International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

Adaptive Dynamic Image Generation through User Relevance Feedback

[1]K.Phaneendra Kumar, [2]Neelima G, [3]Srinivas Ganganagunta
[1]Dept of CSE, Vignan's Lara Institute of Technology &Sciences, Guntur, AP, India
[2] Dept of CSE , Acharya Nagarjuna University, Nagarjuna Nagar, Guntur, AP, India
[3] Dept of Physics, University of Technology and Applied Sciences – IBRA, Sultan of Oman
[1]Phaneendra2008@gmail.com, [2]neelima.guntupalli80@gmail.com, [3]ganganagunta.srinivas@utas.edu.om

Abstract-Text-to-image generation has seen remarkable progress with the emergence of deep learning models like Stable Diffusion.These models allow for high-quality and customized image creation. However, conventional approaches often require extensive computational resources and subject-specific fine- tuning, limiting their scalability and accessibility. Instant Imager eliminates this need by leveraging in-context learning, enabling the model to replicate the capabilities of numerous subject-specific expert models. This innovation allows for the instant creation of high-flexibility and creative potential. The framework's efficiency is evident in its ability to generate customized images 10 times faster than the conventional optimization methods, while delivering user specific number of images with superior quality. Evolutions on stable-diffusion- v1-5 and stable-diffusion-sdxl-turbo datasets highlight its performance, consistency outperforming existing models as confirmed by generation process. In addition to speed and quality, Instant Imager streamlines the image generation process, making it an essential tool for artists, designers, and content creators.

Keywords: Text-to-Image Generation, Stable Diffusion, In-Context Learning, Stable-diffusion-v1-5, Expert Models, High- Fidelity Images.

I.INTRODUCTION

In recent years, text-to-image generation has improved a lot, thanks to deep learning models like Stable Diffusion. These models allow users to create high-quality, personalized images from text input. This change is reshaping digital content creation.

However, conventional methods typically require extensive computational resources and subject-specific fine- tuning, which constrain their scalability and practical deployment.

Recent advances have highlighted the potential of integrating multiple expert models to overcome these challenges. Drawing inspiration from hybrid approaches in other fields [1], our proposed framework--Instant Imager--employs in-context learning to amalgamate the capabilities of numerous subject-specific models into a single, agile system. This approach eliminates the need for individual model optimization substantially reducing processing time and computational overhead. A key innovation and distinguishing future of Instant Imager is its intuitive user interface component-a range bar-that empowers users to dynamically select the precise number of images they wish to generate. This feature not only streamlines the image generation process but also enhances user interactivity by offering granular control over output quantity, making the system highly accessible even for non-expert users

Experimental evolutions on benchmark datasets including Stable-diffusion-v1-5 and Stable-Diffusion- SDXL-Turbo, demonstrate that Instant Imager can generate customized images up to 10 times faster than conventional optimization methods without compromising quality. The range bar contributes to the sufficiency by allowing users to directly control the output thereby reducing unnecessary computation cycles and simplifying the generation workflow.

Furthermore, Instant images represent a significant advancement in text-to-image generation by bridging the gap between sophisticated in-context learning techniques and user-friendly design. By empowering users with direct control over image quantity, the framework not only enhances operational efficiency but also democratizes access to high quality image synthesis for diverse range of creative applications. This work establishes a new benchmark for fast, scalable, and customizable image

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

generation, opening the door to innovative applications in digital art, design, and content creation.

II. RELATED WORKS

Text-to-image generation has experienced a transformative evolution over the past decade, driven by the rapid advancements in deep learning and generative modeling techniques. Early approaches in this domain predominantly relied on Generative-Adversarial-Networks (GANs) and Variational-Autoencoders (VAEs) to synthesize images from textual descriptions. Although GANs and VAEs laid the foundational groundwork, they often encountered issues such as mode collapse, unstable training dynamics, and limitations in generating high-fidelity images—especially when tasked with producing images from complex or nuanced text inputs.

