VAP-Diffusion: Enriching Descriptions with MLLMs for Enhanced Medical Image Generation

Peng Huang[1,2], Junhu Fu[1,2], Bowen Guo[1,2], Zeju Li[1,2], Yuanyuan Wang[1,2], and Yi Guo[1,2✉]

[1] College of Biomedical Engineering, Fudan University, Shanghai 200433, China
[2] Key Laboratory of Medical Imaging Computing and Computer Assisted Intervention of Shanghai, Shanghai 200032, China
guoyi@fudan.edu.cn
https://github.com/YiBaiHP/VAP-Diffusion

Abstract. As the appearance of medical images is influenced by multiple underlying factors, generative models require rich attribute information beyond labels to produce realistic and diverse images. For instance, generating an image of skin lesion with specific patterns demands descriptions that go beyond diagnosis, such as shape, size, texture, and color. However, such detailed descriptions are not always accessible. To address this, we explore a framework, termed Visual Attribute Prompts (VAP)-Diffusion, to leverage external knowledge from pre-trained Multi-modal Large Language Models (MLLMs) to improve the quality and diversity of medical image generation. First, to derive descriptions from MLLMs without hallucination, we design a series of prompts following Chain-of-Thoughts for common medical imaging tasks, including dermatologic, colorectal, and chest X-ray images. Generated descriptions are utilized during training and stored across different categories. During testing, descriptions are randomly retrieved from the corresponding category for inference. Moreover, to make the generator robust to unseen combination of descriptions at the test time, we propose a Prototype Condition Mechanism that restricts test embeddings to be similar to those from training. Experiments on three common types of medical imaging across four datasets verify the effectiveness of VAP-Diffusion.

Keywords: Medical MLLMs · Diffusion Model · Image Generation.

1 Introduction and Motivation

The rapid advancement of deep learning has enabled its clinical applications [1]. However, collecting large annotated medical datasets remains challenging due to high costs and privacy concerns, limiting the performance of deep-learning-based diagnosis systems. Conditional generative models, such as Generative Adversarial Networks(GANs) and Diffusion Models [2, 9], have been explored for data sharing [4, 5], data augmentation [6, 7] and causal inference [8].

Advanced generative models, such as Diffusion Models [11–13], still fall short in realistic and diverse medical image generation, largely due to the variability

of lesions and tissue backgrounds, which stems from the inherent complexity of medical images. Factors like disease status and imaging conditions significantly impact image content, resulting in pronounced distinctions both within and between image classes. Taking skin lesion image generation as an example, even images from the same category (i.e. diagnosed disease) can vary in texture, shape, size, and color (c.f. Fig. 2). We believe that by providing fine-grained descriptions to the generative model, the model can more effectively capture complex data distributions. *By doing so, the image generation tasks are simplified from matching the density of the entire distribution to matching those of separate individual distributions.* Specifically, in a class-conditioned setting, rather than relying solely on single-class conditions (e.g., disease type), we aim to condition the generated images on enriched attributes (e.g., shape, size).

Since demanded detailed descriptions are often unavailable in practice, we turn to Multi-modal Large Language Models (MLLMs) to generate faithful image descriptions as generation guidance. To this end, we propose Visual Attribute Prompts (VAP)-Diffusion, a framework that leverages external knowledge from MLLMs for enhanced medical image generation. Given the remarkable zero-shot capabilities of MLLMs in general image understanding [14] and spatial reasoning [15], we believe the knowledge embedded in MLLMs can provide valuable insights to the generative models, thus improving the realism and diversity of generated images.

Deriving valid descriptions from pre-trained MLLMs is non-trivial because MLLMs are likely to produce inaccurate or fabricated information that is not even related with the input images. This phenomenon is usually referred as hallucination [16]. MLLMs are more likely to generate hallucinations when analyzing medical images, as they are predominantly trained on natural images. Therefore, we carefully design a Visual Attribute Prompt Strategy (VAPS) following Chain-of-Thoughts [16], aiming at eliciting the reasoning potential of MLLMs for medical images. Our method is applicable to a spectrum of common medical images, including dermatologic, colorectal, and chest X-ray images.

