AI Art Assistant
Workflow-Aligned Diffusion

Grant Waldow Alexandra Kissel Tommy Jiang
Arshad Kazi
Department of Computer Science
University of Wisconsin - Madison
gwaldow@wisc.edu, akissel@wisc.edu, cjiang239@wisc.edu, aakazi@wisc.edu

April 21, 2025

1 Introduction

In recent years, the quantity and quality of AI-generated artwork on the internet has exploded to a massive scale as a result of the efficacy of large-scale latent diffusion models. This emerging technology has disrupted job markets, contributed to the spread of political misinformation, and raised ethical questions about training machine learning models on the intellectual property of artists.

This wave of adoption and disruption has occurred primarily through the popularization of text-to-image LDMs such as Stable Diffusion, Midjourney, and DALL-E. These models promise artistic expression for the layman, assistance for artists, and record cost-cutting measures for marketing firms. However, these models do not live up to the first two promises. Text-to-image suffers from an inherent latent ambiguity of text-image pairs. There are a near-infinite number of reasonable image representations for any given prompt- no matter how specific. Therefore, the novelty of these text-to-image LDMs wears off quickly as users realize they cannot put their exact creative vision to the page- even with hundreds of generation attempts.

ControlNet is a model architecture that attaches to a pre-trained text-to-image LDM, and enables additional control images (such as sketches) to be utilized in the diffusion process. Thus, enabling artists and other users to quickly upscale their completed sketches. However, this process requires a completed sketch- as the model learns simply to match canny or HED images with their corresponding image-caption pair. This may be useful in some contexts, but it does not assist whatsoever in the creation of this conditioning sketch, nor does it integrate in the workflow of the user's line art. Instead, the conditioning sketch is expected to supersede the condition sketch- not enhance it. ControlNet is a step in the right direction for diffusion, but it has a critical flaw in its training methodology that we will ameliorate.

In this paper, we introduce a training data augmentation procedure and conditioning hyperparameter that alter the training of a pre-trained line art-to-image Controlnet so as to enable incremental sketch prediction. After training, our model can utilize partial conditioning sketches to predict what the full completed artwork may look like. This can be used as a guide tool for both experienced and inexperienced artists alike to accelerate the sketching and learning process. We create a canvas application tool that enables us to test this latent diffusion-integrated workflow.

2 Related Work

In this section, we will review three key areas relevant to our project: generative models for sketch completion, diffusion models, and ControlNet. These topics provide the foundation for understanding and situating our approach of generating a full painting from a partial sketch and text prompt using diffusion models enhanced by ControlNet.

2.1 Sketch Completion

Generative models have been instrumental in addressing sketch completion tasks, where the goal is to fill in missing or incomplete regions of a sketch. Early approaches primarily treated sketches as raster images and applied image-based models such as Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs). VAEs[20, 30, 21, 13, 38, 34], firstly introduced in [20], were among the first probabilistic generative models capable of encoding input data into a latent space and generating diverse outputs. Despite their stability and probabilistic rigor, VAEs often produced blurry results, leading to the rise of GANs[9, 28, 24, 17, 10], which provided sharper and more realistic outputs. Conditional GAN framework, such as pix2pix [17], excelled at translating input sketches, edges, or semantic label maps into realistic or stylized outputs by combining adversarial loss with pixel-level loss, hence was firstly utilized for sketch completion. Other frameworks such as CycleGAN [40] further advanced the field by enabling image translation where there are no exact supervisions on the output node, demonstrating flexibility in handling data sets where direct pairing of inputs and outputs is not possible. More recent works, such as GauGAN [27], introduced semantic map-based generation to refine contextual relationships in image synthesis, providing improved fidelity and control for sketch completion tasks. These image-based models demonstrated the potential of generative frameworks to create coherent and visually appealing sketches, but often faced challenges in handling diverse and complex inputs.

Beyond treating sketches as images, researchers explored alternative representations, such as vector graphics, which capture the sequential nature of drawing. Vector-based datasets, such as the Quick, Draw! Dataset [11] and the TU-Berlin Sketch Dataset [5], have been widely used to facilitate the training of models like SketchRNN. These datasets provide millions of labeled sketches and their corresponding vector representations, making them valuable resources for developing and benchmarking vector-based sketch completion models. Models like SketchRNN [11] introduced a recurrent approach by representing sketches as sequences of vectorized strokes, enabling the generation of structured and realistic completions. Similarly, SketchGAN [22] leveraged vector-based representations for joint sketch completion and recognition, utilizing cascaded networks to progressively refine incomplete sketches. Other GAN-based models, such as GANSketching [37], have further advanced vector-based approaches by introducing region-specific sketch refinement and stylistic adaptation.

