

Semantic-Guided Diffusion Models for Controllable Artistic Image Editing Without Style Encoders

Rolando Cerutti

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
rolandocerutti@binghamton.edu

Ojas C. Warrior

School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.

ojas.warrior30@oregonstate.edu

Pernilla Asplund

Department of Electrical Engineering and Computer Science, University of Missouri, Columbia, MO, USA.

pernilla.asplund130@missouri.edu

Abstract

Recent advances in generative modeling have made diffusion-based image synthesis a central technology for artistic production, yet controllable stylistic editing remains constrained by style encoders, reference adapters, and extensive per-artist fine-tuning. This paper examines semantic-guided diffusion models for controllable artistic image editing without style encoders, treating the problem as a socio-technical architecture rather than a narrow model-design task. The analysis argues that removing dedicated style encoders can improve composability, auditability, and deployment flexibility, but it also shifts complexity into semantic conditioning, prompt governance, latent-space control, and infrastructure operations. The paper reviews diffusion foundations, text-conditioned generation, and prior editing methods before developing a system-level account of encoder-free stylistic control. It then analyzes architectural trade-offs across semantic grounding, conditioning separation, spatial control, identity preservation, and human-in-the-loop editing. Particular attention is given to deployment constraints, energy and cost sustainability, robustness under adversarial or ambiguous inputs, fairness in cultural representation, and policy implications for provenance, consent, and creative labor. Case illustrations from studio editing, heritage restoration, game asset pipelines, and public-sector communication show that encoder-free diffusion systems are most viable when semantic guidance is treated as an institutional interface rather than a prompt-only convenience. The paper concludes that style-encoder-free diffusion can expand artistic controllability, but only if evaluation, governance, and infrastructure are designed alongside the model. The central contribution is a research agenda for semantic-guided diffusion systems that are controllable, accountable, and sustainable in real creative workflows.

Keywords

diffusion models; artistic image editing; semantic guidance; style-encoder-free control; creative systems; AI governance; deployment infrastructure; fairness; robustness; sustainability.

1. Introduction

Diffusion models have rapidly become a dominant approach for high-fidelity image generation, and their extension into artistic editing has transformed both creative practice and the technical infrastructure surrounding it [3], [5]. Unlike earlier generative paradigms that often required adversarial training or probabilistic latent variable modeling, diffusion systems learn to reverse a gradual noising process and can produce diverse outputs with strong perceptual quality [1], [2], [4]. In artistic workflows, however, generation alone is insufficient. Artists need controllable transformations that preserve composition, identity, cultural context, and intentional style while allowing iterative refinement. The conventional answer has been to introduce style encoders, reference images, adapters, or fine-tuned checkpoints that encode a target aesthetic and inject it into the generative process [10], [11], [12]. These mechanisms are powerful, but they create architectural dependencies that complicate deployment, governance, and cross-domain reuse.

This paper focuses on semantic-guided diffusion models for controllable artistic image editing without style encoders. The phrase without style encoders does not imply that style is ignored. Rather, it means that stylistic control is mediated through semantic structures, language, spatial guidance, attention modulation, and human-in-the-loop constraints instead of a dedicated encoder that maps reference images into a fixed style representation. This distinction is important because style encoders often behave as opaque bottlenecks. They can entangle artist identity, medium, palette, texture, and cultural motif into a single latent vector that is difficult to inspect, edit, or govern. Removing that bottleneck may improve modularity, but it also raises new questions about how semantic guidance is grounded, how conflicts between semantics and aesthetics are resolved, and how systems remain stable across diverse artistic traditions [8], [9], [13].

A system-level perspective is essential. Artistic image editing is not only a model inference problem. It is a workflow that includes dataset curation, prompt authoring, interface design, human review, storage, provenance, compute scheduling, and legal compliance [18], [19], [20]. Style-encoder-free diffusion systems alter the distribution of labor and risk across these layers. For example, if style is expressed through semantic descriptions and spatial controls, then the prompt and condition graph become governance artifacts. If style is instead encoded through fine-tuned weights, then the checkpoint becomes the locus of control and dispute. Each architectural choice redistributes power among artists, platform operators, model providers, and audiences. This paper therefore treats encoder-free stylistic editing as a socio-technical infrastructure problem, not merely as a modeling trick.

