Navigating the Alignment Challenges of Diffusion Models: Insights and Innovations

Essam Jazibiyya , Rasim Dina *, Wasim Ismaeel

Posted Date: 21 January 2025

doi: 10.20944/preprints202501.1502.v1

Keywords: Diffusion Models, AI Alignment; Generative Models; Ethical AI; Fairness in AI; Bias Mitigation; Reinforcement Learning from Human Feedback (RLHF); Model Interpretability; Responsible AI; Safe AI Deployment; Controllability in AI; Alignment Challenges; Stochastic Models; Generative Adversarial Networks (GANs); AI Governance; Human Values in AI; Data Bias; Interdisciplinary Collaboration; AI Ethics; Future of Generative AI

 doi:10.20944/preprints202501.1502.v1

Article

Navigating the Alignment Challenges of Diffusion Models: Insights and Innovations

Essam Jazibiyya, Rasim Dina * and Wasim Ismaeel

Kingdom of Saudi Arabia, KAUST, King Abdullah University of Science and Technology; essam.jazibiyya@kaust.edu.sa (E.J.); wasim.ismaeel@kaust.edu.sa (W.I.)

* Correspondence: rasim.dina@kaust.edu.sa

Abstract: Diffusion models have emerged as a powerful class of generative models, revolutionizing fields such as image synthesis, text-to-image generation, and molecular design. Despite their remarkable capabilities, ensuring that these models are aligned with human values, ethical principles, and societal goals remains a significant challenge. The alignment problem in diffusion models encompasses issues such as safety, fairness, robustness, and controllability, compounded by the stochastic and generative nature of these models. This paper provides a comprehensive exploration of the alignment of diffusion models, beginning with an overview of their foundational principles and applications. We examine the unique challenges posed by their probabilistic outputs, lack of interpretability, and dependence on large-scale, often biased datasets. Existing approaches to alignment, including fine-tuning, reinforcement learning from human feedback, prompt engineering, and post-processing, are analyzed for their strengths and limitations. Building on this foundation, we identify key research gaps and propose future directions, such as the development of scalable alignment techniques, robust evaluation metrics, and interdisciplinary collaboration frameworks. We also highlight the importance of addressing ethical and societal considerations, including bias mitigation, transparency, and equitable access, to ensure the responsible deployment of diffusion models. By addressing these challenges, we aim to foster a new era of generative AI systems that are not only innovative and powerful but also aligned with the values and aspirations of humanity. This work serves as a foundation for advancing the alignment of diffusion models, inspiring further research and collaboration in this critical domain.

Keywords: diffusion models, AI alignment; generative models; ethical AI; fairness in AI; bias mitigation; reinforcement learning from human feedback (RLHF); model interpretability; responsible AI; safe AI deployment; controllability in AI; alignment challenges; stochastic models; generative adversarial networks (GANs); AI governance; human values in AI; data bias; interdisciplinary collaboration; AI ethics; future of generative AI

1. Introduction

In recent years, diffusion models have emerged as one of the most promising paradigms in generative modeling, demonstrating unprecedented capabilities across a wide range of domains. From photorealistic image synthesis and high-fidelity audio generation to advancements in natural language processing, diffusion models have showcased their potential to redefine the boundaries of generative artificial intelligence [1]. These models operate by iteratively denoising samples from a predefined noise distribution, gradually transforming them into data points that closely resemble the target distribution. This iterative process, inspired by principles of stochastic differential equations and thermodynamics, allows diffusion models to capture intricate data distributions with remarkable precision [2]. The rapid advancements in diffusion models have unlocked numerous opportunities for innovation across industries [3]. In creative fields, they are enabling artists and designers to generate content with unprecedented speed and quality. In scientific research, they are being leveraged for tasks such as molecular design and protein structure prediction [4]. In entertainment, they are transforming how games, animations, and interactive media are created [5]. Despite these successes, the increasing

