², Robert H. Gulden², Christopher J. Henry¹

¹Department of Computer Science, University of Manitoba, Winnipeg Manitoba Canada

²Department of Plant Science, University of Manitoba, Winnipeg Manitoba Canada

Contact: sendena@myumanitoba.ca



Introduction

Semantic segmentation, the task of classifying each pixel in an image according to the object it represents, has become a common approach for automated weed detection. Implemented as a deep neural network, advances in semantic segmentation have led to substantial growth in model size and, consequently, in the required training data (1). Although unlabelled crop field images are plentiful, creating segmentation masks requires domain knowledge and a significant amount of human labour. We propose a generative approach to alleviate this data labelling bottleneck. Instead of requiring humans to manually annotate raw images, this project aims to generate realistic field imagery that is controllable by rough sketch-like input image layouts while also generating corresponding segmentation masks.

The objective of this work is to create models capable of generating realistic UAV-based field images adhering to image conditions while remaining extendible to future work co-generating segmentation masks.

Key Findings

- **Large diffusion models can generate realistic UAV field imagery**
- **Scribble annotations are sufficient to control both location and class of plants in generated images**
- **LoRA-tuning is sufficient to integrate image conditions in large diffusion models**

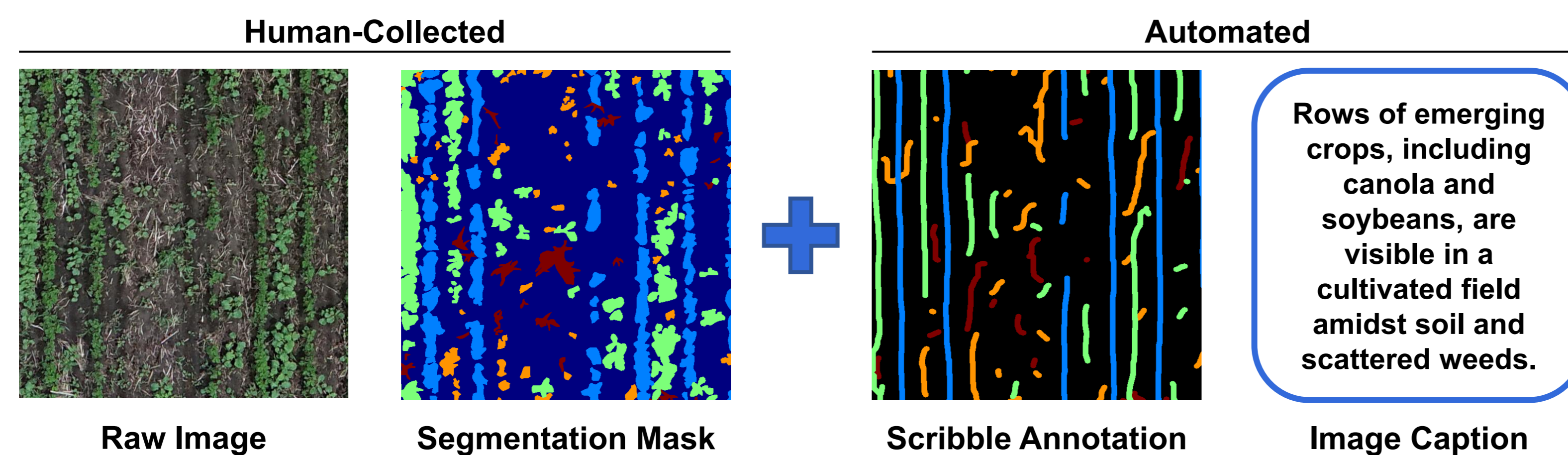


Figure 2. Example of inputs to train the generative model. RGB images and (optional) segmentation masks are the raw inputs. The raw image, along with image metadata, are input into a VLM to create image captions. If present, segmentation masks are translated into scribble annotations through a series of morphological preprocessing steps.