The paper develops this argument through several stages. It first reviews diffusion foundations and related editing methods [1]–[8], [10]–[17]. It then conceptualizes semantic guidance without style encoders and explains how scale-separated conditioning and related approaches change controllability [9]. Subsequent sections analyze architecture, deployment, sustainability, robustness, fairness, and policy. Case illustrations from studio art, heritage, games, and public communication show how encoder-free systems behave under real constraints. The conclusion synthesizes a research agenda for controllable, accountable, and sustainable artistic diffusion systems.

2. Background and Related Work

Diffusion probabilistic models generate images by learning to denoise data across multiple scales, enabling stable training and high sample quality [3], [4]. Latent diffusion models further improve efficiency by operating in a compressed perceptual space, which makes high-resolution synthesis and editing more practical for deployment [5]. These advances have been

complemented by classifier-free guidance, improved architectures, and scalable text conditioning, allowing users to steer generation through natural language [6], [7]. Text-to-image systems such as those built on contrastive language-image pretraining and hierarchical latent generation demonstrated that semantic prompts can control complex visual scenes with remarkable flexibility [7], [8]. This lineage provides the foundation for semantic-guided artistic editing, because language becomes a primary interface for specifying content, mood, composition, and style.

Editing methods have evolved along several lines. Some approaches inject noise into an input image and denoise it under a new prompt, enabling coarse-to-fine transformations [15]. Others perform inpainting or image-to-image translation using diffusion priors, which supports localized edits and restoration tasks [16], [17]. Attention-based editing allows users to manipulate cross-attention maps so that words correspond to regions, thereby improving spatial precision [14]. Control-oriented methods add depth, edges, poses, or segmentation signals to guide structure while retaining generative flexibility [13]. Personalization methods such as textual inversion and DreamBooth fine-tune or learn embeddings for specific subjects or styles [10], [11], while parameter-efficient adaptation reduces the cost of customizing large models [12]. These techniques are effective, but many rely on dedicated encoders, adapters, or learned tokens that can be difficult to audit and expensive to maintain.

The emerging interest in style-encoder-free diffusion stylization responds to these limitations. Instead of compressing style into a separate encoder, such systems condition generation through semantic descriptors, scale-dependent guidance, and multi-level control signals [9]. This approach aligns with broader trends toward modular generative systems, where content, structure, and aesthetics are controlled through distinct interfaces. It also resonates with research on prompt-based editing and attention control, because both attempt to make model behavior more interpretable and editable at the semantic level [14]. However, encoder-free control is not automatically simpler. It requires robust semantic grounding, careful handling of conflicting conditions, and interfaces that help users translate artistic intent into machine-readable guidance. The literature on datasets, model cards, and datasheets further shows that documentation and governance must accompany such systems if they are to be deployed responsibly [18], [20], [21].

3. Semantic Guidance and Style Control Without Encoders

Semantic guidance in encoder-free diffusion systems operates by decomposing artistic intent into multiple layers of meaning. At the highest level, language describes subject matter, mood, genre, period, material, and cultural reference. At intermediate levels, spatial controls specify where elements should appear, how edges should be preserved, and which regions may change. At lower levels, attention and feature modulation influence texture, palette, and rendering behavior without requiring a fixed style vector. Recent work on scale-separated conditioning illustrates how different levels of style and content can be guided at different stages of the diffusion process, allowing stylistic effects to emerge from semantic and structural conditions rather than from a dedicated style encoder [9]. This is significant because it suggests that style can be treated as a relational property of multiple constraints rather than as a single latent object.

The absence of a style encoder changes the nature of controllability. With an encoder, users often select a reference image and trust the system to extract a style representation. This can be convenient, but it hides important decisions about which features are considered stylistic and which are considered content. Without an encoder, users and systems must explicitly

negotiate these boundaries. For example, a request to edit a portrait in the style of a particular movement may need semantic descriptors such as brushwork, color temperature, contrast, and compositional rhythm, alongside spatial constraints that preserve facial identity. The editing system must then reconcile these conditions across diffusion steps. This process is more transparent in principle, but it demands better interfaces, richer semantic ontologies, and stronger evaluation methods.

Semantic guidance also raises the problem of ambiguity. Artistic language is often metaphorical, culturally situated, and historically contingent. A phrase such as luminous melancholy or ceremonial minimalism may be meaningful to a human collaborator but difficult to ground in a model. Encoder-free systems must therefore rely on prompt expansion, semantic retrieval, human curation, or interactive feedback to reduce ambiguity [7], [8]. Some systems may use multimodal embeddings to connect language with visual examples, but this risks reintroducing an encoder-like dependency if not carefully designed [7]. The key architectural question is whether style is represented as a fixed vector, a set of interpretable conditions, or a dynamic negotiation among user intent, model priors, and institutional rules. Encoder-free approaches favor the latter, but they require governance over the semantic vocabulary itself.

