Review

Diffusion Models in Generative AI: Principles, Applications, and Future Directions

Kholoud Hadiyya , Rasim Dina * , Burhanuddin Mokhtar

Posted Date: 7 February 2025

doi: 10.20944/preprints202502.0524.v1

Keywords: Diffusion Models; Generative AI; Probabilistic Modeling; Image Synthesis; Text-to-Image Generation; Audio Synthesis; Molecular Design; Efficient Sampling; Multimodal Generation; Ethical AI; Scalable Architectures

Review

Diffusion Models in Generative AI: Principles, Applications, and Future Directions

Kholoud Hadiyya, Rasim Dina * and Burhanuddin Mokhtar

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

Abstract: Diffusion models have emerged as a powerful class of generative models, offering state-of-the-art performance across a wide range of applications in artificial intelligence. Rooted in probabilistic principles, these models generate data by iteratively refining random noise through a reverse diffusion process, enabling the synthesis of high-quality outputs in domains such as image generation, text-to-image translation, audio synthesis, and molecular design. Compared to earlier generative frameworks like GANs and VAEs, diffusion models excel in stability, diversity, and fidelity, while also supporting flexible conditioning mechanisms for multimodal and domain-specific tasks. Despite their success, diffusion models face several challenges, including high computational costs, scalability to high-dimensional data, and limited interpretability. Additionally, their deployment raises ethical concerns regarding potential misuse, bias, and societal impact. Recent advancements in efficient sampling techniques, hierarchical architectures, and theoretical insights aim to address these limitations, paving the way for broader adoption and impact. This paper provides a comprehensive overview of diffusion models, exploring their underlying principles, key applications, and current limitations. We also highlight future research directions, including the development of scalable and efficient frameworks, integration with emerging technologies, and ethical considerations for responsible deployment. By addressing these challenges, diffusion models have the potential to revolutionize generative AI, driving innovation across diverse fields and contributing to the advancement of artificial intelligence.

Keywords: diffusion models; generative AI; probabilistic modeling; image synthesis; text-to-image generation; audio synthesis; molecular design; efficient sampling; multimodal generation; ethical AI; scalable architectures

1. Introduction

Generative Artificial Intelligence (GenAI) has rapidly evolved into one of the most influential domains in machine learning, enabling the creation of synthetic data that closely resembles real-world distributions. The ability to generate realistic images, coherent text, lifelike audio, and even intricate 3D structures has unlocked a plethora of applications across industries such as healthcare, entertainment, design, and scientific discovery [1]. Central to this transformative capability are generative models, which aim to learn the underlying patterns of data distributions and use this knowledge to produce novel samples [2]. Among these, diffusion models have recently emerged as a powerful and versatile framework for generative tasks, offering significant advantages in quality, stability, and theoretical interpretability. Diffusion models, originally inspired by the mathematical principles of stochastic processes and non-equilibrium thermodynamics, operate through a two-phase mechanism: a forward diffusion process and a reverse generative process. The forward process incrementally corrupts data by adding noise, effectively mapping complex data distributions to a simple prior, such as a Gaussian distribution [3]. The reverse process, which lies at the heart of the model's generative capability, learns to iteratively denoise the corrupted data, reconstructing samples that resemble the original data distribution [4]. This bidirectional framework provides a principled and robust approach to generative modeling, distinguishing diffusion models from earlier paradigms like Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs). Compared to

 doi:10.20944/preprints202502.0524.v1

GANs, which are known for their adversarial training dynamics and susceptibility to issues like mode collapse, diffusion models offer a more stable training procedure. Unlike VAEs, which often struggle to produce high-quality samples due to their reliance on approximate posterior distributions, diffusion models excel at generating samples with high fidelity and diversity. Furthermore, their reliance on explicit likelihood-based training aligns with well-established probabilistic principles, providing a solid theoretical foundation for further advancements [5]. Recent developments in diffusion models have significantly expanded their applicability and performance. Techniques such as score-based generative modeling, improved noise schedules, and hybrid architectures have pushed the boundaries of what these models can achieve. For instance, diffusion models have been employed to generate photorealistic images, synthesize novel molecules for drug discovery, design creative artworks, and even produce high-quality audio for speech and music [6]. These successes underscore the versatility and scalability of diffusion models, making them a cornerstone of modern GenAI research [7]. Despite their impressive capabilities, diffusion models are not without challenges. The computational cost of their iterative denoising process can be substantial, especially when compared to the single-step generation of GANs. Efforts to address this issue, such as optimizing the number of denoising steps or employing efficient sampling techniques, are active areas of research. Additionally, the theoretical understanding of their convergence properties and the design of optimal noise schedules remain open questions that continue to drive innovation in the field [8]. In this paper, we provide a comprehensive exploration of diffusion models in the context of generative AI [9]. We begin by reviewing the theoretical foundations of diffusion processes and their adaptation to generative modeling. We then discuss recent advancements in model architecture, training strategies, and applications. Finally, we address the challenges and future directions for diffusion models, highlighting their potential to shape the next generation of generative technologies [10]. By synthesizing insights from both foundational research and practical implementations, we aim to offer a holistic perspective on the transformative impact of diffusion models in the realm of GenAI [11].

2. Background and Theoretical Foundations

Diffusion models draw their inspiration from concepts in stochastic processes, specifically diffusion processes, which describe the probabilistic evolution of systems over time. To understand their role in generative modeling, it is essential to first examine the mathematical principles underlying these processes and how they are adapted to learn complex data distributions [12].

