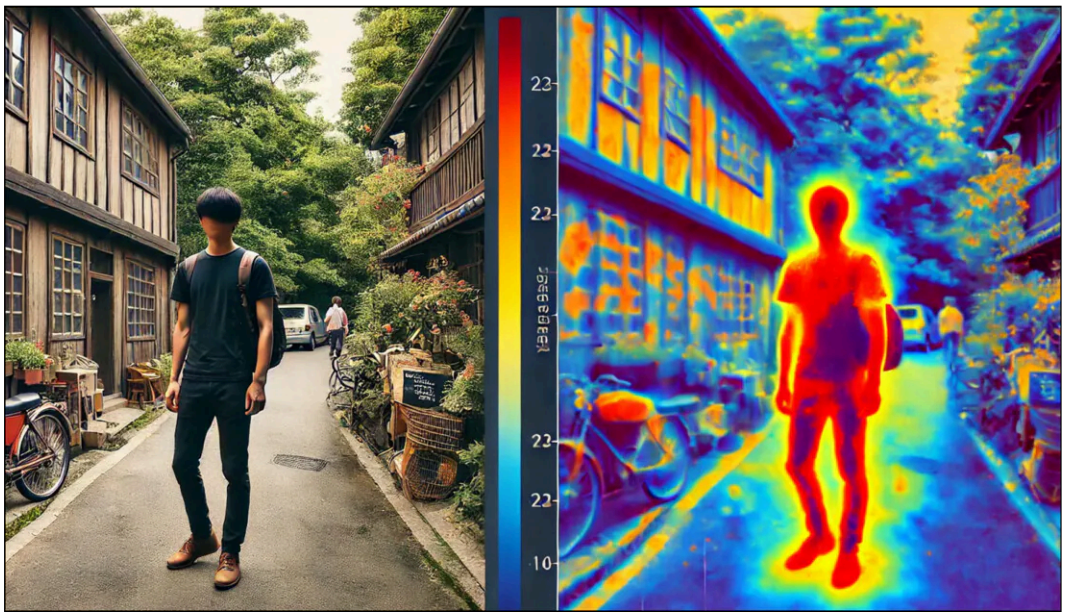
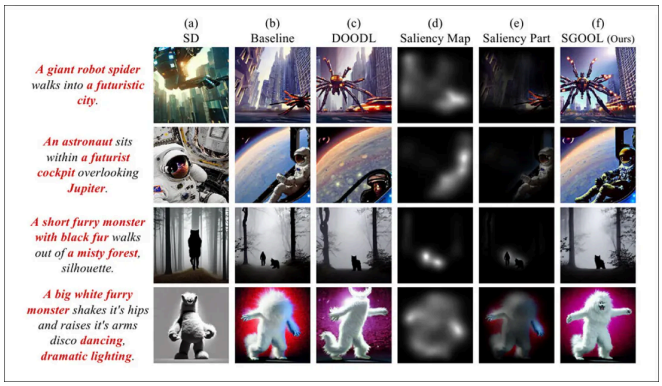
Leveraging Human Attention Can Improve AI-Generated Images

Originally published at https://www.unite.ai/leveraging-human-attention-can-improve-ai-generated-images/



New research from China has proposed a method for improving the quality of images generated by Latent Diffusion Models (LDMs) models such as Stable Diffusion.

The method focuses on optimizing the *salient regions* of an image – areas most likely to attract human attention.



The new research has found that saliency maps (fourth column from left) can be used as a filter, or 'mask', for steering the locus of attention in denoising processes towards areas of the image that humans are most likely to pay attention to. Source: https://arxiv.org/pdf/2410.10257

Traditional methods, optimize the *entire image* uniformly, while the new approach leverages a saliency detector to identify and prioritize more 'important' regions, as humans do.

In quantitative and qualitative tests, the researchers' method was able to outperform prior diffusion-based models, both in terms of image quality and fidelity to text prompts.

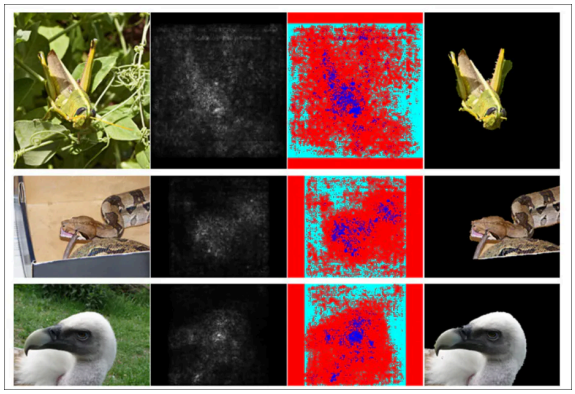
The new approach also scored best in a human perception trial with 100 participants.

