Review

Not peer-reviewed version

Diffusion Models at Scale: Techniques, Applications, and Challenges

Sa'dia Abul-Fazl , Rasim Dina * , Hafez Fairuza

Posted Date: 3 February 2025

doi: 10.20944/preprints202502.0029.v1

Keywords: diffusion models; generative modeling; scalability; noise schedules; sampling efficiency; multimodal applications; high-resolution generation; deep learning; ethical AI; computational efficiency

 doi:10.20944/preprints202502.0029.v1

Article

Diffusion Models at Scale: Techniques, Applications, and Challenges

Sa'dia Abul-Fazl †, Rasim Dina *,† and Hafez Fairuza

Kingdom of Saudi Arabia, KAUST, King Abdullah University of Science and Technology, Saudi Arabia
* Correspondence: rasim.dina@kaust.edu.sa
† Equal contribution

Abstract: Diffusion models have emerged as a powerful class of generative models, offering state-of-the-art performance across various domains such as image synthesis, audio generation, and molecular design. Their unique approach, which involves modeling data distributions through iterative noise addition and denoising processes, has established them as a robust alternative to traditional generative frameworks like GANs and VAEs. However, the scalability of diffusion models—essential for handling high-dimensional data, large-scale datasets, and complex multimodal tasks—poses significant challenges. This survey provides a comprehensive overview of scalable diffusion models, focusing on the innovations that enable their efficient training and sampling. We explore advancements in noise schedules, neural architectures, and sampling acceleration techniques, alongside strategies for training on large-scale datasets and deploying models in resource-constrained environments. Furthermore, we highlight the transformative applications of scalable diffusion models across fields such as creative content generation, healthcare, scientific research, and more. Despite their successes, diffusion models face critical challenges, including computational inefficiency, resource-intensive training, and ethical concerns related to bias and misuse. We discuss these open challenges and outline promising directions for future research, emphasizing the need for interdisciplinary collaboration and task-specific adaptations. By addressing these challenges, scalable diffusion models have the potential to redefine the boundaries of generative modeling, driving innovation and enabling new applications in science, technology, and creative industries. This survey aims to serve as a valuable resource for researchers and practitioners seeking to understand and advance the field of diffusion models.

Keywords: diffusion models; generative modeling; scalability; noise schedules; sampling efficiency; multimodal applications; high-resolution generation; deep learning; ethical AI; computational efficiency

1. Introduction

Generative modeling has been a central focus in machine learning, enabling the creation of synthetic data that mimics the underlying structure of real-world datasets [1]. Among the many generative frameworks proposed, diffusion models have recently garnered significant attention for their ability to generate high-quality data across diverse domains, including image synthesis, audio generation, text-to-image translation, and molecular design [2]. By iteratively refining noisy samples toward desired outputs, diffusion models have demonstrated unprecedented performance, rivaling or surpassing alternatives like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Diffusion models operate on the principle of modeling the process of data destruction and subsequent reconstruction [3]. Specifically, these models learn to reverse a predefined diffusion process, which gradually corrupts data with noise [4]. Through this reversal, they can generate new samples by starting from pure noise and progressively denoising it [5]. This framework, rooted in concepts from stochastic processes and thermodynamics, provides a theoretically sound and flexible approach to generative modeling [6]. Despite their conceptual elegance, early implementations of diffusion models were computationally expensive and struggled to scale efficiently [7]. The scalability of diffusion models has become a critical area of research as modern applications demand the ability to

 doi:10.20944/preprints202502.0029.v1

generate high-resolution outputs, process large-scale datasets, and handle complex data distributions [8]. For instance, in image generation, achieving photo-realistic outputs at high resolutions requires models that are both computationally efficient and robust against the challenges of large-scale data [9]. Similarly, in applications like text-to-image synthesis, scalability is crucial for bridging the gap between diverse modalities and ensuring consistent, high-quality outputs. Recent advancements in both hardware and algorithmic design have played a pivotal role in addressing these challenges [10]. Innovations such as improved noise schedules, advanced neural architectures, and techniques like classifier-free guidance have significantly enhanced the efficiency and quality of diffusion models [11]. Moreover, leveraging large-scale pretraining and distributed computing has enabled these models to scale to unprecedented levels, opening new frontiers in generative modeling [12]. Despite these advancements, several challenges persist [13]. The iterative sampling process inherent to diffusion models is computationally intensive, often requiring hundreds or thousands of steps to produce a single output. This inefficiency becomes particularly problematic when scaling to high-dimensional data or deploying models in real-time applications. Furthermore, the training of diffusion models typically requires vast computational resources, limiting their accessibility to researchers and practitioners with significant infrastructure [14]. Addressing these limitations remains a primary focus of ongoing research. This survey aims to provide a comprehensive overview of scalable diffusion models, examining the methodologies and techniques that have enabled their growth and adaptation to modern challenges [15]. We begin by presenting the foundational concepts of diffusion models, exploring their mathematical underpinnings and operational mechanisms. Following this, we delve into scalability strategies, highlighting innovations in model architecture, training paradigms, and sampling techniques. We also discuss the wide range of applications that have benefited from scalable diffusion models, emphasizing their impact on fields such as computer vision, natural language processing, and scientific discovery [16]. In addition to reviewing the state of the art, we identify key challenges and open research questions in the field of scalable diffusion models [17]. These include improving sampling efficiency, reducing computational costs, and enhancing the adaptability of models to diverse datasets and modalities [18]. By addressing these questions, we hope to inspire future research and contribute to the ongoing advancement of diffusion models [19]. Through this survey, we aim to provide a valuable resource for researchers and practitioners interested in diffusion models and their scalability [20]. By synthesizing recent developments and identifying promising directions for future work, we seek to foster a deeper understanding of this transformative generative modeling framework and its potential to drive innovation across disciplines [21].

2. Background and Fundamentals

To understand the advancements and scalability of diffusion models, it is essential to first grasp their foundational principles [22]. This section provides an overview of the mathematical and theoretical underpinnings of diffusion models, their operational framework, and their relationship to other generative modeling approaches [23].

2.1. Diffusion Processes and Stochastic Modeling

Diffusion models are inspired by the concept of stochastic processes, particularly the diffusion process, which describes the gradual transformation of data into noise over time [24]. Formally, a diffusion process is defined as a continuous-time Markov process where the state evolves according to a stochastic differential equation (SDE) [25]. Let $\mathbf{x}_0$ represent the original data sample [26]. The forward diffusion process adds Gaussian noise to the data over T timesteps, resulting in a noisy sample $\mathbf{x}_T$ [27]. The forward process is defined as:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}),$$

where α_t is a noise schedule that determines the variance of noise added at each timestep [28]. Over multiple iterations, $\mathbf{x}_t$ becomes increasingly noisy, eventually resembling pure Gaussian noise [29].

 doi:10.20944/preprints202502.0029.v1

The reverse process, parameterized by a neural network, aims to reconstruct the original data by denoising $\mathbf{x}_T$ back to $\mathbf{x}_0$ [30]. This process is learned by minimizing a variational bound on the negative log-likelihood of the data [31].