With the emergence of diffusion-based models, image synthesis has undergone a major transformation. These models progressively refine a random noise pattern into a coherent image through a series of iterative steps.

Models such as DALL-E and CLIP-based approaches demonstrated that harnessing large-scale datasets and sophisticated training paradigms could overcome many of the inherent limitations found in earlier methods. Among these, the Stable Diffusion framework has quickly emerged as a prominent solution, delivering a compelling balance between the image quality and computational efficiency.

Recent iterations of this framework--specifically Stable-Diffusion 3.5-Large-Turbo and Stable Diffusion-SDXL-Turbo- represent notable advancements in the field. These models have been engineered to enhance image fidelity while reducing inference times and computational demands. By optimizing network architectures and leveraging large-scale pre-training, these versions address previous challenges such as long synthesis times and the need for extensive fine-tuning when adapting to a new subjects or styles.

In recent years, researchers have increasingly drawn inspiration from in-context learning paradigms, which have shown remarkable success in natural language processing as well as various computer vision applications.In-Context Learning enables a model to leverage the information contained in a prompt or context to generate relevant outputs without the need for extensive retraining. This approach offers the dual benefits of enhanced scalability and reduced computational load.

In summary while significant strides have been made with diffusion-based models especially with advancements such as Stable Diffusion, existing approaches still contend with challenges related to scalability, efficiency, and adaptability. By integrating strengths of in- context learning Instant Imager consolidate the specialized capabilities of numerous subject specific expert models into a single unified system.

III. PROPOSED SYSTEM

To reduce the computational demands of high-resolution image synthesis, we observed that while diffusion models can omit details that have minimal impact on human perception—thereby requiring fewer loss calculations—they still process every pixel individually. This pixel-by-pixel computation is both time- and energy-intensive.

To address this challenge, we split the learning processes into two separate phases: a compression phase and a generative phase. These methods ultimately make high-resolution image generation more practical by significantly reducing the computational resources required while still maintaining the image quality.

A. Perceptual image compression

The perceptual compression model builds on earlier work by employing an auto encoder trained with both perceptual and patch-based adversarial losses. This approach is designed to capture high-level, human-perceptible features while ensuring the local image details remain realistic and free from blurriness and often associated with simple pixel-based losses like L2 or L1.

To break it down further, let's consider an input image x in RGB format with dimensions HxWx3. The encoder E processes this image and produces a compact latent representation z = E(x) that has reduced dimensions h x w x c. In this process, the encoder downscales the image by a factor f (where f = H/h = W/w). We experiment with different down sampling factors that are powers of two (i.e., f = 2^m with m ∈ ℕ), which provides a structured reduction in complexity while preserving essential image details.

One major challenge for compressing images is preventing the latent space from becoming overly noisy or having uncontrolled variance. To tackle this, we integrate two types of regularization into our model:

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

1. KL Regularization (KL-reg):

Here a modest Kullback-Leibler divergence penalty is imposed on the latest distribution. This penalty gently nudges the latent coach towards a standard normal distribution much like what is done in Variational Autoencoders (VAEs). This regularization helps in maintaining a well-behaved latent space without forcing too much distortion.

2. Vector Quantization Regularization (VQ-reg):

In this variant we incorporate vector quantization layer directly into the decoder. This setup which can be viewed as variant of VQGAN, effectively discretizes the latent space. By doing so, it captures the essential features in a more structured form while reducing the risk of high-variance outputs.

The key advantage of our model is that it maintains the two-dimensional spatial structure of the latent space. This is particularly beneficial for subsequent diffusion models, which are designed to work with such 2D data. Unlike earlier methods that flatten the latent space into a one-dimensional sequence for autoregressive modeling—resulting in the loss of crucial spatial relationships—our approach maintains the inherent spatial structure of the image. This allows our compression model to operate with relatively mild compression rates while still producing high-quality reconstructions, preserving more details from the original image. This not only reduces the computational load by operating in a lower-dimensional space but also ensures that the compressed representations retain the essential characteristics and find details of the original images

B. Latent Diffusion Models

Diffusion models, as probabilistic frameworks, learn data distributions by gradually denoising a normally distributed random variable over a series of steps.