Recent advancements have introduced more sophisticated models for sketch completion. Techniques like transformer-based architectures and hybrid models combining GANs and diffusion processes have significantly improved the quality and flexibility of sketch generation. For example, Vision Transformers (ViTs) [4] have been explored for their ability to capture long-range dependencies in generative tasks, including sketch synthesis. Models such as TransGAN [18] extend ViTs to adversarial frameworks, enhancing their application in artistic generation. Meanwhile, a revolution in generative models has slowly begun, gradually replacing GAN-based models with the diffusion models[14].

2.2 Diffusion Model

Diffusion models [14, 26, 36, 3, 35, 15] have undergone significant evolution, establishing themselves as a leading framework for generative modeling. Denoising Diffusion Probabilistic Models (DDPMs) [14] marked the first major milestone. These models define a forward process that gradually adds Gaussian noise to data and a reverse denoising process that recovers the original data distribution. The iterative nature of DDPMs enabled high-quality image synthe-

sis, surpassing GANs in terms of image fidelity and diversity. Subsequent works such as [26], introduced variance scheduling for better efficiency and sampling quality, expanding the applications of DDPMs to conditional generation tasks, such as text-to-image synthesis exemplified by GLIDE [25].

Building upon DDPMs, Song et al. proposed Denoising Diffusion Implicit Models (DDIMs) [35], which introduced a non-Markovian sampling process. This allowed for deterministic sampling and significant acceleration in generating samples. DDIMs retain high-quality outputs while reducing the number of required inference steps, making diffusion models more practical for large-scale tasks. The mathematical framework for continuous data distributions in diffusion models was unified through stochastic differential equations, as detailed in [36].

A revolutionary step in diffusion-based frameworks came with Latent Diffusion Models (LDMs) [31, 3, 33, 29]. LDMs operate in a lower-dimensional latent space, significantly reducing computational costs while preserving output quality. The advanced applications of LDM further highlighted the practicality of diffusion models. Stable Diffusion [31] applied LDMs to high-resolution image generation and multi-modal tasks like text-to-image synthesis, becoming a benchmark in open-source diffusion models. Meanwhile, Imagen [33] demonstrated the strength of diffusion models in text-guided generation, achieving photorealistic results. LDMs have also been extended to generate other modals of data, including sounds [2, 23] and videos [16, 19]. Additionally, ControlNet [39] extended diffusion models by introducing explicit control mechanisms, paving the way for structured tasks like sketch completion and segmentation.

2.3 Controllable Generative Model

Controllable generative models have been a significant focus in deep learning, allowing users to guide the generation process based on specific conditions in order to enhance the accuracy of generated image. As discussed in 2.1, early advancements centered on models like GANs [24, 17] and VAEs [38]. These previous work committed to controlled image synthesis, but limited by fidelity and stability.

Diffusion models has brought a great advancement in image synthesis, but it's hard to control the visual structure of the generated image, since text prompt is less informative than visual information. It's also a major research direction to get diffusion model more controllable. Classifier-Free Diffusion Guidance [15] simplified control mechanisms by training models jointly with and without conditions, making it possible to steer generation with text, class labels, or other inputs. Latent Diffusion Models [31] further advanced this area by performing diffusion in a compressed latent space, significantly improving efficiency while supporting multiple modalities such as text-to-image synthesis. DreamBooth [32] built on this foundation to personalize diffusion models, fine-tuning them on specific subjects using a combination of images and text descriptions. Similarly, Rinon Gal et al. [7] streamlined multi-modal guidance by compressing image-based conditioning into single-word tokens, enabling compact and scalable control.