2.2. Training Objective

The core objective of diffusion models is to learn the reverse diffusion process [32]. This is achieved by optimizing the following loss function:

$$L = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\mathbf{x}_t, t)\|^2\right],$$

where ϵ is the noise added during the forward process, and ϵ_θ is the neural network's prediction of the noise [33]. This objective encourages the model to accurately estimate the noise at each timestep, enabling effective denoising during sampling.

2.3. Relationship to Other Generative Models

Diffusion models share similarities with other generative frameworks, such as GANs and VAEs, but differ in their approach to modeling data distributions. Unlike GANs, which rely on adversarial training, diffusion models are trained using a maximum likelihood-based objective, avoiding issues like mode collapse [34]. Compared to VAEs, diffusion models provide higher-quality samples by directly modeling the data distribution without requiring a latent space [35].

2.4. Sampling Process

The sampling process in diffusion models begins with a random noise vector $\mathbf{x}_T$ sampled from a Gaussian distribution [36]. The model then iteratively applies the learned reverse process to denoise the sample over T timesteps, producing the final output $\mathbf{x}_0$. While this iterative process ensures high-quality outputs, it is computationally intensive, as it typically requires hundreds or thousands of steps [37].

2.5. Scalability Challenges

The iterative nature of diffusion models poses significant challenges for scalability [38]. High-dimensional data, such as high-resolution images or multimodal inputs, increases the computational cost of both training and sampling [39]. Furthermore, achieving competitive performance often requires large-scale datasets and extensive computational resources. Addressing these scalability issues is critical for deploying diffusion models in real-world applications [40]. This foundational understanding sets the stage for exploring the advancements and techniques that enable scalable diffusion models. In the next section, we delve into the methodologies and innovations that have been developed to overcome these challenges [41].

3. Scalability Techniques for Diffusion Models

Scaling diffusion models to handle large datasets, high-dimensional data, and complex tasks requires innovations across multiple dimensions, including model architecture, training paradigms, and sampling efficiency [42]. This section reviews the key techniques and strategies developed to address the scalability challenges of diffusion models.

3.1. Efficient Noise Schedules

The noise schedule, which governs the amount of noise added at each timestep during the forward process, plays a critical role in the performance and efficiency of diffusion models [43]. Traditional linear schedules often require a large number of timesteps to achieve high-quality results. Recent work has explored non-linear noise schedules, such as cosine schedules, which allocate noise more effectively across timesteps. These schedules enable faster convergence and reduce the number of iterations required for sampling, improving scalability [44].

 doi:10.20944/preprints202502.0029.v1

3.2. Improved Neural Architectures

The choice of neural network architecture significantly impacts the scalability of diffusion models [45]. State-of-the-art architectures, such as U-Net, have been widely adopted due to their ability to capture multi-scale features efficiently [46]. Recent advancements include:

- **Cross-Attention Mechanisms:** Incorporating cross-attention layers enables diffusion models to handle multimodal inputs, such as text-to-image tasks, by effectively fusing information from different modalities [47].
- **Hierarchical Models:** Leveraging hierarchical structures allows models to process data at multiple resolutions, reducing computational overhead while maintaining high-quality outputs.
- **Lightweight Architectures:** Designing lightweight networks with fewer parameters reduces memory requirements and training time, making diffusion models more accessible [48].

3.3. Accelerated Sampling Techniques

The iterative nature of the sampling process is a major bottleneck for diffusion models. To address this, researchers have developed techniques to accelerate sampling without compromising quality:

- **Denoising Diffusion Implicit Models (DDIM):** DDIM introduces a deterministic sampling process that reduces the number of timesteps required, enabling faster generation [49].
- **Dynamic Programming Methods:** These methods optimize the reverse process by adaptively selecting timesteps, focusing computational resources where they are most needed [50].
- **Score-Based Methods:** Score-based generative models approximate the gradient of the data distribution, allowing for more efficient sampling through improved step sizes and noise schedules.

3.4. Training on Large-Scale Datasets

The performance of diffusion models improves with the availability of large-scale datasets [51]. However, training on such datasets presents challenges, including memory limitations and extended training times [52]. Solutions include:

- **Distributed Training:** Leveraging distributed computing frameworks enables efficient training across multiple GPUs or TPUs, significantly reducing training time.
- **Curriculum Learning:** Gradually increasing the complexity of training data helps models converge faster and achieve better generalization [53].
- **Synthetic Data Augmentation:** Augmenting datasets with synthetic samples generated by smaller models can bootstrap the training of larger diffusion models [54].

3.5. Compression Techniques for Diffusion Models

The computational demands of diffusion models, particularly for training and inference, present significant challenges in terms of storage, memory, and energy consumption. As these models scale to handle larger datasets and higher resolutions, compression techniques have emerged as essential tools for reducing their resource requirements while maintaining performance. This subsection explores various compression strategies tailored to diffusion models.

3.5.1. Model Pruning

Model pruning involves removing redundant or less significant parameters from a trained diffusion model to reduce its size and computational complexity. Key approaches include:

- **Magnitude-Based Pruning:** Parameters with magnitudes below a certain threshold are removed, simplifying the model without significant loss in performance.
- **Structured Pruning:** Entire layers, neurons, or attention heads are pruned, resulting in a more compact architecture suitable for efficient deployment[55–57].
- **Iterative Pruning and Fine-Tuning:** Pruning is performed iteratively, followed by fine-tuning to recover any lost performance.

 doi:10.20944/preprints202502.0029.v1

3.5.2. Quantization

Quantization reduces the precision of model parameters, such as weights and activations, from high-precision formats (e.g., 32-bit floating point) to lower-precision formats (e.g., 8-bit integers). This technique significantly reduces memory usage and accelerates computation. Common quantization strategies include:

- **Post-Training Quantization:** Quantization is applied after training without additional modifications to the training process.
- **Quantization-Aware Training:** The model is trained with quantization in mind, improving its robustness to reduced precision.
- **Mixed-Precision Techniques:** Different parts of the model are quantized to varying levels of precision based on their sensitivity to quantization errors.

3.5.3. Knowledge Distillation

Knowledge distillation transfers knowledge from a large, computationally expensive diffusion model (the "teacher") to a smaller, more efficient model (the "student"). The student model is trained to mimic the teacher's outputs, often achieving comparable performance with significantly fewer parameters. Key considerations include:

- **Distillation Objectives:** Designing loss functions that align the student's outputs with the teacher's, including logits, intermediate feature maps, or denoising trajectories.
- **Task-Specific Distillation:** Tailoring the distillation process to specific tasks, such as image synthesis or text-to-image generation.
- **Cross-Modal Distillation:** Transferring knowledge from multimodal teacher models to task-specific student models[5,47,56].

