arXiv:2603.07066v1 [cs.CV] 7 Mar 2026

MedSteer: Counterfactual Endoscopic Synthesis via Training-Free Activation Steering

Trong-Thang Pham, Loc Nguyen, Anh Nguyen, Hien Nguyen, Ngan Le

No Institute Given

Abstract. Generative diffusion models are increasingly used for medical imaging data augmentation, but text prompting cannot produce *causal* training data. Re-prompting rerolls the entire generation trajectory, altering anatomy, texture, and background. Inversion-based editing methods introduce reconstruction error that causes structural drift. We propose **MedSteer**, a training-free activation-steering framework for endoscopic synthesis. MedSteer identifies a pathology vector for each contrastive prompt pair in the cross-attention layers of a diffusion transformer. At inference time, it steers image activations along this vector, generating counterfactual pairs from scratch where the *only* difference is the steered concept. All other structure is preserved by construction. We evaluate MedSteer across three experiments on Kvasir v3 and HyperKvasir. On counterfactual generation across three clinical concept pairs, MedSteer achieves flip rates of 0.800, 0.925, and 0.950, outperforming the best inversion-based baseline in both concept flip rate and structural preservation. On dye disentanglement, MedSteer achieves 75% dye removal against 20% (PnP) and 10% (h-Edit). On downstream polyp detection, augmenting with MedSteer counterfactual pairs achieves ViT AUC of 0.9755 versus 0.9083 for quantity-matched re-prompting, confirming that counterfactual structure drives the gain. Code will be made publicly available upon acceptance.

Keywords: Endoscopy · Generative AI · Counterfactual Explanation · Representation Steering · Data Augmentation · Causal AI

1 Introduction

A key challenge in endoscopic image analysis is training pathology detectors that respond to disease-specific features rather than confounding anatomy. A natural solution is to generate anatomy-matched pairs of diseased and healthy images so the detector learns only the targeted concept. Diffusion models can in principle generate such pairs, but text-to-image re-prompting between any two concepts ("polyp" to "normal", "colitis" to healthy mucosa, or dyed to undyed tissue) rerolls the entire generation trajectory, producing a completely different image.

Diffusion-based editing methods [22,31] start from a source image but rely on Denoising Diffusion Implicit Models (DDIM) inversion, an approximation that introduces reconstruction error. RadEdit [27] compounds this with mandatory

mask annotations per region. No workaround closes this gap. Even improved inversion [21] cannot eliminate drift, because re-entering a trajectory is inherently approximate. The requirement that non-targeted structure[1] be exactly preserved, not merely approximated, remains unmet.

We propose **MedSteer**, a training-free framework requiring no fine-tuning, source images, or annotations. It estimates pathology vectors from contrastive prompt pairs in the cross-attention space of a frozen diffusion transformer and steers activations along those vectors at inference time. With a shared noise seed, both images traverse the *same* trajectory, making all non-targeted concept structure identical by construction.

Our contributions are:

1. We introduce a **cosine-similarity-derived activation steering** mechanism that steers clinical concepts by removing only the concept-aligned component of each cross-attention token via a per-token gate. This gate reveals exactly where and when the model steers, providing **built-in spatial interpretability** absent from inversion-based approaches.
2. We demonstrate **inversion-free counterfactual pair generation** from scratch. Both images share the same noise trajectory through a frozen model. MedSteer outperforms existing state-of-the-art editing methods in non-targeted concept structure preservation.

