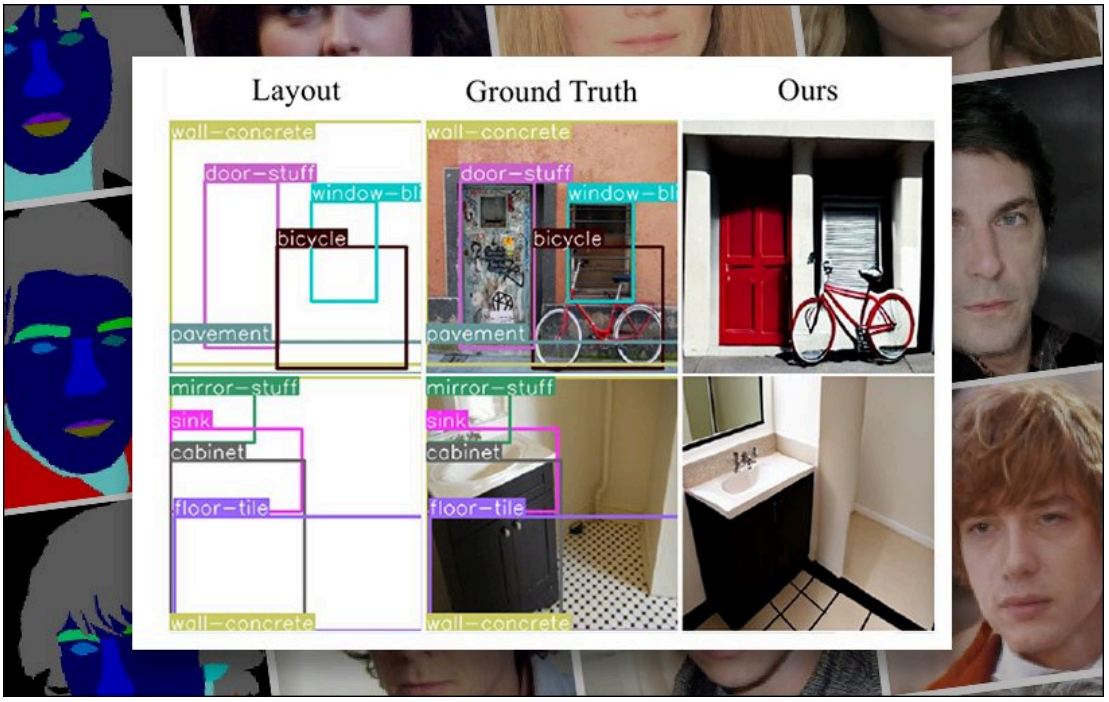
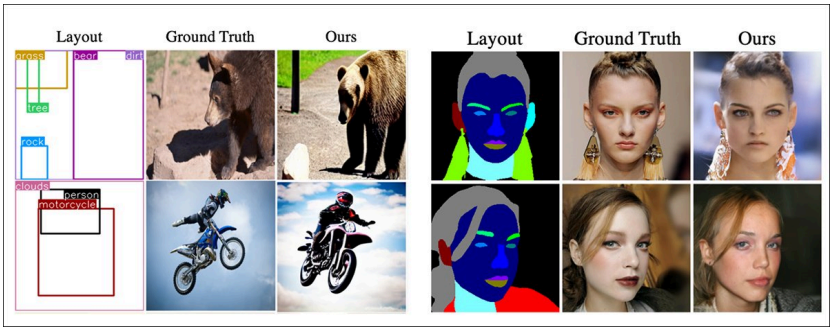
New System Offers Superior Layout Composition for Stable Diffusion

Originally published at https://arquivo.pt/wayback/20240203184218oe_/https://blog.metaphysic.ai/new-system-offers-superior-layout-composition-for-stable-diffusion/

