

Interpretable Diffusion-Based Image Stylization Through Disentangled Content, Texture, and Color Representations

Bimala Dave

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
bimalawork@unh.edu

Abstract

Diffusion-based image stylization has become a central technique for transforming visual content while preserving semantic structure, yet its internal representations remain difficult to inspect, govern, and deploy responsibly at scale. This paper argues that disentangling content, texture, and color representations offers a practical pathway toward interpretable diffusion-based stylization without reducing generative quality. Rather than treating stylization as a monolithic transfer operation, the proposed perspective separates semantic layout, surface appearance, and chromatic distribution into partially independent conditioning pathways that can be diagnosed, audited, and adjusted. The paper develops a system-level architecture that combines diffusion backbones, latent encoders, attention mechanisms, and modular conditioning interfaces. It then examines the structural trade-offs between interpretability and fidelity, controllability and generative diversity, modularity and computational efficiency, and personalization and privacy. The discussion extends into governance, infrastructure, sustainability, robustness, fairness, and policy. Disentangled representations are shown to support model cards, datasheets, provenance tracking, and human-in-the-loop correction, but they also introduce new risks related to representation leakage, uneven performance across cultural visual domains, and hidden technical debt. The paper concludes that interpretable stylization should be treated as a socio-technical infrastructure problem rather than a purely model-centric research problem. A research agenda is outlined for evaluation standards, edge-cloud deployment, carbon-aware scheduling, participatory dataset design, and policy mechanisms that align creative expression with accountability.

Keywords

interpretable diffusion; image stylization; disentangled representations; content texture color; model governance; sustainable deployment; fairness; socio-technical systems.

1. Introduction

Diffusion-based image stylization has rapidly moved from experimental image synthesis to an infrastructural capability embedded in design tools, media production pipelines, e-commerce systems, and public-facing creative platforms. Early diffusion models demonstrated that iterative denoising could produce high-fidelity images by learning to reverse a gradual corruption process, and subsequent work improved sampling quality, efficiency, and conditional control [1], [2], [3]. Latent diffusion models further reduced computational cost by operating in compressed perceptual spaces, enabling high-resolution synthesis with practical memory requirements [4]. At the same time, evaluation studies showed that diffusion models could surpass generative adversarial networks in image fidelity and diversity, establishing them as a dominant paradigm for visual generation [5]. These advances have

made stylization not merely a technical operation but a socially consequential process through which identity, authorship, cultural aesthetics, and platform governance are mediated.

The stylization problem predates diffusion models. Classical neural style transfer demonstrated that convolutional networks could separate content structure from artistic texture, allowing the appearance of one image to be transferred onto the semantic layout of another [6]. Adaptive instance normalization and feature transformation methods later enabled real-time and arbitrary style transfer, making the approach more practical for interactive systems [7], [8]. Diffusion models extend these capabilities by offering stronger generative priors, richer texture synthesis, and greater robustness to complex prompts. However, they also obscure the distinctions that made earlier style transfer interpretable. In many diffusion stylization pipelines, content, texture, color, and semantic attributes are entangled within a shared latent representation or a single conditioning embedding, making it difficult to determine why a particular color palette, spatial arrangement, or material texture was produced.

This paper develops the position that interpretable diffusion-based image stylization can be advanced through disentangled content, texture, and color representations. The argument is not that complete disentanglement is achievable in a strict mathematical sense, but that partial and operational separation can support diagnosis, governance, and more reliable human control. A system that exposes content, texture, and color as distinct conditioning pathways can allow users to inspect which representation influenced a stylistic outcome. It can also allow auditors to test whether a model systematically alters skin tone, architectural style, or cultural ornamentation in ways that violate fairness expectations. Such separation does not automatically guarantee interpretability, but it creates interfaces through which interpretability can be operationalized.

The paper is organized around system-level concerns. Section two reviews related work and conceptual foundations, including diffusion models, neural style transfer, attention mechanisms, and conditional generation. Section three proposes a layered architecture for disentangled stylization that separates semantic content, surface texture, and color distribution while coordinating them through diffusion conditioning. Section four analyzes interpretability mechanisms and representation governance, including model cards, datasheets, and auditability. Section five examines deployment, infrastructure, and sustainability, with attention to technical debt, energy consumption, and edge-cloud trade-offs. Section six discusses robustness, fairness, and policy implications. Section seven considers evaluation, benchmarking, and cross-domain comparisons. Section eight concludes with a forward-looking research agenda.

