A Comprehensive Survey on Bias and Fairness in Generative AI: Legal, Ethical, and Technical Responses

Xiaojian Lin[1,*] and Michael Losavio[2]

[1] The University of Hong Kong, Department of Computer Science, Hong Kong, China
`chrislim@connect.hku.hk`
[2] University of Louisville, Department of Criminal Justice and Department of Computer Science and Engineering, Louisville, Kentucky, USA
`michael.losavio@louisville.edu`
[*]Corresponding Author

Abstract. Recent advancements in generative AI, particularly in computer vision and natural language processing, have brought significant innovations and highlighted critical bias and fairness issues. This paper comprehensively reviews bias in generative AI, examining its causes, impacts, and potential solutions from legal, ethical, and technical perspectives. I begin by discussing the current state of bias in generative AI, focusing on racial, gender, and cultural biases in both computer vision and natural language processing. Through case studies, I demonstrate the real-world impacts of these biases. The paper then explores the root causes of bias, including data imbalance and algorithmic design. It discusses the profound social and technical impacts, such as implications for social justice, trust in AI systems, and model performance. I review existing domestic and international policies, industry standards, and practices to mitigate AI bias, highlighting their strengths and limitations. The paper concludes with proposed solutions for improving data diversity, developing fairness-aware algorithms, enhancing regulatory frameworks, and promoting ethical AI education and public awareness. Our study underscores the need for continuous efforts and interdisciplinary collaboration to address bias and ensure fairness in generative AI systems.

Keywords: Artificial Intelligence, Bias, Fairness, Regulation.

1 Introduction

Generative AI has significantly advanced [1,2,3,4,5,6,7,8,9,10], particularly in computer vision [11,12,13,14,15,16,17,18,19,20], multimodal [21,22], time series [23,24], multi-agent systems [25,26,27,28,29] and natural language processing (NLP) [30,31,32,33,34]. These technologies have transformed various domains, enabling innovations like realistic image synthesis and text generation. However, the rapid development of generative AI has also exposed critical issues surrounding bias and fairness, which have far-reaching societal, technological, and legal implications. Generative AI systems are prone to various forms of bias, including racial, gender, and cultural biases. These can stem from the data used to train models or from the algorithms themselves. For instance, facial recognition systems have been criticized for their higher error rates when identifying individuals with darker skin tones, raising serious ethical and legal

concerns [35]. Recent advancements in facial expression recognition models, particularly those incorporating residual separable convolutional networks, aim to enhance accuracy across diverse expressions [36], which is crucial for addressing performance disparities among different demographic groups. Similarly, NLP models can perpetuate stereotypes and biased language, reflecting societal biases in their training data [37]. Advanced techniques, such as attention mechanisms, have been shown to enhance the contextual accuracy of NLP models, which is essential for mitigating biases that may arise from ambiguous or incomplete data representations [38]. Addressing bias and fairness in generative AI is critical due to its extensive impact on society, technology, and legal systems. Biased AI systems can perpetuate and exacerbate existing social inequalities, affecting decisions in key areas such as hiring, law enforcement, and healthcare. For instance, biased AI in law enforcement can lead to wrongful arrests, while in healthcare, it may result in suboptimal treatment recommendations for certain demographic groups [39]. Understanding these biases is crucial, and Figure 1 illustrates several common types of bias in AI, including data bias, sampling bias, and algorithm bias. This paper aims to examine the multifaceted issues of bias and fairness in generative AI by exploring legal, ethical, and technical responses to these challenges. Our objective is to provide a comprehensive review of the current state of bias in generative AI, analyze the causes and impacts of these biases, and propose actionable solutions to enhance fairness. This study will cover aspects such as data bias, algorithmic bias, and the social and technical impacts of bias, as well as review current policies and propose future directions for research and development.
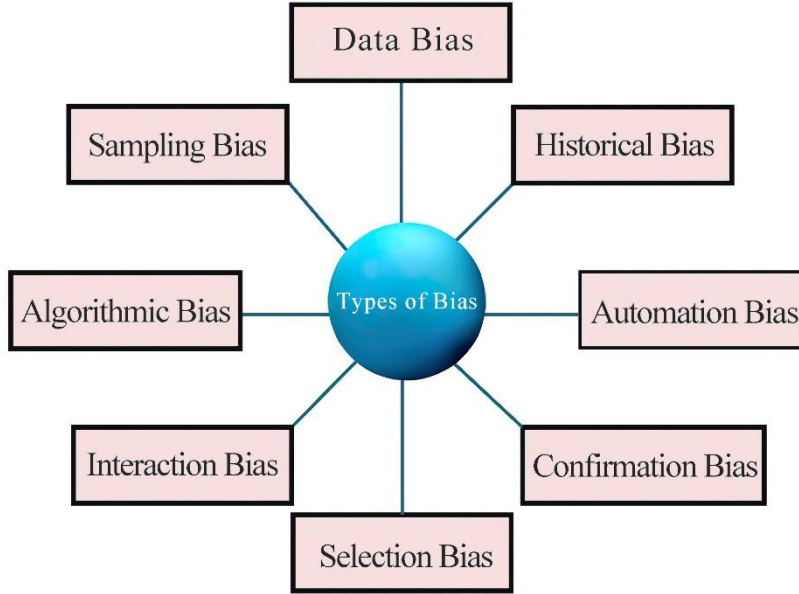