February 21, 2023



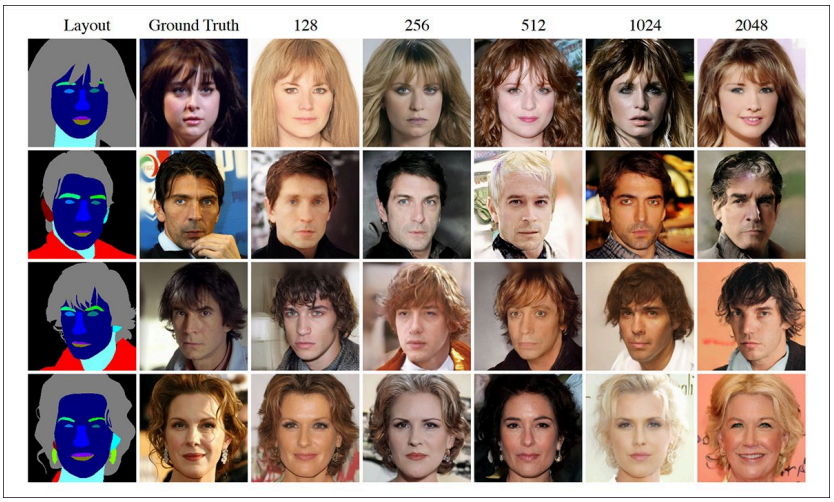
A new research collaboration between the US and China is offering a method for users of Stable Diffusion and other latent diffusion model-based (LDM) networks to create images with more specificity, by composing image layouts in advance. This allows the user to decide in advance where individual components of the image will go, and even to being able to write text prompts for each part of the image.



Scene and face generation with LayoutDiffuse. Source: https://arxiv.org/pdf/2302.08908.pdf

Titled *LayoutDiffuse*, the new approach involves fine-tuning an existing LDM so that it has an additional 'compositional' strategy and schema, allowing users greater control in assembling diverse possible elements into a generated image.

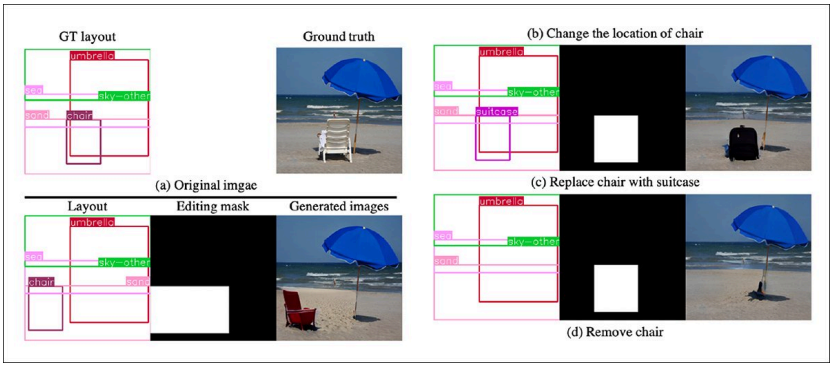
Though most previous attempts to create such a system have focused on scene-based assemblies (such as rooms and interior architecture, exterior scenes, etc.), the new method can use labeled semantic segmentation-style areas to also generate faces, where each particular element of the head has been constrained and predefined.



Generations using LayoutDiffuse, at varying resolutions. The numbers above the columns represent the number of images on which the model at hand was fine-tuned (at 2000 iterations each). As we can see, lower numbers of images produce very usable results, while higher volumes of images can increase diversity, giving the user room to interpose their desired image results, via text-prompts

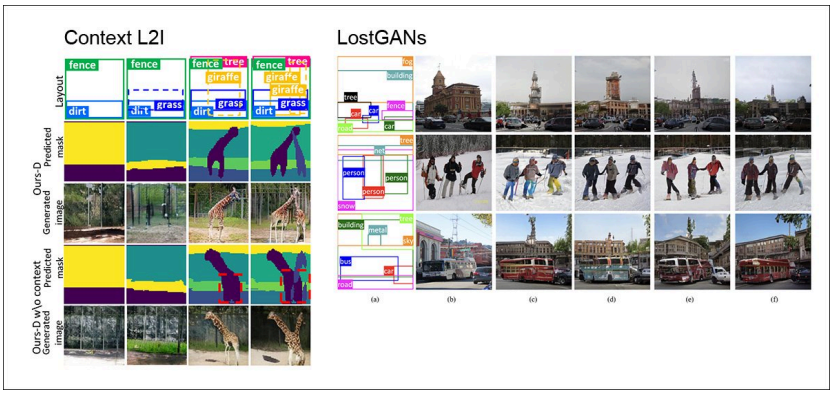
The system shares some common aims with the current 'hot application' of Stable Diffusion, Controlnet, an implementation of the recent Stanford University paper, which uses a variety of model-based semantic methodologies to perform highly-constrained transformations and prior constraints on generated images, almost eliminating the 'random luck' that has often been involved in getting Stable Diffusion to behave more like Photoshop, and less like a 'lucky dip'.

