

Multimodal Conditioning for Text-Guided Diffusion Stylization with Fine-Grained Semantic Control

Desmond Christiansen

Department of Computer Science, University of Central Florida, Orlando, FL, USA.
dchristiansen@ucf.edu

Nicolas Morse

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
nmorse@buffalo.edu

Lope Almansa

School of Computing, Clemson University, Clemson, SC, USA.
lalmansa@clemson.edu

Tola Gospodarek

Department of Computer Science, University of North Texas, Denton, TX, USA.
tolag@unt.edu

Abstract

Text-guided diffusion stylization has become a central capability in generative systems, yet the field still struggles to reconcile expressive style transfer with fine-grained semantic control. This paper examines multimodal conditioning as a system-level design problem rather than a narrow modeling trick. It argues that robust stylization requires coordinated conditioning across text, reference imagery, spatial structure, and semantic maps, with architecture, training, inference, and governance treated as coupled concerns. The discussion begins with the foundations of diffusion models and text-image alignment, then analyzes conditioning mechanisms such as adapters, cross-attention modulation, latent-space guidance, and scale-separated control. Particular attention is given to structural trade-offs among fidelity, editability, latency, memory cost, and interpretability. The paper also considers deployment infrastructure, including latent diffusion serving, edge-cloud partitioning, caching, observability, and failure containment. Because stylization systems can reproduce protected expression, amplify bias, or obscure provenance, the analysis extends to fairness, consent, auditing, and policy. A further contribution is the treatment of sustainability and evaluation as inseparable from system architecture, since compute-intensive sampling and repeated human evaluation shape environmental and institutional costs. The paper concludes that fine-grained semantic control is best understood as a socio-technical property, produced by model design, data practices, interface constraints, and governance regimes acting together. The argument is that multimodal conditioning should be engineered as an accountable infrastructure, not merely as a collection of prompt-engineering techniques.

Keywords

multimodal conditioning; text-guided diffusion; stylization; semantic control; generative systems; diffusion models; deployment; governance.

1. Introduction

Text-guided diffusion stylization sits at the intersection of generative modeling, computer vision, human-computer interaction, and platform governance. Early diffusion research established that iterative denoising can produce high-quality samples by learning to reverse a gradual noise process, first through thermodynamic analogies and later through practical denoising objectives [1], [2], [3]. Subsequent work improved sampling efficiency and image quality, making diffusion competitive with and often superior to generative adversarial networks for synthesis tasks [4], [5]. Latent diffusion then reduced the computational burden by operating in compressed representations rather than pixel space, which enabled high-resolution generation on more accessible hardware [6]. In parallel, multimodal contrastive and autoregressive models demonstrated that natural language could serve as a flexible interface for visual synthesis, allowing users to specify objects, attributes, and compositions through text [7], [8]. Text-to-image systems such as Imagen further showed that large language encoders and diffusion decoders could be combined to achieve photorealism and semantic alignment [9].

Stylization, however, is not simply a matter of adding an aesthetic adjective to a prompt. Style involves texture, color distribution, brush behavior, lighting, composition, period conventions, and cultural references, while semantic control involves preserving or modifying specific objects, regions, and relations. A prompt such as “a city street in the style of a particular movement” may produce a pleasing image, but it offers weak control over which buildings remain, how windows are arranged, or whether a person’s identity is retained. Consequently, recent systems have introduced auxiliary conditioning mechanisms, including spatial control through edges, depth, segmentation, and pose [10], as well as lightweight adapters that inject additional signals without retraining the entire diffusion backbone [11]. These developments shift the problem from prompt engineering toward infrastructure design: a stylization system must route multiple modalities through a shared generative process while maintaining predictability, auditability, and acceptable cost.

This paper argues that multimodal conditioning for text-guided diffusion stylization should be studied as a layered architecture. The lower layers include data curation, text encoders, latent representations, and diffusion schedulers. The middle layers include conditioning encoders, cross-attention modulation, adapter fusion, and region-aware gating. The upper layers include user interfaces, permission models, content provenance, monitoring, and policy enforcement. Each layer introduces trade-offs. Stronger style conditioning can improve aesthetic fidelity but may reduce semantic editability. Finer spatial control can improve user intent but increases annotation and interface complexity. Larger models can improve quality but intensify energy use, latency, and concentration of computational power. Governance mechanisms can reduce harm but may also create barriers for independent creators if they are designed only around large platforms.