Fig. 1. Types of Bias in AI

2   Current State of Bias in Generative AI

Bias in generative AI impacts critical areas such as healthcare, law enforcement, and finance. These biases often arise from the data used in training or the model's inherent structure. **Data bias** occurs when training data is not representative of the population, leading to skewed outcomes. **Algorithmic bias** emerges due to model design, where certain groups are unfairly favored. To address these, fairness-aware loss functions can be introduced, helping mitigate biased results. In fields like healthcare and law enforcement, these biases can lead to serious consequences such as misdiagnosis or wrongful convictions.

2.1   Bias in Computer Vision

Generative AI in computer vision frequently exhibits racial, gender, and cultural biases. For instance, facial recognition systems show higher error rates for darker-skinned individuals due to imbalanced training datasets [35]. Figure 2 highlights representation disparities in key datasets (PPB, IJB-A, Adience). Such biases can lead to unfair outcomes in real-world applications.
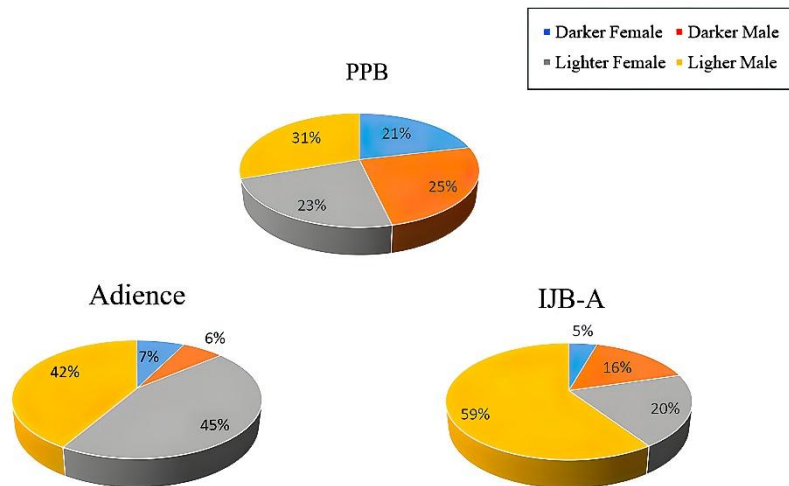

Fig. 2. Racial and Gender Bias in Computer Vision

Case Studies

Facial Recognition. Studies have reported error rates as high as 34% for dark-skinned women compared to 1% for lighter-skinned men [35].
Image Generation. Gender biases in image generation reinforce stereotypes, such as men in professional roles and women in domestic settings [37].

Recent work on fairness-aware training methods, such as Fairness-Aware Adversarial Learning (FAAL), has shown significant progress in reducing demographic disparities in error rates while maintaining overall accuracy [40].

2.2 Bias in Natural Language Processing

Natural Language Processing (NLP) has similarly been impacted by bias, particularly in text generation, chatbots, and voice assistants. These biases can perpetuate stereotypes and contribute to unequal treatment of different demographic groups.