While generated descriptions of training samples can be readily utilized for training, corresponding descriptions are unavailable at test time. Therefore, we propose a Class-Specific Prompt Bank (CSPB) in VAP-Diffusion, which stores descriptions from training samples and randomly retrieves class-specific descriptions during inference. Furthermore, we introduce a Prototype Condition Mechanism (PCM) to ensure the test embeddings are similar to those from training phase, enabling the model to generalize well to unseen description combinations.

In summary, the contributions of VAP-Diffusion are summarized as follows:

- We design VAPS for VAP-Diffusion, which prompts the pre-trained MLLMs to generate accurate and informative descriptions as enriched guidance.
- We propose CSPB along with PCM to ensure that the image generator can be readily used without descriptions or with out-of-distribution descriptions.
- We extensively evaluate VAP-Diffusion on three common medical image datasets and demonstrate that the generated images are more realistic and diverse than those produced by existing algorithms. Moreover, we find that

VAP-Diffusion can boost the performance of downstream classification tasks by up to 11.9% compared to its counterpart.

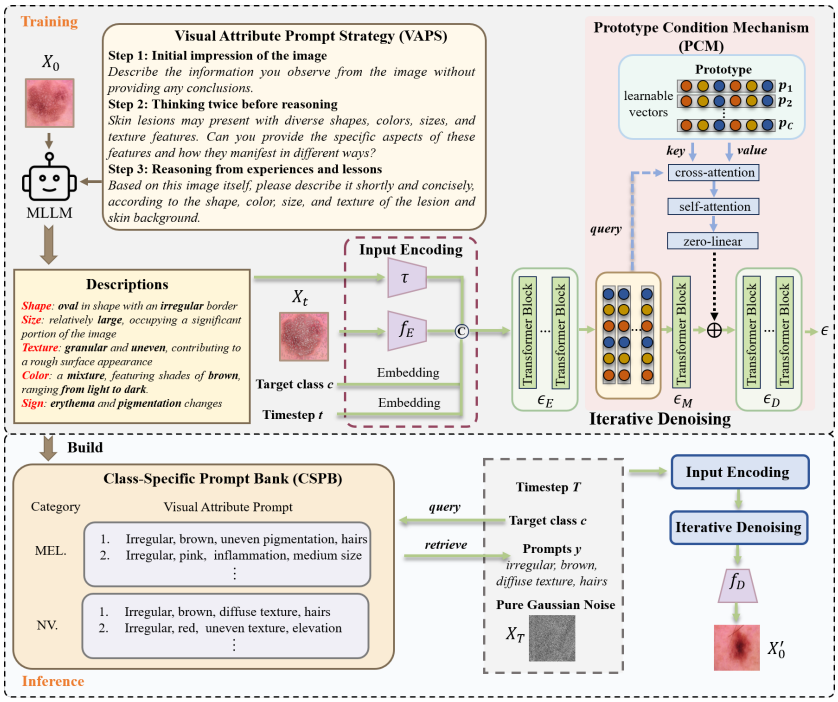

Fig. 1: The overview of VAP-Diffusion. Our framework is built upon a diffusion model and consists of three modules: VAPS, CSPB, and PCM. Together, these modules aim to enhance the generator with enriched descriptions from MLLMs and enable it to work with arbitrary descriptions as conditions during testing.

2 Methodology

We aim to develop a diffusion model capable of generating realistic and diverse medical images, thus enhancing the performance and reliability of computer-aided diagnostic models. As shown in Fig. 1, VAP-Diffusion includes three key components: VAPS, CSPB, and PCM, detailed in the following sections.

2.1 Preliminaries

We build VAP-Diffusion on a state-of-the-art diffusion-based generative model UViT [31]. Diffusion model learns the real data distribution by gradually adding

noise to an input image X_0 and reversing the Markov noising process from purely random noise. During the forward process, for step t, we calculate X_t from X_{t-1} using:

$$q(X_t|X_{t-1}) = \mathcal{N}(X_t|\sqrt{1-\beta_t}X_{t-1}, \beta_t \boldsymbol{I}), \tag{1}$$

where $\mathcal{N}$ denotes the Gaussian distribution, $\boldsymbol{I}$ denotes the identity matrix, $\beta_t \in (0,1)$ is a step-varying hyper-parameter. Specifically, the cumulative transition from X_0 to X_t can be represented as $X_t = \sqrt{\bar{\alpha}_t}X_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ is the cumulative scaling factor.