 doi:10.20944/preprints202501.1502.v1

ubiquity of diffusion models has also raised critical questions about their ethical implications, societal impact, and the mechanisms by which they can be aligned with human values and priorities [6]. Alignment, in the context of machine learning and artificial intelligence, refers to the process of ensuring that a model's outputs and behaviors are consistent with predefined objectives, ethical principles, and societal norms. While alignment has been extensively studied in the context of supervised and reinforcement learning, the probabilistic and generative nature of diffusion models introduces unique challenges [7]. These models are not merely tools for classification or decision-making; they are capable of generating highly realistic, diverse, and creative outputs that can influence human perceptions, decisions, and behaviors [8]. This capability amplifies both their potential benefits and their risks [9]. One of the most pressing challenges in aligning diffusion models is the inherent complexity of controlling their outputs [10]. Unlike deterministic models, which produce predictable results for a given input, diffusion models rely on stochastic processes to sample from learned distributions [11]. This stochasticity makes it difficult to guarantee that the generated outputs will consistently adhere to desired constraints or avoid harmful content [12]. Moreover, the iterative nature of diffusion models introduces additional layers of complexity, as small deviations in early stages of the generation process can propagate and amplify in later stages, leading to unintended outcomes [13]. Another significant challenge lies in addressing the biases and ethical concerns associated with diffusion models [14]. Like other machine learning models, diffusion models are trained on large-scale datasets that often contain implicit biases, stereotypes, and harmful content. These biases can be inadvertently learned and reproduced by the models, leading to outputs that reinforce existing inequalities or propagate harmful narratives [15]. Furthermore, the black-box nature of many diffusion models complicates efforts to audit, interpret, and mitigate these biases, creating a need for more transparent and accountable methodologies. The societal implications of diffusion models extend beyond technical challenges [16]. As these models become increasingly integrated into critical applications, such as healthcare, education, and public policy, their potential to influence human lives and societal structures grows [17]. Ensuring that diffusion models are aligned with ethical principles and societal goals is not merely a technical problem but also a multidisciplinary endeavor that requires input from ethicists, policymakers, social scientists, and other stakeholders [18]. Issues such as trust, accountability, and governance must be addressed to ensure that these technologies are deployed responsibly and equitably. In this paper, we aim to provide a comprehensive exploration of the alignment problem in diffusion models, encompassing its technical, ethical, and societal dimensions. We begin by providing a detailed overview of diffusion models, including their mathematical foundations, training paradigms, and applications. This foundational understanding sets the stage for an in-depth discussion of the alignment challenges unique to these models [19]. We categorize these challenges into three broad areas: technical challenges, such as robustness, controllability, and interpretability; ethical challenges, including bias mitigation, fairness, and transparency; and societal challenges, such as trust, governance, and misuse prevention. To address these challenges, we propose a research agenda that highlights key directions for future work [20]. These include the development of novel fine-tuning techniques to enhance alignment, the integration of explicit alignment objectives into training pipelines, and the design of robust evaluation metrics to assess alignment performance. We also emphasize the importance of interdisciplinary collaboration, advocating for the inclusion of diverse perspectives and expertise in shaping the future of diffusion model research and deployment [21]. By examining the alignment problem in diffusion models through a holistic lens, this paper seeks to contribute to the broader discourse on responsible AI development [22]. Our goal is to inspire researchers, practitioners, and policymakers to prioritize alignment as a central consideration in the development and deployment of diffusion models. Ultimately, we envision a future where diffusion models are not only powerful and innovative but also aligned with the values and aspirations of humanity, driving progress in ways that are ethical, equitable, and beneficial for all [23].

 doi:10.20944/preprints202501.1502.v1

2. Background on Diffusion Models

Diffusion models are a class of generative models that have gained significant attention for their ability to model complex data distributions and generate high-quality outputs. At their core, diffusion models operate by learning a reverse process that transforms noisy data back into samples from the target distribution [24]. This process is inspired by principles of stochastic processes and thermodynamics, specifically leveraging ideas from Gaussian noise perturbation and denoising [25].

2.1. Mathematical Foundations

The foundation of diffusion models lies in the forward and reverse processes. The forward process involves gradually adding noise to data samples over a series of time steps, effectively transforming them into pure noise. Mathematically, the forward process is modeled as a Markov chain:

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

where $\mathbf{x}_t$ represents the data at time step t, β_t is a noise variance schedule, and $\mathcal{N}$ denotes a Gaussian distribution [26]. The reverse process, which the model learns, involves denoising the noisy samples step by step to recover the original data distribution. The reverse process is parameterized as:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)), \tag{2}$$

where μ_θ and Σ_θ are neural network-based functions that approximate the mean and variance of the reverse transition [27]. Training diffusion models involves optimizing a variational lower bound (VLB) on the data likelihood [28]. This objective ensures that the model accurately learns the reverse process, enabling it to generate high-quality samples by iteratively denoising random noise.

