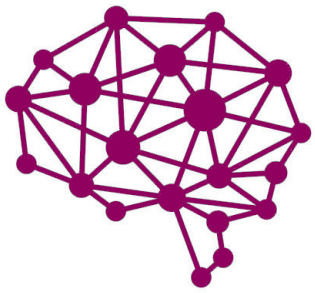
National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

A Unified Framework for Brain Tumour Localisation in MRI Images Using Diffusion Models and Advanced Segmentation Techniques

Diploma Thesis

of

PANAGIOTIS MICHELAKIS



Supervisor: Athanasios Voulodimos
Assistant Professor NTUA

Athens, October 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

A Unified Framework for Brain Tumour Localisation in MRI Images Using Diffusion Models and Advanced Segmentation Techniques

Diploma Thesis

of

PANAGIOTIS MICHELAKIS

Supervisor: Athanasios Voulodimos
Assistant Professor NTUA

Approved by the examination committee on 24th October 2024.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Athanasios Voulodimos	Georgios Stamou	Andreas-Georgios Stafylopatis
Assistant Professor NTUA	Professor NTUA	Professor NTUA

Athens, October 2024

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Panagiotis Michelakis,
Graduate of Electrical and
Computer Engineering,
NTUA

24th October 2024

Περίληψη

Η εφαρμογή της τεχνητής νοημοσύνης (AI) στην ιατρική απεικόνιση έχει βελτιώσει σημαντικά τη διαγνωστική ακρίβεια και αποτελεσματικότητα. Η παρούσα διπλωματική εργασία προτείνει ένα νέο πλαίσιο για την αυτοματοποιημένη ανίχνευση και τμηματοποίηση όγκων του εγκεφάλου σε μαγνητικές τομογραφίες με τη χρήση μοντέλων διάχυσης, του μοντέλου Segment Anything Model (SAM) και του μοντέλου Grounding DINO. Το μοντέλο διάχυσης δημιουργεί αντιπραγματικές εικόνες του υγιούς εγκεφάλου, διευκολύνοντας τον εντοπισμό ανωμαλιών. Το SAM και το Grounding DINO χρησιμοποιούν προτροπές σημείων και κειμένου για την ακριβή τμηματοποίηση των όγκων. Το προτεινόμενο πλαίσιό μας υπερτερεί σταθερά έναντι των μεμονωμένων βασικών μοντέλων, επιδεικνύοντας υψηλές επιδόσεις σε διάφορες μετρικές αξιολόγησης. Παρέχοντας πολλαπλές εικόνες εξόδου, το σύστημα αυτό βοηθά τους ακτινολόγους να λαμβάνουν πιο τεκμηριωμένες αποφάσεις. Αξίζει να σημειωθεί είναι ότι αυτή η μεθοδολογία προορίζεται να βοηθήσει, όχι να αντικαταστήσει, τους επαγγελματίες υγείας, ενισχύοντας τις διαγνωστικές τους ικανότητες και υποστηρίζοντας βελτιωμένα αποτελέσματα για τους ασθενείς.

Λέξεις Κλειδιά

Μοντέλα διάχυσης, αντιπραγματικές εικόνες, κατάτμηση, SAM, GroundingDINO, όγκος εγκεφάλου, μαγνητική τομογραφία

Abstract

The application of artificial intelligence (AI) in medical imaging has significantly enhanced diagnostic accuracy and efficiency. This thesis proposes a novel framework for the automated detection and segmentation of brain tumors in MRI scans using diffusion models, the Segment Anything Model (SAM), and the Grounding DINO model. The diffusion model generates counterfactual images of the healthy brain, facilitating the identification of anomalies. SAM and Grounding DINO use point and text prompts to accurately segment the tumors. Our proposed pipeline consistently outperforms the individual baseline models, demonstrating high performance across various evaluation metrics. By providing multiple output images, this system aids radiologists in making more informed decisions. Crucially, this framework is intended to assist, not replace, medical professionals, enhancing their diagnostic capabilities and supporting improved patient outcomes.

Keywords

Diffusion Models, Counterfactual, Segmentation, SAM, GroundingDINO, Brain Tumour, MRI

to my parents

Acknowledgements

I would like to thank Professor Athanasios Voulodimos, Professor Giorgos Stamou, and Professor Paraskevi Tzouveli for supervising this thesis and for giving me the opportunity to carry it out in the Artificial Intelligence and Learning Systems Laboratory. I also extend my special thanks to Lefteris Tsonis for his guidance, fruitful conversations and the overall excellent collaboration we had over the past year. I would like to thank my parents for their guidance and moral support throughout all these years. Finally, I would like to thank my friends, as I would not be where I am today without their support.

Athens, October 2024

Panagiotis Michelakis

Contents

List of Figures

List of Tables

Κεφάλαιο 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Η τεχνητή νοημοσύνη (ΤΝ) έχει φέρει επανάσταση στην υγειονομική περίθαλψη και θεραπεία, βελτιώνοντας τη διαγνωστική ακρίβεια και την εξατομίκευση των θεραπειών. Στην ιατρική απεικόνιση, τα μοντέλα ΤΝ [10, 33] εντοπίζουν και κατατμούν ανωμαλίες, όπως όγκους, προσφέροντας σημαντική υποστήριξη στη διάγνωση και θεραπεία. Η μαγνητική τομογραφία (MRI) είναι βασική για την απεικόνιση εγκεφαλικών βλαβών, αλλά η χειροκίνητη ανάλυση είναι χρονοβόρα. Στην παρούσα διπλωματική εργασία, προτείνεται ένα πλαίσιο που συνδυάζει μοντέλα διάχυσης για τη δημιουργία αντιφατικών εικόνων υγιούς εγκεφάλου και το SAM και GroundingDINO για ακριβή κατάτμηση όγκων. Το πλαίσιο επιδεικνύει υψηλές επιδόσεις σε διάφορες μετρικές και μπορεί να βοηθήσει τους ακτινολόγους στη λήψη τεκμηριωμένων αποφάσεων, χωρίς να τους αντικαθιστά.

1.2 Πιθανοτικά Μοντέλα Διάχυσης Αποθορυβοποίησης

1.2.1 Έμπνευση από τη Θερμοδυναμική

Η εργασία των Sohl-Dickstein et al. [41] πρωτοπόρησε στο πεδίο των μοντέλων διάχυσης, θέτοντας τις βάσεις για την ανάπτυξη των μεθόδων που ακολούθησαν. Στην εργασία τους, οι συγγραφείς εισήγαγαν τη βασική ιδέα της διάχυσης με έναν προωθητικό μηχανισμό όπου το αρχικό δεδομένο διαβρώνεται σταδιακά με θόρυβο μέσω μιας διακριτής αλυσίδας Markov και, στη συνέχεια, μέσω ενός νευρωνικού δικτύου, το οποίο εκπαιδεύεται να απομακρύνει τον θόρυβο (αποθορυβοποίηση) και να επαναφέρει την αρχική εικόνα. Η εκπαίδευση βασίστηκε στη μεγιστοποίηση της λογαριθμικής πιθανοφάνειας, και η μεθοδολογία ενσωμάτωσε τη μέθοδο Monte Carlo για την ακριβή δειγματοληψία.

1.2.2 Επέκταση των Ιδεών

Παρόλο που οι Sohl-Dickstein et al. [41] έθεσαν τα βασικά θεμέλια για τη διαδικασία διάχυσης, πολλές σημαντικές λεπτομέρειες έλειπαν. Οι Ho et al. [16] ανέπτυξαν περαιτέρω τις αρχικές ιδέες και εισήγαγαν έννοιες όπως το πρόγραμμα μεταβολής της διακύμανσης και

τη διαδικασία δειγματοληψίας DDPM. Ορίζουν επίσης τον μηχανισμό προώθησης ως μια διαδικασία *Markov* όπου σταδιακά προστίθεται θόρυβος στο αρχικό δείγμα. Στη συνέχεια, εκπαιδεύεται ένα νευρωνικό δίκτυο για να προβλέψει τη μέση και τη διακύμανση, προκειμένου να αποκαταστήσει το αρχικό δείγμα κατά τη διάρκεια της διαδικασίας επαναφοράς.

Αναλυτικότερα, παρατίθεται παρακάτω η μαθηματική θεμελίωση για την προαναφερθείσα διαδικασία που εδραιώθηκε από την εν λόγω εργασία των Ho et al.

Η ***προς-τα-εμπρός διαδικασία διάχυσης*** χαρακτηρίζεται από την παρακάτω Μαρκοβιανή διαδικασία:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-b_t}\cdot x_{t-1}, b_t\cdot I), \forall t \in 1,\ldots,T \tag{1.1}$$

όπου **T** είναι ο αριθμός των βημάτων διάχυσης, $b1,\ldots,b_T \in [0,1)$ είναι το πρόγραμμα διακύμανσης , **I** είναι ο ταυτοτικός πίνακας που έχει τις ίδιες διαστάσεις με την εικόνα εισόδου x_0, και $\mathcal{N}(x;\mu,\sigma)$ αντιπροσωπεύει την κανονική κατανομή μέσου όρου μ και συνδιακύμανσης σ που παράγει το x.

Στην ***προς-τα-πίσω διαδικασία διάχυσης***, ξεκινάμε από ένα δείγμα $x_T \sim \mathcal{N}(0,I)$ και παράγουμε νέα δείγματα από το $p(x_0)$ ακολουθώντας τα αντίστροφα βήματα:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\mu(x_t,t),\Sigma(x_t,t)) \tag{1.2}$$

Για να προσεγγίσουμε τα αντίστροφα βήματα, χρησιμοποιούμε ένα νευρωνικό δίκτυο, θ, το οποιό εκπαιδεύεται να προβλέπει το μέσο, μ, και τη συνδιασπορά, Σ, για κάθε βήμα:

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1};\mu_\theta(x_t,t),\Sigma_\theta(x_t,t)) \tag{1.3}$$

Όσον αφορά στα διαταραγμένα δεδομένα, η διακριτή αλυσίδα *Markov* που περιγράφει τη διαδικασία διάχυσης προς τα εμπρός μπορεί να εκφραστεί ως εξής:

$$x_t = \sqrt{1-b_t}x_{t-1} + \sqrt{b_t}\epsilon \tag{1.4}$$

όπου:

$$x_0 \sim p(x_0)$$

$$\epsilon \sim \mathcal{N}(0,I)$$

Αξίζει να σημειωθεί ο παρακάτω μετασχηματισμός ο οποίος μας δίνει η δυνατότητα να λάβουμε ένα δείγμα x_t κατευθείαν από το αρχικό δείγμα x_0:

$$x_t = \sqrt{\bar{a}_t}x_0 + \sqrt{1-\bar{a}_t}\epsilon \tag{1.5}$$

όπου:

$$a_t = 1 - b_t$$

$$\bar{a}_t = \prod_{s=0}^{t} a_s$$

Η αντίστροφη διαδικασία περιγράφεται σύμφωνα με την παρακάτω εξίσωση:

$$x_{t-1} = \mu_\partial(x_t, t) + \sqrt{\Sigma_\partial(x_t, t)}z \tag{1.6}$$

όπου:

$$x_T \sim \mathcal{N}(0, I)$$

$$z \sim \mathcal{N}(0, I)$$

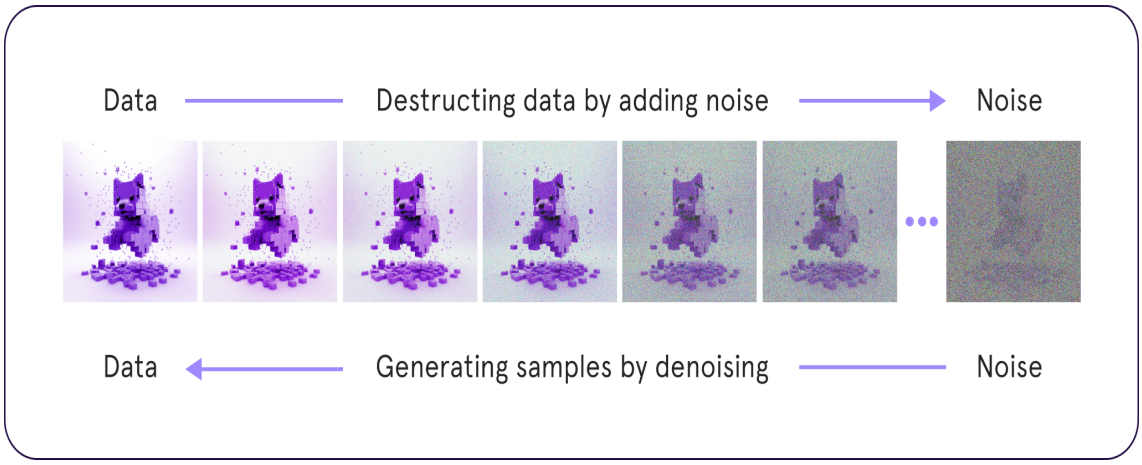


Figure 1.1. *Η προς-τα-εμπρός και η προς-τα-πίσω διαδικασία διάχυσης.*
[1]

Η συνάρτηση απωλειών την χρονική στιγμή t που χρησιμοποιείται για την εκπαίδευση του νευρωνικού δικτύου είναι η εξής:

$$L_{VLB} = L_{t-1} = KL(q(x_{t-1}|x_t, x_0)||p_\partial(x_{t-1}|x_t)) \tag{1.7}$$

όπου το $KL()$ δηλώνει την *Kullback – Leibler* απόκλιση.

Μετά από μία σειρά μαθηματικών απλοποιήσεων, η εν λόγω συνάρτηση κόστους μπορεί να γραφτεί ως εξής:

$$L_{simple} = \mathbb{E}_{t,x_0,\epsilon}\left[||\epsilon - \epsilon_\partial(x_t, t)||_2^2\right] \tag{1.8}$$

Βάσει όλων των παραπάνω, λαμβάνουμε την παρακάτω εξίσωση για τη διαδικασία δειγματοληψίας από την αρχική κατανομή:

$$x_{t-1} = \frac{1}{\sqrt{a_t}}\left[x_t - \frac{1 - a_t}{\sqrt{1 - \bar{a}_t}}\epsilon_\partial(x_t, t)\right] + \sigma_t z \tag{1.9}$$

1.2.3 Βελτιστοποιήσεις και Αποτελέσματα Αιχμής

Παρά το γεγονός ότι τα DDPMs κατάφεραν να παράγουν δείγματα υψηλής ποιότητας, δυσκολεύτηκαν να πετύχουν ανταγωνιστικά αποτελέσματα στη μετρική της λογαριθμικής πιθανοφάνειας. Η εργασία των Dhariwal et al. [29] προτείνει τροποποιήσεις που βελτιστοποιούν αυτή τη μετρική, διατηρώντας ταυτόχρονα υψηλής ποιότητας αποτελέσματα. Οι προτεινόμενες βελτιώσεις περιλαμβάνουν την εκμάθηση της μήτρας διακύμανσης, τη χρήση ενός προγράμματος θορύβου με συνάρτηση συνημίτονου και την εφαρμογή δειγματοληψίας σημασίας για πιο αποτελεσματική εκπαίδευση.

1.2.4 Ταχύτερη Δειγματοληψία με τη Χρήση Μη-Μαρκοβιανών Διαδικασιών

Τα DDPMs μπορούν να παράγουν υψηλής ποιότητας δείγματα, αλλά η διαδικασία διάχυσης απαιτεί πολλές επαναλήψεις για να ολοκληρωθεί η δειγματοληψία. Η εργασία των Song et al. [42] πρότεινε μια γενίκευση της διαδικασίας διάχυσης σε μια μη-Μαρκοβιανή διαδικασία, επιτρέποντας ταχύτερη δειγματοληψία με λιγότερα βήματα. Με τη χρήση του μοντέλου DDIM (Denoising Diffusion Implicit Models), η ταχύτητα της δειγματοληψίας αυξήθηκε σημαντικά χωρίς μεγάλη απώλεια στην ποιότητα των παραγόμενων δειγμάτων.

Στο πλαίσιο αυτό, ο μετασχηματισμός από μια Μαρκοβιανή διαδικασία διάχυσης σε μια πιο γενική μη-Μαρκοβιανή διαδικασία εξαγωγής συμπερασμάτων επιτυγχάνεται με την εισαγωγή του αρχικού δείγματος ως συνθήκη στην εμπρόσθια και την αντίστροφη κατανομή πιθανότητα:

$$q_\sigma(x_{1\dots T}|x_0) = q_\sigma(x_T|x_0)\prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0) \tag{1.10}$$

Η εξίσωση για τα δείγματα κατά την αντίστροφη διαδικασία βρίσκεται μετά την εφαρμογή του κανόνα του *Bayes* και άλλων μαθηματικών τύπων. Αυτή φαίνεται παρακάτω:

$$x_{t-1} = \sqrt{a_t - 1}\tilde{x_0}(t) + \sqrt{1 - a_{t-1} - \sigma_t^2}\cdot\epsilon_\vartheta^{(t)}(x_t) + \sigma_t\epsilon_t \tag{1.11}$$

Όταν $\sigma_t = 0$, η διαδικασία αποκτά ντετερμινιστικό χαρακτήρα, αφού ο συντελεστής της θορύβου, b_t μηδενίζεται. Οι Song et al. αποδεικνύουν στη συνέχεια ότι επιλέγοντας $\sigma_t = 0$ (DDIMs), το μήκος της τροχιάς δειγματοληψίας μειώνεται και επιτυγχάνεται υψηλότερη υπολογιστική αποτελεσματικότητα με ελάχιστες θυσίες στην ποιότητα των παραγόμενων δειγμάτων.

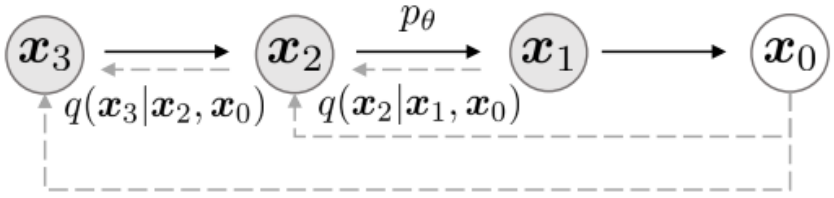


Figure 1.2. *Μη-Μαρκοβιανό Μοντέλο Εξαγωγής Συμπερασμάτων [42]*

1.3 Score Matching με δυναμική Langevin

Αυτή η δεύτερη υποκατηγορία των μοντέλων διάχυσης επικεντρώνεται σε μια διαφορετική διατύπωση της διαδικασίας διάχυσης. Ο πυρήνας αυτών των μοντέλων βασίζεται στη συνάρτηση βαθμολογίας (*Steinscore*) μιας κατανομής πιθανότητας $p(x)$, η οποία υπολογίζεται ως $\nabla_x logp(x)$. Αυτή η ποσότητα δείχνει την κατεύθυνση προς την οποία πρέπει να κινηθούμε από ένα τυχαίο δείγμα x_0 προς ένα δείγμα x_N σε μια περιοχή με υψηλή πυκνότητα. Ο αλγόριθμος που χρησιμοποιείται για αυτή τη διαδικασία ονομάζεται αλγόριθμος δειγματοληψίας *Langevin*.

Σύμφωνα με τον αλγόριθμο δειγματοληψίας *Langevin*, λαμβάνουμε την παρακάτω επαναληπτική διαδικασία:

$$x_t = x_{t-1} + \frac{\epsilon}{2}\nabla_x logp(x_{t-1}) + \sqrt{\epsilon}z_t, \tag{1.12}$$

όπου $z_t \sim N(0, I)$, $\epsilon > 0$ και $x_0 \sim p(x_0)$ (*priordistribution*).

Η συνάρτηση που ελαχιστοποιείται για την εκπαίδευση του νευρωνικού δικτύου είναι η εξής:

$$L_{sm} = \mathbb{E}_{x\sim p(x)}\|s_\partial(x) - \nabla_x logp(x)\|_2^2 \tag{1.13}$$

όπου s_∂ είναι το *SteinScore* και ορίζεται ως:

$$s_\partial(x) \approx \nabla_x logp(x) \tag{1.14}$$

Το κύριο πρόβλημα με την παραπάνω διαδικασία είναι ότι δεν γνωρίζουμε την συνάρτηση βαθμολογίας. Αν την γνωρίζαμε, θα ξέραμε και την αρχική κατανομή πιθανότητας. Προς αντιμετώπιση αυτού του προβλήματος, οι Song et al. πρότειναν την παρακάτω υπο συνθήκη συνάρτηση βαθμολογίας:

$$\nabla_x logp_\sigma(\hat{x}|x) = -\frac{\hat{x} - x}{\sigma^2} \tag{1.15}$$

Έτσι, η συνάρτηση απώλειας γράφεται:

$$L_{dsm} = \frac{1}{L}\sum_{t=1}^{L} \lambda(\sigma_t)\mathbb{E}_{p(x)}\mathbb{E}_{\sigma x\sim p_{\sigma_t}(\hat{x}|x)}\|s_\partial(\hat{x}, \sigma_t) + \frac{\hat{x} - x}{\sigma_t^2}\|_2^2, \tag{1.16}$$

Όσον αφορά τη δειγματοληψία, οι Song et al. πρότειναν μια τροποποιημένη μορφή της δειγματοληψίας *Langevin*, τον αλγόριθμο Annealed Langevin Dynamics. Ο αλγόριθμος αυτός φαίνεται παρακάτω.

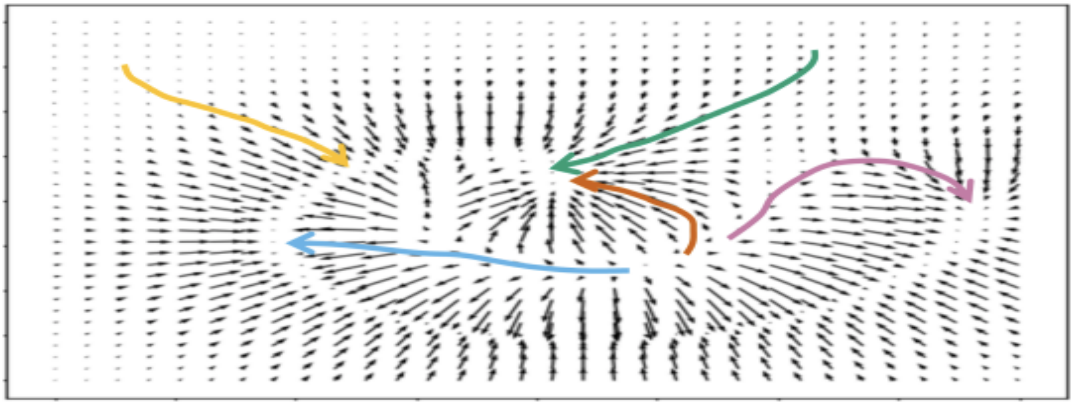


Figure 1.3. *Οπτικοποίηση της τροχιάς με πρόβλεψη του σκορ. Ένα σκορ είναι μια κατεύθυνση για τα επόμενα χρονικά βήματα. Τα δείγματα αποθορυβοποιούνται ως προς την κατεύθυνση σε κάθε θέση. Τα χρώματα αντιπροσωπεύουν τις τροχιές διαφορετικών δειγμάτων. [5]*

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\,\mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

Figure 1.4. *The Annealed Langevin Dynamics Algorithm [43]*

1.4 Στοχαστικές Διαφορικές Εξισώσεις

Η τρίτη υποκατηγορία των μοντέλων διάχυσης αποτελεί μια γενίκευση των δύο προηγούμενων, καθώς εδώ η διατύπωση της διαδικασίας διάχυσης είναι συνεχής και όχι διακριτή. Συγκεκριμένα, η διαδικασία διάχυσης περιγράφεται ως η λύση μιας στοχαστικής διαφορικής εξίσωσης (ΣΔΕ). Αυτή η ιδεά επισημοποιήθηκε από τους Song et al. [44] σε μία πολύ σημαντική εργασία για τον κλάδο ονόματι ***"Score-Based Generative Modeling Through Stochastic Differential Equations"***.

Σε αυτό το πλαίσιο, η προς τα εμπρός ΣΔΕ έχει την εξής μορφή:

$$d\mathbf{x} = \mathbf{f}(x, t)dt + g(t)d\mathbf{w} \tag{1.17}$$

Αυτή η εξίσωση είναι η τυπική εξίσωση Itô ΣΔΕ, όπου $f(x, t)$ είναι ο συντελεστής μετατόπισης, $g(t)$ ο συντελεστής διάχυσης, και dw είναι η διαδικασία Wiener (κίνηση Brownian). Η κίνηση Wiener ορίζεται ως $dw = \epsilon\sqrt{dt}$, όπου ϵ είναι τυχαίος θόρυβος με κατανομή $N(0, I)$.

Ο συντελεστής μετατόπισης έχει σχεδιαστεί έτσι ώστε να μηδενίζει σταδιακά τα δεδομένα x_0, ενώ ο συντελεστής διάχυσης ελέγχει πόσος θόρυβος προστίθεται σε κάθε βήμα. Για να μπορέσουμε να παραγάγουμε δεδομένα από την αρχική κατανομή, πρέπει να αντιστρέψουμε αυτήν τη διαδικασία. Αυτό ακριβώς έδειξε ο Anderson στο άρθρο του, παρουσιάζοντας την αντιστραμμένη εξίσωση ΣΔΕ, η οποία δείχνει πώς μπορούμε να ανακτήσουμε δεδομένα από καθαρό θόρυβο, αφαιρώντας τον όρο διάχυσης που αρχικά προκάλεσε την καταστροφή των δεδομένων. Η εν λόγω ανεστραμμένη ΣΔΕ φαίνεται παρακάτω:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_x log p_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}} \tag{1.18}$$

όπου $d\bar{\mathbf{w}}$ είναι μία διαδικασία Wiener που κυλάει αντίστροφα στον χρόνο.

Οι στοχαστικές διαφορικές εξισώσεις (ΣΔΕς) επιδιώκουν να ενοποιήσουν τα *DDPMs* και τα *SMLDs* κάτω από μια κοινή θεωρητική ομπρέλα. Αυτό επιτυγχάνεται μετατρέποντας τις εξισώσεις της διαδικασίας διάχυσης αυτών των μοντέλων στις συνεχείς αντίστοιχές τους.

Έτσι, η αντίστοιχη ΣΔΕ για τα *DDPMs* είναι:

$$d\mathbf{x} = -\frac{1}{2}b(t)\mathbf{x}dt + \sqrt{b(t)}d\mathbf{w} \tag{1.19}$$

Η αντίστοιχη ΣΔΕ για τα *SMLDs* είναι:

$$d\mathbf{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}}d\mathbf{w} \tag{1.20}$$

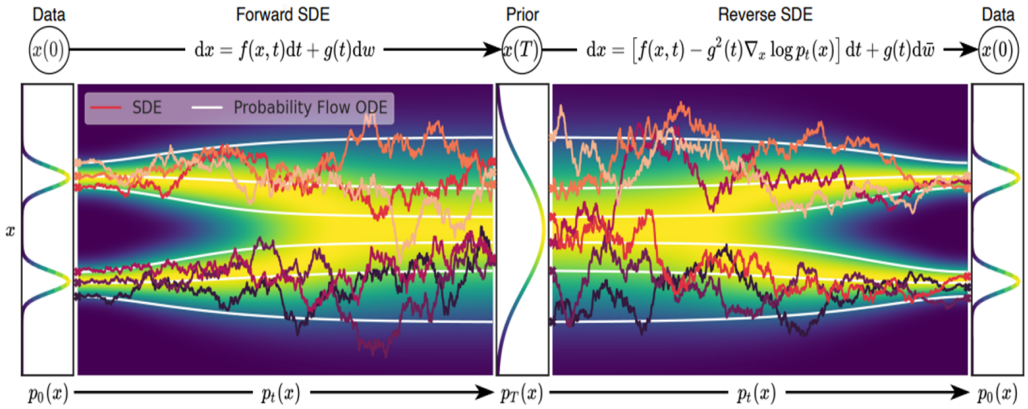


Figure 1.5. *Επισκόπηση της score-based μοντελοποίησης βασισμένης στις ΣΔΕς [44]*

1.5 Παραγωγή Υπό Συνθήκες

Η τρίτη υποκατηγορία των μοντέλων διάχυσης αφορά τη γενίκευση της παραγωγής δειγμάτων υπό συνθήκες (*conditionalgeneration*), δηλαδή τη δημιουργία δειγμάτων βάσει συγκεκριμένων χαρακτηριστικών, όπως είναι οι κειμενικές περιγραφές ή το στυλ ενός άλλου δείγματος. Σε μαθηματικούς όρους, μια συνθήκη y είναι μια επιπλέον είσοδος στο μοντέλο (όπως μια ετικέτα κλάσης ή μια ακολουθία κειμένου) που χρησιμοποιείται για να κατευθύνει τη διαδικασία δημιουργίας δειγμάτων προς μια επιθυμητή κλάση. Αυτή η συνθήκη ενσωματώνεται στην πιθανότητα της αντίστροφης διαδικασίας ως επιπλέον παραμετροποίηση.

Η δημιουργία δειγμάτων υπό συνθήκες προϋποθέτει ότι το μοντέλο δεν προσπαθεί να δειγματοληπτεί από μια γενική κατανομή $p(x)$, αλλά από μια κατανομή της μορφής $p(x|y)$, όπου το y αντιπροσωπεύει την επιπλέον συνθήκη. Οι προηγούμενοι αλγόριθμοι που συζητήθηκαν για τη δειγματοληψία από την κατανομή $p(x)$ πρέπει τώρα να προσαρμοστούν για την κατανομή $p(x|y)$. Ωστόσο, μια απλή υλοποίηση αυτού του είδους μπορεί να οδηγήσει σε περιπτώσεις όπου το μοντέλο αγνοεί ή υποβαθμίζει τη συνθήκη. Για να διορθωθεί αυτό, προτείνεται η χρήση της τεχνικής της **καθοδήγησης** (*Guidance*), η οποία επιτρέπει μεγαλύτερο έλεγχο στη βαρύτητα που αποδίδεται στη συνθήκη. Υπάρχουν δύο κύριοι τύποι καθοδήγησης: η καθοδήγηση με τη χρήση ταξινομητή (*ClassifierGuidance*) και η καθοδήγηση χωρίς ταξινομητή (*Classifier – free Guidance*).

Η πρώτη προσπάθεια για πρακτική παραγωγή δειγμάτων υπό συνθήκες έγινε από τους Dhariwal et al. [7], οι οποίοι πρότειναν τη χρήση ενός ταξινομητή για να κατευθύνουν τη διαδικασία δειγματοληψίας. Αυτή η μέθοδος, που ονομάστηκε Classifier Guidance, χρησιμοποιεί ένα μοντέλο ταξινομητή για να κατευθύνει τη διάχυση στη σωστή κατεύθυνση. Αν και ισχυρή, η συγκεκριμένη μέθοδος απαιτεί τον σχεδιασμό και την εκπαίδευση ενός επιπλέον ταξινομητή, γεγονός που είναι χρονοβόρο και δαπανηρό. Ένα ακόμη πρόβλημα με τη μέθοδο αυτή είναι ότι μπορεί να οδηγήσει σε τυχαίες ή ανεπιθύμητες κατευθύνσεις κατά τη δειγματοληψία.

Αντιμετωπίζοντας αυτά τα προβλήματα, οι Ho et al. [17] πρότειναν τη μέθοδο της καθοδήγησης χωρίς ταξινομητή (*Classifier – freeGuidance*). Σε αυτήν την προσέγγιση, το μοντέλο εκπαιδεύεται ταυτόχρονα τόσο για την παραγωγή δειγμάτων υπό συνθήκες όσο και χωρίς συνθήκες, με το σήμα της συνθήκης να παραλείπεται τυχαία κατά τη διάρκεια της εκπαίδευσης (*dropout*). Αυτή η μέθοδος είναι λιγότερο απαιτητική υπολογιστικά και δίνει στο μοντέλο τη δυνατότητα να παράγει δείγματα και από τις δύο κατανομές ($p(x|y)$ και $p(x)$).

Figure 1.6. *Δείγματα από ένα μοντέλο διάχυσης χωρίς συνθήκη με καθοδήγηση ταξινομητή για την προϋπόθεση της κατηγορίας ¨σκύλος κόργκι¨. Η χρήση της κλίμακας ταξινομητή 1,0 (αριστερά) δεν παράγει πειστικά δείγματα σε αυτή την κατηγορία, ενώ η κλίμακα ταξινομητή 10,0 παράγει πολύ πιο συνεπείς εικόνες με την κατηγορία.[7]*

1.6 Τμηματοποίηση εικόνας και ανίχνευση αντικειμένων

Η τμηματοποίηση εικόνας και η ανίχνευση αντικειμένων είναι κεντρικά θέματα στον τομέα της όρασης υπολογιστών και της μηχανικής μάθησης, με στόχο την κατανόηση και ερμηνεία του περιεχομένου των οπτικών δεδομένων. Αυτά τα καθήκοντα βρίσκουν εφαρμογή σε διάφορους τομείς όπως η αυτόνομη οδήγηση, η ιατρική απεικόνιση, η παρακολούθηση και η ρομποτική.

Η τμηματοποίηση εικόνας περιλαμβάνει την κατάτμηση μιας εικόνας σε πολλαπλά τμήματα ή περιοχές, όπου το καθένα αναπαριστά ένα διαφορετικό αντικείμενο ή μέρος ενός αντικειμένου. Αυτή η διαδικασία βοηθά στην απομόνωση και ανάλυση των διαφόρων στοιχείων της εικόνας, διευκολύνοντας την πιο ακριβή αναγνώριση αντικειμένων και την κατανόηση της σκηνής.

Αντίστοιχα, η ανίχνευση αντικειμένων εστιάζει στον εντοπισμό και την αναγνώριση αντικειμένων μέσα σε μια εικόνα, αναθέτοντάς τους κατηγορίες και καθορίζοντας τα όριά τους με κουτιά περιγράμματος. Η αποτελεσματική ανίχνευση αντικειμένων αντιμετωπίζει προκλήσεις όπως οι διαφορές στην εμφάνιση των αντικειμένων, η κλίμακα, η απόκρυψη και τα σύνθετα υπόβαθρα.

Με την εξέλιξη της τεχνολογίας, οι μέθοδοι αυτές πέρασαν από παραδοσιακές τεχνικές υ-

πολογιστικής όρασης με χαρακτηριστικά που ορίζονται χειροκίνητα σε σύγχρονα μοντέλα μηχανικής μάθησης με τη χρήση νευρωνικών δικτύων. Τα μοντέλα που βασίζονται σε *CNN*, όπως τα Fully Convolutional Networks (FCNs) και $U-Net$, έφεραν σημαντικές βελτιώσεις σε ακρίβεια και αποδοτικότητα. Πιο πρόσφατα, η χρήση των $Transformer-based$ αρχιτεκτονικών έχει μετασχηματίσει ακόμη περισσότερο το τοπίο, με παραδείγματα όπως το Segment Anything Model (SAM) και το GroundingDINO.

Το Segment Anything Model (SAM)[22] είναι ένα ευέλικτο μοντέλο τμηματοποίησης εικόνων που μπορεί να τμηματοποιήσει μια ποικιλία αντικειμένων με βάση διάφορα είδη εισόδων, όπως σημεία, κουτιά ή μάσκες. Η αρχιτεκτονική του βασίζεται σε έναν κωδικοποιητή-αποκωδικοποιητή που εξάγει και ανασυνθέτει χαρακτηριστικά πολλαπλών κλιμάκων από τις εικόνες, επιτρέποντας την ακριβή τμηματοποίηση ακόμη και σε περίπλοκες σκηνές.

Το *GroundingDINO* [25], από την άλλη, συνδυάζει εικόνες και περιγραφές κειμένου για την ανίχνευση αντικειμένων με την αρχιτεκτονική *Transformer*, επιτρέποντας την ανίχνευση αντικειμένων βάσει κειμενικών περιγραφών ή κατηγοριών. Η προσέγγιση αυτή αποδίδει εξαιρετικά αποτελέσματα σε σύνολα δεδομένων όπως το *COCO* [24] και *LVIS* [13], καθιστώντας το *GroundingDINO* μια ισχυρή επιλογή για ανίχνευση αντικειμένων χωρίς προηγούμενη εκπαίδευση.

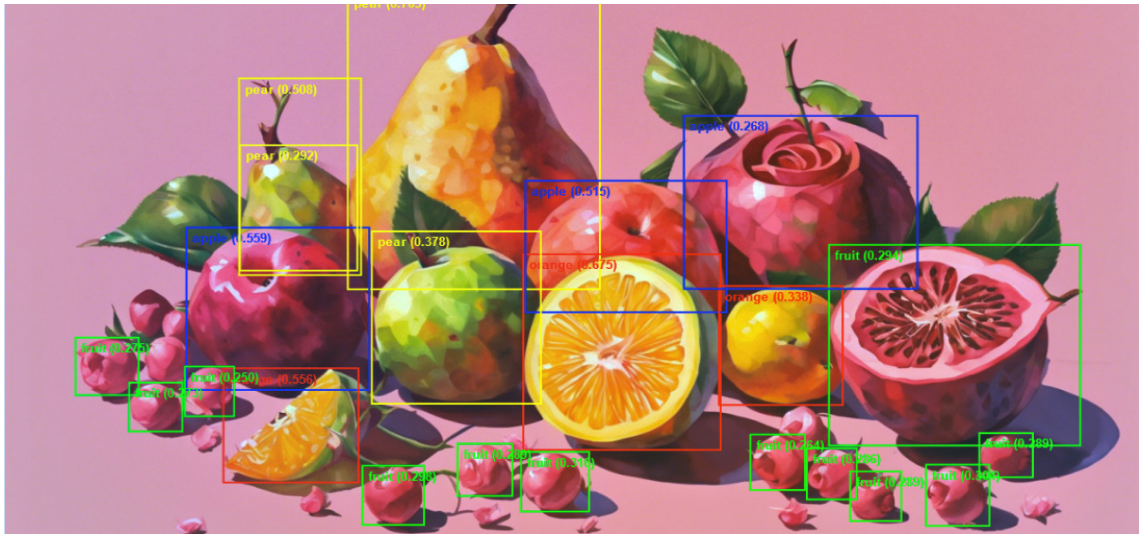


Figure 1.7. *Ανίχνευση Φρούτων σε έναν πίνακα με το GroundingDino. Source: https://www.mlwires.com/grounding-dino-1-5-a-powerful-open-set-object-detection-model/*

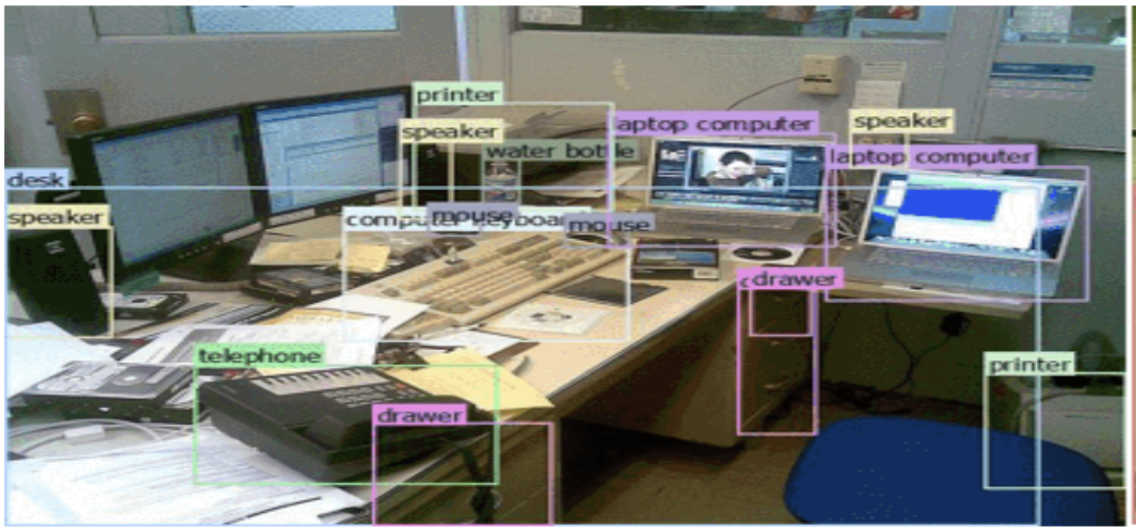


Figure 1.8. *Εντοποισμός αντικειμένων σε ένα γραφείο με το GroundingDino. Source: https://deepdataspace.com/blog/Grounding-DINO-1.5-Pro*

1.7 Κατάτμηση Ιατρικών Εικόνων

Η τμηματοποίηση ιατρικών εικόνων αποτελεί μια από τις πιο κρίσιμες διαδικασίες στην ιατρική ανάλυση και διάγνωση, επιτρέποντας την αυτόματη ανίχνευση και εντοπισμό ανωμαλιών και παθολογικών καταστάσεων σε διάφορα όργανα του ανθρώπινου σώματος. Μέσω της τμηματοποίησης, οι ιατρικές εικόνες μπορούν να αναλυθούν με ακρίβεια, διευκολύνοντας την εξαγωγή πληροφοριών που αφορούν στη διάγνωση ασθενειών, στον σχεδιασμό θεραπευτικών παρεμβάσεων και στην παρακολούθηση της πορείας των ασθενών. Η χρήση της τμηματοποίησης είναι κρίσιμη σε τομείς όπως η ογκολογία, η καρδιολογία και η νευρολογία, καθιστώντας την αναπόσπαστο εργαλείο για την κλινική πρακτική.