Assuming the total number of steps is T, the generative model learns to reverse the forward process using following objective function:

$$\mathcal{L}_D = \mathbb{E}_{X_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), t}\Big[\|\epsilon - \epsilon_\theta(X_t, t)\|_2^2\Big], \tag{2}$$

where ϵ_θ refers to the denoising network. During inference, the trained model ϵ_θ progressively denoises the data by generating X_{t-1} from X_t for $t = T, T-1, \ldots, 1$ until we reconstruct the image X_0' which should be similar to the original images. Notably, we use a pre-trained encoder f_E [29] to map images to the latent space and reconstruct images using a trainable decoder f_D.

2.2 Visual Attribute Prompt Strategy

We explore prompt strategy to guide the pre-trained MLLM to generate accurate descriptions based on pre-defined basic visual attributes such as shape, color, size and texture. Here, we take dermatologic image as example to illustrate VAPS and will share VAPS for other modalities (i.e. colorectal and chest X-ray).

Step 1: Initial impression of the image. First, we aim to obtain a basic understanding of given image X_0. MLLM is directly required to describe the dermatologic image according to template $Question^1$, yielding $t^1 = \text{MLLM}(X_0, Question^1)$. *$Question^1$:"You are an AI visual assistant observing a skin lesion image. Describe the information you observe from the image without providing any conclusions."*

Step 2: Thinking twice before reasoning. Next, we guide MLLMs with $Question^2$, obtaining $t^2 = \text{MLLM}(Question^2)$. Based on predefined visual attributes including texture, size, shape, and color, the MLLM is required to consider the inherent imaging characteristics of chosen medical modalities, instead of direct observation of the input images. The generated response remains objective and independent of t^1, ensuring a robust and unbiased interpretation.

$Question^2$: "Skin lesions may present with diverse shapes, colors, sizes, and texture features. The surrounding skin may also appear normal or show signs of inflammation, pigmentation, or other changes. Can you provide the specific aspects of these features and how they manifest in different ways?"

Step 3: Reasoning from experiences and lessons. Finally, to further mitigate hallucinations in the MLLM, we ask the MLLMs to leverage answers from the previous two stages and generate an accurate and concise summary of the given image, resulting in $t^{mix} = \text{MLLM}(t^1, t^2, X_0, Question^3)$.

$Question^3$: "Based on this image itself, please describe it shortly and concisely, according to the shape, color, size (due to the unavailability of quantitative dimensions, you can just describe their approximate proportion in the whole image), and texture of the lesion and skin background."

We then encode t^{mix} with BiomedCLIP [17] τ to ensure compatibility with the image processing pipeline.

2.3 Class-Specific Prompt Bank

To supplement descriptions for the inference stage, we build CSPB to store the enriched descriptions for each class during training. Specifically, during training, for each image X_0 in the dataset, VAP-Diffusion utilizes a pre-trained MLLM to generate detailed image descriptions y. For a given class c, we make a collection of $\{y_i\}_{i=1}^{n}$, where n represents the number of samples belonging to class c. During inference, for a given class c, a prompt y_i is randomly selected from the prompt bank and used as an additional condition to guide image generation. Notably, we also evaluate our model using a visual attribute prompt bank that differs from the one used during training, ensuring that VAP-Diffusion remains flexible enough to handle challenging cases.

2.4 Prototype Condition Mechanism

To make VAP-Diffusion flexible enough to handle free-text inputs, we propose PCM to enhance the model's robustness against unseen descriptions through prototype matching for each category. Specifically, during the inference stage, given class information and arbitrary descriptions, we aim to regularize the embeddings to remain close to those of training samples from the same class.