2.2. Applications of Diffusion Models

Diffusion models have demonstrated remarkable success across various domains:

- **Image Synthesis:** Models like Denoising Diffusion Probabilistic Models (DDPMs) and improved variants such as Stable Diffusion have set new benchmarks in generating photorealistic images.
- **Text-to-Image Generation:** Systems like DALLE-2 and Imagen leverage diffusion processes to generate high-quality images conditioned on textual descriptions, opening new possibilities for multimodal AI applications.
- **Audio Processing:** Diffusion models have been applied to tasks such as speech synthesis, music generation, and audio denoising, showcasing their versatility in temporal data [29].
- **Scientific Applications:** In fields such as chemistry and biology, diffusion models are being used for molecular design, drug discovery, and protein structure prediction, demonstrating their potential to accelerate scientific innovation.

2.3. Challenges in Diffusion Models

Despite their success, diffusion models present unique challenges that complicate their alignment:

- **Stochastic Nature:** The probabilistic nature of diffusion models makes it difficult to control their outputs with precision, leading to challenges in ensuring alignment with specific objectives [30].
- **Computational Complexity:** The iterative denoising process is computationally expensive, posing challenges for scalability and deployment in resource-constrained environments [31].
- **Bias and Fairness:** Like other machine learning models, diffusion models are susceptible to biases in training data, which can propagate to their outputs, raising ethical concerns [32].
- **Evaluation Metrics:** Assessing the quality and alignment of generated outputs remains an open problem, as existing metrics often fail to capture nuanced aspects of alignment and human values.

This background serves as a foundation for understanding the challenges and opportunities associated with aligning diffusion models [33]. In the following sections, we delve deeper into the alignment problem, exploring its technical, ethical, and societal dimensions.

 doi:10.20944/preprints202501.1502.v1

3. The Alignment Problem in Diffusion Models

As diffusion models become increasingly prevalent in real-world applications, ensuring their alignment with human values, ethical principles, and societal goals has become a critical area of research. The alignment problem refers to the challenge of guiding these models to consistently produce outputs that are safe, unbiased, and aligned with intended objectives, while avoiding harmful or unintended consequences. This section explores the unique aspects of the alignment problem in the context of diffusion models [34].

3.1. Definition and Scope of Alignment

Alignment in diffusion models involves ensuring that their generative outputs meet specific criteria, such as:

- **Safety:** Preventing the generation of harmful, offensive, or inappropriate content [35].
- **Fairness:** Avoiding biases related to race, gender, culture, or other sensitive attributes.
- **Robustness:** Ensuring consistent behavior across diverse inputs and conditions.
- **Controllability:** Providing mechanisms to steer the model's outputs toward desired objectives or constraints [36].

Unlike alignment in deterministic models, which often focuses on task-specific objectives, alignment in diffusion models must account for their probabilistic and generative nature [37]. This introduces unique challenges in defining, measuring, and enforcing alignment.

3.2. Challenges in Aligning Diffusion Models

The alignment problem in diffusion models is shaped by several key challenges, spanning technical, ethical, and societal dimensions [38].

3.2.1. Technical Challenges

Stochasticity and Diversity of Outputs

Diffusion models generate outputs by sampling from a learned distribution, resulting in inherent variability [39]. While this stochasticity enables creative and diverse outputs, it also makes it difficult to ensure that all possible outputs adhere to alignment objectives. Small changes in noise initialization or sampling parameters can lead to significant variations in the generated content [40].

Lack of Interpretability.

The complex architecture and iterative nature of diffusion models make it challenging to interpret how specific features in the training data influence the generated outputs [41]. This lack of transparency hinders efforts to diagnose and address alignment failures [42].

Data Dependence.

Diffusion models are trained on large-scale datasets that often contain biases, stereotypes, and harmful content [43]. These issues can propagate into the model's outputs, necessitating robust mechanisms to detect and mitigate such biases during and after training.

3.2.2. Ethical Challenges

Bias and Fairness

Bias in training data can result in outputs that reinforce stereotypes or marginalize certain groups [44]. For example, text-to-image diffusion models may generate stereotypical depictions of professions or cultural symbols based on biased training data [45].

 doi:10.20944/preprints202501.1502.v1

Harmful Content Generation.

