Preprints.org

Review

Not peer-reviewed version

Making Diffusion Models Practical: A Survey on Acceleration and Optimization

Suzuki Sakurama and Zoe Brian [*]

Posted Date: 11 February 2025

doi: 10.20944/preprints202502.0779.v1

Keywords: diffusion models; efficient sampling; model compression; knowledge distillation; generative modeling; deep learning; hybrid generative models; Latent Diffusion; accelerated inference

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Review

Making Diffusion Models Practical: A Survey on Acceleration and Optimization

Suzuki Sakurama [1] and Zoe Brian [1,2,*]

[1] Graduate School of Engineering, 2-11-16 Yayoi, Bunkyo-ku, Tokyo, 113-8656, Tokyo, Japan
[2] Department of Computer Science, 7-3-1 Kanda-Nihonbashi, Chuo-ku, Tokyo, 103-0027, Tokyo, Japan
[*] Correspondence: zoe.brian@gs.mail.u-tokyo.ac.jp

Abstract: Diffusion models have emerged as a powerful class of generative models, achieving state-of-the-art performance in image, audio, and video synthesis. However, their widespread adoption is hindered by high computational costs, slow inference times, and memory-intensive training. In response, numerous techniques have been proposed to enhance the efficiency of diffusion models while maintaining or improving generation quality. This survey provides a comprehensive review of recent advances in efficient diffusion models. We categorize these approaches into four key areas: (1) accelerated sampling methods, which reduce the number of function evaluations required for inference; (2) efficient model architectures, including lightweight U-Net and transformer variants; (3) knowledge distillation and model compression techniques, such as progressive distillation and pruning; and (4) hybrid generative frameworks that integrate diffusion models with alternative paradigms like GANs and VAEs. Additionally, we discuss open challenges, including the trade-offs between sampling speed and quality, memory-efficient training strategies, and real-time deployment on edge devices. We highlight promising research directions, such as adaptive sampling, hardware-aware optimizations, and self-distilling diffusion models. By providing a structured overview of efficiency-focused advancements, we aim to guide future research toward making diffusion models more practical, scalable, and accessible for real-world applications.

Keywords: diffusion models; efficient sampling; model compression; knowledge distillation; generative modeling; deep learning; hybrid generative models; Latent Diffusion; accelerated inference

1. Introduction

Generative models have witnessed remarkable progress in recent years, with diffusion models emerging as a dominant framework for generating high-quality synthetic data across various modalities, including images, audio, and video. These models operate by simulating a diffusion process, where data is gradually transformed into Gaussian noise and then reconstructed through a learned denoising process [1]. Due to their strong theoretical foundations and empirical success, diffusion models have outperformed alternative generative approaches, such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), in terms of sample quality, diversity, and training stability [2]. As a result, they have been widely adopted in applications such as text-to-image synthesis, inpainting, super-resolution, molecular generation, and speech synthesis [3]. Despite their impressive capabilities, diffusion models suffer from significant computational inefficiencies [4]. One of the primary challenges is the high cost of inference, as generating a single sample requires performing hundreds or even thousands of forward passes through a neural network. This iterative sampling process makes diffusion models orders of magnitude slower than GANs, which can generate samples in a single forward pass [5]. The high computational demands not only limit real-time applications but also make it challenging to deploy these models on edge devices with limited resources [6]. Furthermore, training diffusion models is also resource-intensive, requiring substantial GPU memory and extended

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

training times, which restricts accessibility to large-scale organizations with high-performance computing infrastructure [7]. To address these limitations, researchers have explored multiple strategies to enhance the efficiency of diffusion models [8]. These approaches can be broadly categorized into four key areas:

- **Accelerated Sampling Techniques**: Methods such as improved numerical solvers, stochastic samplers, and adaptive step-size schedules aim to reduce the number of inference steps while maintaining sample quality [9].
- **Efficient Model Architectures**: Lightweight network architectures, low-rank approximations, and pruning techniques help reduce memory and computation requirements without significantly compromising performance [10,11].
- **Knowledge Distillation and Model Compression**: Techniques such as teacher-student training, distillation of diffusion processes into simpler models, and progressive denoising help create compact yet effective diffusion models [12].
- **Hybrid and Alternative Frameworks**: Combining diffusion models with GANs, VAEs, or implicit models seeks to leverage the strengths of each approach to achieve both efficiency and high-quality generation [13].