3.5.4. Parameter Sharing and Factorization

Parameter sharing and factorization techniques reduce the number of unique parameters in a diffusion model by exploiting redundancies [58,59]. Examples include:

- **Weight Sharing:** Sharing weights across different layers or timesteps to reduce model size.
- **Low-Rank Factorization:** Decomposing weight matrices into low-rank approximations, reducing the number of parameters while preserving representational capacity.
- **Tensor Decomposition:** Applying techniques like singular value decomposition (SVD) or Tucker decomposition to compress large parameter tensors.

3.5.5. Sparse Representations

Sparse representations aim to represent model weights and activations with a high degree of sparsity, enabling efficient computation. Techniques include:

- **Sparse Training:** Enforcing sparsity constraints during training to produce inherently sparse models.
- **Post-Training Sparsification:** Applying sparsity-inducing regularizers or thresholding methods to trained models.
- **Dynamic Sparsity:** Adjusting sparsity patterns dynamically during training or inference to optimize performance[60].

3.5.6. Hybrid Compression Techniques

Combining multiple compression techniques can further enhance efficiency while mitigating individual limitations. For example:

- **Pruning and Quantization:** Applying pruning to reduce model size, followed by quantization to accelerate inference.

 doi:10.20944/preprints202502.0029.v1

- **Distillation and Low-Rank Factorization:** Using knowledge distillation to train a smaller model and factorizing its parameters for additional compression.
- **Sparse Quantization:** Combining sparsity with quantization to achieve both storage and computational efficiency.

3.5.7. Challenges and Future Directions

While compression techniques offer significant benefits, several challenges remain in their application to diffusion models:

- **Maintaining Quality:** Ensuring that compression does not degrade the quality of generated outputs, especially for high-resolution or multimodal tasks.
- **Task-Specific Tuning:** Adapting compression techniques to the unique requirements of diffusion models, such as iterative denoising and time-step-dependent operations.
- **Scalability:** Extending compression methods to accommodate larger models and datasets without excessive computational overhead.

Future research could focus on developing compression techniques tailored specifically to the unique characteristics of diffusion models, enabling their deployment in resource-constrained environments while maintaining high performance.

3.6. Hybrid and Modular Approaches

Combining diffusion models with other generative frameworks, such as GANs or VAEs, can enhance scalability [61]. Hybrid approaches leverage the strengths of multiple methods, such as the fast sampling of GANs and the stability of diffusion models. Modular architectures, which decompose the generative process into smaller, independent components, further improve scalability by reducing the complexity of individual modules.

3.7. Hardware and Optimization Advances

Advancements in hardware and optimization techniques have also contributed to the scalability of diffusion models [62]. These include:

- **Mixed-Precision Training:** Utilizing lower-precision formats, such as FP16, reduces memory usage and accelerates training without significant loss in accuracy [63].
- **Custom Hardware Accelerators:** Dedicated accelerators, such as GPUs and TPUs optimized for deep learning workloads, have significantly reduced the computational burden of training and sampling.
- **Gradient Accumulation and Checkpointing:** These techniques optimize memory usage during training, enabling the handling of larger batch sizes and models.

3.8. Scalability in Multimodal and High-Resolution Applications

Scaling diffusion models for multimodal tasks, such as text-to-image generation, and high-resolution data, such as 4K image synthesis, requires task-specific innovations [64]. These include:

- **Multimodal Pretraining:** Training models on diverse datasets spanning multiple modalities improves their ability to generalize across tasks [65].
- **Progressive Resolution Techniques:** Generating data at progressively higher resolutions reduces the computational cost of high-resolution synthesis [66].
- **Guided Diffusion:** Techniques like classifier-free guidance improve the control and quality of generated outputs, especially in multimodal settings [67].

This section highlights the diverse strategies employed to make diffusion models scalable and efficient [68]. These innovations have not only expanded the applicability of diffusion models but have also made them a viable option for real-world applications requiring high performance and

 doi:10.20944/preprints202502.0029.v1

adaptability. In the next section, we explore the state-of-the-art applications of scalable diffusion models and their impact across various domains [69].

4. Applications of Scalable Diffusion Models

The scalability of diffusion models has unlocked their potential for a wide range of applications across diverse fields. From generating realistic images to facilitating scientific discovery, scalable diffusion models have demonstrated their versatility and transformative impact [70]. This section explores some of the most prominent applications enabled by advancements in scalability [17].

4.1. Image Synthesis and Editing

One of the most well-known applications of diffusion models is in image synthesis, where they have achieved state-of-the-art performance in generating high-quality, photorealistic images [71]. Scalable diffusion models are capable of producing high-resolution outputs that rival or surpass those generated by GANs [72]. Key use cases include:

- **High-Resolution Image Generation:** Models like DALL·E 2 and Stable Diffusion generate detailed images at resolutions up to 4K, enabling applications in digital art, design, and entertainment [73].
- **Image Inpainting and Editing:** Diffusion models can seamlessly fill in missing parts of an image or edit existing images with user-specified modifications, making them valuable tools for content creation [74].
- **Style Transfer and Customization:** By conditioning on specific style or content inputs, diffusion models can generate images tailored to user preferences [75].

4.2. Text-to-Image Synthesis

Scalable diffusion models have revolutionized multimodal tasks, particularly text-to-image synthesis. By leveraging cross-attention mechanisms and large-scale pretraining, these models can generate images that align closely with textual descriptions [76]. Applications include:

- **Creative Content Generation:** Artists and designers use text-to-image models to create illustrations, concept art, and visual storytelling elements [77].
- **Advertising and Marketing:** Businesses employ these models to generate customized visuals for advertisements and promotional materials based on specific themes or messages [78].
- **Accessibility and Education:** Text-to-image models enhance accessibility by generating visual aids for educational content or assisting visually impaired individuals in understanding textual information.

4.3. Audio and Speech Generation

Diffusion models have also been adapted for audio synthesis, enabling applications in music, speech, and sound design. Examples include:

- **Speech Synthesis:** Diffusion models generate natural-sounding speech with high fidelity, finding use in virtual assistants, dubbing, and accessibility tools [79].
- **Music Generation:** These models create original compositions or remix existing tracks, aiding musicians and content creators in their workflows [80].
- **Sound Effects Design:** Generating realistic sound effects for films, games, and virtual environments is another emerging application of diffusion-based audio models [81].

4.4. Molecular and Drug Design

In scientific research, scalable diffusion models have shown promise in molecular and drug design by generating novel chemical structures with desired properties. Applications include:

- **Drug Discovery:** Diffusion models assist in identifying potential drug candidates by exploring vast chemical spaces efficiently [82].

 doi:10.20944/preprints202502.0029.v1

- **Protein Design:** These models generate protein structures optimized for specific functions, accelerating advancements in biotechnology and medicine.
- **Material Science:** Generating new materials with tailored properties is another area where diffusion models are being actively explored [83].

