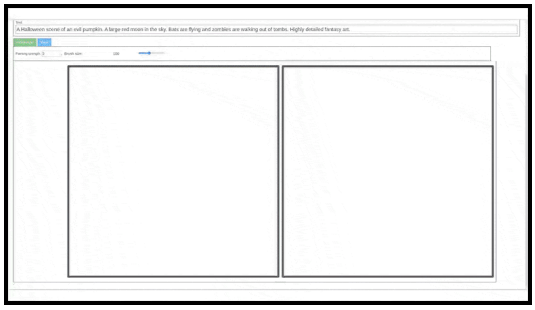
NVIDIA’s eDiffi Diffusion Model Allows ‘Painting With Words’ and More

Originally published at https://www.unite.ai/nvidias-ediffi-diffusion-model-allows-painting-with-words-and-more/

Attempting to make precise compositions with latent diffusion generative image models such as Stable Diffusion can be like herding cats; the very same imaginative and interpretive powers that enable the system to create extraordinary detail and to summon up extraordinary images from relatively simple text-prompts is also difficult to turn off when you’re looking for Photoshop-level control over an image generation.

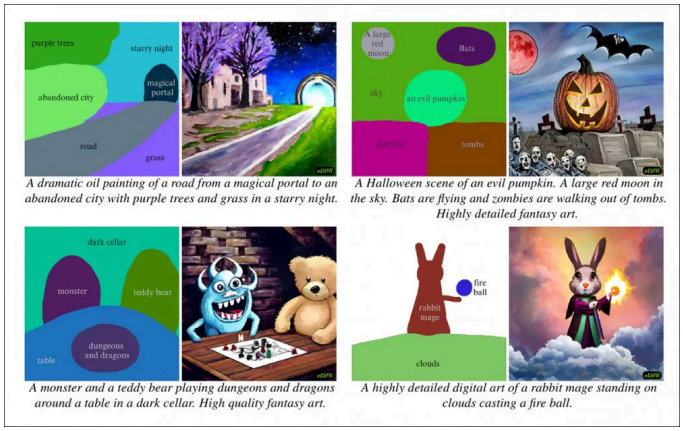
Now, a new approach from NVIDIA research, titled *ensemble diffusion for images* (eDiffi), uses a mixture of multiple embedding and interpretive methods (rather than the same method all the way through the pipeline) to allow for a far greater level of control over the generated content. In the example below, we see a user painting elements where each color represents a single word from a text prompt:



‘Painting with words’ is one of the two novel capabilities in NVIDIA’s eDiffi diffusion model. Each daubed color represents a word from the prompt (see them appear on the left during generation), and the area color applied will consist only of that element. See source (official) video for more examples and better resolution at https://www.youtube.com/watch?v=k6cOx9YjHJc

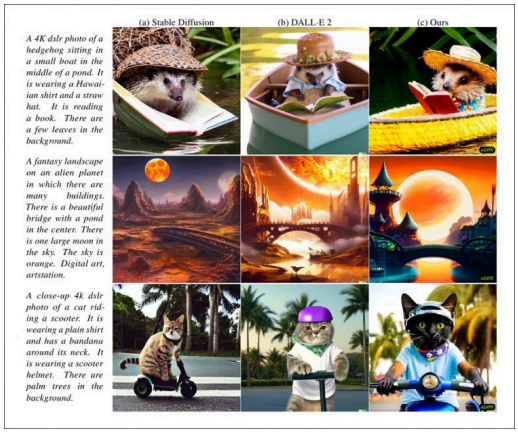
Effectively this is ‘painting with masks’, and reverses the inpainting paradigm in Stable Diffusion, which is based on fixing broken or unsatisfactory images, or extending images that could as well have been the desired size in the first place.

Here, instead, the margins of the painted daub represent the permitted approximate boundaries of just one unique element from a single concept, allowing the user to set the final canvas size from the outset, and then discretely add elements.