This survey provides a comprehensive review of recent advancements aimed at improving the efficiency of diffusion models [14]. We systematically categorize and analyze existing methods, highlighting their advantages, limitations, and trade-offs [15]. Additionally, we discuss emerging trends and potential future research directions in the quest to make diffusion models more computationally feasible while preserving their generative power. By presenting a structured overview of efficiency-improving techniques, this survey aims to serve as a valuable resource for researchers and practitioners working on generative modeling [16]. We hope that our insights will inspire further innovations in the field, ultimately enabling the broader adoption of diffusion models in practical applications, including real-time content generation, interactive AI systems, and resource-constrained deployments [17].

2. Background

Diffusion models are a class of generative models based on iterative denoising processes, inspired by non-equilibrium thermodynamics [18]. They have gained significant attention due to their ability to generate high-quality samples across multiple domains, including images, audio, and 3D data. In this section, we provide an overview of the fundamental principles of diffusion models, including the forward and reverse diffusion processes, mathematical formulations, and key developments leading to their widespread adoption [19].

2.1. Forward and Reverse Diffusion Processes

Diffusion models operate by defining a Markov chain that gradually transforms a data distribution into a simple prior distribution, typically Gaussian noise [20]. This transformation, known as the *forward process*, is formulated as a series of Gaussian transitions:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}), \tag{1}$$

where $\mathbf{x}_t$ represents the corrupted data at timestep t, and β_t is a variance schedule controlling the noise level [21]. To generate samples, the model learns to approximate the *reverse process*, which denoises the data step by step [22]:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{2}$$

The neural network parameterized by θ predicts either the noise component or the clean data sample, enabling gradual reconstruction of realistic data from pure noise.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

2.2. Training Objective

Diffusion models are trained by minimizing a variant of the variational lower bound (ELBO), which can be reformulated as a noise prediction task. The common training objective simplifies to:

$$\mathbb{E}_{\mathbf{x}_0,\boldsymbol{\epsilon},t}\left[\|\boldsymbol{\epsilon}-\boldsymbol{\epsilon}_\theta(\mathbf{x}_t,t)\|^2\right], \tag{3}$$

where $\boldsymbol{\epsilon}$ represents the injected noise, and $\boldsymbol{\epsilon}_\theta$ is the model's predicted noise [23].

2.3. Key Variants of Diffusion Models

Several variants of diffusion models have been proposed to improve their effectiveness and efficiency:

- **Denoising Diffusion Probabilistic Models (DDPMs)**: The original formulation, introduced by Ho et al., demonstrating the capability of diffusion-based generative models [24].
- **Denoising Diffusion Implicit Models (DDIMs)**: A more efficient variant that enables deterministic sampling with fewer steps [25].
- **Score-Based Generative Models (SGMs)**: Leveraging score matching and stochastic differential equations to generalize diffusion processes.
- **Latent Diffusion Models (LDMs)**: Reducing computational complexity by applying diffusion in a lower-dimensional latent space instead of pixel space [26].

2.4. Challenges and Computational Bottlenecks

Despite their success, diffusion models suffer from key limitations, including:

- **Slow Inference**: Generating a single sample requires multiple forward passes, making real-time applications difficult.
- **High Memory Requirements**: Large neural networks with deep architectures demand significant GPU resources [27].
- **Training Complexity**: Optimizing diffusion models requires substantial computational power and time, limiting accessibility [28].

Understanding these challenges is essential before exploring the various efficiency improvements proposed in recent literature [29]. In the next section, we delve into techniques aimed at accelerating sampling, reducing model size, and improving overall efficiency while maintaining generation quality [30].

