Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33774

Balancing Visual Realism and Semantic Consistency in Image Generation: A Review from the Perspective of Practical Failures

Hanai Zhang

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, China
u430004036@mail.uic.edu.cn

Abstract. Generative models have substantially improved visual realism in image editing, image-to-image translation, and text-to-image synthesis. However, in practical applications, visually plausible outputs still often violate task semantics through identity drift, structural distortion, category swaps, and attribute mismatch. This study examines why these failures persist and how existing methods attempt to reduce them. Using a focused literature review, it compares representative studies in diffusion models, conditional GANs, cycle-consistent translation, semantic guidance, and feature-level supervision across major image generation tasks. The analysis shows that realism alone is an inadequate proxy for semantic correctness. Across editing, translation, and text-to-image generation, more reliable results depend on explicit semantic anchors, stronger conditioning, and feature-space constraints that preserve invariant content. The review also finds that current evaluation protocols focus too much on realism metrics and often miss fine-grained semantic errors.

Keywords: Image Generation, Semantic Consistency, Diffusion Models, Image-to-Image Translation, Generative Adversarial Networks

1. Introduction

Recent GAN-based and diffusion-based systems have significantly improved the visual realism of image editing, image-to-image translation, and text-to-image (T2I) generation. Furthermore, studies on inpainting, subject-driven editing, conditional translation, and diffusion synthesis exhibit clear technical progress in image quality, controllability, and diversity. In practice, outputs often look realistic but fail to preserve the source meaning. Common issues include identity drift, structural inconsistencies, object misplacement, and attribute mismatches. Research has improved realism, but semantic failures across tasks remain largely unstudied. This paper focuses on the tension between visual realism and semantic consistency in practical image generation. It also investigates three questions: why semantic inconsistency continues to appear in advanced generative models, how it emerges from the trade-off between realism and semantic preservation, and which preservation strategies are most effective in reducing these failures across editing, translation, and text-to-image synthesis. In order to address these issue, the paper adopts a focused literature review, , highlighting that realism alone is not a sufficient indicator of task success. Through an analysis of representative

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33774

failure patterns and preservation mechanisms, the paper emphasizes the importance of explicit semantic anchors, stronger conditioning, feature-space supervision, and more diagnostic evaluation methods. These observations may inform future work on reliable and controllable image generation.

2. The semantic inconsistency in image editing tasks

2.1. Scene repair and boundary refinement

Image editing requires local changes to remain coherent with preserved content. For example, a common failure is boundary inconsistency, where an edited region looks realistic by itself but does not align semantically with surrounding context. In inpainting, this appears as discontinuous edges, warped horizons, or mismatched illumination, as shown in Figure 1. As a result, such cases show that local realism does not guarantee semantic continuity across edited and unedited regions.

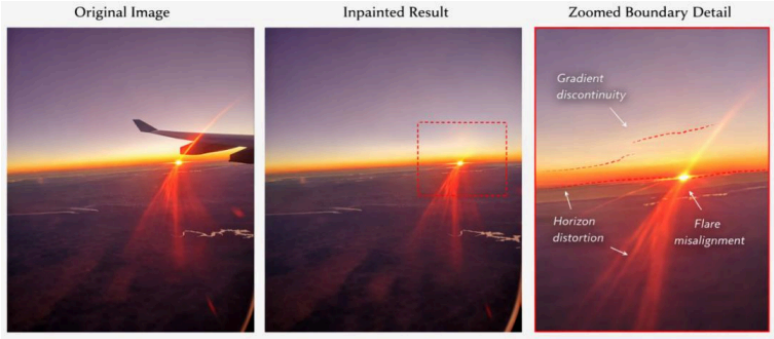

Figure 1. Inpainting failure

In response to these issues, RePaint illustrates the progress and the limitation of diffusion-based inpainting [1]. By repeatedly constraining known pixels during denoising, it enhances the coherence between the generated region and visible context. Still, natural-looking textures are insufficient, as weak contextual alignment can cause structural mismatches. Hence, scene repair should maintain continuity rather than only filling missing pixels.

2.2. Identity shifts and attribute drift

By maintaining the integrity of identity-defining features, semantic reliability in image editing can be maintained while applying the intended modifications. InstructPix2Pix demonstrates the conflict between following instructions and preserving identity, since the model can alter identity-defining features [2]. DreamBooth exhibits a similar issue in subject-driven fine-tuning, where the model can overfit to the new subject and distort the original class prior unless prior-preservation constraints are enforced [3]. These observations indicate that uncontrolled editing can lead to both identity shifts and attribute drift, thus highlighting the need for mechanisms that separate invariant content from editable properties. Consequently, robust semantic controls are essential to ensure edited images match the intended changes without altering the original identity or class.