However, while Controlnet permits users to corral and retain objects and details that Stable Diffusion might otherwise de-prioritize, its semantic segmentation functionality is not yet working fully, and it is a *post facto* solution to compositionality, whereas the new offering trains this per-object capability into the base model.



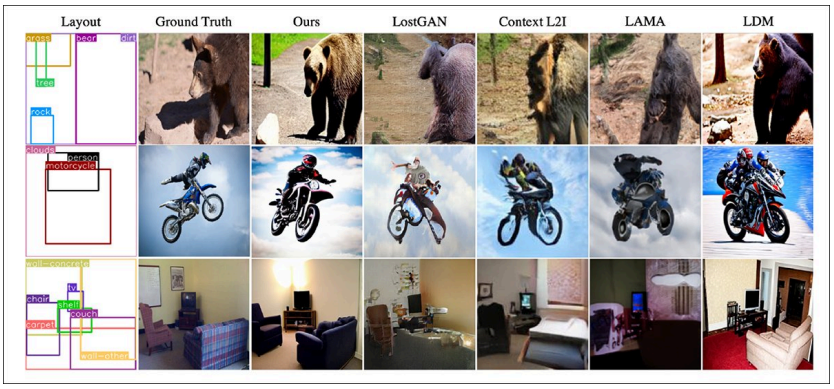
Masks are part of the central architecture of LayoutDiffuse, and can be targeted by custom text-prompts, allowing for movement, alteration and omission, as the user may require, before the image is even generated. This is a contrast to Controlnet, which would still require some level of pixel-based intervention on an Img2Img workflow (i.e., before processing in the latent space), or the complete origination and editing of a 'control' skeleton layout.

Previous works in this area have tended to focus on the older technology of Generative Adversarial Networks (GANs). Among these, tested by the researchers against their new system, are *LostGANs*, *Context-Aware Layout to Image Generation with Enhanced Object Appearance* (Context L2I), and LAMA.



Examples from prior layout composition frameworks Context L2I and LostGANs. Source: https://github.com/iVMCL/LostGANs / https://arxiv.org/pdf/2103.11897.pdf

However, in quantitative and qualitative tests conducted for the new paper, the researchers found that LayoutDiffuse was able to obtain comfortably superior results.



Qualitative results – the final column is a default Stable Diffusion baseline. See below for results of the quantitative test rounds.

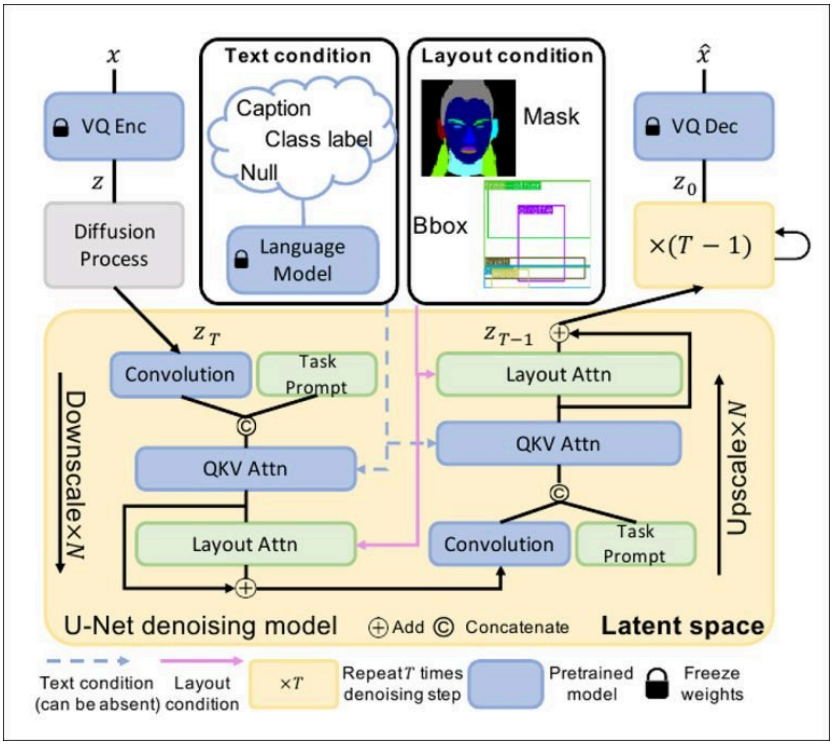
The researchers conclude that LayoutDiffuse achieves state-of-the-art layout capabilities for Stable Diffusion, and more recognizable object depiction, while being more efficient with time and data.