4. Architecture and System-Level Trade-Offs

Architecturally, encoder-free semantic diffusion systems typically separate content preservation, spatial control, and aesthetic modulation. Content preservation may rely on inversion, noise scheduling, or latent consistency constraints [15], [16]. Spatial control may use masks, depth, edges, or segmentation signals [13]. Aesthetic modulation may use language, attention reweighting, or scale-specific conditioning [9], [14]. This separation improves composability because a single system can support multiple artistic tasks without training a new style encoder for each artist or movement. It also improves auditability because each condition can be logged, reviewed, and revised. However, separation introduces coordination costs. Conditions may conflict, especially when a user requests a style that contradicts the source image structure or when multiple semantic descriptors pull the latent representation in different directions.

The trade-off between modularity and coherence is central. A tightly integrated style encoder may produce aesthetically coherent results with less user effort, but it can also entangle content and style in ways that are hard to disentangle. A modular encoder-free system may offer greater control and transparency, but it can produce inconsistent outputs if the conditioning pipeline is poorly calibrated. System designers must therefore decide where to place stabilization mechanisms. These may include semantic conflict resolution, iterative previews, confidence scoring, or human-in-the-loop selection. The architecture must also decide whether style guidance is applied globally, regionally, or progressively across diffusion stages. Scale-separated conditioning suggests that early stages shape composition and late stages refine texture, which means that style controls may need different weights at different resolutions [9]. This is a systems problem as much as a modeling problem, because it affects latency, memory, caching, and user interaction.

Another key trade-off concerns identity and consent. Artistic editing often involves real people, culturally significant objects, or protected heritage sites. Encoder-free systems may make it easier to preserve identity by using explicit spatial and semantic constraints, but they may also make it easier to recombine protected features in unauthorized ways. The architecture must therefore include provenance tracking, permission checks, and edit histories

[20], [21]. If these mechanisms are added only after deployment, they can conflict with the generative pipeline. A better approach is to treat governance as an architectural layer that conditions the diffusion process, records transformations, and enforces policy constraints. This layer may include watermarking, signed metadata, or policy-aware prompt filters, but it must be designed to avoid overblocking legitimate artistic expression.

5. Deployment, Infrastructure, and Sustainability

Deployment realities shape what encoder-free artistic diffusion can achieve. High-quality diffusion systems require substantial compute, memory, and energy, especially when users expect interactive editing at high resolution [5], [18]. Encoder-free systems may reduce the need to store and serve many custom style checkpoints, which can lower storage and model-management costs. Instead of maintaining hundreds of style-specific adapters, a platform can offer a shared semantic control interface with modular conditions. This can simplify updates, improve cache reuse, and reduce the operational burden of per-artist model versions [12]. However, encoder-free systems may increase inference-time conditioning complexity, because multiple signals must be processed and reconciled during generation. The sustainability profile therefore depends on whether reduced storage and model duplication outweigh additional runtime computation.

Infrastructure choices also affect access and equity. Cloud-based diffusion services concentrate compute in large data centers, which may exclude artists with limited bandwidth or budgets [18], [22]. On-device or edge deployment can improve privacy and latency, but it may constrain model size and conditioning complexity. Encoder-free approaches may be attractive for edge systems if they can reuse a single foundation model and express style through lightweight semantic and spatial controls. Yet the interfaces for those controls may require sophisticated user knowledge, unless platforms provide templates, vocabulary support, and community-shared presets. The governance question is whether such presets are treated as neutral tools or as cultural artifacts that require attribution, consent, and compensation. Infrastructure policy therefore intersects directly with creative labor and cultural representation.

Sustainability must be measured across the full lifecycle. Training large diffusion models consumes significant energy, but inference at scale can dominate operational emissions in popular editing platforms [18]. Encoder-free systems that encourage iterative trial and error may increase the number of inference calls, even if they reduce model storage. Designers can mitigate this through preview modes, latent caching, batched edits, and confidence-guided refinement. They can also expose resource costs to users, not as a punitive measure but as a design signal that encourages deliberate editing. A sustainable artistic diffusion system is not simply one that uses less energy per image; it is one whose workflow reduces unnecessary generation while preserving creative agency. This requires metrics that connect model architecture to user behavior, deployment topology, and institutional incentives.

