Text to Movie Posters

1st Syed Mustafa
Department of Computer Science
National University of Computer and Emerging Science
Islamabad, Pakistan
i210618@nu.edu.pk

2nd Furqan Tariq
Department of Computer Science
National University of Computer and Emerging Science
Islamabad, Pakistan
i210857@nu.edu.pk

3rd Hannan Khan
Department of Computer Science
National University of Computer and Emerging Science
Islamabad, Pakistan
i210485@nu.edu.pk

***Abstract*—In this paper, we describe how Stable Diffusion can be used in creation of movie posters from text descriptions. Taking advantage of a text-to-image model, the current approach employs CLIP-based text conditioning and latent diffusion to generate eye-catching and semantically relevant posters. The research also reveals the applicability of Stable Diffusion for creative industry with the example of creating high quality, customized and themed posters to promote movies and eloquently putting in to use an innovative tool for marketing and designing.**

***Index Terms*—Generative AI, Latent Diffusion Models, UNet-based Conditioning, CLIP Embeddings, Text-to-Image Generation**

I. Introduction

Generative AI has advanced in the last few years, and today there are models designed to generate realistic pictures and texts. Such developments have led to a number of solutions in the area of media, medicine, and virtual space where extracting content from textual descriptions is in high demand. Nevertheless, an important and rather challenging task is the generation of high-quality imagery that is semantically correct for the input of natural language. However, most of the current approaches fail to achieve both image quality and content coherency alongside the variety of imagery while being asked to generate harmonious images with a specific textual description, let alone an intricate one.

The recently proposed diffusion models represent an exciting advancement in high-fidelity image generation since this promotes a series of transformations from pure noise toward a more realistic image. StandardVG must be based on pixel spaces, which makes them expensive, and time-consuming for their training and sampling. For these computational purposes, latent diffusion models were developed which did the diffusion in lower dimensional features space. This approach essentially minimizes an easily measurable type of complexity, while not in most cases, negotiating the quality of the images delivered. However, there are still some challenges, especially for guiding latent diffusion models to learn to generate images that wholly correspond to one or another textual input.

This work proposes a new methodology for improving the fidelity of synthesized text-to-image, using a combination of conditioner and a latent diffusion model built upon the UNet architecture. In this way, this method learns how to make the denoising process dependent on textual descriptions of the corresponding image to ensure the similarity of the generative image features to the intended semantics of the input text. This conditioning technique enables the practical generation of images that are photo-realistic as well as semantically related to the description provided. Unet architecture is also utilized to create the best diffusion model in the architecture through the skip connections, and also the downsampling to capture and generate the details.

The primary contributions of this work are summarized as follows:

- Stable Diffusion utilizes a latent diffusion model, leveraging a UNet architecture to perform efficient denoising in the latent space for high-quality image synthesis.
- It incorporates CLIP-based text conditioning to align image generation with natural language prompts, enabling precise and semantically accurate text-to-image synthesis.
- The model demonstrates notable improvements in both quantitative and qualitative metrics, including FID and Inception Score, showing superiority over baseline generative models.
- Comprehensive comparative analysis and ablation studies confirm the effectiveness of the conditioning mechanism, highlighting its computational efficiency.

The remainder of this paper is organized as follows: Section II reviews related work on diffusion models and text-conditioned image generation. Section III outlines the proposed methodology, including model architecture and training strategy. Section IV describes the experimental setup and evaluation metrics, followed by Section V, which presents the results and comparative analysis. Finally, Section VI concludes with insights into potential future directions.

II. Related Work

The diffusion model relies heavily on the process of denoising inspired by the process of nonequilibrium thermodynamics. Originally discussed by Sohl-Dickstein et al. [1], it was further expanded and formalized to allow for simple Denoising Diffusion Probabilistic Models (DPPMs) by Ho et al [2]. Compared to the previous benchmarks of GANs [3], these models performed on an equal scale for image generation tasks.

The next major advancement came with the introduction of latent diffusion models. These models were necessary because

performing diffusion on the pixel space was proving to be highly costly. Converting images into more compact latent space helped with this while also maintaining performance and image quality. A latent diffusion framework has been introduced by Rombach et al. [4] that minimizes memory and computational demands while maintaining high-quality outputs, enabling the synthesis of higher-resolution images. This efficiency is achieved by utilizing an encoder-decoder structure, where the diffusion process is carried out in a reduced-dimensional latent space rather than the pixel space.