2. Related Work and Conceptual Foundations

Diffusion models have their origins in nonequilibrium thermodynamics and were later adapted into deep generative models that learn to denoise data across multiple noise levels [1]. Denoising diffusion probabilistic models formalized a training objective that supports high-quality image generation through iterative refinement [2], while score-based approaches generalized diffusion through stochastic differential equations and continuous-time formulations [3]. Latent diffusion models shifted generation into a lower-dimensional latent space, improving efficiency and enabling text-conditioned synthesis at scale [4]. These developments created the technical foundation for modern stylization systems, but they also concentrated representational power in dense latent spaces that resist direct interpretation.

Neural style transfer provides an important conceptual counterpoint. Gatys and colleagues showed that content and style could be statistically separated using convolutional feature representations [6]. Adaptive instance normalization made arbitrary style transfer more efficient by aligning feature statistics between content and style inputs [7]. Universal style transfer methods further expanded the range of transferable textures and artistic effects through feature transforms [8]. These approaches are often less semantically flexible than diffusion models, but their architecture makes the content-style distinction more visible. Diffusion stylization inherits the generative strength of diffusion while inheriting less of the structural transparency of classical style transfer.

Attention mechanisms and transformer architectures have also shaped the field. The transformer introduced scalable attention for sequence modeling [9], and vision transformers demonstrated that patch-based attention could rival convolutional networks in visual recognition [10]. In diffusion stylization, attention layers mediate relationships between textual prompts, spatial tokens, and stylistic references. This mediation improves compositional control but complicates attribution. A color change may emerge from cross-attention, self-attention, or latent statistics, and a texture shift may result from both prompt semantics and reference-image conditioning. A scale-separated conditioning approach for style-encoder-free diffusion stylization illustrates how conditioning can be organized across scales rather than collapsed into a single style encoder [11]. This insight is important because scale-aware separation can reduce the burden on monolithic style embeddings and create more interpretable control points.

Conditional editing and personalization methods provide additional foundations. SDEdit allows guided image synthesis and editing by adding noise to an input and denoising under new conditions, enabling structure-preserving transformations [12]. ControlNet adds spatial conditioning pathways to diffusion models, improving control over pose, edges, and layout [13]. Textual inversion and DreamBooth enable subject-driven or concept-driven personalization through learned embeddings or fine-tuning [14], [15]. Hierarchical text-conditional generation with CLIP latents connects language and image representations for semantic guidance [16], while contrastive language-image pretraining provides the representational substrate for many text-to-image systems [17]. These methods expand controllability but also intensify governance concerns because personalization can encode identifiable subjects, protected attributes, or proprietary styles.

Interpretability research offers a broader frame. Doshi-Velez and Kim argue that interpretability should be evaluated in relation to the task and audience, not treated as a single intrinsic property [18]. Rudin contends that for high-stakes decisions, interpretable models should be preferred over post hoc explanations of black boxes [19]. Diffusion stylization is rarely a high-stakes decision in the same sense as criminal justice or medicine, but it increasingly affects creative labor, cultural representation, and platform moderation. Interpretability therefore becomes a socio-technical requirement rather than a purely technical luxury. Disentangled content, texture, and color representations can support this requirement by making the stylization process more inspectable at the level of representation.

3. System Architecture for Disentangled Stylization

A system architecture for disentangled diffusion-based stylization can be understood as a layered pipeline rather than a single model. The first layer concerns content representation. A content encoder extracts semantic layout, object boundaries, depth cues, and structural relationships from the input image or from a user-provided sketch. Vision transformer

backbones or convolutional feature extractors can serve this role, depending on the required balance between local detail and global composition [10]. The content representation should preserve spatial relationships while suppressing high-frequency appearance cues, so that later stylization does not confuse semantic structure with surface texture. In practice, content conditioning may be injected through spatial control pathways similar to those used in controllable diffusion [13].

The second layer concerns texture representation. Texture includes material appearance, brushstroke patterns, surface roughness, fabric weave, and stylistic repetition. A texture encoder can be trained on reference patches, artistic corpora, or user-selected texture regions. Unlike content, texture is often stationary across regions, but it may also vary by object class. A stylization system must therefore decide whether texture is global, regional, or object-specific. Scale-separated conditioning is useful here because texture statistics at different scales contribute differently to perceptual style [11]. Coarse scales influence overall artistic feel, while fine scales determine grain, edge quality, and micro-pattern. Exposing these scales as separate conditioning channels improves interpretability because a user or auditor can observe which scale produced a given effect.