2.1. Stochastic Processes and Diffusion

At the core of diffusion models lies the concept of a stochastic process, which represents a sequence of random variables evolving over time [13]. A diffusion process is a specific type of stochastic process that describes the continuous-time dynamics of a system undergoing random perturbations. These processes are governed by stochastic differential equations (SDEs), which can be expressed as:

$$dx = f(x,t)\,dt + g(t)\,dw,$$

where x represents the system state, $f(x,t)$ is the drift coefficient describing deterministic changes, $g(t)$ is the diffusion coefficient governing random noise, and dw is the Wiener process (standard Brownian motion) [14]. Diffusion processes are widely used in physics, finance, and biology to model systems that exhibit random behavior. In the context of generative modeling, the forward diffusion process incrementally adds noise to the data, gradually transforming it into a simple prior distribution, such as a Gaussian [15]. This process can be viewed as solving an SDE with a specific noise schedule [16]. The reverse process, which is the focus of training in diffusion models, learns to invert this transformation, denoising the data step by step to recover samples from the original data distribution [17].

 doi:10.20944/preprints202502.0524.v1

2.2. The Forward and Reverse Processes

The forward diffusion process is defined as a sequence of noisy data distributions $\{q(x_t)\}_{t=0}^{T}$, where t denotes the time step, and x_0 represents the original data. Noise is added iteratively, leading to a progressive corruption of the data:

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

where β_t is the variance schedule controlling the amount of noise added at each step [18]. The reverse process aims to model the reverse-time dynamics, denoising the data from x_T back to x_0. This is achieved by learning a parameterized distribution $p_\theta(x_{t-1}|x_t)$, which is trained to approximate the true reverse conditional distribution $q(x_{t-1}|x_t)$ [19]. The reverse process is expressed as:

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

where μ_θ and Σ_θ are the mean and variance predicted by the model [20].

2.3. Training Objectives

The training of diffusion models involves minimizing the divergence between the forward and reverse processes [21]. A common approach is to use the variational lower bound (VLB) on the negative log-likelihood of the data. The objective can be expressed as:

$$L = \mathbb{E}_q \left[\sum_{t=1}^{T} D_{\mathrm{KL}}(q(x_{t-1}|x_t, x_0) \parallel p_\theta(x_{t-1}|x_t)) \right],$$

where D_{KL} is the Kullback-Leibler divergence [22]. This objective ensures that the learned reverse process closely approximates the true reverse dynamics. In practice, the training process can be simplified by parameterizing the noise prediction directly and optimizing the model to predict the noise added at each step. This reparameterization has been shown to improve training efficiency and stability [23].

2.4. Connections to Other Generative Models

Diffusion models share conceptual similarities with other generative frameworks but differ in their approach and advantages [24]. For instance, while Variational Autoencoders (VAEs) also rely on explicit likelihood-based training, their use of approximate posterior distributions often leads to suboptimal sample quality. On the other hand, Generative Adversarial Networks (GANs) produce high-quality samples but suffer from adversarial training dynamics, such as mode collapse and instability [25]. Diffusion models strike a balance by offering stable training, high sample diversity, and a solid probabilistic foundation [26]. This section has outlined the theoretical principles that underpin diffusion models. The following sections will delve into architectural innovations, practical implementation details, and applications that have propelled diffusion models to the forefront of generative AI research.

3. Architectural Innovations and Advancements

The success of diffusion models in generative tasks can be attributed not only to their robust theoretical foundation but also to a series of architectural innovations that have improved their efficiency, scalability, and performance [27]. This section explores these advancements, focusing on techniques that address the computational challenges and enhance the quality of generated samples.

3.1. Score-Based Generative Modeling

A significant breakthrough in diffusion models came with the introduction of score-based generative modeling, which leverages the score function of the data distribution [28]. The score function,

 doi:10.20944/preprints202502.0524.v1

defined as the gradient of the log probability density $\nabla_x \log q(x)$, provides a measure of how the data density changes in the vicinity of a point x. By estimating the score function at various noise levels, score-based models enable the generation of high-quality samples through iterative denoising [29]. The connection between score-based modeling and diffusion models lies in the equivalence between score matching and reverse-time SDEs [30]. This insight has led to the development of algorithms that directly estimate the score function using neural networks, bypassing the need for explicit likelihood computation and enabling more flexible architectures [31].

3.2. Noise Scheduling and Variance Control

The choice of noise schedule, which determines the amount of noise added at each step of the forward process, plays a crucial role in the performance of diffusion models [32]. Early implementations used simple linear schedules for the noise variance β_t, but recent work has shown that carefully designed schedules can significantly improve sample quality and training stability. Adaptive noise schedules, which dynamically adjust β_t based on the properties of the data and model, have been proposed to strike a balance between effective denoising and computational efficiency [33]. These schedules ensure that the model focuses on challenging regions of the data distribution, leading to more accurate reconstructions.

3.3. Conditional Diffusion Models

Conditional diffusion models extend the framework to allow for controlled generation based on auxiliary information, such as class labels, textual descriptions, or other modalities [34]. This is achieved by incorporating the conditioning signal into the model's architecture, typically through concatenation, attention mechanisms, or cross-modal embeddings [35]. For example, text-to-image generation tasks use conditional diffusion models to synthesize images based on textual prompts [36]. Techniques like classifier-free guidance further enhance the quality of generated samples by modulating the influence of the conditioning signal during the reverse process [37]. This approach has been widely adopted in applications such as DALL·E and Stable Diffusion.

3.4. Efficient Sampling Techniques

One of the primary challenges of diffusion models is the computational cost associated with their iterative denoising process [38]. Generating a single sample often requires hundreds or thousands of denoising steps, making diffusion models slower compared to alternatives like GANs [39]. To address this, several efficient sampling techniques have been developed [40]. Methods such as importance sampling, accelerated reverse processes, and stochastic sampling reduce the number of required steps while maintaining sample quality [41]. Additionally, hybrid approaches that combine diffusion models with other generative frameworks, such as GANs, have been explored to leverage the strengths of both paradigms.

3.5. Hybrid Architectures and Multimodal Extensions

Recent advancements have also focused on hybrid architectures that combine diffusion models with other deep learning frameworks to extend their capabilities. For instance, integrating attention mechanisms and transformer-based architectures has enabled diffusion models to handle multimodal data, such as text, images, and audio, in a unified framework. Multimodal extensions of diffusion models have been particularly impactful in applications like text-to-image generation, where models learn to align textual descriptions with visual features [42]. These architectures often incorporate cross-attention layers, enabling the model to generate coherent and semantically meaningful outputs across modalities [43].

