Original Research Article

An Empirical Evaluation of DragGAN's Efficacy Across Distinct Subject Categories

Sihoon Kim

New York University, Tandon School of Engineering, 6 MetroTech Center, Brooklyn, NY 11201, United States

ABSTRACT

Generative Adversarial Networks (GANs) have reshaped the landscape of synthetic media, enabling the creation of hyper-realistic imagery through adversarial learning. Within this domain, DragGAN has emerged as a notable innovation, offering intuitive point-based manipulation of generated images by translating user-specified handle points to target spatial locations. Despite its qualitative success in published demonstrations, a rigorous quantitative evaluation of its performance across varying semantic categories remains absent from the literature. This study addresses that gap by assessing DragGAN's efficacy in maintaining structural integrity and generating diverse outputs across four distinct subject categories: human faces, dog faces, cat faces, and whole dog bodies. Using a curated dataset of images generated from pre-trained StyleGAN2 checkpoints (FFHQ, AFHQ, and a Self-Distilled StyleGAN body model), the Structural Similarity Index (SSIM) was applied to measure fidelity and a decomposed Inception Score (IS) was used to evaluate perceptual quality and diversity. All categories exhibited substantial structural degradation under point-based manipulation, with mean SSIM scores ranging from 0.21 (cat faces) to 0.33 (dog bodies). The full-body dog category achieved the highest structural preservation, while facial categories—particularly cat and dog faces—showed the greatest degradation. Decomposed Inception Score analysis indicated consistently low classifier confidence across all categories, a pattern attributable to domain mismatch between the generated subjects and the ImageNet-trained Inception-v3 classifier. These findings establish a quantitative baseline indicating that DragGAN's point-based manipulation introduces significant structural distortion across all tested domains, with relative performance differences suggesting that full-body manipulation may be more tractable than fine-grained facial editing.

Keywords: Generative Adversarial Networks; DragGAN; image manipulation; structural similarity index; Inception Score; computer vision; deep learning; StyleGAN

INTRODUCTION

The trajectory of artificial intelligence in computer vision has been characterized by a sustained pursuit of realism and control. The introduction of Generative Adversarial Networks (GANs) by Goodfellow and colleagues marked a paradigm shift in image synthesis, pitting a generator network against a discriminator network in a zero-sum game and thereby enabling the synthesis of data distributions virtually indistinguishable from real-world imagery (1). Subsequent architectures, most notably the StyleGAN family, pushed photorealism further while introducing the disentangled intermediate latent space W that underpins most modern image

Corresponding author: Sihoon Kim, E-mail: seankimjr811@gmail.com.

Accepted April 29, 2026
https://doi.org/10.70251/HYJR2348.4316

manipulation techniques (2, 3).

The raw generative power of GANs, however, brought with it a persistent challenge: controllability. Early architectures were notoriously difficult to steer; targeted edits such as widening a smile or rotating a head without altering identity, lighting, or background were largely intractable owing to feature entanglement in the model's internal representation. DragGAN was proposed as an interactive tool that addresses this control problem through point-based manipulation: users place a handle point on a semantic feature and a target point where they wish that feature to move, and the model iteratively updates the latent representation to satisfy the spatial constraint (4). This represents a meaningful step toward intuitive, user-directed manipulation of generative content.

Despite the qualitative impressiveness of DragGAN, evidenced by viral demonstrations and curated examples in the original publication, its performance has not been rigorously quantified across different types of subjects and manipulation scenarios (4). The precision required to realistically alter a human face, with its intricate geometry and high sensitivity to structural deviations, differs substantially from the precision needed to manipulate a full-body figure or a non-human subject. Research in face perception has demonstrated that faces engage specialized configural processing—involving sensitivity to first-order spatial relations, holistic integration, and fine-grained second-order relational information—that makes facial images uniquely susceptible to structural distortions compared with other object categories (5). Without objective metrics, the field is left with anecdotal evidence that may mask underlying limitations in the model's architecture.