Το σύνολο δεδομένων Brain Tumor Segmentation (BraTS) είναι ένα από τα πιο σημαντικά και ευρέως χρησιμοποιούμενα σύνολα δεδομένων στον τομέα της ανάλυσης ιατρικών εικόνων. Αποτελείται από πολυτροπικές μαγνητικές τομογραφίες (*MRI*) που χρησιμοποιούνται για την τμηματοποίηση όγκων εγκεφάλου, συγκεκριμένα γλοιωμάτων, τα οποία είναι από τους πιο κοινούς και επιθετικούς τύπους όγκων. Οι ακολουθίες *MRI* που περιλαμβάνονται, όπως οι $T1$, $T1Gd$, $T2$ και *FLAIR*, προσφέρουν πολύτιμες πληροφορίες σχετικά με την ανατομία και τις παθολογίες του εγκεφάλου. Αυτό επιτρέπει την ανάπτυξη ακριβών αλγορίθμων τμηματοποίησης, οι οποίοι βοηθούν στην καλύτερη διάγνωση και θεραπευτική προσέγγιση των ασθενών.

Πέρα από το *BraTS*, το Medical Segmentation Decathlon (MSD), που δημοσιεύθηκε το 2018, προσφέρει ένα ακόμα πιο ευρύ πρότυπο για την ανάπτυξη αλγορίθμων τμηματοποίησης ιατρικών εικόνων. Το *MSD* περιλαμβάνει δέκα διαφορετικά σύνολα δεδομένων, καθένα από τα οποία αντιπροσωπεύει διαφορετικές ανατομικές δομές και παθολογίες, όπως όγκοι στον εγκέφαλο, το συκώτι, η καρδιά και άλλα όργανα. Αυτό το πολυδιάστατο σύνολο δεδομένων προσφέρει στους ερευνητές την ευκαιρία να αναπτύξουν αλγόριθμους που μπορούν να προσαρμόζονται και να είναι αξιόπιστοι σε διάφορες ιατρικές εφαρμογές, ενθαρρύνοντας έτσι την καινοτομία στον τομέα της ιατρικής απεικόνισης.

Συμπληρωματικά, το *MONAI*, το οποίο αναπτύχθηκε το 2020 από τη *NVIDIA* και το King's College London, παρέχει μια ολοκληρωμένη και ανοιχτή πλατφόρμα για την ανάπτυξη και αξιολόγηση αλγορίθμων βαθιάς μάθησης σε ιατρικές εικόνες. Προσφέροντας εργαλεία και ροές εργασίας για κάθε στάδιο της διαδικασίας, από τη φόρτωση των δεδομένων μέχρι την εκπαίδευση και την ανάπτυξη μοντέλων, το *MONAI* έχει καθιερωθεί ως ένα βασικό εργαλείο για ερευνητές και κλινικούς που εργάζονται στον τομέα της ιατρικής απεικόνισης. Μέσω του MONAI Model Zoo, οι χρήστες έχουν πρόσβαση σε προ-εκπαιδευμένα μοντέλα που διευκολύνουν την ανάπτυξη νέων αλγορίθμων, ενισχύοντας έτσι τη μετάβαση της έρευνας στην κλινική πράξη.

Με αυτόν τον τρόπο, το *BraTS*, το *MSD* και το *MONAI* αποτελούν θεμελιώδη εργαλεία για την πρόοδο της τμηματοποίησης ιατρικών εικόνων, προσφέροντας πρότυπα δεδομένων και πλατ-

φόρμες που επιτρέπουν τη δημιουργία και αξιολόγηση καινοτόμων και ακριβών μεθόδων.

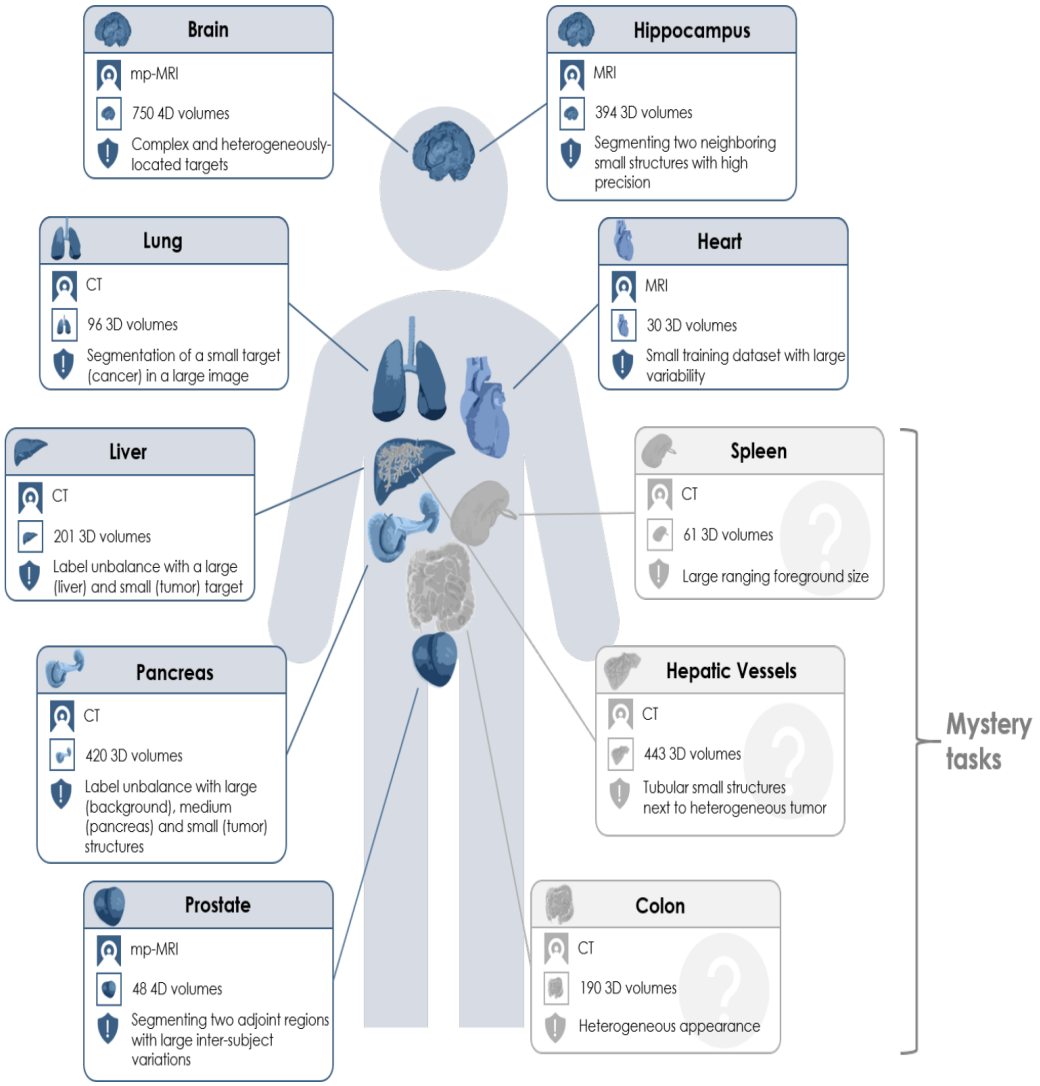


Figure 1.9. *Επισκόπηση των δέκα διαφορετικών εργασιών του Medical Segmentation Decathlon (MSD) [3]*

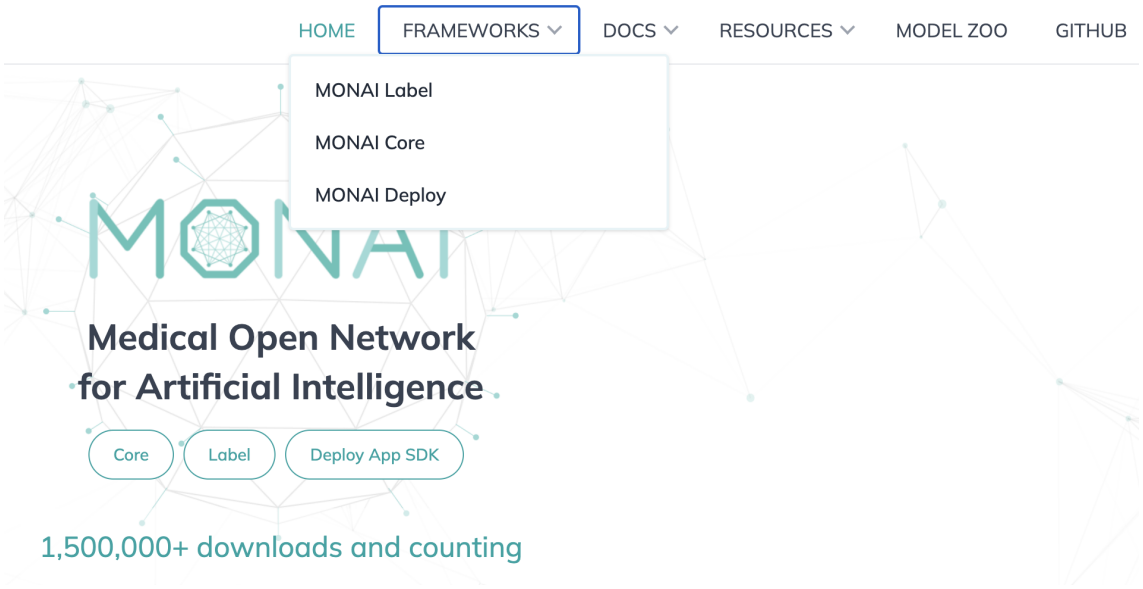


Figure 1.10. *Επισκόπηση του πλαισίου MONAI και των συστατικών του. Source: https://monai.io*

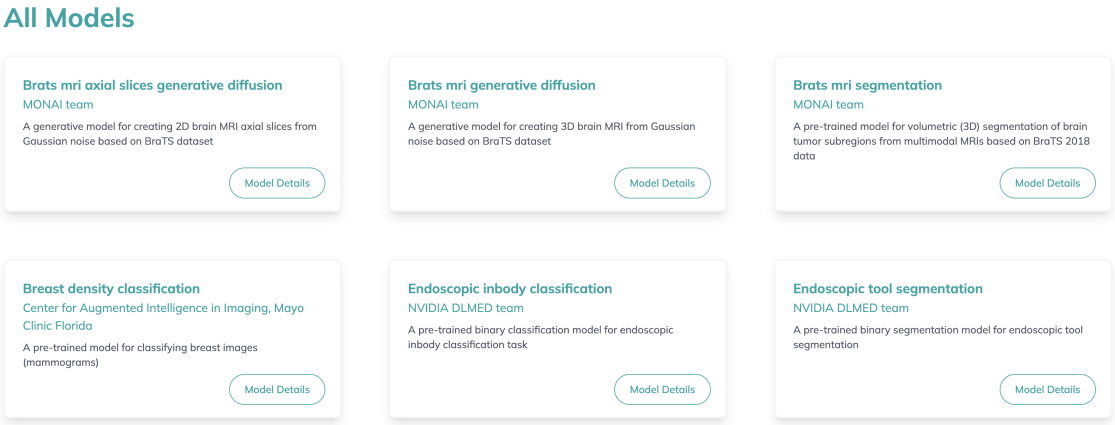


Figure 1.11. *Επισκόπηση ορισμένων από τα διαθέσιμα μοντέλα στο MONAI's Model Zoo. Source: https://monai.io*

1.8 Η Μεθοδολογία μας

Σε αυτό το κεφάλαιο, παρουσιάζουμε τη νέα μας μεθοδολογία για την τμηματοποίηση και τον εντοπισμό βλαβών σε μαγνητικές τομογραφίες (*MRI*) του εγκεφάλου. Βασισμένη στις έννοιες που συζητήθηκαν στα προηγούμενα κεφάλαια, η μεθοδολογία αυτή στοχεύει συγκεκριμένα τις προκλήσεις που σχετίζονται με τον ακριβή εντοπισμό και την τμηματοποίηση εγκεφαλικών βλαβών. Η προσέγγισή μας ενσωματώνει προηγμένες τεχνικές βαθιάς μάθησης, συμπεριλαμβανομένων μοντέλων διάχυσης, μηχανισμών προσοχής και μοντέλων τμηματοποίησης, για τη δημιουργία μιας ισχυρής και αποδοτικής λύσης.

Η μεθοδολογία μας ξεκινά με τη χρήση ενός μοντέλου διάχυσης για τη δημιουργία αντιπραγματικών εικόνων υγιούς εγκεφάλου από *MRI* εγκεφάλων που περιέχουν βλάβες. Αυτές οι αντιπραγματικές εικόνες αναπαριστούν πώς θα έμοιαζε ο εν λόγω εγκέφαλος χωρίς παθολογικές ανωμαλίες. Αυτό το βήμα είναι κρίσιμο καθώς παρέχει ένα σημείο αναφοράς για τον εντοπισμό των βλαβών, αναδεικνύοντας τις αποκλίσεις από την υγιή κατάσταση.

Έπειτα, υπολογίζουμε την διαφορά της αρχικής εικόνας με την αντιπραγματική για να λάβουμε την εικόνα διαφοράς. Σκοπός της εικόνας αυτής είναι να τονίσει τις διαφορές ανάμεσα στις δύο εικόνες.

Μετά τη δημιουργία της εικόνας διαφοράς, εφαρμόζουμε μια σειρά από προκαταρκτικά βήματα επεξεργασίας για την προετοιμασία των δεδομένων για την τμηματοποίηση. Το στάδιο αυτό περιλαμβάνει τον αλλεπάλληλο πολλαπλασιασμό της εικόνας διαφοράς με την αρχική εικόνας με στόχο την ενίσχυση της αντίθεσης της παθογένειας.

Ο πυρήνας της μεθοδολογίας μας αξιοποιεί το μοντέλο Segment Anything Model (SAM) σε δύο ξεχωριστές και ανεξάρτητες ροές εργασιών για την τμηματοποίηση των βλαβών. Στην πρώτη ροή, το *SAM* λαμβάνει ως είσοδο μία σημειακή προτροπή, η οποία συνήθως εισάγεται από έναν γιατρό. Στα πειράματά μας, χρησιμοποιούμε το κέντρο της μάσκας αλήθειας ως σημειακή προτροπή. Το *SAM* έπειτα τμηματοποιεί τη βλάβη βάσει αυτής της προτροπής, παρέχοντας ακριβή εντοπισμό της βλάβης.

Στη δεύτερη ροή, χρησιμοποιούμε ως είσοδο μια κειμενική περιγραφή, η οποία επεξεργάζεται από το μοντέλο *GroundingDINO*. Το *GroundingDINO* δημιουργεί ένα κουτί περιγράμματος γύρω από την περιοχή ενδιαφέροντος, εκτελώντας αποτελεσματικά ανίχνευση αντικειμένων. Αυτό το κουτί περιγράμματος χρησιμοποιείται στη συνέχεια ως είσοδος για το *SAM*, το οποίο τμηματοποιεί τη βλάβη μέσα στην καθορισμένη περιοχή.

Για την αξιολόγηση της απόδοσης της μεθοδολογίας μας, χρησιμοποιούμε διάφορες μετρικές, όπως το Dice score, το Intersection over Union, το AUPRC, την ακρίβεια (*accuracy*), την ανάκληση (*recall*), το *precision*, το F1 score, την απόσταση *Hausdorff* και το Μέσο Συμμετρικό Εμβαδόν Επιφάνειας (*ASSD*). Αυτές οι μετρικές παρέχουν μια ολοκληρωμένη εκτίμηση της ακρίβειας και της αξιοπιστίας των αποτελεσμάτων τμηματοποίησης.

Επιπλέον, πραγματοποιούμε μια μελέτη αφαίρεσης για να κατανοήσουμε τη συμβολή κάθε στοιχείου της μεθοδολογίας μας. Με τη συστηματική αφαίρεση μεμονωμένων στοιχείων και τη μέτρηση της επίπτωσης στην απόδοση, αποκομίζουμε πληροφορίες σχετικά με τη σημασία και την αλληλεπίδραση των διαφόρων στοιχείων στην προσέγγισή μας.

Αυτό το κεφάλαιο προσφέρει μια λεπτομερή ανάλυση κάθε στοιχείου του πλαισίου μας, της πειραματικής ρύθμισης, καθώς και των αποτελεσμάτων της μελέτης αφαίρεσης, αναδεικνύοντας τα δυνατά σημεία και τις πιθανές περιοχές βελτίωσης στην προσέγγισή μας για την τμηματοποίηση και τον εντοπισμό βλαβών σε *MRI* εγκεφάλου.

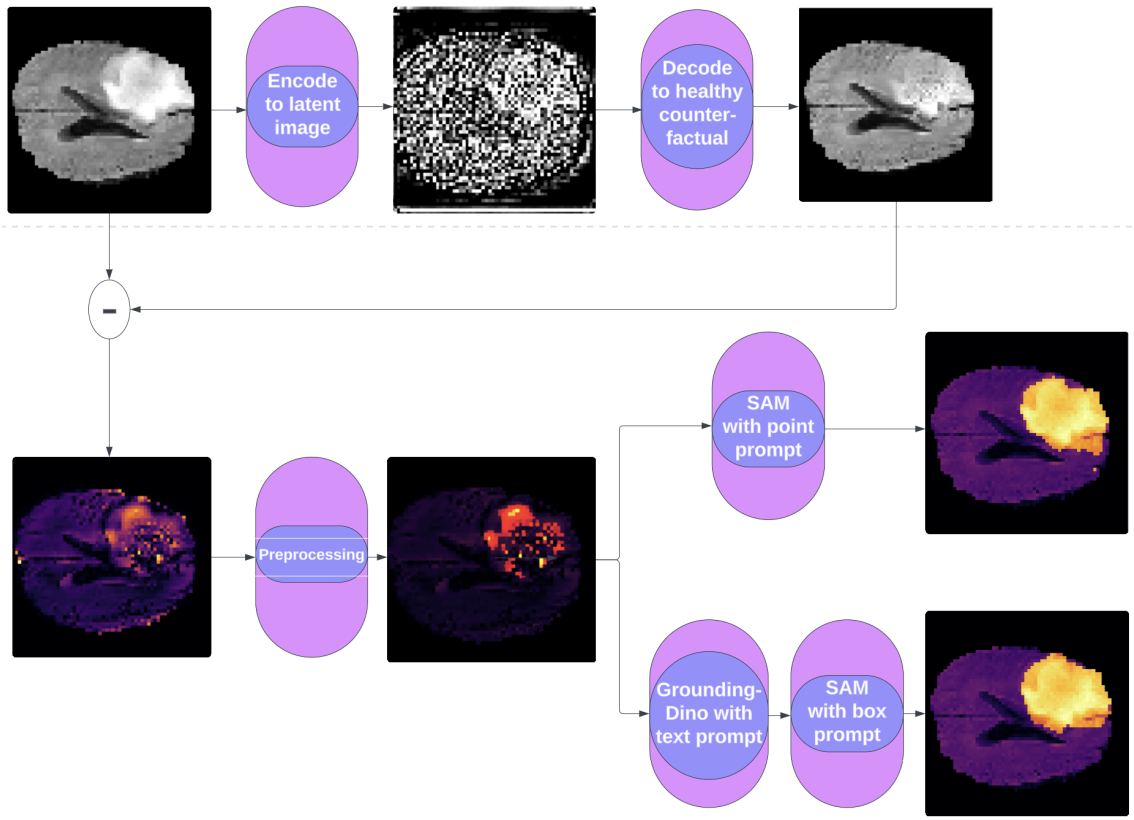


Figure 1.12. *Η μεθοδολογία μας*

1.9 Παραγωγή της Αντιπραγματικής Εικόνας

Στο κεφάλαιο αυτό, παρουσιάζεται με περισσότερες λεπτομέρειες η μεθοδολογία μας για την τμηματοποίηση και τον εντοπισμό αλλοιώσεων σε μαγνητικές τομογραφίες (*MRI*) εγκεφάλου, με έμφαση στη δημιουργία αντιπραγματικών εικόνων που αναδεικνύουν τις διαφορές μεταξύ υγιών και παθολογικών περιοχών. Η διαδικασία στοχεύει στη μετατροπή των εικόνων από την *unhealthy* στην *healthy* κατάσταση, διατηρώντας παράλληλα τα υπόλοιπα χαρακτηριστικά των εικόνων.

Για την εκπαίδευση του μοντέλου, χρησιμοποιείται ένα $U-Net$ που μαθαίνει να προβλέπει τις απαραίτητες αλλαγές για τη δημιουργία υγιών αντιπραγματικών εικόνων από ασθενείς εγκεφάλους. Η διαδικασία αυτή επιτυγχάνεται μέσω της χρήσης του μοντέλου διάχυσης, το οποίο δημιουργεί υγιείς εκδοχές των εικόνων και εντοπίζει τις αλλοιώσεις μέσω της σύγκρισης με τις αρχικές εικόνες.

Κατά τη φάση της δειγματοληψίας, χρησιμοποιείται το μοντέλο *DDIM*, το οποίο επιτρέπει ταχύτερη και πιο ακριβή δειγματοληψία μέσω αντιστρέψιμων μετασχηματισμών. Η διαφορά μεταξύ της αρχικής και της αντιφατικής εικόνας οδηγεί στη δημιουργία ενός *heatmap*, το οποίο τονίζει τις παθολογικές περιοχές για τμηματοποίηση.

Για την καθοδήγηση της διαδικασίας, χρησιμοποιείται η μέθοδος implicit guidance, όπου το μοντέλο εκπαιδεύεται σε συνθήκες με και χωρίς παθολογία, ώστε να μπορεί να παράγει αποτελέσματα που βασίζονται σε συνθήκες (π.χ. υγιείς εγκεφαλικές δομές) με δυναμική ρύθμιση της σημασίας της καθοδήγησης κατά τη δειγματοληψία.

Τέλος, η προσαρμογή της κατάστασης *healthy* γίνεται μέσω ενός μηχανισμού προσοχής (*attention*), ο οποίος ενισχύει την αποτελεσματικότητα της διαδικασίας, βελτιώνοντας την ακρίβεια στην αναγνώριση των αλλοιώσεων.

1.10 Εντοπισμός Παθογένειας

Έχοντας δημιουργήσει την αντιπραγματική εικόνα χωρίς τις παθολογικές περιοχές, υλοποιούμε δύο διαφορετικές προσεγγίσεις τμηματοποίησης και εντοπισμού, που βασίζονται σε προτροπές σημείου και κειμένου, καθώς και η χρήση συνδυασμένων μασκών για καλύτερη κάλυψη των αλλοιώσεων.

Προεπεξεργασία: Πριν τη διαδικασία τμηματοποίησης, η διαφορά εικόνας μεταξύ της αρχικής *MRI* και της *counterfactual* εικόνας ενισχύεται μέσω πολλαπλασιαστικών λειτουργιών. Στόχος είναι να τονιστεί η περιοχή της αλλοίωσης και να μειωθούν τα σφάλματα από άλλες περιοχές που ίσως έχουν επισημανθεί λανθασμένα.

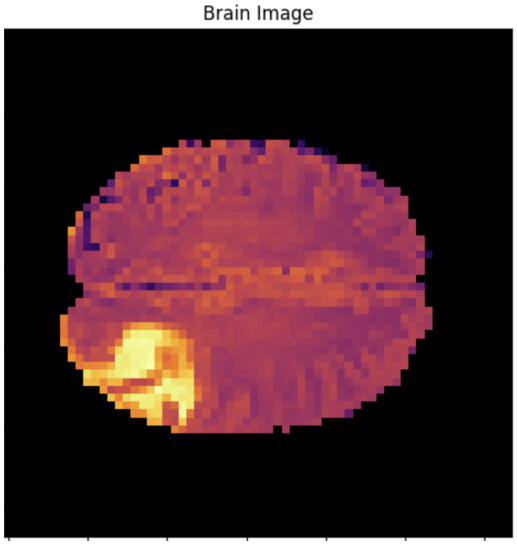


Figure 1.13. *Original brain MRI*

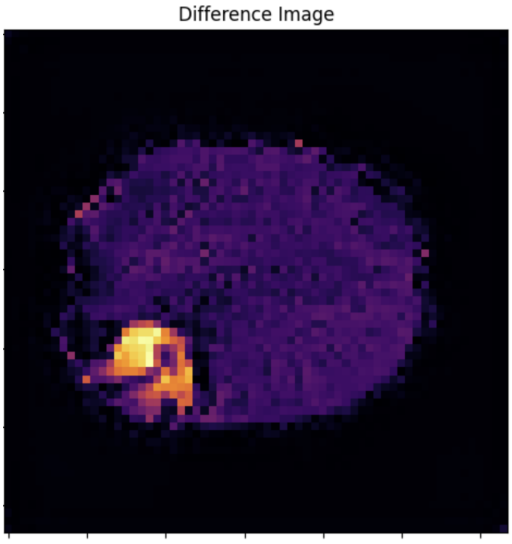


Figure 1.14. *Difference Image*

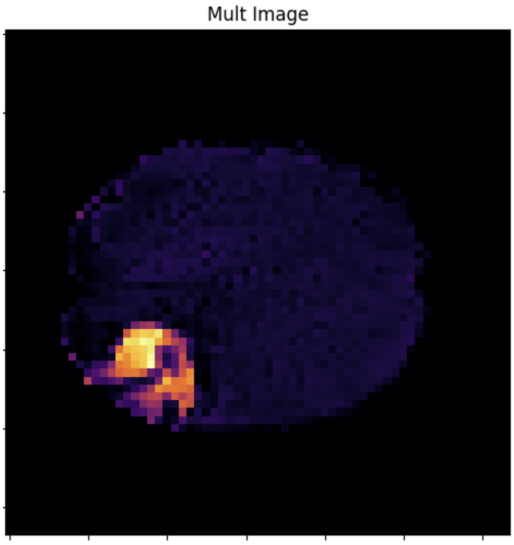


Figure 1.15. *Result of multiplying difference image with the brain MRI image*

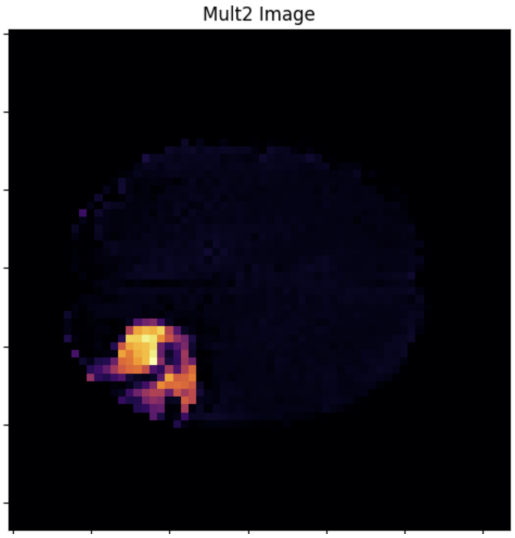


Figure 1.16. *Result of multiplying the difference image with the brain MRI image squared.*

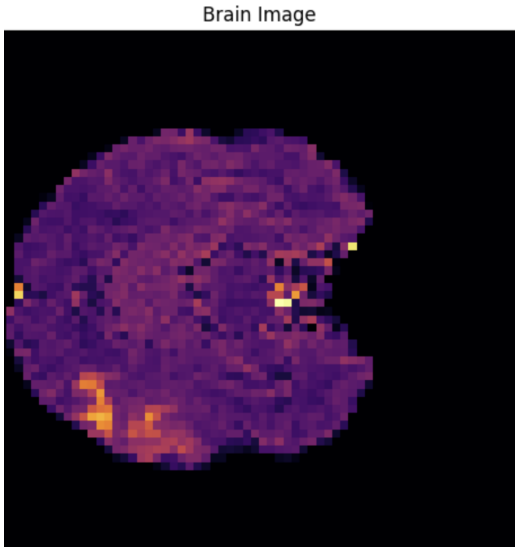


Figure 1.17. *Original brain MRI*

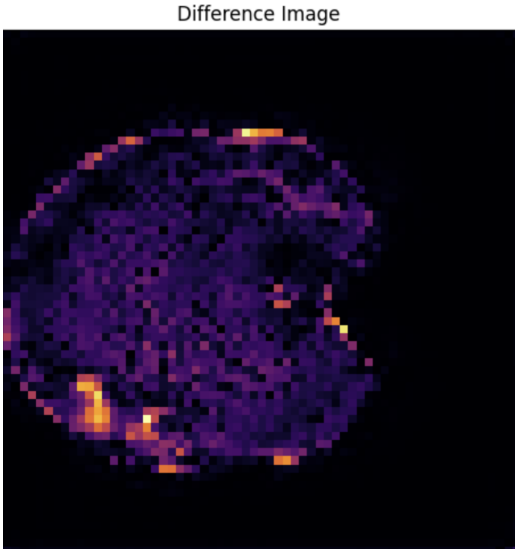


Figure 1.18. *Difference Image*

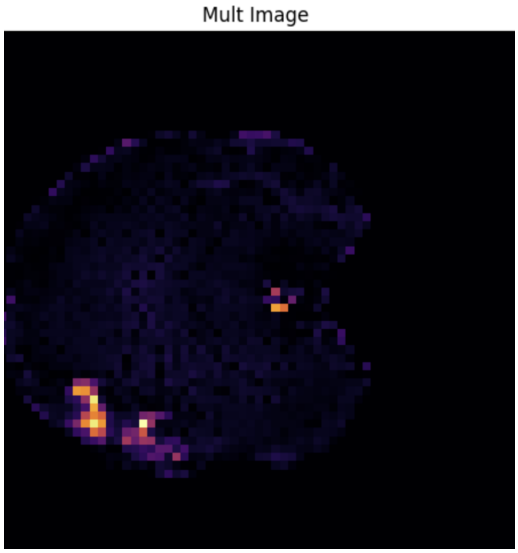


Figure 1.19. *Result of multiplying difference image with the brain MRI image*

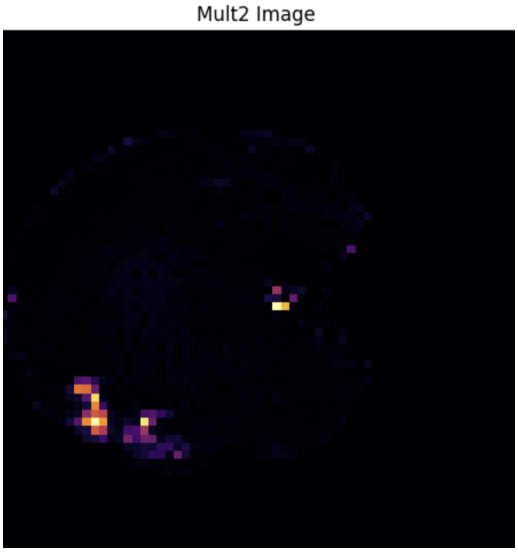


Figure 1.20. *Result of multiplying the difference image with the brain MRI image squared.*

Pipeline με προτροπή σημείου: Στην πρώτη προσέγγιση, χρησιμοποιείται το κέντρο της αλλοίωσης, που ορίζεται από μια πραγματική μάσκα, ως σημείο προτροπής για το μοντέλο *SAM*. Το μοντέλο παράγει διάφορες υποψήφιες μάσκες, από τις οποίες επιλέγεται η μικρότερη για μεγαλύτερη ακρίβεια.

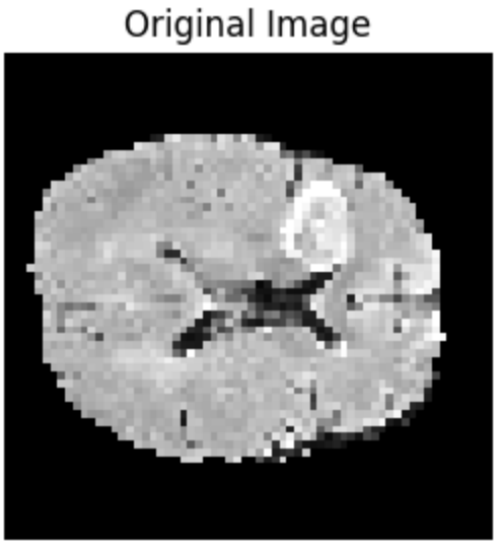


Figure 1.21. *Original brain MRI*

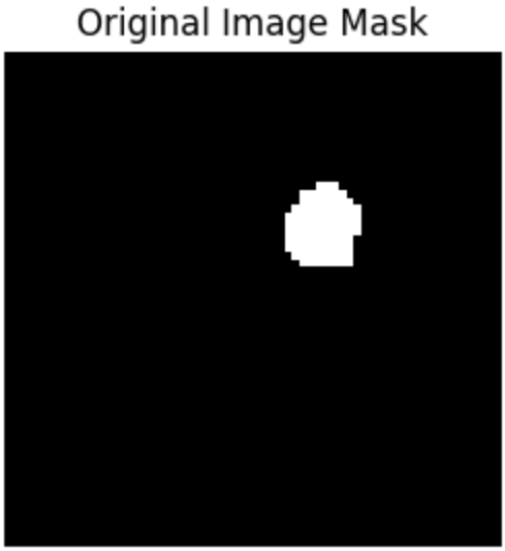


Figure 1.22. *Ground Truth Mask*

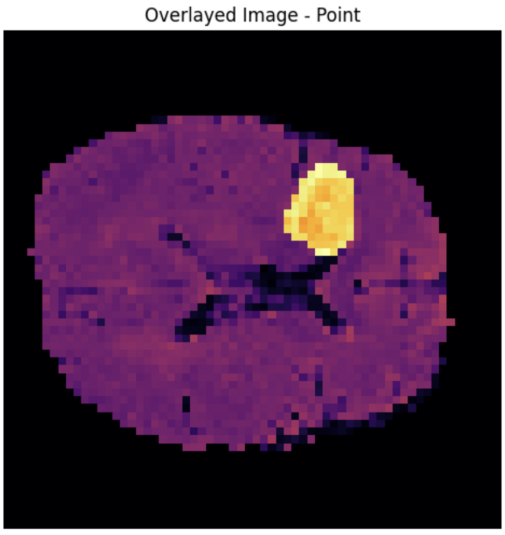


Figure 1.23. *Original brain MRI overlayed with predicted mask*

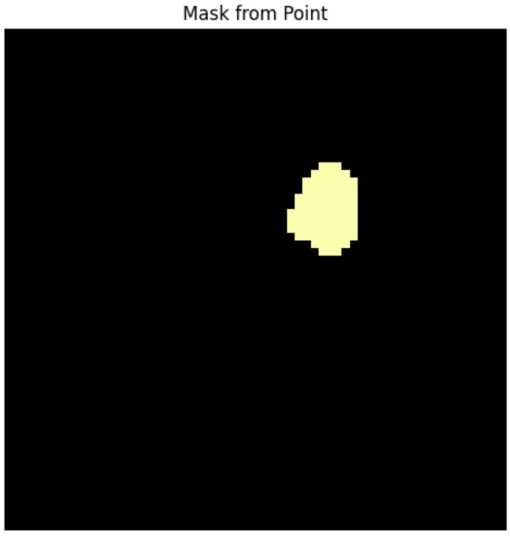


Figure 1.24. *Mask from Point-Prompted Segmentation Pipeline*

Pipeline με προτροπή κειμένου: Στη δεύτερη προσέγγιση, χρησιμοποιείται κείμενο που παρέχεται από έναν γιατρό ή ακτινολόγο για την περιγραφή της αλλοίωσης. Το *GroundingDINO* μοντέλο ανιχνεύει την περιοχή ενδιαφέροντος και παρέχει ένα πλαίσιο (bounding box), το οποίο χρησιμοποιείται ως είσοδος για το *SAM* για την τμηματοποίηση της αλλοίωσης.

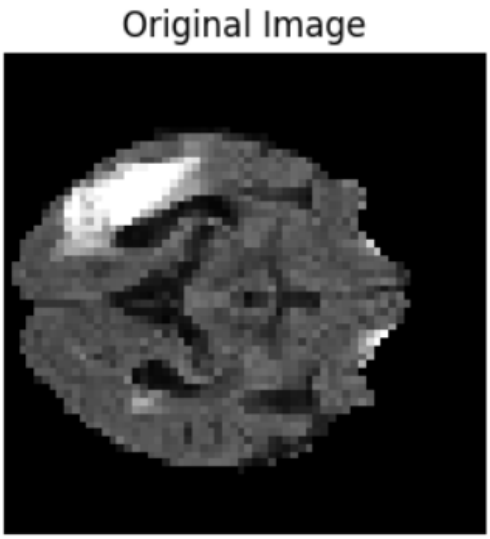


Figure 1.25. *Original brain MRI*

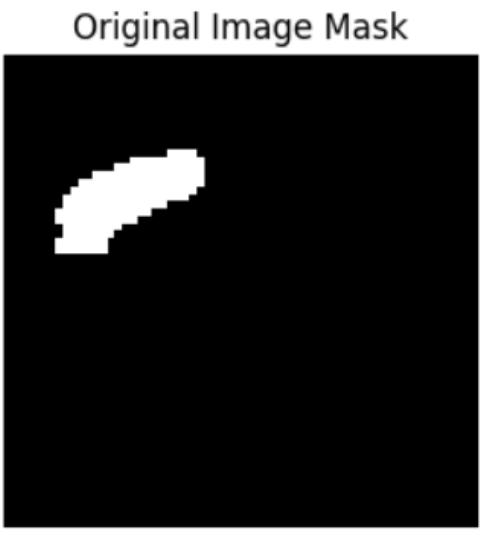


Figure 1.26. *Ground Truth Mask*

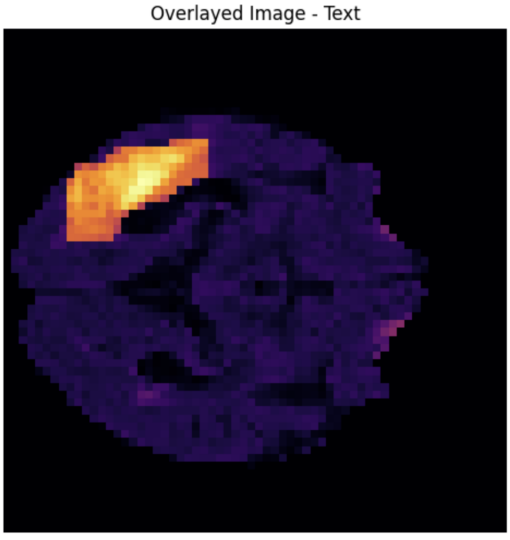


Figure 1.27. *Original brain MRI overlayed with predicted mask*

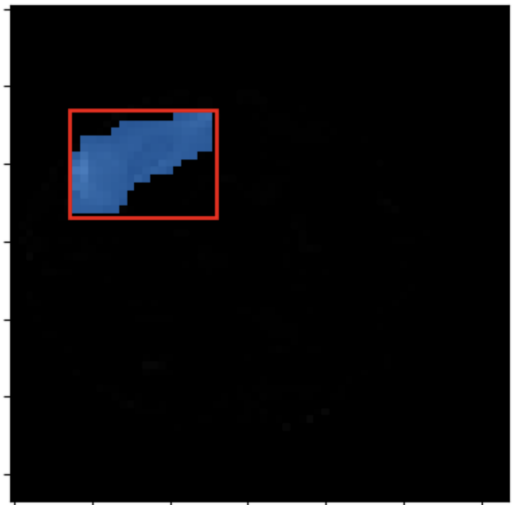


Figure 1.28. *Lesion detection and the corresponding box returned by Grounding DINO*

Συνδυασμός Μασκών: Για κάθε *MRI* εικόνα δημιουργούνται τέσσερις μάσκες: δύο από τις προαναφερθείσες προσεγγίσεις και δύο συνδυαστικές μάσκες από τη διασταύρωση και την ένωση των αρχικών. Αυτή η προσέγγιση διασφαλίζει την κάλυψη όλων των πιθανών παθολογικών περιοχών, ελαχιστοποιώντας τον κίνδυνο να παραληφθούν αλλοιώσεις.