3. Efficient Diffusion Models

While diffusion models have demonstrated remarkable generative capabilities, their high computational cost remains a major limitation [31]. Generating high-quality samples requires performing hundreds or thousands of neural network evaluations, making inference slow and expensive [32]. Additionally, training diffusion models is memory-intensive and computationally demanding, restricting their scalability and deployment in real-world applications [33]. To address these challenges, several techniques have been proposed to improve the efficiency of diffusion models [34]. These approaches can be broadly classified into four categories: accelerated sampling methods, efficient model architectures, knowledge distillation and model compression, and hybrid generative frameworks [35]. In this section, we discuss each of these approaches in detail [36].

3.1. Accelerated Sampling Methods

One of the primary inefficiencies of diffusion models stems from the large number of timesteps required to generate high-quality samples [37]. Traditional sampling techniques rely on slow iterative denoising, making them impractical for real-time applications [38]. Various strategies have been developed to accelerate sampling, including:

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

3.1.1. Denoising Diffusion Implicit Models (DDIMs)

Denoising Diffusion Implicit Models (DDIMs) introduce a non-Markovian diffusion process that enables deterministic sampling [39]. By modifying the reverse diffusion process, DDIMs allow for significantly fewer sampling steps while maintaining high sample quality. The key idea is to reparameterize the reverse process to allow direct sampling of intermediate timesteps:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \tag{4}$$

where $\bar{\alpha}_t$ is a function of the noise schedule. This formulation allows skipping multiple timesteps, leading to a substantial reduction in the number of function evaluations [40].

3.1.2. Higher-Order Solvers and ODE-Based Methods

Score-based generative models can be interpreted as solving a stochastic differential equation (SDE) or an ordinary differential equation (ODE) [41]. Several works have proposed using advanced numerical solvers, such as:

- **Euler and Heun Methods**: Basic numerical solvers that reduce the number of function evaluations [42].
- **Runge-Kutta Methods**: Higher-order solvers that improve accuracy with fewer steps [43].
- **Exponential Integrators**: Methods that approximate the evolution of the diffusion process more efficiently [44].

These techniques reduce the number of sampling steps while maintaining high fidelity in generated samples.

3.1.3. Latent Diffusion Models (LDMs)

Latent Diffusion Models (LDMs) address computational inefficiencies by applying diffusion in a lower-dimensional latent space rather than pixel space [45]. This reduces the overall complexity of the model while preserving high-quality generation [46]. LDMs utilize a pretrained autoencoder to encode images into a compact representation, where the diffusion process is performed:

$$\mathbf{z}_t = \sqrt{\alpha_t}\mathbf{z}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, \tag{5}$$

where $\mathbf{z}_t$ represents the latent representation at timestep t [47]. By operating in the latent space, LDMs achieve significant speedups with reduced memory consumption [48].

3.2. Efficient Model Architectures

Another approach to improving diffusion model efficiency is designing lightweight neural network architectures that reduce computational overhead [49]. Some key advancements include:

3.2.1. Lightweight U-Net Architectures

Most diffusion models utilize U-Net-based architectures for denoising, but their size and depth make them computationally expensive [50]. Several modifications have been proposed to optimize these networks:

- **Channel Reduction**: Using fewer feature channels to decrease memory and computation [51].
- **Depthwise Separable Convolutions**: Reducing parameter count while maintaining expressive power [52].
- **Grouped Convolutions**: Improving efficiency by processing different feature groups separately [53].

3.2.2. Transformer-Based Diffusion Models

Recent research has explored using transformer architectures in place of CNNs for diffusion models [54]. While transformers offer superior scalability and global receptive fields, they can be

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

computationally heavy [55]. Optimized transformer variants, such as sparse attention mechanisms and low-rank approximations, have been proposed to balance efficiency and performance [56,57].