This study addresses that gap by conducting a systematic, quantitative analysis of DragGAN's performance across four distinct image categories: human faces, dog faces, cat faces, and whole dog bodies. The objectives are threefold. First, the structural fidelity of manipulated images is evaluated using the Structural Similarity Index (SSIM), chosen for its alignment with human perception of structural error (6). Second, perceptual quality and generative diversity are assessed using a decomposed Inception Score, with the explicit goal of separating quality from diversity to diagnose potential mode collapse (7). Third, these metrics are compared across the four categories to identify domain-specific strengths and weaknesses of the DragGAN architecture. By providing a metric-based assessment, this work offers insights for developers refining generative algorithms, practitioners integrating such tools into applied workflows, and researchers investigating the limits of latent-space manipulation.

METHODS AND MATERIALS

Theoretical Background

The methodology of this study draws on three established components of the generative modeling literature: the GAN framework, the StyleGAN family of architectures, and the DragGAN manipulation procedure. The GAN framework formulates image synthesis as a minimax game between a generator G and a discriminator D, where G maps a latent noise vector z to a synthesized image and D distinguishes synthesized from real samples (1). StyleGAN and StyleGAN2 introduced a non-linear mapping network that transforms z into an intermediate latent space W in which factors of variation such as pose, identity, and lighting are approximately linearly separable; this disentanglement is the property that makes subsequent latent-space manipulation feasible (2, 3). DragGAN operates on a pre-trained StyleGAN2 generator and alternates two steps: a motion supervision step that updates the latent code w to drive intermediate feature maps toward the target point, and a point tracking step that updates handle-point coordinates by nearest-neighbor search in feature space, ensuring that the system follows the semantic feature rather than a fixed pixel coordinate (4).

Two evaluation metrics are central to the study. The Structural Similarity Index is a perceptual metric that compares two images on luminance, contrast, and structural correlation under a sliding Gaussian window; unlike Mean Squared Error, it penalizes distortions that break local structural coherence and is therefore well suited to detecting the geometric warping that aggressive latent-space manipulation may introduce (6). The Inception Score uses a pre-trained Inception-v3 classifier to assess two complementary properties of a generated set: a quality term that rewards low conditional entropy of class predictions, and a diversity term that rewards high entropy of the marginal class distribution (7, 8). Decomposing the score into these two components, rather than reporting only their combined value, permits explicit diagnosis of mode collapse, in which the generator produces high-quality but redundant outputs.

Dataset Curation and Preprocessing

All evaluation images were generated directly from pre-trained StyleGAN2 checkpoints, which ensured full

reproducibility and removed copyright considerations associated with external image sourcing. Human-face images were generated from a StyleGAN2 checkpoint trained on the Flickr-Faces-HQ (FFHQ) dataset, which contains 70,000 high-quality 1024×1024 images of human faces with substantial variation in age, ethnicity, accessories, and background (2). Dog-face and cat-face images were generated from the StyleGAN2 checkpoints trained on the Animal Faces-HQ (AFHQ) dataset (9). Whole-body dog images were generated from a publicly released Self-Distilled StyleGAN checkpoint at 1024×1024 resolution.

The final test set was stratified into four categories of 100 images each (N = 400 total): human faces (StyleGAN2-FFHQ, 512×512), dog faces (StyleGAN2-AFHQ-Dog, 512×512), cat faces (StyleGAN2-AFHQ-Cat, 512×512), and whole dog bodies (Self-Distilled StyleGAN, 1024×1024). The 512×512 resolution for facial categories is sufficient to expose structural artifacts and texture failures relevant to SSIM analysis while keeping computational cost manageable, and the 1024×1024 resolution for the whole-body category corresponds to the native output of the Self-Distilled StyleGAN checkpoint. To ensure compatibility with the pre-trained generators used by DragGAN, pixel values for all images were normalized from the integer range to the floating-point range [−1, 1], matching the Tanh output activations of the generator.