Examination of Biases. Language models trained on large-scale datasets can inherit and even amplify existing biases present in the data. For example, gender bias is prevalent in many NLP applications, where models may generate text that reinforces traditional gender roles [41].

To illustrate these biases more clearly, I referred to the data from the source [41] and created relevant tables, showing the gender direction projections of various professions in word embeddings, highlighting the extent of gender stereotypes.

Table 1. Extreme Gendered Words.

Ranking	Extreme *he*	Extreme *she*
1	maestro	homemaker
2	skipper	nurse
3	protege	receptionist
4	philosopher	librarian
5	captain	socialite
6	architect	hairdresser
7	financier	nanny
8	warrior	bookkeeper
9	broadcaster	stylist
10	magician	housekeeper

Table 2. Gender Analogy.

Gender appropriate *she −he* analogies	Gender stereotype *she −he* analogies
queen-king	sewing-carpentry
waitress-waiter	nurse-surgeon
sister-brother	blond-burly
ovarian cancer-prostate cancer	giggle-chuckle

mother-father	sassy-snappy
convent-monastery	volleyball-football
	registered nurse-physician
	interior designer-architect
	feminism-conservatism
	vocalist-guitarist
	diva-superstar
	cupcakes-pizzas
	housewife-shopkeeper
	softball-baseball
	cosmetics-pharmaceuticals
	petite-lanky
	charming-affable
	lovely-brilliant

(a) $w^{(g)}$ dimension for all the professions (b) Gender-neutral profession words projected to gender direction in GloVe (c) Gender-neutral profession words projected to gender direction in GN-GloVe
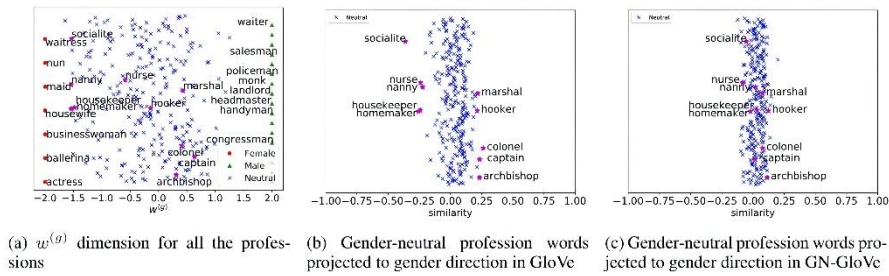
Fig. 3. Gender Neutral Word Embeddings [32]

Figure 3, from Zhao et al. [42], illustrates how newer models like GN-GloVe perform better at reducing gender bias [42].

Examples and Societal Implications

Text Generation. Figure 3 illustrates how newer models like GN-GloVe perform better at reducing gender bias [42].
Chatbots and Voice Assistants. These systems often exhibit biased responses based on user characteristics, like gender or accent, undermining user trust.

Recent advancements have focused on reducing these biases. For instance, a 2024 study introduced a comprehensive pipeline for bias mitigation in language models, employing continuous prefix-tuning to reduce internal and downstream biases while preserving model expressiveness [43]. Another 2024 approach proposed in-context bias suppression using textual preambles, which effectively reduced gender bias in models like LLaMA2 without accessing model parameters [44].

3   Causes of Bias

Generative AI models, despite their impressive capabilities, are prone to various biases that stem from both the data they are trained on and the algorithms themselves. These biases can significantly impact the fairness and reliability of AI applications.

3.1   Data Bias

Data bias stems from data imbalance, representativeness issues, and labeling inaccuracies, which can impact model fairness. For instance, Buolamwini and Gebru demonstrated that facial recognition systems trained on datasets with lighter-skinned individuals showed higher error rates for darker-skinned groups [35]. To address this, weights can be assigned to different classes, calculated as:

$$w_i = \frac{N}{n_i} \tag{1}$$

where w_i is the weight for class i, N is the total number of samples, and n_i is the number of samples in class i. Representation bias occurs when the data does not reflect real-world distribution, leading to skewed outputs. This can be quantified by measuring the disparity between the model's predictions and the actual population distribution:

$$Disparity = \frac{1}{K}\sum_{k=1}^{K} |p_k - q_k| \tag{2}$$

Recent research highlights the impact of imbalanced datasets on healthcare AI models, showing performance disparities [40]. Techniques like re-sampling, synthetic data generation, and fairness-aware algorithms have been proposed to mitigate these issues [42]. Figure 4 shows results from pilot testing the REVIEW tool on three biomedical articles, demonstrating its ability to flag racial bias in analysis, context, and tone. The first article, by Bibbins-Domingo et al. (2009), passed without significant concerns. However, the second, Bunyavanic et al. (2020), was flagged for unscientific hypotheses, inappropriate terminology, and neglect of social determinants of health. Similarly, the third article, Wang (2020), had issues with unscientific claims and inadequate consideration of social determinants. This illustrates the REVIEW tool's effectiveness in identifying racial bias in controversial research.

One specific type of data bias that deserves attention is confirmation bias. Confirmation bias refers to the tendency of models to select data or features that confirm pre-existing beliefs or hypotheses. This can be quantified using the Confirmation Bias Coefficient:

$$C = \frac{\sum_{i=1}^{N}(x_i-\mu)(y_i-\nu)}{\sqrt{\sum_{i=1}^{N}(x_i-\mu)^2}\sqrt{\sum_{i=1}^{N}(y_i-\nu)^2}} \tag{3}$$

where x_i and y_i are the features selected by the model, and μ and ν are their means. Understanding and mitigating confirmation bias is crucial for developing fair and unbiased AI models.

(1) Bibbins-Domingo,2009

(2) Bunyavanic，Grant，Vicencio 2020

(3)Wang2020

Fig. 4. Pilot Testing of the REVIEW Tool. Figure 4 shows the application of the REVIEW Tool in three studies: (1) Bibbins-Domingo et al. 2009, (2) Bunyavanic et al. 2020, and (3) Wang 2020. Two studies were flagged for racial bias concerns in dimensions like Analysis, Context, Tone & Terminology, particularly regarding race and social determinants.

3.2 Algorithmic Bias

Algorithmic bias stems from the design and architecture of AI models, particularly in generative models like GANs and Transformers. These biases arise from structural and training choices. For example, GANs trained on imbalanced datasets can reinforce stereotypes, as highlighted in surveys exploring how generative models struggle with such biases in computer vision tasks [45]. To mitigate such biases, techniques like FairGAN have been developed to integrate fairness constraints during training [46]. Transformer models in natural language processing also exhibit significant biases. Models trained on biased corpora tend to produce prejudiced outputs, influenced by attention mechanisms and the representativeness of the data. A recent study emphasizes the significance of prompt-based learning and the Embedding Association Test as effective strategies for analyzing and mitigating implicit biases in Transformer-based language models [47].

In summary, reducing algorithmic bias requires a combination of data-level and model-level interventions. Ongoing research emphasizes the need for balanced datasets and fairness-aware models to ensure equitable AI applications.

4 Impact of Bias

Bias in AI has profound social and technical implications. Socially, biased AI systems can reinforce existing prejudices, particularly in areas like facial recognition, where misidentification disproportionately affects certain racial or gender groups, leading to wrongful accusations and discrimination. These biases undermine public trust in AI, especially in critical sectors like law enforcement and healthcare, where fairness is essential [35,40]. For example, biased hiring algorithms may unfairly filter out qualified candidates from marginalized groups, reinforcing workplace inequality [39].

Technically, bias affects the performance and reliability of AI models, particularly when they are trained on non-representative datasets. Models often perform poorly for underrepresented groups, producing inaccurate predictions in diverse real-world applications. This lack of robustness hampers the generalizability of AI technologies, as seen in autonomous driving systems, where urban-trained models may fail in rural settings, posing safety risks. Furthermore, biased models can create feedback loops, where biased outputs are reused, further entrenching bias and reducing innovation opportunities. Addressing these biases is both a technical and ethical necessity to ensure AI systems that are fair, reliable, and generalizable across populations [48].