3.6. Scalability and Parallelization

The scalability of diffusion models has been a key factor in their adoption for large-scale generative tasks [44]. Advances in model parallelism, distributed training, and memory-efficient architectures

 doi:10.20944/preprints202502.0524.v1

have enabled the training of diffusion models on massive datasets [45]. Techniques like gradient checkpointing and mixed-precision training further reduce the computational overhead, making these models accessible to a broader range of researchers and practitioners [46].

3.7. Evaluation Metrics and Improvements

The evaluation of generative models has traditionally relied on metrics like Inception Score (IS) and Fréchet Inception Distance (FID) for image generation tasks. Diffusion models have achieved state-of-the-art performance on these benchmarks, often surpassing GANs and VAEs in terms of both quality and diversity [47]. Recent work has also proposed new evaluation metrics tailored to diffusion models, focusing on aspects like sample coherence, conditional fidelity, and computational efficiency [48].

3.8. Applications of Architectural Innovations

The architectural advancements discussed in this section have enabled diffusion models to excel in a wide range of applications [49]. From generating photorealistic images and synthesizing novel molecules to creating high-quality audio and designing 3D structures, these innovations have expanded the scope of what diffusion models can achieve [50]. They have also paved the way for real-world deployments in fields such as content creation, drug discovery, and virtual reality [51]. In the following section, we will explore the practical applications of diffusion models in greater detail, highlighting their transformative impact across various domains [52].

4. Applications of Diffusion Models

The versatility and generative power of diffusion models have made them a cornerstone of modern artificial intelligence, enabling breakthroughs across multiple domains. This section highlights the diverse applications of diffusion models, ranging from creative content generation to scientific research and industrial solutions [53].

4.1. Image Generation and Editing

One of the most prominent applications of diffusion models is in image synthesis [54]. These models have demonstrated the ability to generate photorealistic images with fine-grained details, often surpassing the quality and diversity achieved by other generative frameworks like GANs [55]. Notable examples include models such as DALL·E, Imagen, and Stable Diffusion, which can produce high-resolution images based on textual descriptions [56]. Diffusion models are also widely used in image editing tasks, such as inpainting, super-resolution, and style transfer [57]. By leveraging the reverse diffusion process, these models can seamlessly reconstruct missing regions, enhance image resolution, or apply artistic styles while preserving the original content [58]. Their ability to condition generation on specific features makes them highly effective for personalized and context-aware editing [59].

4.2. Text-to-Image and Multimodal Generation

Text-to-image generation has emerged as a flagship application of diffusion models, enabling the creation of images that align with textual prompts [60]. This capability is powered by conditional diffusion models, which incorporate natural language embeddings into the generative process [61]. Models like CLIP-guided diffusion leverage pre-trained language-image representations to ensure semantic coherence between text and generated images [62]. Beyond text-to-image tasks, diffusion models have been extended to multimodal generation, where multiple data types, such as text, images, and audio, are synthesized in a unified framework [63]. These advancements have facilitated the development of tools for storytelling, virtual reality, and interactive design, where multimodal content plays a crucial role [64].

 doi:10.20944/preprints202502.0524.v1

4.3. Audio and Speech Synthesis

Diffusion models have shown remarkable potential in audio and speech synthesis, generating high-quality waveforms for applications such as text-to-speech (TTS), music generation, and sound design [65]. By modeling the temporal structure of audio signals, diffusion-based methods can produce realistic and coherent outputs [66]. For example, models like WaveGrad and DiffWave utilize diffusion processes to generate speech with natural prosody and clarity. These models have been successfully integrated into TTS systems, enhancing their ability to produce human-like voices. Similarly, diffusion models have been applied to music generation, enabling the creation of original compositions and soundscapes.

4.4. Drug Discovery and Molecular Design

In the field of drug discovery, diffusion models have been employed to generate novel molecular structures with desired properties [67]. By conditioning the generative process on specific chemical or biological criteria, these models can design molecules that satisfy constraints such as drug-likeness, binding affinity, and toxicity [68]. Applications in this domain include de novo drug design, protein structure prediction, and material discovery [69]. Diffusion models have also been used to explore the chemical space more efficiently, accelerating the discovery of compounds for pharmaceutical and industrial use.

4.5. 3D Modeling and Animation

Diffusion models have been extended to 3D data, enabling applications in computer graphics, animation, and virtual reality. These models can generate 3D shapes, textures, and motion sequences, facilitating the creation of realistic and dynamic virtual environments [70]. For instance, diffusion-based methods have been used to synthesize 3D assets for video games, simulate physical phenomena, and design architectural models [71]. Their ability to generate high-quality 3D content has made them valuable tools for creative industries and scientific visualization [72].

4.6. Creative Content Generation

In the realm of creative content, diffusion models have empowered artists, designers, and content creators to push the boundaries of innovation. Applications include generating unique artworks, designing fashion collections, creating video effects, and composing music [73]. The ability to condition generation on user input has enabled interactive tools for creativity, allowing users to co-create with AI systems [74].

4.7. Data Augmentation and Synthetic Data Generation

Diffusion models have been used to augment datasets by generating realistic synthetic samples. This capability is particularly valuable in scenarios where data is scarce or imbalanced, such as medical imaging, autonomous driving, and rare event prediction [75]. Synthetic data generated by diffusion models can improve the performance and robustness of downstream machine learning tasks by providing diverse and high-quality training samples.