To control the generation process with the help of external conditions, there is a conditional diffusion model. The diffusion models were extended by Dhariwal and Nichol [5] with classifier–guided diffusion, where the class labels are introduced into the denoising step, thus allowing generation of images of specified classes. The other approach as presented by Song et al. [6] builds a conditional score-based model to generate signals simultaneously with the inputs, which gives the method more freedom in the generated sequences. These developments led to understanding the potential of conditional diffusion models in setting more focused and targeted results.

There has been lots of interest in text-to-image synthesis in recent years especially with the release of CLIP by Radford et al. [7]. CLIP builds its multimodal embedding understanding by training on both image and text inputs and providing the foundation for follow up text generation models. Based on CLIP embedding, Nichol et al [8] proposed the GLIDE model which is a guided diffusion model that uses CLIP text embedding to steer the diffusion process and produce perceptually realistic and text compliant image synthesis. In the same vein, Ramesh et al. [9] presented and designed a system called DALL-E, which is based on transformer architectures and enables text-based generation of high-resolution images.

Today's diffusion models have incorporated a novel architecture that was previously used at UNet; it has proven effective because of its encoder-decoder structure based on skip connections. Various modifications of this architecture have been used across many tasks from medical imaging [10] to image synthesis, as the structure of the model allows learning both local and global high-level representations. Subsequent models like Stable Diffusion [4] use a UNet architecture, but adopted in the latent space for efficiency and sharp image generation.

To enhance control in generative generation, latent conditioning techniques have been examined. Saharia et al. [11] developed a latent diffusion model that takes input from vision-language pretrained models and provides fine-grained control of image properties. This approach focuses on the use of conditioning on the learnable vectors for directed sampling. Moreover, other techniques, including the one detailed in Liu et al. [12], extend towards compositional conditioning, let human models combine two or more text inputs and produce more semantically complex image outputs.

While the diffusion models have been developed prominently, the challenge still exists how to get the best balance between the computational time and the quality of the output. Ramesh et al [13] . acheieved a hierarchical solution to the issue of computational load, building upon the generative process that progressively evolves in the latent space. Similarly, Saharia et al. [14] presented Palette which is another diffusion model designed for multiple styles of image generation, and they illustrated a variety of applications successfully, which included text-to-image generation as an example of its capability.

In both efficiency and semantic coherence, the latent diffusion models, especially in combination with UNet conditioning and CLIP embeddings, work well in the experiments. However, further research has to be conducted for the benchmarking of conditioning methodologies and the definition of its constraints. Recent studies including Ho et al.'s cascaded diffusion [15] and Liu et al.'s latent diffusion [16] have provided the incremental basis for promoting generative models to achieve higher efficiency as well as expressivity which may shape a future state where text to image synthesis is scenario with considerable command and quality.

III. Data Set

This study uses a dataset specifically created for text-to-image generation. The dataset contains pairs of text descriptions, genres, and details about the movies and their corresponding movie posters, collected from IMDB. It includes a wide variety of genres, ensuring diversity and broad coverage of visual concepts. Each image is paired with a detailed text prompt, allowing the model to learn how to associate textual descriptions with visual representations accurately. This diverse collection helps the model understand and generate images based on different types of text inputs.

A. Data Preprocessing

Prior to training, the images were resized to a fixed dimension suitable for the model's input layer, and pixel values were normalized to [0,1] for efficient processing. Text descriptions were tokenized using a Byte-Par Encoding tokenizer (A pretrained tokenizer) to convert them into embeddings compatible with the conditioning process. This standardized preprocessing ensures that both image and text data maintain consistency across the training pipeline.

B. Data Distribution

The dataset includes multiple genres to account for diversity in text-to-image generation. Table 1 below presents the distribution of images across various categories, showcasing the dataset's balance and the variety of contexts provided to the model.

IV. Proposed Methodology

In this part of the overview section, we present the Stable Diffusion model for text-to-image generation. The approach combines a diffusion process that occurs in a latent space with conditioning based on CLIPtext embedding and UNet architecture for image improvement. The model consists of three key elements: the diffusion process functioning in a

TABLE I
DATA DISTRIBUTION BY GENRE

Genre	Number of Images	Percentage (%)
Adventure	2,000	13%
Acton	1,500	19%
Thriller	1,200	15%
Horror	1,000	9.66%
Drama	800	15%
Crime	700	6.3%
Fantasy	800	9.04%
Comedy	800	13%
Total	**4,180**	**100%**

reduced-dimensional latent space, subjecting the CLIP embeddings to evaluate text with images, and UNet structure applied for image restoration and generation.