2.3. Pixel loss and perceptual distortion

The selection of supervision objective also influences the trade-off between realism and semantic preservation. In particular, pixel-level losses are stable and straightforward to optimize; yet, because they measure images directly in pixel space, they heavily penalize even minor misalignments. When

DOI: 10.54254/2755-2721/2026.GU33774

multiple high-frequency solutions are plausible, the model tends to average them, which suppresses fine detail and produces blur, as shown in Figure 2 [4, 5].

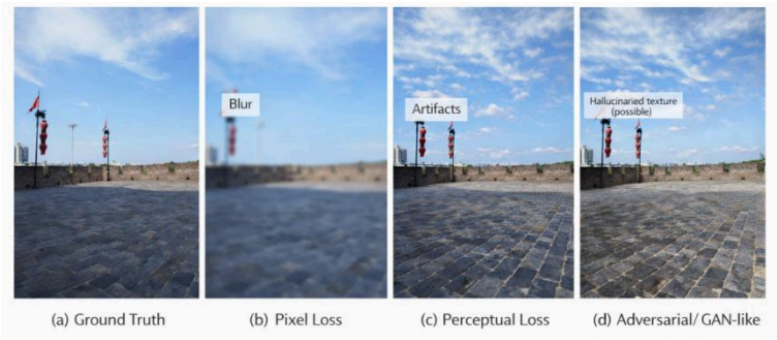

Figure 2. Effects of different losses on visual outcomes

Perceptual loss tackles this problem by comparing deep features instead of raw pixels, helping the model keep higher-level structure and texture [5]. Yet perceptual similarity is still not equivalent to semantic equivalence. It improves visual quality and often sharpens detail, but by itself it does not guarantee stable identity, object relations, or strict correspondence between input and output.

3. The semantic inconsistency in conditional generation and translation

3.1. Image translation under structural constraints

Image translation under structural constraints requires the output to remain aligned with the input while changing appearance. In paired settings, conditional adversarial training provides a strong baseline because the target structure is more explicitly constrained [6]. Nevertheless, an image can appear realistic without maintaining exact correspondence, especially when the objective favors distribution-level realism instead of task-specific consistency. This limitation becomes clearer when pixel, adversarial, and perceptual losses are compared together. Adversarial training often improves sharpness and realism, while perceptual loss helps preserve structure at a higher feature level [5, 6]. However, no single objective fully prevents semantic drift, so extra constraints are often needed to preserve geometry, identity, or region-level consistency.

3.2. Unpaired translation and cycle consistency

Unpaired image-to-image translation is more ambiguous because no exact input-output pairs are available. Cycle consistency helps regularize this setting by requiring a translated image to map back to its source, which discourages arbitrary changes and improves stability [7]. However, cycle consistency does not guarantee semantic correctness. In fact, a translation can be reversible while still mapping incorrect content or relationships, since the model may preserve invertibility without maintaining class identity or precise correspondence. This is why cycle-consistent systems can still generate category swaps, omit certain entities, or produce outputs that are semantically implausible yet visually plausible [7, 8].