6. Robustness, Fairness, and Governance

Robustness in semantic-guided diffusion is multidimensional. The system must withstand ambiguous prompts, adversarial perturbations, distribution shifts, and conflicting conditions. Encoder-free architectures may be more interpretable because style is expressed through explicit semantic and spatial controls, but they are not immune to failure. A prompt may evoke stereotypes, a spatial mask may leak into unintended regions, or a semantic descriptor may be grounded differently across cultures [19]. Robustness testing should therefore include

not only image quality but also semantic fidelity, identity preservation, geographic and cultural variation, and failure recovery. In artistic contexts, failures are not always technical errors. They may be creative surprises, but systems must distinguish between productive ambiguity and harmful misrepresentation.

Fairness concerns arise at several levels. Training datasets often overrepresent certain visual cultures, languages, and aesthetic norms, which can bias what the model considers artistic [18], [19]. Encoder-free systems may reduce dependence on curated style examples, but they still inherit biases from foundation models and semantic embeddings. If style is controlled through language, then linguistic inequities matter. Users who speak underrepresented languages or use non-dominant aesthetic vocabularies may find it harder to achieve desired results. Fairness also concerns labor. Artists whose styles are imitated without consent may face economic harm, even if no style encoder is trained on their work. Governance mechanisms such as provenance metadata, opt-out registries, and licensing agreements are therefore necessary, but they must be technically enforceable and culturally legitimate [20], [21].

Policy implications extend beyond copyright. Regulatory frameworks for artificial intelligence increasingly emphasize transparency, risk management, and human oversight [22]. Encoder-free diffusion systems can support these goals by making conditioning inputs more explicit and by enabling fine-grained edit logs. However, they can also complicate accountability if responsibility is distributed across model providers, platform operators, prompt authors, and users. A policy-aware architecture might require disclosure when an image is generated or substantially edited, maintain provenance records, and provide appeal mechanisms for affected parties. It might also require impact assessments for systems used in heritage, journalism, or public communication [20], [22]. The challenge is to design governance that is robust enough to prevent harm without freezing artistic experimentation.

7. Case Illustrations and Cross-Domain Comparisons

In studio art, encoder-free semantic diffusion can support iterative exploration without locking an artist into a particular style checkpoint. A painter might use semantic descriptors to test color relationships, spatial controls to preserve a composition, and attention modulation to adjust texture. The system can log each condition, allowing the artist to revisit decisions and compare variations [9], [14]. This workflow differs from style transfer because the goal is not to copy a reference but to negotiate a visual language. The main risk is semantic drift, where small changes in prompts produce large aesthetic shifts. Interfaces that visualize condition changes and allow selective rollback become essential.

In heritage restoration, the stakes are higher. Editing a damaged manuscript or architectural image requires historical accuracy, material sensitivity, and cultural permission. Encoder-free systems can encode conservation constraints through spatial masks, semantic descriptions of materials, and policy rules that restrict speculative reconstruction [17], [21]. However, they may also generate plausible but false details. Expert review and provenance documentation are therefore mandatory. Compared with commercial style transfer, heritage workflows prioritize restraint over novelty. This illustrates a broader point: encoder-free controllability is valuable when it supports constraint-based editing rather than unconstrained synthesis.

In game asset pipelines, artists often need many variations of characters, props, and environments while maintaining a coherent visual identity. Encoder-free diffusion could reduce the need for separate style encoders per project by using shared semantic and structural

controls [13]. It could also integrate with version control and asset management systems, where each generated variation is tracked as a branch. The difficulty is consistency across hundreds of assets. Semantic guidance alone may not guarantee the same visual grammar unless the system includes project-level ontologies, reference libraries, and automated consistency checks. Here, the architecture resembles a content management system as much as a generative model.

In public-sector communication, encoder-free editing might be used for accessible visual materials, educational illustrations, or civic campaigns. Fairness and policy compliance are especially important because public institutions must avoid misleading representations and discriminatory imagery [19], [22]. Semantic guidance can help specify inclusive attributes, but it can also encode stereotypes if the vocabulary is not curated. A public agency might therefore require model cards, datasheets, and human review before deployment [20], [21]. Cross-domain comparison shows that the same technical capability produces different risks depending on institutional context. Encoder-free diffusion is not universally better or worse; its value depends on how control, accountability, and expertise are distributed.