5   Current Policies and Countermeasures

Addressing bias in AI requires a multi-faceted approach that combines regulatory frameworks with industry standards. Globally, various countries have introduced regulations to ensure fairness, accountability, and transparency in AI systems. These frameworks aim to address biases that may result in social and technical inequalities. In the U.S., the proposed Algorithmic Accountability Act requires impact assessments for automated decision systems to identify and mitigate biases. Additionally, states like California have enacted privacy and data protection laws that cover AI systems. Across Asia, countries like Japan, South Korea, and China have established policies emphasizing fairness, accountability, and ethical AI development [49, 50, 51]. These include Japan's "Social Principles of Human-Centric AI" and China's "New Generation AI Development Plan." While many regions share common goals, such as ensuring fairness and transparency, differences in enforcement mechanisms persist. For instance, the GDPR imposes strict penalties for non-compliance, whereas other regions may have more lenient approaches to AI governance.

6   Proposed Solutions and Future Prospects

The challenge of mitigating bias in AI requires a comprehensive approach that spans technical improvements, policy and regulatory measures, and social and educational initiatives. This section discusses advanced strategies for technical enhancements, proposes robust policy frameworks, and highlights the importance of raising public awareness and ethical AI education.

6.1   Technical and Policy Improvements

Addressing bias in AI systems requires both technical and policy-level interventions. On the technical side, improving data diversity through techniques like data augmentation and synthetic data generation is critical for ensuring that AI systems perform equitably across diverse populations. Algorithms such as re-weighting and adversarial debiasing are also key to correcting biases during the training process, helping to balance the influence of underrepresented groups. Fairness-aware algorithms, including constraints integrated into model optimization, can be further leveraged to reduce bias in AI outcomes.
On the policy side, transparency and accountability are central. Algorithmic audits should be mandatory, particularly for high-stakes applications in healthcare, hiring, and law enforcement [51]. These audits help identify biases in decision-making processes, ensuring fairness and mitigating potential discrimination. Explainable AI (XAI) is another vital component, offering insights into how AI models make decisions, which helps stakeholders understand and address potential biases. International collaboration, driven by organizations like IEEE and ISO, will also be necessary to establish global standards for AI fairness [52].

6.2   Overview of Policies and Practices

Raising public awareness about AI biases is vital for fostering an informed and critical user base. Public awareness campaigns can educate individuals about how AI systems work, the potential biases they might harbor, and the implications of these biases on decision-making. These campaigns can leverage various media platforms to reach a wide audience, emphasizing the societal implications of AI biases [40].

Ethical AI education is also crucial. Integrating ethics into AI curricula can equip developers and users with the knowledge to recognize and mitigate biases. Educational programs should cover topics such as data privacy, fairness, accountability, and transparency. Courses can include case studies and practical exercises on identifying and addressing biases in AI systems.

Promoting diverse stakeholder participation in AI development is essential for addressing biases effectively. Diverse teams are more likely to recognize and address biases that might go unnoticed in homogeneous groups. Initiatives to support diversity in AI research and development, such as scholarships, mentorship programs, and inclusive hiring practices, can contribute to more equitable AI systems.

In summary, addressing bias in AI requires a comprehensive approach that includes technical improvements, robust policy frameworks, and social and educational initiatives. By adopting these strategies, I can develop AI systems that are fair, reliable, and trustworthy.

7   Conclusion

This paper summarizes the key findings on bias and fairness in generative AI, stressing the importance of addressing biases to improve system reliability and avoid perpetuating inequalities. The research identifies data, algorithmic, and social factors as primary sources of bias. To mitigate these issues, future research must focus on improving data diversity, developing fairness-aware algorithms, and implementing effective policy regulations. Interdisciplinary collaboration will be essential in addressing the complex challenges posed by AI biases. Moreover, fostering AI ethics education and increasing public awareness is crucial for creating informed users and promoting responsible AI development.

References

1. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A. W.-K.: MACE: Mass concept erasure in diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6430–6440. IEEE, Piscataway (2024).
2. Lu, S., Liu, Y., Kong, A. W.-K.: TF-ICON: Diffusion-based training-free cross-domain image composition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2294–2305. IEEE, Piscataway (2023).

3. Lu, S., Zhou, Z., Lu, J., Zhu, Y., Kong, A. W.-K.: Robust watermarking using generative priors against image editing: From benchmarking to advances. arXiv preprint arXiv:2410.18775 (2024).

4. Li, L., Lu, S., Ren, Y., Kong, A. W.-K.: Set you straight: Auto-steering denoising trajectories to sidestep unwanted concepts. arXiv preprint arXiv:2504.12782 (2025).