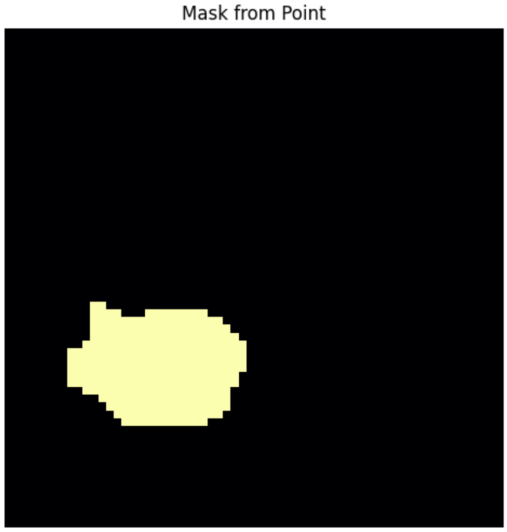


Figure 1.29. *Mask generated by Point-Prompted Pipeline*

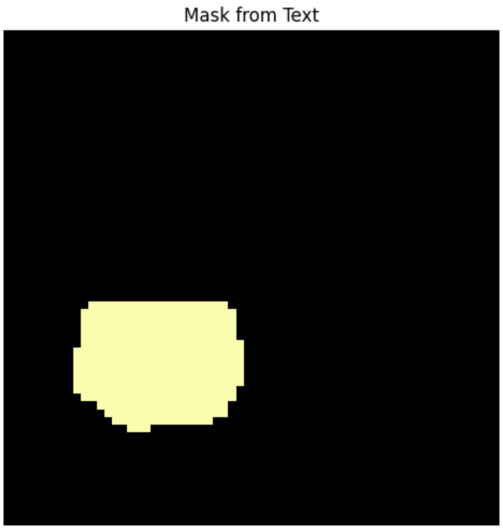


Figure 1.30. *Mask generated by Text-Prompted Pipeline*

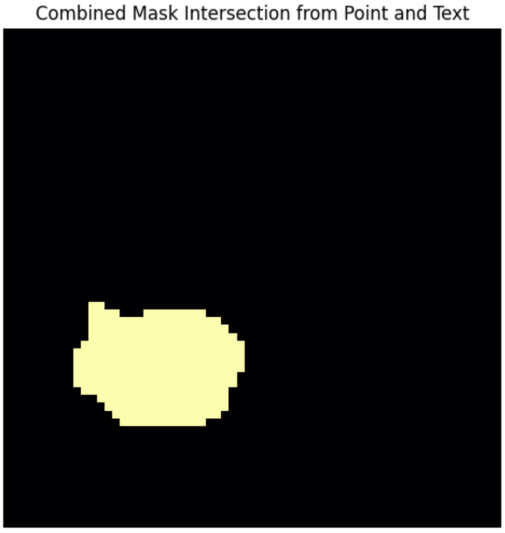


Figure 1.31. *Intersection of generated masks*

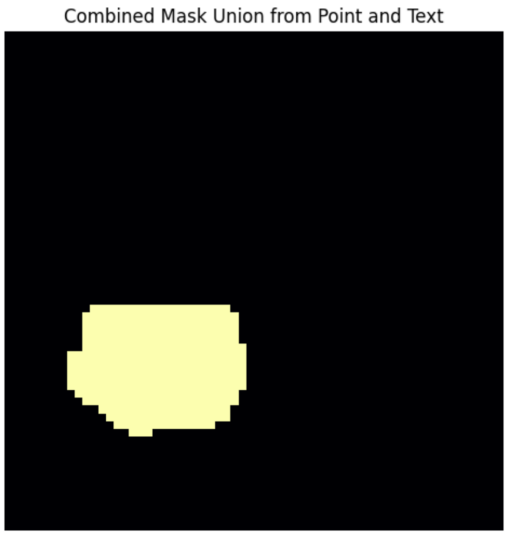


Figure 1.32. *Union of generated masks*

1.11 Μετρικές Αξιολόγησης

Για την αξιολόγηση της απόδοσης του πλαισίου τμηματοποίησής μας, χρησιμοποιούνται διάφορες μετρικές που παρέχουν διαφορετικές οπτικές γωνίες για την ακρίβεια και την αξιοπιστία των αποτελεσμάτων.

Συντελεστής Dice: Μετρά την επικάλυψη μεταξύ της προβλεπόμενης μάσκας και της πραγματικής. Μια τιμή 1 υποδηλώνει τέλεια επικάλυψη, ενώ μια τιμή 0 καμία επικάλυψη.

Intersection over Union (IoU): Υπολογίζει την αναλογία της τομής προς την ένωση των προβλεπόμενων και πραγματικών μασκών, με τιμές από 0 έως 1.

Εμβαδόν Κάτω από την Καμπύλη Precision-Recall (AUPRC) : Αξιολογεί την ισορροπία μεταξύ πρεσισιον και ρεσαλλ, χρήσιμη για μη ισορροπημένα σύνολα δεδομένων.

Precision : Μετρά το ποσοστό των πραγματικών θετικών ανάμεσα σε όλες τις θετικές προβλέψεις.

Recall : Υπολογίζει το ποσοστό των πραγματικών θετικών που εντοπίστηκαν από το σύστημα.

Specificity : Μετρά το ποσοστό των πραγματικών αρνητικών ανάμεσα σε όλα τα αρνητικά παραδείγματα.

F1 Score : Αρμονικός μέσος του *precision* και *recall*, ισορροπώντας την ακρίβεια και την ευαισθησία.

Hausdorff Distance : Μετρά τη μέγιστη απόσταση μεταξύ των ορίων της προβλεπόμενης και της πραγματικής μάσκας, σημαντική για την εξασφάλιση ότι η τμηματοποίηση καλύπτει το σύνολο της βλάβης.

Average Symmetric Surface Distance (ASSD) : Μετρά τη μέση απόσταση μεταξύ των σημείων των ορίων της προβλεπόμενης και πραγματικής μάσκας, προσφέροντας μια πιο συνολική αξιολόγηση της ακρίβειας των ορίων.

Αυτές οι μετρικές εξασφαλίζουν μια ολοκληρωμένη αξιολόγηση της ακρίβειας και της κλινικής σημασίας της προσέγγισής μας στην τμηματοποίηση αλλοιώσεων.

1.12 Πειραματική Διάταξη

Σύνολο Δεδομένων: Για τα πειράματα που διεξήχθησαν σε αυτή τη διπλωματική εργασία, χρησιμοποιήθηκε το σύνολο δεδομένων του Decathlon Task 1, το οποίο επικεντρώνεται σε εγκεφαλικούς όγκους. Το σύνολο αυτό αποτελεί μέρος του Medical Segmentation Decathlon, μιας συλλογής από επισημασμένα ιατρικά δεδομένα σχεδιασμένα για την αξιολόγη-

ση αλγορίθμων τμηματοποίησης. Τα δεδομένα περιλαμβάνουν μαγνητικές τομογραφίες από τέσσερις κατηγορίες (T1, T1Gd, T2 FLAIR). Για τα πειράματά μας, χρησιμοποιήσαμε την κατηγορία FLAIR, η οποία είναι αποτελεσματική στην ανάδειξη ανωμαλιών όπως οι όγκοι. Το σύνολο δεδομένων αποτελείται από 388 τομογραφίες για εκπαίδευση, 96 για επικύρωση και 251 για δοκιμές.

Μετασχηματισμοί Δεδομένων: Πριν την εκπαίδευση, πραγματοποιήθηκαν διάφοροι σημαντικοί μετασχηματισμοί στις εικόνες μαγνητικής τομογραφίας για την προετοιμασία τους. Οι εικόνες προσανατολίστηκαν με το πρότυπο RAS (Right-Anterior-Superior) και επαναδειγματίστηκαν με σταθερή ανάλυση voxel. Στη συνέχεια, τα δεδομένα περικόπηκαν κεντρικά και κανονικοποιήθηκαν σε ένα εύρος μεταξύ 0 και 1. Εφαρμόστηκε τυχαία χωρική περικοπή για την αύξηση δεδομένων, προκειμένου να βελτιωθεί η γενίκευση του μοντέλου.

Αρχιτεκτονική U-Net: Το μοντέλο που χρησιμοποιήθηκε βασίζεται στην αρχιτεκτονική U-Net , ειδικά προσαρμοσμένη για μοντέλα διάχυσης. Το U-Net αποτελείται από έναν κωδικοποιητή και έναν αποκωδικοποιητή με skip connections για τη διατήρηση των χωρικών πληροφοριών. Ο κωδικοποιητής αποτελείται από τέσσερα στάδια με αυξανόμενα πλάτη καναλιών (64, 128, 256, 512), ενώ ο αποκωδικοποιητής ανασυστήνει τις χωρικές διαστάσεις. Σημαντικές βελτιώσεις έγιναν με την προσθήκη μηχανισμών προσοχής, επιτρέποντας στο μοντέλο να εστιάζει σε κρίσιμες περιοχές για ακριβή τμηματοποίηση. Το μοντέλο είναι ενσωματωμένο σε ένα πλαίσιο διάχυσης και χρησιμοποιεί έναν προγραμματιστή (scheduler) για να διαχειρίζεται τη διαδικασία διάχυσης.

Εκπαίδευση: Η διαδικασία εκπαίδευσης περιλάμβανε βασικές παραμέτρους, όπως:

- **Dropout Συνθηκών:** Χρησιμοποιήθηκε ποσοστό 0.15 για την αποτροπή υπερεκπαίδευσης, όπου τυχαία παραλείπονται πληροφορίες κατά τη διάρκεια της εκπαίδευσης.
- **Επαναλήψεις και Batch Size :** Το μοντέλο εκπαιδεύτηκε για 4000 επαναλήψεις με μέγεθος batch 32.
- **Διαστήματα Επικύρωσης:** Κάθε 100 επαναλήψεις πραγματοποιείται επικύρωση για την παρακολούθηση της απόδοσης του μοντέλου.

Υπερπαράμετροι: Οι επιδόσεις του συστήματος επηρεάζονται από βασικές υπερπαραμέτρους, όπως:

- **Text Prompt (TP):** Το κείμενο που χρησιμοποιείται στον αγωγό τμηματοποίησης με προτροπή κειμένου.
- **Βήματα Πολλαπλασιασμού στην Προτροπή Σημείου (PMS):** Αφορά τον αριθμό των φορών που η διαφορά εικόνας πολλαπλασιάζεται με την αρχική εικόνα MRI.
- **Βήματα Πολλαπλασιασμού στην Προτροπή Κειμένου (TMS):** Όπως και το PMS, αλλά για την προτροπή κειμένου.

Πραγματοποιήθηκαν πειράματα με διάφορους συνδυασμούς υπερπαραμέτρων για τη βελτιστοποίηση του πλαισίου. Μερικούς από τους συνδυασμούς που εξετάστηκαν περιλαμβάνουν:

- **TP** = "lesion", **PMS** = 2, **TMS** = 3
- **TP** = "tumour", **PMS** = 1, **TMS** = 2
- **TP** = "anomaly", **PMS** = 3, **TMS** = 3
- **TP** = "lesion", **PMS** = 2, **TMS** = 2

1.13 Πειραματικά Αποτελέσματα

Αφού περιγράψαμε τη μεθοδολογία μας λεπτομερώς στο προηγούμενο κεφάλαιο, παρουσιάζουμε τώρα τα αποτελέσματα που προέκυψαν χρησιμοποιώντας την εν λόγω μεθοδολογία στο Task 1 (Εγκεφαλικός Όγκος) της πρόκλησης Medical Segmentation Decathlon. Αυτό το κεφάλαιο θα παρέχει μια ολοκληρωμένη επισκόπηση της απόδοσης της μεθοδολογία μας σε διάφορες διαμορφώσεις υπερπαραμέτρων.

Τα αποτελέσματα θα παρουσιαστούν σε μορφή πινάκων, με κάθε πίνακα να αντιστοιχεί σε μια μοναδική διαμόρφωση υπερπαραμέτρων. Οι στήλες σε κάθε πίνακα θα αναπαριστούν τις διαφορετικές μεθόδους που χρησιμοποιήθηκαν για τη δημιουργία των μασκών στη μεθοδολογία μας: **Σημείο**, **Κείμενο**, **Τομή** και **Ένωση**. Οι γραμμές θα αντιστοιχούν στις μετρικές αξιολόγησης που χρησιμοποιούνται για την αποτίμηση της απόδοσης της μεθοδολογίας μας. Κάθε μετρική παρέχει μια διαφορετική οπτική στην ακρίβεια και την αξιοπιστία των αποτελεσμάτων τμηματοποίησης.

Εκτός από τα ποσοτικά αποτελέσματα, θα παρέχονται και ορισμένες εικόνες που προέκυψαν από τα πειράματά μας. Αυτές οι εικόνες θα απεικονίζουν την απόδοση της τμηματοποίησης της μεθοδολογίας μας υπό διαφορετικές ρυθμίσεις υπερπαραμέτρων, προσφέροντας μία ολιστική εικόνα για την αποτελεσματικότητα της προσέγγισής μας.

Επιπλέον, θα παρουσιάσουμε τα αποτελέσματα μιας σύντομης μελέτης αφαίρεσης για να διερευνήσουμε τη συμβολή των διαφόρων στοιχείων του πλαισίου μας. Στη μελέτη αυτή, πειραματιζόμαστε με διάφορες αρχιτεκτονικές U-Net, παραλείπουμε εντελώς τα βήματα προεπεξεργασίας και εφαρμόζουμε ακόμη και τον αγωγό SAM-GroundingDINO απευθείας στις αρχικές εικόνες MRI αντί των αντίθετων εικόνων. Αυτή η μελέτη έχει ως στόχο να αναδείξει τη σημασία κάθε στοιχείου και την επίδραση των διαφορετικών αρχιτεκτονικών επιλογών στις μετρικές απόδοσης.

Για τους σκοπούς της ελληνικής περίληψης, παρουσιάζουμε μόνο ένα μέρος των αποτελεσμάτων (λίγες διαμορφώσεις των υπερπαραμέτρων). Αρκετά περισσότεραωαποτελέσματα είναι διαθέσιμα στο αγγλικό κείμενο που ακολουθεί την ελληνική περίληψη. Η επιλεγμένη παρουσίαση αποσκοπεί στην παροχή μιας συνοπτικής επισκόπησης της απόδοσης του

πλαισίου μας, ενώ το πλήρες σύνολο αποτελεσμάτων προσφέρει μια αναλυτικότερη και εκτενέστερη εικόνα των πειραμάτων μας.

Μέσω αυτής της λεπτομερούς παρουσίασης των αποτελεσμάτων, στοχεύουμε να αναδείξουμε τη σταθερότητα και την ευελιξία του πλαισίου μας στην ακριβή τμηματοποίηση εγκεφαλικών όγκων σε εικόνες MRI, ενώ παράλληλα αναδεικνύουμε την επιρροή των διαφόρων υπερπαραμέτρων και των αρχιτεκτονικών επιλογών στις μετρικές απόδοσης.

1.13.1 Lesion - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.821	0.804	0.823
AUPRC	0.761	0.848	0.848	0.785
IoU	0.606	0.731	0.709	0.632
Precision	0.747	0.853	0.942	0.706
Recall	0.762	0.836	0.741	0.857
F1	0.755	0.844	0.830	0.775
Specificity	0.987	0.993	0.998	0.982
Hausdorff	5.919	5.624	5.744	5.785
ASSD	0.450	0.398	0.432	0.418

Table 1.1. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration*

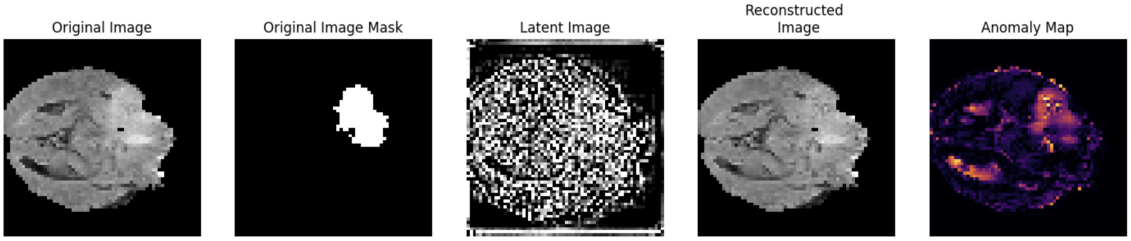


Figure 1.33. *Counterfactual generation*

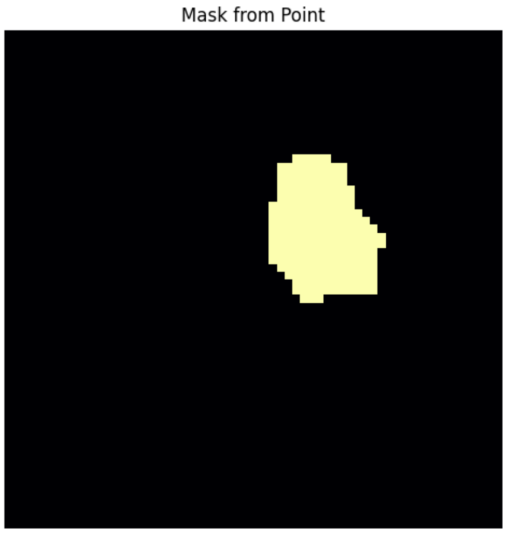


Figure 1.34. *Mask generated by Point-Prompted Pipeline*

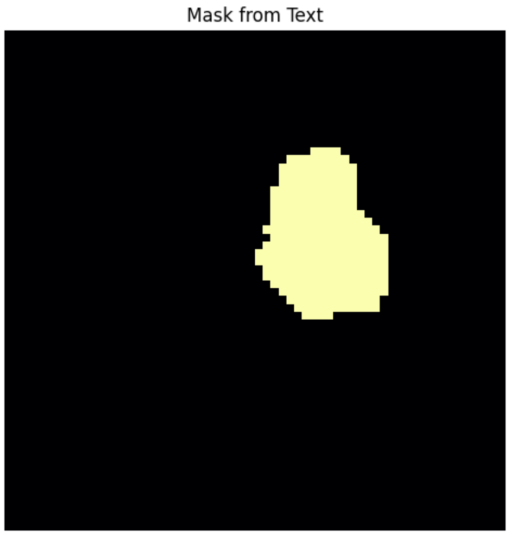


Figure 1.35. *Mask generated by Text-Prompted Pipeline*

1.13.2 Lesion - 2 - 2

	Point	Text	Intersection	Union
Dice	0.799	0.807	0.814	0.792
AUPRC	0.743	0.762	0.789	0.735
IoU	0.580	0.588	0.645	0.539
Precision	0.683	0.640	0.780	0.580
Recall	0.793	0.879	0.788	0.883
F1	0.734	0.741	0.784	0.700
Specificity	0.981	0.974	0.989	0.967
Hausdorff	6.670	6.149	5.710	7.077
ASSD	0.531	0.506	0.429	0.609

Table 1.2. *Evaluation Metrics for the "Lesion - 2 - 2" Configuration*

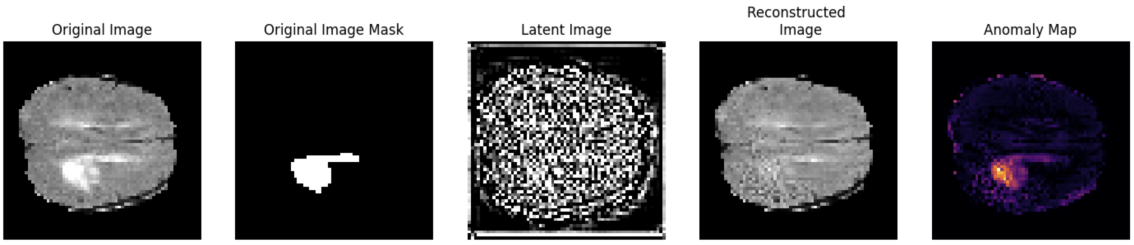


Figure 1.36. *Counterfactual generation*

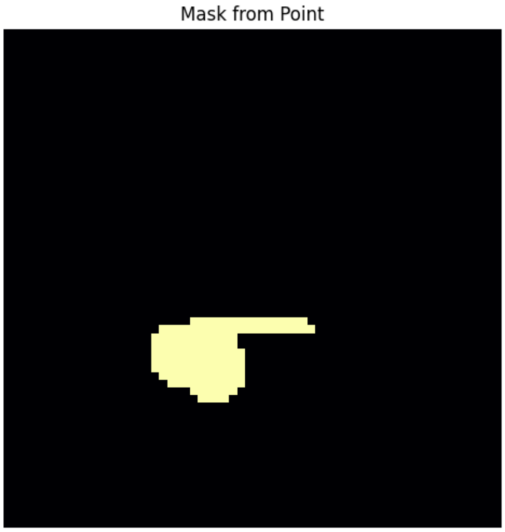


Figure 1.37. *Mask generated by Point-Prompted Pipeline*

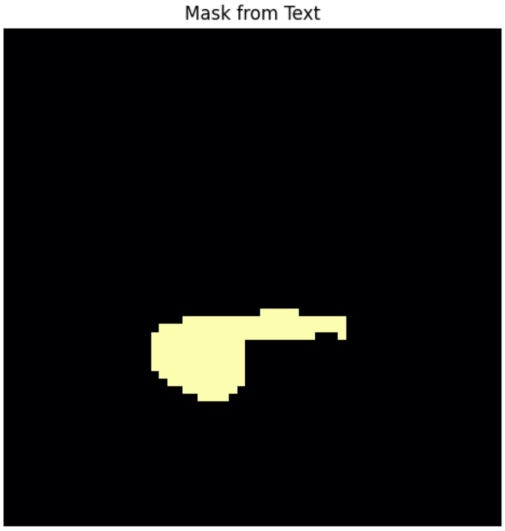


Figure 1.38. *Mask generated by Text-Prompted Pipeline*

1.13.3 Tumour - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.798	0.794	0.810
AUPRC	0.761	0.724	0.842	0.703
IoU	0.606	0.546	0.697	0.498
Precision	0.747	0.621	0.943	0.544
Recall	0.762	0.819	0.728	0.853
F1	0.755	0.706	0.822	0.664
Specificity	0.987	0.974	0.998	0.963
Hausdorff	5.919	6.665	5.861	6.709
ASSD	0.450	0.618	0.462	0.610

Table 1.3. *Evaluation Metrics for the "Tumour - 3 - 3" Configuration*

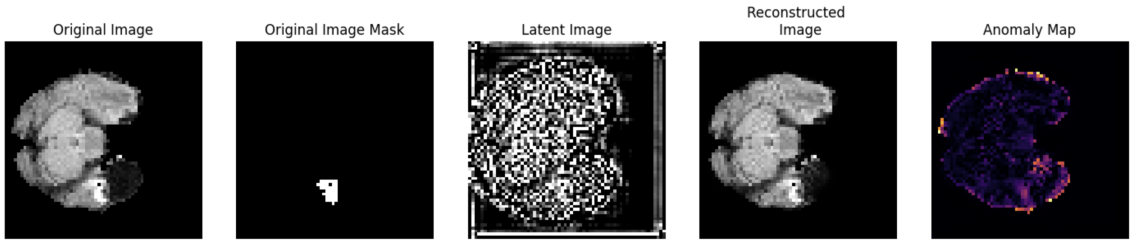


Figure 1.39. *Counterfactual generation*

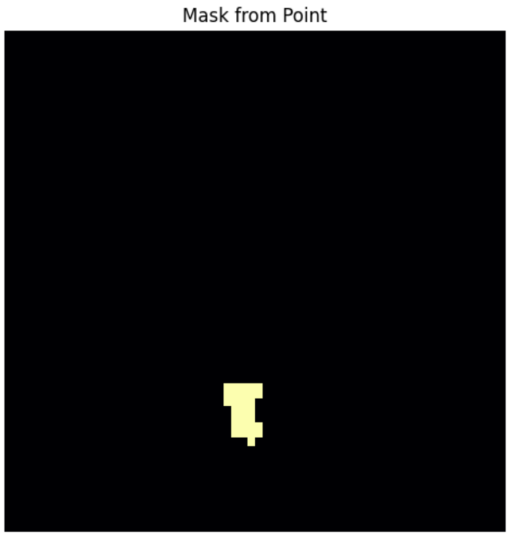


Figure 1.40. *Mask generated by Point-Prompted Pipeline*

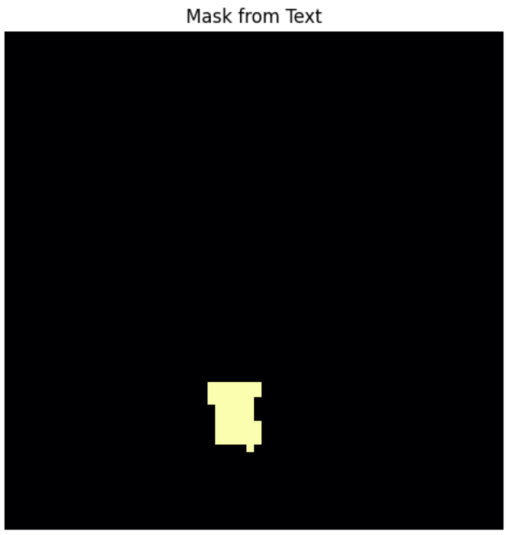


Figure 1.41. *Mask generated by Text-Prompted Pipeline*

1.13.4 Tumour - 1 - 1

	Point	Text	Intersection	Union
Dice	0.774	0.698	0.775	0.697
AUPRC	0.734	0.623	0.735	0.624
IoU	0.566	0.332	0.567	0.333
Precision	0.662	0.346	0.666	0.346
Recall	0.795	0.895	0.793	0.897
F1	0.723	0.500	0.724	0.499
Specificity	0.979	0.912	0.979	0.912
Hausdorff	7.177	10.920	7.039	11.058
ASSD	0.657	1.556	0.652	1.562

Table 1.4. *Evaluation Metrics for the "Tumour - 1 - 1" Configuration*

Figure 1.42. *Counterfactual generation*

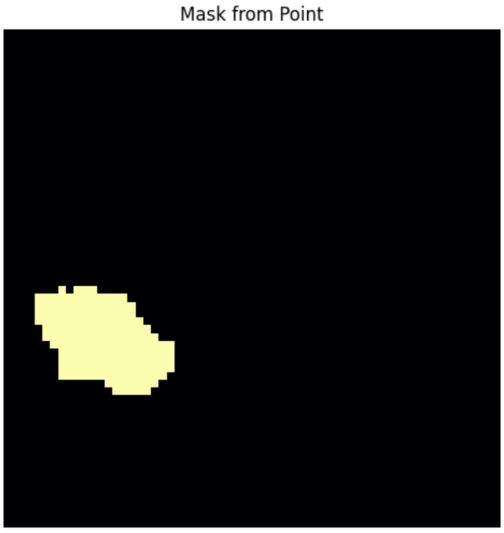


Figure 1.43. *Mask generated by Point-Prompted Pipeline*

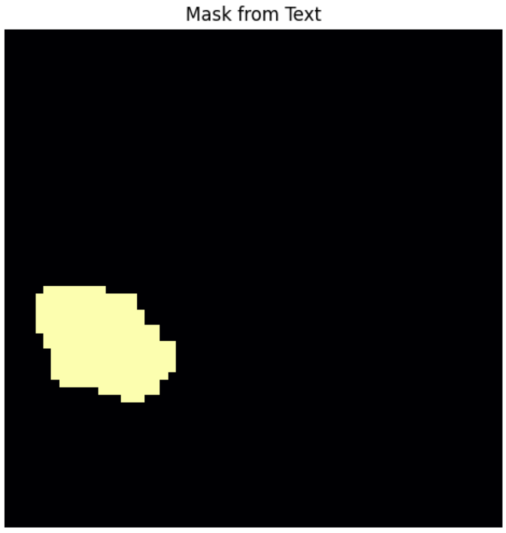


Figure 1.44. *Mask generated by Text-Prompted Pipeline*

1.13.5 Anomaly - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.781	0.776	0.811
AUPRC	0.761	0.707	0.831	0.698
IoU	0.606	0.527	0.676	0.494
Precision	0.747	0.615	0.941	0.544
Recall	0.762	0.787	0.705	0.844
F1	0.755	0.691	0.806	0.662
Specificity	0.987	0.974	0.998	0.963
Hausdorff	5.919	6.948	6.199	6.654
ASSD	0.450	0.685	0.528	0.610

Table 1.5. *Evaluation Metrics for the "Anomaly - 3 - 3" Configuration*

Figure 1.45. *Counterfactual generation*

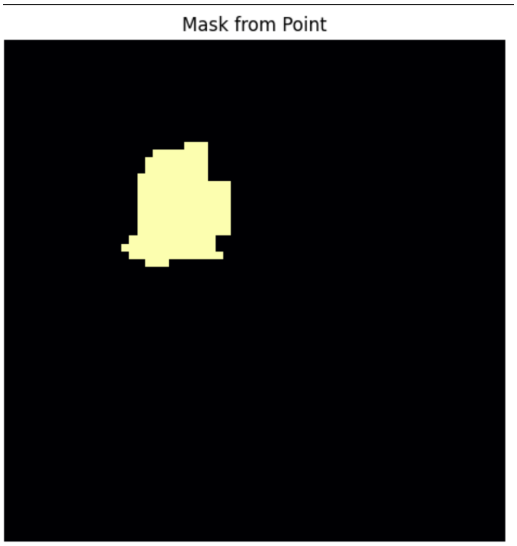


Figure 1.46. *Mask generated by Point-Prompted Pipeline*

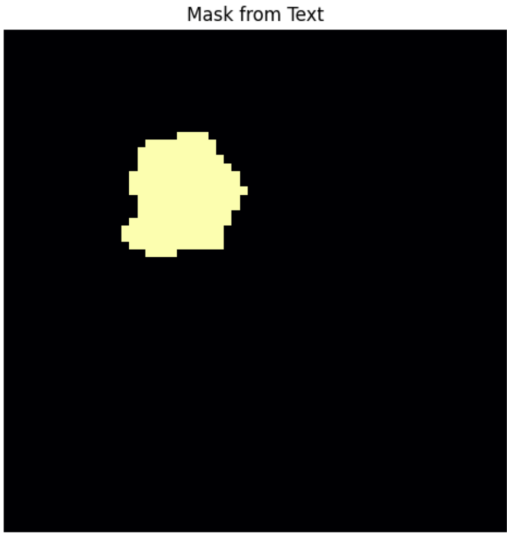


Figure 1.47. *Mask generated by Text-Prompted Pipeline*

1.13.6 Μικρότερο U-Net

Πειραματιζόμαστε με μια μικρότερη αρχιτεκτονική για το $U - Net$ του μοντέλου διάχυσής μας. Συγκεκριμένα, ο κωδικοποιητής μας επεξεργάζεται τις εικόνες μέσω τριών, αντί για τεσσάρων, σταδίων με σταθερό πλάτος καναλιών: 64, 64 και 64 κανάλια. Κάθε στάδιο τώρα περιλαμβάνει ένα, αντί για δύο, *residual block*. Παρακάτω παρέχουμε τα αποτελέσματα της διαμόρφωσης των υπερπαραμέτρων ¨$lesion - 3 - 3$¨.

	Point	**Text**	**Intersection**	**Union**
Dice	0.757	0.778	0.758	0.777
AUPRC	0.778	0.776	0.807	0.766
IoU	0.611	0.624	0.626	0.611
Precision	0.855	0.787	0.944	0.736
Recall	0.681	0.751	0.650	0.783
F1	0.759	0.769	0.770	0.759
Specificity	0.993	0.988	0.998	0.982
Hausdorff	6.838	7.489	7.205	6.971
ASSD	0.414	1.325	0.420	1.535

Table 1.6. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration of the smaller U-Net*

1.13.7 Αφαιρώντας το Μοντέλο Διάχυσης

Σε αυτή την ενότητα, παρουσιάζουμε τα αποτελέσματα που προέκυψαν από την άμεση εφαρμογή των αγωγών τμηματοποίησης με σημειακές και κειμενικές υποδείξεις στις αρχικές εικόνες MPI από το σύνολο δεδομένων Medical Segmentation Decathlon Task 1 (Brain Tumor) , χωρίς τη δημιουργία αντιπραγματικών εικόνων.

	Point	**Text**	**Intersection**	**Union**
Dice	0.337	0.164	0.303	0.218
AUPRC	0.195	0.056	0.125	0.124
IoU	0.114	0.022	0.083	0.052
Precision	0.145	0.049	0.178	0.072
Recall	0.353	0.285	0.364	0.324
F1	0.315	0.223	0.444	0.356
Specificity	0.514	0.422	0.463	0.355
Hausdorff	28.436	39.323	30.147	0.347
ASSD	4.147	7.348	4.736	6.548

Table 1.7. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration without the usage of the Counterfactual*

1.14 Περιπτώσεις Αποτυχίας

Παρά τη συνολική αξιοπιστία του πλαισίου μας, υπάρχουν περιπτώσεις όπου το σύστημα απέτυχε να τμηματοποιήσει σωστά τις βλάβες στις εικόνες *MRI*. Αυτές οι περιπτώσεις αποτυχίας προκύπτουν κυρίως λόγω της παρουσίας πολύ μικρών και δυσδιάκριτων βλαβών, οι οποίες αποτελούν σημαντικές προκλήσεις για την ακριβή ανίχνευση και τμηματοποίηση.

Οι κύριοι λόγοι αποτυχίας της διαδικασίας τμηματοποίησης περιλαμβάνουν:

- **Πολύ μικρές βλάβες:** Οι πολύ μικρές βλάβες συχνά είναι δύσκολο να εντοπιστούν και να τμηματοποιηθούν με ακρίβεια. Το μικρό τους μέγεθος οδηγεί σε ανεπαρκή αντίθεση και προεξοχή στις εικόνες, καθιστώντας δύσκολο για το μοντέλο να τις διαφοροποιήσει από τον περιβάλλοντα ιστό.
- **Χαμηλή ποιότητα των αρχικών εικόνων** ***MRI*****:** Σε ορισμένες περιπτώσεις, οι αρχικές εικόνες *MRI* μπορεί να είναι κακής ποιότητας, με χαμηλή ανάλυση ή υψηλά επίπεδα θορύβου. Αυτό μπορεί να εμποδίσει την ικανότητα του μοντέλου να εντοπίσει και να τμηματοποιήσει σωστά τις βλάβες.

Παρακάτω παρατίθενται παραδείγματα εικόνων όπου το μοντέλο μας απέτυχε να τμηματοποιήσει σωστά τις βλάβες. Κάθε σύνολο εικόνων περιλαμβάνει την αρχική MPI, την αρχική μάσκα εικόνας, την λανθάνουσα εικόνα, την ανακατασκευασμένη εικόνα και τον χάρτη ανωμαλιών. Αυτά τα παραδείγματα καταδεικνύουν τις προκλήσεις που αντιμετώπισε το μοντέλο μας στην ακριβή ανίχνευση πολύ μικρών και δυσδιάκριτων βλαβών.

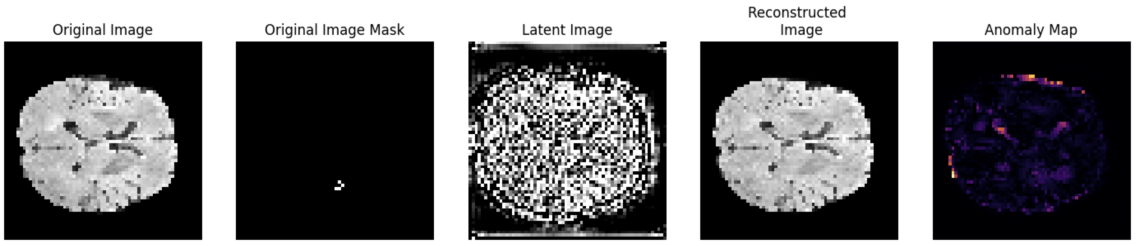


Figure 1.48. *Counterfactual generation*

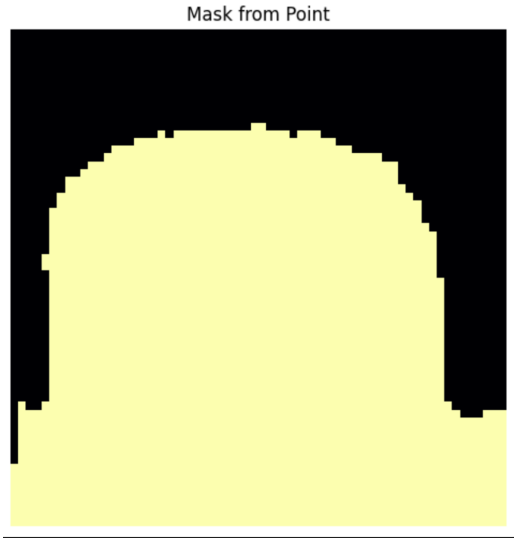


Figure 1.49. *Mask generated by Point-Prompted Pipeline*

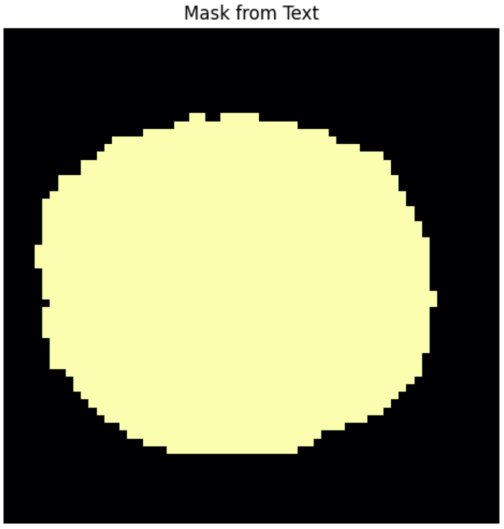


Figure 1.50. *Mask generated by Text-Prompted Pipeline*

1.15 Ερμηνεία Αποτελεσμάτων και Ανάλυση

Σε αυτή την ενότητα, παρουσιάζουμε μια ολοκληρωμένη ανάλυση των πειραματικών αποτελεσμάτων που προέκυψαν από τη χρήση του πλαισίου τμηματοποίησης μας στην εργασία του Εγκεφαλικού Όγκου από την πρόκληση Medical Segmentation Decathlon. Η ανάλυση βασίζεται σε διάφορες διαμορφώσεις υπερπαραμέτρων, όπως διαφορετικές κειμενικές προτροπές, τον αριθμό των βημάτων προεπεξεργασίας, και την εφαρμογή του συνδυασμού SAM-GroundingDINO.

Τα πειράματά μας έδειξαν ότι οι κειμενικές προτροπές ''*lesion*'' και ''*tumour*'' ήταν εξίσου αποτελεσματικές στην καθοδήγηση της διαδικασίας τμηματοποίησης. Αυτές οι προτροπές παρείχαν σταθερά υψηλές επιδόσεις σε όλες τις μετρικές αξιολόγησης. Η προτροπή ''*anomaly*'', αν και επίσης αποτελεσματική, έφερε ελαφρώς χαμηλότερα αποτελέσματα σε σύγκριση με τις ''*lesion*'' και ''*tumor*''. Αυτό υποδεικνύει ότι οι προτροπές που σχετίζονται άμεσα με τη φύση του στόχου (δηλαδή, βλάβη ή όγκος) είναι πιο αποτελεσματικές στη βελτίωση της ακρίβειας της τμηματοποίησης.

Παρατηρήσαμε μια σαφή τάση όπου η αύξηση του αριθμού των βημάτων προεπεξεργασίας και στις δύο μεθόδους τμηματοποίησης (με προτροπή σημείου και κειμένου) οδήγησε σε βελτιωμένα αποτελέσματα. Αυτή η τάση διατηρήθηκε έως ένα συγκεκριμένο σημείο· συγκεκριμένα, η ποιότητα της τμηματοποίησης βελτιώθηκε όσο ο αριθμός των βημάτων προεπεξεργασίας αυξανόταν μέχρι τα τρία. Ωστόσο, όταν ο αριθμός των βημάτων έφτασε τα τέσσερα, η ποιότητα των αποτελεσμάτων άρχισε να μειώνεται, λόγω της υπερβολικής απώλειας πληροφοριών από τη συνεχή πολλαπλασιαστική διαδικασία, η οποία δυσκολεύει την ακρίβεια της τμηματοποίησης των βλαβών. Αντίθετα, η διαμόρφωση με μηδενικά βήματα προεπεξεργασίας (0-0) έδωσε τα χειρότερα αποτελέσματα, υπογραμμίζοντας τη σημασία της κατάλληλης προεπεξεργασίας για τη βελτίωση της απόδοσης τμηματοποίησης.

Η μεθοδολογία μας έδειξε ανθεκτικότητα σε διάφορες διαμορφώσεις, διατηρώντας υψηλή απόδοση ανεξάρτητα από την κειμενική προτροπή που χρησιμοποιήθηκε, εφόσον σχετίζεται με την κατηγορία ¨βλάβη¨ ή ¨όγκος¨. Οι μέθοδοι τμηματοποίησης με προτροπή σημείου και κειμένου παρήγαγαν συγκρίσιμα αποτελέσματα, υποδεικνύοντας ότι και οι δύο μπορούν να χρησιμοποιηθούν αποτελεσματικά στην πράξη. Ομοίως, οι μέθοδοι διασταύρωσης και ένωσης, που συνδυάζουν τα αποτελέσματα από τις δύο μεθόδους, έδειξαν επίσης παρόμοια ποιότητα, επιβεβαιώνοντας περαιτέρω την ανθεκτικότητα της προσέγγισής μας.

Στο πλαίσιο της αφαιρετικής μελέτης, πειραματιστήκαμε με μια μικρότερη διαμόρφωση του μοντέλου διάχυσης, η οποία έφερε ελαφρώς χειρότερη απόδοση σε σχέση με το μεγαλύτερο μοντέλο, αλλά εξακολουθούσε να παράγει καλά αποτελέσματα συνολικά. Αυτό υποδεικνύει ότι, ενώ το μέγεθος του μοντέλου διάχυσης επηρεάζει την απόδοση, το πλαίσιο παραμένει αποτελεσματικό ακόμη και με μικρότερα μοντέλα.

Επιπλέον, αξιολογήσαμε την απόδοση του συνδυασμού SAM-GroundingDINO όταν εφαρ-

μόστηκε απευθείας στις αρχικές εικόνες *MRI*, χωρίς τη δημιουργία αντιπαραδειγματικών εικόνων. Σε αυτό το σενάριο, παρατηρήσαμε σημαντικά χειρότερα αποτελέσματα, αναδεικνύοντας τον κρίσιμο ρόλο του μοντέλου διάχυσης και της διαδικασίας δημιουργίας αντιπαραδειγμάτων στη βελτίωση της ακρίβειας της τμηματοποίησης.