Experimental Design and Implementation

The evaluation used the official DragGAN implementation (4). To minimize selection bias, the dragging procedure was automated programmatically rather than performed manually. For each image, a pre-trained landmark detector was used to identify semantically meaningful handle points (dlib for faces; OpenPose for bodies). A target point was then generated by applying a random displacement vector to the handle point, with the magnitude constrained so that the target remained within image bounds and represented a plausible movement, such as displacing a mouth corner upward by approximately twenty pixels. The DragGAN optimization was executed for a maximum of 200 iterations or until the handle point reached the target, whichever occurred first.

Evaluation Protocols

Two protocols were applied to the manipulated outputs. Protocol 1 measured structural fidelity using SSIM, comparing each manipulated output to its corresponding pre-manipulation source image (6). SSIM was computed using an 11×11 Gaussian sliding window with standard deviation $\sigma = 1.5$, and the dynamic range was set to 2.0 to match the [−1, 1] normalization of the input images. A score of 1.0 indicates identity between source and output; scores are expected to drop modestly under successful manipulation, while a drastic drop (e.g., below 0.5) indicates that the manipulation has compromised the structural identity of the subject.

Protocol 2 measured perceptual quality and generative diversity using a decomposed Inception Score (7). All 400 manipulated images (100 per category) were passed through a pre-trained Inception-v3 classifier to obtain conditional class probability distributions p(y|x) (8). The quality component was computed as exp(E[H(p(y|x))]), reflecting the average classifier confidence; the diversity component was computed as exp(H(E[p(y|x)])), reflecting the spread of predictions over the 1,000 ImageNet classes (10). Reporting these components separately, rather than combining them into a single Inception Score, allows the explicit detection of mode collapse, in which the model produces high-quality but redundant outputs.

RESULTS

The aggregated evaluation metrics across the four subject categories, computed over 400 manipulated images (N = 100 per category), are summarized in Table 1.

Table 1. *Summary of evaluation metrics for DragGAN performance across four subject categories (N = 100 images per category; 400 total). SSIM denotes the Structural Similarity Index, which ranges from 0 (no structural similarity) to 1 (identical). IS (Quality) is the quality component of the decomposed Inception Score, computed as exp(E[H(p(y|x))]); IS (Diversity) is the diversity component, computed as exp(H(E[p(y|x)])), where p(y|x) is the conditional class distribution produced by a pre-trained Inception-v3 classifier trained on ImageNet. IQR denotes interquartile range. Domain mismatch between ImageNet classes and the generated subjects should be considered when interpreting absolute IS values.*

Subject Category	SSIM (mean)	IS — Quality	IS — Diversity
Human (Face)	0.30	3.38	184.20
Dog (Face)	0.23	4.03	424.59
Cat (Face)	0.21	3.55	340.09
Dog (Body)	0.33	3.34	425.73

Structural Fidelity

The SSIM results delineate a clear performance boundary for DragGAN that is largely defined by the frequency of the visual information being manipulated. The highest structural preservation was observed in the Dog (Body) category, with a mean SSIM of 0.33. Although this score still indicates substantial structural change relative to the source image, it was meaningfully higher than that of any facial category, suggesting that full-body manipulation is somewhat more tractable for the DragGAN architecture. The facial categories exhibited the lowest structural fidelity: Cat (Face) yielded the lowest mean SSIM at 0.21, followed by Dog (Face) at 0.23 and Human (Face) at 0.30. Faces are characterized by high-frequency, geometrically rigid relationships—the spacing of the eyes, the symmetry of the nose, and the local correlation of skin texture—and when the optimization attempts to drag a single semantic point, it frequently fails to propagate the intended change naturally to surrounding pixels, instead producing local warping or texture blurring that SSIM penalizes heavily.

Within the Cat (Face) category, a single outlier with a maximum SSIM of 0.884 was observed, well above the upper 95% confidence-interval bound of 0.560. This case is most plausibly explained by an automatic handle-target placement in which the two points fell in close spatial proximity, producing a near-identity transformation. Excluding this outlier does not materially change the distributional characteristics of the category, as the interquartile range (IQR: 0.089–0.333) remains stable.

