Evaluating Concept-guided Visual Counterfactual Generation in Medical Imaging

Anonymous

Anonymous Institution

Abstract. Visual counterfactual explanations (VCEs) offer an intuitive approach to explainability in medical image classification by generating minimally modified images that flip a classifier's prediction, revealing the pathological features that drive its decisions. We evaluate concept-guided VCE generation in which Concept Activation Vectors (CAVs) steer latent representations along clinically meaningful directions across three generative backbones: a domain-specific Medical Autoencoder (AE), and the VAE components of Stable Diffusion v1.5 and Stable Diffusion XL (SDXL) [21]. Counterfactuals are assessed on CheXpert and NIH ChestX-ray8 across five pathologies using Classifier Flip Rate (CFR) and Fréchet Radiomic Distance (FRD), evaluated with two independent classifiers (XRV and Ark+ [22]) under Platt calibration [23]. We find that SDXL achieves the highest concept steering performance, with a mean curtailment CFR of 94% on CheXpert under Ark+. We further demonstrate that previously reported near-zero CFR for Stable Diffusion was a calibration artefact, and that properly calibrated diffusion backbones substantially outperform the Medical AE on all semantic concept metrics.

Keywords: Explainable AI · Visual Counterfactual Explanations · Concept Activation Vectors · Medical Imaging · Chest X-ray · Latent Space Traversal · Stable Diffusion · Medical AutoEncoder · SDXL · Ark+.

1 Introduction

Trust is a fundamental requirement for the adoption of artificial intelligence in high-stakes clinical settings. Deep learning models have demonstrated remarkable performance on medical image classification, in some cases achieving or exceeding the diagnostic accuracy of experienced clinicians [1]. Yet their integration into routine clinical practice remains limited, largely because their internal reasoning is unclear: predictions are issued without any explanation that a clinician can verify, challenge, or learn from. In safety-critical domains such as radiology, the inability to understand a model's decision-making process is a significant barrier to clinical trust and regulatory acceptance.

Explainable AI (XAI) methods aim to bridge this gap, yet standard saliency approaches face serious concerns regarding their reliability as faithful explanations

Code to reproduce this experiment is available here: https://anonymous.4open.science/status/Concept-guided-visual-counterfactuals-C060

[2], [3]. Visual Counterfactual Explanations (VCEs) offer a more principled alternative by identifying the modifications that alter a model's prediction, a framing that aligns with clinical reasoning. However, standard VCEs often operate on binary labels rather than the specific pathological features, such as cardiac enlargement, used by clinicians.

Concept Activation Vectors (CAVs) [4] address this by identifying directions in a latent space corresponding to user-defined clinical concepts. When applied to latent space traversal, these vectors enable the generation of counterfactuals that exaggerate or curtail specific features, making them both visually and clinically interpretable. Building on [5], who demonstrated this approach using a Medical Autoencoder (AE), we address their finding that generative backbone quality is a primary limiting factor.

This work conducts a systematic evaluation of concept-guided VCEs using three distinct backbones: the domain-specific Medical AE, the VAE component of Stable Diffusion v1.5 (SD) [6] and the larger-capacity VAE component of Stable Diffusion XL (SDXL) [21]. Across the CheXpert [8] and NIH ChestX-ray8 [9] datasets, we assess performance using Classifier Flip Rate (CFR), Fréchet Radiomic Distance (FRD) [10], and inter-dataset CAV stability.

The contributions of this paper are as follows:

- We present the first systematic, multi-backbone evaluation of concept-guided VCE generation, comparing a domain-specific Medical AE against the general-purpose latent spaces of SD and SDXL [21] across two datasets and three complementary metrics, with two independent classifiers (XRV and Ark+ [22]).
- We show that, once properly calibrated, general-purpose diffusion backbones substantially outperform the domain-specific autoencoder on semantic concept steering, reframing generative capacity rather than domain specialisation as the primary driver of counterfactual quality.
- We identify and correct a calibration artefact responsible for the previously reported near-zero concept-steering performance of Stable Diffusion, demonstrating that faithful VCE evaluation requires probability calibration and multiple independent classifiers.
- We expose the CAV extraction strategy (LR vs. MD) as a significant, previously underappreciated factor, and characterise the concept-strength versus spatial-fidelity trade-off per pathology, confirming the cross-dataset stability of CAV directions for all backbones.