4.8. Scientific Research and Simulation

In scientific research, diffusion models have been applied to simulate complex phenomena, such as climate patterns, fluid dynamics, and astrophysical processes. By learning the underlying dynamics of these systems, diffusion models can generate realistic simulations that aid in hypothesis testing and decision-making. Additionally, diffusion models have been used in genomics and bioinformatics to analyze and generate biological sequences, enabling advancements in personalized medicine and genetic engineering [76].

 doi:10.20944/preprints202502.0524.v1

4.9. Real-World Deployments and Challenges

The adoption of diffusion models in real-world applications has demonstrated their transformative potential [77]. From powering creative tools like Adobe Firefly to enabling advanced medical imaging solutions, diffusion models are reshaping industries [78]. However, their deployment also presents challenges, such as computational cost, ethical considerations, and the need for robust evaluation metrics [79]. Addressing these challenges is critical to ensuring the responsible and effective use of diffusion models [80]. The applications of diffusion models discussed in this section underscore their versatility and impact across domains [81]. In the following section, we will examine the challenges and limitations of diffusion models, along with potential directions for future research [82].

5. Challenges and Limitations

Despite their remarkable success across various applications, diffusion models face several challenges and limitations that hinder their broader adoption and efficiency [83]. This section outlines these challenges, focusing on computational demands, scalability, and ethical considerations [84].

5.1. Computational Complexity

One of the most significant drawbacks of diffusion models is their computational cost. Generating a single sample typically requires hundreds or thousands of denoising steps, each involving a forward pass through a deep neural network [85]. This iterative process is computationally intensive and time-consuming, making diffusion models slower than alternative generative frameworks like GANs, which generate samples in a single forward pass. Efforts to address this limitation include optimizing the number of denoising steps, designing more efficient neural network architectures, and developing fast sampling techniques [86]. However, achieving a balance between computational efficiency and sample quality remains an ongoing challenge [87].

5.2. Memory and Resource Requirements

Training diffusion models often requires substantial computational resources, including large-scale GPUs or TPUs, significant memory, and extensive training time. The high-dimensional nature of the data and the iterative training process contribute to these resource demands, limiting the accessibility of diffusion models to organizations with significant computational infrastructure. Techniques such as gradient checkpointing, mixed-precision training, and distributed learning have been employed to mitigate these challenges. Nevertheless, resource constraints remain a barrier for smaller research groups and practitioners.

To further alleviate these resource demands, model compression techniques such as pruning, decomposition, and quantization have been explored. **Pruning** involves removing less important weights or neurons from the model, reducing the number of parameters and improving efficiency without significantly compromising performance [88–90]. **Decomposition** techniques, such as low-rank factorization, break down large model matrices into smaller components, leading to reduced memory and computational requirements[68,91–93]. **Quantization** reduces the precision of model parameters, effectively compressing the model while maintaining an acceptable level of accuracy. These techniques, when combined with traditional optimization methods, can significantly improve the efficiency of diffusion models, making them more accessible to a broader range of users and applications [82,94]. However, careful consideration is required to ensure that the trade-offs in performance and quality are minimized, especially in tasks that require high precision or intricate details.

5.3. Scalability to High-Resolution Data

While diffusion models have demonstrated impressive results on medium-resolution data, scaling them to handle high-resolution images, videos, and 3D content poses additional challenges. The increased dimensionality of the data requires more complex architectures and longer training times,

 doi:10.20944/preprints202502.0524.v1

further exacerbating computational demands [95]. Moreover, generating high-resolution content often requires fine-tuning the noise schedule, model parameters, and conditioning mechanisms, which can be labor-intensive and prone to trial-and-error experimentation [96].

5.4. Mode Collapse and Diversity

Although diffusion models generally exhibit better sample diversity than GANs, they are not entirely immune to issues such as mode collapse [97]. In certain cases, the model may fail to capture the full complexity of the data distribution, leading to underrepresented or missing modes in the generated samples [98]. Addressing this issue requires careful design of the training objective, noise schedule, and model architecture [99]. Research into adaptive loss functions and regularization techniques is critical to ensuring that diffusion models generate diverse and representative samples [100].

5.5. Interpretability and Theoretical Understanding

While diffusion models are grounded in probabilistic principles, their iterative nature and complex dynamics make them challenging to interpret. The lack of a comprehensive theoretical understanding of their convergence properties, optimal noise schedules, and parameterization choices poses a barrier to their further development [101]. Efforts to improve interpretability include analyzing the role of the score function, studying the behavior of reverse processes, and exploring connections to other generative frameworks [102]. Advancing theoretical insights will be essential for designing more efficient and reliable diffusion models [103].

5.6. Ethical and Social Implications

As with other generative AI technologies, diffusion models raise ethical concerns related to misuse, bias, and intellectual property [104]. The ability to generate highly realistic content can be exploited for malicious purposes, such as creating deepfakes, spreading misinformation, or infringing on copyrights [105]. Additionally, biases present in the training data can propagate into the generated content, leading to unfair or harmful outcomes. Mitigating these risks requires a multi-faceted approach, including the development of robust content verification tools, the implementation of safeguards to prevent misuse, and the adoption of ethical guidelines for training and deploying diffusion models [106]. Transparency in data collection and model training processes is also crucial to addressing biases and ensuring accountability [107].

5.7. Evaluation Challenges

Evaluating the performance of diffusion models remains a complex task [108]. Traditional metrics such as Inception Score (IS) and Fréchet Inception Distance (FID) provide a measure of sample quality and diversity but do not fully capture other aspects of generative performance, such as conditional fidelity, semantic coherence, and computational efficiency [109]. The development of new evaluation metrics tailored to diffusion models is an active area of research. These metrics should account for the unique characteristics of diffusion-based generation, providing a more comprehensive assessment of model performance [110].

5.8. Open Research Questions

Several open questions remain in the development and application of diffusion models [111]. These include:

- How can we design more efficient and scalable sampling techniques without compromising sample quality [112]?
- What are the optimal noise schedules and parameterization strategies for different data types and applications [113]?
- How can we improve the interpretability and theoretical understanding of diffusion models [114]?

 doi:10.20944/preprints202502.0524.v1

- What safeguards and ethical frameworks are necessary to prevent misuse and ensure responsible deployment [115]?