3.3. Knowledge Distillation and Model Compression

Diffusion models can be made more efficient by compressing large models into smaller, faster ones. Several techniques have been explored:

3.3.1. Teacher-Student Knowledge Distillation

Knowledge distillation transfers knowledge from a large teacher model to a smaller student model [58]. This can be done by training the student to match the teacher's intermediate representations or final outputs [59]. Distillation-based approaches have successfully reduced model size while preserving sample quality [60].

3.3.2. Progressive Distillation

Progressive distillation reduces the number of sampling steps by training a student model to mimic the teacher's results in fewer steps [61]. This allows for significant speedups without sacrificing visual fidelity [62].

3.3.3. Quantization and Pruning

Quantization reduces memory and computational costs by representing model parameters with lower precision (e.g., 8-bit or 4-bit representations) [63]. Pruning removes redundant parameters, reducing the overall model size [64]. These techniques have been applied to diffusion models to improve inference efficiency [65].

3.4. Hybrid Generative Models

Combining diffusion models with other generative frameworks has led to novel architectures that leverage the strengths of each approach [66]. Some notable examples include:

3.4.1. Diffusion-GAN Hybrids

Diffusion models excel in sample quality but are slow, while GANs generate samples quickly but may suffer from mode collapse [67]. Hybrid models combine diffusion models with adversarial training to achieve both efficiency and high-quality generation [68].

3.4.2. Variational Diffusion Models

Integrating VAEs with diffusion models enables efficient latent-space generation, reducing computational requirements while retaining expressive power [69].

3.4.3. Score-Based Energy Models

Score-based energy models combine diffusion models with energy-based learning, leading to improved sampling efficiency and stability [70].

3.5. Summary of Efficiency Improvements

Table 1 provides a summary of the different efficiency improvements and their impact on sampling speed, memory usage, and model complexity [71].

Table 1. Comparison of different efficiency improvements for diffusion models.

Method	**Speedup**	**Memory Reduction**	**Quality Trade-off**
DDIMs	High	Low	Minimal
Latent Diffusion Models	High	High	Minimal
Transformer-Based Models	Medium	Medium	Variable
Knowledge Distillation	High	High	Some
Hybrid Generative Models	Medium	Medium	Variable

In the next section, we discuss open challenges and future research directions in improving the efficiency of diffusion models [72].

4. Open Challenges and Future Directions

Despite significant progress in improving the efficiency of diffusion models, several challenges remain [73]. Many of the existing acceleration techniques introduce trade-offs, such as reduced sample quality, increased training complexity, or loss of flexibility [74]. In this section, we highlight key open challenges and potential future research directions that can further enhance the efficiency, scalability, and applicability of diffusion models [75].

4.1. Balancing Speed and Sample Quality

One of the most critical challenges in efficient diffusion models is the trade-off between sampling speed and output quality [76]. While methods such as DDIMs and higher-order solvers reduce the number of sampling steps, they may introduce artifacts or loss of fine details [77]. Future research could explore adaptive sampling techniques that dynamically adjust step sizes based on sample complexity, ensuring high-quality outputs while minimizing computational cost [78].

4.2. Memory-Efficient Training Strategies

Training diffusion models remains highly resource-intensive, often requiring high-end GPUs with large memory capacities. Techniques such as gradient checkpointing, low-precision training (e.g., mixed-precision or 8-bit training), and efficient tensor operations could be further explored to reduce memory footprint. Additionally, distributed training strategies tailored for diffusion models could help scale training to larger datasets without excessive resource requirements.

4.3. Lightweight Model Architectures

While efforts have been made to design more efficient U-Net and transformer architectures, there is still room for improvement [79]. Future work could investigate:

- **Neural Architecture Search (NAS)** for discovering optimal architectures that balance efficiency and performance [80].
- **Sparse and low-rank models** to reduce the number of parameters while preserving expressive power [11].
- **Modular diffusion networks** that adapt dynamically based on computational constraints [81].

Developing novel architectures tailored for mobile and edge-device deployment remains an open challenge [82].