1.16 Σύγκριση με Υπάρχουσες Εργασίες

Σε αυτό το κεφάλαιο, παρουσιάζουμε μια συγκριτική ανάλυση των αποτελεσμάτων που προέκυψαν από το πλαίσιο μας σε σχέση με άλλες μελέτες που έχουν χρησιμοποιήσει μοντέλα διάχυσης για την τμηματοποίηση και εντοπισμό εγκεφαλικών όγκων. Παρόλο που η σύγκριση αυτή στοχεύει να τοποθετήσει την εργασία μας στο πλαίσιο της υπάρχουσας βιβλιογραφίας, υπάρχουν αρκετές προκλήσεις που εμποδίζουν μια άμεση, ένα προς ένα σύγκριση.

Μία από τις κύριες δυσκολίες είναι η έλλειψη δημόσια διαθέσιμων *benchmarks* από άλλες μελέτες. Πολλές από αυτές δεν αναφέρουν συγκεκριμένες λεπτομέρειες, όπως το υποσύνολο του συνόλου δεδομένων *BraTS* που χρησιμοποίησαν ή την ακριβή αρχιτεκτονική των μοντέλων, συμπεριλαμβανομένων των διαφοροποιήσεων στη διαμόρφωση του $U-Net$. Αυτές οι διαφορές στα δεδομένα και τον σχεδιασμό των μοντέλων μπορούν να επηρεάσουν σημαντικά την απόδοση της τμηματοποίησης, καθιστώντας δύσκολη την άμεση σύγκριση όλων των μετρικών αξιολόγησης.

Επιπλέον, τα πρωτόκολλα εκπαίδευσης, οι υπερπαράμετροι και τα βήματα προεπεξεργασίας συχνά δεν αναφέρονται με συνέπεια στις δημοσιεύσεις, προσθέτοντας ένα επιπλέον επίπεδο πολυπλοκότητας στη σύγκριση. Για παράδειγμα, διαφορές στον αριθμό των βημάτων προεπεξεργασίας, τον τύπο των κειμενικών προτροπών και την ενσωμάτωση μηχανισμών προσοχής στις αρχιτεκτονικές $U-Net$ είναι κρίσιμοι παράγοντες που επηρεάζουν τα αποτελέσματα, αλλά δεν αναφέρονται πάντα πλήρως στα σχετικά έργα.

Παρά αυτές τις προκλήσεις, προσπαθούμε να παρέχουμε μια υψηλού επιπέδου σύγκριση των συνολικών τάσεων απόδοσης, εστιάζοντας στις κοινώς αναφερόμενες μετρικές όπως ο δείκτης *Dice*, το *IoU*, η ακρίβεια και η ανάκληση, κ.α.. Αυτή η ποιοτική σύγκριση θα αναδείξει τα πλεονεκτήματα του πλαισίου μας και τη σχέση του με ή τη βελτίωσή του έναντι υφιστάμενων προσεγγίσεων στην τμηματοποίηση εγκεφαλικών όγκων μέσω μοντέλων διάχυσης.

Στις ακόλουθες παραγράφου, παρουσιάζουμε την απόδοση του πλαισίου μας παράλληλα με τα αποτελέσματα που αναφέρονται σε τρεις σημαντικές παρεμφερείς δημοσιεύσεις, λαμβάνοντας προσεκτικά υπόψη τους προαναφερθέντες περιορισμούς. Παρόλο που μια άμεση σύγκριση με *benchmarks* δεν είναι δυνατή, αυτή η ανάλυση παρέχει πολύτιμες πληροφορίες για τον ευρύτερο αντίκτυπο των μοντέλων διάχυσης στην τμηματοποίηση εγκεφαλικών όγκων.

Wolleb et al.: Η μέθοδος των Wolleb et al. [50] χρησιμοποιεί *DDPMs* για τη δημιουργία αντιθετικών εικόνων, όπου μόνο οι παθολογικές περιοχές τροποποιούνται, δημιουργώντας έναν χάρτη ανωμαλιών για τμηματοποίηση. Σε σύγκριση με τη δική μας μεθοδολογία, η προσέγγισή τους χρησιμοποιεί λιγότερους μηχανισμούς προσοχής και βασίζεται στην καθοδηγούμενη αποθορυβοποίηση με ταξινομητή. Αντίθετα, το δικό μας μοντέλο ενσωματώνει προχωρημένη προεπεξεργασία, μηχανισμούς προσοχής σε πολλά επίπεδα και πιο προηγμένα εργαλεία τμηματοποίησης όπως το *SAM* και το *GroundingDINO*, επιτρέποντάς μας να διατηρούμε υψηλή ακρίβεια τμηματοποίησης με λιγότερα βήματα διάχυσης.

Sanchez et al. : Οι Sanchez et al. [40] χρησιμοποιούν ένα μοντέλο διάχυσης με έμμεση καθοδήγηση και μηχανισμούς προσοχής για τη δημιουργία αντιθετικών εικόνων με σκοπό τον εντοπισμό βλαβών. Αν και και οι δύο μέθοδοι ενσωματώνουν $U-Net$ αρχιτεκτονικές με προσοχή, η δική μας προσέγγιση ξεχωρίζει χρησιμοποιώντας διασταυρούμενη προσοχή και προηγμένα μοντέλα τμηματοποίησης όπως το *SAM* και το *GroundingDINO* μετά την προεπεξεργασία. Ενώ οι Sanchez et al. χρησιμοποιούν δυναμική κανονικοποίηση κατά τη φάση εξαγωγής για να διατηρήσουν την ποιότητα ανασυγκρότησης, η δικιά μας μεθοδολογία χρησιμοποιεί μια διαδικασία πολλαπλασιασμού (προεπεξεργασία) για να ενισχύσει την ορατότητα των βλαβών, οδηγώντας σε πιο ακριβή αποτελέσματα τμηματοποίησης.

Fontanella et al.: Η μέθοδος των Fontanella et al. [11] , "$Dif-fuse$", συνδυάζει ένα *DDPM* με χάρτες επισημάνσεων (saliency maps) για τον εντοπισμό ανωμαλιών στις ιατρικές εικόνες μέσω της δημιουργίας αντιθετικών εικόνων. Σε σύγκριση με τη δική μας προσέγγιση, η "$Dif-fuse$" προσθέτει μηχανισμούς χάρτη επισημάνσεων για τη βελτίωση του εντοπισμού των παθολογικών περιοχών, ενώ η δικιά μας μεθοδολογία χρησιμοποιεί το *SAM* και το *GroundingDINO* για τμηματοποίηση, με έμφαση στην προεπεξεργασία της εικόνας διαφοράς. Παρά το γεγονός ότι η "$Dif-fuse$" επιτυγχάνει αξιόλογα αποτελέσματα, η δικιά μας μεθοδολογία αποδίδει καλύτερα σε ορισμένες ρυθμίσεις, δείχνοντας την αποτελεσματικότητα της προσέγγισής μας στη δημιουργία αντιθετικών εικόνων και τη χρήση προηγμένων τεχνικών τμηματοποίησης.

1.17 Επίλογος

Στην παρούσα διπλωματική εργασία, αναπτύχθηκε και αξιολογήθηκε ένα αξιόπιστο πλαίσιο για την τμηματοποίηση και τον εντοπισμό όγκων στον εγκέφαλο μέσω *MRI* εικόνων. Η προσέγγισή μας ενσωματώνει ένα μοντέλο διάχυσης για τη δημιουργία υγιών αντιπραγματικών εικόνων, ακολουθούμενο από τμηματοποίηση χρησιμοποιώντας τα μοντέλα *SAM* και *GroundingDINO*. Εξετάσαμε διάφορες διαμορφώσεις υπερπαραμέτρων και πραγματοποιήσαμε μια μελέτη αφαίρεσης για να κατανοήσουμε τη σημασία κάθε στοιχείου της διαδικασίας μας.

Τα βασικά ευρήματα περιλαμβάνουν:

- **Αποτελεσματικότητα των Προτροπών Κειμένου:** Τα πειράματα έδειξαν ότι οι προτροπές όπως "*lesion*" και "*tumor*" είναι εξίσου αποτελεσματικές, παρέχοντας υψηλή απόδοση σε όλες τις μετρικές. Η προτροπή "*anomaly*" είχε ελαφρώς χαμηλότερη απόδοση, επισημαίνοντας τη σημασία της επιλογής προτροπών που συνδέονται άμεσα με τη φύση των βλαβών.

- **Επίδραση των Βημάτων Προεπεξεργασίας:** Η αύξηση των βημάτων προεπεξεργασίας βελτίωσε την ποιότητα της τμηματοποίησης, με βέλτιστα αποτελέσματα στα τρία βήματα. Πέρα από αυτό το σημείο, η απόδοση άρχισε να μειώνεται λόγω απώλειας πληροφοριών. Η απουσία προεπεξεργασίας παρήγαγε τα χειρότερα αποτελέσματα, υπογραμμίζοντας τη σημασία της σωστής προεπεξεργασίας.

- **Ανθεκτικότητα της Μεθοδολογίας:** Η μεθοδολογία μας απέδειξε την ανθεκτικότητά του σε διαφορετικές διαμορφώσεις, διατηρώντας υψηλή απόδοση ανεξάρτητα από την προτροπή κειμένου, εφόσον αυτή ήταν σχετική με τις βλάβες. Οι προτροπές σημείων και κειμένου απέδωσαν συγκρίσιμα αποτελέσματα, όπως και οι μέθοδοι διασταύρωσης και ένωσης των αποτελεσμάτων.

- **Μελέτη Αφαίρεσης:** Επιβεβαίωσε τον κρίσιμο ρόλο του μοντέλου διάχυσης και της δημιουργίας αντιθετικών εικόνων για την επίτευξη υψηλής ακρίβειας τμηματοποίησης. Η αφαίρεση του μοντέλου διάχυσης και η απευθείας εφαρμογή του *SAM – GroundingDINO* στις αρχικές *MRI* εικόνες οδήγησε σε σημαντικά χειρότερα αποτελέσματα.

Συνολικά, η εργασία αυτή συμβάλλει στον τομέα της τμηματοποίησης ιατρικών εικόνων, δείχνοντας την αποτελεσματικότητα της ενσωμάτωσης γεννητικών μοντέλων με προηγμένες τεχνικές τμηματοποίησης. Προτείνεται ως ένα αξιόπιστο εργαλείο για την υποστήριξη των ιατρών στη διάγνωση και τον σχεδιασμό θεραπειών.
Μελλοντικές κατευθύνσεις θα μπορούσαν να περιλαμβάνουν:

- **Βελτιωμένες Τεχνικές Προεπεξεργασίας:** Έρευνα για εναλλακτικές μεθόδους προεπεξεργασίας που θα διατηρούν περισσότερες πληροφορίες.

- **Εκτεταμένη Ρύθμιση Υπερπαραμέτρων:** Περαιτέρω εξερεύνηση υπερπαραμέτρων για βελτιστοποίηση της απόδοσης.

- **Εφαρμογή σε Πραγματικό Χρόνο και Κλινικές Δοκιμές:** Δοκιμές της μεθοδολογίας σε κλινικά περιβάλλοντα για την αξιολόγηση της πρακτικής του χρησιμότητας.

Συμπερασματικά, το πλαίσιο μας αποτελεί σημαντική πρόοδο στην αυτόματη τμηματοποίηση όγκων εγκεφάλου σε *MRI* εικόνες, προσφέροντας ένα ευέλικτο και αποτελεσματικό εργαλείο για την υποβοήθηση της ιατρικής διάγνωσης.

Chapter 2

Introduction

The application of artificial intelligence (AI) in medicine [10, 33] has revolutionized healthcare by improving diagnostic accuracy, personalizing treatment plans, and enhancing patient outcomes. In the field of medical imaging, AI models have shown significant potential in detecting and segmenting anomalies such as tumors and lesions, which are critical for early diagnosis and treatment planning. Magnetic Resonance Imaging (MRI) is a vital imaging modality for brain tumors due to its high resolution and contrast. However, the manual analysis of MRI scans is time-consuming and prone to human error. Therefore, automated systems that can accurately locate and segment brain tumors are of paramount importance.

One attempt at such automation is through generative models. Generative models have gained much popularity in recent years due to their ability to generate high-quality images, synthesize music and human-like speech. Among the various generative models that have been proposed over the years - GANs [12], VAEs [21], autoregression transformers [47], flow models [9] - diffusion models have been consistently producing the best results among various fields. They are a class of probabilistic models that gradually inject noise into a given input and then learn to reverse that process with the help of a neural network. They then use said trained neural network to generate new data sampled from the learned initial distribution. The theory behind the class of diffusion models is mathematically oriented and it took many independent contributions to establish their dominance in today's world.

Building on the strengths of diffusion models, we propose a novel framework that leverages diffusion models to produce counterfactual brain MRI images without lesions. These counterfactual images serve as a reference to highlight the presence and extent of lesions in the original scans. For the segmentation task, we employ the Segment Anything Model (SAM) [22] and the Grounding DINO model [25]. SAM is renowned for its ability to segment objects from any image with minimal input, making it highly adaptable to diverse segmentation tasks. Grounding DINO complements SAM by using grounding techniques that enhance the model's capability to localize and identify objects based on point as well as textual prompts.

To evaluate the performance of our framework, we utilize several metrics, including the Dice score, Area Under the Precision-Recall Curve (AUPRC), precision, recall, specificity, F1 score, Hausdorff distance [18], and Average Symmetric Surface Distance (ASSD) [48]. Our framework achieves high performance in all of these metrics outperforming not only the baseline models but numerous similar frameworks in the context of medical segmentation. These metrics provide a comprehensive assessment of the model's accuracy, reliability, and robustness in detecting and segmenting tumors. Additionally, our framework generates multiple output images that highlight different aspects of the segmentation, which can be invaluable for radiologists in making more informed decisions.

It is important to emphasize that this framework is not intended to replace radiologists or doctors. Instead, it is designed to be used as an assistive tool, augmenting the expertise of medical professionals by providing them with detailed, accurate, and easily interpretable imaging results. By integrating advanced AI techniques with clinical practice, we aim to enhance the diagnostic process and support better patient outcomes.

Chapter 3

Fundamental ideas behind Diffusion Models

3.1 Denoising Diffusion Probabilistic Models

3.1.1 Inspiration from Thermodynamics

The work of Sohl-Dicksteain et al. [41] pioneered the field of diffusion models, serving as the seminal inspiration for subsequent research and publications in the area.

In this paper, the authors introduce the main idea behind the class of diffusion models as we know it today. A forward process is defined where the input is gradually corrupted with noise using a discrete Markov chain as well as a backward process where a a neural network is trained to gradually restore the initial input by removing the noise. Moreover, the authors choose the log-likelihood as the training objective and provide a more tractable lower bound to that amount in the form of a sum of Kullback-Leibler divergence and entropies. Finally, via the use of Monte Carlo sampling, it is shown that one can *exactly* sample from the initial data probability distribution.

3.1.2 Extending the Ideas

Although Sohl-Dickstein et al. [41] set the general framework for the diffusion process, they left out many important details regarding the implementation of said process. This theoretical and practical gap was filled by another seminal paper in the field; the work of Ho et al. [16]. In this paper, the authors introduced numerous new concepts expanding on previous work in the field. Issues such as variance schedule, a formal definition of the DDPM sampling procedure, and a more tractable definition of the cost were discussed.

In this context, the *forward diffusion process* is described by the following Markovian process:

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t;\ \sqrt{1-b_t}\cdot x_{t-1}, b_t \cdot I), \forall t \in 1, ..., T \tag{3.1}$$

where **T** is the number of diffusion steps, $b1, ..., b_T \in [0, 1)$ is the variance schedule , **I** is

the identity matrix that has the same dimensions as the input image x_0, and $\mathcal{N}(x; \mu, \sigma)$ represents the normal distribution of mean μ and covariance σ that produces x.

In the *reverse diffusion process*, we start from a sample $x_T \sim \mathcal{N}(0, I)$, and generate new samples from $p(x_0)$ by following the reverse steps:

$$p(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu(x_t, t), \Sigma(x_t, t)) \tag{3.2}$$

In order to approximate the reverse steps, a neural network is trained that receives a noise input and learns to predict the mean and covariance.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)) \tag{3.3}$$

Given that both of the forward and the reverse processes are modeled as Markov chains, the joint probability distribution in each case can be written as follows:

$$q(x_1, x_2, \ldots, x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}) \tag{3.4}$$

$$p_\theta(x_1, x_2, \ldots, x_T) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t) \tag{3.5}$$

Equivalently, regarding the actual perturbed data, the discrete Markov chain that describes the forward diffusion process can be expressed as follows:

$$x_t = \sqrt{1 - b_t} x_{t-1} + \sqrt{b_t} \epsilon \tag{3.6}$$

where:

$$x_0 \sim p(x_0)$$

$$\epsilon \sim \mathcal{N}(0, I)$$

There is also a transformation that can be used in order to obtain a sample x_t directly from the initial sample x_0.

$$x_t = \sqrt{\bar{a}_t} x_0 + \sqrt{1 - \bar{a}_t} \epsilon \tag{3.7}$$

where:

$$a_t = 1 - b_t$$

$$\bar{a}_t = \prod_{s=0}^{t} a_s$$

The reverse process is described by the following equation:

$$x_{t-1} = \mu_\theta(x_t, t) + \sqrt{\Sigma_\theta(x_t, t)} z \tag{3.8}$$

where:

$$x_T \sim \mathcal{N}(0, I)$$

$$z \sim \mathcal{N}(0, I)$$

It is noted that ϵ are standard spherical Gaussian noise terms independent of the *past* values of x and z are standard spherical Gaussian noise terms independent of the *future* values of x. Both of these terms depend on time t, i.e. $\epsilon = \epsilon_t$, $z = z_t$, but we avoid the subscripts in favor of simplicity.

Notice how during the reverse process, additional noise is added to the predicted sample at every time step. This is to ensure that the generation process does not get stuck in modes of the distribution.

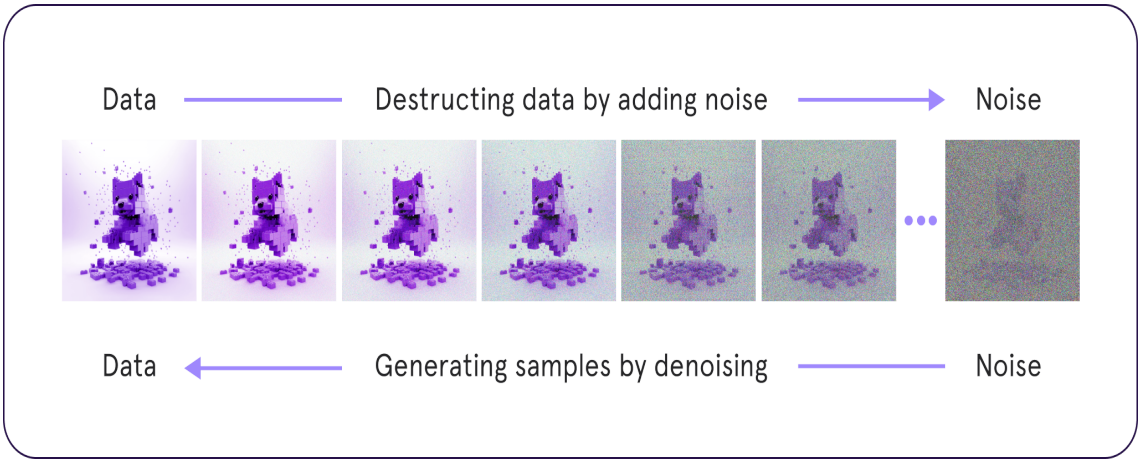


Figure 3.1. *The Forward and Backward Diffusion Process.*
[1]

For our model to produce accurate and high-quality results, we want the reverse joint probability distribution, $p_\partial(x_0, x_1, ..., x_T) = p(x_T) \prod_{t=1}^{T} p_\partial(x_{t-1}|x_t)$ to closely approximate the joint forward probability distribution, $q(x_0, x_1, ..., x_T) = q(x_0) \prod_{t=1}^{T} q(x_t|x_{t-1})$. For this reason, the minimization of the Kullback-Leibler (KL) divergence between these two distributions is used as the objective for training the neural network:

$$KL(q(x_0, x_1, ..., x_T)||p_\partial(x_0, x_1, ..., x_T)) \tag{3.9}$$

$$= -\mathbb{E}_{q(x_0, x_1, ..., x_T)} [log p_\partial(x_0, x_1, ..., x_T)] + const \tag{3.10}$$

$$= \mathbb{E}_{q(x_0, x_1, ..., x_T)} \left[-log p(x_T) - \Sigma_{t=1}^{T} log \frac{p_\partial(x_{t-1}|x_t)}{q(x_t|x_{t-1})} \right] + const \tag{3.11}$$

$$= L_{VLB} + const \tag{3.12}$$

$$\geq \mathbb{E}[-log p_\partial(x_0)] + const \tag{3.13}$$

The term L_{VLB} represents the Variational Lower Bound of the log-likelihood of the data x_0. This is a commonly used objective in neural network training since it closely approximates the log-likelihood of the data and is computationally tractable. The term *const* represents a constant that does not depend on the network's parameters, θ.

After some simplification, the authors arrive at the following form for the L_{VLB}:

$$\mathbb{E}_{q(x_0,x_1,\ldots,x_T)}\left[KL(q(x_T|x_0)||p(x_T)) + \Sigma_{t=1}^{T} KL(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) - logp_\theta(x_0|x_1)\right] \tag{3.14}$$

This quantity can be deconstructed in the following manner:

$$L_0 = -logp_\theta(x_0|x_1) \tag{3.15}$$

$$L_T = KL(q(x_T|x_0)||p(x_T)) \tag{3.16}$$

$$L_{t-1} = KL(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) \tag{3.17}$$

With this formalization, the L_{VLB} can be written as:

$$L_{VLB} = L_T + \Sigma_{t=1}^{T} L_{t-1} + L_0 \tag{3.18}$$

The authors proposed to ignore the terms L_T and L_0; the former because it does not depend on the network's parameters and the latter because they managed to produce better results in practice without it.

Thus, the loss function at time t becomes:

$$L_{VLB} = L_{t-1} = KL(q(x_{t-1}|x_t,x_0)||p_\theta(x_{t-1}|x_t)) \tag{3.19}$$

The distribution $q(x_{t-1}|x_t,x_0)$ is called the ***forward process posterior distribution***:

$$q(x_{t-1}|x_t,x_0) = \mathcal{N}(x_{t-1};\tilde{\mu}_t(x_t,x_0),\tilde{b}_t I) \tag{3.20}$$

where:

$$\tilde{\mu}_t(x_t,x_0) = \frac{\sqrt{\tilde{a}_{t-1}}b_t}{1-\tilde{a}_t}x_0 + \frac{\sqrt{a_t}(1-\tilde{a}_{t-1})}{1-\tilde{a}_t}x_t \tag{3.21}$$

$$\tilde{b}_t = \frac{1-\tilde{a}_{t-1}}{1-\tilde{a}_t}b_t \tag{3.22}$$

It can also be proven that minimizing the KL divergence between two Gaussian distributions when the variance is fixed, is equivalent to minimizing the distance between their means. In this case, the aforementioned training objective can be further simplified to the following:

$$L_{t-1} = \mathbb{E}_{q(x_0,x_1,\dots,x_T)}\left[\frac{1}{2\sigma_t^2}\|\tilde{\mu}_t(x_t,x_0) - \mu_\theta(x_t,t)\|_2^2\right] + const, \tag{3.23}$$

where $\sigma_t^2 = b_t$.

Finally, we can express x_0 as a function of x_t from equation (5) and acquire a simplified version of $\tilde{\mu}_t$:

$$\tilde{\mu}_t(x_t,x_0) = \tilde{\mu}_t(x_t) = \frac{1}{\sqrt{a_t}}\left(x_t - \frac{b_t}{\sqrt{1-\tilde{a}_t}}\epsilon\right) \tag{3.24}$$

Now, since we want $\mu_\theta \approx \tilde{\mu}_t$ and x_t is already available to the input model, the simplified objective becomes:

$$L_{simple} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(x_t,t)\|_2^2\right] \tag{3.25}$$

Equivalently, using equation (7), the above objective can be written:

$$L_{simple} = \mathbb{E}_{t,x_0,\epsilon}\left[\|\epsilon - \epsilon_\theta(\sqrt{\tilde{a}_t}x_0 + \sqrt{1-\tilde{a}_t}\epsilon, t)\|_2^2\right] \tag{3.26}$$

The model now tries to predict the added noise at each step in the forward diffusion process. With the variance matrix fixed and the noise ϵ_θ of each time step available, we can sample from the initial distribution using equation (6):

$$x_{t-1} = \frac{1}{\sqrt{a_t}}\left[x_t - \frac{1-a_t}{\sqrt{1-\tilde{a}_t}}\epsilon_\theta(x_t,t)\right] + \sigma_t z \tag{3.27}$$

3.1.3 Optimization and State-of-the-Art Results

Although DDPMs [41], [16] managed to produce high-fidelity generated samples, they did poorly on achieving competitive log-likelihoods, a popular metric in generative modeling whose optimization is believed to be correlated with the ability of the model to capture all of the modes of the data distribution.

This was highlighted by Dhariwal et al. in their paper [29] where they proposed several improvements to the DDPM to optimize the log-likelihood metric while maintaining high-quality samples, conclusively showing that DDPMs are capable of producing state-

of-the-art results even to datasets of high dimensions such as ImageNet.

In more detail, Dhariwal et al. proposed three modifications to the DDPM algorithm which they then showed optimize the log-likelihood metric.

Firstly, they suggested learning the covariance matrix $\Sigma_\partial(x_t, t)$ instead of fixing it to a constant value like Ho et al. had previously proposed [16]. Noticing that $\Sigma_\partial(x_t, t) = \sigma_t^2 I$ produced similar results for $\sigma_t^2 = b_t$ and $\sigma_t^2 = \tilde{b}_t$, the authors expressed the covariance matrix as follows:

$$\Sigma_\partial(x_t, t) = exp(vlogb_t + (1 - v)log\tilde{b}_t) \tag{3.28}$$

Based on this, they also modified the training objective since the one from equation (25) did not depend on the covariance matrix. For the new hybrid learning objective, they combined equations (25) and (18):

$$L_{hybrid} = L_{simple} + \lambda L_{VLB} \tag{3.29}$$

They noted that $\lambda = 0.001$ worked best in practice to prevent the L_{VLB} term from overwhelming L_{simple}.

The second modification that was proposed was using a cosine noise schedule, instead of a linear one. With a cosine noise schedule, the information is destroyed more slowly in the forward process making the transition from data to noise much smoother.

$$\tilde{a}_t = \frac{f(t)}{f(0)}, f(t) = cos(\frac{t/T + s}{1 + s} \cdot \frac{\pi}{2}) \tag{3.30}$$

The authors finally proposed *importance sampling* :

$$L_{VLB} = \mathbb{E}_{t \sim p_t}\left[\frac{L_t}{p_t}\right] \tag{3.31}$$

where:

$$p_t \propto \sqrt{\mathbb{E}[L_t^2]} \text{ and } \sum p_t = 1$$

Importance sampling is a reformulation of the cost L_{VLB} to make its gradient less noisy allowing for direct optimization of that cost instead of the hybrid one. One practical consideration regarding importance sampling is that the values of $\mathbb{E}[L_t^2]$ are unknown and subject to change during training. For this reason, the authors propose maintaining a history of the previous 10 values for each loss term and updating this history dynamically during training.

3.1.4 Faster Sampling with Implicit Modeling

While DDPMs produce high-quality samples, surpassing other generative models across numerous metrics, the diffusion process requires many iterations to produce said samples. Compared to GANs, which only require a single pass through the network to produce a sample, DDPMs can be more than ***a thousand times slower*** in the generating process.

This problem was thouroughly addressed by Song et al. [42]. In this paper, they propose generalizing the forward diffusion process, which is Markovian in nature in DDPMs, to a non-Markovian process. With this generalization, they are able to show that faster sampling can be achieved in the reverse process by purposefully omitting some steps in the reverse chain. They call these generalized inference models *Denoising Diffusion Implicit Models (DDIMs)*.

The key observation was that the training objective in DDPMs (12) depends only on the marginal distributions $q_\sigma(x_t|x_0)$ and not on the joint distribution$q_\sigma(x_{1\dots T})|x_0$. Based on that, the authors propose alternative non-Markovian inference models with different joint distributions but the same marginals.

In this context, the transformation from a Markovian diffusion process to a more general non-Markovian inference process is achieved by conditioning the forward and reverse transition probability distributions to the initial sample:

$$q_\sigma(x_{1\dots T}|x_0) = q_\sigma(x_T|x_0)\prod_{t=2}^{T} q_\sigma(x_{t-1}|x_t, x_0) \tag{3.32}$$

where $q_\sigma(x_t|x_0) = \mathcal{N}(\sqrt{a_t}x_0, (1-a_t)I)$ for all $t > 1$ and:

$$q_\sigma(x_{t-1}|x_t, x_0) = \mathcal{N}\left(\sqrt{a_{t_1}}x_0 + \sqrt{1-a_{t-1}-\sigma_t^2}\cdot\frac{x_t - \sqrt{a_t}x_0}{\sqrt{1-a_t}}, \sigma_t^2 I\right) \tag{3.33}$$

Based on this, the forward non-Markovian process can be obtained by applying Baye's rule:

$$q_\sigma(x_t|x_{t-1}, x_0) = \frac{q_\sigma(x_{t-1}|x_t, x_0)q_\sigma(x_t|x_0)}{q_\sigma(x_{t-1}|x_0)} \tag{3.34}$$

The introduction of the parameter σ in equation (26) generalizes the reverse (and consequently, the forward) process by allowing us to control its stochasticity. Moreover, for the value $\sigma_t = \sqrt{(1-a_{t-1})/(1-a_t)}\sqrt{1-a_t/a_{t-1}}$ the forward process becomes Markovian and the model becomes the standard DDPM.

In order to define the trainable generative (reverse) process, a prediction about the initial sample, x_0 must be made since the transition probability distributions are conditioned, and thus dependent, on it.

The authors make the following prediction for x_0 given x_t:

$$\tilde{x_0}(t) = \frac{x_t - \sqrt{1 - a_t} \cdot \epsilon_\partial^{(t)}(x_t)}{\sqrt{a_t}} \tag{3.35}$$

The trainable process now becomes:

$$p_\partial^{(t)}(x_{t-1}|x_t) = q_\sigma(x_{t-1}|x_t, \tilde{x_0}(t)), t > 1 \tag{3.36}$$

For t=1: $p_\partial^{(t)}(x_0, x_1) = \mathcal{N}(\tilde{x_0}(1), \sigma_1^2 I)$.

Regarding the actual samples in the reverse process:

$$x_{t-1} = \sqrt{a_t - 1}\tilde{x_0}(t) + \sqrt{1 - a_{t-1} - \sigma_t^2} \cdot \epsilon_\partial^{(t)}(x_t) + \sigma_t \epsilon_t \tag{3.37}$$

When $\sigma_t = 0$, the process becomes deterministic in nature since the coefficient of the noise term, ϵ_t becomes zero. Song et al. then demonstrate that by choosing $\sigma_t = 0$ (DDIMs), the length of the sampling trajectory is decreased and higher computational efficiency is achieved with minimal sacrifice to the quality of the generated samples.

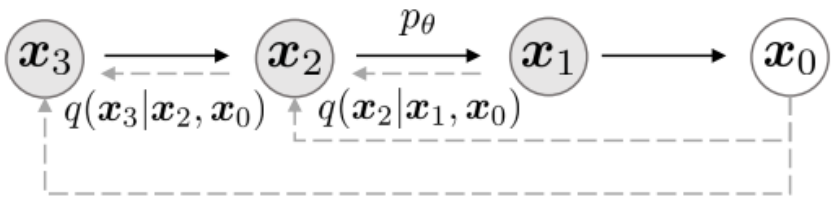


Figure 3.2. *The Non-Markovian Inference Model [42]*

3.2 Score Matching with Langevin Dynamics (SMLDs)

This is the second sub-category of diffusion models which focuses on a different formulation of the diffusion process. At the core of these models lies the (Stein) score function of a probability density $p(x)$ which is given by $\nabla_x logp(x)$. This quantity provides the directions according to which we move from a random sample x_0 towards a sample x_N in a region with high density. The algorithm that is used for this process is called Langevin sampling algorithm.

The theoretical foundation for this category of diffusion models was established in another seminal paper by Song et al. [43]. According to the Langevin sampling algorithm, we obtain the following iterative process:

$$x_t = x_{t-1} + \frac{\epsilon}{2}\nabla_x logp(x_{t-1}) + \sqrt{\epsilon}z_t, \tag{3.38}$$

where $z_t \sim \mathcal{N}(0, I)$, $\epsilon > 0$ and $x_0 \sim p(x_0)$ (prior distribution). In mathematics, this is known as a Langevin Markov chain Monte Carlo (MCMC).

This process allows sampling from a probability distribution $p(x)$ by using just its score function. Although equation (36) might seem strange at first, it is actually similar to equation (24) which we derived for the sampling process in DDPMs. Indeed, if we solve equation (36) for the term x_{t-1} we can directly compare it to equation (24) by the assumption that the quantity $\epsilon_\theta(x_t, t)$ corresponds to the gradient of the data density.

A direct approach would be to train a neural network in order to approximate the score function, i.e. $s_\theta(x) \approx \nabla_x logp(x)$. The logical objective to minimize is the following:

$$L_{sm} = \mathbb{E}_{x\sim p(x)}\|s_\theta(x) - \nabla_x logp(x)\|_2^2 \tag{3.39}$$

There are two problems with this approach, however. First, we do not have the score function in order to use it in the objective. But even if we did have the score function, the manifold hypothesis, which states that real-world high-dimensional data lie on low-dimensional manifolds embedded within the high-dimensional space, would raise some difficulties that are explained by Song et al. [43].

Song et al. [43] tackle both of these problems in their paper. Regarding the manifold hypothesis, they suggest perturbing the dataset with random Gaussian noise makes the data distribution more amenable to score-based generative modeling. This noise will help spread out the distribution making the manifold easier to work with.

This is achieved by first defining a geometric sequence of L noise levels $[\sigma_i]_{i=0}^L$ and then using this sequence to perturb the data:

$$x_t = x_{t-1} + \sqrt{\sigma_t^2 - \sigma_{t-1}^2}\epsilon_t, \epsilon_t \sim \mathcal{N}(0, I) \tag{3.40}$$

The reasoning is that instead of training the model $s_\partial(x)$ to predict the score function directly, we instead train a network $s_\partial(\hat{x}, \sigma_t)$ to estimate the scores of perturbed data distributions. Song at al. [43] called these neural networks a ***Noise Conditional Score Networks (NCSNs)***. The objective now becomes:

$$L_{dsm} = \frac{1}{L}\sum_{t=1}^{L} \lambda(\sigma_t)\mathbb{E}_{p(x)}\mathbb{E}_{\hat{x}\sim p_{\sigma_t}(\hat{x}|x)}\|s_\partial(\hat{x}, \sigma_t) - \nabla_x log p_\sigma(\hat{x})\|_2^2, \tag{3.41}$$

where $\lambda(\sigma_t)$ is a weighting function, $\sigma_1, \sigma_2, ..., \sigma_T$ is the sequence of Gaussian noise scales, $p_{\sigma_1} = p(x_0)$ and $\hat{x}$ are the final perturbed data.

These scores accept a close form for a Gaussian distribution. In particular:

$$\nabla_x log p(x) = -\frac{x}{\sigma^2}, p(x) = \mathcal{N}(0, \sigma^2) \tag{3.42}$$

For the perturbed data, we now acquire a conditional score function:

$$\nabla_x log p_\sigma(\hat{x}|x) = -\frac{\hat{x} - x}{\sigma^2} \tag{3.43}$$

Thus, the aforementioned objective will now utilize this conditional score function to train the neural network:

$$L_{dsm} = \frac{1}{L}\sum_{t=1}^{L} \lambda(\sigma_t)\mathbb{E}_{p(x)}\mathbb{E}_{\sigma x\sim p_{\sigma_t}(\hat{x}|x)}\|s_\partial(\hat{x}, \sigma_t) + \frac{\hat{x} - x}{\sigma_t^2}\|_2^2, \tag{3.44}$$

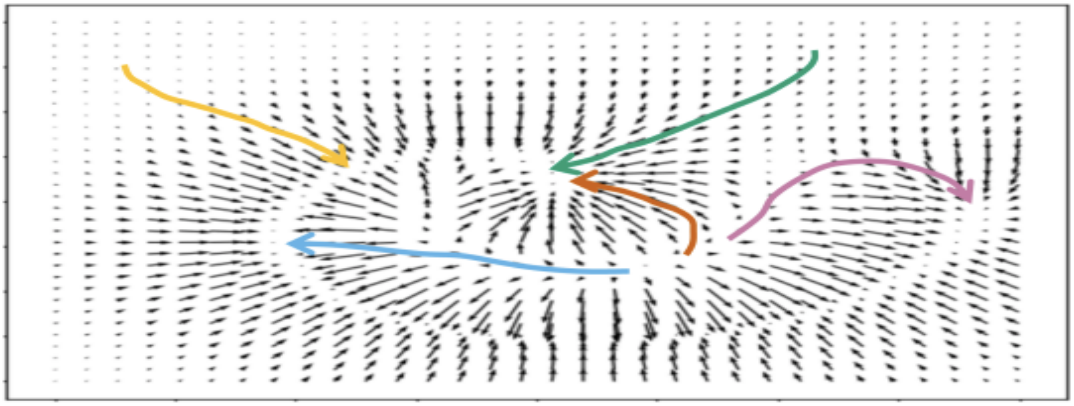


Figure 3.3. *Visualization of the trajectory by predicting score. A score is a direction for next timesteps. Samples are denoised in the direction at each position. Colors represent trajectories of different samples. [5]*

As for sampling, Song et al. [43] proposed a modified version of the Langevin sampling algorithm, ***Annealed Langevin Dynamics***.

The idea is to start Langevin sampling in a highly σ perturbed data distribution with a large time step α_i for a predetermined amount of steps T. Once this is done we do this again but for a slightly less σ perturbed data distribution with a slightly smaller time step α_i. We repeat this process for L noise scales.

Algorithm 1 Annealed Langevin dynamics.

Require: $\{\sigma_i\}_{i=1}^{L}, \epsilon, T$.

Initialize $\tilde{\mathbf{x}}_0$
for $i \leftarrow 1$ to L **do**
  $\alpha_i \leftarrow \epsilon \cdot \sigma_i^2/\sigma_L^2$ ▷ α_i is the step size.
  for $t \leftarrow 1$ to T **do**
    Draw $\mathbf{z}_t \sim \mathcal{N}(0, I)$
    $\tilde{\mathbf{x}}_t \leftarrow \tilde{\mathbf{x}}_{t-1} + \frac{\alpha_i}{2}\mathbf{s}_{\boldsymbol{\theta}}(\tilde{\mathbf{x}}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\,\mathbf{z}_t$
  end for
  $\tilde{\mathbf{x}}_0 \leftarrow \tilde{\mathbf{x}}_T$
end for
return $\tilde{\mathbf{x}}_T$

Figure 3.4. *The Annealed Langevin Dynamics Algorithm [43]*

This formulation of diffusion models produces high-quality results. However, in practice, DDPMs dominate the scene due to their simplicity, more intuitive algorithm, and lack of hyperparameters in the reverse process.

3.3 Stochastic Differential Equations (SDEs)

The third sub-category in the class of diffusion models constitutes a generalization of the previous two since here, the formulation of the diffusion process is continuous rather than discrete. In particular, the diffusion process is described as the solution to a stochastic differential equation (SDE). Song et al. formalized these ideas in their paper titled ***"Score-Based Generative Modeling Through Stochastic Differential Equations"*** [44].

In this context, the forward SDE has the following form:

$$d\mathbf{x} = \mathbf{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w} \tag{3.45}$$

This SDE is the standard Itô SDE where **f(x,**,t) is the drift coefficient, g(t) is the diffusion coefficient and d**w** is the Wiener process. For the standard Wiener process (Brownian motion):

$$d\mathbf{w} = \epsilon\sqrt{dt}, \tag{3.46}$$

where

$$\epsilon \sim \mathcal{N}(0, \mathbf{I})$$

The drift coefficient is designed such that it gradually nullifies the data x_0, while the diffusion coefficient controls how much Gaussian noise is added at each step.

Now, in order to generate data from the initial distribution, we should be able to reverse this process. This is exactly what Anderson showed in his paper [2].

The reverse-time SDE is given by:

$$d\mathbf{x} = [\mathbf{f}(\mathbf{x}, t) - g^2(t)\nabla_x log p_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}} \tag{3.47}$$

where $d\bar{\mathbf{w}}$ is a Wiener process that flows backwards in time.

Intuitively, the reverse-time differential equation shows that we can recover data from pure noise by removing the diffusion term that is responsible for the destruction of the data in the first place.

As mentioned earlier, stochastic differential equations constitute an attempt to unify DDPMs and SMLDs under a common theoretical umbrella. For this, we can transform the equations for the diffusion process of these models into their continuous counterparts in order to bring them in a form that resembles the Itô SDE.

In this context, the corresponding SDE for DDPMs is the following:

$$d\mathbf{x} = -\frac{1}{2}b(t)\mathbf{x}dt + \sqrt{b(t)}d\mathbf{w} \tag{3.48}$$

This SDE yields a process with a fixed variance of one when the initial distribution has unit variance. This is why it is known as a ***Variance Preserving (VP)*** SDE.