The central claim is that fine-grained semantic control is not an isolated model property. It emerges from the interaction of multimodal encoders, conditioning fusion strategies, runtime resource allocation, human feedback, and institutional rules. A system that performs well in a benchmark but lacks provenance logging, consent mechanisms, or accessible controls is not a complete stylization infrastructure. Conversely, a system with strong governance but brittle conditioning will fail in creative and commercial settings. The sections that follow examine foundations, architecture, semantic control, deployment, robustness, fairness, sustainability, and policy implications. The discussion aims to provide a system-level vocabulary for

researchers and practitioners who need to design diffusion stylization systems that are expressive, reliable, and accountable.

2. Foundations and Related Work

Diffusion models learn to generate data by reversing a noise corruption process. The foundational formulation by Sohl-Dickstein and colleagues connected nonequilibrium thermodynamics to generative learning, showing that a sequence of small noise additions could be inverted by a learned Markov chain [2]. Denoising diffusion probabilistic models made this approach practical for image synthesis by introducing improved training objectives and sampling procedures [3]. Denoising diffusion implicit models later demonstrated that deterministic sampling could accelerate generation without sacrificing quality, which is critical for interactive stylization [4]. Dhariwal and Nichol showed that diffusion models could surpass generative adversarial networks on image synthesis benchmarks, establishing diffusion as a dominant paradigm for high-fidelity visual generation [5]. These technical advances are necessary background because stylization inherits the strengths and weaknesses of diffusion sampling: high quality but potentially high latency, stochastic variation, and sensitivity to guidance parameters.

Latent diffusion models changed the deployment economics of generative imaging by moving the denoising process into a compressed latent space learned by an autoencoder [6]. This makes high-resolution synthesis more tractable and enables conditioning mechanisms to operate on compact representations. Text conditioning was popularized by contrastive language-image pretraining, which aligned images and captions in a shared embedding space and provided a robust semantic interface for downstream generation [7]. DALL-E and related autoregressive systems showed that language could guide image creation in zero-shot settings, while Imagen demonstrated the power of large frozen text encoders combined with diffusion decoders [8], [9]. Together, these works established the expectation that users can control generation through natural language, but they also revealed the limitations of text as a sole control channel. Text is ambiguous, underspecified, and poor at expressing precise spatial relations or subtle stylistic mixtures.

Style transfer has its own lineage. Gatys and colleagues used convolutional neural networks to separate content and style statistics, enabling artistic style transfer by optimizing an image to match content features from one image and style statistics from another [19]. Adaptive instance normalization later made arbitrary style transfer faster and more flexible by aligning feature statistics in real time [20]. These methods are important because they frame style as a statistical property distributed across network features, not as a single label. Diffusion stylization inherits this statistical view but embeds it within a generative process that can also follow text. The result is a hybrid control problem: style may be represented by reference images, textual descriptors, or learned embeddings, while content may be represented by prompts, spatial maps, or source images. Multimodal conditioning is the engineering response to this hybridity.

Recent work on controllable generation has produced a diverse toolkit. ControlNet adds spatial conditioning by training a trainable copy of the diffusion model that receives edge, depth, segmentation, or pose inputs [10]. T2I-Adapter introduces lightweight adapters that align external conditions with the diffusion model’s internal features, reducing training cost and preserving the base model’s capabilities [11]. DreamBooth and textual inversion personalize generation by binding subjects or styles to tokens or embeddings [12], [13]. Prompt-to-prompt editing manipulates cross-attention maps to change specific words while

preserving overall structure [14]. Plug-and-play diffusion features and zero-shot image-to-image translation use internal features for training-free editing and translation [15], [16]. InstructPix2Pix learns to follow editing instructions from paired data [17]. These approaches collectively show that fine-grained control is achievable, but they also expose tensions among controllability, generalization, and infrastructure complexity. A system that combines all of them must decide which condition has priority, how conflicts are resolved, and how failures are surfaced to users.

