

Stable Diffusion: A Study of the Text-to-Image Generation System

Om Jaiswal¹, Shrikant Shinde², Venkatesh Hoogar³, Devesh Bhadane⁴, Krishna Ghai⁵

Department of Computer Science and Engineering

MIT Art, Design and Technology University Pune -412201

ommega003@gmail.com¹, shri1506g@gmail.com², venkateshhoogar122@gmail.com³, krishnaghai17@gmail.com⁴,
deveshbhadane10@gmail.com⁵

Abstract—With the popularization of Stable Diffusion (SD) as a prominent Latent Diffusion Model (LDM), text-to-image (T2I) synthesis has become a new industry. This paper reports implementation of Stable Diffusion-based text-to-image generation system with advanced control procedures. Rather than a theoretical survey, the main focus of our work is the deployment system. In particular we introduce FastAPI as a backend framework, develop diffusion libraries for where models are pre-trained, ControlNet which provides structural guidance (Canny, Depth, Pose, Scribble), and a Prompt Intelligence Engine for automated prompt enhancement as a result of a Consistency Manager. The system includes Real-ESRGAN for upscaling and GFPGAN for face enhancement. Critical float16 precision, attention slicing, CPU offloading, xformers, and other optimization techniques ensure stable operation on consumer-grade hardware, providing as little as 4GB GPU memory. This implementation shows that a pre-trained T2I system can be designed for productionready models, not necessitating model training nor complex mathematical implementations.

Index Terms—Stable Diffusion, Latent Diffusion Models, Text-to-Image, Deep Generative Models, Diffusion Models, ControlNet, FastAPI, Computer Vision.

I. INTRODUCTION

A. The Context and Development of Deep Generative Models

Deep generative modeling as a field has experienced explosive evolution from early architectures like Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs) to the state-of-the-art, as shown in Diffusion Models (DMs). Earlier models showed great limitations; VAEs, while containing a structured latent space, often produced overly smooth or blurry outputs, while GANs often experienced problems like mode collapse, causing constraining the diversity of generated samples. The rise of DMs successfully closed the historical gaps we had created, developing a new standard for providing high fidelity to diverse data generation.

B. Motivation for Practical Implementation

Despite the better quality of DMs, their computational overhead cannot be discounted. Latent Diffusion Models (LDMs) were designed to alleviate this computational bottleneck by performing the iterative denoising procedure in a compressed latent space. This paper presents a fully implemented approach of a Stable Diffusion-based T2I generation system. The key

contributions of this work are: (1) a production-ready system architecture using FastAPI and the Diffusers library, (2) Integration of ControlNet for precise structural guidance across four conditioning modalities (Canny edge, depth, pose, and scribble), (3) Implementing a Prompt Intelligence Engine, a prompt enhancement to ensure that output quality improves, (4) a Consistency Manager to maintain character identity over diverse generated images, (5) a Story Generator for narrativeoriented multi-scene image synthesis, and (6) comprehensive optimization techniques enabling operation on 4GB GPU memory.

C. Overview of Implementation of the System

The provided system offers a full pipeline from text prompt to a high-quality generated image. Users interact through web and REST API and submit prompts with optional style presets and ControlNet conditioning images, and generation parameters. The backend is responsible for the Stable Diffusion model using the Diffusers library, optimizes memory-saving methods, and optionally enhances outputs through upscaling and face restoration. The core generative facility is made available by pre-trained Stable Diffusion models leveraged through Hugging Face's Diffusers library via the use of Stable Diffusion 1.5 and 2.1 versions which were loaded as pipeline objects.

II. SYSTEM ARCHITECTURE AND IMPLEMENTATION STACK

A. Backend Framework: FastAPI

FastAPI is used as the backend framework for the system, including asynchronous request handling, automatic OpenAPI documentation, and very high-performance REST API endpoints. FastAPI was chosen because it supports asynchronous operations natively, which are needed to operate the computationally intensive process of diffusion without blocking various other requests. The API exposes endpoints for single image generation, batch generation, story-based generation, and ControlNet guided generation. It is provided by pre-trained Stable Diffusion models accessed through Hugging Face's Diffusers library. Stable Diffusion 1.5 and 2.1 variants, loaded as pipeline objects that encapsulate the complete text-to-image generation workflow. The pipeline integrates four

essential components: (a) Text Encoder: A CLIP-based transformer that converts text prompts into semantic embeddings, (b) VAE (Variational Autoencoder): Compresses images to/from latent space, enabling efficient diffusion, (c) U-Net, which is the denoising network that iteratively removes noise from latent representations, and (d) Scheduler, which controls the denoising process supported by DDIM, DPM-Solver, and Euler schedulers. The system loads pre-trained weights only. This includes no training or fine-tuning.