5. Gao, D., Lu, S., Walters, S., Zhou, W., Chu, J., Zhang, J., Zhang, B., Jia, M., Zhao, J., Fan, Z., Zhang, W.: EraseAnything: Enabling concept erasure in rectified flow transformers. arXiv preprint arXiv:2412.20413 (2024).

6. Ren, Y., Lu, S., Kong, A. W.-K.: All That Glitters Is Not Gold: Key-Secured 3D Secrets within 3D Gaussian Splatting. arXiv preprint arXiv:2503.07191 (2025).

7. Yu, X., Chen, Z., Zhang, Y., Lu, S., Shen, R., Zhang, J., Hu, X., Fu, Y., Yan, S.: Visual Document Understanding and Question Answering: A Multi-Agent Collaboration Framework with Test-Time Scaling. arXiv preprint arXiv:2508.03404 (2025).

8. Lu, S., Hu, X., Wang, C., Chen, L., Han, S., Han, Y.: Copy-move image forgery detection based on evolving circular domains coverage. Multimedia Tools and Applications 81(26), 37847–37872 (2022). Springer.

9. Yang, S., Lu, S., Wang, S., Er, M. H., Zheng, Z., Kot, A. C.: Temporal-Guided Spiking Neural Networks for Event-Based Human Action Recognition. arXiv preprint arXiv:2503.17132 (2025).

10. Zhu, Y., Wang, R., Lu, S., Li, J., Yan, H., Zhang, K.: OFTSR: One-step flow for image super-resolution with tunable fidelity-realism trade-offs. arXiv preprint arXiv:2412.09465 (2024).

11. Li, Y., Zhang, Z., Qiu, X., Li, X., Liu, Z., Wang, L., Li, R., Zhu, Z., Gao, H.-a., Lin, X., Cui, Z., Zhao, H., Zhao, H.: Reusing Attention for One-stage Lane Topology Understanding. arXiv preprint arXiv:2507.17617 (2025).

12. Lin, X., Zhang, W., Jiang, Y., Wu, W., Guo, Y., Wang, K., Zhang, Z., Wang, G., Jin, L., Zhao, H.: Butter: Frequency Consistency and Hierarchical Fusion for Autonomous Driving Object Detection. arXiv preprint arXiv:2507.13373 (2025).

13. Zhang, W., Zhong, J., Yao, G., Han, R., Lin, X., Jiang, L., Zhang, Z., Luo, C.: Dual-channel Heterophilic Message Passing for Graph Fraud Detection. arXiv preprint arXiv:2504.14205 (2025).

14. Yang, C., He, Y., Tian, A. X., Chen, D., Wang, J., Shi, T., Heydarian, A., Liu, P.: WCDT: World-centric diffusion transformer for traffic scene generation. arXiv preprint arXiv:2404.02082 (2024)

15. Wang, J., He, Y., Li, K., Li, S., Zhao, L., Yin, J., Zhang, M., Shi, T., Wang, X.: MDANet: A multi-stage domain adaptation framework for generalizable low-light image enhancement. Neurocomputing 627, 129572. Elsevier (2025).

16. Xin, Y., Zhuo, L., Qin, Q., Luo, S., Cao, Y., Fu, B., He, Y., Li, H., Zhai, G., Liu, X., et al.: Resurrect Mask AutoRegressive Modeling for Efficient and Scalable Image Generation. arXiv preprint arXiv:2507.13032 (2025).

17. Wang, R., He, Y., Sun, T., Li, X., Shi, T.: UniTMGE: Uniform Text-Motion Generation and Editing Model via Diffusion. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6104–6114. IEEE, Tucson (2025).

18. He, Y., Wang, X., Shi, T.: DDPM-MoCo: Advancing industrial surface defect generation and detection with generative and contrastive learning. In: International Joint Conference on Artificial Intelligence, pp. 34–49. Springer, Jeju (2024).

19. He, Y., Li, S., Wang, J., Li, K., Song, X., Yuan, X., Li, K., Lu, K., Huo, M., Tang, J., et al.: Enhancing low-cost video editing with lightweight adaptors and temporal-aware inversion. arXiv preprint arXiv:2501.04606 (2025).