4.5. Creative Applications

The creative industries have embraced scalable diffusion models for a variety of purposes, including:

- **Digital Art and Animation:** Artists use diffusion models to create unique artworks and animations, expanding the possibilities of creative expression.
- **Game Design:** These models generate assets, such as characters, environments, and textures, streamlining the game development process [84].
- **Film and Media Production:** Diffusion models aid in visual effects creation, storyboarding, and content generation for films and media projects.

4.6. Healthcare and Medical Imaging

In healthcare, diffusion models have demonstrated potential in enhancing medical imaging and diagnostics. Applications include:

- **Medical Image Reconstruction:** Diffusion models improve the quality of medical images, such as MRI and CT scans, by denoising and enhancing resolution [85].
- **Anomaly Detection:** These models assist in identifying anomalies in medical images, aiding in early diagnosis and treatment planning.
- **Synthetic Data Generation:** Generating synthetic medical data helps address data scarcity while preserving patient privacy [86].

4.7. Scientific Research and Simulation

Beyond healthcare, diffusion models are increasingly used in scientific research and simulations. Examples include:

- **Climate Modeling:** Generating high-resolution climate simulations to predict weather patterns and study environmental changes [87].
- **Physics Simulations:** Modeling complex physical systems, such as fluid dynamics and particle interactions, with high accuracy [88].
- **Astronomy and Space Exploration:** Enhancing astronomical images and generating realistic simulations of celestial phenomena [89].

4.8. Open Challenges in Applications

While diffusion models have achieved remarkable success across these domains, several challenges remain. These include improving sampling efficiency for real-time applications, ensuring robustness across diverse datasets, and addressing ethical concerns such as misuse and bias in generated content [90]. Overcoming these challenges is critical for maximizing the societal impact of scalable diffusion models. This section highlights the transformative applications of scalable diffusion models, showcasing their versatility and potential [91]. In the next section, we discuss the open challenges and future directions for advancing diffusion models and addressing their current limitations [92].

5. Open Challenges and Future Directions

While scalable diffusion models have achieved significant advancements and demonstrated their versatility across various domains, several challenges and limitations remain [93]. Addressing these issues is essential for further improving their efficiency, applicability, and societal impact. This section outlines key challenges and explores promising directions for future research [94].

 doi:10.20944/preprints202502.0029.v1

5.1. Sampling Efficiency

One of the most prominent challenges in diffusion models is the computational inefficiency of the sampling process [95]. The iterative nature of sampling, often requiring hundreds or thousands of steps, makes real-time applications and deployment in resource-constrained environments challenging [96]. Promising directions for addressing this issue include:

- **Reduced-Step Sampling:** Developing techniques that minimize the number of timesteps required for sampling without compromising output quality, such as improved noise schedules and learned sampling strategies [97].
- **Parallel Sampling:** Exploring methods to parallelize the sampling process, leveraging advancements in hardware accelerators and distributed computing.
- **Hybrid Approaches:** Combining diffusion models with other generative frameworks, such as GANs, to leverage the fast sampling capabilities of alternative methods [98].

5.2. Training Efficiency and Resource Requirements

Training diffusion models on large-scale datasets is computationally expensive, requiring significant memory and time resources. This limits their accessibility to researchers and practitioners with limited computational infrastructure [99]. Future research could focus on:

- **Data-Efficient Training:** Developing training paradigms that require fewer data samples, such as self-supervised learning and transfer learning.
- **Efficient Architectures:** Designing lightweight and modular architectures that reduce memory and computational overhead while maintaining performance [100].
- **Energy Efficiency:** Investigating energy-efficient training methods to reduce the environmental impact of large-scale diffusion models [101].

5.3. Scalability to High-Resolution and Multimodal Tasks

While diffusion models have shown promise in high-resolution and multimodal applications, scaling to these tasks introduces unique challenges. For instance, generating ultra-high-resolution images or handling multiple modalities (e.g., text, image, and audio) simultaneously requires substantial computational power and advanced architectural designs. Potential solutions include:

- **Progressive Resolution Techniques:** Implementing hierarchical or progressive generation strategies to reduce computational costs for high-resolution tasks [102].
- **Unified Multimodal Models:** Developing models that can seamlessly integrate and process multiple data modalities, leveraging shared representations and cross-modal attention mechanisms [103].
- **Task-Specific Adaptations:** Tailoring diffusion models to specific tasks, optimizing their performance for targeted applications.

5.4. Robustness and Generalization

Ensuring that diffusion models are robust and generalize well across diverse datasets is critical for their real-world applicability. Current models can struggle with out-of-distribution data or adversarial perturbations. Research in this area could focus on:

- **Adversarial Robustness:** Enhancing the resilience of diffusion models to adversarial attacks through robust training techniques [104].
- **Domain Adaptation:** Improving the ability of models to generalize across domains with limited or no fine-tuning.
- **Uncertainty Quantification:** Incorporating mechanisms to quantify and manage uncertainty in generated outputs, particularly in high-stakes applications [105].

 doi:10.20944/preprints202502.0029.v1

5.5. Ethical and Societal Considerations

The widespread adoption of diffusion models raises important ethical concerns, including issues related to bias, misuse, and accountability [106]. Addressing these concerns is crucial for ensuring the responsible deployment of these models [107]. Key areas of focus include:

- **Bias Mitigation:** Identifying and mitigating biases in training datasets and model outputs to promote fairness and inclusivity.
- **Content Moderation:** Implementing safeguards to prevent the generation of harmful or malicious content, such as misinformation or explicit imagery.
- **Transparency and Accountability:** Enhancing the interpretability of diffusion models and establishing clear accountability frameworks for their use [108].

5.6. Theoretical Understanding and Interpretability

Despite their empirical success, the theoretical foundations of diffusion models are still being actively explored [109]. A deeper understanding of their behavior, limitations, and connections to other generative frameworks could lead to further improvements [110]. Future research could investigate:

- **Optimization Dynamics:** Analyzing the training and sampling dynamics of diffusion models to identify areas for improvement [111].
- **Connections to Other Frameworks:** Exploring the relationships between diffusion models and other generative approaches, such as energy-based models and normalizing flows.
- **Interpretability Techniques:** Developing tools and methods to interpret the decisions and outputs of diffusion models, particularly in critical applications [112].

5.7. Expanding Applications

While diffusion models have demonstrated success in many domains, there are still unexplored areas where they could have significant impact [113]. Expanding their applications requires interdisciplinary collaboration and task-specific adaptations [114]. Examples include:

- **Healthcare:** Advancing applications in medical imaging, drug discovery, and personalized medicine [115].
- **Education:** Enhancing educational tools through interactive content generation and multimodal learning aids [116].
- **Environmental Science:** Supporting climate modeling, ecological simulations, and sustainable development initiatives [117].