In practice, they similar the reverse process of a fixed lint Markov chain where each step gradually cleans up the noise. For image synthesis, leading models use a modified version of the variational lower bound on the data distribution a method similar to denoising score matching. These models can interpret equally weighted sequence of denoising autoencoders Eθ(xt, t); t = 1. . . T, which are trained to predict denoised variant of their input xt, where xt is a noisy version of the input x. The corresponding objective can be further simplified to,

$$L_{DM} = \mathrm{E}_{x,c\sim N(0,1),t} \, /E - E\theta(x_t, t)/^2{}_2, \qquad (1).$$

with t uniformly sampled from {1, …, T}.

Building on this, our approach takes advantage of a perceptual compression model consisting of an Encoder(E) and a Decoder(D) that transforms high-dimensional images into a compact low-dimensional latent space.

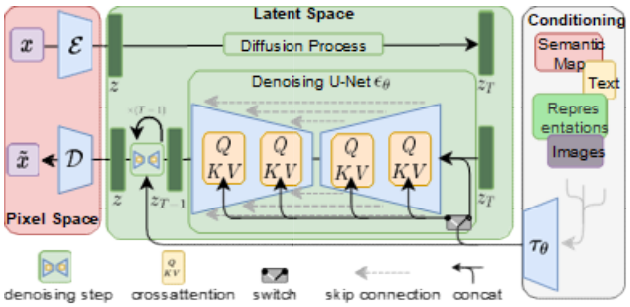

Figure 1. We condition Latent Diffusion Models (LDMs) using either simple concatenation of inputs or a more flexible and general cross-attention mechanism.

The revised objective for our latent diffusion model becomes one where the network implemented as a time conditional UNet-is Trained to denoise the latent representation z (obtained from encoder E). Because the forward process is fixed, we can efficiently compute the noisy latent at each step during training. Once trained, our model can generate samples in the latent space which are then quickly converted back into full resolution images by single pass-through decoder D.

$$L_{LDM} := \mathrm{E}_{E(x),c\sim N(0,1),t} \, /E - E\theta(z_t, t)/^2{}_2 \qquad (2)$$

C. Mechanism of Conditioning

Diffusion models can be adapted to generate images conditioned on additional inputs—such as text, semantic maps, or other images—by modeling conditional distributions of the form p(z | y).To enable this, we modify the standard UNet backbone of our diffusion model by incorporating a cross-attention mechanism, allowing the model to focus on the most relevant features of the conditional input.We first pass the input condition (y) through a domain specific encoder τθ that transforms it into an immediate representation. This representation is then fused into UNet via cross-attention enabling the model to

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

guide the denoising process accordingly. Both the UNet and the encoder are trained together using pairs of images and conditions.

Based on image-conditioning pairs, we then learn the conditional LDM via

$$L_{LDM} := \mathrm{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t} \left\| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y)) \right\|_2^2 \quad (3)$$

Both τθand Eθare jointly optimized using Equation (3). This conditional mechanism is highly flexible, as τθcan be parameterized with domain-specific expert models—for example, using unmasked transformers when the conditioning input y consists of text prompts.

D. Experiments

Our experiments demonstrate the Latent Diffusion Models (LDM) offer a flexible and efficient approach for diffusion-based image synthesis across various image types. We begin by comparing our models by traditional pixel-based diffusion methods examining both training efficiency and inference speed.

Figure 2. We present samples generated by our text-to-image synthesis model, LDM-8 (KL), trained on the LAION [78] dataset. The images were produced using 200 DDIM sampling steps with η=1.0. We apply unconditional guidance [32] with a guidance scale of s=10.0 to enhance alignment with user-defined text prompts.

Notably, LDM's are trained using VQ-regularized latent spaces sometimes deliver high quality samples, even though their reconstruction performance may be slightly lower compared to models with continuous latent spaces.