The ability of diffusion models to generate highly realistic content raises concerns about their potential misuse, such as creating fake images, deepfakes, or misinformation [46]. Ensuring that these models do not contribute to societal harm is a critical aspect of alignment.

3.2.3. Societal Challenges

Trust and Accountability

As diffusion models are integrated into applications that impact human lives, such as healthcare or education, ensuring public trust becomes paramount. Users must have confidence that these models operate in ways that are fair, transparent, and aligned with societal values.

Governance and Regulation

The widespread deployment of diffusion models necessitates the development of governance frameworks and regulatory policies to ensure responsible use. These frameworks must balance innovation with accountability, addressing concerns about misuse and unintended consequences [47].

3.3. Existing Approaches to Alignment

Efforts to align diffusion models have drawn inspiration from various techniques, including:

- **Fine-Tuning:** Adapting pre-trained models to specific tasks or domains using curated datasets that emphasize alignment objectives [48].
- **Prompt Engineering:** Designing prompts or input conditions that guide the model's outputs toward desired behaviors.
- **Post-Processing:** Applying filters or constraints to generated outputs to enforce alignment criteria.
- **Reinforcement Learning from Human Feedback (RLHF):** Leveraging human feedback to iteratively refine model behavior and improve alignment [49].

While these approaches have shown promise, they often address specific aspects of alignment and may not generalize across all use cases [50]. A holistic and interdisciplinary approach is needed to address the multifaceted nature of the alignment problem in diffusion models.

3.4. Research Gaps and Open Questions

Despite progress, several open questions remain in the alignment of diffusion models:

- How can alignment objectives be formally defined and measured in the context of generative models?
- What are the trade-offs between alignment, creativity, and diversity in diffusion model outputs [51]?
- How can alignment techniques be scaled to large, complex models without compromising computational efficiency [52–54]?
- What role should policymakers, ethicists, and other stakeholders play in shaping the alignment of diffusion models?

Addressing these questions requires collaboration across technical, ethical, and societal domains [55]. The following sections explore potential solutions and future directions for advancing the alignment of diffusion models.

4. Approaches to Addressing the Alignment Problem

Addressing the alignment problem in diffusion models requires a combination of technical innovations, ethical considerations, and interdisciplinary collaboration. This section explores various approaches that have been proposed or implemented to align diffusion models with desired objectives and societal values.

 doi:10.20944/preprints202501.1502.v1

4.1. Technical Approaches

4.1.1. Fine-Tuning and Domain Adaptation

Fine-tuning involves adapting pre-trained diffusion models to specific tasks or domains using carefully curated datasets [56]. By incorporating alignment objectives into the fine-tuning process, researchers can guide models to generate outputs that adhere to ethical and societal constraints. Domain adaptation techniques further refine models for specialized applications, ensuring alignment with context-specific requirements [57].

4.1.2. Reinforcement Learning from Human Feedback (RLHF)

Reinforcement Learning from Human Feedback (RLHF) has emerged as a powerful tool for aligning generative models [58]. In this approach, human feedback is used to reward or penalize model outputs, shaping the model's behavior over successive iterations [59]. For diffusion models, RLHF can be employed to fine-tune the denoising process, ensuring that the generated outputs meet alignment criteria.

4.1.3. Prompt Engineering and Conditioning

Prompt engineering involves designing input prompts that steer the model's outputs toward desired behaviors [60]. For diffusion models, this can include conditioning the model on specific attributes, such as style, content, or ethical constraints. Techniques such as classifier guidance or latent space conditioning allow for more granular control over the generation process.

4.1.4. Bias Mitigation Techniques

To address biases inherent in training data, several techniques have been proposed:

- **Data Filtering:** Removing or reweighting biased samples in the training dataset.
- **Adversarial Training:** Introducing adversarial objectives to penalize biased outputs during training [61].
- **Fairness Constraints:** Incorporating fairness objectives directly into the model's loss function.

These techniques aim to reduce the propagation of biases while preserving the model's generative capabilities [62].

4.1.5. Post-Processing and Output Filtering

Post-processing involves applying filters or constraints to the model's outputs after generation [63]. For example, content moderation systems can detect and remove harmful or inappropriate outputs [64]. While effective, this approach is reactive and may not address the root causes of misalignment.