B. Prompt Intelligence Engine

A custom Prompt Intelligence Engine enhances user-provided prompts before generation. The engine automatically appends quality-improving modifiers including lighting specifications (e.g., “cinematic lighting,” “soft diffused lighting”), detail boosters (e.g., “highly detailed,” “8K resolution,” “intricate textures”), and negative prompts (e.g., “blurry, low quality, distorted, deformed”). Style presets give users the ability to select from existing aesthetic categories like Photorealistic, Anime/Cartoon, Cinematic, Fantasy, Cyberpunk, Oil Painting, and Watercolor, each of which corresponds to different prompt modifications.

C. ControlNet for Structural Control

ControlNet is incorporated to ensure very specific spatial and structural control over generated images. The implementation supports four ControlNet conditioning methods: Canny Edge Control takes edge maps from reference images to guide composition and object boundaries; Depth Control (MiDaS) uses depth maps to preserve three-dimensional structure and spatial relationships; Pose Control (OpenPose) allows for skeleton-based human pose specification for character positioning; and Scribble Control respects rough hand-drawn sketches as structural guides. ControlNet works by creating a trainable copy of the U-Net encoder, linked to the frozen original U-Net by zero convolution layers.

D. Consistency Manager

The Consistency Manager targets a key problem in maintaining character identity for sequential image generation across multiple scenes. This implementation relies on two key mechanisms: (1) Seed Locking: Fixed random seeds ensure reproducible noise initialization among generation requests, and (2) Session Tracking: User sessions keep consistent generation parameters (e.g., seed values, guidance scales, and style presets). In the event that we want to create images of the same character in varying poses or environments, the Consistency Manager reuses the same base seed while allowing controlled variation through prompt modifications, achieving visual coherence without needing custom model training.

E. Story Generator

The Story Generator extends the system beyond single-image generation to narrative-based multi-scene synthesis. Users provide a story premise or character description, and the system

automatically generates a sequence of prompts corresponding to story beats. The character consistency of each scene remains constant using the Consistency Manager with different environmental and action descriptions. The produced image sequence can be used for storyboarding, comic art, or visual storytelling development.

F. Upscaling & Face Enhancement

Modules for post-processing improve the quality of generated images: Real-ESRGAN is a GAN-based super-resolution model, which upscales images 2x or 4x while restoring realistic textures and reducing artifacts. GFPGAN provides specialized face restoration that corrects common facial distortions in generated images, such as eye alignment, skin texture, and facial feature coherence. The modules function as optional post-processing steps, applied after diffusion completion.

G. Optimization Techniques for Memory-Constrained Hardware

To run on consumer-grade GPUs (4GB VRAM and above) and achieve these goals, the implementation includes several optimization strategies: float16 Precision reduces memory consumption by approximately 50% compared to float32; Attention Slicing splits attention computation into smaller slices, trading reduced peak memory for slightly increased computation time; CPU Offloading moves idle components to CPU memory, keeping only actively used components on GPU; and xformers Memory-Efficient Attention reduces memory footprint through block-sparse attention patterns. These techniques collectively support stable generation at resolutions up to 768x768 pixels on 4GB GPUs.

H. Environment and Dependencies

It is written in Python 3.10 alongside core dependencies, featuring including PyTorch 2.0+, tensor operations, as well as GPU acceleration, along with the Diffusers 0.24+ pre-trained diffusion pipeline, Transformers for CLIP text encoder implementation, and OpenCV for image preprocessing for ControlNet conditioning inputs. The backend of ControlNet will make use of OpenCV for preprocessing image strings, and Uvicorn with FastAPI on top of our async web framework system to capture data from processing steps. FastAPI Backend handles the user input via FastAPI, Prompt Intelligence, ControlNet, Consistency Manager, Stable Diffusion core; Generated Image.