Similarly, the SDE for SMLDs is given by the following equation:

$$d\mathbf{x} = \sqrt{\frac{d[\sigma^2(t)]}{dt}}d\mathbf{w} \tag{3.49}$$

Contrary to the previous one, this SDE produces a process with exploding variance as $t \to \infty$. Thus, it is called ***Variance Exploding (VE)*** SDE.

In the same manner, the general reverse-time SDE can also be discretized for both DDPMs and SMLDs:

DDPMs:

$$d\mathbf{x} = -\frac{1}{2}b(t)(\mathbf{x}_t - \nabla_{\mathbf{x}_t} log p_t(\mathbf{x}_t))dt + \sqrt{b(t)}d\bar{\mathbf{w}} \tag{3.50}$$

SMLDs:

$$d\mathbf{x} = [-\frac{d[\sigma^2(t)]}{dt}\nabla_{\mathbf{x}_t} log p_t(\mathbf{x}_t)]dt + \sqrt{\frac{d[\sigma^2(t)]}{dt}}d\bar{\mathbf{w}} \tag{3.51}$$

Formal proofs for all the aforementioned results are thoroughly presented by Song et al. [44]. As it is evident, DDPMs and SMLDs constitute discretizations of SDEs.

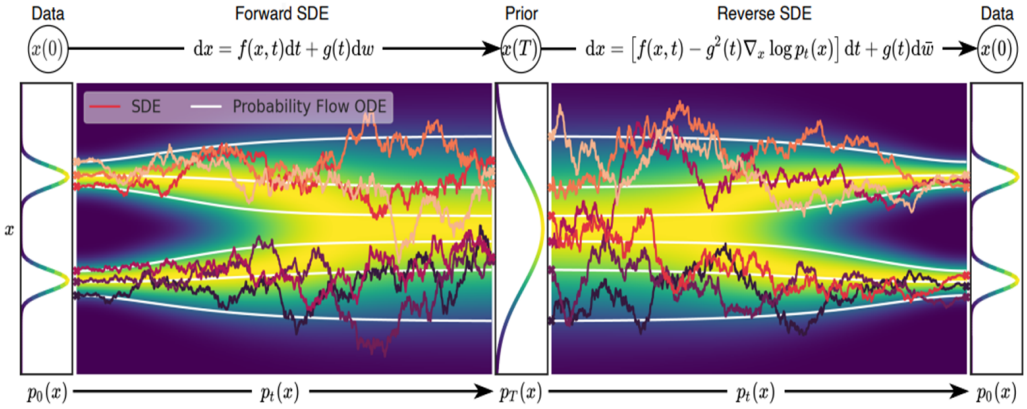


Figure 3.5. *Overview of score-based generative modeling through SDEs [44]*

The sampling procedure can be performed with any numerical method applied to the aforementioned reverse-time SDE. Examples of sampling algorithms that are used in practice are the Euler-Maruyama method as well as the Predictor-Corrector sampler that is discussed by Song et al. [44]. Moreover, it is shown by Song et al. [44] that the reverse Markov chain defined in DDPMs amounts to a numerical SDE solver. In the case of SMLDs, the reverse-time SDE is solved by Annealed Langevin sampling.

Chapter 4

Further Advancements in Diffusion Models

4.1 Conditional Generation - Introduction

All the formulations of diffusion models that have been discussed thus far concern unconditional sample generation. Conditional generation i.e. the generation of samples based on textual description or the style of another sample, is a crucial component of every generative model.

In mathematical terms, a condition, $\boldsymbol{y}$, is an additional input to the model (class label, text sequence, etc...) which is used in order to guide the sampling process towards a desirable class of samples. This condition is incorporated in the probability distribution of the reverse process as an additional conditioning parameter. Equations (3) and (5) and (14) are now written:

$$p_\partial(x_{t-1}|x_t, y) = \mathcal{N}(x_{t-1}; \mu_\partial(x_t, t, y), \Sigma_\partial(x_t, t, y)) \tag{4.1}$$

$$p_\partial(x_1, x_2, ..., x_T|y) = p(x_T) \prod_{t=1}^{T} p_\partial(x_{t-1}|x_t, y) \tag{4.2}$$

$$\mathbb{E}_{q(x_0, x_1, ..., x_T)} \Big[KL(q(x_T|x_0)||p(x_T)) + \Sigma_{t=1}^{T} KL(q(x_{t-1}|x_t, x_0)||p_\partial(x_{t-1}|x_t, y)) - log p_\partial(x_0|x_1, y) \Big] \tag{4.3}$$

An unconditional generative model attempts to sample from a distribution $\mathbf{p(x)}$ while a conditional generative model samples from a distribution $\mathbf{p(x|y)}$ where $\mathbf{y}$ is an additional input (class label, text sequence, etc...). As previously discussed, diffusion models use the score function of these distributions instead in order to steer the sampling process in a desirable direction. So, in conditional generation models the formula for the score function - and the algorithm for the sampling process in general - must now be updated and expressed in terms of $\mathbf{p(x|y)}$.

Based on the above vanilla formulation of conditioning we can now train a neural network such that $\mathbf{x}_\partial(\mathbf{x_t}, \mathbf{t}, \mathbf{y}) \approx \mathbf{x_0}$, $\boldsymbol{\epsilon}_\partial(\mathbf{x_t}, \mathbf{t}, \mathbf{y}) \approx \boldsymbol{\epsilon_0}$ or

$\mathbf{s}_\theta(\mathbf{x_t}, \mathbf{t}, \mathbf{y}) \approx \nabla_{\mathbf{x_t}} \mathbf{logp}(\mathbf{x_t}, \mathbf{y})$ depending on the specific interpretation and implementation.

A major downside of the aforementioned vanilla formulation is that a conditional diffusion model trained in this way may learn to ignore or downplay any given conditioning information. For this reason, another method called **Guidance** is proposed as a way to gain more control over amount of weight the model gives to the conditioning information. There are two types of Guidances: **Classifier Guidance** and **Classifier-free Guidance**.

4.2 Conditional Generation - Classifier Guidance

A first attempt at introducing practical conditional generation in the diffusion model framework was made by Dhariwal et al. in their paper: ***"Diffusion Models Beat GANs on Image Synthesis"*** [7]. In this paper, the authors propose a post-hoc method for conditioning based on a pre-trained classifier. They call this method of conditioning *Classifier Guidance*.

In more detail, Classifier guidance provides a way to steer diffusion sampling in the direction that maximizes the probability of the final sample being classified as a particular class. For this process, an auxiliary model is used, the *classifier*, that predicts $p(y|\mathbf{x_t})$, where y represents an arbitrary input feature, which could be a class label, a textual description of the input, or even a more structured object like a segmentation map or a depth map.

Initially formulated in the context of score functions, we can apply Bayes' rule in order to obtain an expression for the conditional score function:

$$p(x_t|y) = \frac{p(y|x_t)p(x_t)}{p(y)} \implies$$

$$logp(x_t|y) = logp(y|x_t) + logp(x_t) - logp(y) \implies$$

$$\nabla_{x_t} logp(x_t|y) = \nabla_{x_t} logp(y|x_t) + \nabla_{x_t} logp(x_t) \tag{4.4}$$

It is evident that the conditional score function is a sum of the unconditional score function and a conditioning term. Note that the quantity $\nabla_x logp(y|x_t)$ does not itself constitute a score function since the gradient is with respect to x and not y.

For intuition, the direction in which we are moving in the high-dimensional space from noise to data is now given as a vector sum of the direction provided by the original unconditional function and the direction that arises from the condition.

Dhariwal et al. made a modification to equation (45) by adding a parameter, γ, which they called the **guidance scale**:

$$\nabla_{x_t} logp_\gamma(x_t|y) = \nabla_{x_t} logp(x_t) + \gamma \nabla_{x_t} logp(y|x_t) \tag{4.5}$$

The guidance scale, as evident by its name, is responsible for scaling the conditioning term which allows for control over its influence in the generative process.

Obviously, classifier guidance does not explicitly refer to SMLDs but it can also be applied to DDPMs by introducing a conditioning parameter, $\mathbf{y}$, in the predicted noise term. By using the equivalance between the predicted noise of DDPMs and the score function of SMLDs:

$$\nabla_{x_t} log p_\partial(x_t) = -\frac{1}{\sqrt{1-\tilde{\alpha}_t}} \epsilon_\partial(x_t), \tag{4.6}$$

a formula for the conditional predicted noise can be derived:

$$\epsilon_\partial(x_t, t, y) \approx -\sigma_t \nabla_{x_t} log p(x_t|y) \tag{4.7}$$

In fact, the authors provide algorithms for conditional sampling for both DDPMs and DDIMs showcasing the flexibility of their method.

Figure 4.1. *Samples from an unconditional diffusion model with classifier guidance to condition on the class "Pembroke Welsh corgi". Using classifier scale 1.0 (left; FID: 33.0) does not produce convincing samples in this class, whereas classifier scale 10.0 (right; FID: 12.0) produces much more class-consistent images. [7]*

Although a powerful idea, Classifier guidance has **two important limitations** that make it impractical.

Since diffusion models gradually denoise the input in numerous steps, any separate classifier that will be used for guidance must also be able to cope with high noise levels. This would require training a classifier specifically for the purpose of guidance which can be very computationally expensive. The second limitation refers to the fact that most of the information found in the input $\mathbf{x}$ is not relevant to predicting y and thus, the quantity $\nabla_x log p(y|x)$ can yield arbitrary, and even adversarial, directions.

4.3 Conditional Generation - Classifier-free Guidance

The aforementioned shortcomings were tackled by Ho et al. in their paper ***"Classifier-Free Diffusion Guidance"*** [17]. As the name suggests, classifier-free guidance is a

method for guidance that does not require an auxiliary classifier model.

In this form of guidance, a Bayesian classifier is constructed by **combining** a conditional and an unconditional diffusion model. Learning two separate diffusion models could potentially be computationally expensive. For this, we learn both the conditional and unconditional diffusion models together as a **singular conditional model**. The unconditional diffusion model of the singular model can be queried by performing ***conditioning dropout***, i.e. some percentage of the time, the conditioning information, $\mathbf{y}$, is removed (10% - 20% tends to work well in practice).

In this way, the singular model is capable of generating samples both from $\mathbf{p(x_t|y)}$ and $\mathbf{p(x_t)}$ depending on whether the conditioning signal is provided or not. With this capability in mind, equation (50) can be written in the opposite direction as follows:

$$\nabla_{x_t} logp(y|x_t) = \nabla_{x_t} logp(x_t|y) - \nabla_{x_t} logp(x_t) \tag{4.8}$$

The conditioning term has now be expressed as a function of the conditional and unconditional score functions, both of which our singular diffusion model provides. We now substitute this formula in equation (51) for Classifier guidance:

$$\nabla_{x_t} logp_\gamma(x_t|y) = \nabla_{x_t} logp(x_t) + \gamma[\nabla_{x_t} logp(x_t|y) - \nabla_{x_t} logp(x_t))] \implies$$

$$\nabla_{x_t} logp_\gamma(x_t|y) = (1-\gamma)\nabla_{x_t} logp(x_t) + \gamma\nabla_{x_t} logp(x_t|y) \tag{4.9}$$

In mathematics, and specifically in the field of *affine spaces*, this is known as a *barycentric combination* (or *barycentric sum*). For $\gamma = 0$, we recover the unconditional model, and for $\gamma = 1$, we recover the standard conditional model. For $0 < \gamma < 1$, there is a trade-off between the conditioning and the non-conditioning term.

The interesting case is when the above barycentric combination becomes non-convex, i.e. when $\gamma > 1$. Then, the diffusion model prioritizes the conditional score function, while also moving in the opposite direction of the unconditional score function. As a result, samples that explicitly use the conditioning information are favored while the probability of generating samples that do not is significantly reduced.

4.4 Latent Diffusion Models and Conditioning via Cross-Attention

Despite the acceleration that DDIMs provide in the inference process, the diffusion models we have seen thus far operate in pixel space by manipulating tensors of the same size; a process with intrinsic limits. This results in slow inference speed and high computational cost.

All of the aforementioned problems were tackled by Rombach et al. in their seminal

Figure 4.2. *Two sets of samples from OpenAI's GLIDE model, for the prompt 'A stained glass window of a panda eating bamboo.', taken from their paper. Guidance scale 1 (no guidance) on the left, guidance scale 3 on the right. [30]*

paper titled ***"High-Resolution Image Synthesis with Latent Diffusion Models"*** [37].

The main idea behind this paper is to use a pre-trained autoencoder in order to enable the diffusion model to be trained on limited computational resources.

In particular, in the first step of the process, an encoder, $\mathcal{E}$, is used so as to extract a more compact representation of the original input, i.e. map it to a latent space of lower dimension. We will denote the latent-space input with $\mathbf{z}$.

Then, $\mathbf{z}$ is used as the input to the diffusion model. After the completion of the forward process, the latent input will have been transformed into its noisy version, $\mathbf{z_T}$. The $\mathbf{z_T}$ representation is then passed through the neural network (typically a U-Net) which has been trained to predict $\mathbf{z_{t-1}}$ given $\mathbf{z_t}$ at any time step t. After the completion of the reverse process, we get the output of the U-Net, $\mathbf{z_0}$, which is then passed through the pre-trained decoder, $\mathcal{D}$ which maps it from latent space back to the pixel space.

The training objective for the neural network of the diffusion model now becomes:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon,t}\left[||\epsilon - \epsilon_\theta(z_t, t)||_2^2\right] \tag{4.10}$$

Apart from the latent space, this paper also introduced a new method for conditional generation. The authors proposed to integrate additional information directly into the intermediate layers of the U-Net model using a cross-attention mechanism, similar to the Transformer architecture.

For this, an additional domain-specific encoder, τ_θ was used that projects the various modalities upon which we want to condition, $\boldsymbol{y}$, into an intermediate representation: $\mathbf{c} = \tau_\theta(y)$. This representation is then mapped into the intermediate layers of the U-Net according to the following equations:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}}) \cdot V$$

where:

$$Q = W_Q^{(i)} \cdot \varphi_i(z_t) \tag{4.11}$$

$$K = W_K^{(i)} \cdot \mathbf{c} \tag{4.12}$$

$$V = W_V^{(i)} \cdot \mathbf{c} \tag{4.13}$$

In the equations above, $\varphi_i(z_t)$ represents a flattened intermediate representation of the U-Net and the matrices $W_Q^{(i)}, W_K^{(i)}, W_V^{(i)}$ are learnable parameters. Finally, the training objective for the neural network becomes:

$$L_{LDM} = \mathbb{E}_{\mathcal{E}(x),\epsilon,t}\left[\|\epsilon - \epsilon_\theta(z_t, c, t)\|_2^2\right] \tag{4.14}$$

This paper gave the inspiration for the very popular ***"Stable Diffusion"*** image generator AI model which came onto the scene in August of 2022 and fascinated the whole world, academic and not, with its high-quality images and capabilities.

Figure 4.3. *An image created by Stable Diffusion with the prompt "A Water Butterfly". Source: https://www.reddit.com/r/StableDiffusion/*

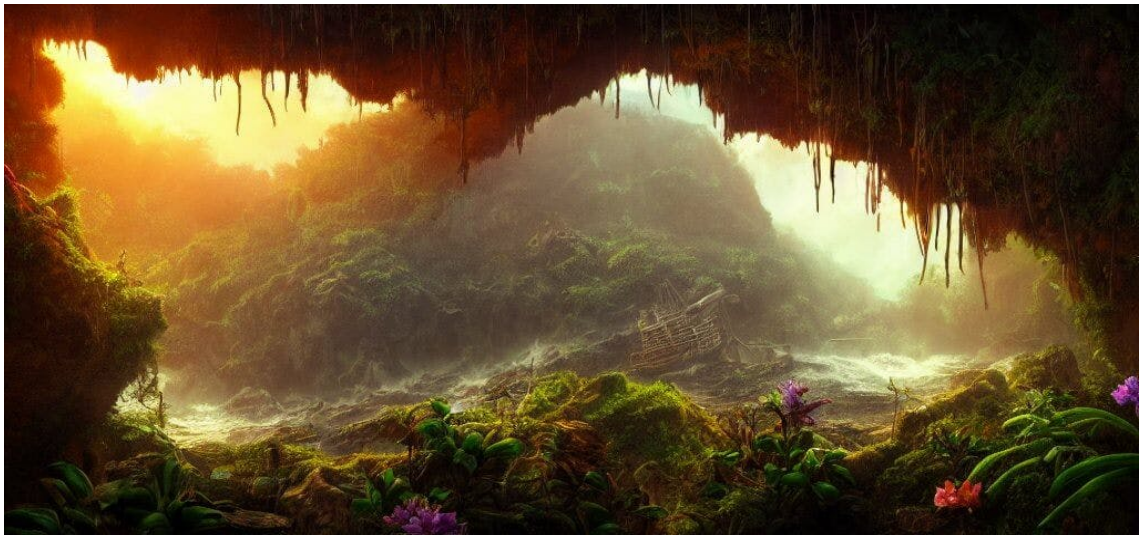


Figure 4.4. *An image created by Stable Diffusion with the prompt "5. A Landscape View Of A River From A Forest Cave ". Source: https://www.reddit.com/r/StableDiffusion/*

Chapter 5

Image Segmentation and Object Detection

5.1 Introduction

Image segmentation and object detection are fundamental tasks in the fields of computer vision and machine learning, aiming to interpret and understand the contents of visual data. These tasks have widespread applications, ranging from autonomous driving and medical imaging to surveillance and robotics.

Image segmentation involves partitioning an image into multiple segments or regions, each representing a different object or part of an object. This process helps in isolating and analyzing various components within an image, facilitating more precise object recognition and scene understanding.

Object detection, on the other hand, focuses on identifying and locating objects within an image. This task not only involves classifying objects into predefined categories but also drawing bounding boxes around each detected object. Effective object detection systems must handle various challenges, such as variations in object appearance, scale, occlusion, and complex backgrounds.

In the early stages, image segmentation and object detection were primarily explored within the realm of traditional computer vision techniques, relying on handcrafted features and simple classifiers [27, 20]. As the field evolved, the advent of deep learning revolutionized these tasks, with Convolutional Neural Networks (CNNs) playing a pivotal role. CNN-based models like Fully Convolutional Networks (FCNs) [26], U-Net [38] , and Faster R-CNN [36] demonstrated significant improvements in accuracy and efficiency. These advancements enabled more sophisticated applications and provided the foundation for modern computer vision systems.

More recently, the introduction of Transformer-based architectures [47] has further transformed the landscape. These models, which excel in capturing long-range dependencies and contextual information, have become the backbone of state-of-the-art solutions. Notable examples in the field include the **Segment Anything Model (SAM)** [22] and the **GroundingDino** model [25], which leverage large-scale datasets and advanced algorithms

to deliver unparalleled performance in image segmentation and object detection tasks respectively.

5.2 Segment Anything Model (SAM)

The Segment Anything Model (SAM) [22] marks a significant leap in image segmentation technology. Designed to be a flexible, all-purpose, zero-shot model, SAM can segment a wide variety of objects in diverse images, thanks to its use of extensive datasets and cutting-edge deep learning techniques.

SAM's versatility comes from its ability to take different types of input prompts—points, boxes, and masks—which guide the segmentation process. This flexibility means SAM can adapt to various segmentation needs, whether it's defining precise object boundaries or segmenting broader regions within an image.

At the heart of SAM is an advanced encoder-decoder architecture. The encoder is tasked with extracting detailed, multi-scale features from the input image through a series of convolutional layers. These layers progressively downsample the image, capturing both spatial hierarchies and contextual information. The decoder then takes these features and generates high-resolution segmentation masks, using upsampling layers and skip connections to merge fine-grained details from earlier layers with the higher-level semantic information from the encoder.

One of the standout features of SAM is its ability to handle different segmentation prompts effectively. For instance, when given point-based prompts, the model generates segmentation masks centered around those points, accurately capturing objects even in cluttered scenes. With box-based prompts, SAM refines the segmentation within the specified bounding box, ensuring precise object boundaries. Mask-based prompts allow SAM to improve and refine existing segmentations, making it a powerful tool for iterative segmentation tasks.

SAM was trained on a large-scale dataset, **SA-1B**, consisting of **eleven million (11M) images** and **one billion (1B) masks**. This extensive training data enables the model to generalize well across different types of objects and scenes. The training process also incorporates advanced data augmentation techniques and regularization methods to boost the model's robustness and performance.

5.3 GroundindDINO

The GroundingDINO model [25] is a significant advancement in the field of object detection. GroundingDINO is designed to address open-set object detection by integrating

Figure 5.1. *Van Gogh's painting titled "Farmhouse in Provence". Source: https://segment-anything.com/*

Figure 5.2. *Van Gogh's painting titled "Farmhouse in Provence" segmented by SAM. Source: https://segment-anything.com/*

text and visual inputs, enabling it to detect and understand objects specified by text descriptions or categories. This capability makes it highly versatile and effective in complex, real-world scenarios.

Just like SAM, GroundingDINO also employs a Transformer-based architecture. The model uses an encoder-decoder structure, where the encoder processes input images to extract detailed, multi-scale features. These features are then integrated with text features using a Feature Enhancer module, which combines visual and textual data into a unified representation. This combined data is fed into the decoder, which refines the object detection boxes to align with the textual input.

The model outputs **900 object bounding boxes** along with similarity scores to the input words. It then selects the boxes with the highest similarity scores above a certain threshold, effectively identifying the objects described by the text. This approach allows GroundingDINO to perform tasks such as referring expression comprehension, where the model identifies objects based on descriptive text prompts (e.g., "the red car").

GroundingDINO is trained on numerous large and diverse datasets, which include annotations for various objects and scenes. It showcased exceptional performance on many popular benchmarks in object detection such as the COCO (Common Objects in Context) dataset [24], the LVIS (Large Vocabulary Instance Segmentation) dataset [13], .This extensive training helps the model generalize well across different categories and environments. The training process incorporates advanced data augmentation techniques and regularization methods to enhance robustness and accuracy.

The model has demonstrated remarkable performance on several benchmarks. For instance, it achieved a 52.5 Average Precision (AP) on the COCO zero-shot detection benchmark without any training data from COCO, and set a new record on the ODinW zero-shot benchmark with a mean AP of 26.1. These results underscore GroundingDINO's ability to detect objects in a zero-shot manner, meaning it can recognize and locate objects it has never seen before during training.

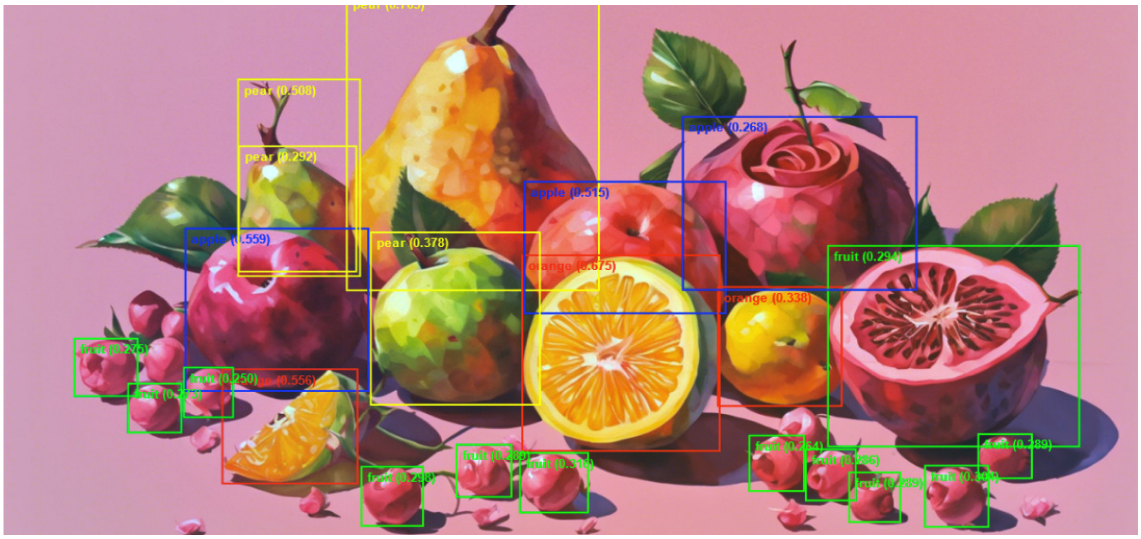


Figure 5.3. *Fruit Detection in a painting by GroundingDino. Source: https://www.mlwires.com/grounding-dino-1-5-a-powerful-open-set-object-detection-model/*

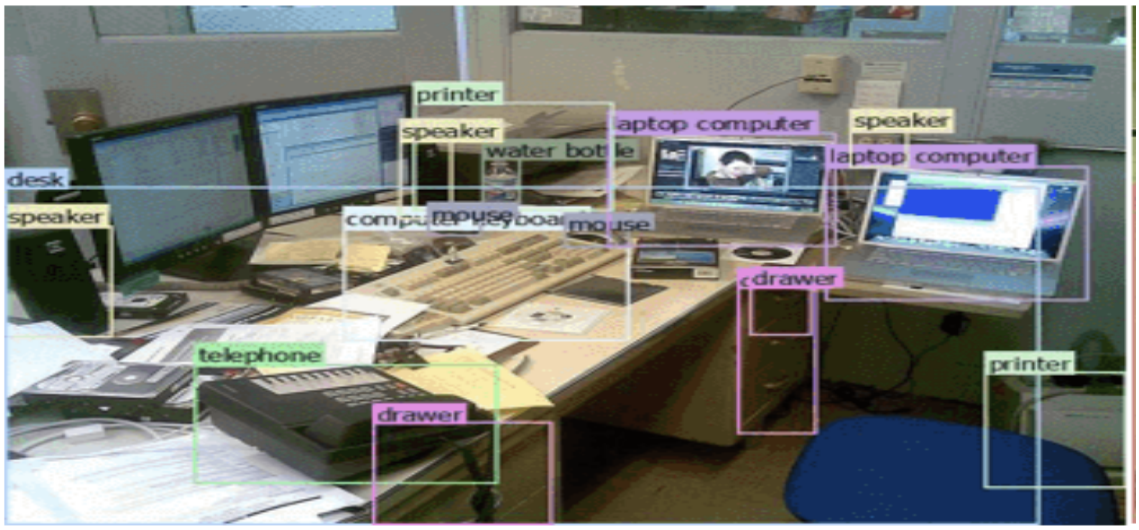


Figure 5.4. *Object Detection in an office by GroundingDino. Source: https://deepdataspace.com/blog/Grounding-DINO-1.5-Pro*

5.4 Medical Image Segmentation

Medical image segmentation is a crucial process in medical imaging that involves partitioning an image into different regions, typically to isolate and analyze anatomical structures or regions of interest. This process is fundamental for various clinical applications, including diagnosis, treatment planning, and the monitoring of disease progression. Accurate segmentation enables clinicians to quantify tissue volumes, detect abnormalities, and guide surgical procedures with precision.

The importance of medical image segmentation lies in its ability to provide detailed and quantitative information about anatomical structures. For instance, in oncology, segmentation helps in identifying the exact location and size of tumors, which is essential for planning radiation therapy and assessing treatment response. In cardiology, segmentation of heart chambers and blood vessels aids in evaluating cardiac function and diagnosing cardiovascular diseases. Moreover, segmentation is pivotal in neuroscience for studying brain anatomy and identifying pathological changes associated with neurological disorders. [49, 46]

However, medical image segmentation poses several challenges. One significant challenge is the variability in anatomical structures across different patients and the presence of noise and artifacts in medical images. These factors can complicate the segmentation process and lead to inaccuracies. Additionally, manual segmentation by radiologists is time-consuming and subject to inter- and intra-observer variability. To overcome these challenges, there has been a growing interest in developing automated segmentation methods using machine learning and deep learning techniques.

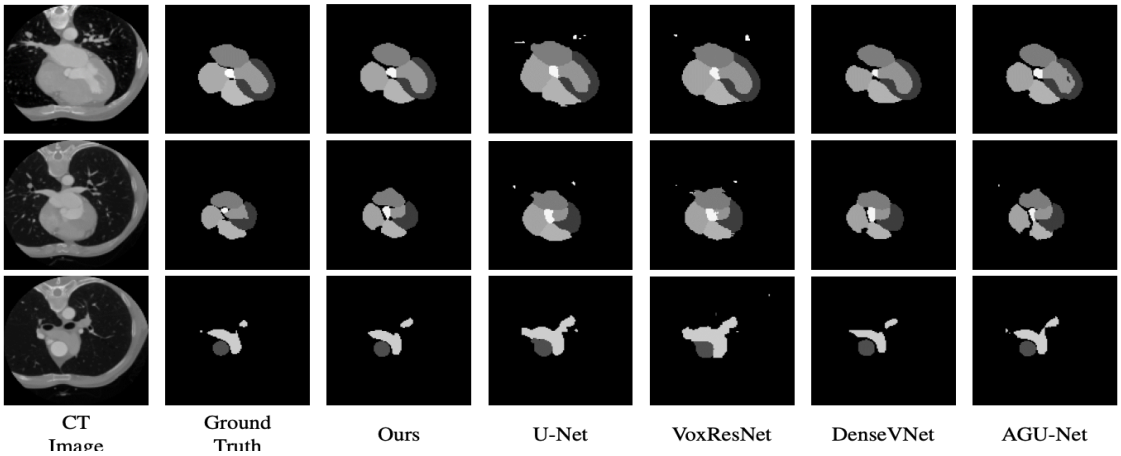


Figure 5.5. *Examples of results of the proposed method of Park et al. [31]*

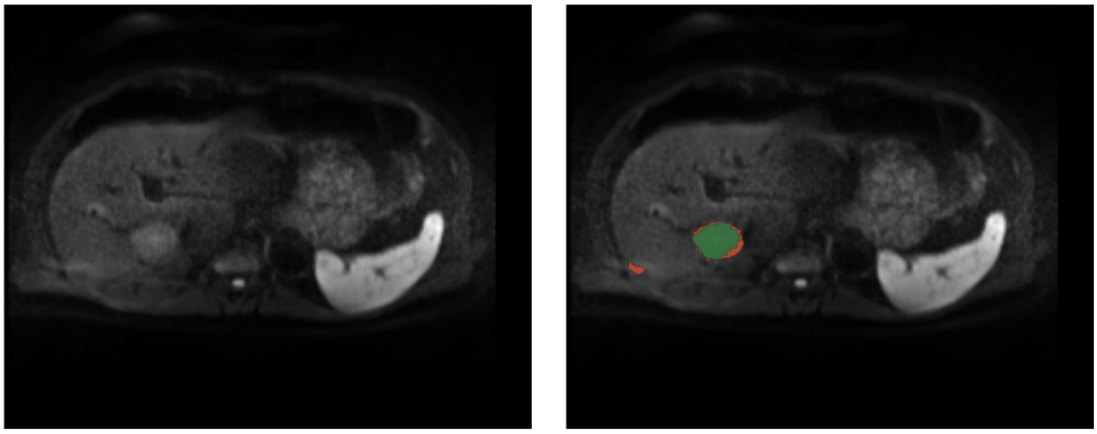


Figure 5.6. *Automatic liver lesion segmentation with the method proposed by Christ et al.[6]*

5.5 Brain Tumour Segmentation

Brain tumor segmentation is a specific application of medical image segmentation that focuses on identifying and delineating tumors in brain MRI scans. This task is particularly challenging due to the complex and heterogeneous nature of brain tumors, which can vary greatly in size, shape, and appearance. Accurate segmentation of brain tumors is critical for diagnosing the type and grade of the tumor, planning surgical interventions, and monitoring the effectiveness of treatments.

The segmentation of brain tumors involves distinguishing tumor tissue from normal brain tissue and other structures. This is essential for radiologists and neurosurgeons to develop precise treatment plans and evaluate patient prognosis. Automated brain tumor segmentation methods, particularly those based on deep learning, have shown promising

results in improving segmentation accuracy and reducing the time required for analysis.

Despite the advancements, brain tumor segmentation still faces several challenges. The variability in tumor appearance across different patients, the presence of edema and necrotic regions, and the low contrast between tumor boundaries and surrounding tissues can hinder accurate segmentation. Furthermore, the scarcity of labeled training data and the need for robust models that generalize well to diverse clinical settings remain significant obstacles [48].

5.6 BraTS Dataset

The Brain Tumor Segmentation (BraTS) dataset is one of the most widely used and influential datasets in the field of medical image analysis. It provides a comprehensive set of multi-modal MRI scans for the segmentation of brain tumors, specifically gliomas, which are among the most common and aggressive types of brain tumors. The dataset includes four types of MRI sequences: T1, post-contrast T1-weighted (T1Gd), T2-weighted (T2), and Fluid-Attenuated Inversion Recovery (FLAIR). These sequences offer diverse and complementary information about brain anatomy and pathology, facilitating robust and accurate tumor segmentation.

The significance of the BraTS dataset lies in its ability to provide a standardized benchmark for the development and evaluation of brain tumor segmentation algorithms. By offering a large collection of annotated scans, BraTS enables researchers to train and validate their models on a diverse set of images, thereby improving the generalizability and robustness of their methods. The dataset includes annotations for different tumor regions, such as the enhancing tumor, the tumor core, and the whole tumor, which are essential for comprehensive tumor analysis and treatment planning.

The BraTS 2021 dataset continues to build on this legacy, offering an even larger and more diverse set of data for training, validation, and testing. The BraTS 2021 dataset includes 1251 MRI scans in total, divided into 1251 for training, 219 for validation, and 530 for testing. This extensive dataset provides a robust foundation for developing advanced segmentation algorithms and has been used in numerous studies and competitions to benchmark the performance of new methods. The dataset's comprehensive nature ensures that models trained on BraTS 2021 are well-equipped to handle the variability and complexity of real-world clinical data, making it an invaluable resource for the medical imaging community.

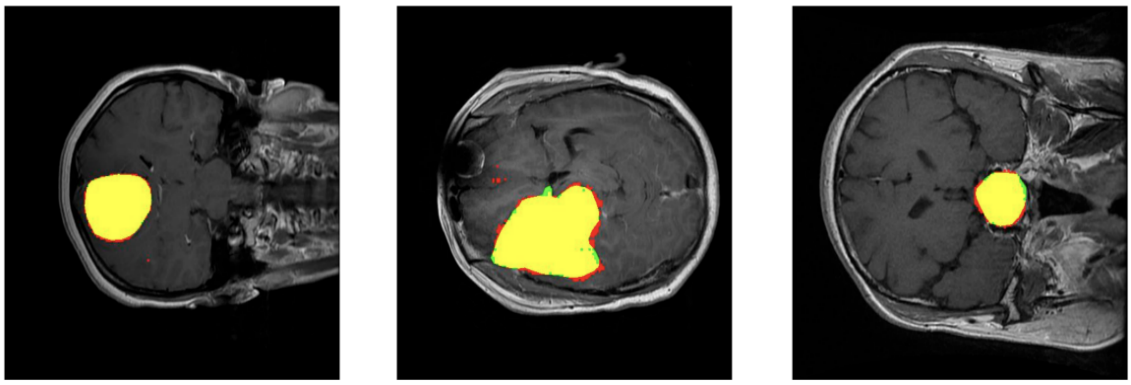


Figure 5.7. *Examples of results of the proposed method of Diaz-Pernaz et al. [8]*

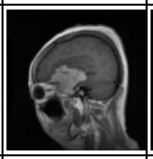
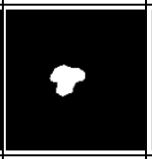
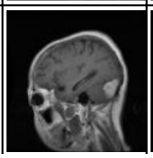
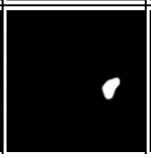
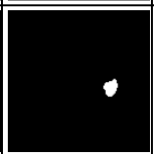
Instance Number	MRI Scan	Ground Truth	U-Net	Attention U-Net	ResUnet	ResUnet++	R2Unet
1.							
2.							

Figure 5.8. *Examples of results of the proposed method of Gupta et al. [14]*

5.7 Medical Segmentation Decathlon

The Medical Segmentation Decathlon (MSD) [3] was launched in 2018 with the purpose of providing a comprehensive benchmark for the development and evaluation of medical image segmentation algorithms. The primary goal of the MSD is to encourage the creation of generalizable and robust segmentation methods that can perform well across a variety of tasks and imaging modalities. By offering a diverse set of datasets, the MSD challenges researchers to develop techniques that are not only accurate but also adaptable to different types of medical data.

The MSD dataset is composed of ten distinct medical imaging datasets, each representing a different anatomical structure or pathology. These include brain tumors, liver tumors, hippocampus, prostate, lung tumors, cardiac structures, pancreas, hepatic vessels, spleen, and colon cancer. Each dataset consists of multi-modal imaging data with corresponding expert annotations, providing a rich and diverse resource for training and evaluating segmentation models. The inclusion of multiple modalities and anatomical regions ensures that models developed using the MSD can be tested for their robustness and versatility across a wide range of medical imaging challenges.

Since its launch, the Medical Segmentation Decathlon has inspired numerous papers and advanced segmentation techniques [15, 19, 45]. It has become a standard benchmark for evaluating new methods, driving innovation in the field of medical image analysis. Techniques such as nnU-Net, which automatically configures itself for any given dataset, were significantly influenced by the challenges posed by the MSD. The decathlon format has encouraged the development of algorithms that are not only high-performing on individual tasks but also capable of generalizing across different types of medical data. This has led to significant advancements in the robustness and applicability of medical image segmentation methods.

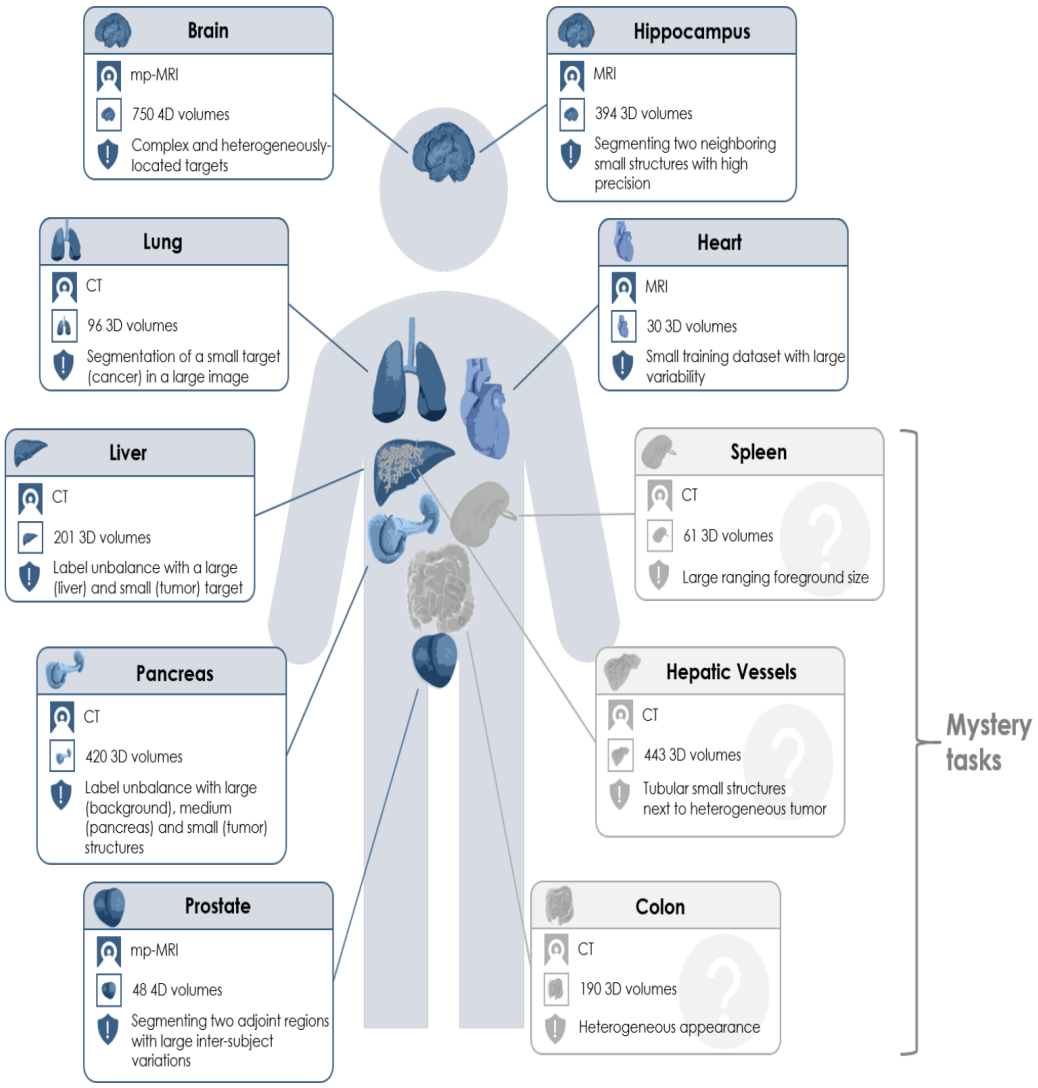


Figure 5.9. *Overview of the ten different tasks of the Medical Segmentation Decathlon (MSD) [3]*

5.8 MONAI

The MONAI (Medical Open Network for AI) framework is an open-source, PyTorch-based library created by NVIDIA and King's College London in April 2020, along with collaboration from academic, clinical, and industry partners [4]. It is designed specifically for healthcare imaging, facilitating the design, development, and deployment of deep learning models in medical imaging applications. MONAI provides researchers and developers with a comprehensive suite of tools and workflows, from data loading and transformation to model training and evaluation. This framework addresses the unique challenges in medical imaging, such as handling diverse data formats, integrating domain-specific preprocessing techniques, and ensuring robust model performance across different imaging modalities.