4.2. Ethical and Societal Approaches

4.2.1. Ethics-by-Design

Ethics-by-design emphasizes the integration of ethical considerations throughout the development lifecycle of diffusion models [65]. This approach involves proactively identifying potential risks and embedding safeguards into the model's architecture, training process, and deployment strategy [66].

4.2.2. Stakeholder Engagement

Engaging stakeholders, including ethicists, policymakers, industry practitioners, and end-users, is critical for ensuring that alignment efforts reflect diverse perspectives and societal needs [67]. Collaborative frameworks can help identify alignment objectives, evaluate trade-offs, and build trust in diffusion model applications [68].

4.2.3. Governance and Regulation

Governance frameworks and regulatory policies play a crucial role in ensuring the responsible use of diffusion models. These frameworks should address issues such as accountability, transparency,

 doi:10.20944/preprints202501.1502.v1

and the prevention of misuse [69]. For example, certification programs or compliance standards could be established to evaluate the alignment of diffusion models before deployment [70].

4.2.4. Public Awareness and Education

Raising public awareness about the capabilities and limitations of diffusion models is essential for fostering informed discussions about their alignment [71]. Educational initiatives can help users understand the risks and benefits of these models, empowering them to make informed decisions about their use.

4.3. Interdisciplinary Collaboration

Aligning diffusion models requires expertise from multiple disciplines, including computer science, ethics, law, sociology, and psychology [72]. Interdisciplinary collaboration can help address complex challenges, such as defining alignment objectives, evaluating societal impact, and designing governance mechanisms. Collaborative research initiatives and cross-sector partnerships can drive innovation and ensure that alignment efforts are both practical and impactful [73].

4.4. Emerging Research Directions

Several emerging research directions hold promise for advancing the alignment of diffusion models:

- **Differentiable Alignment Objectives:** Developing alignment objectives that can be directly optimized during training, enabling seamless integration with the model's learning process.
- **Interactive Alignment Tools:** Creating tools that allow users to interactively guide and evaluate model outputs, enhancing transparency and control [74].
- **Scalable Alignment Techniques:** Designing alignment methods that scale efficiently to large models and datasets, reducing computational overhead [75].
- **Alignment Benchmarks:** Establishing standardized benchmarks and metrics for evaluating the alignment of diffusion models across diverse applications [76].

4.5. Limitations of Current Approaches

While the approaches outlined above represent significant progress, they are not without limitations:

- Many techniques focus on specific aspects of alignment and may not generalize across all use cases [77].
- Trade-offs between alignment, creativity, and diversity remain poorly understood.
- The computational cost of alignment methods can be prohibitive for large-scale models [78].
- Ethical and societal considerations are often treated as secondary to technical objectives, leading to misaligned priorities.

Addressing these limitations requires a concerted effort to develop holistic, scalable, and interdisciplinary solutions. The next section explores future directions and open challenges in achieving aligned and responsible diffusion models.

5. Future Directions and Open Challenges

The alignment of diffusion models remains a dynamic and evolving field, with numerous opportunities for innovation and exploration. This section highlights key future directions and open challenges that must be addressed to ensure the responsible development and deployment of diffusion models [79].

 doi:10.20944/preprints202501.1502.v1

5.1. Advancing Alignment Techniques

5.1.1. Unified Alignment Frameworks

A critical need exists for the development of unified frameworks that integrate alignment objectives into all stages of the diffusion model pipeline, from data preprocessing to post-generation filtering [80]. Such frameworks should enable seamless incorporation of ethical, technical, and societal considerations, fostering a holistic approach to alignment.

5.1.2. Scalable Alignment Solutions

As diffusion models grow in size and complexity, the computational cost of alignment techniques becomes a significant challenge. Future research should focus on designing scalable methods that can efficiently align large-scale models without compromising performance or accessibility [81].

5.1.3. Dynamic and Adaptive Alignment

Static alignment techniques may struggle to address evolving societal norms, emerging risks, or novel use cases. Dynamic and adaptive alignment methods, capable of updating models in response to new information or feedback, represent an important research direction [82]. This could involve continuous learning mechanisms or modular architectures that allow for flexible adjustments [83].