Addressing these challenges will require collaboration across disciplines, combining insights from machine learning, statistics, ethics, and domain-specific expertise [116]. In the following section, we discuss future directions for diffusion models, highlighting promising areas of research and potential solutions to the challenges outlined above [117].

6. Future Directions

The rapid development of diffusion models has opened numerous avenues for future research and innovation [118]. Building on their foundational strengths and addressing current limitations, this section outlines promising directions that could shape the next generation of diffusion-based generative models [119].

6.1. Efficient Sampling and Training

Improving the efficiency of sampling and training processes is a critical area of research. Strategies to reduce the number of denoising steps while maintaining sample quality include:

- **Progressive Distillation:** Iteratively distilling the generative process into fewer steps, enabling faster sampling without significant quality degradation [120].
- **Hybrid Approaches:** Combining diffusion models with other frameworks, such as GANs or autoregressive models, to leverage their complementary strengths [121].
- **Parallelized Architectures:** Designing architectures that support parallel processing of diffusion steps, reducing latency during both training and sampling [122].

These advancements could make diffusion models more accessible for real-time and resource-constrained applications [123].

6.2. Scalability to High-Resolution and Complex Data

Scaling diffusion models to handle high-resolution images, long video sequences, and complex 3D data is another important direction [124]. Promising approaches include:

- **Hierarchical Models:** Leveraging multi-scale architectures to process data at different levels of granularity, improving efficiency and scalability [125].
- **Cross-Modality Conditioning:** Enhancing the ability of models to generate and align complex multimodal data, such as text-conditioned 3D assets or video [126].
- **Sparse Representations:** Incorporating sparse or compressed representations to reduce the memory footprint and computational cost of high-dimensional data [127].

These innovations could enable applications in areas such as virtual reality, video synthesis, and scientific visualization [128].

6.3. Theoretical Advances and Interpretability

Furthering the theoretical understanding of diffusion models is essential for improving their design and reliability. Key research directions include:

- **Optimal Noise Schedules:** Investigating the mathematical properties of noise schedules to identify optimal configurations for various data distributions [129].
- **Stochastic Process Analysis:** Deepening the analysis of reverse-time stochastic differential equations (SDEs) to better understand the generative dynamics [130].
- **Explainable Generative Processes:** Developing tools and frameworks to interpret the intermediate steps of diffusion models, enhancing their transparency and trustworthiness [131].

These efforts could lead to more robust and theoretically grounded generative models.

 doi:10.20944/preprints202502.0524.v1

6.4. Ethical and Responsible AI Practices

As diffusion models become more prevalent, ensuring their ethical use is paramount [132]. Future work in this area should focus on:

- **Bias Mitigation:** Developing methods to identify and reduce biases in training data and generated outputs [133].
- **Content Verification:** Creating tools to detect synthetic content and distinguish it from real data, addressing concerns about misinformation and misuse [134].
- **Transparent Development:** Establishing guidelines for the responsible development and deployment of diffusion models, including open disclosure of training data and algorithms [135].

By prioritizing ethical considerations, researchers can ensure that diffusion models benefit society while minimizing potential harms [136].

6.5. Domain-Specific Applications

Expanding the use of diffusion models in domain-specific applications offers significant potential for innovation. Examples include:

- **Healthcare:** Generating synthetic medical data for training diagnostic models, enhancing privacy and diversity in healthcare datasets.
- **Material Science:** Designing novel materials with desired properties using generative modeling of molecular structures.
- **Education and Creativity:** Creating tools for personalized learning and interactive creative expression, such as AI-assisted art or music generation.

Tailoring diffusion models to specific domains can unlock new possibilities and accelerate progress in these fields [137].

6.6. Integration with Emerging Technologies

The integration of diffusion models with other emerging technologies represents a promising frontier [138]. Potential synergies include:

- **Quantum Computing:** Exploring quantum-inspired diffusion processes to enhance generative capabilities and computational efficiency [139].
- **Edge AI:** Adapting diffusion models for deployment on edge devices, enabling on-device generative capabilities for applications like augmented reality and personalized assistants [140].
- **Federated Learning:** Leveraging federated training paradigms to build diffusion models while preserving data privacy and security [141].

These integrations could broaden the applicability and impact of diffusion models across industries [142].

6.7. Novel Architectures and Frameworks

Finally, the exploration of novel architectures and training frameworks can push the boundaries of what diffusion models can achieve. Directions include:

- **Transformer-Based Diffusion Models:** Combining the representational power of transformers with the generative capabilities of diffusion processes.
- **Dynamic Diffusion Processes:** Investigating adaptive diffusion processes that adjust their dynamics based on data complexity or user-defined criteria.
- **Unsupervised and Few-Shot Learning:** Enhancing the ability of diffusion models to learn from limited or unlabeled data, reducing the reliance on large annotated datasets.

These innovations could redefine the landscape of generative modeling, making diffusion models more versatile and powerful.

 doi:10.20944/preprints202502.0524.v1

6.8. Conclusion

The future of diffusion models is both exciting and challenging, with numerous opportunities for innovation and impact [143]. By addressing current limitations and exploring these promising directions, researchers can unlock the full potential of diffusion models, paving the way for transformative advancements in artificial intelligence and beyond [144].

7. Conclusion

Diffusion models have emerged as a transformative force in the field of generative AI, offering a principled and versatile framework for generating high-quality data across various modalities. Rooted in probabilistic foundations and leveraging iterative refinement processes, these models have demonstrated exceptional performance in applications ranging from image synthesis and text-to-image generation to audio synthesis, molecular design, and beyond [145]. The success of diffusion models lies in their ability to approximate complex data distributions while maintaining sample diversity and fidelity [146]. By systematically denoising data through a reverse diffusion process, these models overcome many limitations of earlier generative frameworks, such as instability in training and mode collapse. Their flexibility to incorporate conditioning mechanisms has further expanded their utility in multimodal and domain-specific tasks [147]. Despite their remarkable capabilities, diffusion models face challenges that must be addressed to unlock their full potential. These include high computational costs, scalability issues, and the need for more interpretable and theoretically grounded frameworks [148]. Moreover, as their adoption grows, ethical considerations surrounding bias, misuse, and societal impact must remain at the forefront of research and development efforts [149]. Looking ahead, the future of diffusion models is ripe with possibilities. Innovations in efficient sampling techniques, scalable architectures, and domain-specific adaptations promise to make these models more accessible and impactful [150]. Integrating diffusion models with emerging technologies such as quantum computing, edge AI, and federated learning could further expand their horizons. Additionally, advancing ethical practices and transparent development will ensure that the benefits of diffusion models are realized responsibly and equitably [151].