The new paper is titled *LayoutDiffuse: Adapting Foundational Diffusion Models for Layout-to-Image Generation*, and comes from six researchers across USC Information Sciences Institute, Shanghai Jiao Tong University and Amazon Web Services.

Approach

There is a notable locus of research effort in this period aimed at increasing the user's ability to impose fine control over latent diffusion image generation. The recent Mixture of Diffusers (MoD) method is a *post facto* composition framework capable, like LayoutDiffuse, of enabling per-section prompts within a composition; an earlier work from Deep Mind attempted to frame compositions by specifying coordinates within the generative area; and Controlnet's predecessor Instruct2Pix facilitated a tight control over composition, though concentrating more on transforming and adhering to the strictures of in-object content, rather than object placement within the composition (which it can nonetheless achieve through semantic segmentation and various other methods).

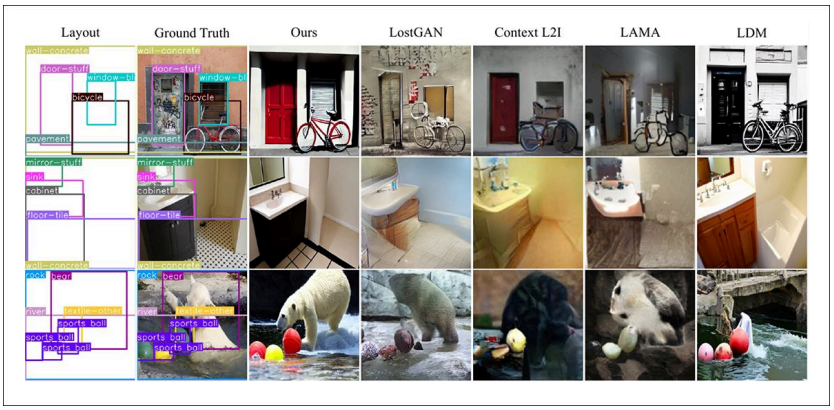
LayoutDiffuse enables layout>image generation through the use of what the researchers name *foundational diffusion models* (FDMs). An FDM is a diffusion model which has been fine-tuned with a novel *neural adapter*.



Conceptual architecture for LayoutDiffuse. Task-adaptive prompts and Layout Attention are imposed into the traditional training/fine-tuning schema, added to Queries/Keys/Values (QKV) Transformers attention layers.

The adapter is comprised of two components: task-adaptive prompts and *layout attention* (see image below). The former refocuses the attention of the generative process towards individual instances which will be placed into the final layout, while task-adaptive prompts are custom vectors which orient the modified DM toward 'layout mode'.

Though the tests for LayoutDiffuse were used on Stable Diffusion, the authors note that the system is quite generic, and can be applied to any latent diffusion model, and can also be used on models which are either conditioned on text (such as Stable Diffusion, whose image/text conditionality was founded on CLIP), or unconditioned.



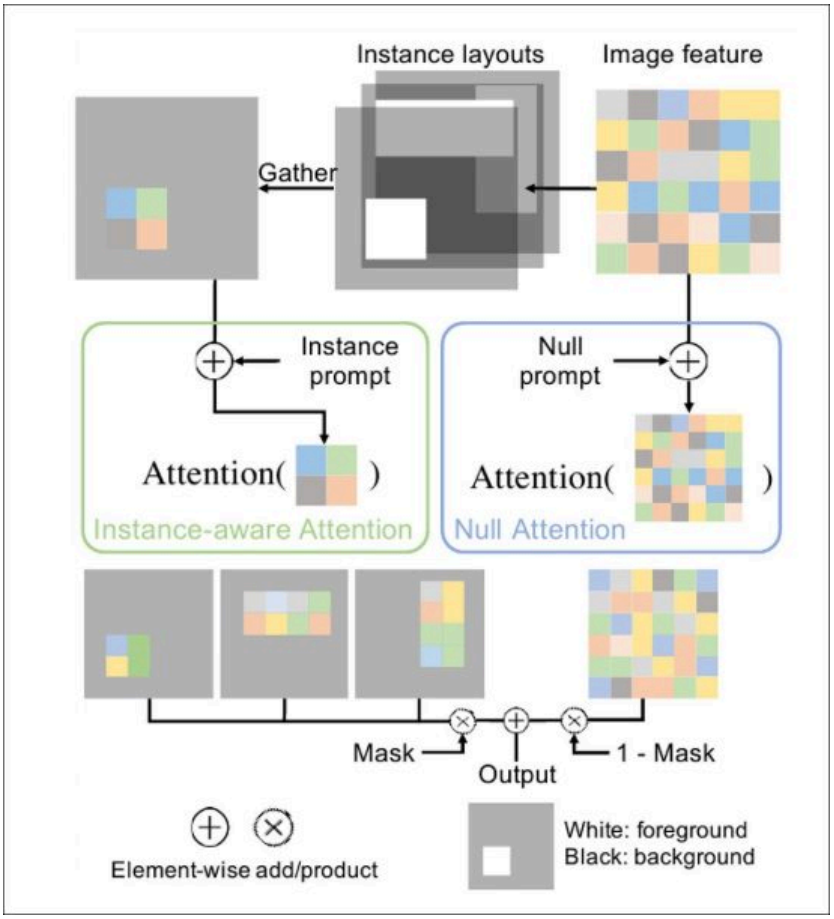
Qualitative results from bounding-box layout>image generation on COCO.