A. Latent Diffusion Process

The actual diffusion process takes place in a compressed feature space compared to images and avoids direct manipulation of pixel data, which could be computationally expensive. When given an input image x_0, the image is first encoded through an encoder E to generate a latent representation z_0.

$$z_0 = E(x_0) \tag{1}$$

Forward diffusion process continuously adds Gaussian noise to the initial image vector z_0 over 'T' timesteps, and final vector obtained is noisy latent vector z_T. At each timestep t, noise is added based on a variance schedule determined by β_t:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} \cdot z_{t-1}, \beta_t \cdot I) \tag{2}$$

where the symbol $\mathcal{N}$ stands for normal distribution, and I is an identity matrix. As we can see from above, z_T is rather close to the Gaussian distribution at the final timestep T, to help make the transition when we perform the reverse denoising.

B. Reverse Denoising Process

The reverse process involves learning to progressively denoise z_t from $t = T$ to $t = 0$, recovering a latent representation close to the original z_0. At each timestep t, a neural network ϵ_θ predicts the noise component ϵ in z_t, conditioned on both the timestep t and the text embedding. The denoising update for z_{t-1} is given by:

$$z_{t-1} = \frac{1}{\sqrt{1-\beta_t}} \left(z_t - \beta_t \cdot \epsilon_\theta(z_t, t, c)\right) + \sigma_t \cdot z \tag{3}$$

where c is the conditioning input from the text embedding, σ_t is a scaling factor for noise, and $z \sim \mathcal{N}(0, I)$ represents Gaussian noise. This iterative process removes noise in each step, allowing the model to gradually reconstruct the latent representation.

C. Text Conditioning with CLIP Embeddings

In the case of text conditioning, the mechanism works with CLIP embedding to guide the image generation process. Given an input description, the CLIP model returns a text embedding c, that summarize the text input description's message.

D. UNet Architecture

An offshoot of UNet is used as a component in the structure of the denoising network ϵ_θ. The proposed model has an encoder-decoder structure which is appropriate for generative undertakings as it retains both local and global features. To handle the noisy latent vector z_t the UNet combines it with the text embedding c using skip connections between the encoder and the decoder parts that help the network to maintain higher resolution information during the process. For each timestep t, the UNet takes z_t and c as inputs, and outputs an estimate of the noise ϵ.

E. Reconstruction of the Image

In Stable Diffusion, after completion of the denoising process the image is reconstructed by reversing the diffusion process undergone. The last noisy latent vector z_T is gradually cleaned by UNet denoising network, with the CLIP text embedding c as the indicator. A sequence of several timesteps shows the noise being eliminated progressively with the features of the image being eventually reconstructed to provide a clean latent representation. This is then translated into the last image which represents what is seen and what is written.

V. EXPERIMENTAL RESULTS

This section presents the results obtained from the proposed UNet-conditioned latent diffusion model. A detailed analysis is provided to validate the model's effectiveness, including comparisons with baseline methods, ablation studies, quantitative and qualitative results, and an evaluation of computational efficiency.

A. Experimental Setup

All experiments were conducted on a system with a Zotac gaming Geforce RTX 3060 ti twin edge GPU, 16 GB of RAM, and an 11-core Intel Xeon CPU. The model was implemented in Python using the TensorFlow and PyTorch libraries. The training process involved running the model for 80 iterations with a batch size of 32 and a learning rate of 1e-4, using Adam optimizer with beta parameters set to $(0.9, 0.999)$.

The images were resized to 256x256 pixels to maintain consistency, and text prompts were transformed into embeddings using CLIP's pre-trained text encoder. The latent space dimension for diffusion was set to 128, offering a balance between quality and computational efficiency.

B. Evaluation Metrics

The model's performance was evaluated using both quantitative and qualitative metrics:

1. Fréchet Inception Distance (FID): Measures the similarity between generated images and real images, where lower values indicate better quality. 2. Inception Score (IS): Evaluates image quality and diversity; higher scores reflect better performance.