Given that the denoising network ϵ_θ consists of encoder blocks ϵ_E, middle blocks ϵ_M and decoder blocks ϵ_D. For each time step t and one sample X_t, we calculate the encoding features as $F_e = \epsilon_E(f_E(X_t), t, c, \tau(y)) \in \mathbb{R}^K$. Based on these, prototypes of all C classes are constructed. For each class c, we produce $\boldsymbol{p}_c$ which contains K elements. To build the connection between prototypes and input class c, $\boldsymbol{p}_c$ is optimized with $\mathcal{L}_{recon}$ to reconstruct F_e at any time step t via a cross-attention layer. We further employ a self-attention layer to enhance the multi-modal representation. To stabilize the training process, we build a linear layer initialized with all zero which can gradually inject the multi-modal priors.

With balancing term α, our objective function for a single optimization is:

$$\mathcal{L}_{VAP} = \mathbb{E}_{X_0, \epsilon \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), t}\left[||\epsilon - \epsilon_\theta(f_E(X_t), t, c, \tau(y))||_2^2\right] + \alpha \mathcal{L}_{recon}. \tag{3}$$

Dataset	Method	FID ↓	IS ↑	Precision ↑	Recall ↑
ISIC2018	StyleGAN	**16.968**	2.967	**0.640**	0.427
	CBDM	59.963	3.066	0.319	0.128
	LDM	53.375	3.491	0.395	0.264
	DiT	57.368	2.974	0.373	0.251
	UViT	52.760	3.353	0.346	0.275
	VAP-Diffusion (Ours)	19.790	**4.404**	0.617	**0.812**
ISIC2019	StyleGAN	**23.725**	4.347	**0.457**	0.360
	CBDM	55.676	4.798	0.329	0.254
	LDM	57.853	4.962	0.341	0.279
	DiT	59.816	5.184	0.281	0.277
	UViT	43.292	5.113	0.372	0.304
	VAP-Diffusion (Ours)	25.232	**5.252**	0.416	**0.496**
ChestXray14	StyleGAN	**36.221**	2.296	0.343	0.418
	CBDM	52.139	2.419	0.276	0.416
	LDM	47.730	2.592	0.343	0.450
	DiT	50.016	2.560	0.282	0.459
	UViT	56.204	2.723	0.295	0.455
	VAP-Diffusion (Ours)	40.487	**2.728**	**0.375**	**0.600**
Colonoscopy Database	StyleGAN	57.871	2.699	0.214	0.341
	CBDM	55.692	4.725	0.334	0.386
	LDM	46.244	4.604	0.464	0.442
	DiT	53.163	4.776	0.334	0.406
	UViT	46.265	4.779	0.395	0.492
	VAP-Diffusion (Ours)	**32.147**	**4.812**	**0.405**	**0.612**

Table 1: Quantitative comparison of image synthesis results with other SOTA algorithms.

Dataset		FID ↓	Recall ↑
ISIC2018	VAP-Diffusion	**26.767**	**0.609**
	w/o VAPS	52.760	0.275
	w/o PCM	28.821	0.582
ISIC2019	VAP-Diffusion	**30.357**	**0.361**
	w/o VAPS	43.292	0.304
	w/o PCM	31.931	0.338
ChestXray14	VAP-Diffusion	**44.318**	**0.509**
	w/o VAPS	56.204	0.455
	w/o PCM	46.335	0.485
Colonoscopy Database	VAP-Diffusion	**32.303**	**0.519**
	w/o VAPS	46.265	0.492
	w/o PCM	37.261	0.501

Table 2: Ablation experiment results for VAP-Diffusion with unseen input texts. The PCM module only functions with the VAPS module and remains inactive in its absence

3 Experiments Results

Datasets. (1) For dermoscopic modality, we use two public datasets including ISIC2018 dataset [18] and ISIC2019 dataset [19]. (2) For chest X-ray images, we choose the ChestXray14 [20], and the dataset is further processed to include only single-label images. (3) For colonoscopic images, we utilize three publicly available datasets [21–23], as well as one private dataset. We split ISIC2018 and ChestXray14 datasets into training and testing sets following official guidelines, while randomly splitting ISIC2019 and colonoscopicd with a ratio of 8:2.

Evaluation Metrics We quantify model's performance based on two criteria: realism and diversity. We evaluate realism with FID [24] and Precision [25], where lower FID and higher Precision indicate better image fidelity. We assess diversity using the IS [26] and Recall [25], with higher values reflecting greater image diversity. For downstream classification tasks, we employ mAUC and F1-score, where higher scores imply better model performance.