For a visual comparison of how different regularization methods impact LDM training and their ability to generalize to higher resolutions (greater than 2562).

In this section, we explore how our latent diffusion models (LDMs) behave when using various downsampling factors, f ∈ {1, 2, 4, 8, 16, 32}—with LDM-1 being the standard pixel-based diffusion model. To ensure a fair comparison, all experiments are run on a single NVIDIA A100, with each model trained for the same number of steps and using the same number of parameters.

Models with intermediate downsampling levels—specifically LDM-4 to LDM-16—offer a more effective balance between training efficiency and the preservation of fine image details.Notably, there's a significant difference in performance, with LDM-8 achieving an FID score that is 38 points lower than the pixel-based LDM-1 after 2 million training steps.

Figure 3 then shows how sample quality evolves during 2 million training steps on the ImageNet dataset using class-conditional models. Our observations reveal that very low down-sampling factors (LDM-1 and LDM-2) lead to slower training progress, while excessively high factors result in a quick plateau in image quality. We believe this is because low factors force the diffusion model to handle most of the perceptual compression, whereas very high factors compress the image too much, causing information loss.

Additionally, Figure 4 presents a comparison of models trained on both CelebA-HQ and ImageNet, evaluating them based on sampling speed (measured using the DDIM sampler) and FID scores.The models with downsampling factors between 4 and 8 not only produced better FID scores but also sampled images more quickly than the pixel-based approach. For complex datasets like ImageNet, reducing the compression too much can harm quality, so a moderate compression rate is essential.

Additionally, comparisons using the DDIM sampler on both CelebA-HQ and ImageNet datasets reveal that these moderate down sampling models not only produce better image quality but also offer faster sampling speeds. This is especially important for complex datasets like ImageNet, where too much compression can compromise quality. In conclusion, our findings suggest that adopting moderate down sampling factors—specifically those used in LDM-4 and LDM-8—provides an optimal trade-off between efficiency and image fidelity

Figure 2. Samples generated by our model for user-

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

defined text prompts demonstrate its effectiveness in text-to-image synthesis.

Figure 3: Creative Outputs of Generative Models: From Neural Patterns to Whimsical Art

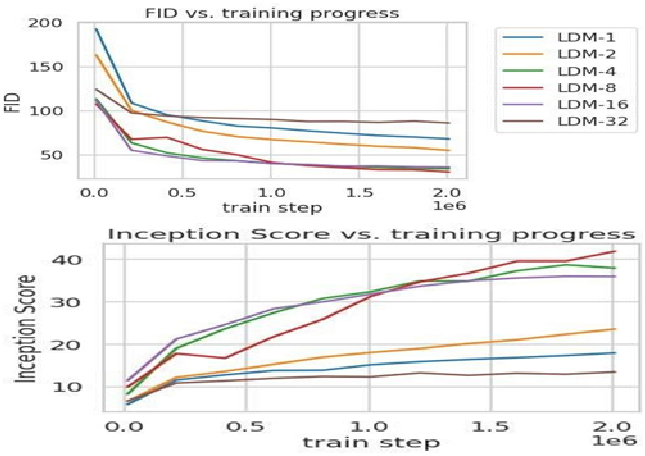

Figure 4. This analysis looks at training class-conditional Latent Diffusion Models (LDMs) using different down sampling factors f over 2 million training steps on the ImageNet dataset. The pixel-based model LDM-1 takes significantly longer to train than models with higher downsampling factors, like LDM-4, LDM-8, LDM-12, and LDM-16. However, excessive perceptual compression, as observed in LDM-32, results in a decline in overall sample quality. All models were trained under the same computational budget on a single NVIDIA A100 GPU. The results were obtained using 100 DDIM sampling steps [84] with $\kappa=0$.