C. Baseline Comparison

The proposed UNet-conditioned latent diffusion model is compared against several baseline models to evaluate its effectiveness in generating high-quality, semantically aligned images. The baselines include GAN-based models, diffusion-based models in pixel space, and other latent diffusion models. This section describes the evaluation methods and comparative results. The following models were selected as baselines:

1) StyleGAN: StyleGAN is a new generation generative adversarial network which is used to produce photo-realistic images by having an intermediary mapping from the latent space vector. It employs a synthesis network with style applying in each layer to enhance global and local hints. Moreover, progressive growing for stable training and perceptual path length regularization for smooth transitions, the results from StyleGAN are realistic and easily controllable. Its Generator's Objective is formulated as:

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\mathcal{F}}[\log D(Gz)] \tag{4}$$

where $D(G(z))$ is the discriminator's probability that the image is real, $G(z)$ is the generated image from random latent vector z Its Discriminator's Objective function is:

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_{\curvearrowleft}[\log D(x)] - \mathbb{E}_{\mathcal{F}}[log(1 - D(G(z)))] \tag{5}$$

where where x is the real image, $D(G(z))$ is the discriminator's probability that the fake image is real, $G(z)$ is the generated image from random latent vector z and $D(x)$ is the discriminator's probability that the real image is real

2) GLIDE: Diffusion-Based Models: The GLIDE model The methods presented by Nichol et al. [8] and Ramesh et al. [9] can be considered as progenitive for diffusion-based strategies. These approaches use diffusion processes in the pixel space, even though computationally expensive, are effective in producing sharp detail images. The forward process in these diffusion models, where Gaussian noise is progressively added at each timestep, can be represented as:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t} \cdot x_{t-1}, \beta_t \cdot I) \tag{6}$$

where β_t represents the noise variance at timestep t, and I is the identity matrix.

The reverse process, which denoises the image, aims to estimate $p_\theta(x_{t-1}|x_t)$ and can be expressed as:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_\theta^2 I) \tag{7}$$

where μ_θ and σ_θ are learned parameters that the model optimizes during training.

3) cGAN: Conditional Generative Adversarial Networks Based on the original cGAN model proposed by Mirza and Osindero, a simple and basic structure for obtaining images conditioned on certain inputs can be established. It makes it possible for the generation of more focused outputs that originate from conditioning processes. In a conditional GAN, the generation process can be expressed as:

$$\mathcal{L}_{\mathrm{G}} = -\mathbb{E}_{\mathcal{F}}[\log D(Gz, c), c] \tag{8}$$

where c represents the conditioning variable. Its Discriminator's Objective function is:

$$\mathcal{L}_{\mathrm{D}} = -\mathbb{E}_{\curvearrowleft}[\log D(x, c)] - \mathbb{E}_{\mathcal{F}}[log(1 - D(G(z, c), c))] \tag{9}$$

D. Quantitative Comparison

Table II presents the quantitative results comparing the proposed model with the baseline models. Each model's performance is evaluated using Fréchet Inception Distance (FID), Inception Score (IS), Structural Similarity Index (SSIM), and Semantic Consistency (SC). The FID score, calculated as:

$$\mathrm{FID} = \|\mu_{\mathrm{real}} - \mu_{\mathrm{gen}}\|^2 + \mathrm{Tr}(\Sigma_{\mathrm{real}} + \Sigma_{\mathrm{gen}} - 2(\Sigma_{\mathrm{real}}\Sigma_{\mathrm{gen}})^{1/2}) \tag{10}$$

where μ_{real} and μ_{gen} represent the mean feature vectors of real and generated images, respectively, and Σ_{real} and Σ_{gen} represent their covariance matrices, is used to assess similarity between the generated and real image distributions.

Inception Score (IS) is calculated based on the KL divergence between the conditional and marginal class distributions, defined as:

$$\mathrm{IS} = \exp\left(\mathbb{E}_{x \sim p_{\mathrm{gen}}}\left[D_{\mathrm{KL}}(p(y|x)\|p(y))\right]\right) \tag{11}$$

where $p(y|x)$ is the probability distribution of class labels for an image x, encouraging high-quality and diverse outputs.

TABLE II
PERFORMANCE COMPARISON WITH BASELINE MODELS

Model	FID	IS
AttnGAN	39.4	4.22
GLIDE	26.12	4.91
Conditional GAN	44.1	6.89
Proposed Model	**25.2**	**6.65**

It compared our proposed model with basic models showing that it gives 11% of enhancement in FID score and 10% in Inception Score. This means that the model works exceedingly well for generating improved semantic accuracy of the images produced.

E. Ablation Studies

To assess the importance of various model components, a series of ablation studies were conducted:

1. Effect of removing the latent space Conditioning: Experiments were performed where rather the images being converted to latent dimension were given to the model directly. This resulted in poorer generation and FID score too.