III. THE MATHEMATICAL CONCEPTS OF LDM

The LDM framework is directly associated with the principles of Denoising Diffusion Probabilistic Models (DDPMs), that which can be analysed against the background of mathematical formalism of SDEs can be considered Stochastic differentials. The fundamental operation of LDMs occurs by a core use of DDPM mechanisms in the compressed latent space.

A. Forward Diffusion Process (Noising)

Forward diffusion model assumes gradual corruption of an original instance of data x_0 with a Markov chain of T adding isotropic Gaussian noise bit by bit. This transformation converts the pristine data x_0 to a pure Noise distribution x_T . The transition probability $q(x_t|x_{t-1})$ is set by a predefined variance schedule, β :

$$q(x_t|x_{t-1}) = N(x_t; \beta_t x_{t-1}, \beta_t \mathbf{I}) \quad (1)$$

where x_t is the noisy data at time t ; $\mathbf{I}$ is the identity matrix. One of the significant properties of the forward process is that the latent state x_t at any time t can be expressed directly in terms of the initial state x_0 .

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$ and $\epsilon \sim N(0, \mathbf{I})$.

B. Reverse Denoising Process (Generation)

The generative step is simply the reversal of the forward process utilizing a neural network (typically the U-Net architecture) parameterized by θ to predict and subtract the noise at each step. This learned reverse process $p_\theta(x_{t-1}|x_t)$ is approximated by a Gaussian distribution.

$$p_\theta(x_{t-1}|x_t) = N(x_{t-1}; \mu_\theta(x_t, t, c), \Sigma_\theta(x_t, t, c)) \quad (3)$$

The main job of the network is to estimate the noise component $\epsilon_\theta(x_t, t, c)$ and thereby compute the reconstruct less noisy state (mean μ_θ and covariance Σ_θ) x_{t-1} . The conditional embedding c has become part of the mean computation which is one of the mean calculation key concepts of Stable Diffusion's conditioned generation.

C. Training Objective and Stochastic Regularization

The DDPM training objective simplifies the optimization problem to reducing the mean-squared error in the predicted noise $\epsilon_\theta(x_t, t, c)$ and the true noise ϵ that were fed to the input data x_t . The objective function in equation 3 is simplified, where the key task is learning the noise predictor:

$$\mathcal{L}(\theta) = \mathbb{E}_{t, x_0, \epsilon} \left[\left\| \epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, t, c) \right\|^2 \right] \quad (4)$$

where ϵ is sampled from $N(0, \mathbf{I})$ and t uniformly sampled from the timesteps $1, \dots, T$. To achieve this naturally stochastic DDPM process through its inherently stochastic nature in latent space, the application of the existence of this inherently stochastic DDPM process in the latent space provides an important advantage by hindering the excessive generation of exaggerated smooth results, a powerful stochastic regulator. This makes it possible for efficient and robust outputs that act as a great stochastically perfect stabiliser.

IV. STABLE DIFFUSION ARCHITECTURE: COMPONENT ANALYSIS

The performance of Stable Diffusion is motivated from coordinated function based on three specialized sub-

components: VAE, the Denoising U-Net conditional and Text Encoder. In a cross-attention mechanism.

A. Variational Autoencoder and Latent Representation

The VAE serves as the perceptual compression system, mapping the high-dimensional model for the image x to a compact latent space of hyperplane representation z , with the compressed operation space fundamental to LDM efficiency. In LDM, the latent space is usually trained with KL-regularization for a coherent and smooth latent manifold, which supports stable sampling and downstream editing tasks. Its symmetric encoder-decoder structure and skip connections efficiently process the spatio-temporal dynamics that arise from the diffusion steps. Empirical research demonstrates that features extracted by U-Net's encoder change very little across the time-steps, capturing stable structural information. In contrast, the decoder features are highly variable highlighting their role in iterative refinement and synthesizing the details through noise prediction.

B. Text Encoding and Cross-Attention Mechanism

Conditioning on the text prompt y begins with a Text Encoder (τ_θ), a CLIP-oriented model frequently trained, which projects y into a semantic vector $c = \tau_\theta(y)$. This vector gets passed through the U-Net by specialized cross-attention layers found in the intermediate resolution blocks of the U-Net.

V. PERFORMANCE BOTTLENECKS AND INFERENCE OPTIMIZATION

Notwithstanding those huge efficiency improvements of a LDM computational paradigm unlike pixel-based diffusion architecture, the iterative nature of the denoising process still creates significant computational requirement, real-time deployment can be a problem and low-latency generation.