Results, Discussion, & Future Work

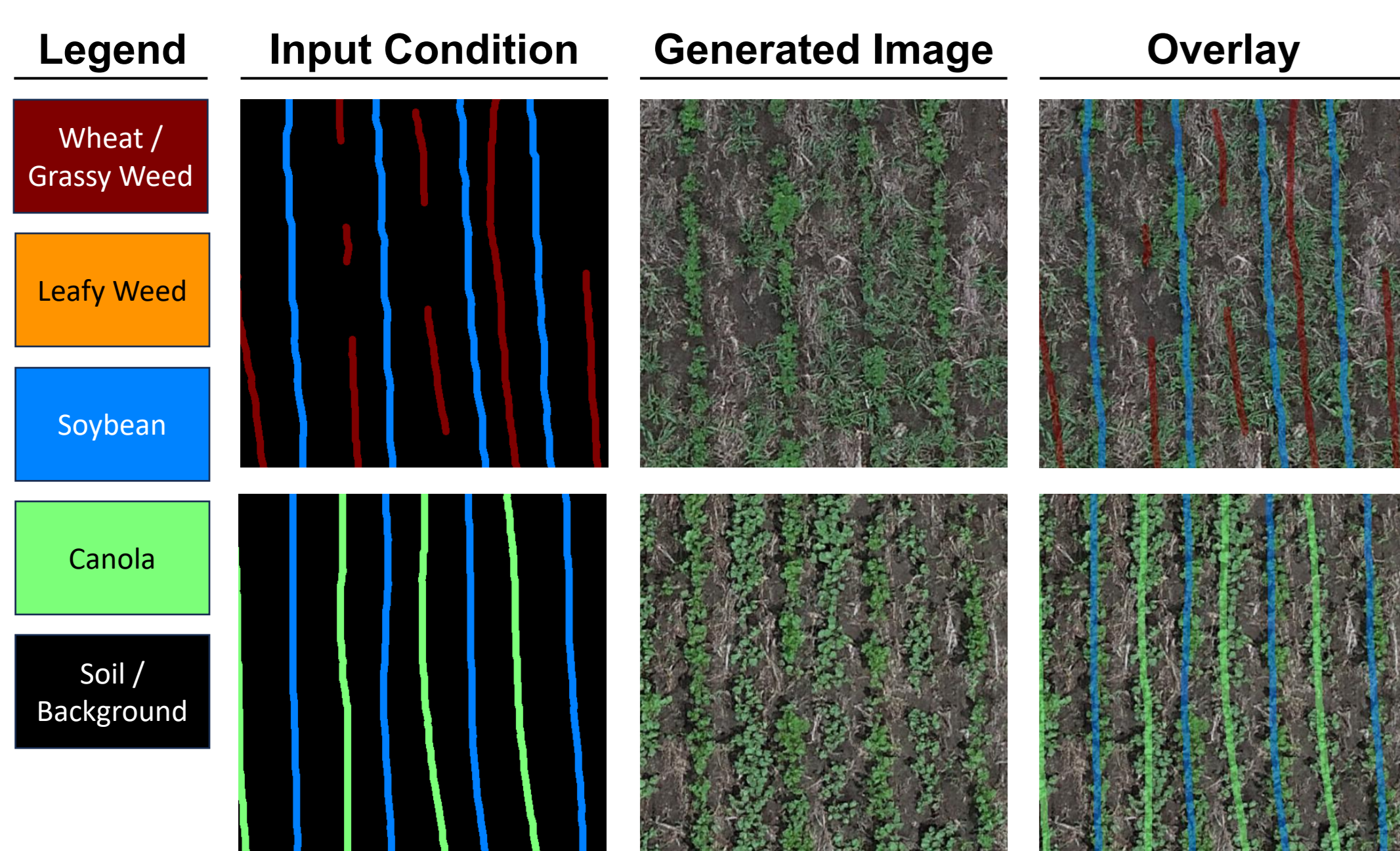


Figure 3. Examples of input conditions, corresponding generated imagery, and an overlay of the two. Plants in the generated images closely adhere to both the class and layout described by the input condition.

References

- 1) Minaee et al. 2022. "Image Segmentation Using Deep Learning: A Survey". 44(7): 3523-3542
- 2) Ho et al. 2020. "Denoising Diffusion Probabilistic Models". NeurIPS.
- 3) Podell et al. 2024. "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis". ICLR.
- 4) Esser et al. 2024. "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis". ICML.
- 5) Hu et al. 2022. "Low-Rank Adaptation of Large Language Models". ICLR.