2 Related Work

Latent space traversal. A prominent line of work generates explanations by traversing the latent space of a generative model, variously using GANs, diffusion processes, or prototype-guided steering to produce sparse or high-fidelity counterfactuals [11–16]. Most directly relevant is the Latent Shift method of Cohen et al. (2021), which decouples the generative and classification components, driving a pre-trained autoencoder with classifier gradients to produce counterfactual

chest X-rays; we build on this paradigm but replace gradient-based steering with concept-based steering via CAVs.

Concept Activation Vectors. [4] introduced CAVs for quantifying the influence of user-defined concepts: training a linear classifier to separate positive and negative concept examples in a model's latent space identifies a direction corresponding to that concept, letting a user ask not just where a model looks but whether it responds to specific semantic features. [18] extend this idea to large language models via In-Context Vectors that steer behaviour by shifting latent representations from demonstration examples, highlighting the broad applicability of concept-based latent steering across modalities. We adopt CAVs as the steering mechanism for latent traversal, enabling targeted manipulation of specific clinical concepts rather than class-level label flipping.

Concept-guided VCE generation and backbones. [5] combined CAVs with latent space traversal to generate concept-guided VCEs for chest X-ray classification, using the ResNet-based Medical AE from TorchXRayVision [19]; they showed CAV directions in the AE latent space are stable across random data subsets and generalisable across CheXpert [8] and NIH [9], with strong localisation for large pathologies such as cardiomegaly but clear AE-decoder limitations for smaller ones such as atelectasis. They identified reconstruction fidelity as the primary bottleneck and explicitly proposed diffusion-based models for future work. [20] similarly noted that pixel-level overlap metrics are poor proxies for human evaluation, motivating a broader suite including CFR and FRD [10]. Generated counterfactual quality is fundamentally bounded by backbone capacity: simple autoencoders blur due to MSE averaging, whereas Latent Diffusion Models [6] pair a VAE encoder-decoder with a denoising diffusion process in a compressed latent space for high-fidelity synthesis at scale. SD [6], pre-trained on LAION-5B [7], provides a rich, semantically structured latent space despite no explicit medical knowledge, while SDXL [21] scales the architecture with a significantly larger UNet backbone and refined VAE, yielding a higher-capacity latent space that may support richer concept directions. Whether such general-purpose latent spaces enable more effective concept-guided steering than a domain-specific medical autoencoder, and how backbone capacity interacts with steering, are the central empirical questions of this work.

3 Methodology

Our method generates visual counterfactuals in two stages across three generative backbones: a domain-specific Medical AE, the VAE component of SD, and the larger-capacity VAE component of SDXL [21].

CAV generation. We construct CAVs in each model's latent space using a linear classifier.

CAV-guided traversal. We then shift latent samples along these vectors to exaggerate or curtail clinical features, and decode them into visual counterfactuals.

Concept Vector Generation. We evaluate two CAV extraction strategies. The **Linear Regression (LR)** method trains a linear classifier in the flattened

embedding space to separate positive and negative representations, taking the weight vector as the concept direction. The **Mean Direction (MD)** method takes the concept vector as the normalised difference of class centroids, requiring no classifier training.

For statistical stability we compute the vectors across 5 independent subsets (500 balanced samples each). Rather than assuming latent space symmetry, we explicitly compute and L2-normalise two distinct directional vectors per concept, $\mathbf{v}_{exaggerate}$ and $\mathbf{v}_{curtail}$.

Visual Explanations. Building on the Latent Shift methodology [17], a sample is encoded into the latent space and its representation shifted by adding the corresponding concept vector. To prevent under- or over-manipulation, a dynamic line-search iteratively shifts the latent until a pretrained classifier registers a predefined target change ($\Delta = 0.15$) in prediction probability, bounded by an early-stopping patience mechanism, before decoding to the counterfactual image. The spatial attribution map is the pixel-wise absolute difference between the baseline reconstruction and the final counterfactual.