In conclusion, diffusion models represent a significant milestone in the evolution of generative AI. By addressing current limitations and embracing new directions, researchers and practitioners can harness their full potential, driving progress across diverse fields and contributing to the broader goal of advancing artificial intelligence for the benefit of society.

References

1. 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**.
2. Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; Wang, L. Mitigating hallucination in large multi-modal models via robust instruction tuning. In Proceedings of the The Twelfth International Conference on Learning Representations, 2023.
3. 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.
4. Yu, F.; Huang, K.; Wang, M.; Cheng, Y.; Chu, W.; Cui, L. Width & depth pruning for vision transformers. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 3143–3151.
5. Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150* **2019**.
6. Peng, Z.; Wang, W.; Dong, L.; Hao, Y.; Huang, S.; Ma, S.; Wei, F. Kosmos-2: Grounding Multimodal Large Language Models to the World. *ArXiv* **2023**, *abs/2306*.
7. ShareGPT. https://sharegpt.com/, 2023.
8. 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**.
9. 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].

 doi:10.20944/preprints202502.0524.v1

10. 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.
11. Ganjdanesh, A.; Kang, Y.; Liu, Y.; Zhang, R.; Lin, Z.; Huang, H. Mixture of Efficient Diffusion Experts Through Automatic Interval and Sub-Network Selection. *arXiv preprint arXiv:2409.15557* **2024**.
12. Hu, S.; Tu, Y.; Han, X.; He, C.; Cui, G.; Long, X.; Zheng, Z.; Fang, Y.; Huang, Y.; Zhao, W.; et al. MiniCPM: Unveiling the Potential of Small Language Models with Scalable Training Strategies. *arXiv preprint arXiv:2404.06395* **2024**.
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, 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].
15. 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].
16. 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**.
17. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *ICLR* **2021**.
18. 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**.
19. 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.
20. Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; Ommer, B. High-resolution image synthesis with latent diffusion models. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
21. 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.
22. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
23. 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].
24. tiny vision language model. https://github.com/vikhyat/moondream, 2024.
25. 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**.
26. 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.
27. Malladi, S.; Gao, T.; Nichani, E.; Damian, A.; Lee, J.D.; Chen, D.; Arora, S. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems* **2023**, *36*, 53038–53075.
28. 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.
29. Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; Komatsuzaki, A. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114* **2021**.
30. 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.
31. Mathew, M.; Karatzas, D.; Jawahar, C. Docvqa: A dataset for vqa on document images. In Proceedings of the Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2021, pp. 2200–2209.
32. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, 2021.
33. Peng, B.; Alcaide, E.; Anthony, Q.; Albalak, A.; Arcadinho, S.; Cao, H.; Cheng, X.; Chung, M.; Grella, M.; GV, K.K.; et al. Rwkv: Reinventing rnns for the transformer era. *arXiv preprint arXiv:2305.13048* **2023**.
34. Chu, X.; Qiao, L.; Zhang, X.; Xu, S.; Wei, F.; Yang, Y.; Sun, X.; Hu, Y.; Lin, X.; Zhang, B.; et al. MobileVLM V2: Faster and Stronger Baseline for Vision Language Model. *arXiv preprint arXiv:2402.03766* **2024**.

 doi:10.20944/preprints202502.0524.v1

35. Dockhorn, T.; Vahdat, A.; Kreis, K. Score-Based Generative Modeling with Critically-Damped Langevin Diffusion. In Proceedings of the International Conference on Learning Representations, 2021.
36. 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**.
37. 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**.
38. Bolya, D.; Fu, C.Y.; Dai, X.; Zhang, P.; Hoffman, J. Hydra attention: Efficient attention with many heads. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 35–49.
39. Chen, J.; Liu, Y.; Li, D.; An, X.; Feng, Z.; Zhao, Y.; Xie, Y. Plug-and-Play Grounding of Reasoning in Multimodal Large Language Models. *arXiv preprint arXiv:2403.19322* **2024**.
40. Wang, Z.; Wang, C.; Xu, X.; Zhou, J.; Lu, J. Quantformer: Learning extremely low-precision vision transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2022**.
41. Yao, Y.; Yu, T.; Wang, C.; Cui, J.; Zhu, H.; Cai, T.; Zhao, W.; Zhang, K.; Hong, Y.; Li, H.; et al. MiniCPM-V 2.0: An Efficient End-side MLLM with Strong OCR and Understanding Capabilities. https://github.com/OpenBMB/MiniCPM-V, 2024.
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. Li, W.; Wang, X.; Xia, X.; Wu, J.; Xiao, X.; Zheng, M.; Wen, S. Sepvit: Separable vision transformer. *arXiv preprint arXiv:2203.15380* **2022**.
44. Zhang, L.; Zhang, L.; Shi, S.; Chu, X.; Li, B. Lora-fa: Memory-efficient low-rank adaptation for large language models fine-tuning. *arXiv preprint arXiv:2308.03303* **2023**.
45. 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**.
46. 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*.
47. Salimans, T.; Ho, J. Progressive Distillation for Fast Sampling of Diffusion Models. In Proceedings of the International Conference on Learning Representations, 2022.
48. 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.
49. 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].
50. Fang, A.; Jose, A.M.; Jain, A.; Schmidt, L.; Toshev, A.; Shankar, V. Data filtering networks. *arXiv preprint arXiv:2309.17425* **2023**.
51. 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.
52. 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*.
53. 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**.
54. 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.
55. Lin, B.; Tang, Z.; Ye, Y.; Cui, J.; Zhu, B.; Jin, P.; Zhang, J.; Ning, M.; Yuan, L. Moe-llava: Mixture of experts for large vision-language models. *arXiv preprint arXiv:2401.15947* **2024**.
56. 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*.
57. Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. Llama 2: Open Foundation and Fine-Tuned Chat Models, 2023, [arXiv:cs.CL/2307.09288].

