DALL-E Brain : Génération d'images IRM cérébrales 2D T1-w, T2-w et FLAIR à partir de descriptions textuelles

Souhail EL-ALLALY[1] Thomas GRENIER[1] Chantal REVOL-MULLER[1]

[1]INSA Lyon, CREATIS UMR 5220, U1294, Villeurbanne, France

Résumé – Ce projet explore l'application des Modèles de Diffusion Latents (LDMs) pour générer des images IRM cérébrales synthétiques mais réalistes en pondération T1, T2 et FLAIR à partir de descriptions textuelles. Pour entraîner notre modèle, nous avons construit un jeu de données comprenant plus de 40 000 scans IRM de sujets sains, chacun associé à des descriptions textuelles spécifiques à chaque coupe. Ces annotations comprennent les structures anatomiques, l'âge, le sexe et les informations d'imagerie, incluant le plan de coupe et la séquence IRM. Nous proposons un LDM pour gérer des descriptions longues avec un vocabulaire spécialisé en imagerie médicale. Les performances du modèle sont évaluées à l'aide du MS-SSIM pour l'autoencodeur variationnel (VAE) et du FID pour les images générées par diffusion. Les résultats expérimentaux montrent que, sur la seule base de descriptions textuelles, notre méthode permet de générer des scans IRM réalistes, soulignant ainsi le potentiel des LDMs pour la synthèse d'images médicales.

Abstract – This project explores the application of Latent Diffusion Models (LDMs) for generating synthetic yet realistic T1-weighted, T2-weighted, and FLAIR MRI brain images from textual prompts. To train our model, we constructed a dataset of over 40,000 MRI scans of healthy subjects, each paired with slice-specific textual descriptions. These annotations include anatomical structures, age, sex, and imaging details such as slice plane and MRI sequence. To enhance textual conditioning, we modified an LDM to handle long prompts with a vocabulary specialized for the medical domain. Model performance is evaluated using MS-SSIM for the VAE component and FID for the diffusion-generated images. Experimental results demonstrate that using only textual descriptions, our method can generate realistic MRI scans, highlighting the potential of LDMs for medical imaging synthesis.

1 Introduction

Les récents progrès des modèles vision-langage, tels que CLIP [7], ont permis la création de représentations textuelles et visuelles unifiées, donnant naissance à des modèles génératifs texte-image comme DALL-E et les Modèles de Diffusion Latents (LDMs) [8]. Bien que ces modèles excellent dans la synthèse d'images réalistes, leur application à l'imagerie médicale reste limitée en raison des exigences de précision anatomique et de validation par des experts. Plusieurs approches ont été explorées pour la génération d'images médicales conditionnées par du texte, notamment les modèles basés sur des transformers [1], le conditionnement vision-langage [11] et l'apprentissage contrastif [10]. Ces méthodes reposent cependant sur des ensembles de données appariées texte-image de taille limitée, rendant la génération d'images médicales pilotée par le texte particulièrement difficile.

2 Construction de la base de données image/texte

Le développement d'un modèle d'IA génératif requiert un jeu de données d'entraînement large et diversifié. Dans cette étude, nous avons besoin de coupes IRM cérébrales de sujets sains avec leurs descriptions anatomiques, mais les bases de données publiques annotées sont rares. Toutefois, ces descriptions peuvent être générées automatiquement via la segmentation des structures cérébrales, permettant d'accéder aux volumes anatomiques. Nous avons constitué un jeu de données image/-texte en combinant IBSR[1](18 IRM T1 segmentées) et des IRM non segmentées issues de OASIS[2] (77 IRM T1), IXI[3] (64 IRM T2) et Kirby[4] (42 IRM FLAIR).

2.1 Génération d'IRM labellisées

Pour augmenter le nombre d'IRM labellisées, nous avons adopté une segmentation rapide et efficace. Les outils automatisés comme SynthSeg et FreeSurfer étant trop coûteux en calcul, nous avons privilégié un recalage affine 3D permettant d'extraire le contenu anatomique des coupes IRM.