ControlNet [39] marks a pivotal development in controllable diffusion models, offering a robust framework for multi-modal guidance. It addresses the limitations of earlier approaches by introducing a modular structure that appends trainable control layers duplicated from the frozen pre-trained diffusion model. This design eliminates the need for retraining the entire model, making it computationally efficient and highly adaptable. What sets ControlNet apart is its ability to seamlessly integrate diverse guidance signals—such as edge maps, depth maps, human poses, and semantic segmentation maps—within a unified framework. This flexibility allows users to manipulate both structural and semantic aspects of the generated content simultaneously, ensuring precise control over the output. Based on that, ControlNet is also deployed as the content extractor in diffusion-based style-transfer models [1, 12, 8]. By enabling multi-modal guidance, ControlNet significantly enhances the practical utility of diffusion models in applications such as interactive image editing.

3 Technical Approach

In this section, we will share the technical details behind the setup of our experiments. We used Stable Diffusion v1.5 as the base latent diffusion model onto which ControlNet was attached. This ControlNet implementation was sourced from Huggingface's Diffusers Github. This 2-year-old model is antiquated compared to state-of-the-art models due to the rapid improvement of LDMs. However, the principles of our method still apply here and produce clear interpretable results. There is nothing preventing this method from being applied to newer LDM foundation models as long as they are supported by ControlNet. Our model training modifications have two elements. First, we mutate the conditioning images from the dataset with masks. Second, we introduce hyper-parameter λ with the aim of having it control the guidance strength of the conditioning image.

3.1 Data Preparation

The training data for our method consisted of conditioning images sourced from the dataset LAION aesthetics sketch dataset [6]. To simulate partial sketches and emulate real-world user inputs, we applied the following pre-processing steps to these conditioning images:

- **Inversion**: The HED images in the dataset had white lines representing the edges, so we inverted the grayscale image colors to obtain something that looked like sketch lines
- **Negative Masking**: Each conditioning image was randomly masked with 1–3 negative circular masks of varying sizes. These masks created "windows" that revealed selected portions of the scene, approximating partial sketches.

These pre-processing steps introduced a diverse range of masked conditioning images into the training dataset, enhancing the model's ability to generalize and handle incomplete inputs effectively. Figure 1 illustrates examples of the masked conditioning images generated through this process.

Figure 1: Examples of masked conditioning images. Starting from original line art sketches sourced from the LAION aesthetics sketch dataset [6], random circular masks were applied to simulate partial sketches. This augmentation process mimics real-world sketching scenarios.

3.2 Conditioning Strength Parameter

We introduce a hyperparameter λ, hypothesizing that it will enable us to control the strength of the sketch conditioning in the same manner that context-free-guidance (CFG) controls prompt conditioning strength. This is achieved by scaling the vectorized conditioning image by the value of the hyperparameter.

$$\hat{\varepsilon} = \text{U-Net}(z_t, t, \text{ControlNet}(c))$$

$$c' = \lambda \cdot c$$

$$\hat{\varepsilon} = \text{U-Net}(z_t, t, \text{ControlNet}(c'))$$

These equations represent the simple modification of scaling our vectorized input to the ControlNet c by λ in relation to the network's overall noise prediction. During training, this value is randomly selected from a range of [0.5-1.5] to ensure the model is robust to variations at inference time.

3.3 GUI Application

In order to test user experience and assistive power of our model, we needed to create a GUI drawing application that harbors our model. This GUI was created using PyQt5. It does not support xformers acceleration due to a lack of Windows support. This application has a drawing canvas for the user to sketch on, a text box for the user to enter their prompt, and a "draw" button which triggers our model to generate an image conditioned on those user inputs. This generated image is then inserted under their sketch in the drawing environment at 50% opacity so that the user can still see their sketch. This process can be repeated until the user is satisfied with their drawing. Image generation takes approximately 6 seconds on a NVIDIA RTX 3060 GPU only with other typical consumer hardware.

The code for this application as well as the code for the model is provided in a GitHub repository and linked at the end of this paper.

4 Experiments

In this section, we evaluate the performance of our ControlNet-based model for incremental sketch prediction under various conditioning scenarios. The primary goal of these experiments is to assess how adjustments in training data and conditioning influence the system's ability to generate intermediate sketches that align with artistic intent.

We fine-tuned the Stable Diffusion 1.5 ControlNet model on the modified dataset. The training was conducted for 27,000 steps with a batch size of 8, requiring approximately 5 hours on an NVIDIA A100 GPU with 40 GB of VRAM. The experiments are compared against a baseline LineArt ControlNet to validate the proposed enhancements.