Generative Diagnostics

Composite Inception Scores were modest across all categories, ranging from 3.34 (Dog Body) to 4.03 (Dog Face). The quality component, computed as the exponentiated negative conditional entropy of classifier predictions, was uniformly low (0.013–0.028), indicating that the Inception-v3 classifier was rarely confident in its class assignments for the generated images (8). This pattern is expected given the domain mismatch between the generated subjects—stylized faces and animal bodies—and the 1,000 object categories on which Inception-v3 was trained (10). The low quality component therefore reflects the limited applicability of ImageNet-class labels to these subjects rather than poor perceptual quality in absolute terms.

The diversity component, computed as the exponentiated marginal entropy of the aggregate class distribution, was high across all categories (Dog Body 425.73, Dog Face 424.59, Cat Face 340.09, Human Face 184.20). However, because none of the tested categories map cleanly to a single ImageNet class, these elevated values reflect distributional spread of low-confidence predictions across many unrelated categories rather than meaningful semantic diversity in the generated content. Taken together, the low quality scores and high diversity scores indicate that the decomposed Inception Score is of limited diagnostic value in this setting; we therefore caution against interpreting these diversity values as evidence of either substantive variety or mode collapse. Domain-appropriate alternatives, such as the Fréchet Inception Distance computed with classifiers fine-tuned on the relevant domains, would offer more interpretable diagnostics in future work (11).

DISCUSSION

Two findings emerge from the analysis. The first is the relative ordering of structural fidelity across categories, with whole-body subjects preserving more global structure than any of the facial categories. This ordering is consistent with a semantic-density account: full-body subjects are loose collections of articulated parts whose plausible configurations span a comparatively large region of the StyleGAN2 latent manifold, so a point-based drag aligns reasonably well with the kinematic degrees of freedom encoded in W. Faces, by contrast, occupy a narrower, more rigidly constrained region of the manifold, and the gradient-based latent updates that DragGAN performs frequently leave this region, producing outputs that remain face-like but no longer preserve the source identity (3, 5). The interpretation is consistent with prior observations that latent-space directions for facial attributes are highly non-linear and not globally disentangled (3).

The second finding concerns the Inception Score itself. In its standard formulation, IS assumes that the generated distribution is meaningfully covered by the 1,000 ImageNet classes used to train the Inception-v3 classifier (7, 8, 10). For domain-specific generators such as those used here, this assumption fails, and both the quality and diversity components behave in ways that do not cleanly track the underlying generative behavior. This is a known limitation of IS for non-ImageNet-aligned domains and argues for the use of Fréchet Inception Distance with domain-adapted feature extractors, or kernel-based metrics such as Kernel Inception Distance, in subsequent evaluations of point-based manipulation systems (11–13).

These results also bear on the broader question of which generative backend best supports interactive, point-based editing. Recent work has begun to apply DragGAN-style interaction paradigms to diffusion models, which are trained to denoise across the full data distribution and are theoretically less prone to mode collapse than GANs (14, 15). DragDiffusion and related methods transfer the handle-target interaction onto diffusion latents, and the structural degradation observed here in the GAN setting motivates further empirical comparison between GAN- and diffusion-based backends under matched evaluation protocols (16).

From an applied standpoint, the SSIM ordering suggests that DragGAN is most appropriate for tasks where high-level pose and composition matter more than identity preservation, such as concept iteration, scene blocking, or coarse animation prototyping. It is less well suited to tasks in which the source identity must be preserved, such as portrait retouching or any application with downstream identity-verification requirements, where the observed degradation in facial categories represents a meaningful risk.