Processus de recalage et de segmentation

Nous avons effectué un recalage affine 3D en niveaux de gris des images IRM non labellisées (OASIS, IXI, Kirby) sur les images de référence IBSR, en utilisant le cadre ANTs[5]. Avant le recalage, le crâne a été supprimé pour garantir un alignement multimodal précis. La même transformation affine a été appliquée aux images et aux masques de segmentations associés, mais en adaptant le type d'interpolation. Ce processus a généré 18×77 nouvelles IRM annotées en pondération T1, 18×64 en pondération T2 et 18×42 en pondération FLAIR. L'application de la transformation inverse a permis de recaler IBSR sur OASIS, IXI et Kirby, augmentant ainsi encore la taille du jeu de données (Fig. 1).

Génération d'atlas

Pour renforcer la cohérence anatomique, nous avons créé trois

[1]https://www.nitrc.org/projects/ibsr

[2]https://www.oasis-brains.org/data

[3]https://brain-development.org/ixi-dataset

[4]https://www.nitrc.org/frs/?group_id=313

[5]https://andysbrainbook.readthedocs.io/en/latest/ANTs/ANTs_Overview.html

nouveaux atlas à l'aide d'un vote majoritaire entre les 18 labels IBSR recalés. Cela a permis de générer l'atlas OASIS (77 IRM T1 annotées), l'atlas IXI (64 IRM T2 annotées) et l'atlas Kirby (42 IRM FLAIR annotées). En modifiant successivement l'atlas de référence et les bases de données à recaler, nous avons encore augmenté le nombre d'images annotées. Au total, cette procédure a permis de constituer un jeu de données de 40 600 volumes IRM annotés en 3D, représentant environ 25,6 millions de coupes axiales, sagittales et coronales (Fig. 2).

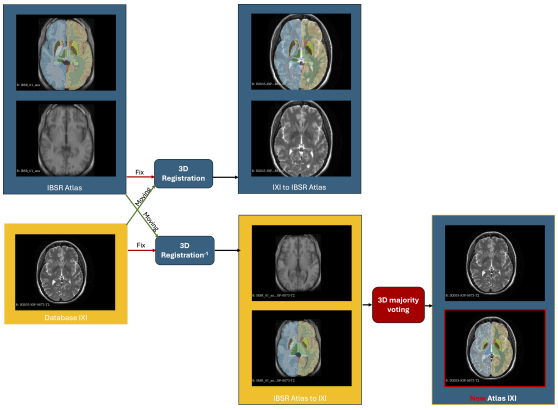


FIGURE 1 : Création d'IRM annotées à partir de l'atlas IBSR et génération de nouveaux atlas.

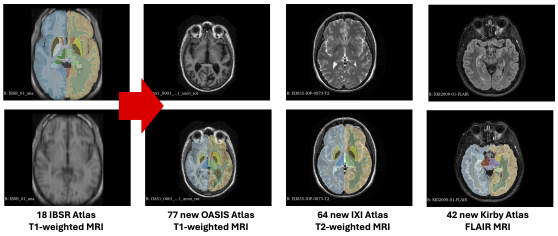


FIGURE 2 : Atlas générés à partir des recalages IBSR.

2.2 Génération des descriptions textuelles

L'atlas IBSR possède une table de correspondance reliant la valeur numérique des labels dans les images segmentées au nom des structures anatomiques correspondantes. Les métadonnées telles que la taille des voxels, le type de séquence IRM, ainsi que le sexe et l'âge du sujet sont disponibles pour les images mobiles et fixes. À partir des segmentations, nous avons généré automatiquement des données tabulaires pour chaque coupe IRM décrivant les structures anatomiques et leurs volumes exportables au format CSV. Enfin, des descriptions en langage naturel ont été produites par synthèse textuelle à base de modèles de phrases prédéfinis. Pour introduire de la variabilité, une vingtaine de modèles de phrases ont été aléatoirement sélectionnés pour décrire chaque coupe (Fig. 3). Le jeu de données final, généré via notre pipeline de traitement, comprend 40 600 scans IRM annotés de sujets sains, couvrant plusieurs modalités d'IRM. Chaque volume est associé aux descriptions anatomiques textuelles de chacune de ses coupes. Cela aboutit à un ensemble de données étendu contenant environ 26 millions de paires image/texte, offrant ainsi une ressource riche et bien annotée pour l'entraînement de modèles d'IA générative.