The third layer concerns color representation. Color is not merely a low-level statistic; it carries cultural meaning, brand identity, emotional tone, and accessibility implications. A color representation module can encode palettes, luminance distributions, saturation ranges, and chromatic relationships. In a disentangled architecture, color conditioning should be separable from texture conditioning. Otherwise, a request for warmer tones may inadvertently alter material texture, and a request for a painterly texture may shift skin tones in ways that are difficult to predict. Separating color also enables targeted correction, such as preserving identity-related color attributes while applying stylistic palettes to background regions. This separation is particularly important for fairness, since color manipulation can disproportionately affect underrepresented skin tones and cultural visual conventions.

The fusion layer coordinates content, texture, and color conditioning within the diffusion denoising process. One design uses cross-attention adapters that inject each representation at selected timesteps and spatial resolutions. Another design uses latent concatenation with learned gating, allowing the model to weight representations dynamically. A third design uses modular adapters trained separately and composed at inference time. Each design embodies a trade-off. Attention-based fusion offers flexibility but can be difficult to audit because influence is distributed across layers. Gating offers clearer control signals but may reduce generative diversity. Modular composition supports interpretability and reuse but can create unexpected interactions when modules are combined outside their training distribution. The architecture should therefore include diagnostic instrumentation that records which conditioning pathway was active at each stage.

A central trade-off is between interpretability and fidelity. Strong disentanglement can constrain the model so much that it loses the rich, emergent texture and color blending that make diffusion stylization compelling. Conversely, a highly entangled model may produce beautiful results but resist explanation and correction. The goal is not maximal separation but actionable separation. Content, texture, and color should be separated enough that a human can identify the source of a stylistic decision, test counterfactuals, and intervene without retraining the entire system. This requires evaluation metrics that measure not only visual quality but also controllability, counterfactual consistency, and auditability. The architecture must also support personalization through methods such as textual inversion or subject-driven

fine-tuning, but it should isolate personalized embeddings so that they do not contaminate shared content or color representations [14], [15].

4. Interpretability Mechanisms and Representation Governance

Interpretability in diffusion stylization should be understood at multiple levels. At the local level, users need to know why a specific region received a particular texture or color. At the global level, platform operators need to understand whether a model systematically favors certain aesthetic norms. At the institutional level, regulators and cultural organizations need evidence that stylization systems respect copyright, consent, and representation. Disentangled representations support these levels by creating named control points. A content map, texture code, and color palette can each be logged, versioned, and compared across model updates. This logging transforms stylization from an opaque event into a traceable process.

Model cards and datasheets provide established governance instruments that can be adapted to stylization systems [20], [21]. A model card for a diffusion stylizer should document the training data, known biases, intended uses, out-of-scope uses, and performance across demographic and cultural groups. A datasheet should describe how style references were collected, whether artists consented, and what licensing constraints apply. Disentangled representations make these documents more meaningful because they allow claims to be tied to specific modules. For example, a system can report that its color module was evaluated on skin-tone preservation, while its texture module was evaluated on cultural ornamentation fidelity. Without separation, such claims are difficult to verify.

Representation governance also involves access control. Content representations may encode identity or location, texture representations may encode proprietary artistic signatures, and color representations may encode brand palettes. If these representations are stored or transmitted across services, they become assets that require security, retention limits, and privacy protection. A governance framework should specify which representations can be logged, for how long, and under what conditions. It should also address model inversion risks, where a representation could be used to reconstruct sensitive training examples. These risks are not unique to diffusion models, but they become more salient when stylization systems are deployed in creative platforms that collect user images and stylistic references.

Interpretability mechanisms should include counterfactual controls. A user should be able to hold content constant while varying texture, hold texture constant while varying color, and observe the resulting changes. This capability supports both creative exploration and auditability. An auditor can test whether changing a color prompt alters facial geometry, or whether changing a texture reference changes object identity. Such tests require the system to expose intermediate representations or at least provide reproducible conditioning controls. The system should also provide uncertainty indicators when representations conflict, such as when a texture reference implies high-frequency detail but the content resolution cannot support it. Interpretability is therefore not only about explanation after the fact but about designing interfaces that make the system’s internal commitments visible during use.

5. Deployment, Infrastructure, and Sustainability

Deploying interpretable diffusion stylization at scale involves infrastructure choices that shape both capability and accountability. Cloud-based deployment offers centralized model updates, large-scale compute, and consistent governance controls, but it raises latency, bandwidth, and privacy concerns. Edge deployment reduces data transfer and enables real-time interaction, but it constrains model size and may limit the availability of disentangled

modules. A hybrid architecture can place content extraction and lightweight color adjustment on the edge while running texture synthesis or high-resolution diffusion in the cloud. This division must be designed carefully because splitting representations across locations creates synchronization challenges and new attack surfaces.