4 Evaluation

4.1 Experimental Setup

Each CAV linear classifier uses an 80/20 stratified train/test split with no feature scaling, following the reference implementation [4]. The line search terminates when any of three conditions is met: the probability shift exceeds the target delta of 0.15 from the backbone reconstruction baseline, the patience limit of three consecutive wrong-direction steps is reached, or the maximum alpha is reached. The target delta of 0.15 is set slightly above the CFR flip threshold of 0.1 to account for the difference between the backbone reconstruction baseline used during line search and the original image baseline used in CFR evaluation.

4.2 Datasets

CheXpert [8] is a large-scale chest X-ray dataset containing 224,316 frontal and lateral radiographs from 65,240 patients collected at Stanford University Medical Center. Labels are automatically extracted from radiology reports, giving binary labels for 14 pathologies including Cardiomegaly, Edema, Consolidation, Atelectasis and Pleural Effusion. We use frontal view images only, following standard practice. Counterfactual evaluation uses 50 positive cases per pathology for curtailment and 50 negative cases for exaggeration, selected to ensure label confidence.

NIH ChestX-ray8 [9] comprises 108,948 frontal-view chest X-rays from 32,717 unique patients collected at the NIH Clinical Center. Disease labels for 14 pathologies are taken from associated radiology reports. Notably, this dataset forms part of the training data for the TorchXRayVision Medical AE [19], meaning

NIH images are in distribution for this backbone, a factor that contributes to the stronger AE reconstruction and lower FRD scores observed on this dataset compared to CheXpert.

4.3 Evaluation Metrics

For each backbone and dataset combination, we generate counterfactuals in both the exaggeration direction (amplifying the target pathology concept) and the curtailment direction (suppressing it).

We assess the quality and semantic validity of the generated counterfactuals using three complementary metrics: inter-dataset CAV Stability measures the generalisation of our concept vectors, while FRD and CFR assess the visual explanations directly.

CAV Stability is measured as the cosine similarity between the CheXpert and NIH concept vectors for each pathology and backbone, providing a measure of how well the learned concept directions generalise across datasets.

Fréchet Radiomic Distance measures the distance between the radiomic features of each counterfactual and its corresponding source image. Lower FRD indicates a more conservative edit relative to the baseline; higher FRD reflects a larger concept-driven modification.

Classifier Flip Rate measures the proportion of counterfactuals that produce a probability shift greater than 0.1 from the baseline, evaluated separately for exaggeration and curtailment directions. A flip is considered successful when the predicted probability for the target pathology increases by 0.1 for exaggeration, or decreases by 0.1 for curtailment. We evaluate CFR under two classifiers: the TorchXRayVision DenseNet (XRV) [19] and the Ark+ chest X-ray classifier [22]. We apply Platt scaling [23] to both classifiers to correct systematic probability offsets in the raw logits before threshold evaluation.

Counterfactuals are generated for n=50 cases per direction, pathology, and dataset (positive for curtailment, negative for exaggeration). For each backbone we evaluate both the LR and MD CAV extraction methods.

5 Experimental results and discussion

5.1 Generalization of Concept Vector

(i) Cosine Similarity (CAV Stability). CAV directions generalise meaningfully across datasets for all backbones, replicating the cross-dataset stability of [5]. For the AE and SD backbones, inter-dataset cosine similarities (Table 1) lie substantially above the random baseline across all five pathologies, confirming that the learned directions capture dataset-independent structure. The AE leads for four of five pathologies (peak: Cardiomegaly 0.234), consistent with its domain-specific latent space encoding pathology-relevant structure more directly; SD leads only for Effusion (0.145 vs. 0.123). SDXL (LR method) falls substantially below AE and SD for Cardiomegaly but is comparable for Consolidation (0.059)

Table 1. Inter-dataset CAV cosine similarity between CheXpert and NIH concept vectors. Random baseline shows expected similarity under null alignment. SDXL values use the LR method, scaled variant.