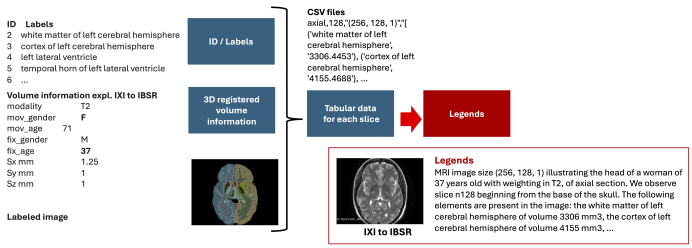


FIGURE 3 : Exemple de paire coupe-description générée.

3 Encodeur/Décodeur d'images

3.1 Autoencodeur Variationnel (VAE)

Un VAE est un encodeur-décodeur conçu pour apprendre une représentation latente compacte des données d'entrée. L'encodeur projette une image dans un espace latent de plus faible dimension en estimant une distribution de probabilité avec un vecteur de moyenne μ et un vecteur d'écart-type σ. Le décodeur reconstruit une image à partir d'un échantillon latent, garantissant que la représentation conserve les caractéristiques essentielles des données [5]. Dans notre modèle génératif, le VAE agit comme un mécanisme de compression, encodant les coupes IRM dans l'espace latent pour le processus de diffusion (Fig. 5). Après débruitage, le décodeur du VAE reconstruit l'image dans le domaine des pixels.

Fonction de perte

Étant donnée une image d'entrée $\mathbf{x} \in \mathbb{R}^{H\times W\times 1}$, l'encodeur la projette dans un espace latent de dimensions réduites $\mathbf{z} \in \mathbb{R}^{h\times w\times c}$, $\mathbf{z} = E(\mathbf{x})$. Le décodeur reconstruit ensuite l'image depuis cette représentation latente, $\tilde{\mathbf{x}} = D(\mathbf{z}) = D(E(\mathbf{x}))$. L'entraînement du VAE est guidé par la fonction de perte (Eq. 1) :

$$\mathcal{L}_{VAE} = \min_{E,D} \max_{\psi} \Big(\mathcal{L}_{\text{rec}}\left(\mathbf{x}, D\left(E\left(\mathbf{x}\right)\right)\right) - \mathcal{L}_{\text{adv}}\left(D\left(E\left(\mathbf{x}\right)\right)\right) + \log D_{\psi}\left(\mathbf{x}\right) + \mathcal{L}_{\text{reg}}\left(\mathbf{x}; E, D\right) \Big) \quad (1)$$

avec $\mathcal{L}_{\text{rec}}$ la perte de reconstruction perceptuelle mesurant la différence entre les images originales et reconstruites, $-\mathcal{L}_{\text{adv}} + \log D_{\psi}(\mathbf{x})$ la perte adversariale pénalisant les solutions triviales et $\mathcal{L}_{\text{reg}}$ le terme de régularisation minimisant la divergence KL entre l'espace latent et une distribution gaussienne standard.

Entraînement du VAE

Le VAE a été entraîné sur 40 000 coupes IRM axiales, réparties en un ensemble de train (24 000 images), de validation (6 000 images) et de test (10 000 images).

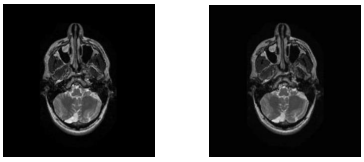
Hyperparamètres : nombre d'époques = 100, taille de batch = 26, GPU = 32GB, taille d'image = $(256, 256)$, dimension latente = $(64, 64, 1)$, durée totale d'entraînement $\sim$ 40 heures.

3.2 Résultats du VAE

La performance du VAE est évaluée en comparant qualitativement et quantitativement les images reconstruites $\tilde{\mathbf{x}} = D(E(\mathbf{x}))$ avec les entrées originales $\mathbf{x}$.

Résultats qualitatifs

La Fig. 4 présente une comparaison entre une coupe IRM originale (4a) et sa reconstruction par le VAE (4b). La forte similarité visuelle confirme que le VAE préserve efficacement les structures anatomiques essentielles.



(a) Originale (b) Reconstruite

FIGURE 4 : Encodage et reconstruction d'une coupe IRM par le VAE.

Résultats quantitatifs

Pour évaluer objectivement la fidélité de la reconstruction, nous utilisons l'Indice de Similarité Structurelle Multi-Échelle (MS-SSIM) [9], qui mesure la similarité des images en tenant compte de la luminance, du contraste et des structures.

Le tableau 1 présente les scores MS-SSIM sur deux échantillons de 10 000 coupes IRM prélevés respectivement dans les ensembles de train et de test. Les scores élevés (~ 0.98) confirment que le VAE reconstruit fidèlement les images tout en préservant leurs détails anatomiques.