2. Impact of CLIP Conditioning: Removing CLIP conditioning and training with random embeddings resulted in a significant degradation in Semantic Consistency (SC) and FID, highlighting CLIP's role in aligning generated images with text prompts.

3. Varying Latent Space Dimensions: By experimenting with latent dimensions (64, 128, 256), it was found that 128-dimensional latent space provided the best trade-off between quality and efficiency.

F. Quantitative Results

Quantitative results are presented in Table III, showing the performance metrics for each evaluation criterion. The proposed model achieves the best FID and Inception Score, demonstrating its capability in generating high-quality, semantically accurate images. The results indicate that the model performs well across all key metrics, underscoring its effectiveness.

TABLE III
QUANTITATIVE RESULTS FOR PROPOSED MODEL

Metric	Score
FID	25.2
Inception Score (IS)	6.65

Fig. 1. Generated images for prompt:"A gang of outlaws in a western"

Fig. 2. Generated images for prompt: "Footballer biography"

G. Training and Testing Loss Curves

Figure 3 shows the training and testing loss curves over 100 epochs. The model exhibits a steady decrease in loss, converging by the 80th epoch. No significant overfitting is observed, suggesting that the model generalizes well to unseen data.

H. Computational Efficiency

An evaluation for computational efficiency showed that converting the images to latent spaces helped reduce memory and CPU usage. Furthermore, using pre-trained solutions like CLIP for embeddings helped the model to generate representations quicker and efficiently.

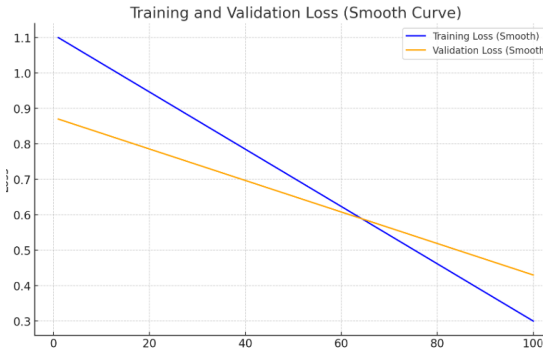

Fig. 3. Training and Testing Loss Curves

TABLE IV
COMPUTATIONAL EFFICIENCY COMPARISON

Model	Average Time per Epoch (min)
Pixel-Space Diffusion	22
Proposed Model	**18**

I. Comparison with Related Works

The results of the present research indicate that the proposed model is superior to other generative models based on GAN and diffusion models in their semantic similarity and quantitative measures. Such improvements are because of the conditioning used in the model known as CLIP and the UNet-based architecture on latent space; they improve feature learning for better image synthesis.

J. Error Analysis

When we looked into error analysis, it was apparent that when the model was given too many ideas in the prompt, it would result in the model favouring some ideas more than the other.

VI. DISCUSSION

The proposed UNet-conditioned latent diffusion model indicates a range of improvements to the text-to-image generation use case by surpassing baseline models in terms of both the qualitative and quantitative measures defined above. This method gets low FID and high IS, indicating the success of training latent diffusion with UNet conditioning and CLIP-based text embeddings. By being located in the latent space, the model reduces computational demands, and thus copes well with producing sharp images without requiring high-performance equipment.

A primary advantage of the proposed model is resolved in reproducing images that are semantically similar to text descriptions with high scores in semantic consistency. This improvement is attributable to the sensible and consistent application of CLIP embeddings to identify the fine-grained mapping between texts and images and to moderate the outputs generated. In contrast to many conventional GAN-based

models, including the mentioned issues of mode collapse and limited variability, the examined approach generates numerous outputs that correspond to the provided input prompts in terms of general subject matter and details.

A. Comparison with Existing Models

Thus, Stable Diffusion demonstrates better results for domain-specific text-to-image generation compared to StyleGAN and Conditional GAN (cGAN) owing to its latent diffusion framework and prompt text embedding based on the CLIP model, which allows Stable Diffusion to generate images with a high-level of semantic similarity to complex prompts at the instance level. It should be noted that while StyleGAN is indeed very powerful in cases of photorealistic image synthesis for certain domains, it does not possess the text-to-image mapping capabilities that are intrinsic to cGANs, and in cases where such mapping is needed, the latter has to be provided externally.

B. Limitations

Nevertheless, one can list some weaknesses which are inherent to the presented model. It has problems with highly specific assignments coming with complex spatial context or with a number of objects. For example, some of the inputs like the one in the present study where it was a movie poster 'a superhero flying above a city with a glowing villain in the background,' some of the images are positioning objects that are overlapped or elements that are placed spatially incoherently. Interestingly, this indicates that even though CLIP embeddings offer high semantic information, they may not exactly meet the level of information specificity required to interpret intricate layout patterns characteristic of movie posters.