3. Multimodal Conditioning Architecture

A multimodal conditioning architecture for diffusion stylization typically begins with a latent diffusion backbone because pixel-space denoising is too expensive for interactive use [6]. The backbone includes an encoder that maps images into latents, a denoising network that predicts noise or velocity, a scheduler that controls the sampling trajectory, and a decoder that reconstructs the final image. Conditioning enters through multiple pathways. Text is encoded by a language model or a multimodal contrastive encoder and injected through cross-attention layers [7], [10]. Reference style images may be encoded by a separate vision encoder or converted into style embeddings through inversion or adapter modules [13], [15]. Spatial conditions such as segmentation masks, depth maps, scribbles, or edge maps are encoded by task-specific networks and fused at different resolutions [10], [11]. The architecture must decide where each modality enters, how strongly it influences each layer, and how conflicts are arbitrated.

One important design choice is whether conditioning is global, regional, or scale-separated. Global conditioning applies a style or semantic signal to the entire image, which is simple but imprecise. Regional conditioning uses masks or attention gating to apply different conditions to different areas, which supports fine-grained control but requires users or upstream models to provide spatial information. Scale-separated conditioning distributes control across resolution levels, allowing coarse semantics to be governed at low resolutions and fine textures at high resolutions. The required reference [18] addresses this problem by proposing scale-separated conditioning for style-encoder-free diffusion stylization, which is relevant because it highlights how style and content can be disentangled through architectural separation rather than through a single monolithic encoder. In system terms, scale separation is attractive because it matches the hierarchical nature of image generation: low-frequency structure stabilizes composition, while high-frequency layers refine texture. However, scale separation also complicates debugging because a failure in one scale may not be visible until later sampling steps.

Cross-attention modulation is another central mechanism. In text-to-image diffusion, cross-attention maps connect text tokens to spatial regions, and editing methods can manipulate these maps to preserve or alter specific concepts [14]. For stylization, cross-attention can be used to bind style descriptors to texture regions while binding semantic tokens to object regions. Yet cross-attention is not a transparent interface. Attention maps are high-dimensional, layer-dependent, and sensitive to prompt phrasing. A production system may need to log attention statistics, expose region-level controls, and provide rollback when a conditioning signal causes semantic drift. Adapters offer a complementary approach. Instead of modifying the base model, adapters learn to project external conditions into the diffusion network’s feature space [11]. This modularity is valuable for deployment because different adapters can be loaded for different styles, domains, or control modalities without retraining the backbone. It also introduces versioning and compatibility problems: an adapter trained for

one checkpoint may degrade with another, and the system must manage a matrix of model, adapter, and scheduler versions.

The architecture must also handle modality conflict. A text prompt may request a modern city, while a style reference may depict historical architecture. A segmentation map may specify a person in the foreground, while a style embedding may impose a pose or color palette that obscures the person. A system can resolve conflicts through weighted fusion, priority rules, or user-facing clarification. Weighted fusion is computationally convenient but semantically opaque. Priority rules are easier to audit but may frustrate users when their intent is complex. Interactive clarification can improve outcomes but increases latency and UI burden. The appropriate choice depends on the deployment context. A consumer creative tool may favor permissive fusion and quick previews, while a medical illustration or architectural visualization system may require strict semantic constraints and explicit conflict warnings. Thus, multimodal conditioning is not only a model design problem; it is a decision architecture for resolving competing user intentions.

4. Fine-Grained Semantic Control and Structural Trade-offs

Fine-grained semantic control refers to the ability to modify specific objects, attributes, relations, and regions while preserving others. This is difficult because diffusion models entangle semantics and appearance in their latent representations. Personalization methods such as DreamBooth and textual inversion show that subject identity and style can be encoded into model parameters or token embeddings [12], [13]. However, these methods often trade flexibility for fidelity. A personalized model may reproduce a subject faithfully but struggle to place it in novel contexts. A textual inversion embedding may capture a style but leak content from the reference images. Prompt-to-prompt editing and cross-attention control provide more surgical edits by manipulating attention maps during sampling [14]. They can preserve background structure while changing a word, but they require careful tuning and may fail when edits are spatially complex. Plug-and-play and zero-shot translation methods offer training-free alternatives by reusing internal features [15], [16]. They are attractive for rapid deployment but may be less reliable for domains with strict constraints.