5.8. Future Outlook

The future of diffusion models lies in their ability to scale efficiently while maintaining versatility and robustness [118]. By addressing the challenges outlined above, diffusion models can continue to evolve as a cornerstone of generative modeling, driving innovation across diverse fields and applications. This section highlights the open challenges and future directions for scalable diffusion models [119]. By addressing these issues, researchers and practitioners can unlock the full potential of this transformative technology, paving the way for new applications and advancements in generative modeling [120].

6. Conclusion

Scalable diffusion models have emerged as a powerful class of generative models, demonstrating remarkable capabilities across a wide range of applications, from image synthesis to scientific discovery. Their unique approach to modeling data distributions through iterative noise and denoising processes has positioned them as a versatile and robust alternative to traditional generative frameworks, such as GANs and VAEs [121]. This survey has explored the core principles of diffusion models, the techniques enabling their scalability, and their transformative applications. Key innovations, including efficient noise schedules, improved neural architectures, and accelerated sampling methods, have significantly

 doi:10.20944/preprints202502.0029.v1

enhanced the performance and accessibility of these models [122]. Furthermore, their success in high-resolution, multimodal, and domain-specific tasks underscores their potential to address complex real-world challenges [123]. Despite their advancements, diffusion models face several challenges, including computational inefficiencies, resource-intensive training, and ethical considerations [124]. Addressing these challenges will require continued research into sampling efficiency, data-efficient training paradigms, and methods for ensuring robustness and fairness [125]. Additionally, expanding their theoretical understanding and exploring new applications will be critical for unlocking their full potential [126]. As the field progresses, diffusion models are poised to play a pivotal role in shaping the future of generative modeling [127]. Their scalability, adaptability, and ability to integrate with other frameworks make them a promising tool for innovation across diverse domains [128]. By addressing existing limitations and fostering interdisciplinary collaboration, diffusion models can continue to drive advancements in science, technology, and creative industries [129].

In conclusion, scalable diffusion models represent a significant milestone in the evolution of generative modeling. Their impact extends beyond technical achievements, offering new opportunities for solving pressing societal challenges and enabling creative exploration. As researchers and practitioners build on the foundations laid by this field, diffusion models are set to redefine the boundaries of what is possible in generative AI.

References

1. Han, Y.; Zhang, C.; Chen, X.; Yang, X.; Wang, Z.; Yu, G.; Fu, B.; Zhang, H. ChartLlama: A Multimodal LLM for Chart Understanding and Generation, 2023, [arXiv:cs.CV/2311.16483].
2. Yu, L.; Xiang, W. X-pruner: explainable pruning for vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24355–24363.
3. Lo, K.M.; Liang, Y.; Du, W.; Fan, Y.; Wang, Z.; Huang, W.; Ma, L.; Fu, J. m2mKD: Module-to-Module Knowledge Distillation for Modular Transformers. *arXiv preprint arXiv:2402.16918* **2024**.
4. Zhao, W.; Han, Y.; Tang, J.; Wang, K.; Song, Y.; Huang, G.; Wang, F.; You, Y. Dynamic Diffusion Transformer. *arXiv preprint arXiv:2410.03456* **2024**.
5. Liu, Y.; Yang, H.; Dong, Z.; Keutzer, K.; Du, L.; Zhang, S. Noisyquant: Noisy bias-enhanced post-training activation quantization for vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20321–20330.
6. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* **2023**.
7. Cha, J.; Kang, W.; Mun, J.; Roh, B. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742* **2023**.
8. Liang, Y.; Ge, C.; Tong, Z.; Song, Y.; Wang, J.; Xie, P. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800* **2022**.
9. Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; Rohrbach, M. Towards vqa models that can read. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8317–8326.
10. Marin, D.; Chang, J.H.R.; Ranjan, A.; Prabhu, A.; Rastegari, M.; Tuzel, O. Token pooling in vision transformers. *arXiv preprint arXiv:2110.03860* **2021**.
11. Gao, P.; Zhang, R.; Liu, C.; Qiu, L.; Huang, S.; Lin, W.; Zhao, S.; Geng, S.; Lin, Z.; Jin, P.; et al. SPHINX-X: Scaling Data and Parameters for a Family of Multi-modal Large Language Models. *arXiv preprint arXiv:2402.05935* **2024**.
12. Azizi, S.; Mustafa, B.; Ryan, F.; Beaver, Z.; Freyberg, J.; Deaton, J.; Loh, A.; Karthikesalingam, A.; Kornblith, S.; Chen, T.; et al. Big self-supervised models advance medical image classification. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 3478–3488.
13. Zhai, X.; Mustafa, B.; Kolesnikov, A.; Beyer, L. Sigmoid loss for language image pre-training. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 11975–11986.
14. Li, J.; Li, D.; Savarese, S.; Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In Proceedings of the International conference on machine learning. PMLR, 2023, pp. 19730–19742.