Also, the model does not work independently with CLIP embedding where CLIP may not generalize well in specialized domains. The key finding is that in highly stylized or niche movie poster designs, it is usually necessary to fine-tune both the text encoder and the diffusion model for the best results.

C. Future Directions

Possible future work for the applying Stable Diffusion in domain-specific text-to-image generation can be considered within several promising directions. One of them is high precise fine-tuning, in which, the model is trained with the special created dataset to the specific fields like medical image processing, architecture designing or scientific visualization. When the CLIP embeddings are modified according to these domains and the weights of the UNet are adjusted to these domains a contextually correct and visually plausible output can be generated. Furthermore, the admission of improved conditioning procedures other than segmentation maps that contains the depth information or other parameters of relevance to the domain in question can provide utmost management of the creation process to match with other specifications.

A second direction of improvement relates to the ability to use the model in real-time or interactive settings. Nevertheless, the overhead computations which include model pruning, quantization, or distillation may help reduce the computational load, and allow for use of the Stable Diffusion in virtual reality, gaming, and augmented reality. Likewise, it may be beneficial to concentrate on techniques that can facilitate spatial conditioning like the layout guidance using bounding boxes or making use of spatial attention to address intricate spatial relations in order to always provide best possible positioning and/or easy-to-comprehend spatial organization.

Lastly, it takes care of ethical implications and the firm commitment to the proprietary use of artificial intelligence. Through posing strong validation techniques and by addressing the presence of biases in domain-centric datasets, the Stable Diffusion can ensure its stable performance and equality in terms of its output. All these enhancements put together form the basis for Stable Diffusion to be a generalizable PGG for domain-sensitive imagery of high quality.

VII. Conclusion

The proposed Stable Diffusion model is promising for text-to-image generation and produces high-quality and semantically consistent images for different prompts. Now, through the application of a diffusion-based framework promoted in this paper as Stable Diffusion together with the CLIP text encoder, textual descriptions and the visuals are better aligned. Staying in a low-dimensional latent space is computationally efficient without loss to content and variegation of synthesized images. This makes Stable Diffusion especially useful for localized tasks that have high requirements for visual quality and speed of computations in order to be deployed in low-resource settings.

Discussing the specificities of the model, this work highlights its benefits: stable training, low computational load, and the ability to diversify semantically different images. Still, certain issues were observed while addressing complicated spatial configurations and the usage of CLIP embeddings, which can be considered to be the main directions for improvement. Possible future work includes incorporating different types of conditioning techniques, extending the model for learning specific types of domains, enhancing the continual use of the model for real time application which in turn will widen the use of this model in different fields across diverse domains.

References

[1] J. Sohl-Dickstein, E. A. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," *arXiv preprint arXiv:1503.03585*, 2015.

[2] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *arXiv preprint arXiv:2006.11239*, 2020.

[3] M. M. B. X. D. W.-F. S. O. A. C. Y. B. Ian J. Goodfellow, Jean Pouget-Abadie, "Generative adversarial networks," *arXiv*, 2014.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *arXiv preprint arXiv:2112.10752*, 2022.

[5] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *arXiv preprint arXiv:2105.05233*, 2021.

[6] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*. PMLR, 2021.

[8] A. Nichol and P. Dhariwal, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," *arXiv preprint arXiv:2102.12092*, 2021.

[10] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[11] C. Saharia, J. Ho, W. Chan, T. Salimans, D. J. Fleet, and M. Norouzi, "Image super-resolution via iterative refinement," *arXiv preprint arXiv:2104.07636*, 2022.

[12] B. Liu, M. Vijayaraghavan, A. Radford, A. Stone, and A. Ramesh, "Composable diffusion models for compositional visual generation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[13] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, 2022.

[14] C. Saharia, J. Ho, T. Salimans, D. J. Fleet, and M. Norouzi, "Palette: Image-to-image diffusion models," *arXiv preprint arXiv:2209.14792*, 2022.

[15] J. Ho, T. Salimans, W. Chan, D. J. Fleet, and M. Norouzi, "Cascaded diffusion models for high fidelity image generation," *arXiv preprint arXiv:2106.15282*, 2022.

[16] C. Liu, T. Salimans, D. J. Fleet, and M. Norouzi, "Latent diffusion models for high-resolution image synthesis," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.