The core trade-offs can be understood along several dimensions. Fidelity concerns how closely the output matches the style reference and semantic specification. Editability concerns how easily users can change one aspect without disturbing others. Controllability concerns the precision and predictability of the mapping from user input to output. Diversity concerns the range of plausible outputs for a given prompt. Latency concerns the time required for generation. Compute cost concerns the memory and processing resources consumed. Interpretability concerns the degree to which users and auditors can understand why a particular output was produced. These dimensions often conflict. Increasing style fidelity may reduce diversity and editability because the model overfits to reference statistics. Increasing spatial control may reduce naturalness because hard masks create boundary artifacts. Increasing diversity may reduce semantic precision because the model explores variations that violate constraints. A mature system should expose these trade-offs rather than hide them behind a single quality score.

Region-aware conditioning is one response to this problem. Instead of applying a style globally, the system can assign different style embeddings or text prompts to different regions. This can be implemented through masked cross-attention, regional adapters, or latent blending. The advantage is that users can specify that a sky should follow one style while a building follows another. The disadvantage is that region boundaries may become visible,

especially when style statistics differ sharply. The system must also decide how to handle soft boundaries, occlusion, and lighting consistency. If the style reference for the sky has a warm sunset palette while the building style has cool shadows, the generated image may look physically incoherent. Advanced systems may need a global harmonization pass or a learned compositor. This adds another model to the pipeline, increasing latency and creating new failure modes. Thus, fine-grained control is not free; it redistributes complexity from the user prompt to the system architecture.

Text ambiguity is another persistent challenge. Natural language is excellent for conveying high-level intent but poor at specifying precise spatial arrangements, material properties, and stylistic mixtures. Users may say “cyberpunk but subtle” or “the style of a watercolor sketch with photorealistic lighting.” These requests are underdetermined. A multimodal system can compensate by accepting reference images, sketches, color palettes, and region masks. However, each added modality increases the interface burden and the risk of conflicting instructions. The design challenge is to make multimodal control accessible without requiring users to become annotation experts. Progressive disclosure, where simple prompts are augmented with optional controls, is one strategy. Another is constraint-based generation, where users specify hard and soft constraints and the system reports which constraints were satisfied. Such interfaces require backend support for constraint checking, which in turn requires semantic segmentation, object detection, and style metrics. The result is a socio-technical system that combines generative modeling with interactive constraint management.

5. Deployment Infrastructure and Runtime Governance

Deployment infrastructure determines whether fine-grained stylization is practical. Latent diffusion reduces the cost of high-resolution generation, but interactive use still demands careful resource management [6]. A service may run multiple diffusion models, adapters, and auxiliary encoders. Each request may require text encoding, style encoding, spatial condition encoding, denoising, and decoding. If the system supports region-aware control, it may need to run additional segmentation or detection models. If it supports iterative editing, it may need to cache latents, attention maps, and intermediate features. These requirements create a complex serving stack. The system must schedule GPU resources, manage model versions, handle adapter compatibility, and isolate failures so that one malformed request does not degrade the entire service. Observability is essential: operators need to know which model versions were used, which adapters were active, how long each stage took, and whether any safety filters or content policies were triggered.

Edge-cloud partitioning is one architectural response to latency and privacy concerns. Some operations, such as prompt encoding or lightweight style projection, can run on client devices. Heavy denoising may run in the cloud. This partitioning can reduce bandwidth and protect sensitive user images. However, it creates synchronization challenges and may expose intermediate representations to different trust domains. A system that sends latents to the cloud may still leak information about the source image. A system that performs all computation locally may preserve privacy but limit model size and quality. The appropriate balance depends on the use case. A mobile sketching app may prioritize responsiveness and privacy, while a professional studio may prioritize quality and accept cloud processing under a contractual data governance regime. Infrastructure design is therefore inseparable from trust and policy.

Runtime governance includes rate limiting, content filtering, watermarking, provenance logging, and audit trails. For stylization, provenance is particularly important because outputs

may imitate living artists, protected characters, or culturally significant motifs. A robust system should record the input prompt, reference images, model version, adapter version, random seed, and any human edits. This record supports reproducibility and allows auditors to investigate complaints. Watermarking can signal that an image is synthetic, but watermarking is not a complete solution because it can be removed or may not survive aggressive post-processing. Provenance standards and cryptographic signing offer stronger guarantees, but they require ecosystem coordination. The system must also decide how to handle opt-out requests from artists whose styles are represented in training data or reference libraries. These decisions are not purely technical; they involve legal interpretation, platform policy, and community norms. A deployment that ignores them may face reputational, regulatory, and ethical risks.