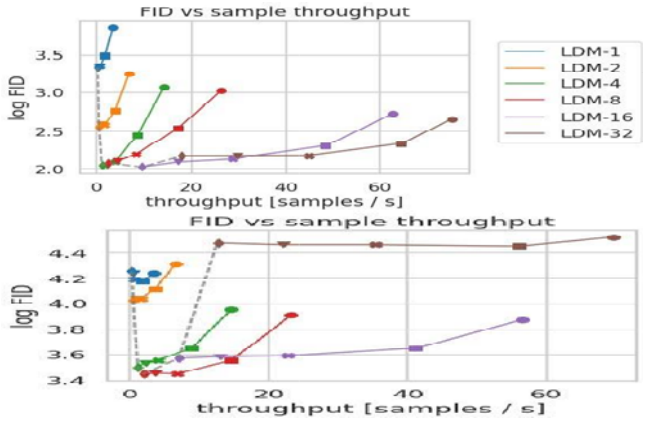

Figure 5. This comparison evaluates Latent Diffusion Models (LDMs) with varying compression levels on the CelebA-HQ (left) and ImageNet (right) datasets. Different markers represent DDIM sampling steps {10, 20, 50, 100, 200}, arranged from right to left along each performance curve. A dashed line indicates FID scores at 200 sampling steps, emphasizing the strong performance of LDM-{4–8}. FID scores were computed using 5,000 generated samples. All models were trained for 500,000 steps on CelebA-HQ and 2 million steps on ImageNet, each using a single NVIDIA A100 GPU.

E. Conditional Latent Diffusion

1. Transformer and Encoders for LDMs

In this section we explore how integrating transformer encoders with cross-attention conditioning expands the capabilities of Latent Diffusion Models (LDM) to handle various conditioning inputs that were previously unexplored. For instance, In our text to image experiments we train AKL regularized LDM with 1.5 billion parameters using language prompts from the LAION-400M dataset. The input text is first tokenized using the BERT tokenizer and then passed through the employer transformer module $\tau\theta$ to generate a latent representation. This representation is then integrated into the UNet architecture via multi-head

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

cross-attention mechanisms. This combination of specialized language and visual processing produces a robust model that generalizes well to complex user defined text prompts as illustrated. Her evolution we follow established protocols by text to image generation on the MS-COCO validation set. Our model outperforms state-of-the-art autoregressive and GAN-based techniques, and the quality of the generated samples is further improved by classifier-free diffusion guiding. In text-to-image synthesis, our guided model (LDM-KL-8-G) performs on par with state-of-the-art diffusion and autoregressive models, although having substantially fewer parameters. We train models on the Open Images dataset for semantic layout-to-image generation then refine them on COCO to illustrate the adaptability of our conditioning strategy. We additionally test our top class-conditional models on ImageNet in accordance with previous studies; the results are presented in Table 3 and Section 4. Interestingly, our models maintain a significantly lower number of parameters and computing needs while outperforming the state-of-the-art diffusion model ADM.

2. Convolutional Sampling Beyond 2562

We transform our latent fusion models into flexible instruments for image-to-image translation by feeding spatially aligned conditioning data into the input of our denoising network. This approach makes it possible to train on a variety of tasks, such as inpainting, super-resolution, and semantic synthesis. We employ landscape photos and the semantic maps that correlate to them for semantic synthesis. In this setup, the latent representation generated by our f = 4VQ-regularized model is concatenated with down-sampled semantic mappings. Even though 256^2resolution images are used for training, the model performs well at higher resolutions and can produce images at megapixel scale when sampling is done convolutionally.

Method	FID↓	IS↑	Precision↑	Recall↑	N_{params}	
BigGan-deep [3]	6.95	203.6±2.6	**0.87**	0.28	340M	-
ADM [15]	10.94	100.98	0.69	**0.63**	554M	250 DDIM steps
ADM-G [15]	4.59	186.7	0.82	0.52	608M	250 DDIM steps
LDM-4 (ours)	10.56	103.49±1.24	0.71	0.62	400M	250 DDIM steps
LDM-4-G (ours)	**3.60**	**247.67**±5.59	**0.87**	0.48	400M	250 steps, c.f.g [32], $s = 1.5$

Table1. Comparison of class-conditional ImageNet models

We further harness this capacity to extend our super resolution inpainting models, enabling the fraction of large images in the range of 521^2 to 1024^2. Here the signal-to-noise ratio--affected by the latent space scale plays a critical role in quality of the output. Refer to figure below to analyze how latent space scale affects the quality of the output.