2 Related Work

Diffusion-Based Editing and Medical Counterfactual Generation. Training-based editing approaches fine-tune the denoiser or auxiliary networks [14,13,16], while training-free methods manipulate attention maps [4,24], improve inversion fidelity [21,12,3], or use masks to localise edits [1,6]. In the medical domain, counterfactual and pathology-removal methods [25,29,9,11] and RadEdit [27] all follow the same generate $\rightarrow$ invert $\rightarrow$ edit pipeline, inheriting the fundamental limitation that DDIM inversion is an approximation and background drift is bounded but never eliminated. *MedSteer avoids this pipeline by generating both images from the same noise seed, guaranteeing non-targeted structure preservation.*
Concept Control via Latent Directions. Recent work shows that diffusion activations encode semantically meaningful linear directions [16,23,30,31] that can steer generation. SDID [17] and SAeUron [7] require extensive training, while concept-erasure methods fine-tune weights [10,15] or Low-Rank Adaptation (LoRA) adapters [20,19], all prohibitive when labelled medical data is scarce. None has been applied to medical imaging, where class labels encode *entangled* visual attributes that text cannot decompose, and utility must be validated on downstream clinical tasks rather than perceptual metrics alone. *MedSteer addresses this gap in that it requires no training, operates on entangled medical class labels, and validates utility on a downstream detection task.*

[1] By *non-targeted* we mean all structure unrelated to the steered concept: anatomy, texture, and background that should remain identical between the pair.

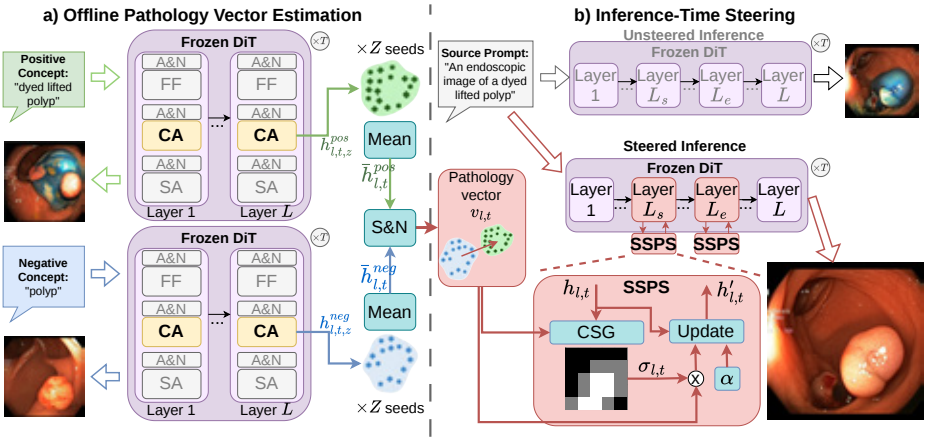

Fig. 1: MedSteer method pipeline. **A&N**: Add & Norm. **FF**: Feed Forward. **CA**: Cross-Attention. **SA**: Self-Attention. **(a) Offline Pathology Vector Estimation:** CA features $h_{l,t,z}$ are collected from the frozen DiT for positive and negative prompts across Z random seeds. A *Mean* step yields $\bar{h}^{pos}_{l,t}$ and $\bar{h}^{neg}_{l,t}$. a *S&N* (Subtract & Normalize) step then produces the unit pathology vector $v_{l,t}$ (dyed lifted polyp → polyp). **(b) Inference-Time Steering:** *Unsteered Inference* runs the frozen DiT unmodified. *Steered Inference* applies Spatially Selective Pathology Steering (SSPS) at layers $l \in \{L_s, \ldots, L_e\}$ across all T denoising steps. Inside SSPS, a *CSG* (Cosine-similarity gate) produces the per-token score $\sigma_{l,t}$, which is scaled by α and fed to an *Update* step that subtracts the aligned component from $h_{l,t}$, yielding the counterfactual activation $h'_{l,t}$. While both inference branches are shown together here, only the desired branch is executed in practice.

3 Methodology

In this section, we describe MedSteer through three stages: the backbone and intervention point (Sec. 3.1), Pathology Vector Estimation (Sec. 3.2), and Spatially Selective Pathology Steering (SSPS) (Sec. 3.3).