6. Robustness, Fairness, Governance, and Policy

Robustness in multimodal stylization means maintaining acceptable behavior under distribution shift, adversarial inputs, and unexpected combinations of modalities. A model trained on natural photographs may fail on diagrams, scanned artwork, or low-light images. A style adapter trained on one artist may produce artifacts when applied to another domain. A prompt injection in a reference image or metadata field may override safety filters. Robustness testing should therefore include cross-domain stylization, conflicting modality tests, and stress tests with unusual aspect ratios, textual overlays, and degraded inputs. The system should fail gracefully: when confidence is low, it should ask for clarification, fall back to a safer mode, or refuse the request with an explanation. Graceful failure is a governance feature because it prevents silent harm and builds user trust.

Fairness concerns arise at multiple levels. First, training data may underrepresent certain cultures, skin tones, body types, and artistic traditions. As a result, stylization may default to dominant aesthetic norms or produce stereotypical outputs. Second, personalization methods can be used to imitate individuals without consent, including artists, performers, and private persons [12], [13]. Third, region-aware control can be used to manipulate specific parts of an image in ways that reinforce harmful stereotypes, such as altering a person’s skin tone or facial features. Fourth, access to fine-grained control may be uneven. Large studios and well-resourced researchers can train adapters and curate datasets, while individual artists may rely on opaque hosted services. Datasheets for datasets and model cards are useful instruments for transparency, but they must be specific and actionable to address these asymmetries [21], [22]. Without meaningful documentation, users cannot know what a model was trained on, what it is likely to fail at, or how to contest harmful outputs.

Policy implications extend beyond model documentation. Copyright and moral rights regimes vary across jurisdictions, and stylization systems may operate across borders. A system that is lawful in one country may violate another’s personality rights or cultural heritage protections. Platforms may respond with blanket bans on certain style prompts, but broad bans can suppress legitimate artistic experimentation and marginalized voices. A more nuanced approach combines provenance, licensing, consent workflows, and dispute resolution. For example, artists could register style preferences, licensing terms, or opt-out signals that are respected by compliant systems. Auditors could inspect logs to determine whether a system complied with those signals. Researchers could study how users perceive stylized outputs and whether disclosure changes trust. These measures require collaboration among technologists, legal experts, artists, and community representatives. Governance is not an external constraint on diffusion stylization; it is part of the system’s operating environment.

7. Evaluation and Sustainability

Evaluation of multimodal stylization cannot rely on a single metric. Fréchet inception distance and similar distributional measures capture some aspects of image quality, but they do not measure semantic control, style faithfulness, or user satisfaction. Human evaluation is essential, but it is expensive, subjective, and difficult to reproduce. A robust evaluation framework should combine automated metrics, human studies, and task-specific benchmarks. Automated metrics might assess object presence, attribute consistency, spatial alignment, and style similarity. Human studies might assess perceived intent satisfaction, aesthetic quality, and cultural appropriateness. Task-specific benchmarks might evaluate editing precision by asking whether a specified region changed while others remained stable. The evaluation should also report failure cases and uncertainty, not only average scores. Without such transparency, benchmark results can create a false sense of capability.

Sustainability is a growing concern because diffusion sampling is compute-intensive. Latent diffusion reduces cost, but high-resolution generation, multi-adapter fusion, and iterative editing can still consume significant energy [6]. The environmental footprint depends on hardware, data center efficiency, batch size, and usage patterns. A system that generates many previews for each final image may have a larger footprint than its final outputs suggest. Developers should measure energy per request, support caching of intermediate latents, and design interfaces that discourage wasteful generation. They should also consider the lifecycle cost of model training, adapter fine-tuning, and frequent model updates. Sustainability is not only an environmental issue; it is also an economic one. Compute-intensive stylization favors well-funded actors and can exclude independent creators. Efficient architectures and open-weight models can broaden access, but they may also lower barriers to misuse. The policy challenge is to encourage efficiency and accessibility without ignoring safety.