8. Evaluation and Operational Metrics

Evaluating encoder-free artistic diffusion requires metrics that go beyond perceptual similarity. Technical metrics such as image quality, prompt alignment, and structural preservation remain useful, but they cannot fully capture artistic intent [5], [7]. A system may produce a high-quality image that fails to respect a cultural constraint or that subtly alters identity. Evaluation should therefore include semantic fidelity across multilingual prompts, spatial accuracy, consistency across iterative edits, and human judgments from domain experts [9], [14]. It should also measure failure modes, such as condition conflict, style leakage, and unintended content insertion. These metrics must be tied to deployment decisions, not just benchmark papers.

Operational metrics matter for sustainability and governance. Inference latency, energy per edit, cache hit rate, and storage overhead determine whether a system is practical for small studios or public institutions [18]. Auditability metrics, such as the completeness of provenance logs and the percentage of edits with permission checks, determine whether a system can satisfy policy requirements [20], [22]. Fairness metrics should track representation across cultural groups, language support, and access to editing capabilities. Finally, user agency metrics can assess whether artists feel in control of the system or merely reactive to model suggestions. A multi-metric evaluation framework helps prevent optimization for narrow visual quality at the expense of social and environmental outcomes.

9. Limitations and Future Directions

The main limitation of encoder-free semantic diffusion is that semantic control is only as good as its grounding. Language and spatial conditions cannot fully capture embodied, tacit, or culturally specific knowledge. Future systems may combine semantic guidance with interactive learning, where artists teach the model through corrections rather than through fixed encoders [10], [12]. This could preserve modularity while improving personalization. Another direction is the development of shared ontologies for artistic concepts, co-designed by artists, cultural institutions, and technologists. Such ontologies could make style-encoder-free control more precise, but they would also raise questions about who defines artistic categories and how those definitions evolve.

Future work should also address infrastructure equity. Encoder-free diffusion could enable local, community-owned creative tools if models and interfaces are optimized for lower-resource environments. Alternatively, it could concentrate power in platforms that control the foundation models and semantic vocabularies. Policy design will influence which path dominates [22]. Research is needed on federated or decentralized approaches to conditioning, provenance, and model updates. Finally, sustainability requires lifecycle assessment that includes training, inference, user behavior, and hardware replacement. Without such assessment, claims about efficiency may be misleading.

10. Conclusion

Semantic-guided diffusion models for controllable artistic image editing without style encoders offer a promising alternative to encoder-centric and checkpoint-heavy workflows. By expressing style through semantic, spatial, and scale-separated conditions, these systems can improve modularity, auditability, and cross-domain reuse [9]. They also align with human-in-the-loop editing, attention control, and constraint-based restoration [13], [14], [17]. However, removing the style encoder does not remove complexity. It relocates complexity into semantic grounding, condition coordination, interface design, infrastructure, and governance. The success of such systems depends on whether they are designed as socio-technical infrastructures that support artistic agency, cultural respect, and accountable deployment.

A research agenda for encoder-free artistic diffusion should therefore integrate model architecture with evaluation, policy, and sustainability. It should develop interpretable semantic controls, robust conflict resolution, provenance mechanisms, and fairness metrics. It should also study how artists and institutions actually use these systems in studio, heritage, game, and public-sector contexts. The central conclusion is that style-encoder-free diffusion can expand controllable artistic editing, but only when technical design is paired with governance and infrastructure choices that make control meaningful, equitable, and sustainable.

References

1. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2672–2680.
2. Kingma, D. P., & Welling, M. (2014). Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.
3. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
4. 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*.
5. 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.
6. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. *Advances in Neural Information Processing Systems*, 34, 8780–8794.

7. 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.
8. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*.
9. Zhang, Jingtao, et al. "Scale-separated conditioning for style-encoder-free diffusion stylization." *arXiv preprint arXiv:2608.19719* (2026).
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2022). Textual inversion: An image is worth one word. *arXiv preprint arXiv:2208.01618*.
11. 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.
12. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *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. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross attention control. *International Conference on Learning Representations*.
15. 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*.
16. Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11461–11471.
17. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., & Norouzi, M. (2022). Palette: Image-to-image diffusion models. *ACM SIGGRAPH 2022 Conference Proceedings*, 1–10.
18. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., & Jitsev, J. (2022). LAION-5B: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35, 25278–25294.
19. 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.

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. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence. Publications Office of the European Union.