

Frequency-Decoupled Diffusion for Robust Image-to-Image Translation Under Complex Lighting Conditions

Clas Burman

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

Tomas Ridge

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

Abstract

Image-to-image translation systems increasingly operate in environments where illumination is non-uniform, mixed in color temperature, temporally unstable, or physically ambiguous. Under such conditions, conventional diffusion models often entangle scene content, material properties, and lighting cues, producing translations that are visually plausible in controlled benchmarks but brittle in deployment. This paper develops a system-level account of frequency-decoupled diffusion for robust image-to-image translation under complex lighting. The proposed conceptual architecture separates low-frequency illumination and chromatic structure from high-frequency texture, edge, and material detail, then conditions the generative process through coordinated but partially independent pathways. This separation is not merely a signal-processing convenience; it shapes model robustness, controllability, failure modes, computational cost, and governance. The discussion examines architectural trade-offs, deployment constraints across edge and cloud infrastructures, sustainability implications, fairness risks related to skin tone and cultural color interpretation, and policy questions surrounding provenance and consent. Case illustrations from autonomous driving, medical imaging, film production, e-commerce, remote sensing, and augmented reality show that lighting robustness is domain-dependent and cannot be reduced to a single benchmark. The paper argues that frequency-decoupled diffusion offers a credible route toward more stable and auditable translation systems, but only if evaluation, infrastructure, and governance are designed together. The conclusion emphasizes that robust image-to-image translation under complex lighting is a socio-technical problem, not only a generative modeling problem.

Keywords

diffusion models, image-to-image translation, complex lighting, frequency decoupling, robust systems, fairness, deployment, governance.

1. Introduction

Image-to-image translation has moved from narrow domain conversion toward general-purpose visual editing, restoration, relighting, and simulation. Early conditional adversarial methods demonstrated that a learned mapping could transform sketches, labels, or photographs into alternate visual domains, while cycle-consistent approaches reduced the need for paired supervision [4], [5]. Diffusion models later improved sample quality, training stability, and conditional controllability, enabling text-guided editing, inpainting, super-

resolution, and instruction-based modification [1], [3], [6], [7], [8], [9]. Yet complex lighting remains a persistent obstacle. A single scene may contain direct sunlight, deep shadow, colored bounce light, specular highlights, atmospheric haze, and mixed indoor sources. These factors alter low-level pixel statistics, high-level semantic interpretation, and the apparent identity of objects. A model that treats all visual variation as content variation will over-correct, hallucinate, or preserve the wrong lighting cues.

Frequency-decoupled diffusion responds to this problem by treating illumination and texture as related but not identical generative responsibilities. Low-frequency components often encode broad illumination gradients, color casts, and global tone, while high-frequency components carry edges, pores, fabric weave, sensor noise, and fine material structure. When these are entangled in a single latent representation, a model asked to change lighting may also change identity, geometry, or texture. When they are partially separated, the system can condition illumination changes on one pathway while preserving or transferring detail through another. This idea connects to broader work on diffusion conditioning, personalization, and efficient adaptation [10], [11], [12], [13], [15]. However, the systems implications are more important than any single architectural block. Frequency decoupling changes latency, memory pressure, interpretability, failure diagnostics, and the points at which human review can intervene.

This paper therefore adopts a system-level orientation. It asks how frequency-decoupled diffusion behaves when embedded in real infrastructures, from mobile capture to cloud rendering to clinical or automotive pipelines. It examines structural trade-offs between fidelity and robustness, between controllability and automation, and between local processing and centralized governance. It also considers fairness, because lighting interacts with skin tone, cultural color expectations, and historical biases in photographic technology. Finally, it considers sustainability, because high-resolution diffusion is computationally expensive and may be deployed at scales where energy and carbon become material design constraints. The central claim is that robust translation under complex lighting requires coordinated innovation in representation, architecture, evaluation, deployment, and policy.

2. Related Work and Conceptual Foundations

Diffusion probabilistic models generate data by reversing a gradual noising process, and their iterative refinement has proven effective for high-fidelity synthesis [1], [2]. Latent diffusion reduces computational cost by operating in a compressed representation, enabling high-resolution image synthesis and editing at more practical scales [3]. These advances made diffusion a default choice for many image-to-image tasks. Conditional GANs and cycle-consistent models remain conceptually important because they framed translation as a domain-mapping problem and highlighted the difficulty of unpaired supervision [4], [5]. Instruction-based editing and stochastic differential editing methods further showed that users expect flexible, prompt-driven control rather than fixed domain labels [6], [7]. Inpainting and super-resolution demonstrated that diffusion can reconstruct missing or degraded regions with strong perceptual quality, though fidelity and hallucination remain contested [8], [9].