Pathology	AE		SD		SDXL	
	CheX↔NIH ↑	Random	CheX↔NIH ↑	Random	CheX↔NIH ↑	Random
Cardiomegaly	**0.234**	0.001	0.173	0.000	0.083	0.000
Edema	**0.161**	0.008	0.093	0.001	0.020	0.009
Effusion	0.123	0.001	**0.145**	0.001	0.060	−0.005
Consolidation	**0.102**	0.002	0.097	0.002	0.059	0.011
Atelectasis	**0.085**	−0.006	0.039	0.003	0.018	−0.004

Table 2. Mean FRD per backbone, variant, and method, averaged over curtail and exaggerate directions (n=100 images). Lower FRD indicates less deviation from the source image.

Backbone	Variant	CheXpert ↓		NIH ↓	
		LR	MD	LR	MD
AE	–	**3.78**	5.01	**3.72**	4.44
SD	scaled	7.40	5.48	7.46	4.95
	unscaled	6.38	**4.97**	6.43	**4.51**
SDXL	scaled	8.09	6.57	7.63	5.95
	unscaled	6.29	5.63	6.06	4.98

and Effusion (0.060), suggesting its larger general-purpose latent space encodes concept directions that are somewhat less dataset-invariant than the domain-specific AE.

5.2 Visual Explanations

(i) Fréchet Radiomic Distance FRD quantifies how much the radiomic texture of a counterfactual departs from its source image on a per-pair basis; lower values indicate a more conservative edit, while higher values reflect a larger concept-driven modification. FRD (Table 2) reveals a clear backbone–method interaction. The Medical AE produces the most conservative counterfactuals regardless of method (LR: 3.78 CheXpert, 3.72 NIH), reflecting the limited generative capacity of the AE decoder rather than superior radiological fidelity. Among diffusion backbones, MD consistently yields lower FRD than LR, with SD (unscaled, MD) achieving the lowest diffusion-model FRD (4.97 CheXpert; 4.51 NIH), comparable to the AE under MD. SDXL produces higher FRD than SD under both methods and both scale variants, indicating its larger-capacity latent space drives more aggressive concept-directed edits. The scaled variants of both SD and SDXL show notably higher LR FRD than their unscaled counterparts, suggesting latent scaling amplifies the spatial extent of the concept shift. This motivates unscaled variants when image fidelity is a priority and scaled variants when maximal concept strength is needed.

(ii) Classifier Flip Rate. CFR measures the proportion of generated counterfactuals that produce a classifier probability shift of at least 0.1, serving as the primary indicator of whether concept-guided steering successfully alters the model's decision in the intended direction. *Ark+/Platt results (Table 3, top).* The Medical AE achieves modest curtailment (mean CFR: 21.2% CheXpert; 32.4% NIH) and limited exaggeration (11.2% CheXpert; 20.0% NIH), consistent with its blurry reconstructions producing concept shifts below the calibrated threshold for most pathologies; its strongest individual result is Cardiomegaly curtailment on NIH (46%), where the in-distribution AE latent space supports a more structured concept direction.

SD (scaled, MD method) substantially exceeds the AE, with mean curtailment CFR of 88.0% (CheXpert) and 63.2% (NIH) and mean exaggeration of 72.0% and 81.2% respectively. SDXL (scaled, MD) further outperforms SD on curtailment (CheXpert: 94.0%; NIH: 92.4%), peaking at Cardiomegaly (98%), Consolidation (98%), and Pleural Effusion (98%) curtailment on CheXpert. SDXL exaggeration under Ark+ is somewhat lower than its curtailment (72.8% CheXpert; 78.0% NIH), reflecting an asymmetry in Ark+'s sensitivity to concept amplification versus suppression that is less pronounced in XRV.