 doi:10.20944/preprints202502.0029.v1

15. Meng, C.; He, Y.; Song, Y.; Song, J.; Wu, J.; Zhu, J.Y.; Ermon, S. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In Proceedings of the International Conference on Learning Representations, 2022.
16. Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Luo, Z.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; et al. Generative multimodal models are in-context learners. *arXiv preprint arXiv:2312.13286* **2023**.
17. Sauer, A.; Boesel, F.; Dockhorn, T.; Blattmann, A.; Esser, P.; Rombach, R. Fast high-resolution image synthesis with latent adversarial diffusion distillation. *arXiv preprint arXiv:2403.12015* **2024**.
18. Li, Y.; Yuan, G.; Wen, Y.; Hu, J.; Evangelidis, G.; Tulyakov, S.; Wang, Y.; Ren, J. Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems* **2022**, *35*, 12934–12949.
19. Song, Y.; Ermon, S. Generative modeling by estimating gradients of the data distribution. In Proceedings of the Proceedings of the 33rd International Conference on Neural Information Processing Systems, 2019, pp. 11918–11930.
20. Watson, D.; Chan, W.; Ho, J.; Norouzi, M. Learning fast samplers for diffusion models by differentiating through sample quality. In Proceedings of the International Conference on Learning Representations, 2022.
21. Chavan, A.; Shen, Z.; Liu, Z.; Liu, Z.; Cheng, K.T.; Xing, E.P. Vision transformer slimming: Multi-dimension searching in continuous optimization space. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 4931–4941.
22. Zhou, B.; Hu, Y.; Weng, X.; Jia, J.; Luo, J.; Liu, X.; Wu, J.; Huang, L. TinyLLaVA: A Framework of Small-scale Large Multimodal Models. *arXiv preprint arXiv:2402.14289* **2024**.
23. Liu, L.; Ren, Y.; Lin, Z.; Zhao, Z. Pseudo Numerical Methods for Diffusion Models on Manifolds. In Proceedings of the International Conference on Learning Representations, 2022.
24. Schwenk, D.; Khandelwal, A.; Clark, C.; Marino, K.; Mottaghi, R. A-okvqa: A benchmark for visual question answering using world knowledge. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 146–162.
25. Luhman, E.; Luhman, T. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388* **2021**.
26. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. In Proceedings of the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.
27. Chung, H.; Sim, B.; Ye, J.C. Come-closer-diffuse-faster: Accelerating conditional diffusion models for inverse problems through stochastic contraction. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12413–12422.
28. Zhu, Y.; Zhu, M.; Liu, N.; Ou, Z.; Mou, X.; Tang, J. LLaVA-phi: Efficient Multi-Modal Assistant with Small Language Model. *arXiv preprint arXiv:2401.02330* **2024**.
29. Zhang, Q.; Chen, Y. Fast Sampling of Diffusion Models with Exponential Integrator. In Proceedings of the International Conference on Learning Representations, 2023.
30. Dai, W.; Li, J.; Li, D.; Tiong, A.M.H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.N.; Hoi, S. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in Neural Information Processing Systems* **2024**, *36*.
31. Liu, X.; Gong, C.; et al. Flow Straight and Fast: Learning to Generate and Transfer Data with Rectified Flow. In Proceedings of the The Eleventh International Conference on Learning Representations, 2022.
32. Yin, T.; Gharbi, M.; Zhang, R.; Shechtman, E.; Durand, F.; Freeman, W.T.; Park, T. One-step diffusion with distribution matching distillation. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 6613–6623.
33. Lepikhin, D.; Lee, H.; Xu, Y.; Chen, D.; Firat, O.; Huang, Y.; Krikun, M.; Shazeer, N.; Chen, Z. Gshard: Scaling giant models with conditional computation and automatic sharding. *arXiv preprint arXiv:2006.16668* **2020**.
34. Pan, Z.; Zhuang, B.; Huang, D.A.; Nie, W.; Yu, Z.; Xiao, C.; Cai, J.; Anandkumar, A. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching. *arXiv preprint arXiv:2402.14167* **2024**.
35. Zhao, Y.; Xu, Y.; Xiao, Z.; Jia, H.; Hou, T. MobileDiffusion: Instant Text-to-Image Generation on Mobile Devices, 2024, [arXiv:cs.CV/2311.16567].
36. Tu, T.; Azizi, S.; Driess, D.; Schaekermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards generalist biomedical ai. *NEJM AI* **2024**, *1*, AIoa2300138.
37. Yu, S.; Chen, T.; Shen, J.; Yuan, H.; Tan, J.; Yang, S.; Liu, J.; Wang, Z. Unified visual transformer compression. *arXiv preprint arXiv:2203.08243* **2022**.
38. Liu, Z.; Wang, Y.; Han, K.; Zhang, W.; Ma, S.; Gao, W. Post-training quantization for vision transformer. *Advances in Neural Information Processing Systems* **2021**, *34*, 28092–28103.

 doi:10.20944/preprints202502.0029.v1

39. Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; Zhu, J. DPM-solver: a fast ODE solver for diffusion probabilistic model sampling in around 10 steps. In Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 5775–5787.
40. Chen, Z.; Ma, X.; Fang, G.; Tan, Z.; Wang, X. AsyncDiff: Parallelizing Diffusion Models by Asynchronous Denoising, 2024, [arXiv:cs.CV/2406.06911].
41. Valipour, M.; Rezagholizadeh, M.; Kobyzev, I.; Ghodsi, A. Dylora: Parameter efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. *arXiv preprint arXiv:2210.07558* **2022**.
42. Shi, B.; Wu, Z.; Mao, M.; Wang, X.; Darrell, T. When Do We Not Need Larger Vision Models? *arXiv preprint arXiv:2403.13043* **2024**.
43. Chen, J.; Yu, Q.; Shen, X.; Yuille, A.; Chen, L.C. ViTamin: Designing Scalable Vision Models in the Vision-Language Era, 2024, [arXiv:cs.CV/2404.02132].
44. Dong, X.; Zhang, P.; Zang, Y.; Cao, Y.; Wang, B.; Ouyang, L.; Zhang, S.; Duan, H.; Zhang, W.; Li, Y.; et al. InternLM-XComposer2-4KHD: A Pioneering Large Vision-Language Model Handling Resolutions from 336 Pixels to 4K HD. *arXiv preprint arXiv:2404.06512* **2024**.
45. Zhu, W.; Hessel, J.; Awadalla, A.; Gadre, S.Y.; Dodge, J.; Fang, A.; Yu, Y.; Schmidt, L.; Wang, W.Y.; Choi, Y. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems* **2024**, *36*.
46. Wu, Q.; Liu, Y.; Zhao, H.; Kale, A.; Bui, T.; Yu, T.; Lin, Z.; Zhang, Y.; Chang, S. Uncovering the disentanglement capability in text-to-image diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 1900–1910.
47. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the International Conference on Machine Learning, 2024.
48. Chen, X.; Fang, H.; Lin, T.Y.; Vedantam, R.; Gupta, S.; Dollár, P.; Zitnick, C.L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* **2015**.
49. Luo, S.; Tan, Y.; Patil, S.; Gu, D.; von Platen, P.; Passos, A.; Huang, L.; Li, J.; Zhao, H. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv preprint arXiv:2311.05556* **2023**.
50. Chen, T.; Cheng, Y.; Gan, Z.; Yuan, L.; Zhang, L.; Wang, Z. Chasing sparsity in vision transformers: An end-to-end exploration. *Advances in Neural Information Processing Systems* **2021**, *34*, 19974–19988.
51. Yuan, Z.; Xue, C.; Chen, Y.; Wu, Q.; Sun, G. Ptq4vit: Post-training quantization for vision transformers with twin uniform quantization. In Proceedings of the European conference on computer vision. Springer, 2022, pp. 191–207.
52. Chen, M.; Peng, H.; Fu, J.; Ling, H. Autoformer: Searching transformers for visual recognition. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 12270–12280.
53. Shang, Y.; Yuan, Z.; Xie, B.; Wu, B.; Yan, Y. Post-training quantization on diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 1972–1981.
54. Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122* **2023**.
55. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
56. Saleh, B.; Elgammal, A. Large-scale classification of fine-art paintings: Learning the right metric on the right feature. *arXiv preprint arXiv:1505.00855* **2015**.
57. Brock, A.; Donahue, J.; Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In Proceedings of the International Conference on Learning Representations, 2019.
58. Fan, Y.; Lee, K. Optimizing DDPM Sampling with Shortcut Fine-Tuning. In Proceedings of the International Conference on Machine Learning, 2023, pp. 9623–9639.
59. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
60. Chen, Y.H.; Sarokin, R.; Lee, J.; Tang, J.; Chang, C.L.; Kulik, A.; Grundmann, M. Speed is all you need: On-device acceleration of large diffusion models via gpu-aware optimizations. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4651–4655.