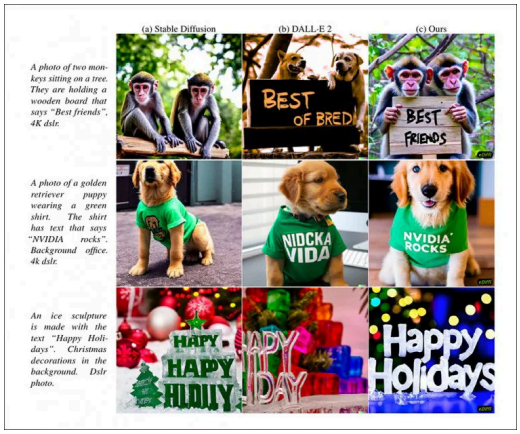
Examples from the new paper. Source: https://arxiv.org/pdf/2211.01324.pdf

The variegated methods employed in eDiffi also mean that the system does a far better job of including every element in long and detailed prompts, whereas Stable Diffusion and OpenAI's DALL-E 2 tend to prioritize certain parts of the prompt, depending either on how early the target words appear in the prompt, or on other factors, such as the potential difficulty in disentangling the various elements necessary for a complete but comprehensive (with respect to the text-prompt) composition:



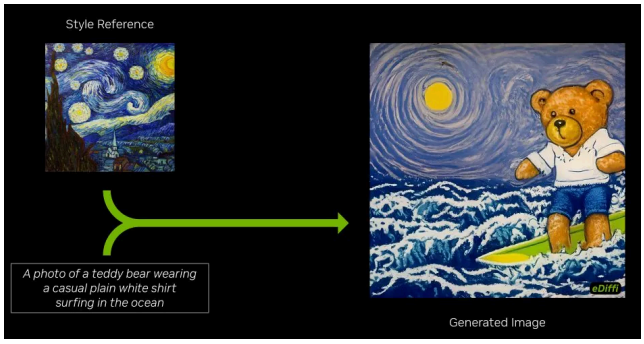
From the paper: eDiffi is capable of iterating more thoroughly through the prompt until the maximum possible number of elements have been rendered. Though the improved results for eDiffi (right-most column) are cherry-picked, so are the comparison images from Stable Diffusion and DALL-E 2.

Additionally, the use of a dedicated T5 text-to-text encoder means that eDiffi is capable of rendering comprehensible English text, either abstractly requested from a prompt (i.e. *image contains some text of [x]*) or explicitly requested (i.e. *the t-shirt says 'Nvidia Rocks'*):



Dedicated text-to-text processing in eDiffi means that text can be rendered verbatim in images, instead of being run only through a text-to-image interpretive layer than mangles the output.

A further fillip to the new framework is that it's possible also to provide a single image as a style prompt, rather than needing to train a DreamBooth model or a textual embedding on multiple examples of a genre or style.

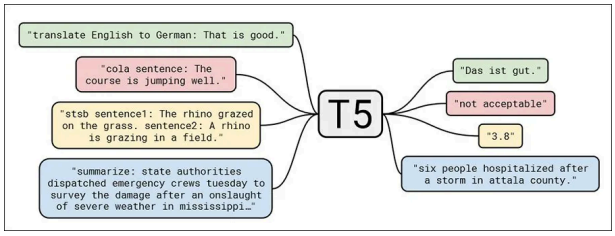


Style transfer can be applied from a reference image to a text-to-image prompt, or even an image-to-image prompt.

The new paper is titled *eDiffi: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers*, and

The T5 Text Encoder

The use of Google's Text-to-Text Transfer Transformer (T5) is the pivotal element in the improved results demonstrated in eDiffi. The average latent diffusion pipeline centers on the association between trained images and the captions which accompanied them when they were scraped off the internet (or else manually adjusted later, though this is an expensive and therefore rare intervention).

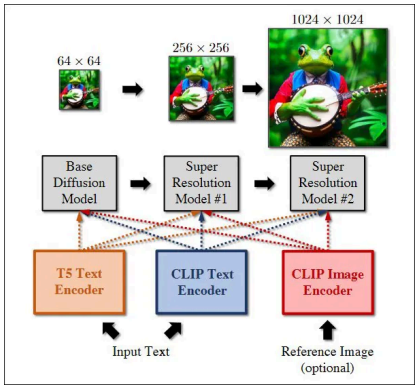


From the July 2020 paper for T5 – text-based transformations, which can aide the generative image workflow in eDiffi (and, potentially, other latent diffusion models). Source: https://arxiv.org/pdf/1910.10683.pdf

By rephrasing the source text and running the T5 module, more exact associations and representations can be obtained than were trained into the model originally, almost akin to *post facto* manual labeling, with greater specificity and applicability to the stipulations of the requested text-prompt.