Materials & Methods

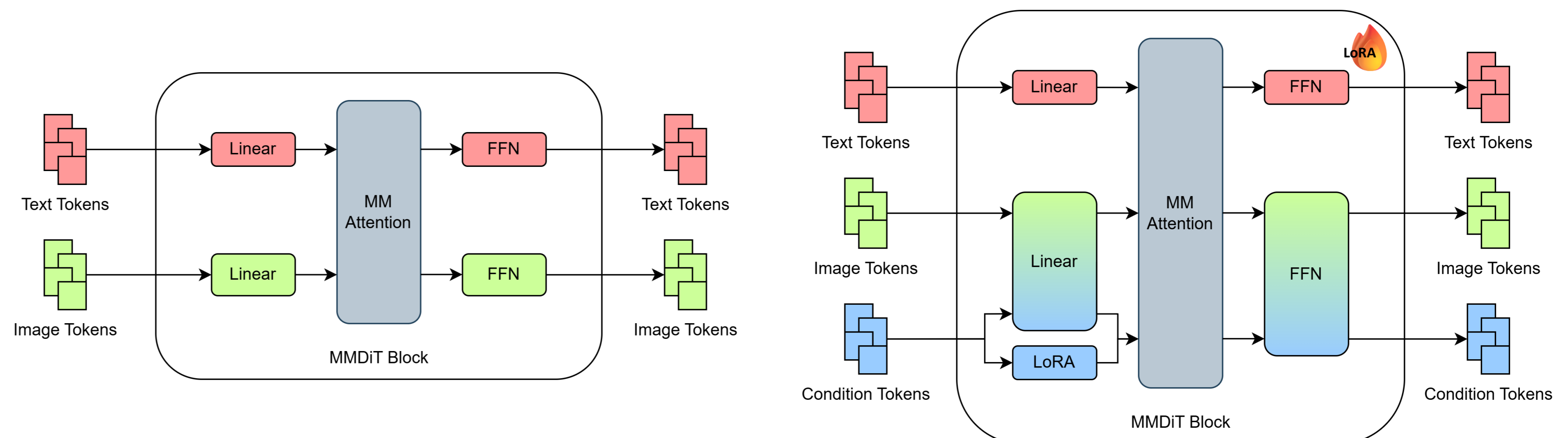


Figure 1. Left: Architecture of the MMDiT block in the base FLUX.1-dev diffusion model. Separate streams process image and text tokens, and the separate modalities only interact in multi-modal attention blocks. Right: Architecture of the proposed MMDiT block. An additional stream processes image condition tokens. This additional stream shares base weights with the image token stream but includes a trainable LoRA adapter for attention projections. The flame icon indicates that all components are trainable using LoRA.

A generative neural network is a model which produces new synthetic data (2). Diffusion models are a type of generative model that, once trained, create new sample images based on random input, and are considered the leading approach for synthetic image generation. Large, pretrained, general-purpose generative models are referred to as foundation models, and their diverse pretraining makes them ideal as a base for domain-specific fine-tuning (3). Control over generated imagery is exerted through input conditions, and text conditions are considered standard for foundation models (3). Images are required in each training step of a diffusion model, whereas conditions are optionally present for each training step.

The multi-modal diffusion transformer (MMDiT)-based FLUX.1-dev is selected as the foundation model due to its state-of-the-art image fidelity among open-weight models. The standard MMDiT block separately processes streams of text and image tokens (4). This is modified to include a third stream for image conditions via a LoRA adapter (5) on the linear projection layer of the attention mechanism, exclusively used by image condition tokens. All layers in all transformer blocks are additionally trained using LoRA.

During training, our model requires up to three inputs: a raw image, an image caption, and optionally, a scribble annotation. Human image annotation is time-consuming, therefore automated methods are used to generate conditions. A vision-language model (VLM), Qwen2.5-VL-32B-Instruct, is prompted with an image and basic information regarding the seeded crop and generates unique image captions. This reduces the human captioning burden from one caption per image to one caption per field. Semantic segmentation masks are manually created for a subset of images and are converted to scribble annotations through morphological preprocessing steps, including per-class dilation, erosion, and skeletonization.

The images used were captured by a UAV flown at a height of 12 meters. Imaged plots consist of alternating rows of soybean and wheat, or soybean and canola.

- Fine-tuning pure text-to-image generative models allows for coarse control of the resulting imagery, however much finer control can be exerted using image conditions. Scribble annotations allow a user to specify the locations of objects in synthetic imagery without specifying shape or size, making scribble annotations simple to generate.
- Generated images closely adhere to both the layout and classes provided by the scribble annotation. The resulting imagery does not show any significant artifacts resulting from the scribble annotations, such as visible lines in locations corresponding to scribbles.
- Best results were obtained by fine-tuning for 100,000 steps with a learning rate of 1e-6 using the AdamW optimizer. Models were trained with 4 NVIDIA H100 GPUs for approximately 4 hours using a batch size of 2.

- Beyond changes to the MMDiT blocks of FLUX.1-dev, analogous changes were experimented with on the single transformer blocks. This modification resulted in minor artifacts in locations corresponding to scribbles and a slight reduction in image fidelity.
- Relatively few annotated images are required to train the modified model to integrate image conditions. 4693 raw images are used in training, however only 95 images (2.02%) have corresponding scribble annotations. Training data sampling was biased such that there was a 50% chance that a selected training sample contained an image condition.
- This work will be extended to simultaneously generate segmentation masks for the synthetic data using similar architectural modifications.

Funding