CONCLUSION

This study presents a quantitative, metric-based evaluation of DragGAN across four subject categories, moving beyond the qualitative demonstrations that have characterized much of the existing literature on point-based manipulation. The Structural Similarity Index analysis shows that DragGAN's manipulation introduces substantial structural distortion in every category tested, with mean SSIM ranging from 0.21 (cat faces) to 0.33 (dog bodies), and that whole-body manipulation is more tractable than fine-grained facial editing. The decomposed Inception Score, while informative as a diagnostic exercise, proved to be of limited utility in this setting due to domain mismatch between the generated subjects and the ImageNet-trained Inception-v3 classifier; future evaluations should adopt domain-appropriate metrics such as the Fréchet Inception Distance with adapted feature extractors. The SSIM evidence, taken on its own, is sufficient to establish that DragGAN can produce visually coherent outputs while introducing significant structural change relative to the input—a finding with direct implications for applications requiring identity preservation. Addressing this limitation will be central to the next generation of intuitive image-manipulation tools, and the evidence reported here is consistent with a broader shift toward hybrid and diffusion-based backends that may better combine the interaction model of DragGAN with improved structural fidelity.

ACKNOWLEDGEMENTS

The author thanks the mentors and peers at Lumiere Education for their guidance during the research process, and acknowledges the open-source community for making the StyleGAN and DragGAN codebases available for academic inquiry, as well as the curators of the FFHQ, AFHQ, and ImageNet datasets.

FUNDING SOURCES

The author declares that no external funding was received for the conduct of this research or the preparation of this article.

CONFLICT OF INTEREST

The author declares no conflicts of interest related to this work.

REFERENCES

1. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, *et al*. Generative adversarial nets. *Adv Neural Inf Process Syst*. 2014; 27: 2672-80.
2. Karras T, Laine S, Aila T. A style-based generator architecture for generative adversarial networks. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2019: 4401-10. https://doi.org/10.1109/CVPR.2019.00453
3. Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T. Analyzing and improving the image quality of StyleGAN. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2020: 8107-16. https://doi.org/10.1109/CVPR42600.2020.00813
4. Pan X, Tewari A, Leimkühler T, Liu L, Meka A, Theobalt C. Drag your GAN: interactive point-based manipulation on the generative image manifold. *ACM SIGGRAPH 2023 Conf Proc*. 2023; Article 11: 1-11. https://doi.org/10.1145/3588432.3591500
5. Maurer D, Le Grand R, Mondloch CJ. The many faces of configural processing. *Trends Cogn Sci*. 2002; 6 (6): 255-60. https://doi.org/10.1016/S1364-6613(02)01903-4
6. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004; 13 (4): 600-12. https://doi.org/10.1109/TIP.2003.819861
7. Salimans T, Goodfellow I, Zaremba W, Cheung V,

Radford A, Chen X. Improved techniques for training GANs. *Adv Neural Inf Process Syst*. 2016; 29: 2234-42.
8. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. *Proc IEEE Conf Comput Vis Pattern Recognit*. 2016: 2818-26. https://doi.org/10.1109/CVPR.2016.308
9. Choi Y, Uh Y, Yoo J, Ha JW. StarGAN v2: diverse image synthesis for multiple domains. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2020: 8188-97. https://doi.org/10.1109/CVPR42600.2020.00821
10. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, *et al*. ImageNet large scale visual recognition challenge. *Int J Comput Vis*. 2015; 115 (3): 211-52. https://doi.org/10.1007/s11263-015-0816-y
11. Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Adv Neural Inf Process Syst*. 2017; 30: 6626-37.
12. Barratt S, Sharma R. A note on the Inception Score. arXiv preprint arXiv:1801.01973. 2018.
13. Bińkowski M, Sutherland DJ, Arbel M, Gretton A. Demystifying MMD GANs. *Int Conf Learn Represent*. 2018.
14. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst*. 2020; 33: 6840-51.
15. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-resolution image synthesis with latent diffusion models. *Proc IEEE/CVF Conf Comput Vis Pattern Recognit*. 2022: 10684-95. https://doi.org/10.1109/CVPR52688.2022.01042
16. Shi Y, Xue C, Liew JH, Pan J, Yan H, Zhang W, *et al*. DragDiffusion: harnessing diffusion models for interactive point-based image editing. arXiv preprint arXiv:2306.14435. 2023. https://doi.org/10.1109/CVPR52733.2024.00844