Natural Selection

Saliency, the ability to prioritize information in the real world and in images, is an essential part of human vision.

A simple example of this is the increased attention to detail that classical art assigns to important areas of a painting, such as the face, in a portrait, or the masts of a ship, in a sea-based subject; in such examples, the artist's attention converges on the central subject matter, meaning that broad details such as a portrait background or the distant waves of a storm are sketchier and more broadly representative than detailed.

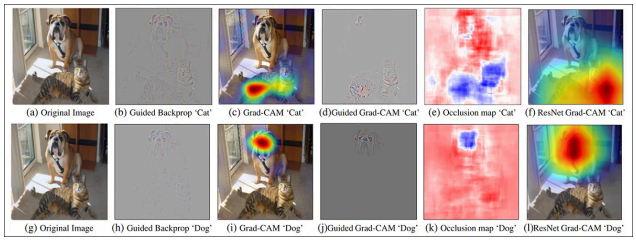
Informed by human studies, machine learning methods have arisen over the last decade that can replicate or at least approximate this human locus of interest in any picture.



Object segmentation (semantic segmentation) can be an aide in individuating facets of an image, and developing corresponding saliency maps. Source: https://arxiv.org/pdf/1312.6034

In the run of research literature, the most popular saliency map detector over the last five years has been the 2016 Gradient-weighted Class Activation Mapping (Grad-CAM) initiative, which later evolved into the improved Grad-CAM++ system, among other variants and refinements.

Grad-CAM uses the gradient activation of a semantic token (such as 'dog' or 'cat') to produce a visual map of where the concept or annotation seems likely to be represented in the image.



Examples from the original Grad-CAM paper. In the second column, guided backpropagation individuates all contributing features. In the third column, the semantic maps are drawn for the two concepts 'dog' and 'cat'. The fourth column represents the concatenation of the previous two inferences. The fifth, the occlusion (masking) map that corresponds to the inference; and finally, in the sixth column, Grad-CAM visualizes a ResNet-18 layer. Source: https://arxiv.org/pdf/1610.02391

Human surveys on the results obtained by these methods have revealed a correspondence between these mathematical individuations of key interest points in an image, and human attention (when scanning the image).

SGOOL

The new paper considers what saliency can bring to text-to-image (and, potentially, text-to-video) systems such as Stable Diffusion and Flux.

When interpreting a user's text-prompt, Latent Diffusion Models explore their trained latent space for learned visual concepts that correspond with the words or phrases used. They then parse these found data-points through a denoising process, where random noise is gradually evolved into a creative interpretation of the user's text-prompt.

At this point, however, the model gives *equal attention to every single part of the image*. Since the popularization of diffusion models in 2022, with the launch of OpenAI's available Dall-E image generators, and the subsequent open-sourcing of Stability.ai's Stable Diffusion framework, users have found that 'essential' sections of an image are often under-served.

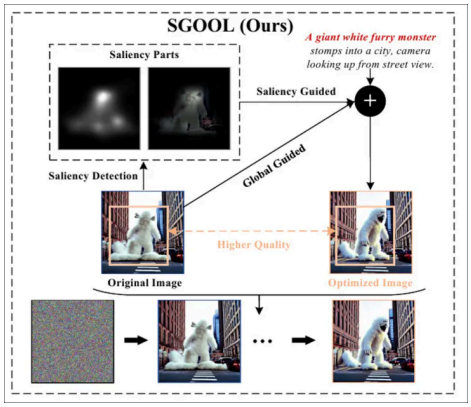
Considering that in a typical depiction of a human, the person's face (which is of maximum importance to the viewer) is likely to occupy no more than 10-35% of the total image, this democratic method of attention dispersal works against both the nature of human perception and the history of art and photography.

When the buttons on a person's jeans receive the same computing heft as their eyes, the allocation of resources could be said to be non-optimal.

Therefore, the new method proposed by the authors, titled *Saliency Guided Optimization of Diffusion Latents* (SGOOL), uses a saliency mapper to increase attention on neglected areas of a picture, devoting fewer resources to sections likely to remain at the periphery of the viewer's attention.

Method

The SGOOL pipeline includes image generation, saliency mapping, and optimization, with the overall image and saliency-refined image jointly processed.



Conceptual schema for SGOOL.

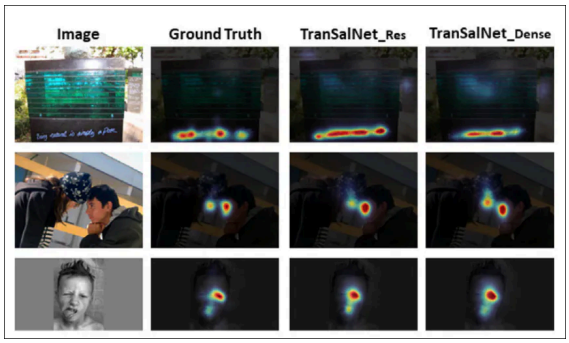
The diffusion model's latent embeddings are optimized directly with fine-tuning, removing the need to train a specific model. Stanford University's Denoising Diffusion Implicit Model (DDIM) sampling method, familiar to users of Stable Diffusion, is adapted to incorporate the secondary information provided by saliency maps.

The paper states:

> *'We first employ a saliency detector to mimic the human visual attention system and mark out the salient regions. To avoid retraining an additional model, our method directly optimizes the diffusion latents.*
>
> *'Besides, SGOOL utilizes an invertible diffusion process and endows it with the merits of constant memory implementation. Hence, our method becomes a parameter-efficient and plug-and-play fine-tuning method. Extensive experiments have been done with several metrics and human evaluation.'*

Since this method requires multiple iterations of the denoising process, the authors adopted the Direct Optimization Of Diffusion Latents (DOODL) framework, which provides an invertible diffusion process – though it still applies attention to the entirety of the image.

To define areas of human interest, the researchers employed the University of Dundee's 2022 TransalNet framework.



Examples of saliency detection from the 2022 TransalNet project. Source: https://discovery.dundee.ac.uk/ws/portalfiles/portal/89737376/1_s2.0_S0925231222004714_main.pdf

The salient regions processed by TransalNet were then cropped to generate conclusive saliency sections likely to be of most interest to actual people.

The difference between the user text and the image has to be considered, in terms of defining a loss function that can determine if the process is working. For this, a version of OpenAI's Contrastive Language–Image Pre-training (CLIP) – by now a mainstay of the image synthesis research sector – was used, together with consideration of the estimated semantic distance between the text prompt and the global (non-saliency) image output.

The authors assert:

> *'[The] final loss [function] regards the relationships between saliency parts and the global image simultaneously, which helps to balance local details and global consistency in the generation process.*
>
> *'This saliency-aware loss is leveraged to optimize image latent. The gradients are computed on the noised [latent] and leveraged to enhance the conditioning effect of the input prompt on both salient and global aspects of the original generated image.'*

Data and Tests

To test SGOOL, the authors used a 'vanilla' distribution of Stable Diffusion V1.4 (denoted as 'SD' in test results) and Stable Diffusion with CLIP guidance (denoted as 'baseline' in results).

The system was evaluated against three public datasets: CommonSyntacticProcesses (CSP), DrawBench, and DailyDallE*.

The latter contains 99 elaborate prompts from an artist featured in one of OpenAI's blog posts, while DrawBench offers 200 prompts across 11 categories. CSP is composed of 52 prompts based on eight diverse grammatical cases.

For SD, baseline and SGOOL, in the tests, the CLIP model was used over ViT/B-32 to generate the image and text embeddings. The same prompt and random seed was used. The output size was 256×256, and the default weights and settings of TransalNet were employed.

Besides the CLIP score metric, an estimated Human Preference Score (HPS) was used, in addition to a real-world study with 100 participants.

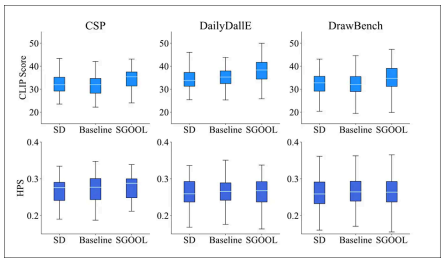
Metrics	CLIP Score				HPS			
	CSP	DailyDallE	DrawBench	Average	CSP	DailyDallE	DrawBench	Average
SD	31.67	31.67	32.15	31.83	0.2650	0.2573	0.2595	0.2606
Baseline	31.61	35.20	31.92	32.91	0.2705	0.2622	0.2627	0.2651
SGOOL	**34.84**	**38.11**	**34.93**	**35.96**	**0.2778**	**0.2630**	**0.2632**	**0.2680**

Quantitative results comparing SGOOL to prior configurations.

In regard to the quantitative results depicted in the table above, the paper states:

> *'[Our] model significantly outperforms SD and Baseline on all datasets under both CLIP score and HPS metrics. The average results of our model on CLIP score and HPS are 3.05 and 0.0029 higher than the second place, respectively.'*

The authors further estimated the box plots of the HPS and CLIP scores in respect to the previous approaches:



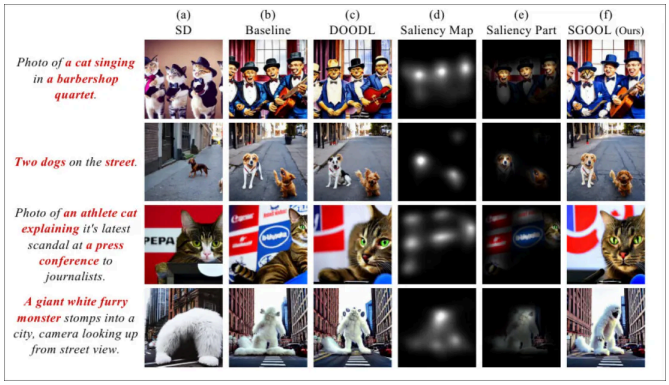
Box plots for the HPS and CLIP scores obtained in the tests.

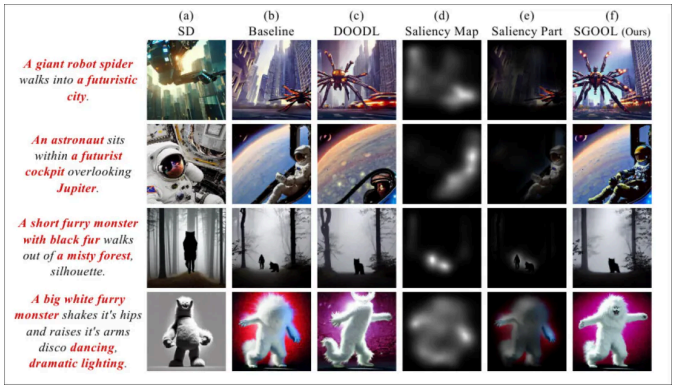
They comment:

> *'It can be seen that our model outperforms the other models, indicating that our model is more capable of generating images that are consistent with the prompts.*
>
> *'However, in the box plot, it is not easy to visualize the comparison from the box plot due to the size of this evaluation metric at [0, 1]. Therefore, we proceed to plot the corresponding bar plots.*
>
> *'It can be seen that SGOOL outperforms SD and Baseline on all datasets under both CLIP score and HPS metrics. The quantitative results demonstrate that our model can generate more semantically consistent and human-preferred images.'*

The researchers note that while the baseline model is able to improve the quality of image output, it does not consider the salient areas of the image. They contend that SGOOL, in arriving at a compromise between global and salient image evaluation, obtains better images.

In qualitative (automated) comparisons, the number of optimizations was set to 50 for SGOOL and DOODL.





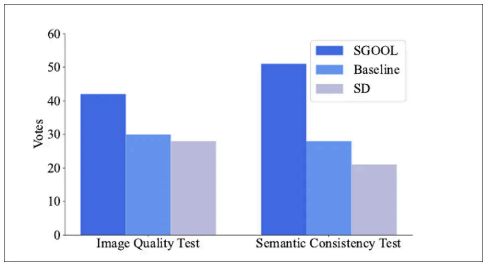
Qualitative results for the tests. Please refer to the source paper for better definition.

Here the authors observe:

> *'In the [first row], the subjects of the prompt are "a cat singing" and "a barbershop quartet". There are four cats in the image generated by SD, and the content of the image is poorly aligned with the prompt.*
>
> *'The cat is ignored in the image generated by Baseline, and there is a lack of detail in the portrayal of the face and the details in the image. DOODL attempts to generate an image that is consistent with the prompt.*
>
> *'However, since DOODL optimizes the global image directly, the persons in the image are optimized toward the cat.'*

They further note that SGOOL, by contrast, generates images that are more consistent with the original prompt.

In the human perception test, 100 volunteers evaluated test images for quality and semantic consistency (i.e., how closely they adhered to their source text-prompts). The participants had unlimited time to make their choices.



Results for the human perception test.

As the paper points out, the authors' method is notably preferred over the prior approaches.

Conclusion

Not long after the shortcomings addressed in this paper became evident in local installations of Stable Diffusion, various bespoke methods (such as After Detailer) emerged to force the system to apply extra attention to areas that were of greater human interest.

However, this kind of approach requires that the diffusion system initially go through its normal process of applying equal attention to every part of the image, with the increased work being done as an extra stage.

The evidence from SGOOL suggests that applying basic human psychology to the prioritization of image sections could greatly enhance the initial inference, without post-processing steps.

* *The paper provides the same link for this as for CommonSyntacticProcesses.*

First published Wednesday, October 16, 2024

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

Discover More