Conditioning mechanisms are central to robust translation. Control-oriented diffusion can inject structural guidance such as edges or depth, while personalization methods allow subject or style adaptation through fine-tuning, textual inversion, or low-rank adaptation [10], [11], [12], [13]. Vision-language models provide semantic supervision and open-vocabulary alignment, but they also introduce ambiguity because language often underspecifies lighting [14]. Recent work on scale-separated conditioning for style-encoder-free diffusion stylization

suggests that separating spatial scales can improve style transfer without relying on a dedicated style encoder [15]. That line of work is relevant because lighting is inherently multi-scale: global illumination, local shadow, and micro-texture operate at different spatial frequencies and should not be forced through a single conditioning channel.

From a conceptual standpoint, frequency decoupling is a form of inductive bias. It asserts that not all pixels should be edited with equal freedom. Illumination may be global and smooth, while material detail may be stable and identity-bearing. A system that respects this hierarchy can expose controls that match human reasoning: adjust the light, preserve the face. It can also generate more interpretable failure modes. If the low-frequency pathway drifts, the image may become globally tinted or incorrectly exposed. If the high-frequency pathway drifts, the image may lose texture or introduce unnatural sharpness. This separation supports diagnosis, audit, and selective human correction. It also creates new risks, because incorrect frequency decomposition may cause halo artifacts, color bleeding, or inconsistent shadows. The architecture is therefore not automatically superior; it is a design commitment with specific operational consequences.

3. Frequency-Decoupled Diffusion Architecture

A frequency-decoupled diffusion system can be understood as a coordinated set of generative pathways rather than a single monolithic denoiser. An input image is decomposed into multiple frequency bands or scale-aware representations. The low-frequency representation captures global illumination, color balance, exposure, and broad shading. The mid-frequency representation captures object boundaries, shadow transitions, and material grouping. The high-frequency representation captures texture, fine edges, noise, and micro-structure. Each representation can be encoded into a latent space and processed by a diffusion backbone conditioned on task variables such as target lighting, time of day, weather, or artistic intent. The pathways are not fully independent; a consistency mechanism ensures that changes in illumination respect geometry and that texture transfer does not violate object identity.

The conditioning layer is where system behavior is governed. A lighting token bank or attribute vector can specify color temperature, intensity, directionality, and softness. A content-preservation controller can constrain identity-sensitive regions such as faces, text, medical lesions, or product logos. A material-aware module can estimate reflectance and roughness cues so that highlights are modified without destroying surface appearance. Cross-pathway attention allows the low-frequency illumination state to inform high-frequency rendering, while a residual bridge preserves original detail when the requested change is small. This design supports graded intervention: a user may relight an entire scene, correct a color cast, or adjust only specular highlights. The architecture also permits staged deployment, where low-frequency correction runs on-device and high-frequency refinement runs in the cloud.

Structurally, frequency decoupling introduces trade-offs. Multiple pathways can increase parameter count, memory traffic, and synchronization overhead. If the bands are processed sequentially, latency may rise; if processed in parallel, hardware utilization may improve but peak memory may increase. The decomposition itself must be robust to compression, sensor noise, and mixed lighting. Poor decomposition can create visible seams or inconsistent shadows. A system designer must therefore decide whether decoupling is fixed, learned, or adaptive. Fixed filters are interpretable and cheap but may fail on unusual scenes. Learned decomposition is flexible but harder to audit. Adaptive decomposition can allocate capacity

based on scene complexity but requires confidence estimation and fallback logic. These choices affect not only image quality but also operational predictability.

4. Robustness Under Complex Lighting

Complex lighting includes mixed color temperatures, strong directional sources, colored bounce, deep shadows, specular surfaces, atmospheric scattering, and low-light noise. In such conditions, the same object can produce very different image statistics. A robust translation system must distinguish changes in illumination from changes in identity, geometry, and material. Frequency-decoupled diffusion helps because global color and exposure shifts tend to appear in low-frequency components, while identity and texture remain more stable in higher frequencies. This separation allows the model to condition on lighting without rewriting every pixel. It also allows preservation constraints to be applied more precisely, reducing unintended changes to faces, text, or diagnostic features.

Robustness, however, is not equivalent to visual smoothness. A model can produce a pleasing relighting result while altering shadow direction in a way that contradicts scene geometry. It can remove a shadow that should remain or introduce a highlight that implies a different material. Robustness therefore requires physical plausibility, temporal consistency, and semantic preservation. In video or augmented reality, flicker and shadow swimming become critical. In autonomous driving, incorrect relighting may obscure pedestrians or lane markings. In medical imaging, color and contrast changes may alter diagnostic interpretation. Frequency decoupling can support robustness by isolating lighting transformations, but it must be paired with geometry-aware constraints, temporal regularization, and uncertainty estimation.