 doi:10.20944/preprints202502.0524.v1

58. 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.
59. 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**.
60. Vahdat, A.; Kautz, J. NVAE: a deep hierarchical variational autoencoder. In Proceedings of the Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020, pp. 19667–19679.
61. Cao, J.; Ye, P.; Li, S.; Yu, C.; Tang, Y.; Lu, J.; Chen, T. MADTP: Multimodal Alignment-Guided Dynamic Token Pruning for Accelerating Vision-Language Transformer, 2024, [arXiv:cs.CV/2403.02991].
62. Du, Y.; Yang, M.; Dai, B.; Dai, H.; Nachum, O.; Tenenbaum, J.B.; Schuurmans, D.; Abbeel, P. Learning universal policies via text-guided video generation. In Proceedings of the Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023, pp. 9156–9172.
63. 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**.
64. Xu, Y.; Zhao, Y.; Xiao, Z.; Hou, T. Ufogen: You forward once large scale text-to-image generation via diffusion gans. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 8196–8206.
65. Yuan, Z.; Shang, Y.; Zhou, Y.; Dong, Z.; Xue, C.; Wu, B.; Li, Z.; Gu, Q.; Lee, Y.J.; Yan, Y.; et al. LLM Inference Unveiled: Survey and Roofline Model Insights, 2024, [arXiv:cs.CL/2402.16363].
66. Lu, P.; Bansal, H.; Xia, T.; Liu, J.; Li, C.; Hajishirzi, H.; Cheng, H.; Chang, K.W.; Galley, M.; Gao, J. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255* **2023**.
67. 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.
68. 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.
69. Wei, H.; Kong, L.; Chen, J.; Zhao, L.; Ge, Z.; Yu, E.; Sun, J.; Han, C.; Zhang, X. Small Language Model Meets with Reinforced Vision Vocabulary. *arXiv preprint arXiv:2401.12503* **2024**.
70. 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**.
71. Gupta, A.; Dollar, P.; Girshick, R. Lvis: A dataset for large vocabulary instance segmentation. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 5356–5364.
72. Vats, A.; Raja, R.; Jain, V.; Chadha, A. The Evolution of Mixture of Experts: A Survey from Basics to Breakthroughs **2024**.
73. 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].
74. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* **2023**.
75. 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.
76. Li, Y.; Wang, C.; Jia, J. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models, 2023, [arXiv:cs.CV/2311.17043].
77. Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966* **2023**.
78. 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.
79. 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**.
80. Wang, A.; Chen, H.; Lin, Z.; Zhao, S.; Han, J.; Ding, G. CAIT: Triple-Win Compression towards High Accuracy, Fast Inference, and Favorable Transferability For ViTs. *arXiv preprint arXiv:2309.15755* **2023**.

 doi:10.20944/preprints202502.0524.v1

81. Yan, H.; Liu, X.; Pan, J.; Liew, J.H.; Liu, Q.; Feng, J. Perflow: Piecewise rectified flow as universal plug-and-play accelerator. *arXiv preprint arXiv:2405.07510* **2024**.
82. Oquab, M.; Darcet, T.; Moutakanni, T.; Vo, H.; Szafraniec, M.; Khalidov, V.; Fernandez, P.; Haziza, D.; Massa, F.; El-Nouby, A.; et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193* **2023**.
83. Tang, Y.; Han, K.; Wang, Y.; Xu, C.; Guo, J.; Xu, C.; Tao, D. Patch slimming for efficient vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12165–12174.
84. 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.
85. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* **2023**.
86. Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; Elhoseiny, M. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478* **2023**.
87. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042* **2023**.
88. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
89. 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**.
90. 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].
91. 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**.
92. Lee, J.; Lee, Y.; Kim, J.; Kosiorek, A.; Choi, S.; Teh, Y.W. Set transformer: A framework for attention-based permutation-invariant neural networks. In Proceedings of the International conference on machine learning. PMLR, 2019, pp. 3744–3753.
93. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
94. 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.
95. 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.
96. 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*.
97. Zhao, H.; Zhang, M.; Zhao, W.; Ding, P.; Huang, S.; Wang, D. Cobra: Extending Mamba to Multi-Modal Large Language Model for Efficient Inference. *arXiv preprint arXiv:2403.14520* **2024**.
98. 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**.
99. He, B.; Li, H.; Jang, Y.K.; Jia, M.; Cao, X.; Shah, A.; Shrivastava, A.; Lim, S.N. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding, 2024, [arXiv:cs.CV/2404.05726].
100. 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].
101. 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.