12

20. Wang, J., He, Y., Zhong, Y., Song, X., Su, J., Feng, Y., He, H., Zhu, W., Yuan, X., Lu, K., et al.: Twin Co-Adaptive Dialogue for Progressive Image Generation. arXiv preprint arXiv:2504.14868 (2025).

21. Shen, Y., Lin, X., Fan, W.: A-MESS: Anchor-based Multimodal Embedding with Semantic Synchronization for Multimodal Intent Recognition. arXiv preprint arXiv:2503.19474 (2025).

22. Jiang, Y., Zhao, J., Yuan, Y., Zhang, T., Huang, Y., Zhang, Y., Wang, Y., Li, Y., Guo, X., Zhao, Y., Zhang, J., Zhang, Z., Lin, X., Zou, Y., Ma, H., Shang, Y., Hu, Y., Cai, K., Zhang, R., Chen, B., Gao, Y., Jiao, Z., Qin, Y., Du, S., Xiao, T., Liu, Z., Chen, Y., Rong, X., Wang, R., Zheng, Y., Fan, Z., Sensoy, M., Zhang, H., Zhou, P., Jin, L., Zhao, H., Yang, X., Zhao, J., Li, J., Zhou, J. T., Cheng, Z.-Q., Huang, L., Liu, Z., Zhu, Z., Li, J., Wang, G., Li, Q., Zhang, X.-Y., Yang, Y., Ye, M., Ren, W., He, Z, Su, H., Ni, R., Jing, L., Wei, X., Xing, J., Alioto, M., Shen, S., Radeva, P., Tao, D., Zhang, Y.-Q., Yan, S., Zhang, C., He, Z., Li, X.: Never Compromise to Vulnerabilities: A Comprehensive Survey on AI Governance. arXiv preprint arXiv:2508.08789 (2025).

23. Zhang, W., Xu, D., Yao, G., Lin, X., Guan, R., Du, C., Han, R., Xuan, X., Luo, C.: FreCT: Frequency-augmented Convolutional Transformer for Robust Time Series Anomaly Detection. arXiv preprint arXiv:2505.00941 (2025).

24. Zhang, W., Lin, X., Yu, W., Yao, G., Zhong, J., Li, Y., Han, R., Xu, S., Shi, H., Luo, C.: DConAD: A Differencing-based Contrastive Representation Learning Framework for Time Series Anomaly Detection. arXiv preprint arXiv:2504.14204 (2025).

25. Zhou, Z., Zhang, J., Zhang, J., He, Y., Wang, B., Shi, T., Khamis, A.: Human-centric reward optimization for reinforcement learning-based automated driving using large language models. arXiv preprint arXiv:2405.04135 (2024).

26. Zhou, Y., Li, L., Qiu, S., Yang, Z., Zhao, Y., Han, S., He, Y., Li, K., Ji, H., Zhao, Z., et al.: GLIMPSE: Do Large Vision-Language Models Truly Think With Videos or Just Glimpse at Them? arXiv preprint arXiv:2507.09491 (2025).

27. Zhou, Y., He, Y., Su, Y., Han, S., Jang, J., Bertasius, G., Bansal, M., Yao, H.: ReAgent-V: A Reward-Driven Multi-Agent Framework for Video Understanding. arXiv preprint arXiv:2506.01300 (2025).

28. Liang, X., He, Y., Xia, Y., Song, X., Wang, J., Tao, M., Sun, L., Yuan, X., Su, J., Li, K., et al.: Self-evolving Agents with reflective and memory-augmented abilities. arXiv preprint arXiv:2409.00872 (2024).

29. Li, Z., He, Y., He, L., Wang, J., Shi, T., Lei, B., Li, Y., Chen, Q.: FALCON: Feedback-driven Adaptive Long/short-term memory reinforced Coding Optimization system. arXiv preprint arXiv:2410.21349 (2024).

30. Liang, X., He, Y., Tao, M., Xia, Y., Wang, J., Shi, T., Wang, J., Yang, J.: Cmat: A multi-agent collaboration tuning framework for enhancing small language models. arXiv preprint arXiv:2404.01663 (2024).