Fairness is also implicated. Lighting algorithms have historically been tuned to lighter skin tones, and color correction can distort darker skin tones, exacerbate bias, or erase cultural color meanings. A frequency-decoupled system can allocate preservation controls to skin regions, but those controls must be designed with diverse data and participatory evaluation. Otherwise, the model may treat certain complexions as lighting noise to be normalized. Similar concerns arise in product photography, where color accuracy affects commerce, and in remote sensing, where atmospheric correction affects environmental interpretation. Robustness under complex lighting is thus not only a technical metric; it is a question of whose visual reality the system is optimized to represent.

5. System-Level Trade-Offs and Deployment Considerations

Deploying frequency-decoupled diffusion requires balancing quality, latency, energy, privacy, and cost. High-resolution diffusion is computationally intensive, and multiple frequency pathways can increase that burden. Cloud deployment offers elastic capacity and centralized updates but introduces bandwidth, latency, and privacy risks. Edge deployment reduces data transfer and supports real-time interaction but constrains memory and battery. A common pattern is hybrid: lightweight low-frequency correction and preview generation occur on-device, while high-frequency refinement and final rendering occur in the cloud. This pattern can improve user experience, but it also creates synchronization challenges and inconsistent behavior when connectivity is poor. The system must degrade gracefully rather than fail abruptly.

Infrastructure choices also affect sustainability. Data centers consume water and electricity, and generative workloads can be significant at scale. Frequency decoupling may reduce total computation if it avoids full-resolution processing for global lighting changes, but it may also add overhead if all pathways run unconditionally. The sustainable design is adaptive: allocate

high-frequency computation only when the content requires it, cache stable representations, and compress intermediate signals. Edge processing can reduce transmission energy but may increase device turnover if it accelerates battery degradation. These trade-offs require lifecycle analysis, not only per-inference benchmarks. Policy incentives that reward efficiency and transparency may influence which architectures become dominant.

Operational governance is equally important. A relighting system used in news, advertising, or insurance must document whether an image was synthetically altered. Provenance metadata and watermarking can help, but they must survive compression and editing. Access control and audit logs are needed for sensitive domains such as medicine and surveillance. Frequency-decoupled models may offer some interpretability because lighting and texture changes are partially separable, but that interpretability is not automatic. System operators need dashboards that show which pathway changed, how confident the model is, and whether human review is required. Deployment is therefore not a final engineering step; it is a continuous socio-technical process.

6. Governance, Fairness, and Policy Implications

Governance for image-to-image translation under complex lighting must address synthetic media, consent, bias, and accountability. Relighting can make a person appear to be somewhere they were not, at a time they did not attend, or under conditions that alter perception of emotion and health. Even when the transformation is subtle, it can affect trust. Policies that require disclosure of synthetic modification are therefore relevant, but they must be technically feasible. Provenance standards, signed capture metadata, and content credentials can help, yet they depend on adoption across devices, platforms, and editing tools. Frequency-decoupled diffusion does not solve these problems, but its modularity can make it easier to document which attributes were changed and to preserve original signals for audit.

Fairness requires attention to both data and deployment. Training datasets often underrepresent complex lighting for darker skin tones, non-Western architectural materials, and everyday scenes in low-resource regions. A model trained predominantly on studio lighting may interpret natural shade or indoor mixed light as an error. Frequency decoupling can help preserve identity while adjusting illumination, but only if the training distribution supports diverse appearances. Evaluation should include skin tone diversity, cultural color expectations, gender presentation, age, and disability-related visual cues. Participatory design can identify harms that standard benchmarks miss. Governance should also consider labor: content moderators, annotators, and reviewers may be exposed to harmful material or asked to make subjective judgments without adequate support.

Policy implications extend to competition and infrastructure. If robust relighting is available only through large cloud providers, smaller creators and public institutions may face dependency and cost barriers. Open standards for lighting metadata, model cards, and evaluation protocols can reduce lock-in. Regulation may need to distinguish between beneficial correction, such as improving visibility in low light, and deceptive alteration, such as changing the apparent time of an event. Because lighting is context-sensitive, rules based only on pixel-level thresholds are likely to fail. A more durable approach combines provenance, transparency, auditability, and domain-specific oversight. Frequency-decoupled systems can contribute to this agenda by making transformations more inspectable, but governance must not treat architectural elegance as a substitute for accountability.

7. Case Illustrations and Cross-Domain Comparisons

Autonomous driving illustrates the safety stakes. A perception pipeline may rely on image translation to normalize day, night, rain, and shadow conditions. Frequency decoupling can adjust global illumination while preserving lane markings, pedestrians, and signs. However, incorrect shadow removal may erase cues used by human drivers or other models. Temporal consistency is essential, and uncertainty must trigger conservative behavior. Medical imaging presents a different profile. Color and contrast are diagnostically meaningful, so relighting must be constrained and validated. A frequency-decoupled model may help separate illumination artifacts from tissue texture, but clinical deployment requires regulatory review, traceability, and strong human oversight.