A. Typical Workload Characterization and Bottlenecking

Workload analysis repeatedly identifies the attention operation within transformer blocks in the U-Net as the main computational bottleneck. This is about 45% of the combined inference time, this is particularly important as image resolution heightens, that require customized optimization strategies.

B. Algorithmic Acceleration Techniques

Not only does research reduce steps from knowledge distillation, but it emphasizes computing costs per step. Feature reuse strategies take advantage of the empirical stability of the encoder's feature representations against similar time-steps. By omitting encoder computation at the intermediate step and reuse of the previous output, this approach achieves substantial speedups while preserving high generative quality. Conceptual Burden Distribution in Stable Diffusion Inference. Although attention instructions typically make up the majority of the total

inference time, most prominently at higher resolutions, which serves as a prime target for algorithmic and hardware acceleration.

VI. ADVANCED DIFFUSION SAMPLING AND SCHEDULERS

Although the DDPM framework is robust, its dependence on hundreds or thousands of steps to produce high-fidelity outputs is a performance bottleneck. Advanced sampling techniques address this by dramatically lowering needed size of the denoising iterations. The most important breakthrough, allowing sampling to run at a substantially faster pace, is the formulation of Song's Denoising Diffusion Implicit Models (DDIM) et al. DDIM is a rewrite of the diffusion process as a non-Markovian process, which means that there is deterministic generation rather than a stochastic Markov Chain. This deterministic nature allows the model to bypass large parts of the reverse trajectory, although it does not sacrifice the quality of the final process. The DDIM sampling process uses a nonMarkovian inverted step $p_{\theta}(x_{t-1}|x_t, x_0)$ which depends not just on the current noisy latent state x_t but also on the first data x_0 (informally inferred by the noise predictor ϵ_{θ}). The update rule simplifies to:

With our deterministic DDIM, research moved towards methods using numbers—diffusion models can easily be interpreted as solving a particular class of Ordinary Differential Equations (ODEs) for $\sigma_t = 0$. This perspective allows for a convenient and effective high-order ODE solver, like the DPMSolver series, which produces similar or better fidelity to the DDIM with only 10–20 sampling steps. These solvers treat the process of sampling as finding the solution of an ODE and proceed stepwise using numerical integration to navigate through time effectively and efficiently.

VII. EVALUATION FRAMEWORKS AND BENCHMARKING

LDMs are complex but require metrics for evaluation that validly describe quality and semantic congruence.

A. Conventional Metrics and Critique of FID

Traditional Frechet Inception Distance (FID) is much criticized as an unreliable primary metric for modern T2I models. Its limitations are due to three major issues: (1) Its dependence on the older Inception-v3 network produces feature representations that are ineffective at handling the multi-layered and rich work from recent T2I models. (2) FID works under wrong assumptions about normality of feature distribution. (3) It is affected by small sample complexity and yields poor results.

B. Emerging Evaluation Paradigms

To solve for the limits of FID, new metrics that are more efficient and converge with human perception and semantic

fidelity are gaining traction. CMMD (CLIP Maximum Mean Discrepancy) is presented as a more robust and dependable alternative. It leverages richer CLIP embeddings that are obtained from the same SD vision-language models for conditioning, making them inherently more pertinent to T2I fidelity. Structural precision benchmarks measure model utility between scientific and engineering domains with precisely engineered question sets that necessitate high degrees of structural accuracy. Image hallucination metrics like I-HallA employ question-answering systems in response to generated images to analyze semantic similarity.

VIII. APPLICATIONS AND DOMAIN-SPECIFIC EXTENSIONS

The computational efficiency and strong generative capacity of Stable Diffusion greatly increase its potential for generalization across diverse applications. The successful performance through modalities highlights the effective implementation of diffusion on an abstract latent representation as opposed to raw data.

A. Biomedical and Signal Processing

In biomedical informatics, diffusion models are becoming important for synthesis, reconstruction, and segmentation tasks. In medical imaging, specialized LDMs prove more efficient over traditional generative models in parameters such as FID, SSIM, precision, and recall. Beyond visual data, the LDM's key principles have been extended to non-image modalities. BioDiffusion does a remarkable job generating high-fidelity biomedical time-series signals (e.g., EEG and ECG), handling data privacy and noise-induced issues.