DOI: 10.54254/2755-2721/2026.GU33774

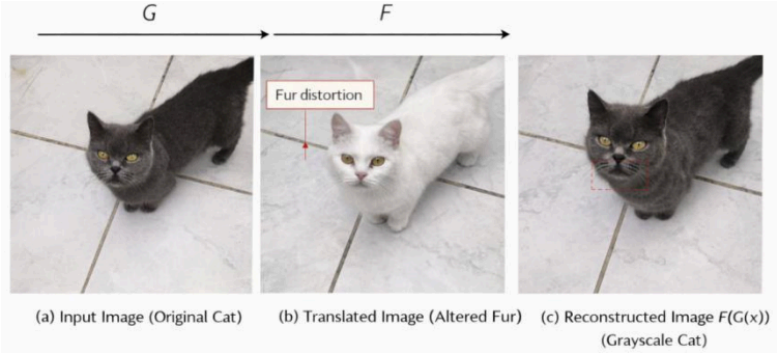

Figure 3. Cycle consistency satisfied but semantic mismatch example

3.3. Semantic constraints and example guidance

To reduce this gap, later work adds clearer semantic guidance. For instance, Sem-GAN incorporates segmentation-based consistency so that translated regions are encouraged to preserve class identity rather than only overall plausibility [8]. This directly addresses the fact that cycle consistency alone cannot ensure correct mode mapping in complex scenes. However, when the target domain is multi-modal, semantic consistency by itself still does not select a unique valid output. In this context, example guidance is useful, as it specifies the target appearance through a reference image while allowing the source image to maintain layout and content. In EGSC-IT, exemplar guidance and feature masks reduce style leakage across regions and support more semantically stable many-to-many translation [9]. EGSC further reduces local inconsistency by separating content and style more carefully, so that style cues are less likely to leak across regions [9]. Thus, semantic constraints prevent meaning drift, whereas example guidance helps select a specific realistic output mode. The broader lesson is that realism must be paired with either semantic anchors or stronger appearance control if translation is expected to preserve both structure and intended meaning.

3.4. Blur and distortion in text-to-image generation

Text-to-image generation often achieves strong global realism, yet it can still fail to preserve local semantics. Common errors include blur, missing or malformed object parts, and unstable geometry. These failures are significant because they reduce not only perceptual quality but also the semantic reliability of generated scenes. Specifically, blur often occurs when multiple fine-detail solutions are plausible, causing the model to average uncertain high-frequency content [10, 11]. In contrast, distortion is more closely related to unstable control during sampling: though stronger classifier-free guidance can improve prompt alignment, it may also amplify artifacts and deform geometry [12]. Thus, additional spatial or semantic constraints, like stronger conditioning or control mechanisms, are required when naturalness and structural correctness must be preserved simultaneously [13].

4. Mechanism-level analysis of semantic preservation strategies

4.1. External semantic anchors and segmentation priors

External semantic anchors represent one effective approach for constraining generation In particular, segmentation priors, layout controls, and other spatial cues tell the model which objects or regions should stay stable during generation. Within diffusion systems, ControlNet injects spatial cues such

DOI: 10.54254/2755-2721/2026.GU33774

as edges, depth, pose, or segmentation maps into the denoising process to guide geometry directly instead of relying only on text or appearance cues [13].

4.2. Feature space supervision and distillation

A second class of methods works in feature space. Pixel reconstruction and GAN objectives are useful for realism, but feature-level supervision constrains higher-level representations more directly. This is why perceptual losses, semantic feature matching, or teacher-guided distillation often improve semantic preservation even when they do not optimize pixel similarity exactly [5, 14, 15]. The central idea is that intermediate representations should remain consistent for the semantic properties that matter. This idea can be summarized as a feature-matching loss:

$$L_{\text{feat}} = \sum_l \lambda_l \left\| \phi_l(\hat{x}) - \phi_l(x_{\text{ref}}) \right\|_2^2 \tag{1}$$

where the generated image is encouraged to match the reference in intermediate feature space rather than only in pixel space. In practice, this helps preserve structure, identity-related cues, and task-relevant content more reliably than pixel comparison alone. By shaping the internal feature space instead of just the final pixels, these methods preserve identity, structure, and task-relevant content while enabling visually realistic synthesis.

4.3. Representation decomposition and latent space editing

A third strategy is to decompose representation and control. Many editing failures occur because identity, structure, and style are entangled in one latent representation, so changing one factor unintentionally changes others. To address this, methods that separate content from style, or that expose more interpretable controls, reduce this problem by keeping stable information in one part of the representation and editable information in another [9, 11, 15].

5. Existing challenges and future directions

5.1. Semantic inconsistencies and root causes

Semantic inconsistency is a recurring system-level problem rather than an isolated defect in one model family. Across editing, translation, and text-to-image synthesis, the same pattern reappears: objectives that optimize realism can still allow identity drift, structural distortion, incorrect relations, or unstable correspondence. First, training objectives usually reward perceptual plausibility more directly than semantic equivalence. Pixel, adversarial, and diffusion losses all improve realism, but none of them alone guarantees that the meaning implied by the task will be preserved [5, 10, 13]. Second, entangled representations make it hard to change one factor without affecting others. When identity, layout, and style are coupled, edits can spill into unrelated content, and fine-tuning on a subject can shift class priors [2, 3]. Third, weak conditioning leaves the mapping underdetermined. Cycle consistency, text prompts, and some spatial controls can reduce ambiguity, but semantically wrong outputs may still look plausible unless the condition enforces stronger invariances or the model gets explicit semantic anchors [7, 8, 13]. These causes show that reliable generation needs explicit preservation in objectives, representations, conditioning, and sampling, instead of assuming semantics will arise automatically from realism.