4.1 Ablation Study

To evaluate the contribution of our proposed modifications, we conducted an ablation study to compare the performance of models trained with and without the hyper-parameter λ.

- **Full Method**: Incorporates both masked conditioning images and the λ hyper-parameter. The hyper-parameter λ, which scales the guidance strength of the conditioning input, is sampled randomly from the range [0.5-1.5] during training. These fixed values were chosen to evaluate the robustness and consistency of the model's performance at varying levels of guidance strength.
- **No-λ Method**: Includes masked conditioning images but keeps the guidance strength fixed throughout training by omitting the λ hyper-parameter.

The models were evaluated on a test set of line art sketches with varying levels of completion. For qualitative analysis, we visualized incremental sketch predictions to examine the plausibility and coherence of the outputs relative to the input sketches.

Result: Figures 2 and 3 show the outputs of the full method and the no-λ method, respectively. Incorporating λ led to more realistic and context-aware completions, as evidenced by smoother transitions and improved alignment with the provided sketches. While the base model had the highest fidelity in the full-sketch case (row 1), it often failed to imagine the parts missing from the partial sketches. For instance, missing the carriage in row 3 of Figures 2. Our method performed much better in this partial-sketch task.

4.2 Dynamic Use Application

In addition to the creation of the model, we created a bare-bones GUI application to rapidly test the model in it's intended user environment. This allows the user to create a partial sketch and rapidly generate an image conditioned on their partial sketch and prompt using our model. This is essential for testing the assistive efficacy of our model.

Figure 2: Results using the full method with masked conditioning images (**Col. 1**) and the λ hyper-parameter set to 0.5, 1, and 1.5 (**Col. 3**, **Col. 4**, and **Col. 5**, respectively) during training compared to the baseline ControlNet LineArt model (**Col. 2**). The model exhibits plausible incremental sketch predictions, with smooth transitions and improved coherence across varying levels of guidance strength.

Result: This application proved to work well for the use of the model. The user sketch is used to guide the model's generation through the ControlNet, and the generated image is used by the user to guide their incrementally adjusted sketch. Figure 4 shows that the model is capable of generating a full scene that is consistent with the prompt even from a very simple incomplete sketch image. We can see from Figures 5 and 6 that the model is capable of imagining beyond the conditioning image to form an image coherent with both the sketch and prompt. For example, when given "Motorcycle jump off a ramp" and a sketch of a ramp, the model correctly imagines a motorcycle jumping off of the user-drawn ramp.

Figure 3: Results using the no-λ method (**Col. 3**) compared to the baseline ControlNet LineArt model (**Col. 2**). The absence of dynamic guidance strength led to outputs that are less robust to incomplete inputs, with reduced coherence compared to the full method.

5 Conclusion

This project introduces a novel approach to enhancing the ControlNet framework by integrating a data augmentation pipeline and a dynamic conditioning mechanism to enable incremental sketch-to-image generation using latent diffusion models. By applying negative circular masking to conditioning images and incorporating a guidance hyperparameter, we enable the model to extrapolate partial sketches into completed artworks. This methodology addresses critical limitations in existing sketch-based conditioning techniques by improving robustness to incomplete inputs and providing a controllable mechanism for guided image synthesis.

We demonstrated the efficacy of our approach using Stable Diffusion v1.5 as the base latent diffusion model, underscoring its potential for integration with more advanced diffusion models. Despite constraints imposed by the older architecture, the proposed method is adaptable to contemporary and future latent diffusion frameworks, enabling improved generalization and scalability. Furthermore, the development of a canvas application illustrates the practical implementation of our workflow, offering a platform for evaluating and utilizing the incremental sketch prediction capabilities in real-world artistic and creative contexts.

While this work advances the capabilities of diffusion models for sketch completion, there remain avenues for further exploration, such as refining augmentation strategies, stabilizing training, optimizing hyperparameter scaling, and leveraging state-of-the-art foundation models for enhanced performance. By addressing these challenges, future research can extend the applicability and precision of the proposed system, further advancing the intersection of generative modeling and creative design.

6 Code

The project code is available Here.