TABLE 1 : Scores MS-SSIM pour l'évaluation du VAE.

	Train (10,000)	**Test (10,000)**
MS-SSIM [0 -1] ↑	0.984	0.977

4 LDM conditionné par texte

Nous nous appuyons sur le modèle de diffusion latent (LDM) 2D de MONAI[6], dont le tutoriel n'aborde pas le conditionnement par texte [6]. Ce modèle de base se concentre uniquement sur la génération d'images. Nous modifions le code original en intégrant le conditionnement textuel[7].

Architecture

Les LDMs fonctionnent dans un espace latent compressé plutôt que dans l'espace image. Notre architecture (Fig. 5) se compose (i) d'un VAE qui encode les coupes IRM d'entrée en une représentation latente, (ii) d'un encodeur de texte qui transforme les descriptions textuelles en embeddings, et (iii) d'un modèle de diffusion opérant dans l'espace latent, affinant progressivement des représentations bruitées en représentations cohérentes.

Extraction des embeddings avec un tokenizer

Pour mettre en oeuvre le conditionnement par texte, nous utilisons un tokenizer WordPiece et un modèle Transformer préentraîné sur la littérature biomédicale. Le texte d'entrée T_{orig} est tokenisé, puis converti en une représentation vectorielle contextualisée. Ces embeddings sont utilisés comme conditionnement dans le modèle de diffusion, qu'ils proviennent d'un modèle entraîné contrastivement tel que PubMedCLIP[8] [2] ou d'un modèle de langage comme BioMedBERT[9] [3]. Embeddings $=$ tokenizer$(T_{\text{orig}}) \in \mathbb{R}^{T \times d_t}$ avec T le nombre de tokens et $d_t = 768$ la dimension des embeddings contextualisés de chaque token.

Mécanisme de cross-attention

Pour aligner les embeddings textuels avec les représentations latentes visuelles, nous intégrons un mécanisme de cross-attention dans le LDM. Dans ce cadre, les queries $Q \in \mathbb{R}^{B \times N_{\text{spatial}} \times C_{\text{feat}}}$ proviennent des caractéristiques latentes du U-Net, avec B la taille de batch, $N_{\text{spatial}} = h' \times w'$ et $C_{\text{feat}} = 768$. Les keys et les values $K, V \in \mathbb{R}^{B \times T \times 768}$ sont issues des embeddings textuels. Le mécanisme d'attention suit la formulation : $\text{attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_t}}\right) V$.

Ce mécanisme injecte du contenu sémantique issu du texte dans les caractéristiques visuelles latentes, guidant le processus de débruitage vers des représentations cohérentes avec le texte. Le U-Net est configuré pour appliquer le mécanisme de cross-attention sur ses deux derniers niveaux.

Comparaison des encodeurs de texte

Nous évaluons deux encodeurs de texte : PubMedCLIP et BioMedBERT. PubMedCLIP est limité à 77 tokens, ce qui restreint sa capacité à gérer des descriptions anatomiques longues. En revanche, BioMedBERT peut traiter jusqu'à 512 tokens, offrant une représentation textuelle plus riche et détaillée. De plus, BioMedBERT atteint des performances comparables avec seulement la moitié des époques d'entraînement requises par PubMedCLIP. Compte tenu de ces avantages, nous sélectionnons BioMedBERT comme encodeur de texte principal.

Entrainement

L'entraînement du LDM consiste à minimiser la fonction objectif suivante (Eq. 2) :

$$\mathcal{L}_{LDM} = \mathbb{E}_{\epsilon(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right] \quad (2)$$

où ϵ représente le bruit ajouté et $\epsilon_\theta(z_t, t, \tau_\theta(y))$ le bruit prédit par le U-Net à l'instant t. Le LDM a été entraîné sur 40 000 coupes IRM axiales avec leurs descriptions associées : 24 000 images pour l'entraînement, 6 000 images pour la validation et 10 000 images pour le test.

Hyperparamètres : VAE pré-entraîné + LDM conditionné avec BiomedBERT, 100 époques , taille de batch = 40, GPU 32GB, taille d'image : $(256, 256)$, dimension latente : $(64, 64, 1)$, temps d'entraînement total : 14 jours, 22 heures, 9 minutes.