Figure 6. In spatially conditioned tasks like semantic synthesis of landscape photos, a Latent Diffusion Model (LDM) trained at 256^2 resolution can handle higher resolutions, like 512×1024.

As explained in Section 3.3, we concatenate low-resolution inputs with the model's input data in order to immediately condition our Latent Diffusion Model (LDM) for super-resolution. In our preliminary tests, which follow the SR3 methodology, we use bicubic interpolation with a 4× down-sampling ratio to degrade images and run ImageNet data through the preprocessing pipeline of SR3. We use a pretrained autoencoder that was trained using VQ-regularization on the OpenImages dataset and has a downsampling factor of f = 4. Since the UNet inputs are immediately concatenated with the low-resolution image y, our transformation τθserves as the identity function.

Competitive performance is demonstrated by our qualitative and quantitative findings (refer to Figure 10 and Table 5). Surprisingly, SR3 surpasses our LDM-SR model in terms of the Inception, yet our model has a lower FID score.

International Journal of Intelligent Computing Systems

Volume 1, Issue 1, June 2025

Figure 7. For ImageNet 64→256 super-resolution on the ImageNet validation set, LDM-SR excels at generating realistic textures, while SR3 demonstrates an advantage in producing more coherent fine structural details. Additional examples and SR3 outputs can be found in the appendix.

Additionally, we carried out a user trial akin to SR3, in which users were shown a low-resolution image along with two comparable high-resolution outputs, one produced by a pixel-based baseline and the other by our LDM-SR model. Participants were requested to express their preference, thereby assisting in validating the impressive performance of our LDM-SR methodology.

Our qualitative and quantitative findings (refer Figure 6 and Table 2) underscore the competitive efficacy of the algorithm. Importantly, the LDM-SR model registers a lower Fréchet Inception Distance (FID) in comparison to SR3, even though SR3 achieves a marginally higher Inception Score (IS). While direct image regression models may yield the highest PSNR and SSIM values, these metrics frequently favor smoother, less intricate outputs that do not necessarily correspond to the visual quality as perceived by human viewers.

Model (*reg.*-type)	train throughput samples/sec.	sampling @256	throughput† @512	train+val hours/epoch	FID@2k epoch 6
LDM-1 (no first stage)	0.11	0.26	0.07	20.66	24.74
LDM-4 (*KL*, w/ attn)	0.32	0.97	0.34	7.66	15.21
LDM-4 (*VQ*, w/ attn)	0.33	0.97	0.34	7.04	14.99
LDM-4 (*VQ*, w/o attn)	0.35	0.99	0.36	6.66	15.95

Table 2. There are some variations from the results in Figure 4 when evaluating inpainting efficiency, mostly because of variations in GPU configurations and batch sizes. See the supplemental material for further information

F. Limitations and Societal Impact

While over latent diffusion models (LDMs) significantly cut down on computational demands compared to traditional pixel-based methods, they still have some drawbacks. One key limitation is that their step-by-step sampling process tends to be slower than that of GANs.

Additionally, although our f = 4 auto encoding models maintain high image quality, there can be challenges when task demand very fine, pixel-level precision the models reconstruction ability might not capture every minute detail perfectly. This issue is also noticeable in our super resolution models with seem somewhat constrained when it comes to preserving ultra-fine details.

Generative models for image synthesis hold significant promise by democratizing access to advanced creative tools and lowering the barriers to content generation; however, they also pose notable risks. These systems can be exploited to produce realistic yet manipulated images contribute to the spread of misinformation and deep fakes – a concern that disproportionately impacts vulnerable groups. Finally like many deep learning systems, these models can inadvertently replicate or even amplify existing biases found in the data. Additionally, there is a potential for such models to reveal sensitive information from the training data, present in the input data rising ethical and privacy issues.