Rethinking the Foundations

The process of creating the foundational diffusion model involves fine-tuning the entire Stable Diffusion model, with the exception of the VQ-VAE and text encoders.

This, the authors state, allows the system to become more compositionally-oriented even without the imposition of their new neural adapter, albeit the effectiveness is greatly increased with use of the latter.

The new functionality is essentially a 'retro-fitting' of the core Stable Diffusion architecture: the additional layout attention layers are added *after* the Queries/Keys/Values (QKV) Transformers attention layers, and are imposed as residual blocks with an output layer initialized at zero. Therefore these layers alone are trained 'from scratch', in the context of a completed Stable Diffusion model which required weeks of training across thousands of GPUs.



Workflow for the Layout Attention layer. Null attention (top right), the authors note, is also used in Classifier Free Guidance (CFG), a central instrumentality in Stable Diffusion, which allows the system to modulate how closely the generated image will adhere to the user's text-prompt. The 'masks' visualized above represent areas of concentration for object instances that will appear in the image, and also illustrate where in the image the objects are to be placed.

Effectively this approach means that all the convolutional and QKV attention layers are initialized with pretrained weights from the original model.

The novel Layout Attention layer (see image above) adds two new wrinkles to the standard architecture of Stable Diffusion: an instance prompt, which is a learnable class embedding covering each category in the trained data, and task-adaptive prompts.

In the former, each region in the image gets an additional token to carry the desired instrumentality, while the part of the image which is to be disregarded (in terms of the sub-object being considered) is handled with a *null prompt* (the grey areas in the image above), which produces no

(manifested) output.

This system also accommodates Classifier-Free Guidance (CFG), an essential control mechanism in Stable Diffusion, which determines the extent to which the user's specified prompt is enacted, and the extent to which the system is permitted a 'free hand' to broadly interpret the generated image.

Likewise, task-adaptive prompts, which indicate to the system that a layout schema should be used, are interposed into the pre-existing QKV attention layers.

Training and Tests

LayoutDiffuse was fine-tuned with the Adam optimizer for 30, 60 and 120 epochs at a learning rate of 3×10^{-5} on three datasets: CelebA-Mask, COCO Stuff and Visual Genome (VG). The training took place on NVIDIA A10 Tensor Core GPUs, for 10 hours (CelebA-Mask, 200 epochs) and 50 hours (COCO and VG, 60 and 120 epochs) respectively. The captions from COCO were adapted for the text-to-image model, and the caption-free VG dataset was used with comma-separated class labels (i.e., *'car'*, *'tree'*, *'building'*).

In addition to the baseline Stable Diffusion model (denoted as 'LDM' in results), LayoutDiffuse was tested against the aforementioned prior methods, meaning that a comparison was made to GAN-based and VQ-VAE+AR-based methodologies. Additional methods tested were TAMING, TwFA, and OC-GAN.

Tests conducted were complex and varied. Below we see the results from a test on bounding-box layout-to-image functionality, where LayoutDiffuse scores best under Fréchet Inception Distance (FID), Classification Accuracy Score (CAS), and Inception Score (IS), and also leads the board in a qualitative 50-user study.