Film and media production use relighting as a creative tool. Artists may want to change time of day or match shots while preserving actor identity and set texture. Frequency decoupling aligns with established color grading workflows, where global tone and local detail are adjusted separately. E-commerce product photography requires accurate color and material representation, because returns and trust depend on visual fidelity. Remote sensing uses atmospheric correction and shadow normalization across seasons and sensor conditions, where low-frequency corrections must respect physical models. Augmented reality needs real-time consistency between virtual objects and real illumination, making latency and tracking critical. Across these domains, the same architecture must be tuned to different definitions of robustness, fairness, and acceptable failure.

8. Evaluation, Sustainability, and Operationalization

Evaluation for frequency-decoupled diffusion should move beyond perceptual similarity and user preference. Metrics should assess lighting plausibility, shadow consistency, identity preservation, text preservation, temporal stability, and uncertainty calibration. Benchmark datasets should include mixed lighting, low light, high dynamic range, diverse skin tones, and domain-specific edge cases. Human studies should be stratified and context-aware, because a result acceptable in entertainment may be unacceptable in medicine or journalism. Red-team testing should probe adversarial lighting, compression artifacts, and distribution shift. Because the system may be deployed in interactive settings, evaluation must also measure latency, failure recovery, and cognitive load on users who must correct outputs.

Operationalization requires monitoring after deployment. Lighting conditions drift with seasons, locations, sensors, and user behavior. A model that performs well in one city may fail in another due to architectural color, atmospheric haze, or local norms. Continuous evaluation can detect drift, but it must respect privacy. Federated learning and on-device adaptation offer partial solutions, though they introduce governance complexity. Sustainability metrics should include energy per successful edit, not only energy per inference, because failed edits cause retries. Caching and adaptive computation can reduce waste. Documentation should record model version, training data provenance, decomposition strategy, and known limitations. These practices transform frequency decoupling from a laboratory architecture into an accountable infrastructure.

9. Future Directions

Future work should explore adaptive frequency allocation, physics-informed lighting representations, and uncertainty-aware control. Instead of fixed bands, a model could learn which scales are safe to edit based on scene semantics and user intent. Physics-based rendering cues could constrain shadows and highlights, while uncertainty estimates could route difficult cases to human review. Federated and on-device training could improve

privacy and local relevance, but they require secure aggregation and fair contribution models. Foundation models may provide broad priors, yet domain-specific fine-tuning will remain necessary for safety-critical applications. Standards for lighting metadata and provenance could enable interoperability across tools, making it easier to audit transformations.

Another direction is participatory governance. Communities affected by synthetic media should help define acceptable use, especially for identity, culture, and historical representation. Researchers can contribute by building interfaces that expose frequency-level edits, not only final images. Such interfaces can support informed consent and creative control. Finally, sustainability must be treated as a first-class objective. Efficient architectures, renewable-powered infrastructure, and lifecycle accounting will determine whether large-scale relighting is compatible with environmental goals. Frequency-decoupled diffusion is promising precisely because it offers a modular path toward efficiency and control, but its benefits depend on how the surrounding system is designed.

10. Conclusion

Frequency-decoupled diffusion provides a coherent framework for robust image-to-image translation under complex lighting. By separating low-frequency illumination from high-frequency texture and identity-bearing detail, it allows generative models to change light without unnecessarily rewriting content. This separation supports controllability, interpretability, and graceful degradation, and it aligns with established practices in color grading, computational photography, and physical rendering. However, the approach is not a complete solution. It introduces architectural complexity, potential decomposition artifacts, and new dependencies on data diversity and governance. Its value depends on how it is evaluated, deployed, and regulated. Robust translation under complex lighting is a socio-technical challenge that spans model design, infrastructure, fairness, sustainability, and policy. The most credible path forward combines frequency-aware generative architectures with transparent provenance, inclusive evaluation, adaptive computation, and accountable human oversight. In that combined form, frequency-decoupled diffusion can contribute to image-to-image systems that are not only visually convincing but also stable, equitable, and trustworthy in real-world conditions.

References

1. Ho, J., Jain, A., & Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840–6851.
2. 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*.
3. 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.
4. Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1125–1134.
5. Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2223–2232.

6. Brooks, T., Holynski, A., & Efros, A. A. (2023). InstructPix2Pix: Learning to follow image editing instructions. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18392–18402.
7. 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*.
8. 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.
9. Saharia, C., Ho, J., Chan, W., Salimans, T., Fleet, D. J., & Norouzi, M. (2022). Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4), 4713–4726.
10. 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.
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. 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*.
13. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
14. 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 International Conference on Machine Learning*, 8748–8763.
15. Zhang, Jingtao, et al. "Scale-separated conditioning for style-encoder-free diffusion stylization." *arXiv preprint arXiv:2608.19719* (2026).