5.2. Improving Evaluation and Metrics

5.2.1. Alignment Benchmarks

The lack of standardized benchmarks for evaluating alignment in diffusion models hinders progress and comparability across studies [84]. Developing robust benchmarks that capture diverse alignment objectives, such as fairness, safety, and controllability, is essential for advancing the field [85].

5.2.2. Human-Centric Evaluation

Quantitative metrics often fail to capture the nuanced and subjective aspects of alignment, such as cultural sensitivity or ethical considerations. Incorporating human-centric evaluation methods, such as user studies or expert reviews, can provide richer insights into model behavior and alignment quality [86].

5.3. Mitigating Bias and Ethical Risks

5.3.1. Bias-Aware Training Pipelines

Future work should explore the design of training pipelines that explicitly account for biases in the data and model. This could involve techniques such as bias-aware loss functions, adversarial debiasing, or the integration of fairness constraints during training.

5.3.2. Ethical Auditing and Transparency

Developing tools and methodologies for auditing diffusion models is critical for ensuring accountability and transparency [87]. Ethical auditing frameworks should enable stakeholders to assess the alignment of models with societal values and regulatory requirements.

5.4. Interdisciplinary Collaboration

5.4.1. Cross-Disciplinary Research Initiatives

Aligning diffusion models is not solely a technical problem but a multidisciplinary challenge. Future efforts should prioritize collaboration between computer scientists, ethicists, social scientists, and policymakers [88]. Cross-disciplinary research initiatives can help address complex issues, such as defining alignment objectives, evaluating societal impacts, and designing governance mechanisms [89].

 doi:10.20944/preprints202501.1502.v1

5.4.2. Policy and Governance Frameworks

The development of policy and governance frameworks tailored to diffusion models is a pressing need. These frameworks should address issues such as accountability, misuse prevention, and equitable access, ensuring that diffusion models are deployed responsibly and ethically [90].

5.5. Addressing Societal and Global Impacts

5.5.1. Global Equity in AI Deployment

Diffusion models have the potential to exacerbate existing inequalities if their benefits are not equitably distributed [91]. Future research should explore strategies for promoting global equity in the development and deployment of diffusion models, such as open-access initiatives or capacity-building programs in underserved regions [92].

5.5.2. Long-Term Risks and Safety

The long-term risks associated with diffusion models, including their potential misuse or unintended societal consequences, warrant careful consideration [93]. Research into robust safety mechanisms, fail-safe systems, and long-term monitoring strategies is essential for mitigating these risks [94].

5.6. Open Questions

Despite progress, several open questions remain:

- How can alignment objectives be formalized in a way that balances technical feasibility with ethical and societal considerations [95]?
- What are the trade-offs between alignment, creativity, and diversity in generative outputs, and how can these be optimized?
- How can alignment techniques be adapted to new and emerging applications of diffusion models?
- What role should international collaboration play in developing standards and policies for aligned diffusion models [96]?

5.7. Vision for the Future

The alignment of diffusion models represents a critical frontier in the development of responsible AI [97]. By addressing the challenges and exploring the opportunities outlined in this section, researchers and practitioners can ensure that diffusion models are not only powerful and innovative but also aligned with the values and aspirations of humanity [98]. Achieving this vision requires a commitment to interdisciplinary collaboration, ethical responsibility, and the pursuit of equitable and inclusive AI systems [99]. In the next section, we summarize the key insights and contributions of this work, reflecting on the broader implications of aligning diffusion models for the future of AI [100].

6. Conclusion

The rapid advancements in diffusion models have opened new frontiers in generative AI, enabling applications that span creative industries, scientific research, and beyond. However, these advancements come with significant responsibilities. Ensuring that diffusion models are aligned with human values, ethical principles, and societal goals is a critical challenge that demands urgent attention.

In this work, we have explored the fundamentals of diffusion models, highlighting their unique characteristics and widespread applications. We have delved into the alignment problem, identifying the technical, ethical, and societal challenges that arise in ensuring that these models generate outputs that are safe, fair, and aligned with intended objectives. By examining existing approaches, including fine-tuning, reinforcement learning, prompt engineering, and ethical frameworks, we have outlined the progress made in addressing alignment issues while acknowledging the limitations of current methods.

 doi:10.20944/preprints202501.1502.v1

Looking forward, we have identified key future directions for research, emphasizing the need for scalable alignment techniques, robust evaluation metrics, and interdisciplinary collaboration. The development of unified alignment frameworks, dynamic and adaptive methods, and global governance mechanisms will be essential for addressing the complexities of aligning diffusion models. Additionally, we have underscored the importance of promoting equity, transparency, and accountability in the deployment of these models to ensure their benefits are distributed fairly and their risks are mitigated effectively.