4.4. Fast and Adaptive Sampling Methods

Existing accelerated samplers often rely on fixed schedules and step sizes, which may not be optimal for all data distributions [83]. Future research could focus on:

- **Adaptive step-size solvers** that adjust the number of steps dynamically during sampling.
- **Neural samplers** that learn efficient sampling trajectories conditioned on the target distribution [84].
- **Hybrid samplers** combining stochastic and deterministic steps for improved trade-offs [85].

These approaches could further reduce inference time while maintaining or even improving sample quality [86].

4.5. Efficient Conditional Diffusion Models

Many real-world applications, such as text-to-image generation (e.g., Stable Diffusion) and controllable synthesis, require conditioning mechanisms [87]. However, incorporating conditioning signals often increases model complexity [88]. Future directions could include:

- **Efficient cross-attention mechanisms** for reducing the cost of conditioning inputs [89].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

- **Sparse and structured conditioning** to optimize memory and computation usage.
- **Multi-modal conditioning** that balances flexibility and efficiency [90].

Improving conditional diffusion models could make them more practical for applications like real-time interactive generation [91].

4.6. Self-Distillation and Online Learning

Current distillation methods require training a separate student model, which adds extra overhead [92]. Self-distillation techniques, where a model distills its knowledge progressively during training, could be a promising direction [93]. Additionally, online learning frameworks that allow diffusion models to adapt to new data distributions without full retraining could improve efficiency in dynamic environments.

4.7. Hybrid Generative Frameworks

While diffusion models have shown superior sample quality compared to GANs and VAEs, combining them with other generative paradigms could lead to more efficient models [94]. Future research could explore:

- **Diffusion-GAN hybrids** that leverage adversarial training for faster sampling [95].
- **Energy-based diffusion models** that integrate energy-based learning to improve stability [96].
- **Latent-variable diffusion models** that operate in more compact representation spaces [97].

These hybrid approaches could bridge the gap between efficiency and high-quality generation [98].

4.8. Hardware Optimization and Deployment

Efficient deployment of diffusion models on real-world hardware remains an open challenge [99]. Future directions could include:

- **Custom hardware accelerators** optimized for diffusion-based generative models [100].
- **Optimization for mobile and edge devices** through model compression and pruning [101].
- **Efficient inference frameworks** such as TensorRT, ONNX, and specialized AI chips [102].

Optimizing diffusion models for hardware efficiency could significantly expand their applicability.

4.9. Theoretical Understanding of Efficiency in Diffusion Models

Most efficiency improvements have been driven by empirical findings rather than theoretical insights [103]. A deeper theoretical understanding of diffusion processes, sampling dynamics, and optimization techniques could guide the development of more principled and generalizable efficiency improvements [104]. Key research directions include:

- **Understanding the fundamental speed limits** of diffusion-based sampling [105].
- **Mathematical analysis of sampling trajectories** and optimal noise schedules [106].
- **Connections between diffusion models and other generative paradigms** for theoretical unification [107].

Developing a stronger theoretical foundation could lead to more efficient and interpretable diffusion models [108].

4.10. Summary of Future Directions

Table 2 summarizes key open challenges and future research opportunities [109].

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

Table 2. Summary of open challenges and future research directions in efficient diffusion models.

Challenge	Potential Research Directions
Sampling Speed vs [110]. Quality	Adaptive solvers, learned step sizes, hybrid samplers
Training Efficiency	Gradient checkpointing, mixed-precision, distributed training
Model Architecture	NAS, sparse networks, efficient transformers
Conditional Diffusion	Optimized attention, sparse conditioning, multi-modal inputs
Distillation and Learning	Self-distillation, progressive training, online adaptation
Hybrid Generative Models	GAN-diffusion hybrids, energy-based models, latent-space diffusion
Hardware and Deployment	AI accelerators, mobile optimization, inference frameworks
Theoretical Foundations	Optimal sampling theory, diffusion-GAN connections

Addressing these challenges will be crucial for making diffusion models more practical, scalable, and efficient [111]. In the next section, we conclude our survey with a summary of key findings and final thoughts [112].