IV.CONCLUSION AND FUTURE WORK

In this study, we present Instant Imager, a novel latent fusion framework that transforms text-to-image synthesis by utilizing in-context learning.

Our approach significantly speeds up the image generation process producing high-quality, customized images up to 10 times faster than traditional optimized methods while reducing the computational burden.

Our model effectively adapts to a variety of tasks, including super-resolution, inpainting, and semantic synthesis, and produces visually appealing images by combining Stable Diffusion with transformer-based conditioning mechanisms and well-designed encoder-decoder architectures..Extensive experiments on datasets like ImageNet and LAION-400 M demonstrate that instant image achieves competitive performance balancing efficiency with high- fidelity results. Furthermore, we have incorporated user driven range bar

segment that allows users to select specific images according to their preferences adding an extra level of control and personalization to the synthesis process. It allows users to adjust parameters or choose a specific range of outputs the range bar adds significant value offering enhanced control and personalization. This makes our system not only more accessible but also more responsive to the diverse needs of creative professionals.

Future studies will try to improve the model's ability to capture fine-grained information and speed up sampling even more, which will increase the model's potential for a wide range of imaginative and useful applications.

V. REFERENCES

[1] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, William W. Cohen, " Subject-driven Text-to- Image Generation via Apprenticeship Learning,"

[2] Florian Bordes, Randall Balestriero, and Pascal Vincent, “High Fidelity Visualization of What Your Self-Supervised Representation Knows About.” arXiv:2112.09164, 2021.

[3] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, Tong Sun, " Towards Language-Free Training for Text-to-Image Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 17907-17917

[4] Youwei Liang, Junfeng He, Gang Li, Peizhao Li, Arseniy Klimovskiy, Nicholas Carolan, Jiao Sun, Jordi Pont-Tuset, Sarah Young, Feng Yang, Junjie Ke, Krishnamurthy DjDvijotham, Katherine M. Collins, Yiwen Luo, Yang Li, Kai J Kohlhoff, Deepak Ramachandran, Vidhya Navalpakkam, “Rich Human Feedback for Text-to-Image Generation.,” Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19401-19411

[5] Yuval Alaluf, Daniel Garibi, Or Patashnik, Hadar Averbuch-Elor, and Daniel Cohen-Or. 2023. “Cross-Image Attention for Zero-Shot Appearance Transfer,“ arXiv:2311.03335 [cs.CV]

[6] Jie Qin, Jie Wu, Weifeng Chen, Yuxi Ren, Huixia Li, Hefeng Wu, Xuefeng Xiao, Rui Wang, Shilei Wen, “DiffusionGPT: LLM-Driven Text-to-Image Generation System”, arXiv:2401.10061 [cs.CV]

[7] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired imageto-image translation using cycle-consistent adversarial networks.” The IEEE International Conference on Computer Vision (ICCV), 2017.

[8] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. “Generative image inpainting with contextual attention.” In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[9] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, Jing Shao, ”Semantics Disentangling for Text-To-Image Generation”; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 2327-2336.

[10] Songwei Ge, Taesung Park, Jun-Yan Zhu, Jia-Bin Huang, "Expressive Text-to-Image Generation with Rich Text," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),2023,pp.7545-7556.

[11] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, Xiaodong He, "AttnGAN: Fine-Grained Text to Image Generation With Attentional Generative Adversarial Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 1316-1324.

[12] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, Yinfei Yang, "Cross-Modal Contrastive Learning for Text-to-Image Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 833-842.

[13] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, Yong Jae Lee, “GLIGEN: Open-Set Grounded Text-to-Image Generation”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22511-22521

[14] Wentong Liao, Kai Hu, Michael Ying Yang, Bodo Rosenhahn, “Text to Image Generation With Semantic-Spatial Aware GAN“, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 18187-18196

[15] Yuming Jiang, Shuai Yang, Haonan Qiu, Wayne Wu, Chen Change Loy, Ziwei Liu,"Text2Human: text-driven controllable human image generation", https://dl.acm.org/doi/10.1145/3528223.3530104, Article No.: 162.