The alignment of diffusion models represents not only a technical challenge but also a profound societal opportunity. By aligning these models with human values and aspirations, we can unlock their full potential to drive innovation, enhance creativity, and address pressing global challenges. Achieving this vision will require a concerted effort from researchers, practitioners, policymakers, and communities, fostering a future where generative AI serves as a force for good.

In conclusion, the alignment of diffusion models is a multifaceted and evolving field that holds the promise of transforming the relationship between AI and society. By addressing the challenges and seizing the opportunities presented by this field, we can pave the way for a new era of responsible and impactful AI systems. We hope that this work serves as a foundation for further exploration and inspires collective efforts to align diffusion models with the values that define our shared humanity.

References

1. Li, Z.; Yang, B.; Liu, Q.; Ma, Z.; Zhang, S.; Yang, J.; Sun, Y.; Liu, Y.; Bai, X. Monkey: Image Resolution and Text Label Are Important Things for Large Multi-modal Models, 2024, [arXiv:cs.CV/2311.06607].
2. 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.
3. Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A.C.; Lo, W.Y.; et al. Segment anything. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 4015–4026.
4. 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**.
5. 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.
6. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *International journal of computer vision* **2015**, *115*, 211–252.
7. 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**.
8. 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**.
9. 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].
10. Jie, S.; Tang, Y.; Ding, N.; Deng, Z.H.; Han, K.; Wang, Y. Memory-Space Visual Prompting for Efficient Vision-Language Fine-Tuning, 2024, [arXiv:cs.CV/2405.05615].
11. Zhang, Q.; Tao, M.; Chen, Y. gDDIM: Generalized denoising diffusion implicit models. In Proceedings of the International Conference on Learning Representations, 2023.
12. 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**.
13. Shazeer, N. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150* **2019**.
14. 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**.
15. 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.

 doi:10.20944/preprints202501.1502.v1

16. 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.
17. 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.
18. 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**.
19. Jiang, A.Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D.S.; de las Casas, D.; Hanna, E.B.; Bressand, F.; et al. Mixtral of Experts, 2024, [arXiv:cs.LG/2401.04088].
20. 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.
21. 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.
22. 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**.
23. 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].
24. Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Muyan, Z.; Zhang, Q.; Zhu, X.; Lu, L.; et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238* **2023**.
25. Wang, H.; Wang, Y.; Ye, Y.; Nie, Y.; Huang, C. Elysium: Exploring Object-level Perception in Videos via MLLM, 2024, [arXiv:cs.CV/2403.16558].
26. Hudson, D.A.; Manning, C.D. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 6700–6709.
27. 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**.
28. 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**.
29. Zhang, Q.; Chen, Y. Fast Sampling of Diffusion Models with Exponential Integrator. In Proceedings of the International Conference on Learning Representations, 2023.
30. 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**.
31. Zniyed, Y.; Nguyen, T.P.; et al. Enhanced network compression through tensor decompositions and pruning. *IEEE Transactions on Neural Networks and Learning Systems* **2024**.
32. 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.
33. 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**.
34. Hao, Z.; Guo, J.; Jia, D.; Han, K.; Tang, Y.; Zhang, C.; Hu, H.; Wang, Y. Learning efficient vision transformers via fine-grained manifold distillation. *Advances in Neural Information Processing Systems* **2022**, *35*, 9164–9175.
35. 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.
36. 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**.
37. 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.