3.1 Backbone and Intervention Point

MedSteer builds on PixArt-α [5], a Diffusion Transformer (DiT) [26] with N=28 transformer layers. Each layer contributes self-attention, cross-attention (CA), and feed-forward terms to the residual stream. The CA output is the sole conduit through which textual semantics reach image tokens, making it the natural intervention point. A direction estimated from contrastive text prompts can be subtracted there to suppress a clinical concept without touching anatomy encoded elsewhere. We therefore intervene on CA outputs at layers $l \in \{L_s, \ldots, L_e\}$ across all denoising timesteps, with the layer range determined empirically in Sec. 4.5.

3.2 Pathology Vector Estimation

Contrastive prompt design. A clinical concept P is a pathological class label (e.g., Polyp, Normal Cecum, Ulcerative Colitis, Normal Z-line, Esophagitis, Dyed Lifted Polyp). For each P, we assemble Z positive/negative prompt pairs $\{(T_{pos}^{(z)}, T_{neg}^{(z)})\}$ that differ only in the presence of P (e.g., "`[context] of a dyed lifted polyp`" vs. "`[context] of a polyp`"), with context phrasing (e.g., "An endoscopic image") and random seed varied across pairs to marginalise out texture and viewpoint variation.

Pathology vector. For each prompt pair, we perform Z forward passes through the frozen model. *Mean*: the CA features $h_{l,t,n}$ are averaged over all Z seeds and all spatial tokens, yielding $\bar{h}_{l,t}^{pos}, \bar{h}_{l,t}^{neg} \in \mathbb{R}^{d_l}$ per layer l and timestep $t \in \{1, \ldots, T\}$. *Subtract & Normalize*: following the mean-difference principle [32], the L2-normalised contrastive difference gives the pathology vector $v_{l,t} = \frac{\bar{h}_{l,t}^{pos} - \bar{h}_{l,t}^{neg}}{\|\bar{h}_{l,t}^{pos} - \bar{h}_{l,t}^{neg}\|_2}$. This yields a unit vector per layer and timestep that captures the shared semantic difference attributable to P. The per-timestep parameterization is necessary because the diffusion process encodes concepts differently across denoising steps, with early timesteps establishing global structure and later steps refining fine-grained details. Since $v_{l,t}$ is concept-specific rather than image-specific, it is computed once offline and reused across all inference calls at no additional cost.

3.3 Spatially Selective Pathology Steering (SSPS)

Counterfactual setup. Unlike inversion-based methods, MedSteer requires no source image, no mask annotation, and no fine-tuning: both images are generated from scratch using the same noise seed z. Both branches share the same prompt T_{pos}. While *Unsteered Inference* runs the frozen DiT unmodified, *Steered Inference* inserts SSPS at layers $l \in \{L_s, \ldots, L_e\}$ at every denoising timestep to suppress the target pathology. Because both branches share the same seed z, any structural difference between the paired outputs arises solely from the SSPS intervention, making the comparison a true minimal-edit counterfactual.

Cosine-similarity gate (CSG). Because pathological changes are spatially focal, a global scalar subtraction would over-suppress non-target tokens. We score each token's alignment with the pathology concept via $\cos(h_{l,t}, v_{l,t}) = \langle h_{l,t}, v_{l,t} \rangle / \|h_{l,t}\|_2$, where $h_{l,t}$ denotes the activation of any visual token at layer l, timestep t. We deliberately drop the $1/\|h_{l,t}\|_2$ normalizer, arriving at the dot product $\langle h_{l,t}, v_{l,t} \rangle$, so that tokens expressing the concept more strongly receive proportionally stronger steering: $\sigma_{l,t} = \max\big(\langle h_{l,t}, v_{l,t} \rangle, 0\big)$. The $\max(\cdot, 0)$ operator ensures that only tokens positively aligned with $v_{l,t}$ are modified, and tokens with zero or negative alignment pass through unchanged. We then update each token's activation by subtracting the pathology-aligned component scaled by α (the *Update* step in Fig. 1), leaving all orthogonal components (anatomy, texture, viewpoint) intact: $h'_{l,t} = h_{l,t} - \alpha\, \sigma_{l,t}\, v_{l,t}$, where α is the steering strength.