One of the key features of MONAI is its emphasis on reproducibility and standardization in medical imaging research. The framework includes standardized pipelines for common medical imaging tasks, such as segmentation, classification, and detection, which can be easily customized and extended. Additionally, MONAI supports a variety of imaging modalities, including MRI, CT, and ultrasound, making it a versatile tool for a wide range of medical imaging applications. The integration with PyTorch ensures compatibility with state-of-the-art deep learning models and allows for seamless incorporation into existing research workflows. Furthermore, the MONAI Model Zoo provides pre-trained models and scripts for various medical imaging tasks, accelerating the development and deployment of new models.

The impact of MONAI is evident in its widespread adoption and the numerous studies that have utilized the framework. Researchers have used MONAI to develop robust and accurate models for tasks such as tumor segmentation, organ delineation, and disease detection [23, 32, 34, 35]. The framework's comprehensive documentation, active community support, and continuous development have made it a go-to resource for medical imaging researchers worldwide. By providing a standardized and flexible platform, MONAI has significantly advanced the field of medical imaging and facilitated the translation of deep learning research into clinical practice.

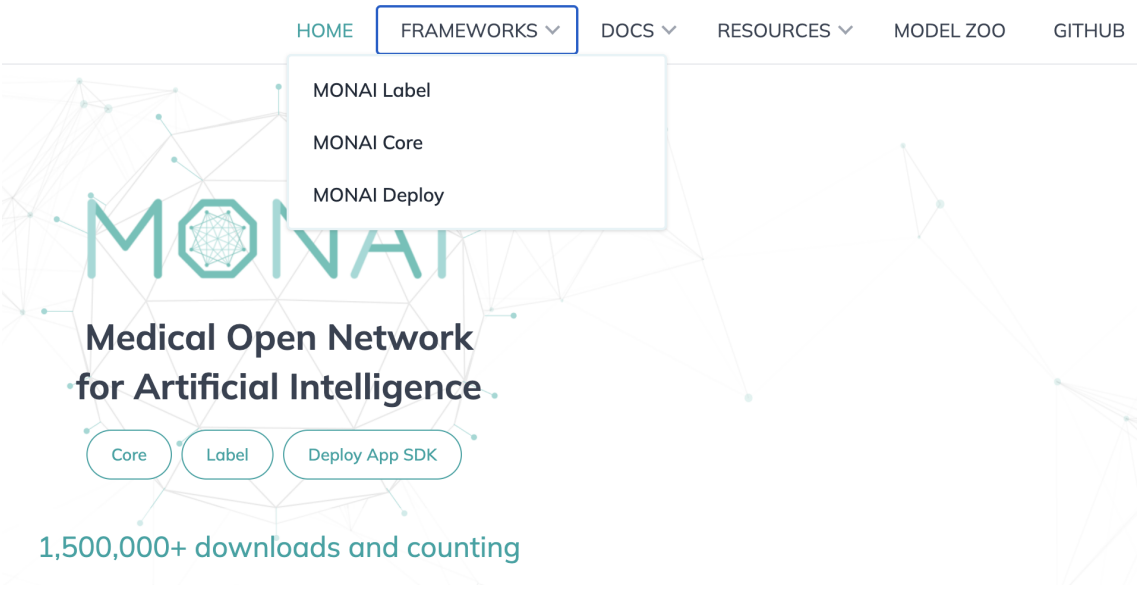


Figure 5.10. *Overview of the MONAI framework and its components. Source: https://monai.io*

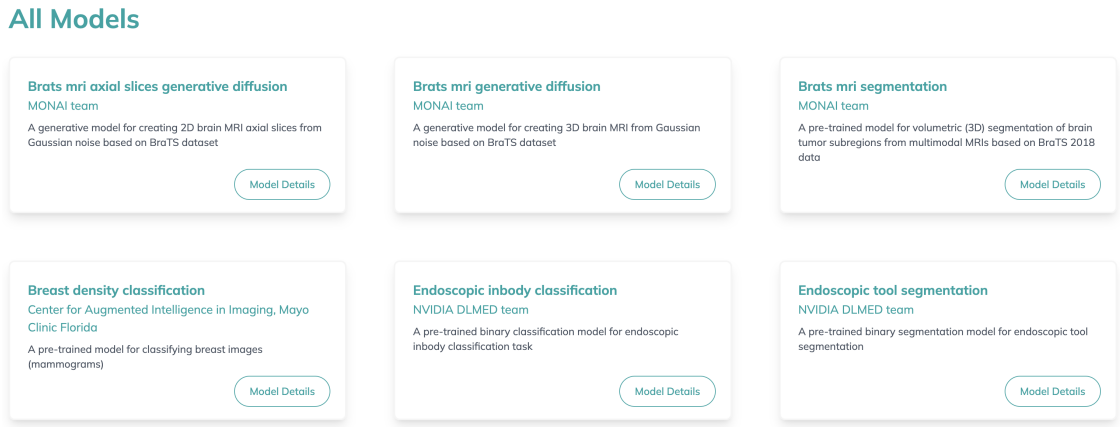


Figure 5.11. *Overview of some of the models available in MONAI's Model Zoo. Source: https://monai.io*

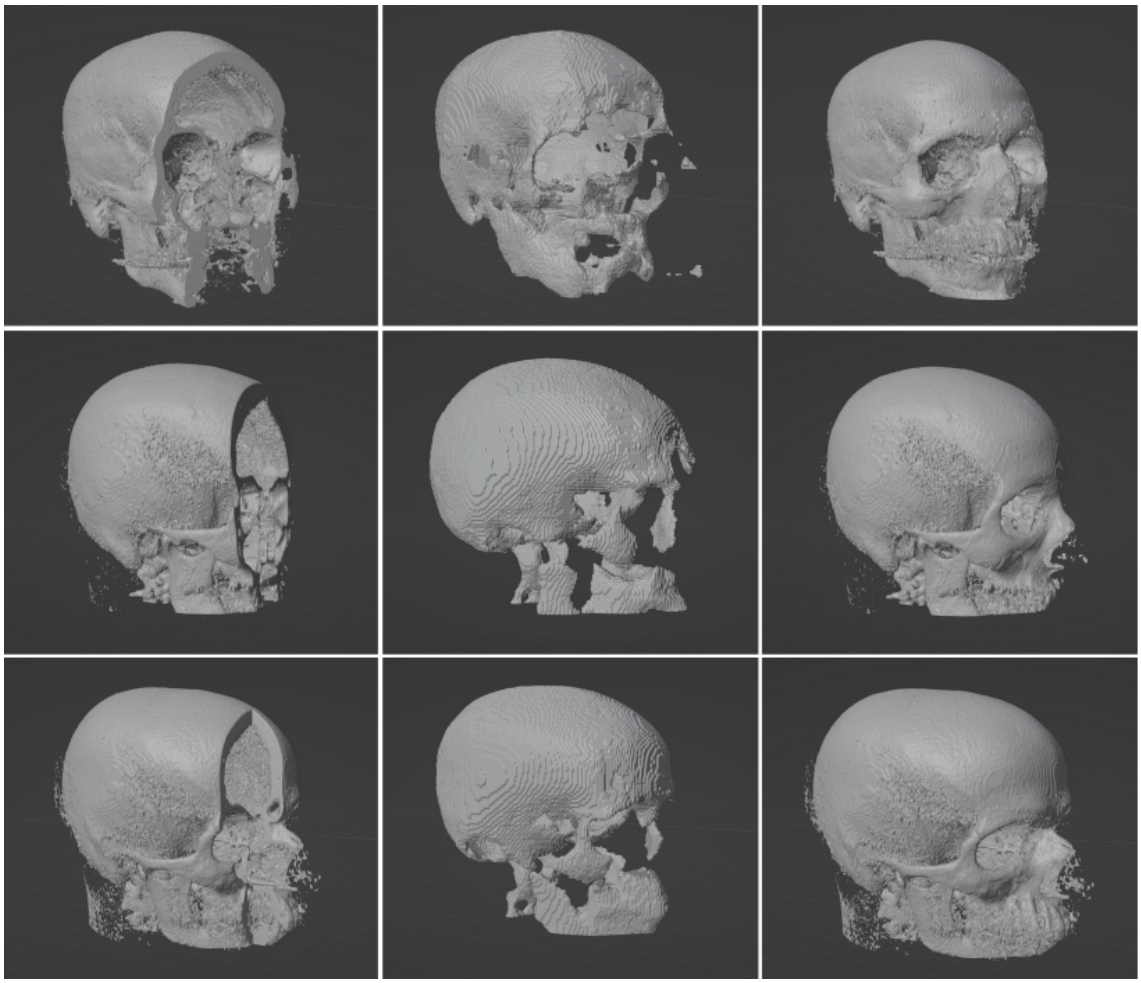


Figure 5.12. *Reconstruction of facial defects on the MUG500+ dataset, using MONAI's pre-trained model [23]*

Chapter 6

Our Framework

6.1 General Overview of the Framework

In this chapter, we introduce our novel framework for the segmentation and localization of lesions in brain MRI images. Building on the concepts discussed in the previous chapters, this framework specifically targets the challenges of accurately identifying and segmenting brain lesions. Our approach integrates advanced deep learning techniques, including diffusion models, attention mechanisms, and segmentation models, to create a robust and efficient solution.

Our framework begins with the use of a diffusion model to generate healthy counterfactual images from brain MRIs that contain lesions. These counterfactual images represent what the brain would look like without any pathological abnormalities. This step is crucial as it provides a baseline for identifying and localizing the lesions by highlighting deviations from the healthy state.

Following the generation of these counterfactual images, we apply a series of preprocessing steps to prepare the data for segmentation. This includes standardization, noise reduction, and any necessary transformations to enhance the quality and consistency of the images.

The core of our framework utilizes the Segment Anything Model (SAM) in two distinct and independent pipelines for lesion segmentation. In the first pipeline, SAM receives a point prompt, typically inserted by a doctor. For our experimental purposes, we use the centroid of the ground truth mask as the point prompt. SAM then segments the lesion based on this prompt, providing precise localization of the lesion.

In the second pipeline, we utilize a text prompt as input, which is processed by the GroundingDINO model. GroundingDINO generates a bounding box around the area of interest, effectively performing object detection. This bounding box is then used as input for SAM, which segments the lesion within the specified region.

To evaluate the performance of our framework, we employ several metrics, including

Dice score, Intersection over Union, AUPRC, precision, recall, specificity, F1 score, Hausdorff distance, and Average Symmetric Surface Distance (ASSD). These metrics provide a comprehensive assessment of the accuracy and reliability of our segmentation results.

Additionally, we conduct an ablation study [28] to understand the contribution of each component within our framework. By systematically removing individual components and measuring the performance impact, we gain insights into the significance and interplay of the various elements in our approach.

This chapter provides a detailed exploration of each component of our framework, the experimental setup, and the results of our ablation study, highlighting the strengths and potential areas for improvement in our approach to brain MRI lesion segmentation.

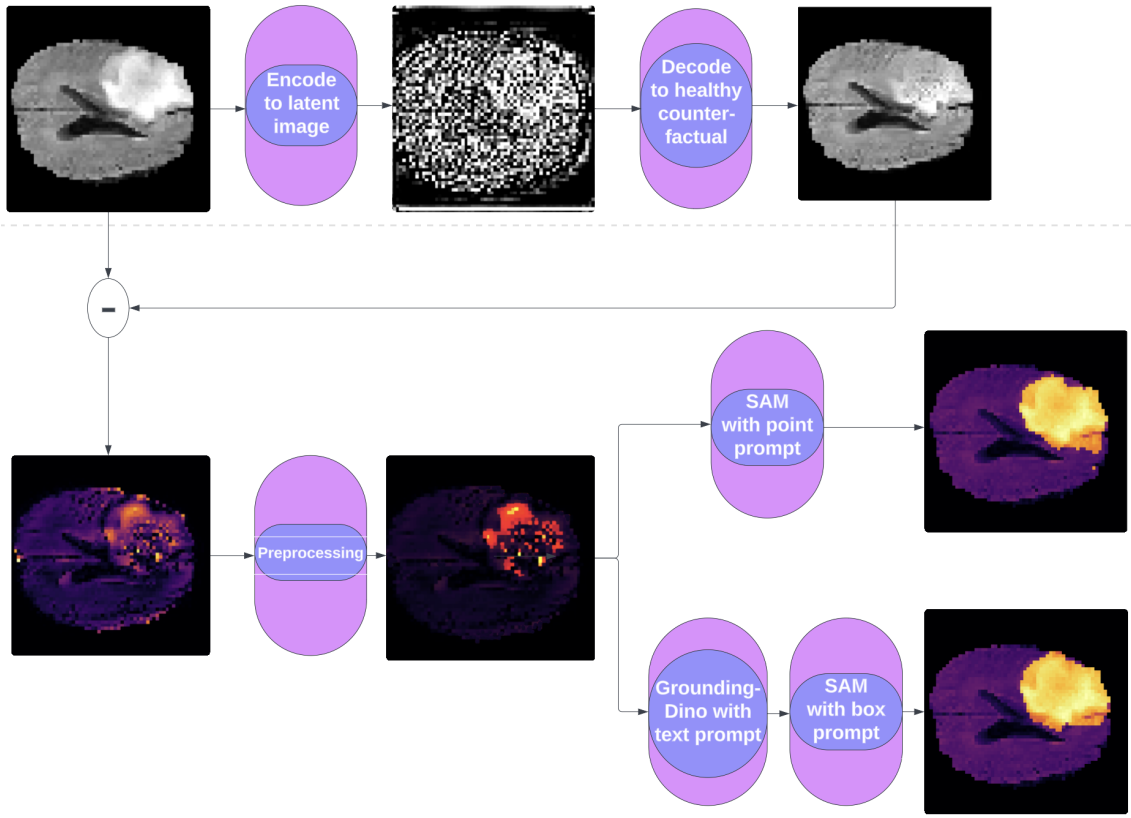


Figure 6.1. *Our Framework*

6.2 Counterfactual Diffusion Generation

Our approach focuses on transforming an input image from the unhealthy domain to the healthy domain during inference, while preserving all other characteristics of the image. Specifically, we aim to identify and modify the key features indicative of lesions, such as those in brain tumor datasets.

This precise adjustment, referred to as counterfactual generation in causal literature, allows us to pinpoint the minimal changes needed to convert an unhealthy image to a healthy one. Once our model $\partial(x_t, c, t)$ is trained on imaging data with $\mathbf{c} \in$ [healthy,unhealthy], we can manipulate the input image between these domains at inference by generating counterfactuals, inspired by methodologies outlined in [39, 40, 50].

6.2.1 Training

As mentioned in the first chapter, in order to achieve optimal solutions for $\nabla_x log p_t(\mathbf{x})$ we train our model, ∂, to approximate $\nabla_x log p_t(\mathbf{x}_t|\mathbf{x}_0)$. This process involves using a conditional denoising U-Net, $\epsilon_\partial(x_t, c, t)$ to control the synthesis process via input c. During training, the goal is to learn a ∂^* that minimizes the expectation of the squared error between $\partial(x_t, c, t)$ and the noise term (Equation 2.25). This is represented by the equation:

$$\partial^* = argmin \quad \mathbb{E}_{\mathbf{x}_0,t,\epsilon}[\|\epsilon_\partial(\mathbf{x}_t, \mathbf{c}, t) - \epsilon\|_2^2] \tag{6.1}$$

, where $\mathbf{x}_t = \sqrt{a_t}\mathbf{x}_0 + \sqrt{1 - a_t}\epsilon$, with $\mathbf{x}_0 \sim p_{data}$ being a sample from the (training) data distribution, $t \sim U(0, T)$ and $\epsilon \sim \mathcal{N}(0, I)$.

6.2.2 Inference

Once our model is trained, sample generation starts with $x_T \sim \mathcal{N}(\prime, \mathcal{I})$ and iteratively samples from the reverse process using the diffusion model. We employ the Denoising Diffusion Implicit Models (DDIM) method, which allows for deterministic mapping from latent variables to images. The DDIM formulation provides two main benefits: near-invertible mapping between x_T and x_0 and efficient sampling with fewer iterations, even with the same diffusion discretization. This is achieved by selecting different under-sampling times within the interval $[0, T]$.

This structured approach ensures that our model efficiently learns the transformation needed for accurate counterfactual image generation, which is crucial for precise lesion localization.

6.2.3 Estimating the Difference Image with Counterfactual Diffusion

To generate counterfactual images, we first encode the input image into a latent space by iteratively applying a reverse process for L iterations using an unconditional model. We then decode this latent representation while applying an intervention to the conditioning

c to denote a "healthy" state. This decoding is done by applying the reverse process with implicit guidance and attention conditioning.

The difference between the original and counterfactual images is averaged along the channel dimension to create a heatmap, which highlights the lesion (unhealthy features) for segmentation. Dynamic normalization is applied throughout the inference process to maintain consistency and accuracy.

6.2.4 Implicit Guidance

Generating counterfactuals using a classifier to guide the diffusion process, which involves training an additional model on noisy images, has proven effective [39, 50]. In our approach, we utilize implicit guidance for counterfactual generation. Here, a single diffusion model is trained on both conditional and unconditional objectives by randomly omitting the conditioning variable c during training. The omitted conditioning is denoted as $c = \emptyset$, leading to conditional $\epsilon_\partial(\mathbf{x}_t, c, t)$ and unconditional $\epsilon_\partial(\mathbf{x}_t, \emptyset, t)$ predictions. During sampling, these predictions are combined using a guidance scale w, resulting in $\epsilon_\partial(\mathbf{x}_t, \mathbf{c}, t) = w\epsilon_\partial(\mathbf{x}_t, \mathbf{c}, t) + (1 - w)\epsilon_\partial(\mathbf{x}_t, \emptyset, t)$.

6.2.5 Conditioning

To generate counterfactuals effectively, conditioning during the decoding process is essential. As a baseline, we use adaptive group normalization (AdaGroup), which has been effective in Denoising Probabilistic Models (DPMs). However, for counterfactual generation, normalization alone is insufficient. We enhance conditioning by incorporating a conditional attention mechanism inspired by text-to-image generation methods.

We preprocess the conditioning variable c using an encoder τ_ϕ which projects c into an intermediate representation. This representation is further projected to match the dimensionality of each attention layer within the model and concatenated to the attention context at each layer.

Our approach employs a U-Net with an attention layer that implements softmax, $(\frac{QK_{\mathbf{c}}^T)}{\sqrt{d}}V_{\mathbf{c}}$. The values for $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ are derived from the previous convolutional layer, with $\tau_\phi(\mathbf{c})$ concatenated to $\mathbf{K}$ and $\mathbf{V}$ before the attention layer, forming $\mathbf{K_c} = concat([\mathbf{K}, \tau_\phi(\mathbf{c})])$ and $\mathbf{V_c} = concat([\mathbf{V}, \tau_\phi(\mathbf{c})])$ [30]. This conditional attention mechanism significantly improves the effectiveness of counterfactual generation.

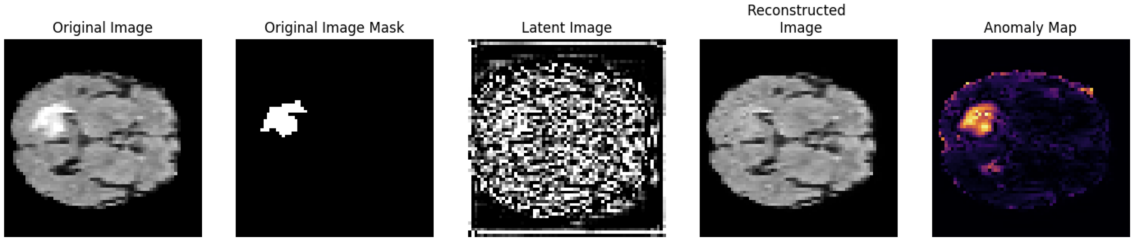


Figure 6.2. *Counterfactual generation process - Example 1*

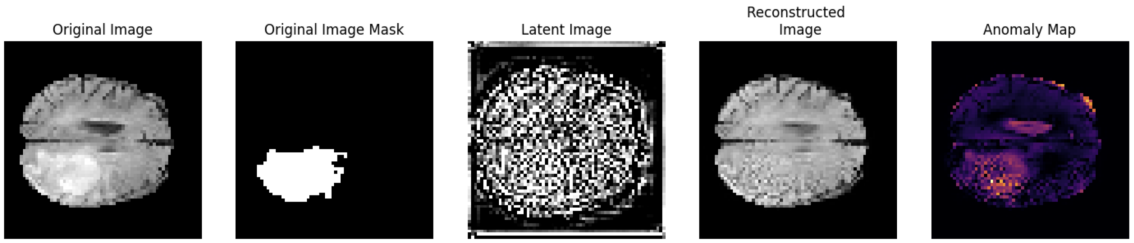


Figure 6.3. *Counterfactual generation process - Example 2*

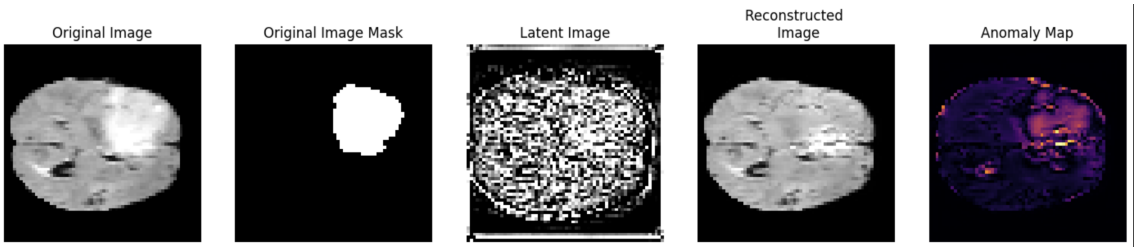


Figure 6.4. *Counterfactual generation process - Example 3*

6.3 Lesion Segmentation

Once the counterfactual image is generated, a difference image is produced by subtracting the counterfactual from the original MRI. However, this difference image may not always be of optimal quality due to several reasons. Firstly, the lesion area might not be sufficiently highlighted, making it challenging to distinguish from the rest of the brain. Secondly, other regions apart from the lesion may also be highlighted, leading to potential misinterpretations.

These issues can arise due to various factors, such as the diffusion model's inability to accurately inpaint the lesion regions and generate a completely healthy counterfactual. Additionally, poor quality MRI slices with unclear lesions can contribute to suboptimal difference images. Addressing these shortcomings is crucial for accurate lesion segmentation, which is the focus of the following sections.

6.3.1 Preprocessing

In the preprocessing stage, our goal is to enhance the quality of the difference image generated by subtracting the counterfactual from the original MRI. This is achieved through a series of multiplicative operations designed to accentuate the lesion area while suppressing non-lesion regions that might have been inadvertently highlighted.

The first step involves multiplying the difference image by the original MRI image. This step ensures that the bright areas corresponding to the lesion remain prominent if the diffusion model has accurately inpainted the lesion region. Simultaneously, other areas that might have been highlighted by accident in the difference image become darker.

To further enhance the contrast and clarity of the lesion area, we repeat this multiplication process multiple times. Specifically, the product of the first multiplication is again multiplied by the original MRI image, and this process is repeated as needed. The number of times this multiplication is performed is a hyperparameter of our framework, allowing for fine-tuning based on the quality of the initial difference image and the characteristics of the dataset.

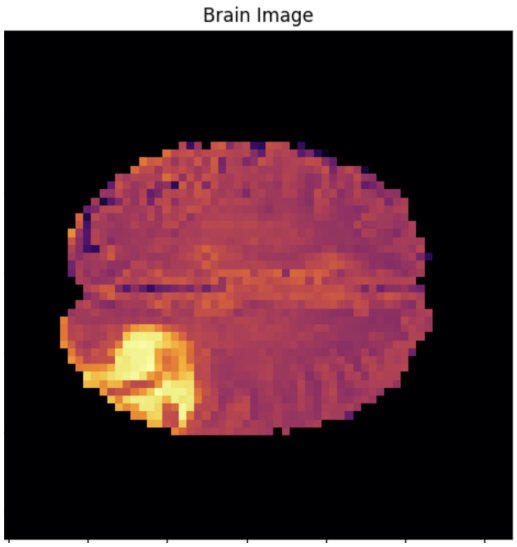


Figure 6.5. *Original brain MRI*

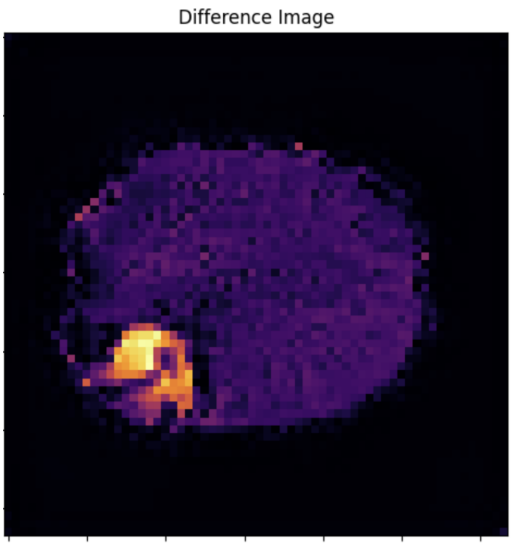


Figure 6.6. *Difference Image*

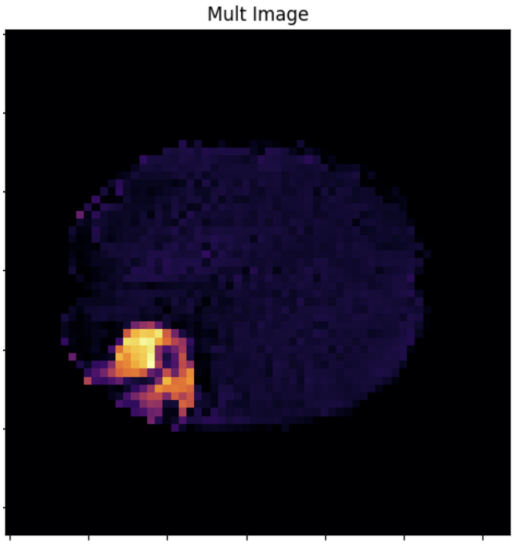


Figure 6.7. *Result of multiplying difference image with the brain MRI image*

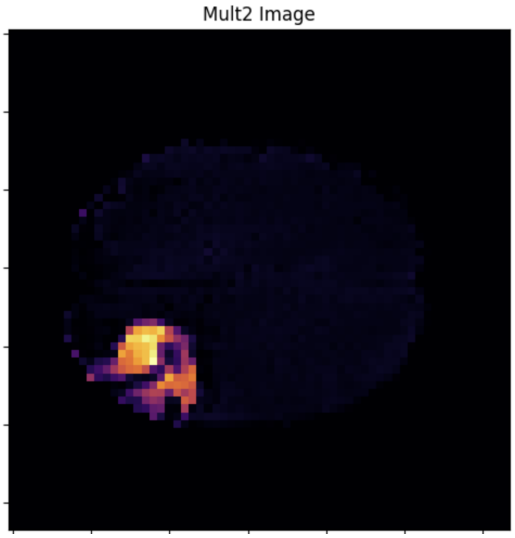


Figure 6.8. *Result of multiplying the difference image with the brain MRI image squared.*

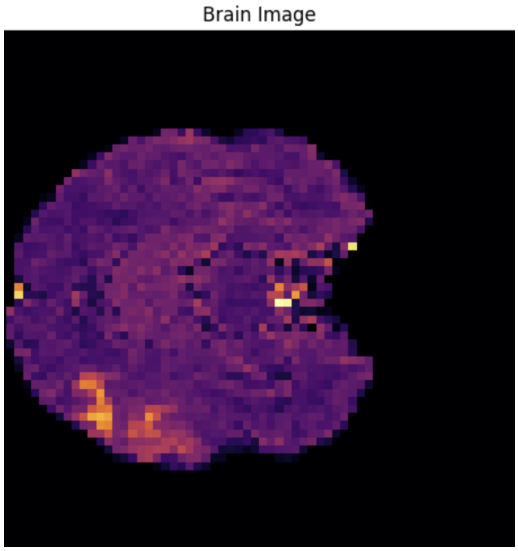


Figure 6.9. *Original brain MRI*

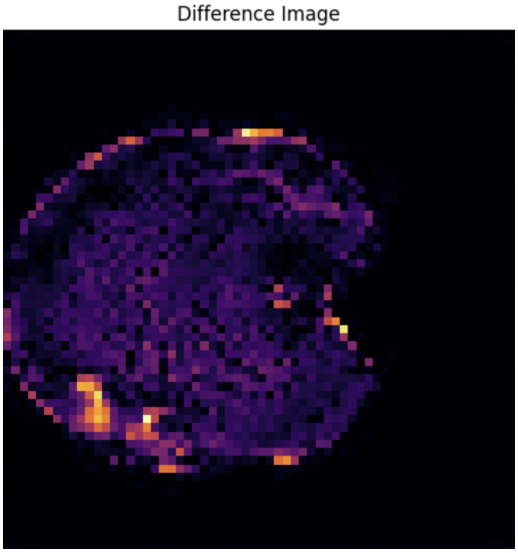


Figure 6.10. *Difference Image*

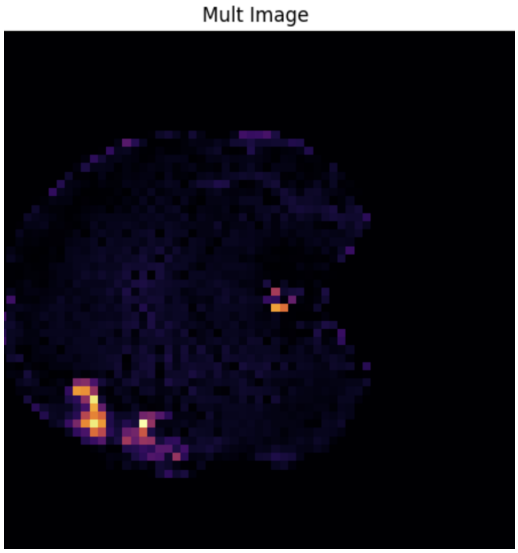


Figure 6.11. *Result of multiplying difference image with the brain MRI image*

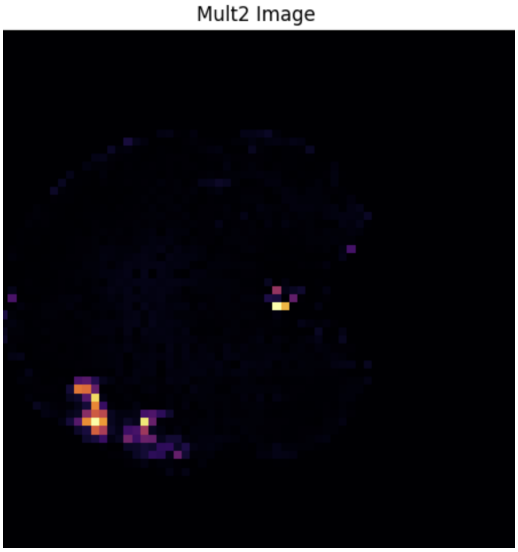


Figure 6.12. *Result of multiplying the difference image with the brain MRI image squared.*

6.3.2 Point-Prompted Segmentation Pipeline

In the first segmentation pipeline, the processed image obtained after the multiplication steps is fed into the Segment Anything Model (SAM). The point used as a prompt for SAM can be provided by a doctor or radiologist. For our testing purposes, we use the centroid of the ground truth mask as the point prompt.

The centroid is found using the following method. First, the input image is binarized to distinguish the white pixels (representing the lesion) from the background. The coordinates of these white pixels are identified, and their centroid is calculated by taking the mean of their positions. This centroid is then rounded to the nearest integer. If the rounded centroid does not correspond to a white pixel, the nearest white pixel to the centroid is selected.

Upon receiving the point prompt, SAM generates several candidate masks. These masks are initially sorted by their area, with the smallest mask being selected as the primary candidate. This approach is based on the observation that SAM tends to produce larger masks with lower confidence. Consequently, the smallest mask is more likely to be accurate.

To further refine the selection, we implement an additional check. If the area of the second smallest mask is less than three times that of the smallest mask, we select the second smallest mask instead. This heuristic addresses potential information loss during the preprocessing stage and ensures that the chosen mask more comprehensively captures the lesion. This method has been shown to yield better results in our experiments, effectively balancing precision and comprehensiveness in lesion segmentation.

In summary, the point-prompted segmentation pipeline leverages the precise location information provided by the centroid of the ground truth mask, combined with a robust mask selection process, to achieve accurate and reliable lesion segmentation.

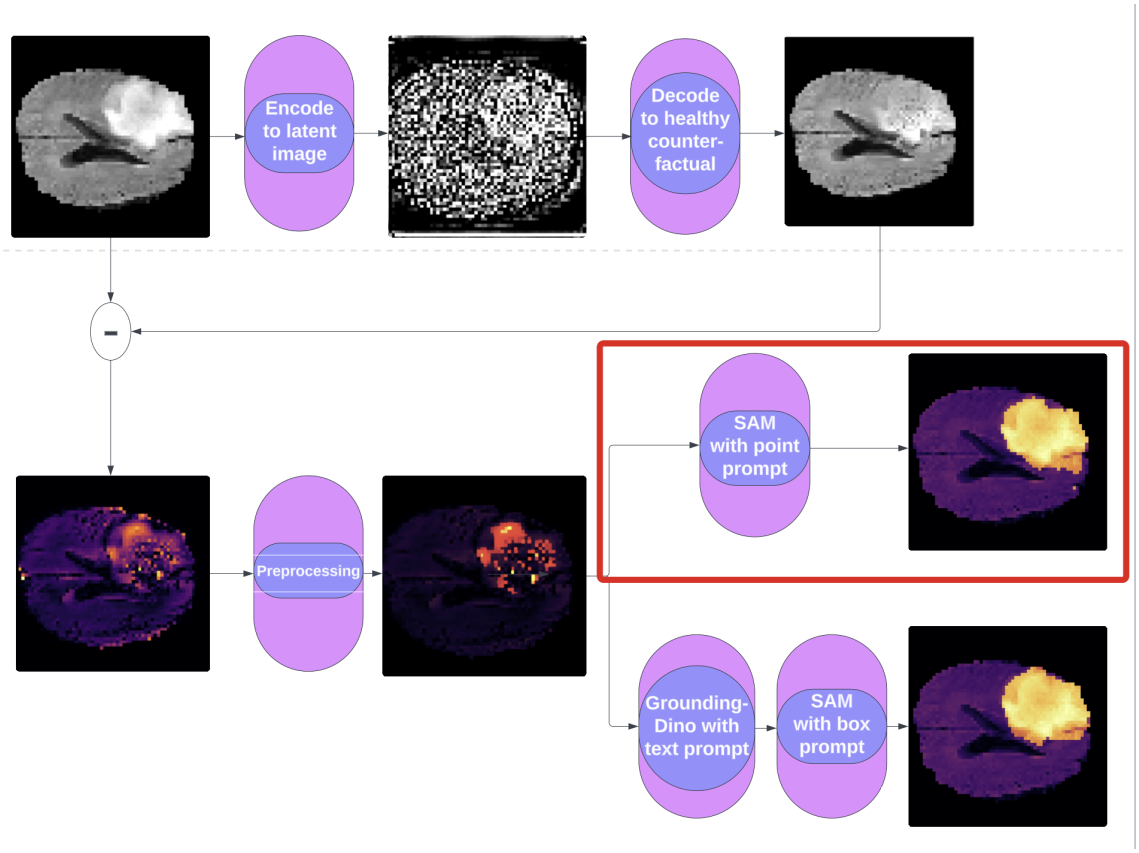


Figure 6.13. *Point-Prompted Pipeline*

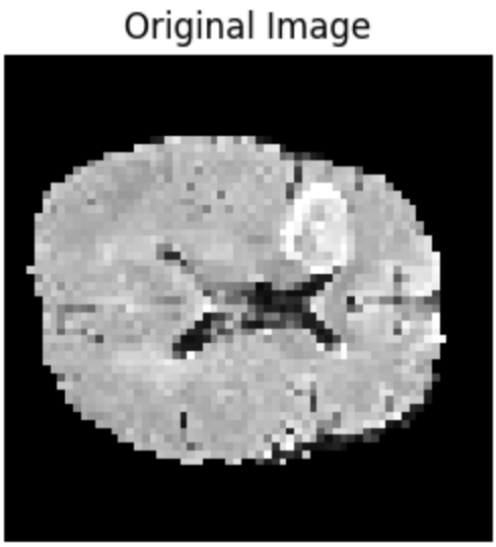


Figure 6.14. *Original brain MRI*

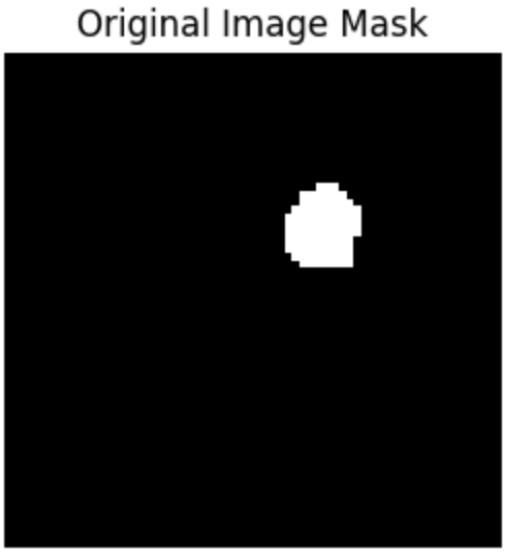


Figure 6.15. *Ground Truth Mask*

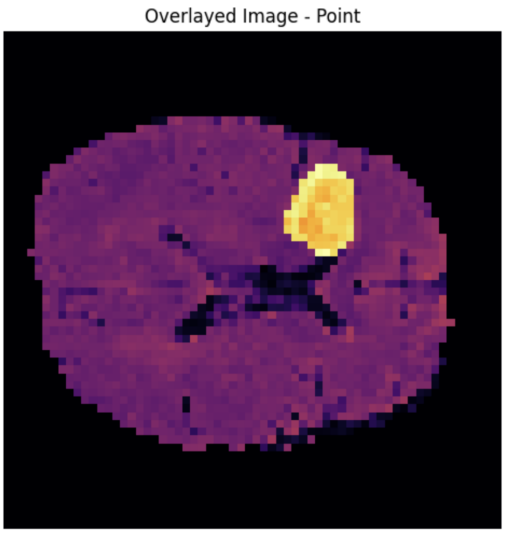


Figure 6.16. *Original brain MRI overlayed with predicted mask*

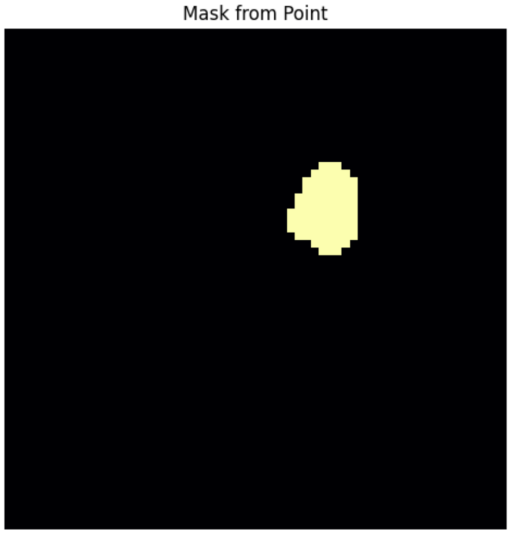


Figure 6.17. *Mask from Point-Prompted Segmentation Pipeline*

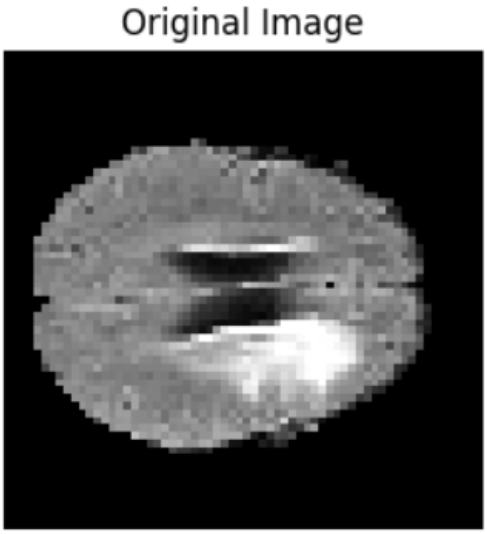


Figure 6.18. *Original brain MRI*

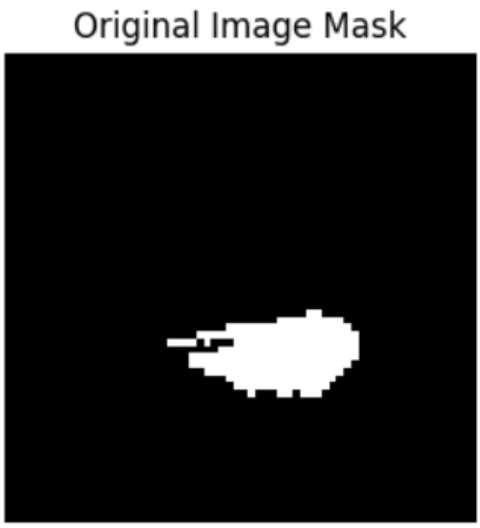


Figure 6.19. *Ground Truth Mask*

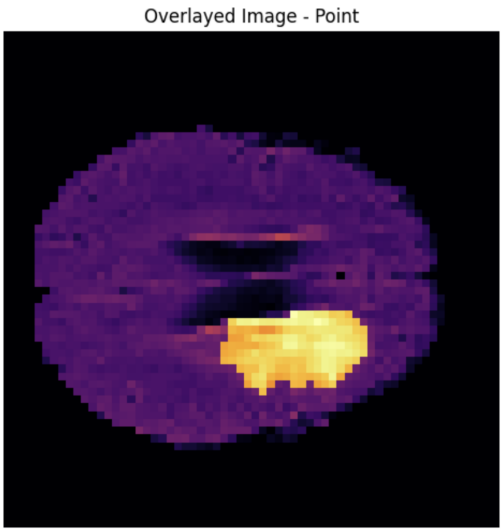


Figure 6.20. *Original brain MRI overlayed with predicted mask*

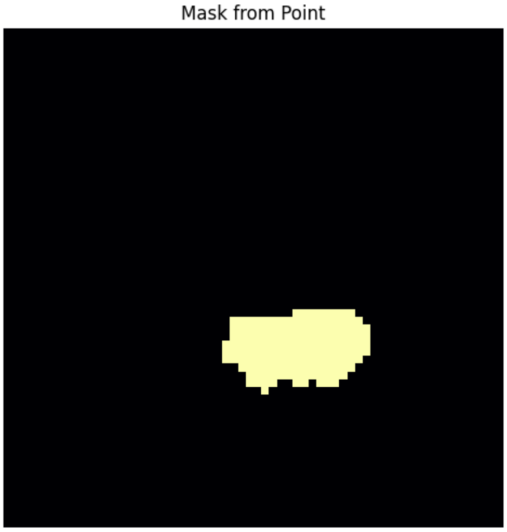


Figure 6.21. *Mask from Point-Prompted Segmentation Pipeline*

6.3.3 Text-Prompted Segmentation Pipeline

In the second segmentation pipeline, a text prompt provided by a doctor or radiologist is utilized. This text prompt is input into the GroundingDINO model, which performs object detection and returns several bounding boxes corresponding to the prompt. Similar to the previous pipeline, we select the bounding box with the smallest area, as GroundingDINO tends to produce larger boxes with lower confidence.