3.1 Comparison with Other SOTA Methods: Image Synthesis

We compared VAP-Diffusion with five advanced image generation algorithms (StyleGAN [27], CBDM [28], LDM [29], DiT [30], and UViT [31]. As shown in Table 1, VAP-Diffusion consistently achieves the highest IS and Recall on all datasets. Notably, on the Colonoscopy dataset, VAP-Diffusion achieves the best

Dataset	Classification Model	Ratio of Real Images					
		mAUC			F1-score		
		1%	10%	100%	1%	10%	100%
ISIC2018	Densenet-121	0.555	0.866	0.964	0.212	0.414	0.710
	+ synthetic StyleGAN	0.913	0.913	0.965	0.526	0.576	0.741
	+ synthetic VAP-Diffusion (Ours)	**0.923**	**0.943**	**0.978**	**0.645**	**0.697**	**0.753**
	maxViT	0.555	0.867	0.964	0.412	0.413	0.713
	+ synthetic StyleGAN	0.912	0.926	0.964	0.522	0.570	0.727
	+ synthetic VAP-Diffusion (Ours)	**0.923**	**0.940**	**0.971**	**0.638**	**0.687**	**0.744**
ChestXray14	Densenet-121	0.527	0.684	0.788	/	/	/
	+ synthetic StyleGAN	0.610	0.684	0.790	/	/	/
	+ synthetic VAP-Diffusion (Ours)	**0.725**	**0.766**	**0.802**	/	/	/
	maxViT	0.515	0.683	0.779	/	/	/
	+ synthetic StyleGAN	0.607	0.684	0.784	/	/	/
	+ synthetic VAP-Diffusion (Ours)	**0.718**	**0.756**	**0.799**	/	/	/
Colonoscopy Database	Densenet-121	0.583	0.879	0.923	0.272	0.803	0.907
	+ synthetic StyleGAN	0.903	0.904	0.925	0.702	0.859	0.918
	+ synthetic VAP-Diffusion (Ours)	**0.923**	**0.935**	**0.946**	**0.744**	**0.880**	**0.940**
	maxViT	0.568	0.869	0.919	0.272	0.803	0.896
	+ synthetic StyleGAN	0.896	0.899	0.923	0.692	0.859	0.914
	+ synthetic VAP-Diffusion (Ours)	**0.918**	**0.931**	**0.943**	**0.736**	**0.879**	**0.930**

Table 3: Comparison of downstream task results with StyleGAN. ISIC2019 was not adopted for testing as it does not provide official test set.

Recall (0.612) and IS (4.812), while maintaining a competitive FID (32.417), indicating its ability to generate high-quality and diverse images.

Additionally, as shown in Fig. 2, compared to StyleGAN, VAP-Diffusion is capable of generating more realistic and diverse medical images that closely resemble real medical images in terms of texture and structural details. Especially, for dermatologic images, VAP-Diffusion is able to generate complex medical data with visible hairs or rough surface textures.

3.2 Ablation Experiments

We conduct ablation experiments on the test sets to evaluate the effectiveness of VAPS and PCM, as well as the robustness of VAP-Diffusion when handling unseen descriptions. As shown in Table 2, VAP-Diffusion consistently achieves competitive FID and Recall scores across all datasets, even with unseen texts. Of note, on ISIC2018, removing VAPS leads to a significant increase in FID from 26.767 to 52.760 and a sharp decrease in Recall from 0.609 to 0.275, reflecting the importance of VAPS in generating realistic and diverse medical images.

3.3 Comparison with StyleGAN: Downstream Tasks

To verify the effectiveness of VAP-Diffusion in downstream tasks, we conducted downstream classification experiments on three modalities using DenseNet [32]