Methods	FID↓		CAS↑		Inception Score↑		User Preference↑	
	COCO	VG	COCO	VG	COCO	VG	Quality	Align Fidelity
LostGAN [41]	42.55	47.62	30.33	28.81	18.01±0.50	14.10±0.38	5.22%	2.50%
OCGAN* [42]	41.65	40.85	-	-	-	-	-	-
Context L2I [6]	29.56	119.74	32.63	20.09	18.57±0.54	6.71±0.30	1.10%	2.50%
LAMA [21]	31.12	31.63	30.52	31.75	14.32±0.58	10.79±0.66	1.98%	3.21%
Taming* [11]	33.68	19.14	-	-	-	-	-	-
TwFA* [46]	22.15	17.74	-	-	24.25±1.04	25.13±0.66	-	-
LDM [31]	24.60	25.27	43.48	34.78	26.11±0.88	20.59±0.41	20.88%	24.38%
LayoutDiffuse (Ours)	**20.27**	**15.96**	**50.02**	**40.41**	**32.07±0.80**	**26.53±0.41**	**62.36%**	**61.88%**

The most involved round of tests, evaluating alignment fidelity through algorithmic and human-tested methodologies.

Of these results, the authors declare:

'We notice that LayoutDiffuse achieves SoTA on all three metrics and is the most preferred algorithm in human evaluation. LayoutDiffuse achieves a significant improvement on CAS and IS, suggesting that the objects generated by our method are more similar to real objects (higher CAS) and are more diversified and distinguishable (higher IS).'

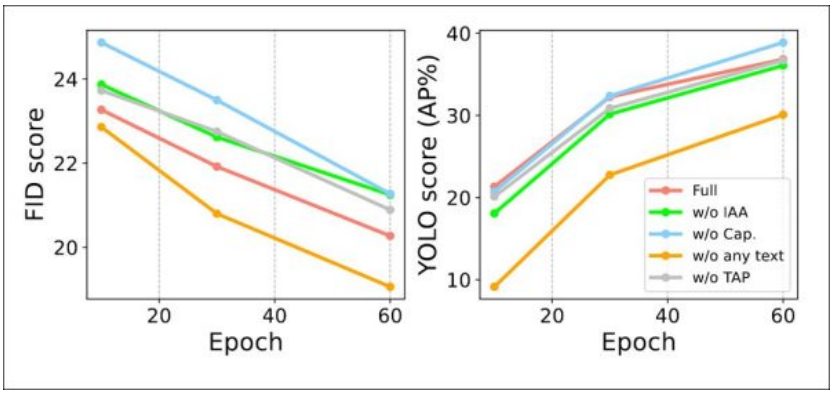
Additionally, the system was tested for mask-to-image functionality (i.e., image-to-image), this time against NVIDIA SPADE, pSp and Palette, using FID and mean Intersection of Union (mIoU):

Methods	FID↓	mIoU(%)↑
SPADE [27]	29.86	60.37
pSp [30]	62.78	53.95
Palette (@500 epoch) [33]	33.82	60.96
LDM (@200 epoch) [31]	19.44	64.17
LayoutDiff. (@30 epoch)	**8.11**	**67.05**

Here again, LayoutDiffuse obtained the highest scores, with the authors commenting:

'[The] diffusion-based methods achieve better performance. [Palette], though being a diffusion-based method, does not perform as well as LayoutDiffuse and LDM, which can largely be due to the fact that it is trained directly in RGB pixel space, which is more difficult. Compare to LDM, LayoutDiffuse achieves better performance with much [fewer] training epochs.'

Finally a test was conducted for 'scene recognizability', this time using YOLO Score and SceneFID. For this, the authors trained a ResNet-101 on generated crops of objects:



Here, the authors state:

'Our method outperforms the best non-diffusion baseline by a large margin. When comparing to LDM, LayoutDiffuse still obtains more than 4%/6% improvement regarding AP/AR, illustrating the efficiency and effectiveness of our method.'

Conclusion

LayoutDiffuse offers a method of compositionality that's hindered by one major factor – it would need to either be adopted by the originators of the checkpoints (in the case of Stable Diffusion, Stability.ai), or else would need collective and probably crowdfunded efforts to generate community-driven layout-aware versions of the latest models – all of which would need to be updated periodically as the checkpoints evolve and advance.

Though the fine-tuning requirements are orders of magnitude lower than the almost-inconceivable prospect of training on LAION from scratch, the resources required are not obtainable by hobbyists or enthusiasts, who may for the time being prefer such *post facto* layout solutions as will become available in this frenetic and productive period in latent diffusion research and development.