 doi:10.20944/preprints202502.0029.v1

61. Changpinyo, S.; Sharma, P.; Ding, N.; Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3558–3568.
62. Xu, C.; Sun, Q.; Zheng, K.; Geng, X.; Zhao, P.; Feng, J.; Tao, C.; Jiang, D. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244* **2023**.
63. Ye, Q.; Xu, H.; Ye, J.; Yan, M.; Hu, A.; Liu, H.; Qian, Q.; Zhang, J.; Huang, F.; Zhou, J. mPLUG-Owl2: Revolutionizing Multi-modal Large Language Model with Modality Collaboration, 2023, [arXiv:cs.CL/2311.04257].
64. Shang, Y.; Cai, M.; Xu, B.; Lee, Y.J.; Yan, Y. LLaVA-PruMerge: Adaptive Token Reduction for Efficient Large Multimodal Models, 2024, [arXiv:cs.CV/2403.15388].
65. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PMLR, 2021, pp. 8748–8763.
66. Wang, J.; Fang, J.; Li, A.; Yang, P. PipeFusion: Displaced Patch Pipeline Parallelism for Inference of Diffusion Transformer Models, 2024, [arXiv:cs.CV/2405.14430].
67. Wang, G.; Liu, J.; Li, C.; Ma, J.; Zhang, Y.; Wei, X.; Zhang, K.; Chong, M.; Zhang, R.; Liu, Y.; et al. Cloud-Device Collaborative Learning for Multimodal Large Language Models. *arXiv preprint arXiv:2312.16279* **2023**.
68. Gupta, A.; Gu, A.; Berant, J. Diagonal state spaces are as effective as structured state spaces. *Advances in Neural Information Processing Systems* **2022**, *35*, 22982–22994.
69. Huang, L.; Wu, S.; Cui, Y.; Xiong, Y.; Liu, X.; Kuo, T.W.; Guan, N.; Xue, C.J. RAEE: A Training-Free Retrieval-Augmented Early Exiting Framework for Efficient Inference. *arXiv preprint arXiv:2405.15198* **2024**.
70. Li, K.; He, Y.; Wang, Y.; Li, Y.; Wang, W.; Luo, P.; Wang, Y.; Wang, L.; Qiao, Y. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355* **2023**.
71. Papa, L.; Russo, P.; Amerini, I.; Zhou, L. A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, p. 1–20. https://doi.org/10.1109/tpami.2024.3392941.
72. Guo, Y.; Yang, C.; Rao, A.; Liang, Z.; Wang, Y.; Qiao, Y.; Agrawala, M.; Lin, D.; Dai, B. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In Proceedings of the International Conference on Learning Representations, 2024.
73. Liu, X.; Zhang, X.; Ma, J.; Peng, J.; et al. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
74. Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems* **2024**, *36*.
75. Heo, B.; Yun, S.; Han, D.; Chun, S.; Choe, J.; Oh, S.J. Rethinking spatial dimensions of vision transformers. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 11936–11945.
76. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. In Proceedings of the International Conference on Machine Learning, 2023, pp. 32211–32252.
77. Zhang, J.; Peng, H.; Wu, K.; Liu, M.; Xiao, B.; Fu, J.; Yuan, L. Minivit: Compressing vision transformers with weight multiplexing. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12145–12154.
78. Liu, Z.; Mao, H.; Wu, C.Y.; Feichtenhofer, C.; Darrell, T.; Xie, S. A convnet for the 2020s. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 11976–11986.
79. Zhang, L.; Hu, A.; Xu, H.; Yan, M.; Xu, Y.; Jin, Q.; Zhang, J.; Huang, F. TinyChart: Efficient Chart Understanding with Visual Token Merging and Program-of-Thoughts Learning. *arXiv preprint arXiv:2404.16635* **2024**.
80. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.
81. He, Y.; Liu, L.; Liu, J.; Wu, W.; Zhou, H.; Zhuang, B. PTQD: accurate post-training quantization for diffusion models. In Proceedings of the Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, pp. 13237–13249.

 doi:10.20944/preprints202502.0029.v1

82. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 6904–6913.
83. Gurari, D.; Li, Q.; Stangl, A.J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; Bigham, J.P. Vizwiz grand challenge: Answering visual questions from blind people. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3608–3617.
84. Lu, H.; Liu, W.; Zhang, B.; Wang, B.; Dong, K.; Liu, B.; Sun, J.; Ren, T.; Li, Z.; Yang, H.; et al. DeepSeek-VL: Towards Real-World Vision-Language Understanding, 2024, [arXiv:cs.AI/2403.05525].
85. Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281* **2023**.
86. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the Proceedings of the IEEE/cvf conference on computer vision and pattern recognition, 2019, pp. 3195–3204.
87. Xu, R.; Yao, Y.; Guo, Z.; Cui, J.; Ni, Z.; Ge, C.; Chua, T.S.; Liu, Z.; Sun, M.; Huang, G. LLaVA-UHD: an LMM Perceiving Any Aspect Ratio and High-Resolution Images, 2024, [arXiv:cs.CV/2403.11703].
88. Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.L.; Murphy, K. Generation and comprehension of unambiguous object descriptions. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 11–20.
89. Kar, O.F.; Tonioni, A.; Poklukar, P.; Kulshrestha, A.; Zamir, A.; Tombari, F. BRAVE: Broadening the visual encoding of vision-language models. *arXiv preprint arXiv:2404.07204* **2024**.
90. Luo, S.; Tan, Y.; Huang, L.; Li, J.; Zhao, H. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378* **2023**.
91. Zhang, P.; Zeng, G.; Wang, T.; Lu, W. TinyLlama: An Open-Source Small Language Model, 2024, [arXiv:cs.CL/2401.02385].
92. Yuan, Z.; Li, Z.; Sun, L. Tinygpt-v: Efficient multimodal large language model via small backbones. *arXiv preprint arXiv:2312.16862* **2023**.
93. Lin, Z.; Lin, M.; Lin, L.; Ji, R. Boosting Multimodal Large Language Models with Visual Tokens Withdrawal for Rapid Inference, 2024, [arXiv:cs.CV/2405.05803].
94. Choromanski, K.; Likhosherstov, V.; Dohan, D.; Song, X.; Gane, A.; Sarlos, T.; Hawkins, P.; Davis, J.; Belanger, D.; Colwell, L.; et al. Masked language modeling for proteins via linearly scalable long-context transformers. *arXiv preprint arXiv:2006.03555* **2020**.
95. Meng, C.; Rombach, R.; Gao, R.; Kingma, D.; Ermon, S.; Ho, J.; Salimans, T. On distillation of guided diffusion models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 14297–14306.
96. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.
97. Kazemzadeh, S.; Ordonez, V.; Matten, M.; Berg, T. Referitgame: Referring to objects in photographs of natural scenes. In Proceedings of the Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 787–798.
98. Qiao, Y.; Yu, Z.; Guo, L.; Chen, S.; Zhao, Z.; Sun, M.; Wu, Q.; Liu, J. VL-Mamba: Exploring State Space Models for Multimodal Learning. *arXiv preprint arXiv:2403.13600* **2024**.
99. Jie, S.; Tang, Y.; Ding, N.; Deng, Z.H.; Han, K.; Wang, Y. Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning, 2024, [arXiv:cs.CV/2405.05615].
100. Liu, H.; Tam, D.; Muqeeth, M.; Mohta, J.; Huang, T.; Bansal, M.; Raffel, C. Few-Shot Parameter-Efficient Fine-Tuning is Better and Cheaper than In-Context Learning. *arXiv preprint arXiv:2205.05638* **2022**.
101. Lin, C.; Peng, B.; Li, Z.; Tan, W.; Ren, Y.; Xiao, J.; Pu, S. Bit-shrinking: Limiting instantaneous sharpness for improving post-training quantization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 16196–16205.
102. Ren, S.; Gao, Z.; Hua, T.; Xue, Z.; Tian, Y.; He, S.; Zhao, H. Co-advise: Cross inductive bias distillation. In Proceedings of the Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, 2022, pp. 16773–16782.
103. Li, Y.; Bubeck, S.; Eldan, R.; Giorno, A.D.; Gunasekar, S.; Lee, Y.T. Textbooks Are All You Need II: phi-1.5 technical report, 2023, [arXiv:cs.CL/2309.05463].
104. Zhao, B.; Wu, B.; Huang, T. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087* **2023**.