Machine learning systems accumulate hidden technical debt through entangled dependencies, configuration drift, data pipeline fragility, and undocumented feedback loops [22]. Disentangled stylization can reduce some debt by modularizing responsibilities, but it can also introduce new debt if modules are independently versioned without integration testing. A texture module updated for one artistic domain may degrade performance on another. A color module tuned for accessibility may conflict with a brand-specific palette. Deployment infrastructure must therefore include continuous evaluation, rollback mechanisms, and representation-level monitoring. The system should track not only output quality but also changes in content preservation, texture consistency, and color fidelity.

Sustainability is a first-order concern. Training and serving diffusion models require substantial energy, and the environmental cost of generative systems has been documented in natural language processing and broader machine learning contexts [23]. Interpretable stylization does not automatically reduce energy use; in fact, multiple encoders and conditioning pathways can increase inference cost. A sustainable design should therefore consider model distillation, quantization, sparse attention, cached content representations, and carbon-aware scheduling. For example, non-urgent batch stylization can be scheduled when renewable energy availability is higher, while interactive editing can use smaller distilled models. Sustainability also involves lifecycle thinking: frequent retraining for new styles may be replaced by modular adapters that can be updated with less computation.

Large-scale generative infrastructure also raises concerns about concentration of power, labor displacement, and the extraction of cultural and artistic value [24]. A small number of platforms may control the models, datasets, and distribution channels for stylization. Disentangled representations can support more participatory governance if they allow independent artists and communities to contribute texture modules or color palettes under transparent terms. However, modularity alone does not ensure equitable benefit sharing. Policy mechanisms such as licensing standards, attribution requirements, and revenue-sharing agreements are needed to connect technical architecture with economic fairness. Deployment should therefore be evaluated not only by throughput and latency but by who can participate, who benefits, and who bears the costs of computational and cultural externalities.

6. Robustness, Fairness, and Policy Implications

Robustness in diffusion stylization refers to the system’s ability to preserve content, maintain texture coherence, and respect color constraints under distribution shift, adversarial inputs, and unusual compositions. A disentangled architecture can improve robustness by isolating failure modes. If color preservation fails, the color module can be examined. If content structure collapses, the content encoder or spatial conditioning pathway can be tested. However, separation can also create brittleness when modules are combined in novel ways. A texture module trained on paintings may produce artifacts when applied to photographic skin, and a color module trained on product photography may distort natural landscapes. Robustness testing should therefore include cross-domain combinations, not only in-domain benchmarks.

Fairness concerns are particularly acute because stylization affects visual representation. Color transfer can lighten or darken skin tones, texture transfer can exoticize or appropriate cultural patterns, and content preservation can reinforce stereotypical compositions. A system that separates color from content should be able to measure whether skin-tone preservation varies across groups. A system that separates texture from content should be able to detect whether specific cultural motifs are disproportionately applied to certain subjects. These measurements require disaggregated evaluation data and participatory input from affected communities. Fairness is not achieved by removing color or texture controls; it is achieved by making their effects visible and contestable.

Safety and policy considerations extend beyond bias. Generative stylization can be used for impersonation, non-consensual editing, misinformation, and copyright infringement. Research on concrete problems in artificial intelligence safety emphasizes specification, robustness, and unintended consequences in deployed systems [25]. In stylization, specification problems arise when users request a style that conflicts with consent or licensing. Robustness problems arise when adversarial inputs bypass content preservation safeguards. Unintended consequences arise when a stylized image is circulated without context or provenance. Policy responses may include watermarking, content credentials, disclosure requirements, and platform-level audit trails. Disentangled representations can support provenance by recording which style modules were applied and under what conditions, but they cannot resolve normative disputes about ownership and fair use on their own.

A policy-aware architecture should distinguish between permissible and impermissible representations. Some texture or color modules may be restricted because they encode protected cultural heritage or trademarked designs. Others may require consent from the artist who provided the reference. The system should support representation-level licensing, revocation, and expiration. If a style module is withdrawn, deployed models should be able to disable it without retraining the entire diffusion backbone. This capability aligns technical modularity with legal and ethical governance. It also creates a need for standardized metadata that travels with representations across services, so that downstream users can understand the conditions attached to a stylization result.