4 Experiments

4.1 Experimental Setup

Datasets and model. We use Kvasir v3 [28] (8,000 gastrointestinal endoscopy images, eight classes) for generative training, pathology vector construction, and all generative evaluations. HyperKvasir [2] provides the held-out test set for downstream polyp detection. The generative backbone is PixArt-α [5] fine-tuned with LoRA (rank r=64) on Kvasir.

Pathology vector construction. For each concept pair (T_{pos}, T_{neg}) we generate N=50 images per prompt and apply the method in Sec. 3 across layers $l \in \{8, \ldots, 16\}$ and all timesteps t to obtain $v_{l,t}$ for each setting.

Baselines and metrics. We compare **Re-prompting** (i.e. generate another image using a different prompt with the same random seed), and two DDIM-inversion-based editing methods: **Plug-and-Play (PnP)** [31] and **h-Edit** [22] (RadEdit [27] is excluded as it requires per-image mask annotations). A pre-trained oracle classifier (area under the ROC curve, AUC = 0.98) measures *concept flip rate* (fraction of outputs no longer classified as the source class) and *confidence shift* Δp (mean drop in source-class probability). Background preservation is assessed via Bg-LPIPS [34], Bg-SSIM, and Bg-PSNR (each restricted to the background, i.e., non-lesion, region isolated by UNet++ [35] (Dice = 0.912)) for counterfactual experiments. Dye disentanglement additionally reports results under Segformer [33] (Dice = 0.921).

4.2 Downstream Polyp Detection

In the first experiment, we assess whether MedSteer-generated augmentations improve downstream polyp detection. We train binary polyp detectors (ConvNeXt [18] and ViT [8]) under five augmentation methods: **Real-only** (20 images/class), **Re-prompting**, **PnP**, **h-Edit**, and **MedSteer**. Each augmention method adds N=1,000 synthetic images so that differences reflect augmentation *structure*, not *quantity*. Since HyperKvasir is an extension of Kvasir, we remove all exact and near-duplicate matches from the test set before evaluation to ensure a strictly out-of-distribution assessment. Table 1 shows that MedSteer achieves

Table 1: Downstream polyp detection on HyperKvasir (N=1,000 synthetic images per condition). CF: counterfactual generation.

		ConvNeXt		**ViT**	
Methods	**CF?**	**F1** ↑	**AUC** ↑	**F1** ↑	**AUC** ↑
Real-only	✗	$0.8977_{\pm.019}$	$0.8469_{\pm.023}$	$0.9062_{\pm.022}$	$0.8942_{\pm.026}$
Re-prompting	✗	$0.9038_{\pm.016}$	$0.8714_{\pm.020}$	$0.9147_{\pm.018}$	$0.9083_{\pm.022}$
PnP	✓	$0.9174_{\pm.014}$	$0.9086_{\pm.017}$	$0.9437_{\pm.013}$	$0.9518_{\pm.017}$
h-Edit	✓	$0.9097_{\pm.015}$	$0.8902_{\pm.019}$	$0.9295_{\pm.016}$	$0.9312_{\pm.020}$
MedSteer	✓	$\mathbf{0.9263}_{\pm.012}$	$\mathbf{0.9341}_{\pm.015}$	$\mathbf{0.9591}_{\pm.011}$	$\mathbf{0.9755}_{\pm.013}$