The authors explain:

> *'In most existing works on diffusion models, the denoising model is shared across all noise levels, and the temporal dynamic is represented using a simple time embedding that is fed to the denoising model via an MLP network. We argue that the complex temporal dynamics of the denoising diffusion may not be learned from data effectively using a shared model with a limited capacity.*
>
> *'Instead, we propose to scale up the capacity of the denoising model by introducing an ensemble of expert denoisers; each expert denoiser is a denoising model specialized for a particular range of noise [levels]. This way, we can increase the model capacity without slowing down sampling since the computational complexity of evaluating [the processed element] at each noise level remains the same.'*



Conceptual workflow for eDiffi.

The existing CLIP encoding modules included in DALL-E 2 and Stable Diffusion are also capable of finding alternative image interpretations for text related to user input. However they are trained on similar information to the original model, and are not used as a separate interpretive layer in the way that T5 is in eDiffi.

The authors state that eDiffi is the first time that both a T5 and a CLIP encoder have been incorporated into a single pipeline:

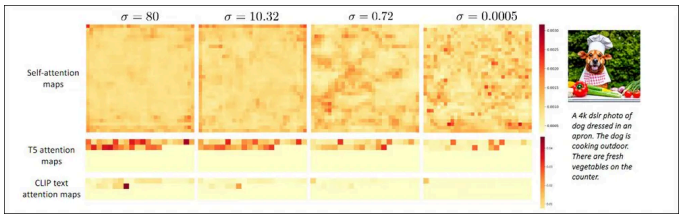
> *'As these two encoders are trained with different objectives, their embeddings favor formations of different images with the same input text. While CLIP text embeddings help determine the global look of the generated images, the outputs tend to miss the fine-grained details in the text.*
>
> *'In contrast, images generated with T5 text embeddings alone better reflect the individual objects described in the text, but their global looks are less accurate. Using them jointly produces the best image-generation results in our model.'*

Interrupting and Augmenting the Diffusion Process

The paper notes that a typical latent diffusion model will begin the journey from pure noise to an image by relying solely on text in the early stages of the generation.

When the noise resolves into some kind of rough layout representing the description in the text-prompt, the text-guided facet of the process essentially drops away, and the remainder of the process shifts towards augmenting the visual features.

This means that any element that was not resolved at the nascent stage of text-guided noise interpretation is difficult to inject into the image later, because the two processes (text-to-layout, and layout-to-image) have relatively little overlap, and the basic layout is quite entangled by the time it arrives at the image augmentation process.

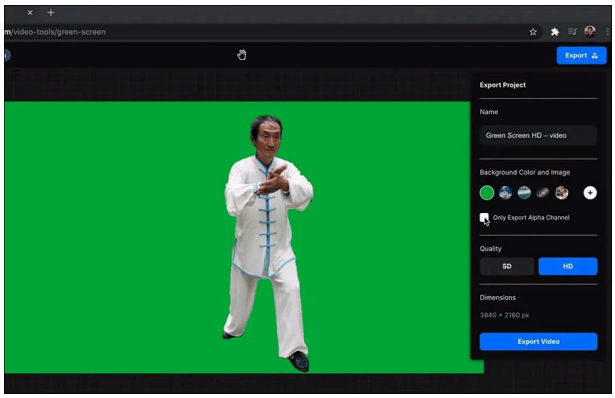


From the paper: the attention maps of various parts of the pipeline as the noise>image process matures. We can see the sharp drop-off in CLIP influence of the image in the lower row, while T5 continues to influence the image much further into the rendering process.

Professional Potential

The examples at the project page and YouTube video center on PR-friendly generation of meme-tastic cute images. As usual, NVIDIA research is playing down the potential of its latest innovation to improve photorealistic or VFX workflows, as well as its potential for improvement of deepfake imagery and video.