XRV/Platt results (Table 3, bottom). With calibrated XRV, SD and SDXL curtailment patterns are broadly consistent with Ark+, though relative rankings differ. XRV exaggeration CFR is split: SD leads on CheXpert (83.6% vs. SDXL 80.4%), while SDXL leads on NIH (88.8% vs. SD 86.0%), with no single backbone dominating both datasets. A notable exception is the Medical AE on NIH, where XRV/Platt yields substantially higher CFR than Ark+/Platt (51.6% vs. 32.4% curtailment; 51.2% vs. 20.0% exaggeration); we attribute this to NIH being in-distribution for both the AE and the XRV classifier, producing a latent space better aligned with XRV's decision boundary than with Ark+'s.

Pathology-level variation. Across both classifiers and all backbones, Consolidation and Edema are the most amenable pathologies for concept-guided steering, with SD and SDXL regularly achieving CFR $\geq 80\%$ in both directions. Atelectasis remains the most resistant for exaggeration (AE: 2%; SD: 24%; SDXL: 32%), consistent with its heterogeneous spatial presentation making exaggeration vectors less stable. LR generally yields higher curtailment CFR under Ark+, while MD is more competitive under XRV, pointing to a classifier-dependent interaction with the CAV extraction strategy.

6 Conclusion

We presented the first systematic evaluation of concept-guided visual counterfactual generation across generative backbones of differing capacity and domain: a domain-specific Medical AE and the general-purpose latent spaces of SD and SDXL. By assessing each backbone with two independent, Platt-calibrated classifiers (XRV and Ark+) over complementary semantic and fidelity metrics, we directly address the open question of whether generative backbone quality limits concept-guided counterfactuals.

Table 3. CFR (%) for both evaluation classifiers under Platt calibration (MD method). Top block: Ark+/Platt; bottom block: XRV/Platt. Curtailment uses n=50 positive cases; exaggeration uses n=50 negative cases per pathology. Bold = highest CFR in each dataset–direction cell. Exg = Exaggerate; Crt = Curtail. Platt calibration corrects the systematic offset in raw TorchXRayVision logits;

	Medical AE				Stable Diffusion (scaled)				SDXL (scaled)			
Pathology	CheXpert		NIH		CheXpert		NIH		CheXpert		NIH	
	Exg	Crt	Exg	Crt	Exg	Crt	Exg	Crt	Exg	Crt	Exg	Crt
Ark+ / Platt calibration												
Atelectasis	2	4	28	24	24	88	90	84	**32**	**90**	**66**	**96**
Cardiomegaly	30	52	32	46	76	92	58	72	**46**	**98**	**64**	**94**
Consolidation	14	20	6	24	**96**	78	88	46	98	**98**	**80**	**88**
Edema	6	20	14	34	86	**94**	74	48	**96**	86	**92**	**94**
Pleural Eff.	4	10	20	34	78	**88**	96	66	**92**	**98**	**88**	**90**
Mean	*11.2*	*21.2*	*20.0*	*32.4*	*72.0*	*88.0*	*81.2*	*63.2*	*72.8*	*94.0*	*78.0*	*92.4*
XRV / Platt calibration												
Atelectasis	0	4	68	32	50	36	80	32	36	46	**88**	16
Cardiomegaly	18	44	56	76	90	84	82	72	**92**	**84**	**90**	**72**
Consolidation	8	26	40	34	**98**	82	**94**	58	**100**	**90**	96	**44**
Edema	0	0	28	68	88	72	80	68	**86**	**70**	**76**	**44**
Pleural Eff.	10	46	64	48	92	76	94	48	**88**	**86**	**94**	42
Mean	*7.2*	*24.0*	*51.2*	*51.6*	*83.6*	*70.0*	*86.0*	*55.6*	*80.4*	*75.2*	*88.8*	*43.6*

Our central finding is that, once properly calibrated, high-capacity general-purpose diffusion latent spaces steer clinical concepts more effectively than a domain-specific autoencoder, despite having no medical pretraining. This reframes generative capacity, rather than domain specialisation, as the primary lever for concept-guided explanation quality. Equally, our results carry a methodological lesson: a backbone's apparent failure can be an artefact of miscalibration rather than a genuine limitation, underscoring that faithful evaluation of explainability methods demands calibrated, multi-classifier assessment. We further identify the concept-vector extraction strategy as a significant and previously underappreciated factor in counterfactual quality.