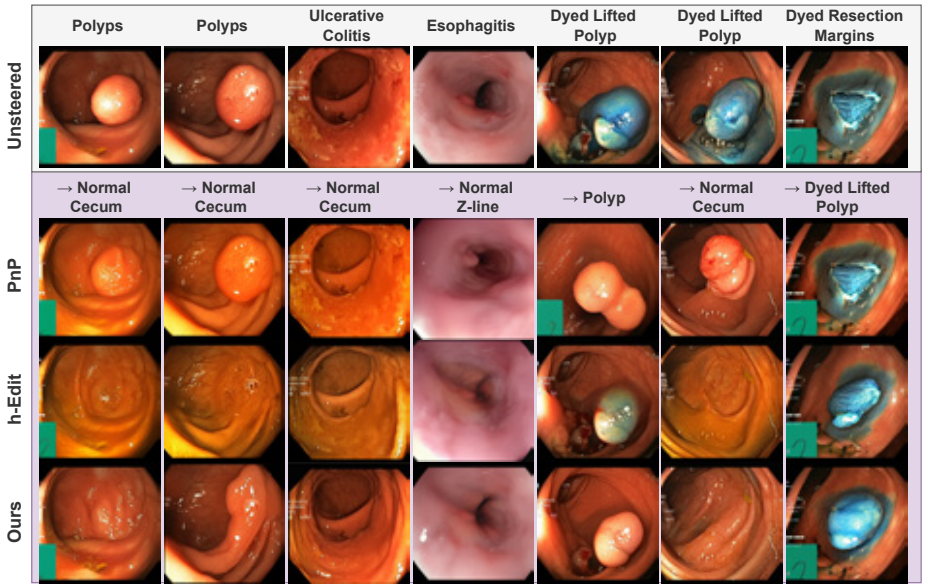

Fig. 2: Qualitative comparison across concept pairs. Rows: Unsteered, PnP, h-Edit, and MedSteer (Ours). Columns show Polyp → Normal Cecum, Ulcerative Colitis → Normal Cecum, Esophagitis → Normal Z-line, and dye steerings.

the best AUC under both backbones. This proves that paired supervision forces detectors to learn useful pathology features that generalize well into OOD setting.

4.3 Counterfactual Generation Across Clinical Concepts

In the second experiment, we evaluate concept steering quality across multiple clinical concept pairs, measuring concept flip rate and background preservation. We evaluate three concept pairs: (A) Polyp ↔ Normal Cecum, (B) Ulcerative Colitis ↔ Normal Cecum, and (C) Esophagitis ↔ Normal Z-line (N=1,000 images, α=±2.5). Background metrics are reported for pair (A) only.

MedSteer leads on pair (A): Flip = 0.800, Δp=0.721, Bg-LPIPS = 0.1449, and best Bg-SSIM and Bg-PSNR. Inversion-based methods accumulate DDIM reconstruction error, degrading both concept-change rate and background fidelity. Pairs (B) and (C) yield flip rates of 0.925 and 0.950, confirming generalisation across clinical concepts and anatomical sites (Table 2 and Fig. 2).

4.4 Dye Disentanglement

In the third experiment, we test disentanglement of co-occurring attributes (polyp morphology and dye staining) when text-based disentanglement is impossible. The "Dyed Lifted Polyp" class encodes two co-occurring attributes (polyp morphology and indigo carmine staining) with no "undyed" counterpart in the training

Table 2: Counterfactual generation (UC: Ulcerative Colitis, Esoph.: Esophagitis). Bg metrics are not applicable for pairs (B) and (C): the polyp segmentor cannot produce valid masks for diffuse mucosal inflammation or esophageal anatomy. Re-prompting is excluded because it rerolls the trajectory and produces different anatomy.

	(A) Polyp → Norm. Cecum					(B) UC → Normal Cecum		(C) Esoph. → Normal Z-line	
Method	**Flip ↑**	**Δp ↑**	**Bg-LPIPS ↓**	**Bg-SSIM ↑**	**Bg-PSNR ↑**	**Flip ↑**	**Δp ↑**	**Flip ↑**	**Δp ↑**
PnP	0.627	0.543	0.1518	0.8506	19.32	0.682	0.617	0.718	0.651
h-Edit	0.714	0.638	0.1565	0.8222	19.51	0.741	0.663	0.796	0.712
MedSteer	**0.800**	**0.721**	**0.1449**	**0.8951**	**22.95**	**0.925**	**0.902**	**0.950**	**0.928**

Table 3: Dye disentanglement under two background segmentors. DDR = dye detection rate (lower = dye removed). Eff. LPIPS = LPIPS/(1−DDR) (↓). Eff. SSIM = (1−DDR)×SSIM, Eff. PSNR = (1−DDR)×PSNR (↑).