5 Résultats et discussion

La qualité des images générées est évaluée à l'aide de la Fréchet Inception Distance (FID), qui mesure la distance entre les distributions de caractéristiques des images IRM synthétiques et réelles dans un espace d'embedding basé sur InceptionV3 [4]. Nous évaluons 1 000 images IRM synthétiques générées à partir des descriptions des ensembles d'entraînement et de validation, ainsi que 1 000 images synthétiques issues de descriptions du jeu de test, en les comparant à 10 000 images IRM réelles de chaque ensemble respectif. La Fig. 6 illustre des coupes axiales 2D générées à l'aide de notre LDM conditionné par texte. La première ligne correspond au premier prompt (P1) et la deuxième au second prompt (P2). Selon les descriptions textuelles fournies, le modèle distingue correctement les types de séquences IRM (T1-w vs. T2-w).

Le tableau 2 présente les scores FID en comparant 1 000 coupes IRM synthétiques à 10 000 coupes réelles. Le modèle atteint des scores FID faibles, indiquant une grande fidélité et une cohérence structurelle.

TABLE 2 : Scores FID des coupes IRM générées par le LDM.

	Train/Val (1 000 vs 10 000)	Test (1 000 vs 10 000)
FID ↓	**16.9**	**17.9**

[6] https://github.com/Project-MONAI/tutorials/tree/main/generation/2d_ldm

[7] https://gitlab.in2p3.fr/chantal.muller/dalle-brain

[8] https://huggingface.co/flaviagiammarino/pubmed-clip-vit-base-patch32

[9] https://huggingface.co/microsoft/BiomedNLP-BiomedBERT-base-uncased-abstract-fulltext

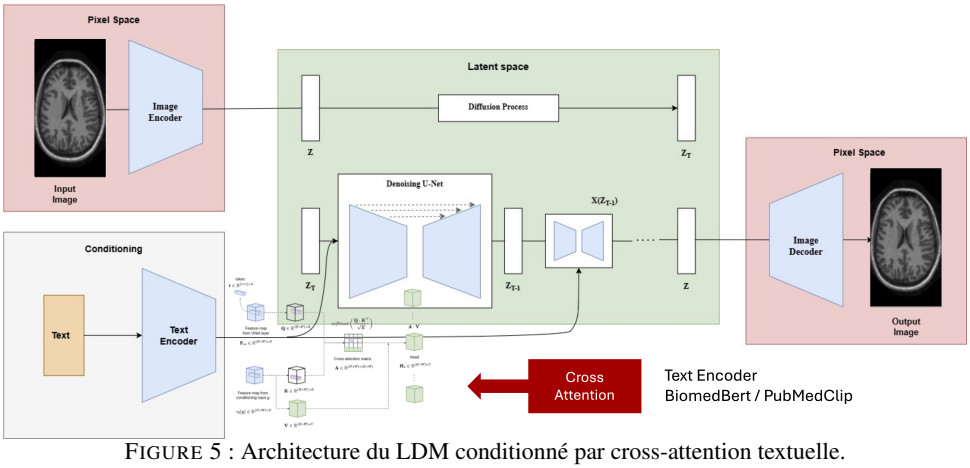
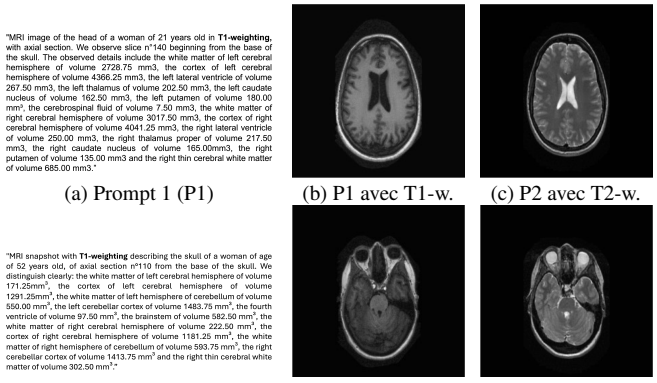


FIGURE 5 : Architecture du LDM conditionné par cross-attention textuelle.



(a) Prompt 1 (P1) (b) P1 avec T1-w. (c) P2 avec T2-w.

(d) Prompt 2 (P2) (e) P2 avec T1-w. (f) P2 avec T2-w.

FIGURE 6 : Coupes IRM T1-w. and T2-w. MRI générées par prompt.