 doi:10.20944/preprints202502.0029.v1

105. Li, Y.; Zhang, Y.; Wang, C.; Zhong, Z.; Chen, Y.; Chu, R.; Liu, S.; Jia, J. Mini-Gemini: Mining the Potential of Multi-modality Vision Language Models. *arXiv preprint arXiv:2403.18814* **2024**.
106. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.
107. Xue, S.; Liu, Z.; Chen, F.; Zhang, S.; Hu, T.; Xie, E.; Li, Z. Accelerating Diffusion Sampling with Optimized Time Steps. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8292–8301.
108. Yu, C.; Chen, T.; Gan, Z.; Fan, J. Boost vision transformer with gpu-friendly sparsity and quantization. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22658–22668.
109. Zhao, W.; Bai, L.; Rao, Y.; Zhou, J.; Lu, J. UniPC: a unified predictor-corrector framework for fast sampling of diffusion models. In Proceedings of the Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, pp. 49842–49869.
110. ShareGPT. https://sharegpt.com/, 2023.
111. Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; Chang, B. An Image is Worth 1/2 Tokens After Layer 2: Plug-and-Play Inference Acceleration for Large Vision-Language Models, 2024, [arXiv:cs.CV/2403.06764].
112. Wang, H.; Wang, Y.; Ye, Y.; Nie, Y.; Huang, C. Elysium: Exploring Object-level Perception in Videos via MLLM, 2024, [arXiv:cs.CV/2403.16558].
113. Zheng, H.; Nie, W.; Vahdat, A.; Azizzadenesheli, K.; Anandkumar, A. Fast sampling of diffusion models via operator learning. In Proceedings of the International conference on machine learning, 2023, pp. 42390–42402.
114. LAION. Gpt-4v dataset. https://huggingface.co/datasets/laion/gpt4v-dataset, 2023.
115. Wang, W.; Chen, W.; Qiu, Q.; Chen, L.; Wu, B.; Lin, B.; He, X.; Liu, W. Crossformer++: A versatile vision transformer hinging on cross-scale attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2023**.
116. Abdin, M.; Jacobs, S.A.; Awan, A.A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Behl, H.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, 2024, [arXiv:cs.CL/2404.14219].
117. Li, Z.; Sun, M.; Lu, A.; Ma, H.; Yuan, G.; Xie, Y.; Tang, H.; Li, Y.; Leeser, M.; Wang, Z.; et al. Auto-vit-acc: An fpga-aware automatic acceleration framework for vision transformer with mixed-scheme quantization. In Proceedings of the 2022 32nd International Conference on Field-Programmable Logic and Applications (FPL). IEEE, 2022, pp. 109–116.
118. Salimans, T.; Ho, J. Progressive Distillation for Fast Sampling of Diffusion Models. In Proceedings of the International Conference on Learning Representations, 2022.
119. Hu, Z.; Lan, Y.; Wang, L.; Xu, W.; Lim, E.P.; Lee, R.K.W.; Bing, L.; Poria, S. LLM-Adapters: An Adapter Family for Parameter-Efficient Fine-Tuning of Large Language Models. *arXiv preprint arXiv:2304.01933* **2023**.
120. Fayyaz, M.; Koohpayegani, S.A.; Jafari, F.R.; Sengupta, S.; Joze, H.R.V.; Sommerlade, E.; Pirsiavash, H.; Gall, J. Adaptive token sampling for efficient vision transformers. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 396–414.
121. Ye, Q.; Xu, H.; Xu, G.; Ye, J.; Yan, M.; Zhou, Y.; Wang, J.; Hu, A.; Shi, P.; Shi, Y.; et al. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality, 2023, [arXiv:cs.CL/2304.14178].
122. Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125* **2023**.
123. Driess, D.; Xia, F.; Sajjadi, M.S.M.; Lynch, C.; Chowdhery, A.; Ichter, B.; Wahid, A.; Tompson, J.; Vuong, Q.; Yu, T.; et al. PALM-E: An Embodied Multimodal Language Model. *arXiv preprint arXiv:2303.03378* **2023**.
124. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* **2023**.
125. Lv, K.; Yang, Y.; Liu, T.; Gao, Q.; Guo, Q.; Qiu, X. Full Parameter Fine-tuning for Large Language Models with Limited Resources. *arXiv preprint arXiv:2306.09782* **2023**.
126. Lyu, Z.; Xu, X.; Yang, C.; Lin, D.; Dai, B. Accelerating diffusion models via early stop of the diffusion process. *arXiv preprint arXiv:2205.12524* **2022**.
127. Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; Lin, D. Sharegpt4v: Improving large multi-modal models with better captions. *arXiv preprint arXiv:2311.12793* **2023**.

 doi:10.20944/preprints202502.0029.v1

128. Dockhorn, T.; Vahdat, A.; Kreis, K. GENIE: higher-order denoising diffusion solvers. In Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 30150–30166.
129. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. Springer, 2014, pp. 740–755.