 doi:10.20944/preprints202502.0524.v1

102. 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.
103. Jiang, S.; Zheng, T.; Zhang, Y.; Jin, Y.; Liu, Z. MoE-TinyMed: Mixture of Experts for Tiny Medical Large Vision-Language Models, 2024, [arXiv:cs.CV/2404.10237].
104. 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.
105. Luhman, E.; Luhman, T. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388* **2021**.
106. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In Proceedings of the International Conference on Learning Representations, 2022.
107. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. In Proceedings of the International Conference on Machine Learning, 2023, pp. 32211–32252.
108. Xu, Y.; Zhang, Z.; Zhang, M.; Sheng, K.; Li, K.; Dong, W.; Zhang, L.; Xu, C.; Sun, X. Evo-vit: Slow-fast token evolution for dynamic vision transformer. In Proceedings of the Proceedings of the AAAI Conference on Artificial Intelligence, 2022, Vol. 36, pp. 2964–2972.
109. Ding, Y.; Qin, H.; Yan, Q.; Chai, Z.; Liu, J.; Wei, X.; Liu, X. Towards accurate post-training quantization for vision transformer. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5380–5388.
110. Zhou, Q.; Sheng, K.; Zheng, X.; Li, K.; Sun, X.; Tian, Y.; Chen, J.; Ji, R. Training-free transformer architecture search. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 10894–10903.
111. 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.
112. 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**.
113. 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.
114. 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**.
115. Mescheder, L.; Geiger, A.; Nowozin, S. Which training methods for GANs do actually converge? In Proceedings of the International conference on machine learning, 2018, pp. 3481–3490.
116. Renggli, C.; Pinto, A.S.; Houlsby, N.; Mustafa, B.; Puigcerver, J.; Riquelme, C. Learning to merge tokens in vision transformers. *arXiv preprint arXiv:2202.12015* **2022**.
117. 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.
118. Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; Zhao, R. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195* **2023**.
119. 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.
120. Xiao, Z.; Kreis, K.; Vahdat, A. Tackling the Generative Learning Trilemma with Denoising Diffusion GANs. In Proceedings of the International Conference on Learning Representations, 2022.
121. He, M.; Liu, Y.; Wu, B.; Yuan, J.; Wang, Y.; Huang, T.; Zhao, B. Efficient Multimodal Learning from Data-centric Perspective. *arXiv preprint arXiv:2402.11530* **2024**.
122. 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**.

 doi:10.20944/preprints202502.0524.v1

123. 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.
124. Jolicoeur-Martineau, A.; Li, K.; Piché-Taillefer, R.; Kachman, T.; Mitliagkas, I. Gotta go fast when generating data with score-based models. *arXiv preprint arXiv:2105.14080* **2021**.
125. Yu, Y.Q.; Liao, M.; Wu, J.; Liao, Y.; Zheng, X.; Zeng, W. TextHawk: Exploring Efficient Fine-Grained Perception of Multimodal Large Language Models. *arXiv preprint arXiv:2404.09204* **2024**.
126. Zhang, W.; Lin, T.; Liu, J.; Shu, F.; Li, H.; Zhang, L.; Wanggui, H.; Zhou, H.; Lv, Z.; Jiang, H.; et al. HyperLLaVA: Dynamic Visual and Language Expert Tuning for Multimodal Large Language Models, 2024, [arXiv:cs.AI/2403.13447].
127. 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.
128. Zong, Z.; Ma, B.; Shen, D.; Song, G.; Shao, H.; Jiang, D.; Li, H.; Liu, Y. MoVA: Adapting Mixture of Vision Experts to Multimodal Context, 2024, [arXiv:cs.CV/2404.13046].
129. 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].
130. 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**.
131. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* **2021**.
132. Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; Cao, Y. Eva: Exploring the limits of masked visual representation learning at scale. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 19358–19369.
133. Yang, L.; Zhang, Z.; Song, Y.; Hong, S.; Xu, R.; Zhao, Y.; Zhang, W.; Cui, B.; Yang, M.H. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys* **2023**, *56*, 1–39.
134. Berthelot, D.; Autef, A.; Lin, J.; Yap, D.A.; Zhai, S.; Hu, S.; Zheng, D.; Talbott, W.; Gu, E. Tract: Denoising diffusion models with transitive closure time-distillation. *arXiv preprint arXiv:2303.04248* **2023**.
135. Team, G.; Mesnard, T.; Hardin, C.; Dadashi, R.; Bhupatiraju, S.; Pathak, S.; Sifre, L.; Rivière, M.; Kale, M.S.; Love, J.; et al. Gemma: Open Models Based on Gemini Research and Technology, 2024, [arXiv:cs.CL/2403.08295].
136. 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**.
137. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W.X.; Wen, J.R. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355* **2023**.
138. Ho, J.; Salimans, T. Classifier-Free Diffusion Guidance. In Proceedings of the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications, 2021.
139. Javaheripi, M.; Bubeck, S.; Abdin, M.; Aneja, J.; Bubeck, S.; Mendes, C.C.T.; Chen, W.; Del Giorno, A.; Eldan, R.; Gopi, S.; et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog* **2023**.
140. 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.
141. 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**.
142. Wu, Q.; Ye, W.; Zhou, Y.; Sun, X.; Ji, R. Not All Attention is Needed: Parameter and Computation Efficient Transfer Learning for Multi-modal Large Language Models. *arXiv preprint arXiv:2403.15226* **2024**.
143. Hinck, M.; Olson, M.L.; Cobbley, D.; Tseng, S.Y.; Lal, V. LLaVA-Gemma: Accelerating Multimodal Foundation Models with a Compact Language Model. *arXiv preprint arXiv:2404.01331* **2024**.
144. Wu, K.; Zhang, J.; Peng, H.; Liu, M.; Xiao, B.; Fu, J.; Yuan, L. Tinyvit: Fast pretraining distillation for small vision transformers. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 68–85.
145. Chen, Z.; Ma, X.; Fang, G.; Tan, Z.; Wang, X. AsyncDiff: Parallelizing Diffusion Models by Asynchronous Denoising, 2024, [arXiv:cs.CV/2406.06911].
146. Ainslie, J.; Lee-Thorp, J.; de Jong, M.; Zemlyanskiy, Y.; Lebrón, F.; Sanghai, S. GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints, 2023, [arXiv:cs.CL/2305.13245].

 doi:10.20944/preprints202502.0524.v1

147. 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.
148. 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**.
149. DeepSeek-AI. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954* **2024**.
150. Leviathan, Y.; Kalman, M.; Matias, Y. Fast inference from transformers via speculative decoding. In Proceedings of the International Conference on Machine Learning. PMLR, 2023, pp. 19274–19286.
151. Kuznedelev, D.; Kurtić, E.; Frantar, E.; Alistarh, D. CAP: Correlation-Aware Pruning for Highly-Accurate Sparse Vision Models. *Advances in Neural Information Processing Systems* **2024**, *36*.