Once the smallest bounding box is identified, it serves as the input prompt for the Segment Anything Model (SAM). SAM then performs segmentation based on this prompt, either on the original difference image or one of the multiplied versions generated during preprocessing. The choice of which image to use for segmentation is a hyperparameter of our framework, allowing for flexibility based on the specific requirements of the task and the quality of the difference image.

This pipeline leverages the strength of GroundingDINO in accurately detecting objects based on textual descriptions and the precision of SAM in segmenting the identified regions. By integrating these models, we achieve a robust and flexible approach to lesion segmentation, adaptable to various input conditions and prompts. This method enhances the accuracy and reliability of segmentation, providing valuable insights for clinical applications.

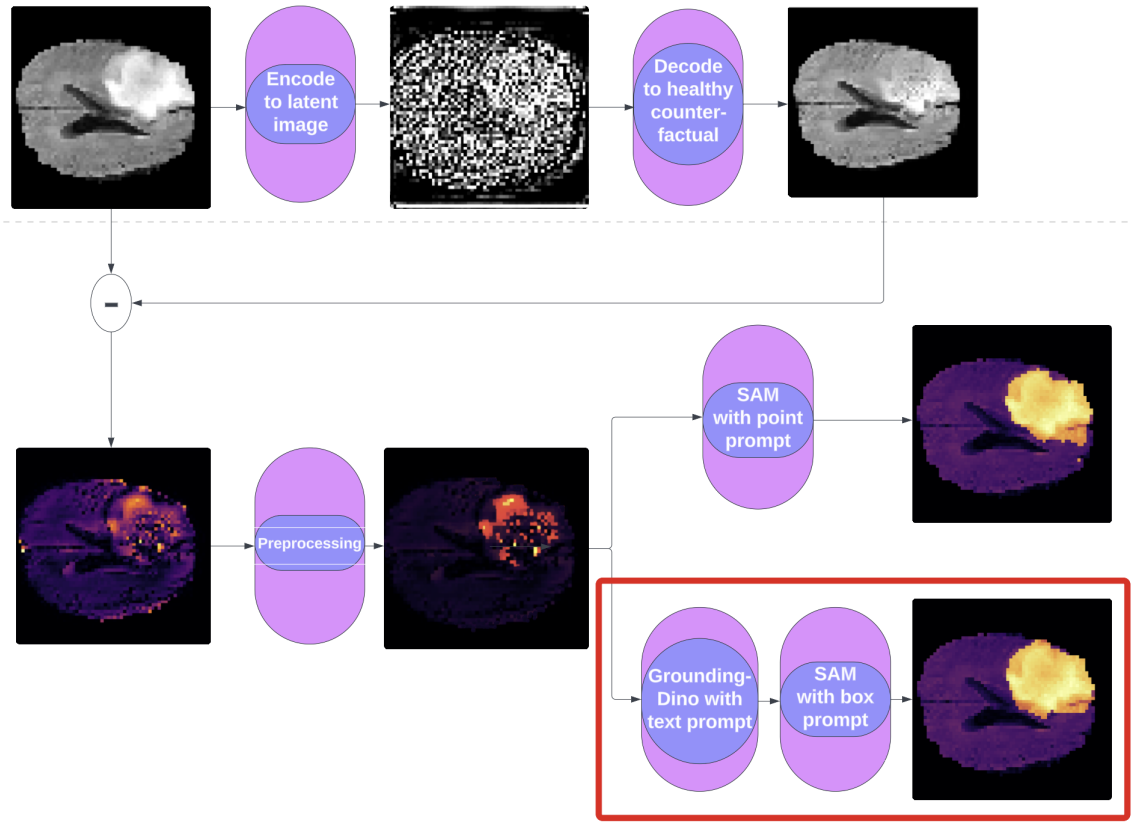


Figure 6.22. *Text-Prompted Pipeline*

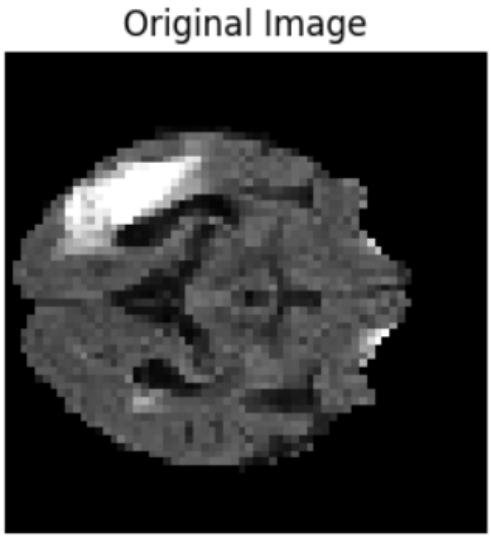


Figure 6.23. *Original brain MRI*

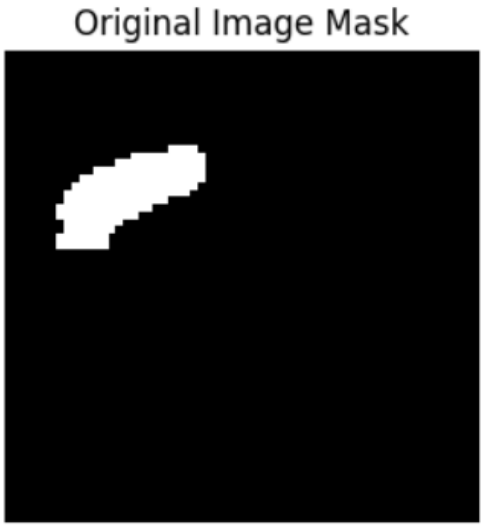


Figure 6.24. *Ground Truth Mask*

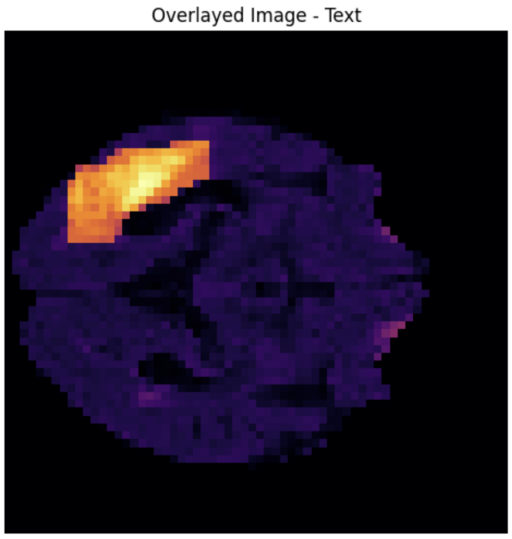


Figure 6.25. *Original brain MRI overlayed with predicted mask*

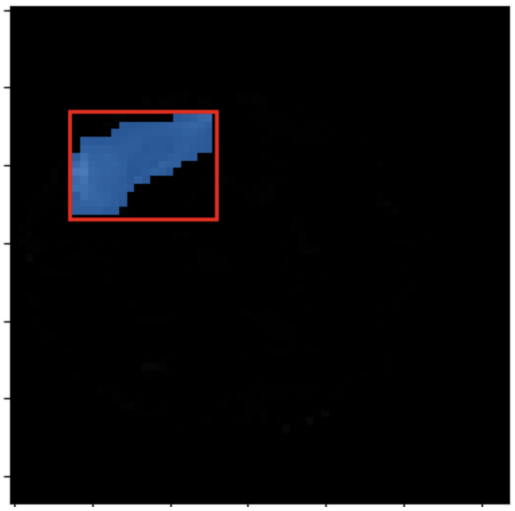


Figure 6.26. *Lesion detection and the corresponding box returned by Grounding DINO*

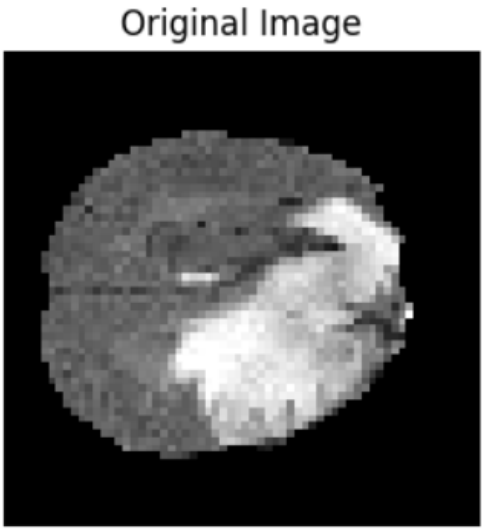


Figure 6.27. *Original brain MRI*

Figure 6.28. *Ground Truth Mask*

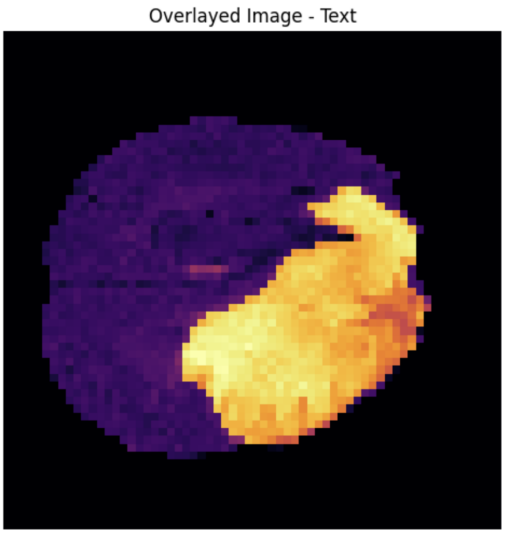


Figure 6.29. *Original brain MRI overlayed with predicted mask*

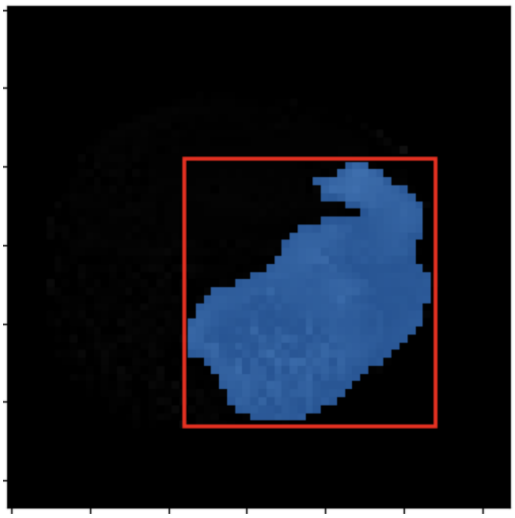


Figure 6.30. *Lesion detection and the corresponding box returned by Grounding DINO*

6.3.4 Intersection and Union of Generated Masks

Our framework generates four distinct masks for each input MRI image: the mask from the point-prompted pipeline, the mask from the text-prompted pipeline, and two additional masks derived from the intersection and union of the initial two masks. The rationale behind generating these combined masks is to leverage the strengths of both segmentation approaches. In medical diagnostics, it is generally preferable to err on the side of caution, meaning a false positive (an incorrect identification of a lesion) is less detrimental than a false negative (a missed lesion). By considering both the intersection and union of the masks, we ensure comprehensive coverage of the lesion, thus minimizing the risk of missing any pathological regions.

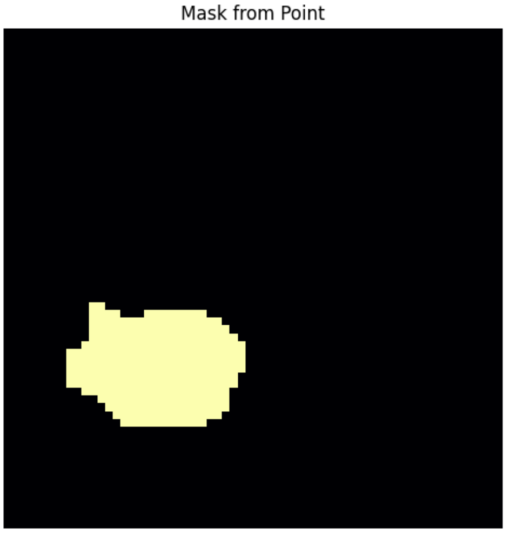


Figure 6.31. *Mask generated by Point-Prompted Pipeline*

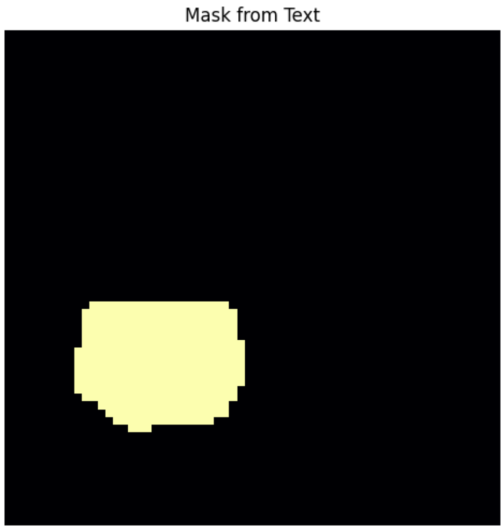


Figure 6.32. *Mask generated by Text-Prompted Pipeline*

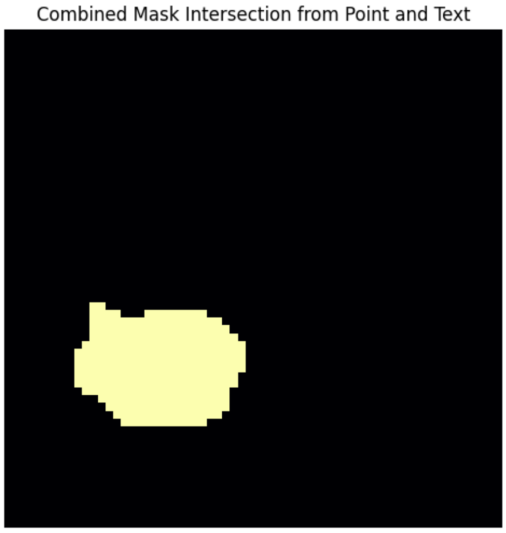


Figure 6.33. *Intersection of generated masks*

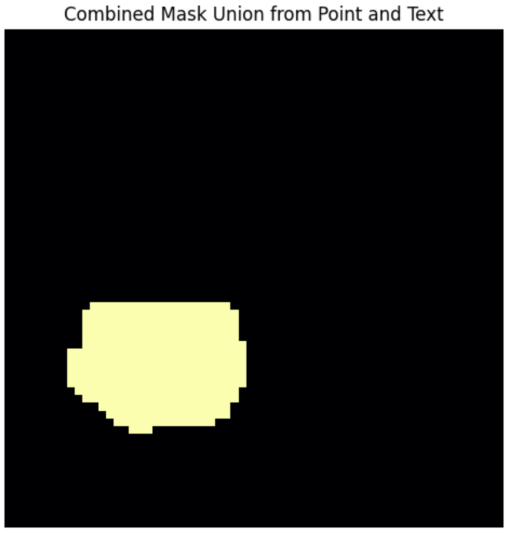


Figure 6.34. *Union of generated masks*

6.4 Evaluation Metrics

To evaluate the performance of our segmentation framework, we employ a variety of metrics. Each metric offers a different perspective on the accuracy and reliability of the segmentation results, and their combined use provides a comprehensive evaluation.

6.4.1 Dice Coefficient

The Dice coefficient, or Dice similarity index, measures the overlap between the predicted mask and the ground truth. It is calculated as:

$$Dice = \frac{2 \times |A \cap B|}{|A| + |B|} \tag{6.2}$$

where A is the predicted mask and B is the ground truth mask. A Dice score of 1 indicates perfect overlap, while a score of 0 indicates no overlap.

6.4.2 Intersection over Union (IoU)

IoU, also known as the Jaccard index, quantifies the ratio of the intersection to the union of the predicted and ground truth masks. It is given by:

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{6.3}$$

IoU values range from 0 to 1, with higher values indicating better segmentation accuracy.

6.4.3 Area Under the Precision-Recall Curve (AUPRC)

AUPRC evaluates the trade-off between precision and recall across different thresholds. It is particularly useful for imbalanced datasets. A higher AUPRC value indicates better model performance in distinguishing between classes.

6.4.4 Precision

Precision measures the proportion of true positives among all positive predictions. It is calculated as:

$$Precision = \frac{TP}{TP + FP} \tag{6.4}$$

where TP is the number of true positives and FP is the number of false positives.

6.4.5 Recall

Recall, or sensitivity, measures the proportion of true positives among all actual positives. It is given by:

$$Recall = \frac{TP}{TP + FN} \tag{6.5}$$

where FN is the number of false negatives.

6.4.6 Specificity

Specificity measures the proportion of true negatives among all actual negatives. It is calculated as:

$$Specificity = \frac{TN}{TN + FP} \tag{6.6}$$

where TN is the number of true negatives.

6.4.7 F1 Score

The F1 score is the harmonic mean of precision and recall, providing a single metric that balances both:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6.7}$$

It is particularly useful when precision and recall are both important.

6.4.8 Hausdorff Distance

The Hausdorff distance measures the maximum distance between the boundary points of the predicted and ground truth masks. It is defined as:

$$Hausdorff(A, B) = max\{sup_{a \in A} inf_{b \in B} d(a, b), sup_{b \in B} inf_{a \in A} d(a, b)\} \tag{6.8}$$

where the function d(*,*) represents the Euclidean distance.

This metric is crucial in medical imaging as it highlights the worst-case discrepancy between boundaries, ensuring that the segmentation captures the true extent of the lesion.

6.4.9 Average Symmetric Surface Distance (ASSD)

ASSD measures the average distance between the boundary points of the predicted and ground truth masks. It is calculated by averaging the distances from each point on one boundary to the nearest point on the other boundary, and vice versa. ASSD is significant in medical contexts because it provides a more holistic measure of boundary accuracy, reducing the impact of outliers compared to the Hausdorff distance.

These metrics collectively provide a robust framework for evaluating the effectiveness of our lesion segmentation approach, ensuring that both accuracy and clinical relevance are thoroughly assessed.

6.5 Experimental Setup

6.5.1 Dataset

For the experiments conducted in this thesis, we utilized the Decathlon Task 1 dataset, specifically focusing on brain tumors. This dataset is part of the Medical Segmentation

Decathlon, a comprehensive collection of annotated medical imaging datasets designed to benchmark the performance of segmentation algorithms.

This data comprises magnetic resonance (MR) imaging from four sequences T1, post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (FLAIR) for each patient. T. For our experiments, we specifically used the FLAIR modality, which is particularly effective in highlighting abnormalities such as tumors.

The dataset consists of 388, 96 and 251 brain MRIs for training, validation, and testing respectively.

6.5.2 Data Transformations

In preparing our dataset for training, we applied several crucial transformations to the brain MRI images to ensure they are standardized and suitable for model input. Initially, the images and their corresponding labels were loaded into memory to ensure the data was correctly retrieved for subsequent processing. Following this, the channel dimension of the images and labels was adjusted to be the first dimension, aligning with the expectations of many deep learning frameworks.

To standardize the orientation of the images, we reoriented the data to the RAS (Right-Anterior-Superior) standard. This was followed by resampling the images to a uniform voxel spacing, ensuring consistent spatial resolution across all samples.

Next, the images and labels were center-cropped to a fixed region of interest, focusing on the most relevant parts of the brain. To enhance the contrast and brightness consistency of the images, we scaled the intensity values based on specified percentiles, normalizing the images to a range between 0 and 1.

For data augmentation, we performed random spatial cropping, providing different image crops during training to improve the model's generalization capabilities. Finally, we ensured that the resultant data structures were compatible for further processing.

These preprocessing steps were essential in preparing the brain MRI images for training, enhancing the robustness and accuracy of our segmentation models by ensuring that the input data was well-standardized and augmented.

6.5.3 U-Net Architecture

The model employed in our framework is based on a U-Net architecture, tailored specifically for diffusion models. The U-Net model, a convolutional neural network (CNN) known for its effectiveness in image segmentation, features an encoder-decoder structure with skip connections to preserve spatial information.

The encoder of the U-Net model progressively reduces the spatial dimensions of the input image while increasing its depth to extract high-level features. Our encoder processes the images through four stages with increasing channel widths: 64, 128, 256, and 512 channels. Each stage comprises two residual blocks, enhancing the network's ability to learn complex patterns.

The decoder then reconstructs the spatial dimensions while reducing the depth, using up-sampling and convolutional layers. Skip connections link corresponding encoder and decoder layers, ensuring that spatial details are retained and incorporated into the final segmentation output.

Attention mechanisms are integrated at specific levels within the network to enhance feature focus. In our implementation, attention layers are added at the second, third, and fourth levels of the network. This allows the model to concentrate on significant regions of the image that are crucial for accurate segmentation. The attention mechanism is further refined with cross-attention, which integrates additional contextual information provided through conditioning.

Conditioning is implemented to incorporate external information, such as class labels, into the segmentation process. This is achieved using an embedding layer with a dimension of 64, which projects the conditioning information into a space that can be effectively combined with the image features. The conditioning information is integrated into the model via cross-attention, ensuring that it influences the segmentation process at multiple levels.

Our U-Net model is integrated within a diffusion framework, utilizing a scheduler to manage the diffusion process. The model is configured to process 2D spatial dimensions, with one input and one output channel. The diffusion process involves iterative steps where noise is added and then progressively removed from the input image, guided by the model parameters. The DiffusionInferer orchestrates this process during inference, ensuring high-quality segmentation results from noisy inputs.

Training the model involves the Adam optimizer, which adjusts the learning rate dynamically to ensure efficient convergence. Both the U-Net model parameters and the embedding layer parameters are optimized concurrently, with a learning rate set to 1×10^{-5}. This allows for precise fine-tuning of the complex parameters within the network.

This detailed U-Net architecture, augmented with targeted attention mechanisms and conditioning, forms the backbone of our segmentation framework. It provides a robust and precise solution for accurately segmenting lesions in brain MRI images.

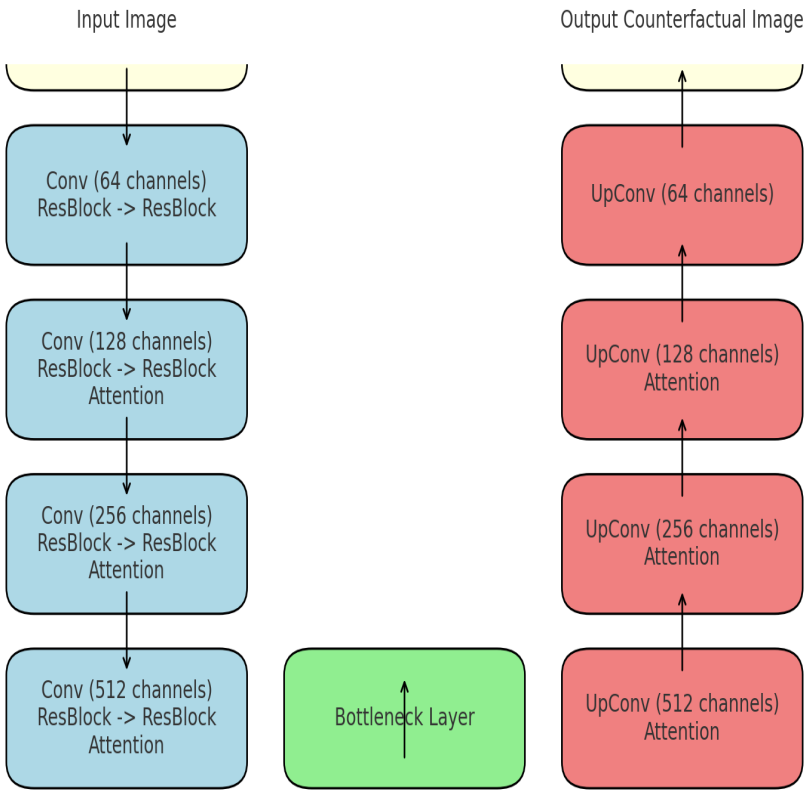


Figure 6.35. *Approximate sketch of our U-Net's architecture. Source: Image by author.*

6.5.4 Training

The training process for our model involves several key parameters and procedures to ensure efficient and effective learning.

Condition Dropout: A dropout rate of 0.15 is used for conditioning to prevent overfitting by randomly omitting conditioning information during training.

Iterations and Batch Size: The model is trained for 4000 iterations with a batch size of 32, balancing computational efficiency and gradient estimation accuracy.

Validation Interval: Validation is performed every 100 iterations to monitor the model's performance on unseen data and adjust hyperparameters as needed.

Data Loading:

- The **training dataset** is loaded using a DataLoader with batch size 32, shuffling, and 4 worker threads to ensure efficient data retrieval.
- The **validation dataset** is similarly loaded but without shuffling to maintain data order during evaluation.

Gradient Scaling: Gradient scaling is applied using GradScaler to manage mixed precision training, optimizing computational resources and training speed.

Performance Tracking: Iterative loss values for both training and validation are recorded at each iteration to track the model's learning progress and convergence. This structured training approach, combining dropout for regularization, frequent validation, efficient data loading, and gradient scaling, ensures that the model learns effectively while minimizing overfitting and computational overhead.

6.5.5 Hyperparameters

Our system's performance is influenced by several key hyperparameters. These include the text prompt used in the text-prompted segmentation pipeline, the number of multiplication steps in the preprocessing phase, and parameters within the point-prompted segmentation pipeline.

Text Prompt (TP): The specific text input used in the text-prompted segmentation pipeline, provided by a doctor or radiologist, to guide the GroundingDINO model in identifying relevant regions in the MRI images.

Number of Multiplication Steps in the Point-Prompted Pipeline (PMS): This refers to the number of times the difference image is multiplied with the original MRI during preprocessing. This hyperparameter affects the clarity and contrast of the lesion area,

which is critical for accurate segmentation in the Point-Prompted Segmentation Pipeline.

Number of Multiplication Steps in the Text-Prompted Pipeline (TMS): This refers to the number of times the difference image is multiplied with the original MRI during preprocessing. This hyperparameter affects the clarity and contrast of the lesion area, which is critical for accurate segmentation in the Text-Prompted Segmentation Pipeline.

We will conduct experiments with different combinations of these hyperparameters to evaluate their impact on our performance metrics, such as Dice score, IoU, AUPRC, precision, recall, specificity, F1 score, Hausdorff distance, and ASSD. This will help us optimize the framework for the best possible segmentation accuracy.

Some of the combinations we are going to experiment with are listed below:

- **TP** = "lesion", **PMS** = 2, **TMS** = 3
- **TP** = "tumour", **PMS** = 1, **TMS** = 2
- **TP** = "anomaly", **PMS** = 3, **TMS** = 3
- **TP** = "lesion", **PMS** = 2, **TMS** = 2

Chapter 7

Experimental Results

Having described our framework in detail in the previous chapter, we now present the results obtained using our system on Task 1 (Brain Tumor) of the Medical Segmentation Decathlon challenge. This chapter will provide a comprehensive overview of the performance of our framework across various hyperparameter configurations.

The results will be presented in the form of tables, with each table corresponding to a unique hyperparameter configuration. The columns in each table will represent the different methods used to generate masks in our framework: Point, Text, Intersection, and Union. The rows will denote the evaluation metrics used to assess the performance of our framework. These metrics include Dice score, IoU, AUPRC, precision, recall, specificity, F1 score, Hausdorff distance, and ASSD. Each metric provides a different perspective on the accuracy and reliability of our segmentation results.

The rows in the tables will denote the evaluation metrics used to assess the performance of our framework. These metrics include Dice score, IoU, AUPRC, precision, recall, specificity, F1 score, Hausdorff distance, and ASSD. Each metric provides a different perspective on the accuracy and reliability of our segmentation results. Each one of our framework's hyper-parameter configurations will be presented in each own subsection in the form of ***"TP - PMS - TMS"***.

In addition to the quantitative results, several images obtained from our experiments will be provided. These images will illustrate the segmentation performance of our framework under different hyperparameter settings, offering visual insight into the effectiveness of our approach.

Furthermore, we will present the results of a short ablation study to investigate the contribution of different components of our framework. In this study, we experiment with various U-Net architectures, omit the preprocessing steps altogether, and even apply the SAM-GroundingDINO pipeline directly to the original MRI images instead of the counterfactuals. This ablation study aims to highlight the importance of each component and the impact of different architectural choices on the performance metrics.

Through this detailed presentation of results, we aim to demonstrate the robustness and flexibility of our framework in accurately segmenting brain tumors in MRI images, while highlighting the influence of different hyperparameters and architectural choices on the performance metrics.

7.1 Lesion - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.821	0.804	0.823
AUPRC	0.761	0.848	0.848	0.785
IoU	0.606	0.731	0.709	0.632
Precision	0.747	0.853	0.942	0.706
Recall	0.762	0.836	0.741	0.857
F1	0.755	0.844	0.830	0.775
Specificity	0.987	0.993	0.998	0.982
Hausdorff	5.919	5.624	5.744	5.785
ASSD	0.450	0.398	0.432	0.418

Table 7.1. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration*

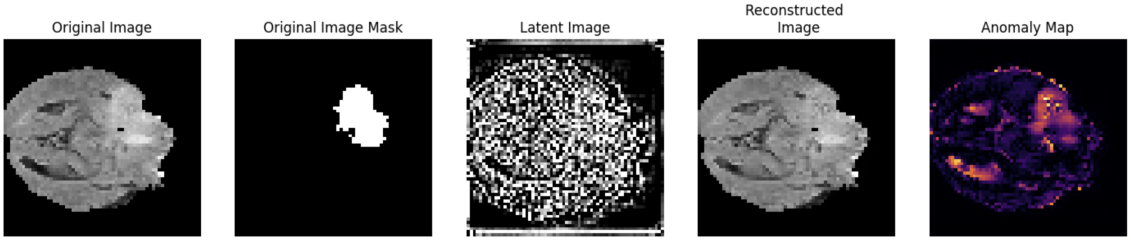


Figure 7.1. *Counterfactual generation*

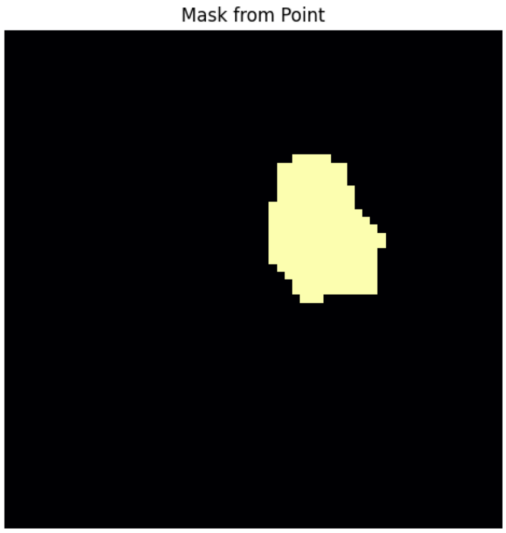


Figure 7.2. *Mask generated by Point-Prompted Pipeline*

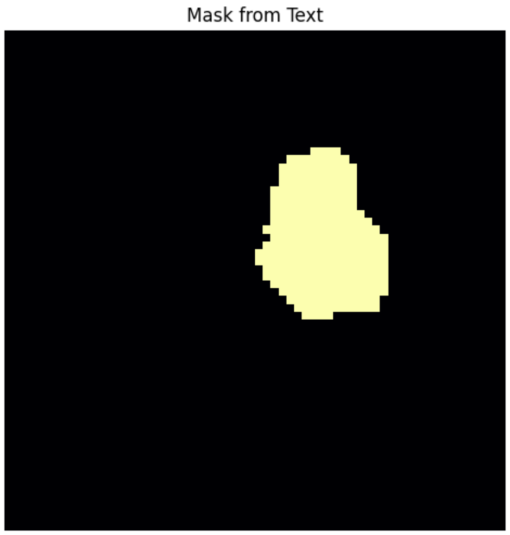


Figure 7.3. *Mask generated by Text-Prompted Pipeline*

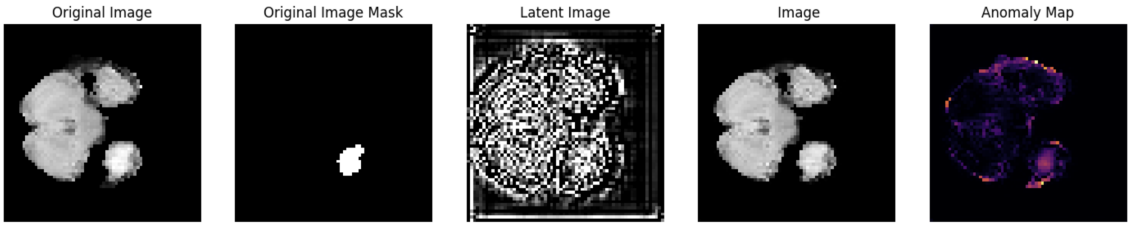


Figure 7.4. *Counterfactual generation*

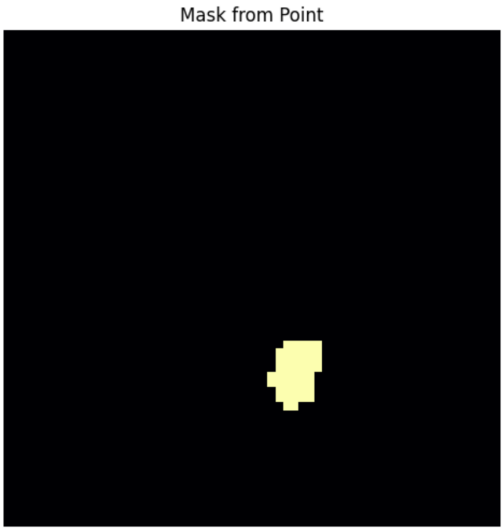


Figure 7.5. *Mask generated by Point-Prompted Pipeline*

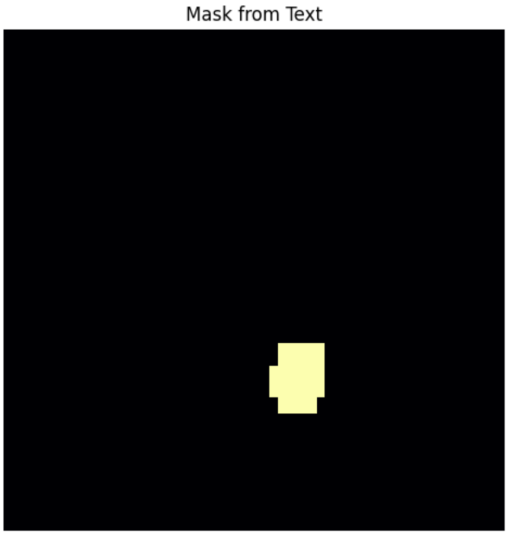


Figure 7.6. *Mask generated by Text-Prompted Pipeline*

7.2 Lesion - 2 - 2

	Point	Text	Intersection	Union
Dice	0.799	0.807	0.814	0.792
AUPRC	0.743	0.762	0.789	0.735
IoU	0.580	0.588	0.645	0.539
Precision	0.683	0.640	0.780	0.580
Recall	0.793	0.879	0.788	0.883
F1	0.734	0.741	0.784	0.700
Specificity	0.981	0.974	0.989	0.967
Hausdorff	6.670	6.149	5.710	7.077
ASSD	0.531	0.506	0.429	0.609

Table 7.2. *Evaluation Metrics for the "Lesion - 2 - 2" Configuration*

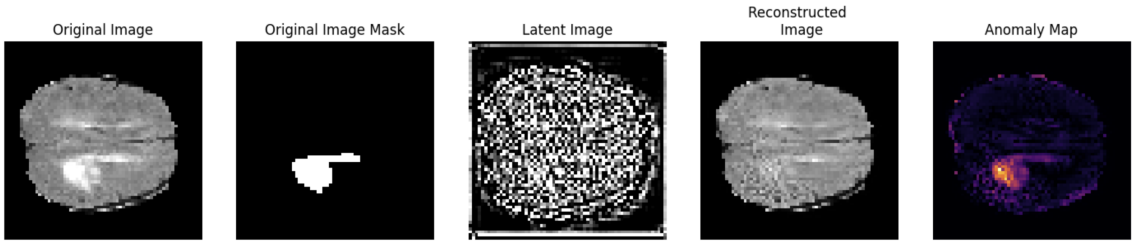


Figure 7.7. *Counterfactual generation*

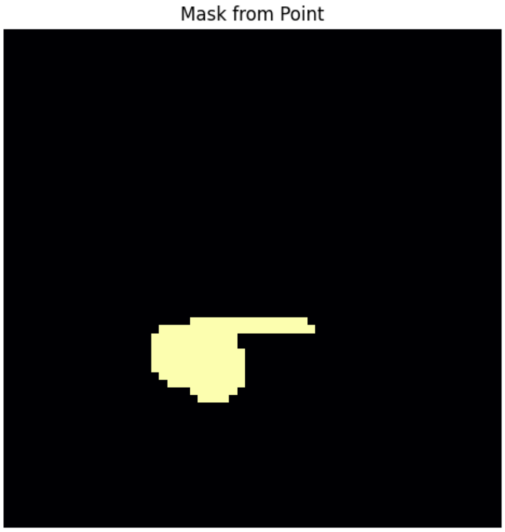


Figure 7.8. *Mask generated by Point-Prompted Pipeline*

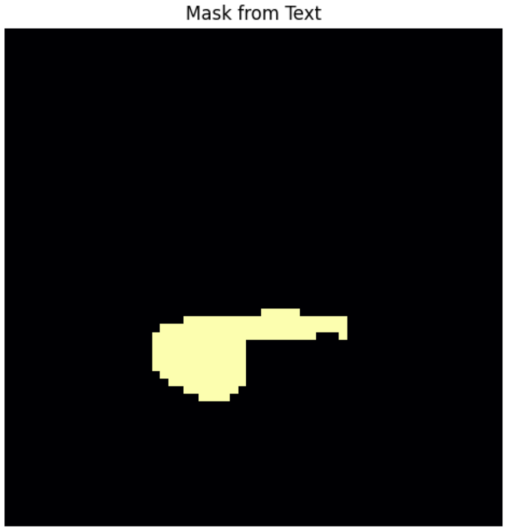


Figure 7.9. *Mask generated by Text-Prompted Pipeline*

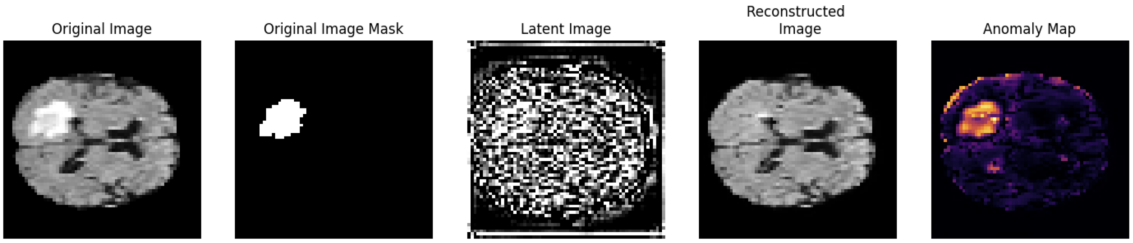


Figure 7.10. *Counterfactual generation*

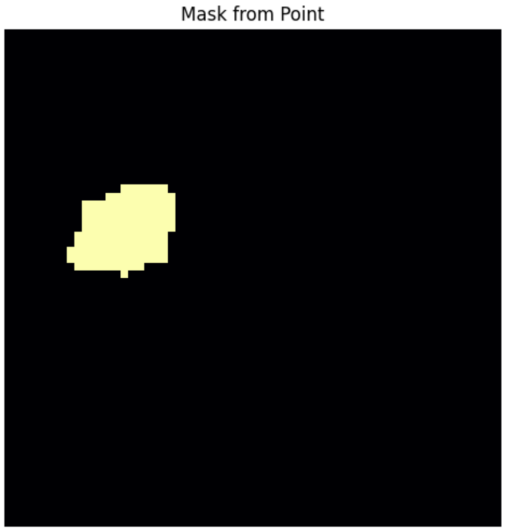


Figure 7.11. *Mask generated by Point-Prompted Pipeline*

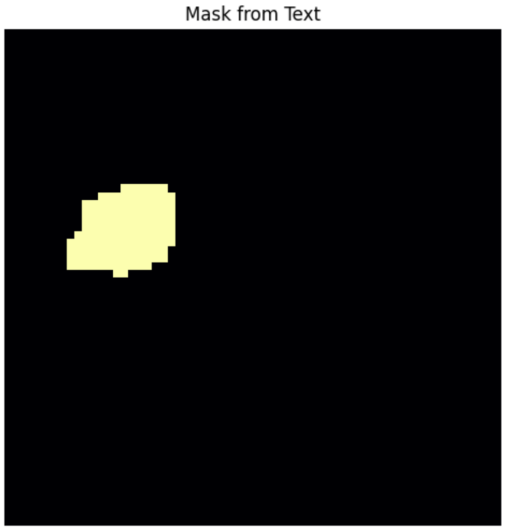


Figure 7.12. *Mask generated by Text-Prompted Pipeline*

7.3 Lesion - 1 - 1

	Point	Text	Intersection	Union
Dice	0.774	0.617	0.775	0.616
AUPRC	0.734	0.586	0.735	0.586
IoU	0.566	0.251	0.568	0.251
Precision	0.662	0.257	0.665	0.257
Recall	0.795	0.910	0.794	0.910
F1	0.723	0.401	0.724	0.401
Specificity	0.979	0.864	0.980	0.864
Hausdorff	7.177	14.482	7.043	14.616
ASSD	0.657	2.332	0.651	2.337