5. Conclusion

Diffusion models have emerged as a powerful class of generative models, capable of producing high-quality samples across various domains, including images, audio, and video [113]. Their theoretical foundations, training stability, and expressive capabilities have made them a dominant choice for generative modeling [114]. However, their practical deployment remains challenging due to high computational costs, slow inference times, and memory-intensive training requirements [115]. In this survey, we explored various approaches aimed at improving the efficiency of diffusion models [116]. We categorized these methods into four major areas: (1) accelerated sampling techniques, (2) efficient model architectures, (3) knowledge distillation and model compression, and (4) hybrid generative frameworks. Accelerated sampling methods, such as DDIMs, higher-order solvers, and latent diffusion models, have significantly reduced the number of function evaluations required for generation [117]. Lightweight architectures and transformer-based diffusion models have provided alternative ways to optimize computational efficiency [118]. Knowledge distillation, pruning, and quantization techniques have enabled compression of large diffusion models into smaller, faster variants [119]. Finally, hybrid approaches that integrate diffusion models with GANs, VAEs, or energy-based methods have demonstrated promising results in improving both efficiency and sample quality [120]. Despite these advancements, several challenges remain open [121]. The trade-off between sampling speed and generation quality continues to be a major bottleneck, requiring more adaptive and theoretically grounded approaches [122]. Training diffusion models efficiently remains an expensive process, and further research into memory-efficient training techniques and distributed learning strategies is necessary [123]. Moreover, optimizing diffusion models for real-time deployment on mobile devices, edge computing systems, and specialized hardware accelerators is an exciting research direction that could expand their practical usability [124].

Looking forward, the field of efficient diffusion models is rapidly evolving, with new techniques emerging at a fast pace. Advances in theoretical understanding, novel architectural designs, and improved training paradigms will continue to push the boundaries of what is possible with diffusion-based generative models. We hope that this survey serves as a valuable resource for researchers and practitioners, inspiring further innovations that make diffusion models more accessible, scalable, and efficient.

References

1. 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.
2. Liu, H.; Li, C.; Li, Y.; Lee, Y.J. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* **2023**.
3. Fan, Y.; Lee, K. Optimizing DDPM Sampling with Shortcut Fine-Tuning. In Proceedings of the International Conference on Machine Learning, 2023, pp. 9623–9639.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

4. 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].
5. Buciluǎ, C.; Caruana, R.; Niculescu-Mizil, A. Model compression. In Proceedings of the Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, 2006, pp. 535–541.
6. Lu, C.; Zhou, Y.; Bao, F.; Chen, J.; Li, C.; Zhu, J. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models. *arXiv preprint arXiv:2211.01095* **2022**.
7. 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.
8. 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].
9. 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.
10. 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.
11. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
12. 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.
13. Fedus, W.; Zoph, B.; Shazeer, N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* **2022**, *23*, 1–39.
14. Du, D.; Gong, G.; Chu, X. Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey. *arXiv preprint arXiv:2405.00314* **2024**.
15. 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**.
16. Gu, A.; Goel, K.; Ré, C. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396* **2021**.
17. Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; Joulin, A. Emerging Properties in Self-Supervised Vision Transformers. In Proceedings of the Proceedings of the International Conference on Computer Vision (ICCV), 2021.
18. 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.
19. 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*.
20. Karras, T.; Aittala, M.; Laine, S.; Aila, T. Elucidating the design space of diffusion-based generative models. In Proceedings of the Proceedings of the 36th International Conference on Neural Information Processing Systems, 2022, pp. 26565–26577.
21. 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**.
22. Xu, S.; Li, Y.; Ma, T.; Zeng, B.; Zhang, B.; Gao, P.; Lv, J. TerViT: An efficient ternary vision transformer. *arXiv preprint arXiv:2201.08050* **2022**.
23. Song, Y.; Dhariwal, P.; Chen, M.; Sutskever, I. Consistency Models. In Proceedings of the International Conference on Machine Learning, 2023, pp. 32211–32252.
24. Dhariwal, P.; Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems* **2021**, *34*, 8780–8794.
25. 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.
26. 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