	UNet++				Segformer		
Method	**Eff. LPIPS ↓**	**Eff. SSIM ↑**	**Eff. PSNR ↑**	**DDR ↓**	**Eff. LPIPS ↓**	**Eff. SSIM ↑**	**Eff. PSNR ↑**
h-Edit	0.4540	0.0957	2.78	0.900	0.5190	0.0952	2.73
PnP	0.3021	0.1872	4.93	0.800	0.3235	0.1865	4.82
MedSteer	**0.2027**	**0.6584**	**15.57**	**0.250**	**0.2193**	**0.6521**	**15.18**

distribution, making text-based disentanglement impossible. We construct a dye pathology vector from T_{pos}="dyed lifted polyps" and T_{neg}="polyps". Because the pair differs only in dye presence, the resulting vector is orthogonal to polyp morphology by construction, suppressing only the dye attribute while preserving polyp structure. MedSteer achieves a dye detection rate (DDR, fraction of outputs still classified as dyed) of 0.250, substantially outperforming h-Edit (DDR = 0.900) and PnP (DDR = 0.800) (Table 3 and Fig. 2). MedSteer also natively supports compositional edits (e.g., Dyed Lifted Polyp → Normal Cecum) with strict anatomical consistency.

4.5 Ablation Studies

Table 4 sweeps layer windows at fixed α=2.5, showing that layers 8–16 are the semantic formation zone. Windows entirely outside this range yield near-zero flip, and a width-12 window at 0–12 achieves only 37.7% flip vs. 70.0% for 8–20, confirming a position effect. Table 4 shows flip rises sharply from α=2.0 (51.3%) to α=2.5 (80.0%) and degrades at α=3.0 (70.0%). The pathology vector stabilises after ∼50 seeds (Table 4). Per-token cosine similarity scores $\sigma_{l,t}$ provide a built-in spatial explanation, with the activation footprint being broad at early diffusion steps and contracting to sparse patches by the final steps.

Table 4: Ablation studies (Polyp → Normal Cecum). **(a)** L_s–L_e ablation (α=2.5, N=50), †W = window width, i.e. number of layers steered. **(b)** Steering strength (layers 8–16, N=50). **(c)** Number of seeds (layers 8–16, α=2.5).

(a) L_s–L_e ablation

Layers	W†	Flip ↑	Δp ↑
0–8	8	0.025	0.018
8–16	**8**	**0.800**	**0.721**
12–20	8	0.013	0.009
0–12	12	0.377	0.396
8–20	12	0.700	0.623

(b) Steering strength

α	Flip ↑	Δp ↑
0.5	0.012	0.008
1.0	0.038	0.021
2.0	0.513	0.516
2.5	**0.800**	**0.721**
3.0	0.700	0.625

(c) Number of seeds

N	Flip ↑	Δp ↑
30	0.762	0.689
50	**0.800**	**0.721**
80	0.793	0.715
100	0.791	0.712

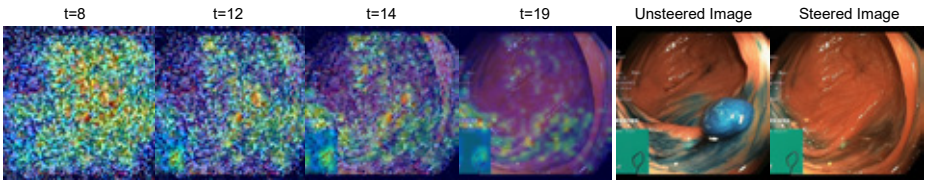

Fig. 3: *Left*: Per-token cosine similarity maps $\sigma_{8,t}$ at layer 8 for selected diffusion steps $t \in \{8, 12, 14, 19\}$ (left to right). Warmer colours indicate stronger alignment with the pathology vector. *Right*: corresponding unsteered (dyed lifted polyps) and steered (normal cecum) endoscopy images.