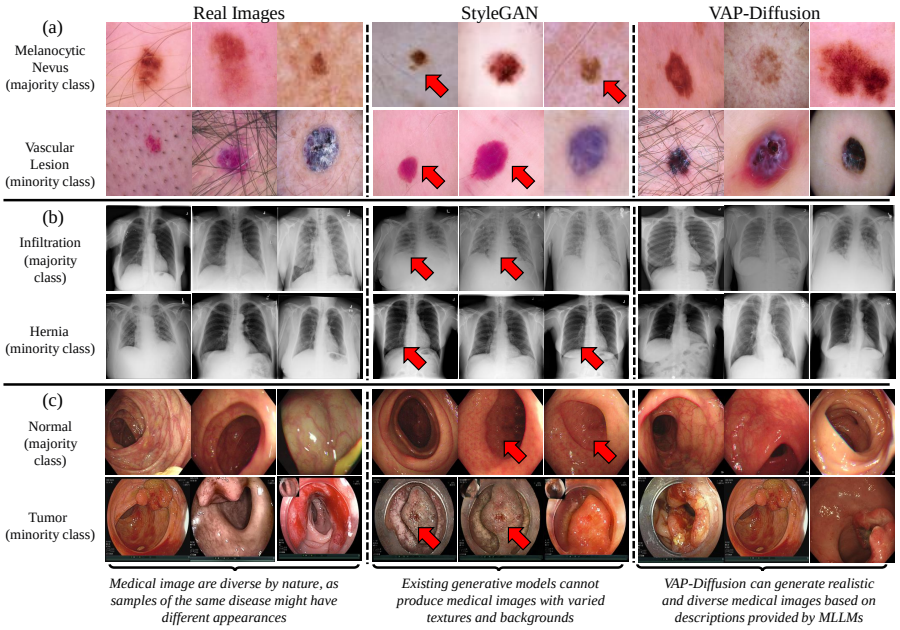

Fig. 2: Visual comparison of image synthesis results with StyleGAN. (a) Dermatologic images; (b) Chest X-ray images; (c) Colonoscopic images. Red arrows indicate repetitive content in StyleGAN-generated images, highlighting its potential limitations in producing diverse data.

and maxViT [33]. As shown in Table 3, VAP-Diffusion consistently outperforms both baseline models and those augmented with StyleGAN-generated synthetic data, particularly when only 1% or 10% of real data is available. Specifically, on the ISIC2018 dataset with 1% real data, VAP-Diffusion achieves an mAUC of 0.923 and an F1-score of 0.645, compared to 0.913 and 0.526 for StyleGAN on Densenet-121. This demonstrates the ability of VAP-Diffusion to generate high-quality synthetic data that effectively supplements limited real data. We observe similar improvements on ChestXray14 and the Colonoscopy database.

4 Conclusion

It is difficult for diffusion models to learn complex data distributions in medical scenes with class conditions only. In this paper, we introduce VAP-Diffusion, a novel medical image generative model which successfully exploits the pre-trained MLLM to provide effective guidance for image generation. The key innovation focuses on an efficient Chain-of-Thoughts-based prompt strategy, enabling the MLLM to generate grounded visual attribute prompts. Additionally, we propose a prompt bank along with class prototypes to leverage the multi-modal priors learned during training, assisting diffusion inference. VAP-Diffusion achieves the

best generation results on four datasets from three medical modalities and outperforms its counterparts by a large margin on downstream classification tasks.

Acknowledgments. This study was funded by the National Natural Science Foundation of China (Grant No 62371139), and Shanghai Municipal Education Commission (Grant No. 24KNZNA09).