 doi:10.20944/preprints202501.1502.v1

38. Fan, Y.; Lee, K. Optimizing DDPM Sampling with Shortcut Fine-Tuning. In Proceedings of the International Conference on Machine Learning, 2023, pp. 9623–9639.
39. 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].
40. 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**.
41. 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**.
42. 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**.
43. 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**.
44. Wang, G.; Liu, J.; Li, C.; Ma, J.; Zhang, Y.; Wei, X.; Zhang, K.; Chong, M.; Zhang, R.; Liu, Y.; et al. Cloud-Device Collaborative Learning for Multimodal Large Language Models. *arXiv preprint arXiv:2312.16279* **2023**.
45. 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].
46. 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.
47. Fang, A.; Jose, A.M.; Jain, A.; Schmidt, L.; Toshev, A.; Shankar, V. Data filtering networks. *arXiv preprint arXiv:2309.17425* **2023**.
48. 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.
49. Papa, L.; Russo, P.; Amerini, I.; Zhou, L. A Survey on Efficient Vision Transformers: Algorithms, Techniques, and Performance Benchmarking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **2024**, p. 1–20. https://doi.org/10.1109/tpami.2024.3392941.
50. 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**.
51. 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**.
52. 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**.
53. Zniyed, Y.; Nguyen, T.P.; et al. Efficient tensor decomposition-based filter pruning. *Neural Networks* **2024**, *178*, 106393.
54. 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*.
55. 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].
56. Zhu, M.; Zhu, Y.; Liu, X.; Liu, N.; Xu, Z.; Shen, C.; Peng, Y.; Ou, Z.; Feng, F.; Tang, J. A Comprehensive Overhaul of Multimodal Assistant with Small Language Models. *arXiv preprint arXiv:2403.06199* **2024**.
57. 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.
58. Li, Y.; Xu, S.; Zhang, B.; Cao, X.; Gao, P.; Guo, G. Q-vit: Accurate and fully quantized low-bit vision transformer. *Advances in neural information processing systems* **2022**, *35*, 34451–34463.
59. Shi, B.; Wu, Z.; Mao, M.; Wang, X.; Darrell, T. When Do We Not Need Larger Vision Models? *arXiv preprint arXiv:2403.13043* **2024**.
60. 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.
61. Du, D.; Gong, G.; Chu, X. Model Quantization and Hardware Acceleration for Vision Transformers: A Comprehensive Survey. *arXiv preprint arXiv:2405.00314* **2024**.

62. Pan, Z.; Zhuang, B.; Huang, D.A.; Nie, W.; Yu, Z.; Xiao, C.; Cai, J.; Anandkumar, A. T-stitch: Accelerating sampling in pre-trained diffusion models with trajectory stitching. *arXiv preprint arXiv:2402.14167* **2024**.
63. 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**.
64. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451* **2020**.
65. 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.
66. 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.
67. 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**.
68. Zhang, J.; Peng, H.; Wu, K.; Liu, M.; Xiao, B.; Fu, J.; Yuan, L. Minivit: Compressing vision transformers with weight multiplexing. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12145–12154.
69. 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.
70. 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.
71. Esser, P.; Kulal, S.; Blattmann, A.; Entezari, R.; Müller, J.; Saini, H.; Levi, Y.; Lorenz, D.; Sauer, A.; Boesel, F.; et al. Scaling rectified flow transformers for high-resolution image synthesis. In Proceedings of the International Conference on Machine Learning, 2024.
72. 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.
73. 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**.
74. 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*.
75. 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.
76. Christopher, J.K.; Bartoldson, B.R.; Kailkhura, B.; Fioretto, F. Speculative Diffusion Decoding: Accelerating Language Generation through Diffusion. *arXiv preprint arXiv:2408.05636* **2024**.
77. 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.
78. 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].
79. Cai, W.; Jiang, J.; Wang, F.; Tang, J.; Kim, S.; Huang, J. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204* **2024**.
80. 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.
81. 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*.
82. 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**.
83. 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].

84. 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.
85. Yu, L.; Xiang, W. X-pruner: explainable pruning for vision transformers. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24355–24363.
86. 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**.
87. 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.
88. 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.
89. 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.
90. 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**.
91. 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].
92. 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.
93. 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].
94. 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].
95. Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual Instruction Tuning. In Proceedings of the NeurIPS, 2023.
96. 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.
97. Chu, X.; Qiao, L.; Lin, X.; Xu, S.; Yang, Y.; Hu, Y.; Wei, F.; Zhang, X.; Zhang, B.; Wei, X.; et al. Mobilevlm: A fast, reproducible and strong vision language assistant for mobile devices. *arXiv preprint arXiv:2312.16886* **2023**.
98. 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].
99. 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**.
100. Xiao, J.; Li, Z.; Yang, L.; Gu, Q. BinaryViT: Towards Efficient and Accurate Binary Vision Transformers. *arXiv preprint arXiv:2305.14730* **2023**.