Les scores FID indiquent une forte capacité de généralisation . La première colonne présente les résultats pour les prompts vus pendant l'entraînement, tandis que la seconde colonne évalue les prompts jamais rencontrés auparavant. Malgré cette nouveauté, le modèle maintient des scores FID bas (16.9 pour les prompts connus, 17.9 pour les prompts inédits), ce qui démontre sa capacité à générer des coupes IRM réalistes à partir de descriptions textuelles inédites.

Nos résultats confirment que le LDM conditionné par texte synthétise des coupes IRM réalistes tout en conservant la cohérence anatomique. Les scores FID faibles suggèrent que le modèle apprend des représentations pertinentes, produisant des images proches des scans IRM réels. L'évaluation qualitative (Fig. 6) met également en évidence sa capacité à générer des coupes IRM diversifiées à partir de descriptions textuelles. Le modèle capture avec précision les structures anatomiques décrites dans les prompts et différencie correctement les types de séquences IRM. Bien que ces résultats soient prometteurs, certaines limites subsistent. Le FID mesure la similarité statistique, mais n'évalue pas entièrement la pertinence clinique des images générées. Les travaux futurs incluront une évaluation par des radiologues experts afin de valider la cohérence anatomique et pathologique des images synthétiques.

6 Conclusion

Nous avons introduit un LDM conditionné par texte pour la génération d'IRM cérébrales synthétiques, en exploitant des descriptions anatomiques comme information de conditionnement. Grâce à l'intégration des embeddings BiomedBERT et d'un mécanisme de cross-attention, le LDM génère des coupes IRM réalistes et anatomiquement cohérentes, contribuant ainsi à pallier la pénurie de données en imagerie médicale. Nos travaux futurs incluront une évaluation par des radiologues experts pour valider la précision anatomique, l'extension du modèle à la génération d'IRM 3D et l'intégration de conditions pathologiques afin de simuler l'évolution de maladies.

Références

[1] A. G. BARRETO, J. M. de OLIVEIRA, F. N. B. GOIS, P. C. CORTEZ et V. H. C. de ALBUQUERQUE : A new generative model for textual descriptions of medical images using transformers enhanced with convolutional neural networks. *Bioengineering*, 10(9):1098, 2023.

[2] Se. ESLAMI, G. de MELO et C. MEINEL : Does clip benefit visual question answering in the medical domain as much as it does in the general domain ? *arXiv preprint arXiv :2112.13906*, 2021.

[3] Y. GU, R. TINN, H. CHENG, M. LUCAS, N. USUYAMA, X. LIU, T. NAUMANN, J. GAO et H. POON : Domain-specific language model pretraining for biomedical natural language processing, 2020.

[4] S. JAYASUMANA, S. RAMALINGAM, A. VEIT, D. GLASNER, A. CHAKRABARTI et S. KUMAR : Rethinking fid : Towards a better evaluation metric for image generation. *In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9307–9315, 2024.

[5] D. P. KINGMA : Auto-encoding variational bayes. *arXiv preprint arXiv :1312.6114*, 2013.

[6] W. HL. PINAYA, M.S. GRAHAM, E. KERFOOT, P-D. TUDOSIU, J. DAFFLON, V. FERNANDEZ, P. SANCHEZ, J. WOLLEB, P.F. DA COSTA, A. PATEL *et al.* : Generative ai for medical imaging : extending the monai framework. *arXiv preprint arXiv :2307.15208*, 2023.

[7] A. RADFORD, J. W. KIM, C. HALLACY, A. RAMESH, G. GOH, S. AGARWAL, G. SASTRY, A. ASKELL, P. MISHKIN, J. CLARK *et al.* : Learning transferable visual models from natural language supervision. *In International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] R. ROMBACH, A. BLATTMANN, D. LORENZ, P. ESSER et B. OMMER : High-resolution image synthesis with latent diffusion models. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[9] Z. WANG, E. P. SIMONCELLI et A. C. BOVIK : Multiscale structural similarity for image quality assessment. *In The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003.

[10] Z. WANG, Z. WU, D. AGARWAL et J. SUN : Medclip : Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv :2210.10163*, 2022.

[11] X. XING, J. NING, Y. NAN et G. YANG : Deep generative models unveil patterns in medical images through vision-language conditioning. *arXiv preprint arXiv :2410.13823*, 2024.