27. Jolicoeur-Martineau, A.; Piché-Taillefer, R.; Mitliagkas, I.; des Combes, R.T. Adversarial score matching and improved sampling for image generation. In Proceedings of the International Conference on Learning Representations, 2021.
28. 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.
29. 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*.
30. Li, Y.; Hu, J.; Wen, Y.; Evangelidis, G.; Salahi, K.; Wang, Y.; Tulyakov, S.; Ren, J. Rethinking vision transformers for mobilenet size and speed. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 16889–16900.
31. 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**.
32. 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**.
33. 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**.
34. Lin, J.; Yin, H.; Ping, W.; Lu, Y.; Molchanov, P.; Tao, A.; Mao, H.; Kautz, J.; Shoeybi, M.; Han, S. Vila: On pre-training for visual language models. *arXiv preprint arXiv:2312.07533* **2023**.
35. 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**.
36. 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**.
37. Li, Y.; Wang, C.; Jia, J. LLaMA-VID: An Image is Worth 2 Tokens in Large Language Models, 2023, [arXiv:cs.CV/2311.17043].
38. 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.
39. 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.
40. 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.
41. Xu, H.; Ye, Q.; Yan, M.; Shi, Y.; Ye, J.; Xu, Y.; Li, C.; Bi, B.; Qian, Q.; Wang, W.; et al. mPLUG-2: A Modularized Multi-modal Foundation Model Across Text, Image and Video, 2023, [arXiv:cs.CV/2302.00402].
42. 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.
43. Le, P.H.C.; Li, X. BinaryViT: pushing binary vision transformers towards convolutional models. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 4664–4673.
44. 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**.
45. Sauer, A.; Lorenz, D.; Blattmann, A.; Rombach, R. Adversarial diffusion distillation. *arXiv preprint arXiv:2311.17042* **2023**.
46. DeepSeek-AI. DeepSeek LLM: Scaling Open-Source Language Models with Longtermism. *arXiv preprint arXiv:2401.02954* **2024**.
47. 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**.
48. Sidorov, O.; Hu, R.; Rohrbach, M.; Singh, A. Textcaps: a dataset for image captioning with reading comprehension. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16. Springer, 2020, pp. 742–758.
49. 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.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

50. 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**.
51. 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**.
52. 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**.
53. 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**.
54. 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.
55. 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.
56. Shi, B.; Wu, Z.; Mao, M.; Wang, X.; Darrell, T. When Do We Not Need Larger Vision Models? *arXiv preprint arXiv:2403.13043* **2024**.
57. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
58. Chen, C.; Borgeaud, S.; Irving, G.; Lespiau, J.B.; Sifre, L.; Jumper, J. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318* **2023**.
59. 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.
60. Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502* **2023**.
61. Shao, Z.; Ouyang, X.; Gai, Z.; Yu, Z.; Yu, J. Imp: An emprical study of multimodal small language models, 2024.
62. 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**.
63. Ordonez, V.; Kulkarni, G.; Berg, T. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems* **2011**, *24*.
64. 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.
65. 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.
66. 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.
67. 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**.
68. 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**.
69. 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**.
70. 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**.
71. He, Y.; Lou, Z.; Zhang, L.; Liu, J.; Wu, W.; Zhou, H.; Zhuang, B. BiViT: Extremely Compressed Binary Vision Transformers. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 5651–5663.
72. Luhman, E.; Luhman, T. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388* **2021**.
73. 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.