4.6 Interpretability

The per-token cosine similarity score $\sigma_{l,t}$ can be rendered directly as a spatial map, yielding a built-in visualisation of *where* the model steers at each diffusion step. As Fig. 3 shows, the activation footprint covers nearly the entire image at early steps and becomes progressively sparser toward the final steps, reflecting that fewer tokens require adjustment once the concept has been established. Inversion-based baselines have no analogous interpretable gate.

5 Conclusion

We presented **MedSteer**, a training-free steering framework for medical image synthesis. It extracts pathology vectors from contrastive prompt pairs in the cross-attention space of a frozen diffusion transformer, then steers activations via cosine similarity to produce counterfactuals, with no source images, DDIM inversion, mask annotations, or retraining. On Kvasir v3 and HyperKvasir, MedSteer outperforms baselines on downstream polyp detection, counterfactual generation, and dye disentanglement. Additionally, per-token cosine similarity provides built-in spatial interpretability at each diffusion step. Future work will extend to 3D volumetric data, video endoscopy sequences, and cross-institutional deployment.

References

1. Avrahami, O., Lischinski, D., Fried, O.: Blended diffusion for text-driven editing of natural images. In: CVPR. pp. 18208–18218 (2022)
2. Borgli, H., Thambawita, V., Smedsrud, P.H., Hicks, S., Jha, D., Eskeland, S.L., Randel, K.R., Pogorelov, K., Lux, M., Nguyen, D.T.D., et al.: Hyperkvasir, a comprehensive multi-class image and video dataset for gastrointestinal endoscopy. Scientific data **7**(1), 283 (2020)
3. Brack, M., Friedrich, F., Kornmeier, K., Tsaban, L., Schramowski, P., Kersting, K., Passos, A.: Ledits++: Limitless image editing using text-to-image models. In: CVPR. pp. 8861–8870 (2024)
4. Cao, M., Wang, X., Qi, Z., Shan, Y., Qie, X., Zheng, Y.: Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465 (2023)
5. Chen, J., Yu, J., Ge, C., Yao, L., Xie, E., Wu, Y., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis (2023)
6. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: DiffEdit: Diffusion-based semantic image editing with mask guidance (2022)
7. Cywiński, B., Deja, K.: SAeuron: Interpretable concept unlearning in diffusion models with sparse autoencoders. In: ICML (2025)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fontanella, A., Mair, G., Wardlaw, J., Trucco, E., Storkey, A.: Diffusion models for counterfactual generation and anomaly detection in brain images, `http://arxiv.org/abs/2308.02062`
10. Gandikota, R., Materzynska, J., Fiotto-Kaufman, J., Bau, D.: Erasing concepts from diffusion models. In: ICCV (2023)
11. Gu, Y., Yang, J., Usuyama, N., Li, C., Zhang, S., Lungren, M.P., Gao, J., Poon, H.: Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys. arXiv preprint arXiv:2310.10765 (2023)
12. Huberman-Spiegelglas, I., Kulikov, V., Michaeli, T.: An edit friendly ddpm noise space: Inversion and manipulations. In: CVPR. pp. 12469–12478 (2024)
13. Kawar, B., Zada, S., Lang, O., Tov, O., Chang, H., Dekel, T., Mosseri, I., Irani, M.: Imagic: Text-based real image editing with diffusion models. In: CVPR. pp. 6007–6017 (2023)
14. Kim, G., Kwon, T., Ye, J.C.: Diffusionclip: Text-guided diffusion models for robust image manipulation. In: CVPR. pp. 2426–2435 (2022)
15. Kumari, N., Zhang, B., Wang, S., Shechtman, E., Zhang, R., Zhu, J.: Ablating concepts in text-to-image diffusion models. In: ICCV (2023)
16. Kwon, M., Jeong, J., Uh, Y.: Diffusion models already have a semantic latent space. In: ICLR (2023)
17. Li, H., Shen, C., Torr, P., Tresp, V., Gu, J.: Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In: CVPR (2024)
18. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)