B. Engineering and Scientific Modeling

The second sub-domain is the process, which introduces necessary structure as well as stochasticities onto the latent space generating visually realistic realizations and statistically consistent with highly complex physical structures. The robustness of SD has utility in forensic science in generating facial images of suspects based on witness descriptions.

IX. HARDWARE ACCELERATION AND IMPLEMENTATION

Diffusion models were more or less forced by this nature into operation. Substantial research has been done on hardware and implementation strategies beyond algorithmic optimization.

A. Dedicated AI Accelerators

The efficient execution of LDMs rests on the efficient processing of the U-Net's high-volume convolutional processing and attention operations. Custom AI accelerators give the user a view on the possibility to reduce power consumption and achieve top-tier performance in generative workloads while offering up to 50% energy savings compared to generalpurpose GPUs.

B. Quantization and Pruning Techniques

To launch such device deployment on edge devices and consumer machines, extreme quantization (e.g., to 4-bit or 8-bit precision) and pruning techniques are indispensable. Research indicates that pruning the model size reduces it. The result is that inference with minimal quality degradation, aggregates applied in time. Aggressive quantization often requires a complex post-training calibration or hardware-aware retraining.

X. ADVANCED FINE-TUNING AND MODEL CUSTOMIZATION

For industrial use, model customization techniques are more critical than ever to maintain good quality and personalization. Pretrained LDMs learn new visual concepts, styles, and limited datasets.

A. Parameter-Efficient Fine-Tuning (PEFT)

Low-Rank Adaptations (LoRA) is the popular PEFT method that alleviates the memory cost of full fine-tuning. By freezing the pre-trained weights and injecting small, trainable rank decomposition matrices into transformer building blocks, LoRA decreases trainable parameters by orders of magnitude. This allows for quick, GPU-friendly tweaking of specific styles or subjects.

B. Injecting New Concepts: Textual Inversion and DreamBooth

To transfer special ideas or pieces into the LDM vocabulary, two main methods are used: Textual Inversion tunes a new token in the Text Encoder's vocabulary to represent a new image, simultaneously DreamBooth fine-tunes layers of the Denoising U-Net and optimizes a new Text Encoder embedding to connect a unique identifier with a subject.

C. Structural and Conditional Control

If explicit spatial or structural control is demanded, ControlNet uses a parallel, trainable copy of the U-Net encoder with distinct dedicated zero convolution layers connected to the original U-Net. This allows the model to accept auxiliary inputs, for example, Canny edges or human pose skeletons, allowing for accurate control of the generated image's composition and structure. Recent study keeps taking the architectural limits forward. Areas of LDMs for more efficiency, scalability, and integration with the new language modeling paradigm.

XI. EMERGING ARCHITECTURAL DIRECTIONS

A. Transformer-based Diffusion Models (DART)

Architectures such as DART (Denoising Autoregressive Transformer) remove the convolutional U-Net entirely, considering the latent image representation as a sequence of tokens. This allows the model to take advantage of great

advances in scalability, context modeling, and long-range dependency handling for large language models (LLMs).

B. Contextualized Diffusion (ContextDiff)

ContextDiff refines the LDM framework by explicitly embedding integration of a history context buffer and an attention mechanism designed to monitor and help enforce consistency between sequences of images or complex textual instructions. This is essential for applications such as text-to-video, in which standard LDMs commonly suffer from temporal incoherence.

C. Synthesis-Understanding Interaction (ArtAug)

The ArtAug framework implements an interactive mechanism where the generated image is dynamically evaluated by a vision-language understanding component. If the image fails to meet certain semantic requirements, the understanding module provides feedback to modify the original prompt or suggest subsequent diffusion steps.

XII. ETHICAL, LEGAL, AND SOCIETAL IMPLICATIONS (ELSI)

The accessibility and photorealism of systems like Stable Diffusion require a deeper analysis of their ethical, legal, and societal footprint.

A. Copyright, Training Data, and Legal Challenges

A large portion of the time, the training process of modern generative AI models relies on gigantic datasets with sometimes copyrighted material. The risk of memorization is still a major legal liability, as a model that directly reproduces training data. Lawsuits filed by artists vs LDM creators call for better transparency of data sourcing and the establishment of fair compensation mechanisms.