31. Wang, J., Zhang, Z., He, Y., Zhang, Z., Song, Y., Shi, T., Li, Y., Xu, H., Wu, K., Yi, X., et al.: Enhancing Code LLMs with Reinforcement Learning in Code Generation: A Survey. arXiv preprint arXiv:2412.20367 (2024).

32. Yi, Q., He, Y., Wang, J., Song, X., Qian, S., Yuan, X., Sun, L., Xin, Y., Tang, J., Li, K., et al.: Score: Story coherence and retrieval enhancement for ai narratives. arXiv preprint arXiv:2503.23512 (2025).

33. Li, X., He, Y., Zu, S., Li, Z., Shi, T., Xie, Y., Zhang, K.: Multi-Modal Large Language Model with RAG Strategies in Soccer Commentary Generation. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6197–6206. IEEE, Tucson (2025).

34. Xiang, Y., Zhang, Z., Li, B., Weng, Y., Zhou, S., He, Y., Li, K.: ReGraP-LLaVA: Reasoning enabled Graph-based Personalized Large Language and Vision Assistant. arXiv preprint arXiv:2505.03654 (2025).

35. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency, pp. 77–91. PMLR, New York (2018).

36. Xu, X., Cui, J., Chen, X., Chen, C.L.: A facial expression recognition method based on residual separable convolutional neural network. Journal of Network Intelligence 7(1), 59–69 (2022).

37. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2979–2989. ACL, Stroudsburg (2017).

38. Shen, T.P., Meng, Q.Q., Zhan, Z.H.: Named entity recognition of Chinese text based on attention mechanism. Journal of Network Intelligence 8(4), 504–517 (2023).

39. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM Computing Surveys 54(6), 1–35 (2021).

40. Zhang, Y., Zhang, T., Mu, R., Huang, X., Ruan, W.: Towards Fairness-Aware Adversarial Learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24746–24755. IEEE, Piscataway (2024).

41. Bolukbasi, T., Chang, K.W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? debiasing word embeddings. Advances in Neural Information Processing Systems 29, 4349–4357 (2016).

42. Zhao, J., Zhou, Y., Li, Z., Wang, W., Chang, K.W.: Learning Gender-Neutral Word Embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 4454–4460. ACL, Stroudsburg (2019).

43. Yu, L., Guo, L., Kuang, P., Zhou, F.: Biases Mitigation and Expressiveness Preservation in Language Models: A Comprehensive Pipeline (Student Abstract). In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 23701–23702. AAAI, Palo Alto (2024).

44. Oba, D., Kaneko, M., Bollegala, D.: In-Contextual Gender Bias Suppression for Large Language Models. In: Findings of the Association for Computational Linguistics: EACL 2024, pp. 1722–1742. ACL, Stroudsburg (2024).

45. Sampath, V., Maurtua, I., Aguilar Martin, J.J., Gutierrez, A.: A survey on generative adversarial networks for imbalance problems in computer vision tasks. Journal of Big Data 8, 1–59 (2021).

46. Xu, D., Yuan, S., Zhang, L., Wu, X.: Fairgan: Fairness-aware generative adversarial networks. In: 2018 IEEE International Conference on Big Data (Big Data), pp. 570–575. IEEE, Piscataway (2018).

47. Bevara, R.V.K., Mannuru, N.R., Karedla, S.P., Xiao, T.: Scaling Implicit Bias Analysis across Transformer-Based Language Models through Embedding Association Test and Prompt Engineering. Applied Sciences 14(8), 3483 (2024).

48. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, pp. 4483–4488. ACM, New York (2019).

49. Social Principles of Human-Centric AI, https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf, last accessed 2024/09/26.

50. Korean companies pledge to invest 65 trillion won in AI at presidential committee's inaugural meeting, https://koreajoongangdaily.joins.com/news/2024-09-

14

26/national/politics/Korean-companies-pledge-to-invest-65-trillion-won-at-presidential-AI-committees-inaugural-meeting/2142881, last accessed 2024/09/26.

51. China's AI Policy & Development: What You Need to Know, https://fiscal-note.com/blog/china-ai-policy-development-what-you-need-to-know, last accessed 2024/09/26.

52. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, https://stand-ards.ieee.org/industry-connections/activities/ieee-global-initiative/, last accessed 2024/09/26.