The required reference [18] is relevant here because scale-separated conditioning suggests that architectural choices can reduce reliance on large style encoders and repeated fine-tuning, potentially lowering compute overhead for stylization. This does not mean that all efficiency problems are solved, but it illustrates how conditioning design is connected to sustainability. A system that separates coarse and fine control may reuse low-resolution latents, apply style at selected scales, and avoid retraining a full model for each new aesthetic. Such strategies are promising, but they require careful evaluation to ensure that efficiency gains do not come at the cost of semantic control or robustness. Sustainability and quality must be assessed together, not traded off silently.

8. Conclusion

Multimodal conditioning for text-guided diffusion stylization with fine-grained semantic control is a layered socio-technical problem. At the model level, it requires diffusion backbones, text encoders, style representations, spatial conditioners, adapters, and attention modulation mechanisms. At the system level, it requires scheduling, versioning, caching, observability, and failure isolation. At the governance level, it requires provenance, consent, fairness auditing, and policy enforcement. These layers are tightly coupled. A decision about conditioning strength affects user control, failure modes, and auditability. A decision about model size affects latency, energy use, and access. A decision about data curation affects style diversity, cultural representation, and legal risk. Treating these as separate concerns leads to brittle systems that perform well in demonstrations but fail in practice.

The paper has argued that fine-grained semantic control should be understood as an emergent property of architecture, infrastructure, and institutions. Text prompts alone are insufficient for precise stylization, but adding more modalities does not automatically solve the problem. Each additional modality introduces encoding costs, conflict resolution challenges, and governance obligations. Scale-separated conditioning, region-aware adapters, and cross-attention control are valuable techniques, but they must be integrated into a coherent system that can explain its behavior, respect user constraints, and account for its costs. The required reference [18] illustrates one direction in which architectural separation can reduce dependence on monolithic style encoders, yet it also reinforces the need for systematic evaluation across quality, robustness, and efficiency.

Future work should focus on three priorities. First, researchers should develop evaluation methods that measure semantic control, style faithfulness, and failure modes in realistic workflows rather than isolated benchmarks. Second, engineers should design deployment stacks that make provenance, consent, and auditing first-class features rather than bolted-on compliance tools. Third, policymakers and platform operators should support governance frameworks that protect artists and communities without freezing experimentation. Diffusion stylization will continue to improve in visual quality, but its legitimacy will depend on whether it can be controlled, understood, and contested. The goal is not only to generate beautiful images. The goal is to build generative infrastructures that are expressive, robust, fair, and accountable.

References

1. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.
2. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. In Proceedings of the 32nd International Conference on Machine Learning (pp. 2256–2265). PMLR.
3. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. In Advances in Neural Information Processing Systems (Vol. 33, pp. 6840–6851).
4. Song, J., Meng, C., & Ermon, S. (2021). Denoising diffusion implicit models. In International Conference on Learning Representations.
5. Dhariwal, P., & Nichol, A. (2021). Diffusion models beat GANs on image synthesis. In Advances in Neural Information Processing Systems (Vol. 34, pp. 8780–8794).
6. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10684–10695).
7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.
8. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., ... Sutskever, I. (2021). Zero-shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning (pp. 8821–8831). PMLR.

9. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., ... Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 36479–36494).
10. Zhang, L., Rao, A., & Agrawala, M. (2023). Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 3836–3847).
11. Mou, C., Wang, X., Xie, L., Wu, Y., Zhang, J., Qi, Z., & Shan, Y. (2024). T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 5, pp. 4296–4304).
12. 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 22500–22510).
13. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A. H., Chechik, G., & Cohen-Or, D. (2023). An image is worth one word: Personalizing text-to-image generation using textual inversion. In *International Conference on Learning Representations*.
14. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., & Cohen-Or, D. (2023). Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations*.
15. Tumanyan, N., Geyer, M., Bagon, S., & Dekel, T. (2023). Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1921–1930).
16. Parmar, G., Singh, K. K., Zhang, R., Li, Y., Lu, J., & Zhu, J. Y. (2023). Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings* (Article 11, pp. 1–11).
17. Brooks, T., Holynski, A., & Efros, A. A. (2023). InstructPix2Pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 18392–18402).
18. Zhang, Jingtao, et al. "Scale-separated conditioning for style-encoder-free diffusion stylization." *arXiv preprint arXiv:2608.19719* (2026).
19. Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2414–2423).
20. Huang, X., & Belongie, S. (2017). Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1501–1510).
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. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229).
23. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623).
24. Crawford, K. (2021). *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press.