19. Lu, S., Wang, Z., Li, L., Liu, Y., Kong, A.W.: MACE: mass concept erasure in diffusion models. In: CVPR (2024)
20. Lyu, M., Yang, Y., Hong, H., Chen, H., Jin, X., He, Y., Xue, H., Han, J., Ding, G.: One-dimensional adapter to rule them all: Concepts, diffusion models and erasing applications. In: CVPR (2024)
21. Mokady, R., Hertz, A., Aberman, K., Pritch, Y., Cohen-Or, D.: Null-text inversion for editing real images using guided diffusion models. In: CVPR. pp. 6038–6047 (2023)
22. Nguyen, T., Do, K., Kieu, D., Nguyen, T.: h-edit: Effective and flexible diffusion-based editing via doob's h-transform. CVPR (2025)
23. Park, Y., Kwon, M., Choi, J., Jo, J., Uh, Y.: Understanding the latent space of diffusion models through the lens of riemannian geometry. In: NeurIPS (2023)
24. Parmar, G., Kumar Singh, K., Zhang, R., Li, Y., Lu, J., Zhu, J.Y.: Zero-shot image-to-image translation. In: ACM SIGGRAPH. pp. 1–11 (2023)
25. Pawlowski, N., Castro, D.C., Glocker, B.: Deep structural causal models for tractable counterfactual inference. In: NeurIPS. vol. 33, pp. 857–869 (2020)
26. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: ICCV (2023)
27. Pérez-García, F., Bond-Taylor, S., Sanchez, P.P., van Breugel, B., Castro, D.C., Sharma, H., Salvatelli, V., Wetscherek, M.T.A., Richardson, H., Lungren, M.P., Nori, A., Alvarez-Valle, J., Oktay, O., Ilse, M.: RadEdit: Stress-testing biomedical vision models via diffusion image editing. In: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (eds.) ECCV. pp. 358–376. Springer Nature Switzerland, Cham (2025)
28. Pogorelov, K., Randel, K.R., Griwodz, C., Eskeland, S.L., de Lange, T., Johansen, D., Spampinato, C., Dang-Nguyen, D.T., Lux, M., Schmidt, P.T., et al.: Kvasir: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference. pp. 164–169 (2017)
29. Sanchez, P., Kascenas, A., Liu, X., O'Neil, A.Q., Tsaftaris, S.A.: What is healthy? generative counterfactual diffusion for lesion localization. In: MICCAI Workshop on Deep Generative Models. pp. 34–44. Springer (2022)
30. Si, C., Huang, Z., Jiang, Y., Liu, Z.: Freeu: Free lunch in diffusion u-net. In: CVPR (2024)
31. Tumanyan, N., Geyer, M., Bagon, S., Dekel, T.: Plug-and-play diffusion features for text-driven image-to-image translation. In: CVPR. pp. 1921–1930 (2023)
32. Wu, Z., Arora, A., Geiger, A., Wang, Z., Huang, J., Jurafsky, D., Manning, C.D., Potts, C.: Axbench: Steering llms? even simple baselines outperform sparse autoencoders. arXiv preprint arXiv:2501.17148 (2025)
33. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. vol. 34, pp. 12077–12090 (2021)
34. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018)
35. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: International workshop on deep learning in medical image analysis. pp. 3–11. Springer (2018)