74. 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**.
75. 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.
76. Salimans, T.; Ho, J. Progressive Distillation for Fast Sampling of Diffusion Models. In Proceedings of the International Conference on Learning Representations, 2022.
77. 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.
78. Kingma, D.P.; Salimans, T.; Poole, B.; Ho, J. Variational diffusion models. In Proceedings of the Proceedings of the 35th International Conference on Neural Information Processing Systems, 2021, pp. 21696–21707.
79. Watson, D.; Ho, J.; Norouzi, M.; Chan, W. Learning to efficiently sample from diffusion probabilistic models. *arXiv preprint arXiv:2106.03802* **2021**.
80. 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.
81. 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*.
82. Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; Huang, J. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204* **2024**.
83. Li, W.; Wang, X.; Xia, X.; Wu, J.; Xiao, X.; Zheng, M.; Wen, S. Sepvit: Separable vision transformer. *arXiv preprint arXiv:2203.15380* **2022**.
84. Zhao, B.; Wu, B.; Huang, T. Svit: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087* **2023**.
85. 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**.
86. Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models, 2024, [arXiv:cs.CV/2306.13394].
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. 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**.
89. 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**.
90. 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.
91. Song, J.; Meng, C.; Ermon, S. Denoising Diffusion Implicit Models. In Proceedings of the International Conference on Learning Representations, 2021.
92. Chen, X.; Cao, Q.; Zhong, Y.; Zhang, J.; Gao, S.; Tao, D. Dearkd: Data-efficient early knowledge distillation for vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12052–12062.
93. Luo, G.; Zhou, Y.; Ren, T.; Chen, S.; Sun, X.; Ji, R. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems* **2024**, *36*.
94. 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**.
95. Xiao, J.; Li, Z.; Yang, L.; Gu, Q. BinaryViT: Towards Efficient and Accurate Binary Vision Transformers. *arXiv preprint arXiv:2305.14730* **2023**.
96. 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.
97. Zhu, B.; Lin, B.; Ning, M.; Yan, Y.; Cui, J.; Wang, H.; Pang, Y.; Jiang, W.; Zhang, J.; Li, Z.; et al. LanguageBind: Extending Video-Language Pretraining to N-modality by Language-based Semantic Alignment. *arXiv preprint arXiv:2310.01852* **2023**.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

98. Gong, C.; Wang, D. NASViT: Neural architecture search for efficient vision transformers with gradient conflict-aware supernet training. *ICLR Proceedings 2022* **2022**.
99. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* **2017**, *123*, 32–73.
100. 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.
101. Fang, A.; Jose, A.M.; Jain, A.; Schmidt, L.; Toshev, A.; Shankar, V. Data filtering networks. *arXiv preprint arXiv:2309.17425* **2023**.
102. 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.
103. 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.
104. Cha, J.; Kang, W.; Mun, J.; Roh, B. Honeybee: Locality-enhanced projector for multimodal llm. *arXiv preprint arXiv:2312.06742* **2023**.
105. Wan, Z.; Wang, X.; Liu, C.; Alam, S.; Zheng, Y.; Liu, J.; Qu, Z.; Yan, S.; Zhu, Y.; Zhang, Q.; et al. Efficient Large Language Models: A Survey, 2024, [arXiv:cs.CL/2312.03863].
106. 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.
107. 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].
108. Byeon, M.; Park, B.; Kim, H.; Lee, S.; Baek, W.; Kim, S. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset, 2022.
109. Zhang, H.; Li, X.; Bing, L. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858* **2023**.
110. 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*.
111. 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**.
112. 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*.
113. 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.
114. 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**.
115. Gagrani, M.; Goel, R.; Jeon, W.; Park, J.; Lee, M.; Lott, C. On Speculative Decoding for Multimodal Large Language Models, 2024, [arXiv:cs.CL/2404.08856].
116. 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].
117. 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**.
118. Dong, P.; Lu, L.; Wu, C.; Lyu, C.; Yuan, G.; Tang, H.; Wang, Y. PackQViT: Faster Sub-8-bit Vision Transformers via Full and Packed Quantization on the Mobile. *Advances in Neural Information Processing Systems* **2024**, *36*.
119. 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**.
120. 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.
121. 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**.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 11 February 2025 doi:10.20944/preprints202502.0779.v1

122. 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.
123. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752* **2023**.
124. 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].

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.