These contributions advance concept-level counterfactuals as a route to interpretable, clinically aligned model explanations, but also expose a clear limit: spatially diffuse pathologies such as Atelectasis remain the most resistant to reliable concept steering under single-direction concept vectors regardless of backbone. Future work should pursue multi-direction or compositional concept representations, domain fine-tuning of high-capacity backbones, and clinician-in-the-loop studies to establish whether strong automated concept steering translates into explanations that clinicians find trustworthy and actionable.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Bibliography

[1] Holmberg, L.: Towards benchmarking explainable artificial intelligence methods. ArXiv abs/2208.12120 (2022)

[2] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, vol. 31 (2018)

[3] Viviano, J.D., Simpson, B., Dutil, F., Bengio, Y., Cohen, J.P.: Saliency is a possible red herring when diagnosing poor generalization. arXiv preprint arXiv:1910.00199 (2019)

[4] Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., et al.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In: International Conference on Machine Learning, pp. 2668–2677. PMLR (2018)

[5] Maksudov, B., Curran, K., Mileo, A.: Towards generating more interpretable counterfactuals via concept vectors: a preliminary study on chest X-rays. In: AICS (2025)

[6] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2022)

[7] Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)

[8] Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., et al.: CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 590–597 (2019)

[9] Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2097–2106 (2017)

[10] Konz, N., Osuala, R., Verma, P., et al.: Fréchet Radiomic Distance (FRD): A versatile metric for comparing medical imaging datasets (2025). https://github.com/RichardObi/frd-score

[11] Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., Lee, S.: Counterfactual visual explanations. In: International Conference on Machine Learning (ICML) (2019)

[12] Chang, C.H., Creager, E., Goldenberg, A., Duvenaud, D.: Explaining image classifiers by counterfactual generation. In: International Conference on Learning Representations (ICLR) (2019)

[13] Boreiko, V., Augustin, M., Croce, F., Berens, P., Hein, M.: Sparse visual counterfactual explanations in image space. In: German Conference on Pattern Recognition (GCPR) (2022)

[14] Van Looveren, A., Klaise, J.: Interpretable counterfactual explanations guided by prototypes. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML), pp. 650–665. Springer (2021)

[15] Atad, M., Dmytrenko, V., Li, Y., Zhang, X., Keicher, M., Kirschke, J., Wiestler, B., Khakzar, A., Navab, N.: CheXplaining in style: Counterfactual explanations for chest X-rays using StyleGAN. arXiv preprint arXiv:2207.07553 (2022)

[16] Augustin, M., Boreiko, V., Croce, F., Hein, M.: Diffusion visual counterfactual explanations. In: Advances in Neural Information Processing Systems (NeurIPS) (2022)

[17] Cohen, J.P., Brooks, R., En, S., Zucker, E., Pareek, A., Lungren, M.P., Chaudhari, A.: Gifsplanation via latent shift: A simple autoencoder approach to counterfactual generation for chest X-rays. In: Medical Imaging with Deep Learning, pp. 74–104. PMLR (2021)

[18] Liu, S., Ye, H., Xing, L., Zou, J.: In-context vectors: Making in-context learning more effective and controllable through latent space steering. In: Proceedings of the 41st International Conference on Machine Learning. JMLR (2024)

[19] Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: TorchXRayVision: A library of chest X-ray datasets and models. In: Medical Imaging with Deep Learning (2022)

[20] Naveed, S., Stevens, G., Robin-Kern, D.: An overview of the empirical evaluation of explainable AI (XAI): A comprehensive guideline for user-centered evaluation in XAI. Applied Sciences (2024)

[21] Podell, D., English, Z., Lacey, K., et al.: SDXL: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)

[22] Zhou, J., et al.: Ark: Anatomy-guided semi-supervised learning for echocardiography segmentation. GitHub repository (2023). https://github.com/jlianglab/Ark/

[23] Platt, J.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In: Smola, A., Bartlett, P., Schölkopf, B., Schuurmans, D. (eds.) Advances in Large Margin Classifiers, pp. 61–74. MIT Press (1999)