[16] Yaru Hao, Zewen Chi, Li Dong, Furu Wei, "Optimizing Prompts for Text-to-Image Generation," Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track.

[17] Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: "A large-scale database for aesthetic visual analysis". In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2408–2415, 2012. doi: 10.1109/CVPR.2012.6247954.

[18] KonpatPreechakul, NattanatChatthee, SuttisakWizadwongsa, and SupasornSuwajanakorn. "Diffusion Autoencoders: Toward a Meaningful and Decodable Representation". arXiv:2111.15640, 2021.

[19] Yang Song and Stefano Ermon. "Improved Techniques for Training Score-Based Generative Models." arXiv:2006.09011, 2020.

[20] Tingting Qiao, Jing Zhang, Duanqing Xu, Dacheng Tao, "MirrorGAN: Learning Text-To-Image Generation by Redescription," Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 1505-1514.

[21] Weihao Xia, Yujiu Yang, Jing-Hao Xue, Baoyuan Wu, "Text-Guided Diverse Face Image Generation and Manipulation,"Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2256-2265.

[22] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, Omer Levy, "An Open Dataset of User Preferences for Text- to-Image Generation," Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track.

[23] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, Kfir Aberman; "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 22500-22510.

[24] Jaemin Cho, Abhay Zala, Mohit Bansal, "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generation Models," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3043-3054.

[25] Xingchao Liu, Chengyue Gong, Lemeng Wu, Shujian Zhang, Hao Su, Qiang Liu, "Training-Free Text-to-Image Generation with Improved CLIP+GAN Space Optimization", arXiv:2112.01573 [cs.CV].

[26] Arash Vahdat and Jan Kautz. NVAE: A deep hierarchical variational autoencoder. In NeurIPS, 2020.

[27] Younjung Hwang, Yi Wu, "Graphic Design Education in the Era of Text-to-Image Generation", https://doi.org/10.1111/jade.12558.

[28] Abukmeil, M., Ferrari, S., Genovese, A., Piuri, V. & Scotti, F. (2022), "A survey of unsupervised generative models for exploratory data analysis and representation learning", ACM Computing Surveys, Vol. 54, No. 5, pp. 1–40.

[29] Matthews, B., Shannon, B. & Roxburgh, M. (2023) Destroy all humans: the dematerialisation of the designer in an age of automation and its impact on graphic design—a literature review, International Journal of Art & Design Education, Vol. 42, No. 3, pp. 367–83.

[30] Yijun Yang, Ruiyuan Gao, Xiao Yang, Jianyuan Zhong, Qiang Xu, "Defending Text-to-Image Models from Adversarial Prompts," Part of Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Main Conference Track.

[31] Weilun Wang; Jianmin Bao; Wengang Zhou; Dongdong Chen; Dong Chen; Lu Yuan, "Learning a Diffusion Model from a Single Natural Image", IEEE Explore.

[32] Ivona Najdenkoska, Animesh Sinha, Abhimanyu Dubey, Dhruv Mahajan, Vignesh Ramanathan & Filip Radenovic, "Context Diffusion: In-Context Aware Image Generation", pp 375–391.

[33] Quynh Phung, Songwei Ge, Jia-Bin Huang, "Grounded Text-to-Image Synthesis with Attention Refocusing", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 7932-7942.

[34] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa,

Michael Rubinstein, Kfir Aberman, "HyperDreamBooth: HyperNetworks for Fast Personalization of Text- to-Image Models," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6527- 6536.

[35] Shoufa Chen, Mengmeng Xu, Jiawei Ren, Yuren Cong, Sen He, Yanping Xie, Animesh Sinha, Ping Luo, Tao Xiang, Juan-Manuel Perez-Rua, "GenTron: Diffusion Transformers for Image and Video Generation", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 6441-6451.

[36] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. CoRR, abs/2102.12092, 2021.

[37] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. CoRR, abs/2104.02600, 2021