In the examples, a novice or amateur user scribbles rough outlines of placement for the specific element, whereas in a more systematic VFX workflow, it could be possible to use eDiffi to interpret multiple frames of a video element using text-to-image, wherein the outlines are very precise, and based on, for instance figures where the background has been dropped out via green screen or algorithmic methods.



Runway ML already provides AI-based rotoscoping. In this example, the 'green screen' around the subject represents the alpha layer, while the extraction has been accomplished via machine learning rather than algorithmic removal of a real-world green screen background. Source: https://twitter.com/runwayml/status/1330978385028374529

Using a trained DreamBooth character and an image-to-image pipeline with eDiffi, it's potentially possible to begin to nail down one of the bugbears of *any* latent diffusion model: temporal stability. In such a case, both the margins of the imposed image and the content of the image would be 'pre-floated' against the user canvas, with temporal continuity of the rendered content (i.e. turning a real-world Tai Chi practitioner into a robot) provided by use of a locked-down DreamBooth model which has 'memorized' its training data – bad for interpretability, great for reproducibility, fidelity and continuity.

Method, Data and Tests

The paper states the eDiffi model was trained on 'a collection of public and proprietary datasets', heavily filtered by a pre-trained CLIP model, in order to remove images likely to lower the general aesthetic score of the output. The final filtered image set comprises 'about one billion' text-image pairs. The size of the trained images is described as with 'the shortest side greater than 64 pixels'.

A number of models were trained for the process, with both the base and super-resolution models trained on AdamW optimizer at a learning rate of 0.0001, with a weight decay of 0.01, and at a formidable batch size of 2048.

The base model was trained on 256 NVIDIA A100 GPUs, and the two super-resolution models on 128 NVIDIA A100 GPUs for each model.

The system was based on NVIDIA's own Imaginaire PyTorch library. COCO and Visual Genome datasets were used for evaluation, though not included in the final models, with MS-COCO the specific variant used for testing. Rival systems tested were GLIDE, Make-A-Scene, DALL-E 2, Stable Diffusion, and Google's two image synthesis systems, Imagen and Parti.

In accordance with similar prior work, zero-shot FID-30K was used as an evaluation metric. Under FID-30K, 30,000 captions are extracted randomly from the COCO validation set (i.e. not the images or text used in training), which were then used as text-prompts for synthesizing images.

The Frechet Inception Distance (FID) between the generated and ground truth images was then calculated, in addition to recording the CLIP score for the generated images.

Model	# of params	Zero-shot FID ↓
GLIDE [49]	5B	12.24
Make-A-Scene [15]	4B	11.84
DALL·E 2 [55]	6.5B	10.39
Stable Diffusion [57]	1.4B	8.59
Imagen [61]	7.9B	7.27
Parti [83]	20B	7.23
eDiffi-Config-A	6.8B	7.35
eDiffi-Config-B	7.1B	7.26
eDiffi-Config-C	8.1B	7.11
eDiffi-Config-D	9.1B	**7.04**

Results from the zero-shot FID tests against current state-of-the-art approaches on the COCO 2014 validation dataset, with lower results better.

In the results, eDiffi was able to obtain the lowest (best) score on zero-shot FID even against systems with a far higher number of parameters, such as the 20 billion parameters of Parti, compared to the 9.1 billion parameters in the highest-specced eDiffi model trained for the tests.

Conclusion

NVIDIA's eDiffi represents a welcome alternative to simply adding greater and greater amounts of data and complexity to existing systems, instead using a more intelligent and layered approach to some of the thorniest obstacles relating to entanglement and non-editability in latent diffusion generative image systems.

There is already discussion at the Stable Diffusion subreddits and Discords of either directly incorporating any code that may be made available for eDiffi, or else re-staging the principles behind it in a separate implementation. The new pipeline, however, is so radically different, that it would constitute an entire version number of change for SD, jettisoning some backward compatibility, though offering the possibility of greatly-improved levels of control over the final synthesized images, without sacrificing the captivating imaginative powers of latent diffusion.

First published 3rd November 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai