UKRI AI Centre for Doctoral Training in Digital Healthcare

Imperial College London

Research Tutorial Report

Hierarchical Text-Conditional Image Generation with CLIP Latents

Ramesh et al., 2022

Noura Ezaz-Nikpay
February 11, 2025

1 Paper Summary

The paper "Hierarchical Text-Conditional Image Generation with CLIP Latents", published as an arXiv pre-print in 2022 [1], proposes a novel architecture for text-to-image generation. Although it has had a high impact, with over 6700 citations to date, it remains as a preprint. The authors, Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, were researchers at OpenAI at the time of publishing. Their proposed architecture (unCLIP) succeeds at generating photorealistic and semantically rich images from text. By using CLIP image and text embeddings [2], unCLIP is a two-stage hierarchical model that consists of a prior and a decoder that allow for text-based image manipulation. In comparison to precedent models, unCLIP is able to generate diverse images that retain style and semantic meaning, exhibiting less of a fidelity-diversity trade-off. This research tutorial report focuses on the paper in the context of its relevance to AI for healthcare.

1.1 Background: AI architectures

Using Constrastive Language-Image Pre-Training (CLIP) [2], which learns a joint representation space for text and images, this paper proposes a two-stage architecture: a prior that produces CLIP image embeddings conditioned on a text caption, followed by a decoder that generates the output image conditioned on the CLIP image embedding produced by the prior. Before detailing the specific architectures that constitute the unCLIP model, there are two key architecture types to first understand: CLIP and stable diffusion models.

1.1.1 CLIP

The paper builds on a precedent model, CLIP [2], which enabled the translation between image and text modalities. CLIP connects text and images through a joint representation space, such that (image, text) pairs are mapped to similar locations. It is trained using a contrastive learning approach that takes batches of (image, text) pairs and maximises the similarity between the embeddings of the correct pairs and minimising that of the incorrect pairs. In this way, visual concepts are associated with their textual counterparts. unCLIP uses the pre-trained CLIP model, which can perform well on unseen tasks without requiring task-specific training (zero-shot prediction, as indicated in Fig. 1.)

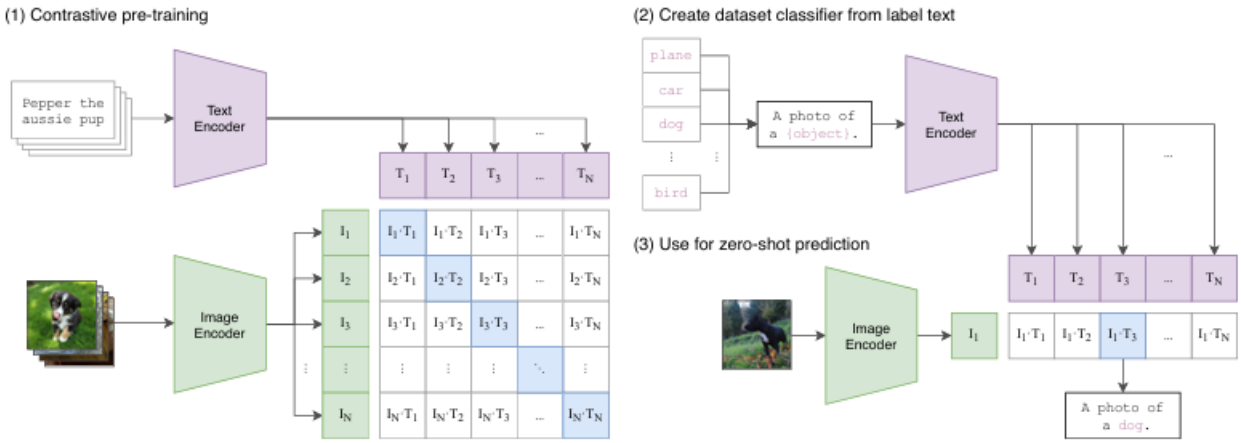

Figure 1: Summary of CLIP, which jointly trains an image and a text encoder that can then be used to predict the textual class of an image. [2]

1.1.2 Stable diffusion models

Diffusion models, exemplified in Fig. 2, are a class of generative models. By learning to predict the original data from the noised versions, a diffusion model generates samples that resemble the data it was trained on. Stable diffusion models are a specific subtype that operate in a lower-dimensional latent space instead of in the original data space. This increases efficiency as well as computational cost and memory requirements. After diffusion, the processed latent rep is then decoded back into the original data space. In the case of unCLIP, the stable diffusion models used gradually add Gaussian noise to the original data over multiple time-steps, from which the model needs to predict the original unnoised data. During inference, the model starts with a randomly sampled noise vector, resulting in high-quality and diverse samples can be created.

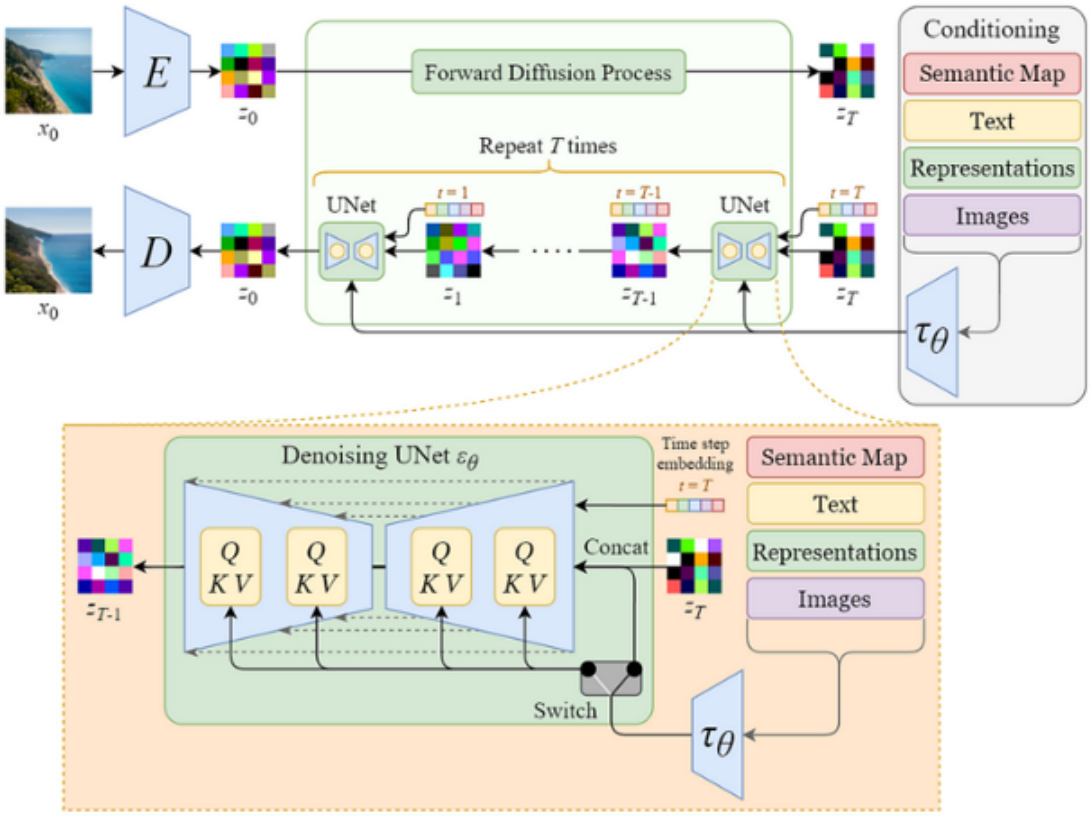

Figure 2: Example of stable diffusion model architecture, with a UNet used for denoising. An image is encoded and noise is added T times in the forward diffusion process. In the reverse process, the noised input is iteratively denoised, using conditioning information. [3]

1.2 Method: unCLIP

Using an initial frozen CLIP representation space, unCLIP consists of two stages: the prior and the decoder. Each uses distinct architectures, each component of which will be subsequently described. The overarching hierarchical prior-decoder structure is based on equation 1, where x is the image, y is the caption, and z_i is the CLIP image embedding.

$$P(x|y) = P(x, z_i|y) = P(x|z_i, y)P(z_i|y) \tag{1}$$

So, to sample from $P(x|y)$, it is possible to first sample z_i using the prior $P(x|z_i)$ and then sample x using the decoder $P(x|z_i, y)$.

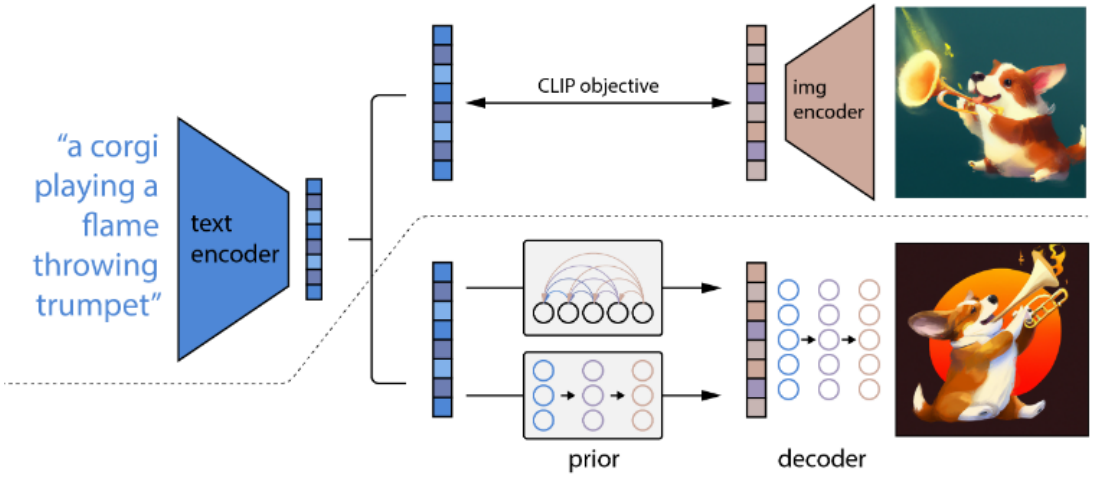

Figure 3: The unCLIP model. Above the dotted line, the CLIP training process is shows. Below the line, the two stage text-to-image generation process is shown [1].

The prior produces CLIP image embeddings conditioned on captions and the CLIP text embedding (which is merely a deterministic function of the caption). The paper introduces two alternatives for a prior: a stable diffusion model and an autoregressive model. For both of these priors, during training the conditioning information was dropped 10% of the time to improve robustness. The image embedding produced by the prior is then used by the decoder to generate the final image that corresponds to the initial caption.

1.2.1 Diffusion prior

Fig. 2 shows a general diffusion model framework. In the case of the prior, the image is the CLIP image embedding, with a Gaussian forward diffusion process. Instead of the UNet, the prior uses a decoder-only transformer with masked self-attention to ensure the model can only attend to elements that precede the current position in the input sequence. This means the value for each dimension of the unnoised CLIP image embedding is predicted sequentially, conditioning each prediction on the previously generated dimensions and the input sequence (consisting of the encoded caption, the CLIP text embedding, a timestep embedding and the noised CLIP image embedding). Although the paper does not specify the reason for sequential dimension prediction, it may be that it allows for a more controlled and systematic generation process during inference.

The diffusion model is trained to directly predict the unnoised image embedding, using a mean squared error loss. To improve quality during sampling, two samples of the image embedding is generated and the one with the higher dot product with the text embedding is chosen. A higher dot product indicates a stronger alignment between the two (as they are in a joint representation space), suggesting that the caption more accurately describes the generated image.

1.2.2 Autoregression prior

Principle Component Analysis (PCA) is used to reduce dimensionality with minimal information loss, from 1024 to 319 dimensions. These principle components, ordered by decreasing eigenvalue magnitude, are quantised into 1024 discrete levels i.e. each continuous value within the 319 dimensions is mapped to one of the 1024 discrete levels. A transformer with masked self-attention then predicts this resulting sequence of discrete codes autoregressively. Although

the choice for discretising the values is not discussed explicitly, it may be because continuous sequence modeling is more challenging and hence harder to train. This is unlike the diffusion prior where the slow addition of noise lends itself to a continuous modeling approach.

In contrast to the diffusion prior, the autoregressive prior also conditions on the quantised dot product between the text and image embedding by prepending a token representing this dot product to the input sequence. This explicitly encodes the desired level of image-caption alignment, encouraging the generation of CLIP image embeddings that are semantically consistent with the inputted caption.

1.2.3 Decoder

The decoder generates an image conditioned on the CLIP image embeddings, by modifying the 3.5 billion parameter GLIDE diffusion model [4] to incorporate CLIP image embeddings with optional conditioning on the text captioning. To increase the resolution, the resulting image is up-sampled twice (from 64x64 to 256x256, then 256x256 to 1024x1024) using ADMNet [5], a diffusion model designed for up-sampling tasks. The up-sampling is made more robust through corrupting the conditioning images during training.

To improve sample quality, classifier-free guidance, where CLIP image embeddings and the caption are dropped with a certain probability so that the model learns to generate images from less informative conditioning signals.

1.3 Results

The paper demonstrates the performance of unCLIP in a range of different ways, such as the importance of the prior in terms of diversity, photorealism and caption similarity, using both human evaluation and the Fréchet inception distance (FID). FID is a measure of diversity that compares the distribution of generated images with the distribution of a set of real images. It does not always match human judgment, motivating the use of human evaluation alongside it. They investigate the aesthetic quality and, in line with standard practices, evaluate FID on the MS-COCO validation set which it was not directly trained on. Compared to other zero-shot models like GLIDE, unCLIP achieves a new state-of-the-art FID of 10.39 when sampling with the diffusion prior. They find that the diffusion prior generally performs better, being both more computationally efficient and producing higher-quality samples.

1.3.1 Fidelity and Diversity trade-off

It is interesting to focus on a particular result, namely that unCLIP avoids the diversity-fidelity trade-off compared to GLIDE. GLIDE directly uses a caption to guide the diffusion process using a text encoder, which makes it very sensitive to the guidance scale (a parameter that controls how strongly the model adheres to the caption). Fidelity therefore tends to come at the cost of diversity because the guidance scale directly controls the whole generation process, including the semantic interpretation of the caption. In contrast, in unCLIP the guidance scale predominantly affects the decoder's refinement of details within the abstracted semantic framework the CLIP image embedding defines. This means that, even at higher guidance levels (higher fidelity), higher diversity can be achieved. In the context of healthcare, retaining fidelity is crucial. If diversity of the data comes at the cost of accuracy then there would be a negative downstream effect on the models trained on the synthetic data.

1.3.2 Image manipulation

unCLIP has functionality that is not merely limited to text-to-image generation. It allows for the creation of variations of an input image (whilst preserving style and semantic meaning) by first encoding it with CLIP and then decoding with a diffusion model (of variable stochasticity).

Fig. 4 demonstrates how this also allows for the seamless blending of two images. Similarly, given that CLIP embeds images and texts to the same latent space, language-guided image manipulation is possible through interpolating CLIP text and image embeddings. This increases the control of the diversity of the space that is traversed, which is salient to data augmentation.

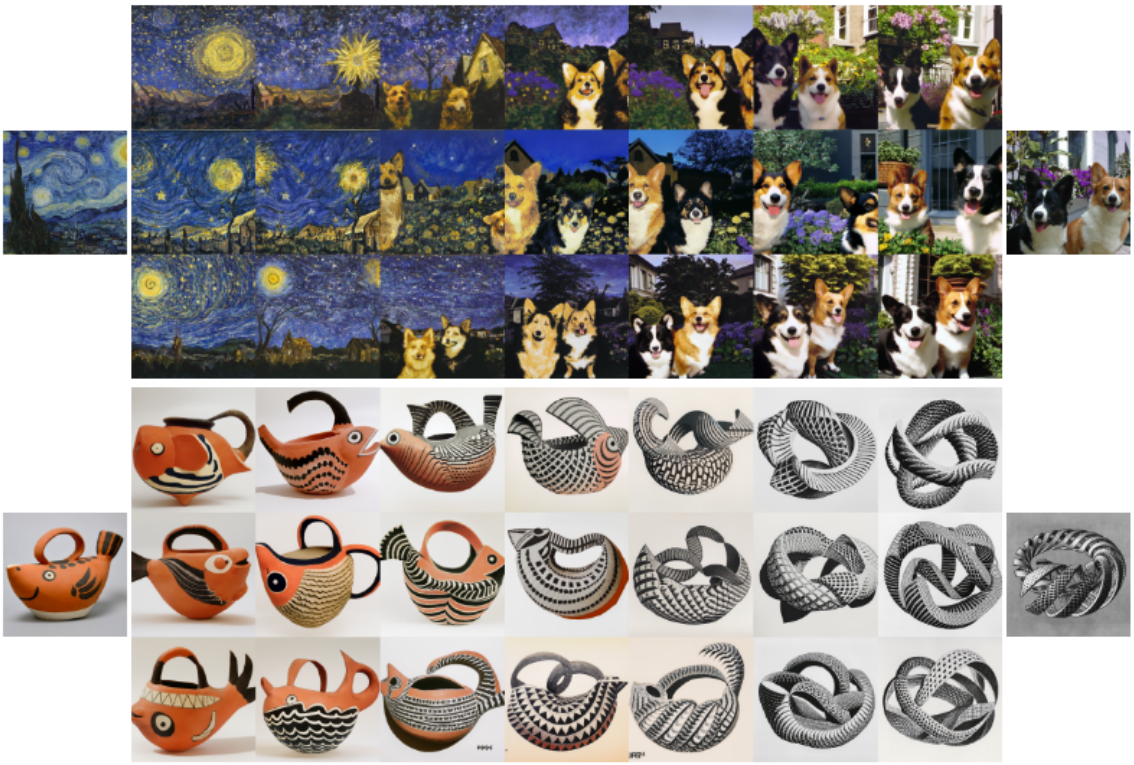

Figure 4: Variations between two images generated by interpolating their CLIP image embeddings. [1]

2 Critique and Discussion

2.1 Relevance to healthcare

Generative modeling is valuable in the context of data augmentation, including for healthcare applications [6] gathering high-quality and high-quantity training data that effectively covers the space is a significant challenge. AI advances that contribute to the potential for synthetic generation of novel medical imaging therefore have valuable practical applications to healthcare. Other valuable functionalities enabled includes the possibility to enhance low-resolution images and translate between different modalities, such as through image-to-image translation [7]. unCLIP is one such advance in AI generative modeling, allowing the generation of photorealistic and diverse images from text, as well as producing diverse but faithful variations of an image by retaining semantics and style. However, this method was not developed with healthcare

in mind and so the risks associated with it need to be considered carefully in the context of healthcare applications.

2.2 Risks specific to healthcare

Although unCLIP offers exciting opportunities for text-informed image augmentation for healthcare, there are also significant limitations. In the context of possible healthcare applications, these limitations are significant enough to prevent the use of unCLIP as it is presented in the paper.

One such limitation is that unCLIP is poor at binding attributes to each other, performing worse than GLIDE at this. In medical imaging, this could mean that it fails to correctly bind attributes like size, shape and location of a tumour to each other, rendering it unusable for informing diagnosis or treatment. This plays a role in unCLIP struggling to produce coherent text, which would affect annotations on medical images. Similarly, for complex scenes, the level of detail is too low. A higher base resolution would be needed, which increases the cost.

Another key limitation is the lack of information about how unCLIP learns biases in the training data. Understanding this is crucial for fairness in healthcare and ensuring no patient group is underrepresented or discriminated against. It is also less possible to identify the outputs as AI generated, which would mean they could be mistaken for the real patient data.

2.3 Impact

Even though the paper exists only as a preprint from 2022, it already has over 6600 citations. These papers range from subject driven generation that binds a unique identifier to a specific subject [8] to text-to-3D image generation [9] and text-to-video generation [10]. It has therefore clearly had significant impact within computer vision, for example it is deployed in OpenAI's DALL.E2 model. There does not, however, seem to yet be a specific healthcare application use-case. The risks, mentioned above, will first need to be mitigated and addressed. However, if and when synthetic medical image data does become widespread in being able to train AI-powered healthcare systems, it will be off the back of advancements such as unCLIP.

3 Acknowledgments

This work was supported by UK Research and Innovation [UKRI AI Centre for Doctoral Training in Digital Healthcare number EP/Y030974/1].

References

[1] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[3] AIHGF. Stable diffusion. https://www.aiuai.cn/aifarm2096.html, 2023. Accessed: 2024-12-13.

[4] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2022.

[5] Xiaofei Zhou, Kunye Shen, and Zhi Liu. Admnet: Attention-guided densely multi-scale network for lightweight salient object detection. *IEEE Transactions on Multimedia*, 26:10828–10841, 2024.

[6] Jinzhuo Wang, Kai Wang, Yunfang Yu, Yuxing Lu, Wenchao Xiao, Zhuo Sun, Fei Liu, Zixing Zou, Yuanxu Gao, Lei Yang, et al. Self-improving generative foundation model for synthetic medical image generation and clinical applications. *Nature Medicine*, pages 1–9, 2024.

[7] Khaled ELKarazle, Valliappan Raman, Patrick Then, and Caslon Chua. *How Generative AI Is Transforming Medical Imaging: A Practical Guide*, pages 371–385. Springer International Publishing, Cham, 2024.

[8] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, June 2023.

[9] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan LI, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 8406–8441. Curran Associates, Inc., 2023.

[10] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15954–15964, October 2023.