DOI: 10.54254/2755-2721/2026.GU33774

5.2. Evaluation gaps and methodological limitations

Current evaluation practice still underestimates semantic failure because it rewards realism more consistently than correctness. A method can score well on standard benchmarks but still drift in identity, swap attributes, or break object relations in ways that affect users. Three key limits stand out. First, metrics often do not match task goals: FID measures realism, CLIP-based scores give broad alignment signals, and structure metrics cover only part of the correspondence [16]. As shown in Table 1, these metrics reflect different aspects of output quality, but none fully captures semantic consistency. This means that strong benchmark performance does not necessarily imply reliable preservation of identity, non-target attributes, or spatial relations. Second, single aggregate scores hide which failure type remains unsolved. Third, benchmarks and prompts often underspecify the invariances that users actually expect systems to preserve.

Table 1. Evaluation metric comparison

Metric	Measures	Limitation
FID	Visual realism	Not semantic
CLIP Score	Text alignment	Surface-level
IoU	Structure match	Limited scope

Evaluation should combine realism metrics with explicit semantic tests, guided human judgment, and sensitivity analysis. Instead of asking only if an output looks good, it should check whether identity stayed stable, non-target attributes were preserved, and spatial relations remain correct.

6. Conclusion

This paper has argued that the central challenge in practical image generation is not realism by itself, but the reliability of semantic preservation under editing and generation constraints. Across image editing, conditional translation, and text-to-image synthesis, outputs can look convincing while still drifting in identity, structure, or relation-level meaning. The results show that more reliable systems rely on explicit semantic anchors, feature-space supervision, and better representation control. In addition, it indicates that evaluation must become more diagnostic, as realism-oriented metrics can conceal persistent semantic errors. Thus, future progress depends on treating semantic preservation as a primary objective during both training and sampling, rather than treating it as a by-product of visual fidelity.

References

[1] Lugmayr, A., Danelljan, M., Romero, A., Yu, F., Timofte, R., & Van Gool, L. (2022). RePaint: Inpainting using denoising diffusion probabilistic models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

[2] 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 (CVPR).

[3] 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. arXiv preprint. https: //arxiv.org/abs/2208.12242

[4] Wycoff, N., Binois, M., & Wild, S. M. (2021). Sequential learning of active subspaces. Journal of Computational and Graphical Statistics, 30(4), 1224-1237.

[5] Johnson, J., Alahi, A., & Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In B. Leibe et al. (Eds.), Computer Vision – ECCV 2016 (pp. 694–711). Springer. https: //arxiv.org/abs/1603.08155

DOI: 10.54254/2755-2721/2026.GU33774

[6] Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-1134).
[7] Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2020). Unpaired image-to-image translation using cycle-consistent adversarial networks. arXiv preprint. https: //arxiv.org/abs/1703.10593
[8] Cherian, A., & Sullivan, A. (2019). Sem-GAN: Semantically-consistent image-to-image translation. In 2019 ieee winter conference on applications of computer vision (wacv) (pp. 1797-1806). IEEE.
[9] Ma, L., Jia, X., Georgoulis, S., Tuytelaars, T., & Van Gool, L. (2018). Exemplar guided unsupervised image-to-image translation with semantic consistency. arXiv preprint arXiv: 1805.11145.
[10] Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E., Seyed Ghasemipour, S. K., Karagol Ayan, B., Mahdavi, S. S., Gontijo-Lopes, R., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. In Advances in Neural Information Processing Systems (NeurIPS). https: //arxiv.org/abs/2205.11487
[11] 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 (CVPR). https: //arxiv.org/abs/2112.10752
[12] Ho, J., & Salimans, T. (2022). Classifier-free diffusion guidance. arXiv preprint. https: //arxiv.org/abs/2207.12598
[13] 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).
[14] Zhu, X., Hu, H., Lin, S., & Dai, J. (2019). Deformable convnets v2: More deformable, better results. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9308-9316).
[15] Huang, X., Li, Y., Poursaeed, O., Hopcroft, J., & Belongie, S. (2017). Stacked generative adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 5077-5086).
[16] 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. In International conference on machine learning (pp. 8748-8763). PmLR.