B. Deepfakes, Misinformation, and Trustworthiness

LDMs' ability to synthesize hyperrealistic forms of media greatly echoes the current media credibility crisis. The rise of high-quality deepfakes poses a serious potential social damage, such as emotional distress, financial fraud, and influencing public opinion. Mitigation efforts need to create suitable detection tools, public education initiatives, and also digital watermarking, and research on digital watermarking and provenance tracking.

C. Algorithmic Bias and Ethical Governance

LDMs can inherit and spread societal biases in their enormous training datasets. Global initiatives like the EU's AI Act and IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems emphasize fundamental concepts of human well-being, fairness, accountability, and transparency (FAIR).

XIII. CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Models & Tools to Leverage the Latent Diffusion Model paradigm to implement Stable Diffusion is quite a breakthrough in text-to-image synthesis. This paper outlined a practical implementation of a Stable Diffusion system with FastAPI backend, Diffusers library integration, ControlNet structural guidance, Prompt Intelligence Engine, Consistency Manager, Story Generator, and memory-constrained hardware optimization methodologies. Significant challenges remain:

- 1) Computational Intensity: Despite acceleration techniques, latency for high-resolution output is still a bottleneck.
- 2) Structural Fidelity and Hallucination: Models are ongoing to grapple with compositional tasks and verifiable precision.
- 3) Ethical and Legal Vulnerabilities: Facing the risks of memorization, deepfake, and algorithmic bias call for governance integration.
- 4) Multi-Modal Extension: Extension to video generation and 3D scene creation need new latent space definitions.

XIV. REFERENCES

- [1] M. Elasmri, O. Elharrouss, S. Al-Maadeed, and H. Tairi, "Image generation: A review," *Neural Processing Letters*, Feb. 2022. [Online]. Available: <https://doi.org/10.1007/s11063-022-10777-x>
- [2] S. Ramzan, M. M. Iqbal, and T. Kalsum, "Text-to-image generation using deep learning," *Engineering Proceedings*, vol. 20, p. 16, Jul. 2022. [Online]. Available: <https://doi.org/10.3390/engproc2022020016>
- [3] A. Ramesh et al., "Zero-shot text-to-image generation," arXiv:2102.12092, Feb. 2021. [Online]. Available: <https://arxiv.org/abs/2102.12092>
- [4] Y. Lim, H. Choi, and H. Shim, "I-HallA: Evaluating Image Hallucination...", arXiv:2409.12784, Jul. 2024. [Online]. Available: <https://arxiv.org/abs/2409.12784>
- [5] Z. Duan et al., "ArtAug: Enhancing Text-to-Image Generation...", arXiv:2412.12888v2, Dec. 2024. [Online]. Available: <https://arxiv.org/abs/2412.12888>
- [6] L. Yang et al., "ContextDiff: Contextualized Diffusion Model...", arXiv:2402.16627v3, Feb. 2024. [Online]. Available: <https://arxiv.org/abs/2402.16627>
- [7] J. Gu et al., "DART: Denoising Autoregressive Transformer...", arXiv:2410.08159v2, Oct. 2024. [Online]. Available: <https://arxiv.org/abs/2410.08159>
- [8] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," arXiv:1312.6114, Dec. 2013. [Online]. Available: <https://arxiv.org/abs/1312.6114>
- [9] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional GANs," *Proc. ICLR*, 2016.
- [10] J. Wu et al., "Learning a probabilistic latent space of object shapes...", *NeurIPS*, pp. 82–92, 2019.
- [11] R. D. Hjelm et al., "Learning deep representations by mutual information estimation...", *Proc. ICLR*, 2019.
- [12] S. D. McDermott and M. W. Mahoney, "Adaptive importance sampling for diffusion-based generative models," arXiv:2102.02760, 2021.
- [13] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," *Proc. ICLR*, 2018.
- [14] Y. Zhang et al., "Self-supervised learning for image synthesis and manipulation," *Proc. CVPR*, 2020.
- [15] C. Ledig et al., "Photo-realistic single image super-resolution using a GAN," *Proc. CVPR*, 2017.
- [16] A. Brock, J. Donahue, and K. Simonyan, "Understanding and improving interpolation in autoencoders...", *Proc. ICLR*, 2019.
- [17] A. Brock et al., "Generative and discriminative voxel modeling with CNNs," arXiv:1608.04236, 2016.