7. Evaluation, Benchmarking, and Cross-Domain Comparisons

Evaluating disentangled stylization requires metrics that go beyond perceptual quality. Content preservation can be measured through semantic segmentation agreement, depth consistency, or feature similarity. Texture fidelity can be measured through style loss, patch statistics, or human preference for material coherence. Color fidelity can be measured through palette distance, luminance preservation, and skin-tone consistency. Interpretability can be measured through counterfactual consistency, representation ablation, and user studies that assess whether participants can predict the effect of changing a specific module. Governance can be measured through auditability, documentation completeness, and time-to-correction when a harmful style is identified.

Cross-domain comparisons are essential because stylization behaves differently across photography, illustration, architectural rendering, fashion, and cultural heritage documentation. Classical style transfer methods often excel at artistic texture transfer but may not preserve semantic details under complex prompts [6], [7], [8]. Diffusion models offer stronger generative priors and better handling of ambiguous references [2], [4]. Conditional control methods improve spatial precision but may require additional training data and computational overhead [13]. Personalization methods increase user attachment and creative

control but intensify privacy and consent risks [14], [15]. A fair benchmark should therefore report performance by domain, input type, and user group rather than presenting a single aggregate score.

Benchmarking should also include infrastructure and sustainability metrics. Latency, throughput, memory footprint, energy per image, and carbon intensity vary substantially across deployment configurations. A model that achieves high visual quality on a centralized cluster may be unsuitable for edge devices or low-bandwidth regions. A model that uses multiple encoders may be more interpretable but less sustainable. Evaluation reports should therefore include trade-off profiles that allow deployers to choose configurations aligned with their governance and environmental priorities. This is especially important for public institutions, educational platforms, and community archives that may have limited compute but strong accountability requirements.

Finally, evaluation should be participatory. Artists, cultural historians, accessibility experts, and affected communities should be involved in defining what counts as successful stylization. A color transformation that is aesthetically pleasing to one group may be culturally inappropriate to another. A texture transfer that is technically accurate may still violate consent. Participatory evaluation can identify harms that automated metrics miss and can help prioritize which representations require stronger governance. It can also inform the design of user interfaces that make disentangled controls understandable to non-experts. Interpretability is not only a property of models; it is a relationship between systems, institutions, and publics.

8. Conclusion

Interpretable diffusion-based image stylization through disentangled content, texture, and color representations offers a pragmatic path toward more accountable generative systems. By separating semantic structure, surface appearance, and chromatic distribution, stylization architectures can expose control points that support diagnosis, correction, personalization, and governance. This separation does not eliminate the complexity of diffusion models, but it makes complexity more navigable. It allows users to ask which representation produced a stylistic effect, auditors to test whether that effect varies unfairly across groups, and policymakers to connect representation-level controls with licensing and consent requirements. The architecture also introduces new trade-offs in fidelity, efficiency, modularity, and security that must be managed through careful system design and continuous evaluation.

The broader argument is that interpretable stylization is a socio-technical infrastructure problem. It involves not only model architecture but also data collection, deployment, energy use, labor, cultural representation, and policy. Model cards, datasheets, provenance metadata, and participatory benchmarks are as important as attention mechanisms and latent encoders. Sustainability and fairness should be treated as design constraints rather than afterthoughts. Robustness should be evaluated across domains, not only on standard datasets. Governance should support representation-level revocation, licensing, and audit trails. Future research should develop standardized metrics for disentanglement that correlate with human understanding, build carbon-aware serving systems for modular diffusion, and create participatory frameworks for cultural style governance. If pursued with these commitments, interpretable diffusion stylization can expand creative possibility while strengthening accountability in the visual systems that increasingly mediate public and private life.

References

1. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning*, 2256–2265.
2. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
3. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., & Poole, B. (2021). Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*.
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.
5. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794.
6. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423.
7. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings of the IEEE International Conference on Computer Vision*, 1501–1510.
8. Li, Y., Fang, C., Yang, J., Wang, Z., Lu, X., & Yang, M.-H. (2017). Universal style transfer via feature transforms. *Advances in Neural Information Processing Systems*, 30, 386–396.
9. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*.
11. Zhang, Jingtao, et al. "Scale-separated conditioning for style-encoder-free diffusion stylization." *arXiv preprint arXiv:2608.19719* (2026).
12. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J.-Y., & Ermon, S. (2022). SDEdit: Guided image synthesis and editing with stochastic differential equations. *International Conference on Learning Representations*.
13. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3836–3847.
14. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.
15. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., & Aberman, K. (2023). DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation.

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 22500–22510.

16. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. arXiv preprint arXiv:2204.06125.
17. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning, 8748–8763.
18. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.
19. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215.
20. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229.
21. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
22. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503–2511.
23. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 3645–3650.
24. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 610–623.
25. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.