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

References

1. Dayarathna, S., Islam, K.T., Uribe, S., Yang, G., Hayat, M., Chen, Z.: Deep learning based synthesis of MRI, CT and PET: Review and analysis. Medical Image Analysis 92, 103046 (2024).
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM 63(11), 139–144 (2020).
3. Mirza, M.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
4. Han, T., Nebelung, S., Haarburger, C., Horst, N., Reinartz, S., Merhof, D., Kiessling, F., Schulz, V., Truhn, D.: Breaking medical data sharing boundaries by using synthesized radiographs. Science Advances 6(49), eabb7973 (2020).
5. Özbey, M., Dalmaz, O., Dar, S.U.H., Bedel, H.A., Özturk, Ş., Güngör, A., Çukur, T.: Unsupervised medical image translation with adversarial diffusion models. TMI (2023).
6. Ktena, I., Wiles, O., Albuquerque, I., Rebuffi, S.-A., Tanno, R., Roy, A.G., Azizi, S., Belgrave, D., Kohli, P., Cemgil, T., et al.: Generative models improve fairness of medical classifiers under distribution shifts. Nature Medicine, 1–8 (2024).
7. Müller-Franzes, G., Niehues, J.M., Khader, F., Arasteh, S.T., Haarburger, C., Kuhl, C., Wang, T., Han, T., Nolte, T., Nebelung, S., et al.: A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. Scientific Reports 13(1), 12098 (2023).
8. Huang, P., Gao, X., Huang, L., Jiao, J., Li, X., Wang, Y., Guo, Y.: Chest-Diffusion: A Light-Weight Text-to-Image Model for Report-to-CXR Generation. In: ISBI, pp. 1–5 (2024).
9. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: NeurIPS, vol 33, pp. 6840–6851 (2020)
10. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)
11. Han, K., Xiong, Y., You, C., Khosravi, P., Sun, S., Yan, X., Duncan, J.S., Xie, X.: Medgen3d: A deep generative framework for paired 3D image and mask generation. In: MICCAI, pp. 759–769 (2023).
12. Khader, F., Mueller-Franzes, G., Arasteh, S.T., Han, T., Haarburger, C., Schulze-Hagen, M., Schad, P., Engelhardt, S., Baessler, B., Foersch, S., et al.: Medical diffusion: denoising diffusion probabilistic models for 3D medical image generation. arXiv preprint arXiv:2211.03364 (2022)

13. Bluethgen, C., Chambon, P., Delbrouck, J.-B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest X-ray images. Nature Biomedical Engineering, 1–13 (2024).
14. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: ICML, pp. 12888–12900 (2022).
15. Chen, B., Xu, Z., Kirmani, S., Ichter, B., Sadigh, D., Guibas, L., Xia, F.: SpatialVLM: Endowing vision-language models with spatial reasoning capabilities. In: CVPR, pp. 14455–14465 (2024)
16. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: NeurIPS, vol 35, pp. 24824–24837 (2022)
17. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: BiomedCLIP: A multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
18. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the International Skin Imaging Collaboration (ISIC). arXiv preprint arXiv:1902.03368 (2019)
19. Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al.: BCN20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
20. 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: CVPR, pp. 2097–2106 (2017)
21. Li, K., Fathan, M.I., Patel, K., Zhang, T., Zhong, C., Bansal, A., Rastogi, A., Wang, J.S., Wang, G.: Colonoscopy polyp detection and classification: Dataset creation and comparative evaluations. PLOS ONE, 16(8), e0255809 (2021)
22. Misawa, M., Kudo, S., Mori, Y., Hotta, K., Ohtsuka, K., Matsuda, T., Saito, S., Kudo, T., Baba, T., Ishida, F., et al.: Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). Gastrointestinal Endoscopy, 93(4), 960–967 (2021)
23. Mesejo, P., Pizarro, D., Abergel, A., Rouquette, O., Beorchia, S., Poincloux, L., Bartoli, A.: Computer-aided classification of gastrointestinal lesions in regular colonoscopy. IEEE Transactions on Medical Imaging (TMI), 35(9), 2051–2063 (2016)
24. 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. In: NeurIPS, vol 30 (2017)
25. Sajjadi, M.S.M., Bachem, O., Lucic, M., Bousquet, O., Gelly, S.: Assessing generative models via precision and recall. In: NeurIPS, vol 31 (2018)
26. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X.: Improved techniques for training GANs. In: NeurIPS, vol 29 (2016)
27. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: NeurIPS, vol 34, pp. 852–863 (2021)
28. Qin, Y., Zheng, H., Yao, J., Zhou, M., Zhang, Y.: Class-balancing diffusion models. In: CVPR, pp. 18434–18443 (2023)
29. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022)

30. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: CVPR, pp. 4195–4205 (2023)
31. Bao, F., Nie, S., Xue, K., Cao, Y., Li, C., Su, H., Zhu, J.: All are worth words: A ViT backbone for diffusion models. In: CVPR, pp. 22669–22679 (2023)
32. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR, pp. 4700–4708 (2017)
33. Tu, Z., Talebi, H., Zhang, H., Yang, F., Milanfar, P., Bovik, A., Li, Y.: MaxViT: Multi-axis vision transformer. In: ECCV, pp. 459–479 (2022)