7 Figures

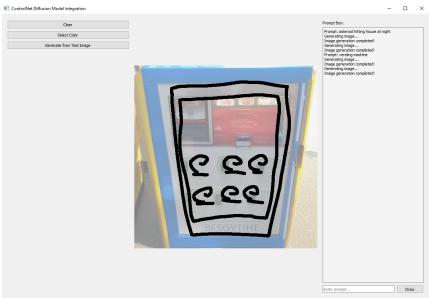

Figure 4: Example output generated by the model from the prompt "vending machine" and a simple partial sketch of a vending machine.

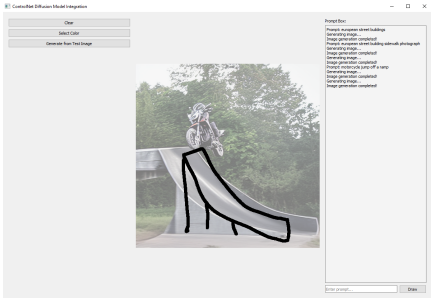

Figure 5: Example output generated by the model from the prompt "motorcycle jump off a ramp" and a simple partial sketch of a ramp.

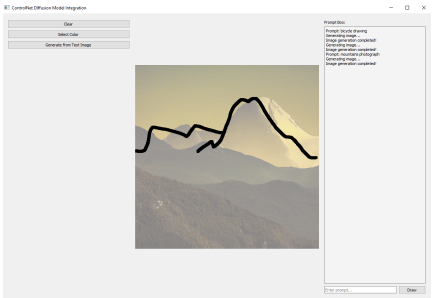

Figure 6: Example output generated by the model from the prompt "mountains photograph" and a simple partial sketch of a mountain range.

References

[1] Jingwen Chen, Yingwei Pan, Ting Yao, and Tao Mei. Controlstyle: Text-driven stylized image generation using diffusion priors. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7540–7548, 2023. 3

[2] Nanxin Chen, Yu Zhang, Heiga Zen, Ron J Weiss, Mohammad Norouzi, and William Chan. Wavegrad: Estimating gradients for waveform generation. *arXiv preprint arXiv:2009.00713*, 2020. 3

[3] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 2, 3

[4] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[5] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31(4):1–10, 2012. 2

[6] Erico25 and contributors. Laion aesthetics sketch dataset. https://huggingface.co/datasets/erico25/laion_aesthetics_sketch, 2024. Dataset for generating lineart sketches from LAION images. 4

[7] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[8] Junyao Gao, Yanchen Liu, Yanan Sun, Yinhao Tang, Yanhong Zeng, Kai Chen, and Cairong Zhao. Styleshot: A snapshot on any style. *arXiv preprint arXiv:2407.01414*, 2024. 3

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014. 2

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 2

[11] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 2

[12] Amir Hertz, Andrey Voynov, Shlomi Fruchter, and Daniel Cohen-Or. Style aligned image generation via shared attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4775–4785, 2024. 3

[13] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017. 2

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2

[15] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 2, 3

[16] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neu-*

ral Information Processing Systems, 35:8633–8646, 2022. 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2, 3

[18] Yifan Jiang, Shiyu Chang, and Zhangyang Wang. Transgan: Two pure transformers can make one strong gan, and that can scale up. *Advances in Neural Information Processing Systems*, 34:14745–14758, 2021. 2

[19] 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*, pages 15954–15964, 2023. 3

[20] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[21] Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive flow. *Advances in neural information processing systems*, 29, 2016. 2

[22] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5830–5839, 2019. 2

[23] Jinglin Liu, Chengxi Li, Yi Ren, Feiyang Chen, and Zhou Zhao. Diffsinger: Singing voice synthesis via shallow diffusion mechanism. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 11020–11028, 2022. 3

[24] Mehdi Mirza. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 2, 3

[25] 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. *arXiv preprint arXiv:2112.10741*, 2021. 3

[26] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 2, 3

[27] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[28] Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 2

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 3

[30] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. 2

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[32] 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*, pages 22500–22510, 2023. 3

[33] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 3

[34] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in neural information processing systems*, 29, 2016. 2

[35] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2, 3

[36] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 3

[37] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch your own gan. In *Proceedings of the*

IEEE/CVF International Conference on Computer Vision, pages 14050–14060, 2021. 2

[38] Xinchen Yan, Jimei Yang, Kihyuk Sohn, and Honglak Lee. Attribute2image: Conditional image generation from visual attributes. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 776–791. Springer, 2016. 2, 3

[39] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3

[40] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017. 2