Table 7.3. *Evaluation Metrics for the "Lesion - 1 - 1" Configuration*

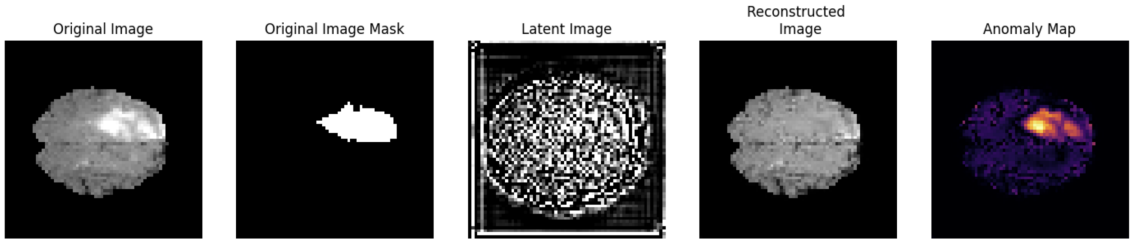


Figure 7.13. *Counterfactual generation*

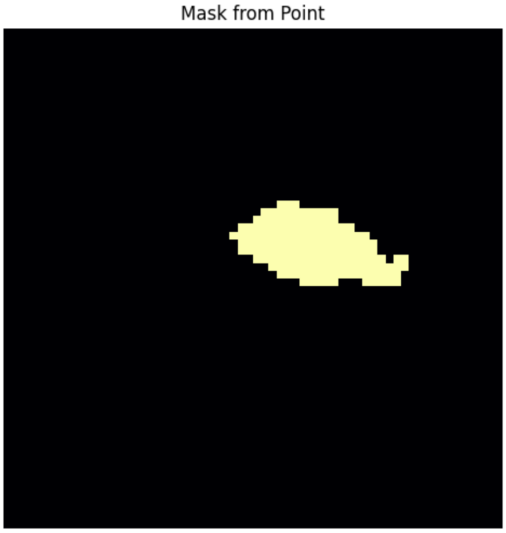


Figure 7.14. *Mask generated by Point-Prompted Pipeline*

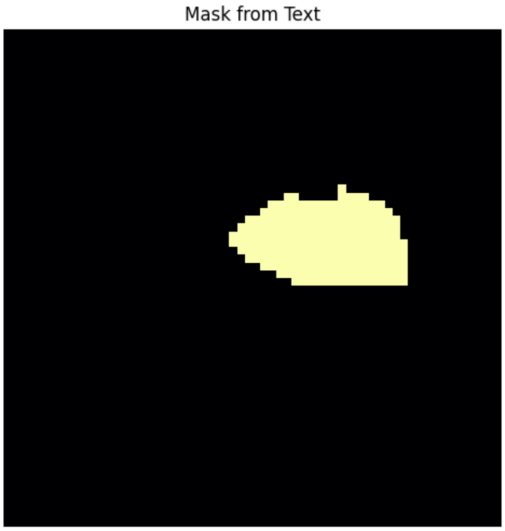


Figure 7.15. *Mask generated by Text-Prompted Pipeline*

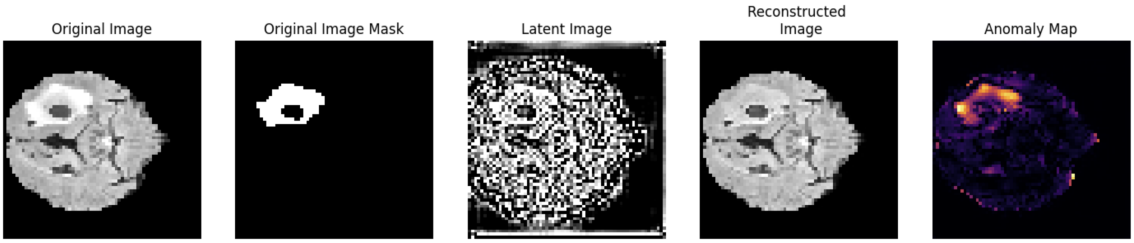


Figure 7.16. *Counterfactual generation*

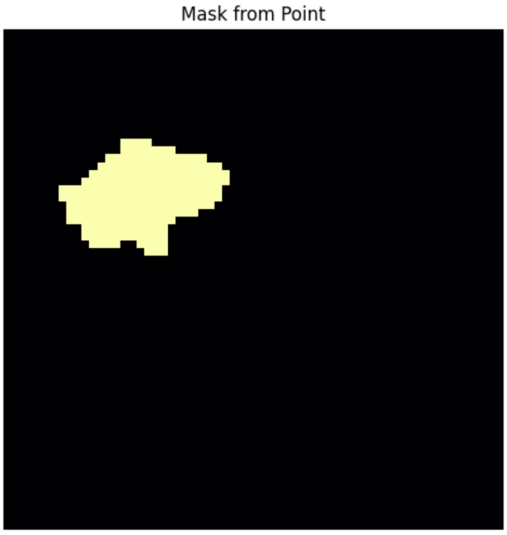


Figure 7.17. *Mask generated by Point-Prompted Pipeline*

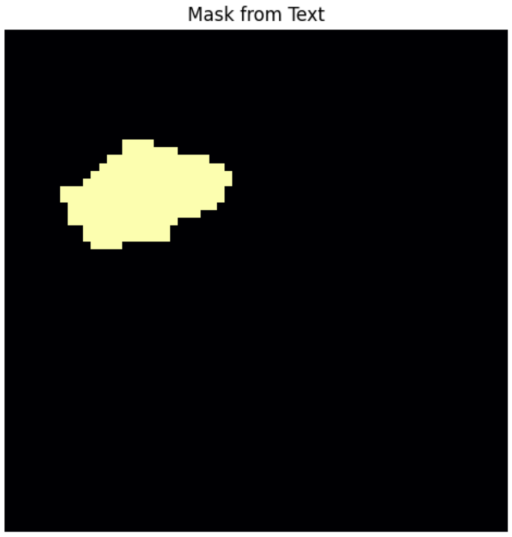


Figure 7.18. *Mask generated by Text-Prompted Pipeline*

7.4 Lesion - 1 - 2

	Point	Text	Intersection	Union
Dice	0.774	0.807	0.797	0.785
AUPRC	0.734	0.762	0.778	0.734
IoU	0.566	0.588	0.629	0.538
Precision	0.662	0.640	0.754	0.579
Recall	0.795	0.879	0.792	0.882
F1	0.723	0.740	0.772	0.699
Specificity	0.979	0.974	0.987	0.967
Hausdorff	7.177	6.149	6.127	7.192
ASSD	0.657	0.506	0.472	0.690

Table 7.4. *Evaluation Metrics for the "Lesion - 1 - 2" Configuration*

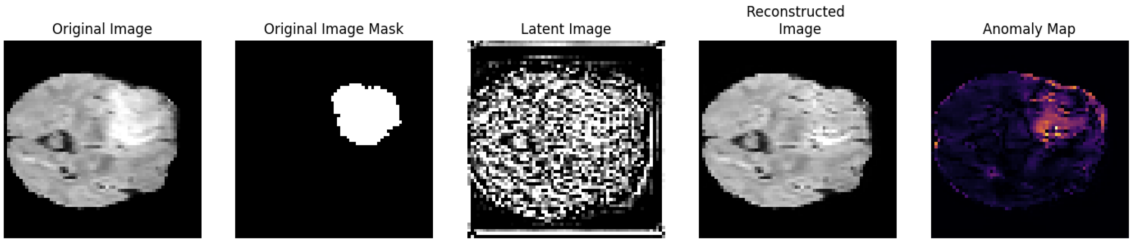


Figure 7.19. *Counterfactual generation*

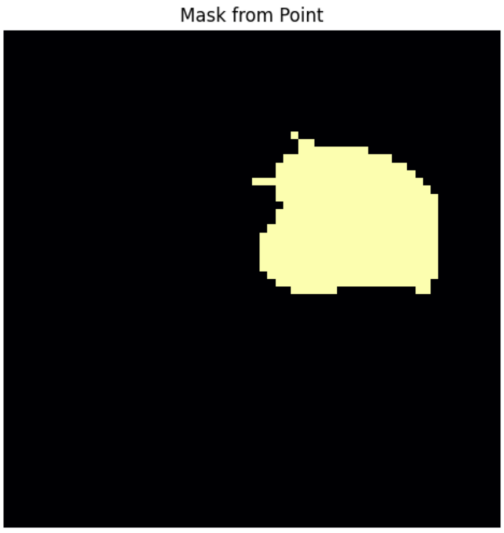


Figure 7.20. *Mask generated by Point-Prompted Pipeline*

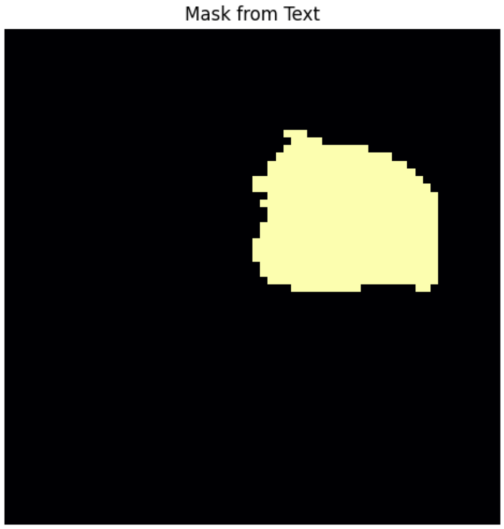


Figure 7.21. *Mask generated by Text-Prompted Pipeline*

Figure 7.22. *Counterfactual generation*

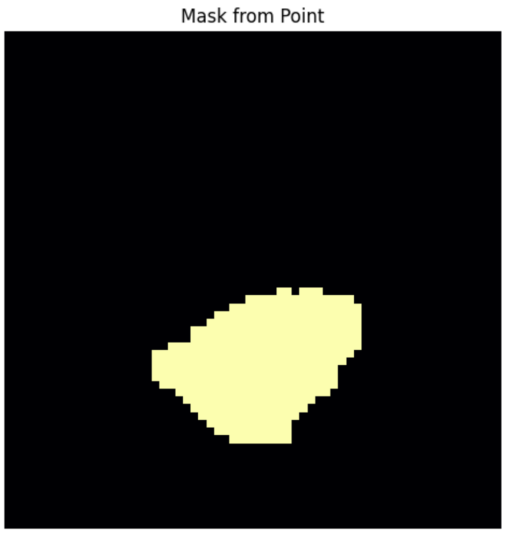


Figure 7.23. *Mask generated by Point-Prompted Pipeline*

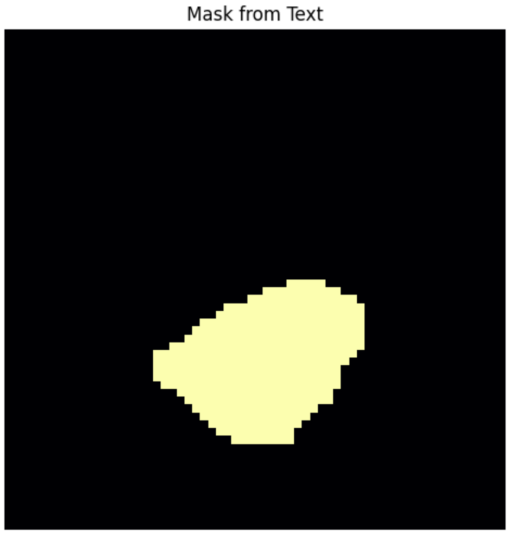


Figure 7.24. *Mask generated by Text-Prompted Pipeline*

7.5 Lesion - 0 - 0

	Point	Text	Intersection	Union
Dice	0.676	0.305	0.691	0.290
AUPRC	0.583	0.565	0.596	0.564
IoU	0.328	0.139	0.351	0.136
Precision	0.358	0.139	0.386	0.136
Recall	0.797	0.989	0.797	0.990
F1	0.494	0.244	0.520	0.240
Specificity	0.926	0.684	0.934	0.676
Hausdorff	10.627	25.292	9.817	26.103
ASSD	1.475	4.872	1.327	5.021

Table 7.5. *Evaluation Metrics for the "Lesion - 0 - 0" Configuration*

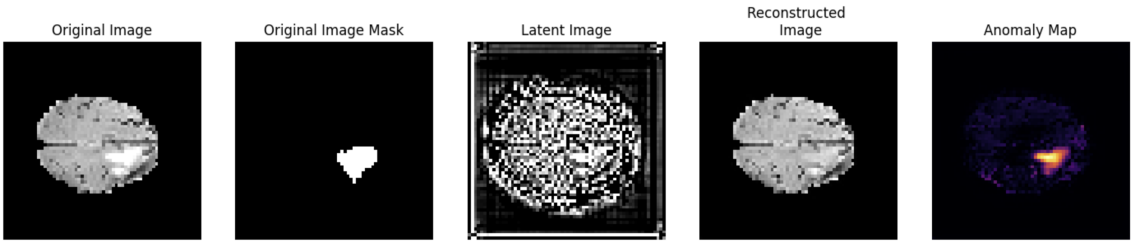


Figure 7.25. *Counterfactual generation*

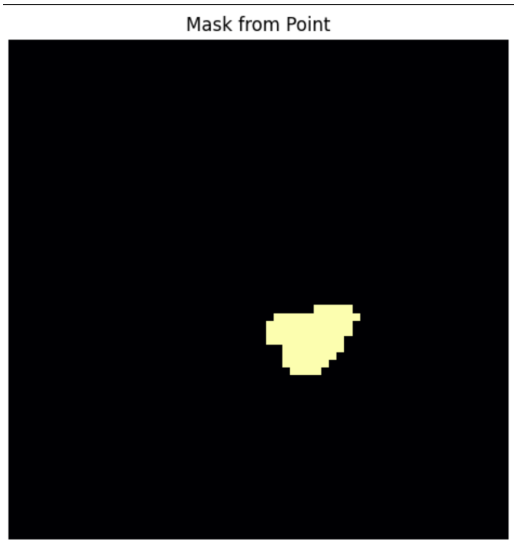


Figure 7.26. *Mask generated by Point-Prompted Pipeline*

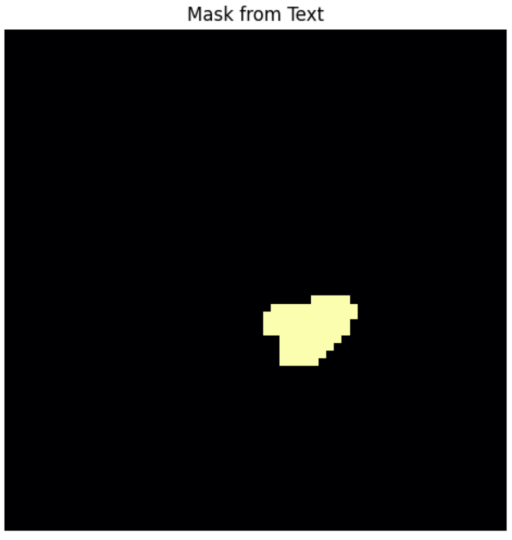


Figure 7.27. *Mask generated by Text-Prompted Pipeline*

Figure 7.28. *Counterfactual generation*

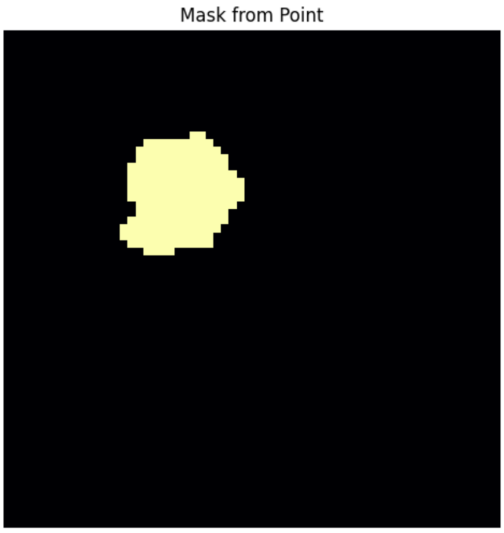


Figure 7.29. *Mask generated by Point-Prompted Pipeline*

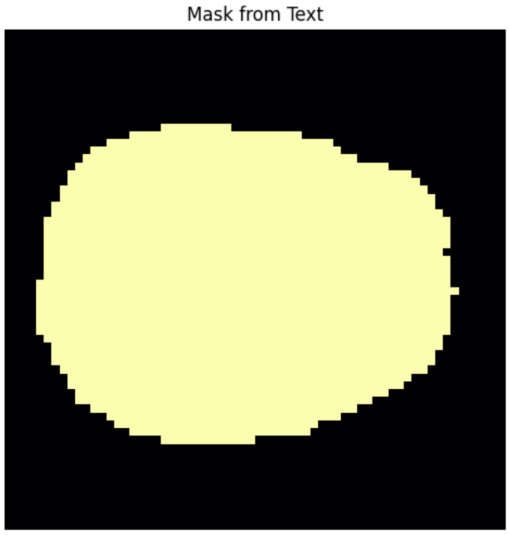


Figure 7.30. *Mask generated by Text-Prompted Pipeline*

7.6 Tumour - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.798	0.794	0.810
AUPRC	0.761	0.724	0.842	0.703
IoU	0.606	0.546	0.697	0.498
Precision	0.747	0.621	0.943	0.544
Recall	0.762	0.819	0.728	0.853
F1	0.755	0.706	0.822	0.664
Specificity	0.987	0.974	0.998	0.963
Hausdorff	5.919	6.665	5.861	6.709
ASSD	0.450	0.618	0.462	0.610

Table 7.6. *Evaluation Metrics for the "Tumour - 3 - 3" Configuration*

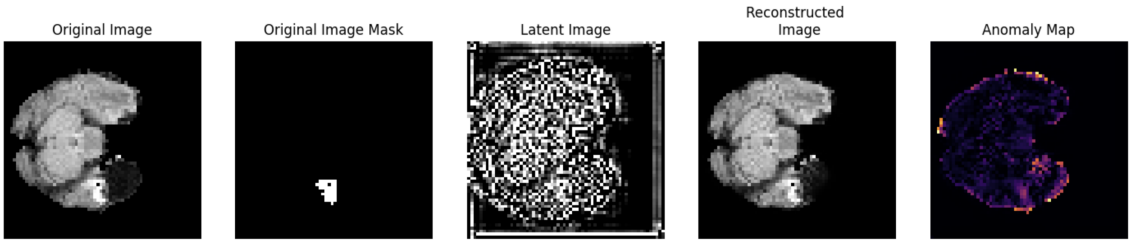


Figure 7.31. *Counterfactual generation*

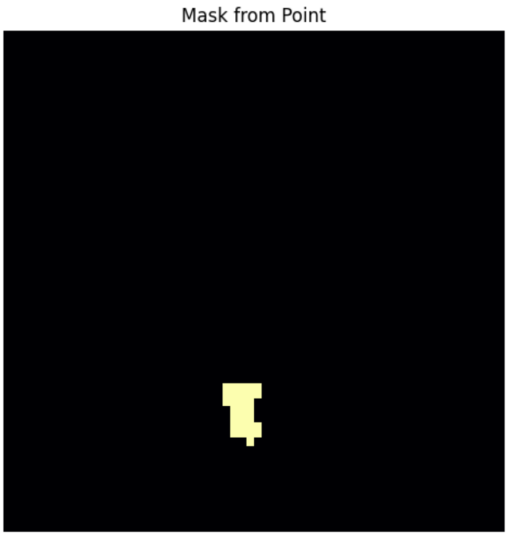


Figure 7.32. *Mask generated by Point-Prompted Pipeline*

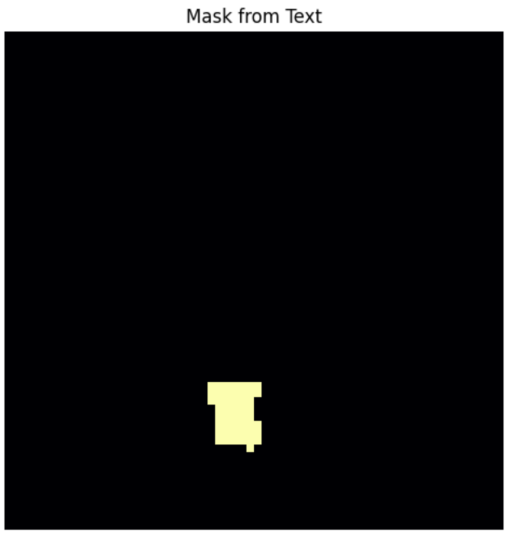


Figure 7.33. *Mask generated by Text-Prompted Pipeline*

Figure 7.34. *Counterfactual generation*

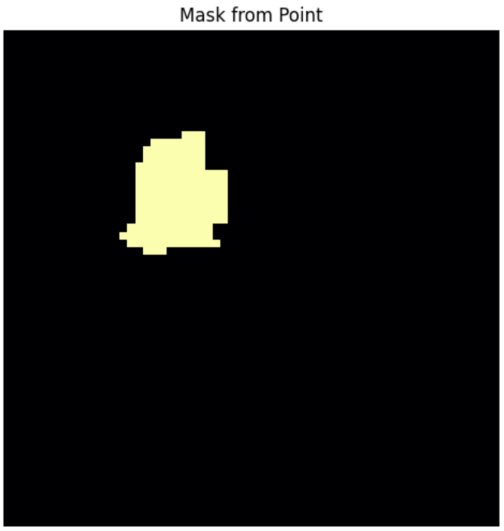


Figure 7.35. *Mask generated by Point-Prompted Pipeline*

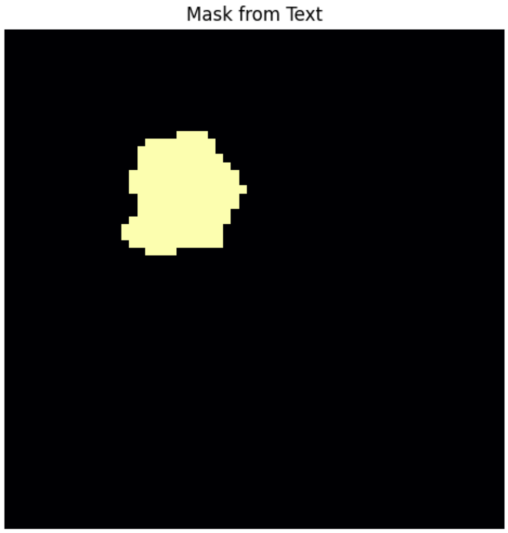


Figure 7.36. *Mask generated by Text-Prompted Pipeline*

7.7 Tumour - 2 - 2

	Point	Text	Intersection	Union
Dice	0.799	0.807	0.813	0.794
AUPRC	0.743	0.794	0.851	0.735
IoU	0.580	0.647	0.721	0.541
Precision	0.683	0.727	0.923	0.586
Recall	0.793	0.853	0.768	0.878
F1	0.734	0.785	0.838	0.702
Specificity	0.981	0.983	0.997	0.968
Hausdorff	6.670	6.045	5.698	6.985
ASSD	0.531	0.485	0.414	0.603

Table 7.7. *Evaluation Metrics for the "Tumour - 2 - 2" Configuration*

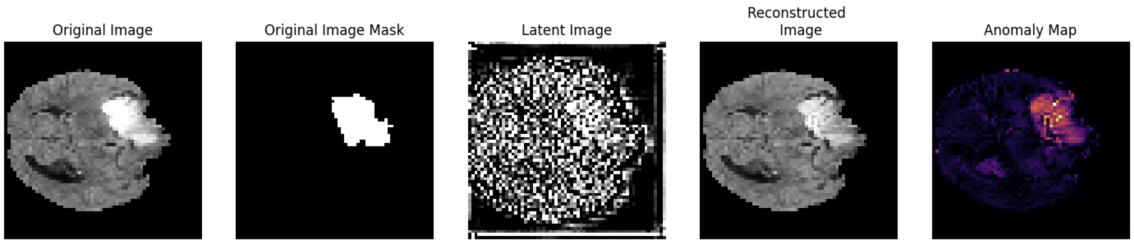


Figure 7.37. *Counterfactual generation*

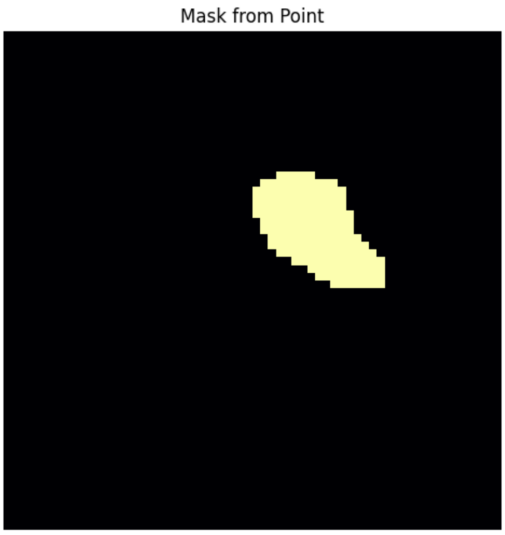


Figure 7.38. *Mask generated by Point-Prompted Pipeline*

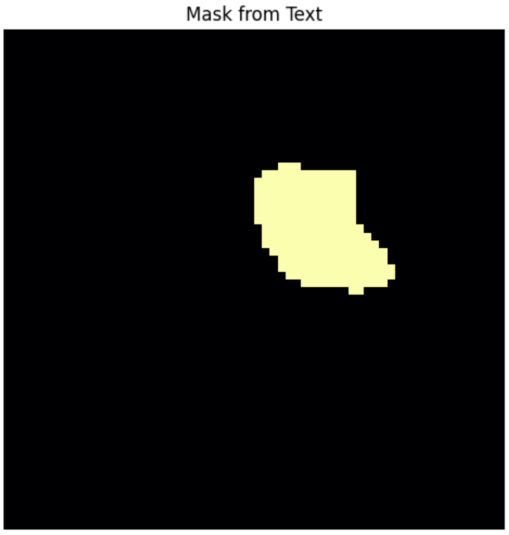


Figure 7.39. *Mask generated by Text-Prompted Pipeline*

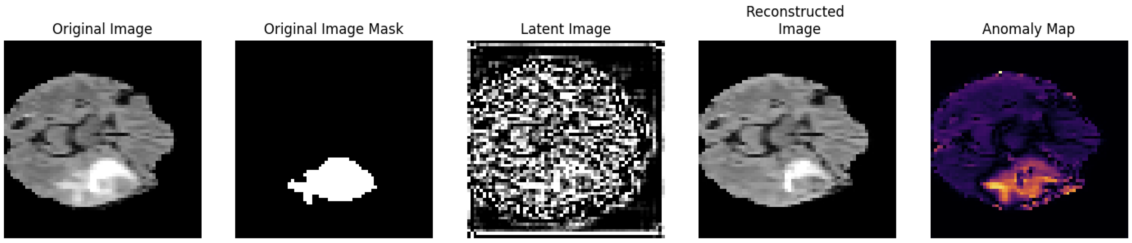


Figure 7.40. *Counterfactual generation*

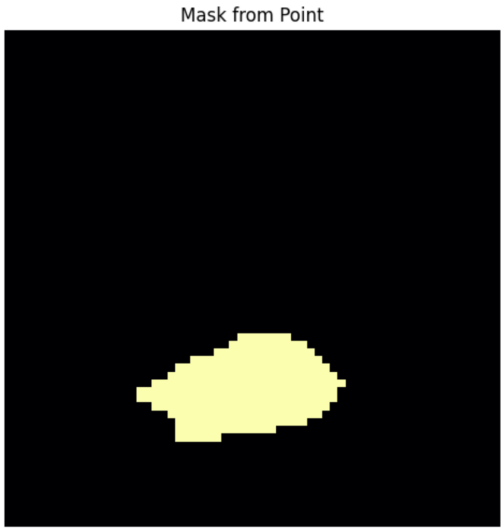


Figure 7.41. *Mask generated by Point-Prompted Pipeline*

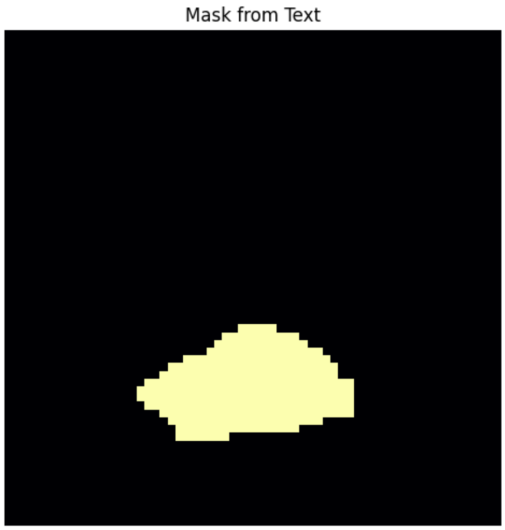


Figure 7.42. *Mask generated by Text-Prompted Pipeline*

7.8 Tumour - 1 - 1

	Point	Text	Intersection	Union
Dice	0.774	0.698	0.775	0.697
AUPRC	0.734	0.623	0.735	0.624
IoU	0.566	0.332	0.567	0.333
Precision	0.662	0.346	0.666	0.346
Recall	0.795	0.895	0.793	0.897
F1	0.723	0.500	0.724	0.499
Specificity	0.979	0.912	0.979	0.912
Hausdorff	7.177	10.920	7.039	11.058
ASSD	0.657	1.556	0.652	1.562

Table 7.8. *Evaluation Metrics for the "Tumour - 1 - 1" Configuration*

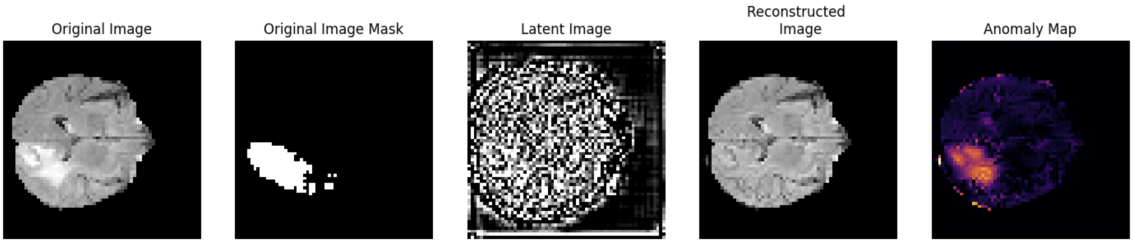


Figure 7.43. *Counterfactual generation*

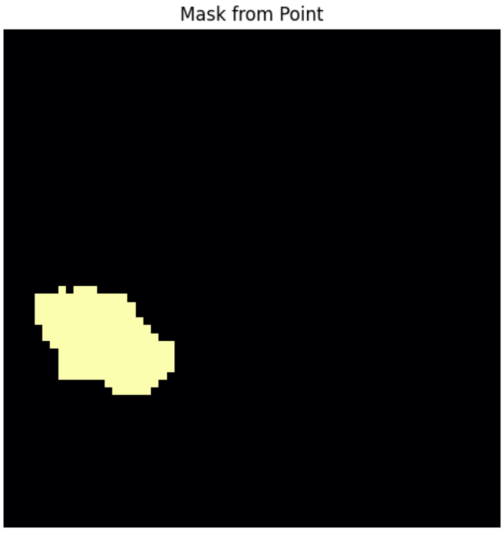


Figure 7.44. *Mask generated by Point-Prompted Pipeline*

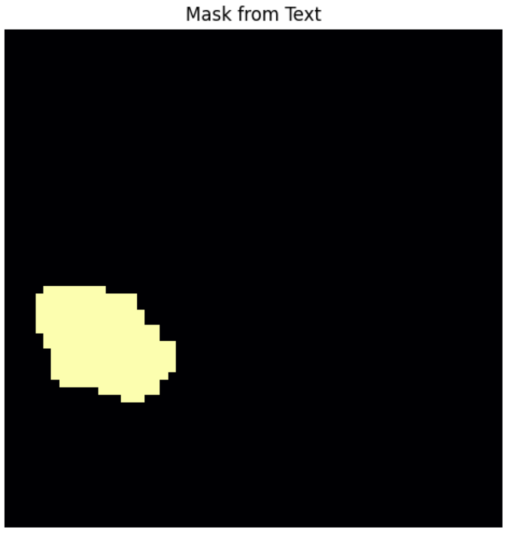


Figure 7.45. *Mask generated by Text-Prompted Pipeline*

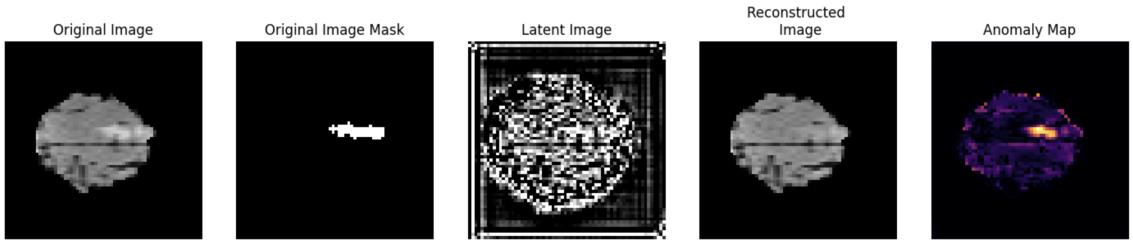


Figure 7.46. *Counterfactual generation*

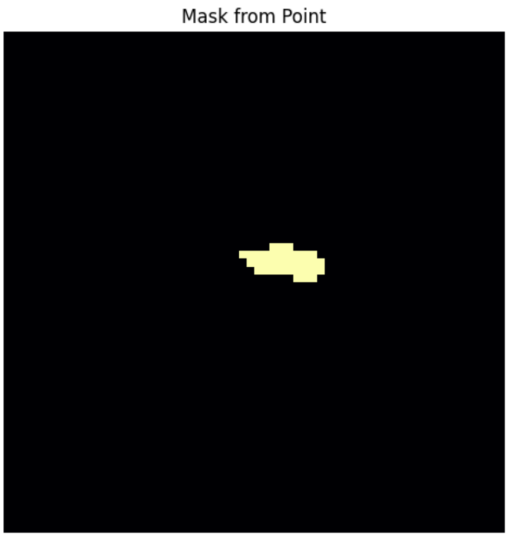


Figure 7.47. *Mask generated by Point-Prompted Pipeline*

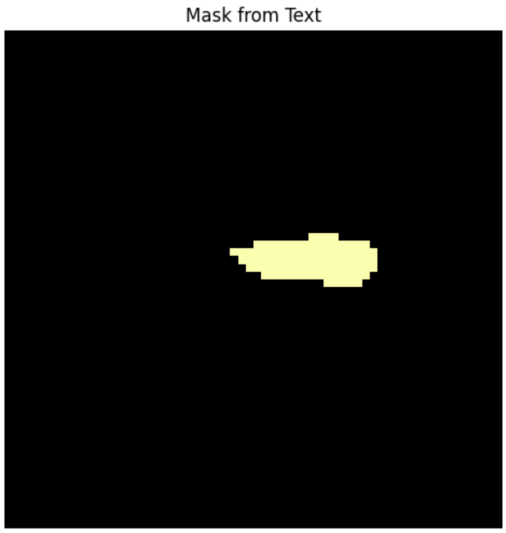


Figure 7.48. *Mask generated by Text-Prompted Pipeline*

7.9 Tumour - 3 - 1

	Point	Text	Intersection	Union
Dice	0.806	0.698	0.806	0.698
AUPRC	0.761	0.623	0.781	0.620
IoU	0.606	0.333	0.633	0.327
Precision	0.747	0.346	0.790	0.340
Recall	0.762	0.895	0.761	0.896
F1	0.755	0.499	0.775	0.492
Specificity	0.987	0.912	0.990	0.910
Hausdorff	5.919	10.920	5.828	11.011
ASSD	0.450	1.556	0.432	1.574

Table 7.9. *Evaluation Metrics for the "Tumour - 3 - 1" Configuration*

Figure 7.49. *Counterfactual generation*

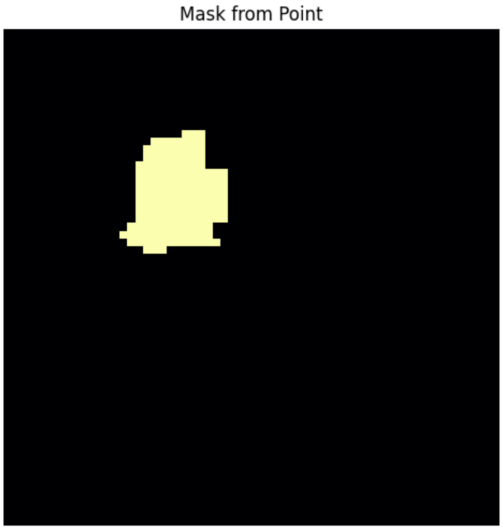


Figure 7.50. *Mask generated by Point-Prompted Pipeline*

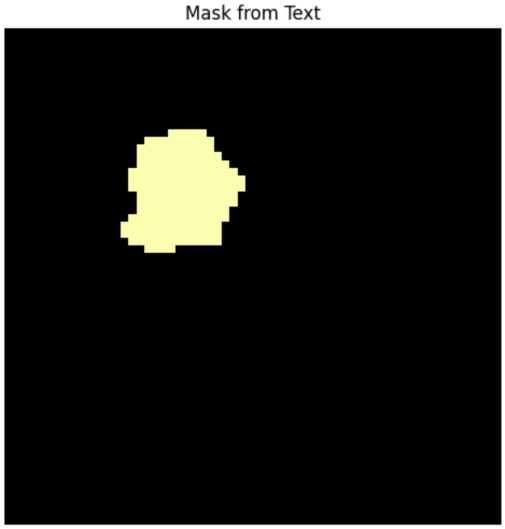


Figure 7.51. *Mask generated by Text-Prompted Pipeline*

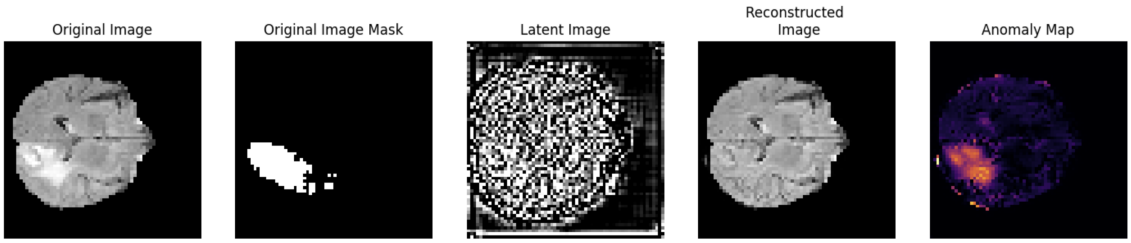


Figure 7.52. *Counterfactual generation*

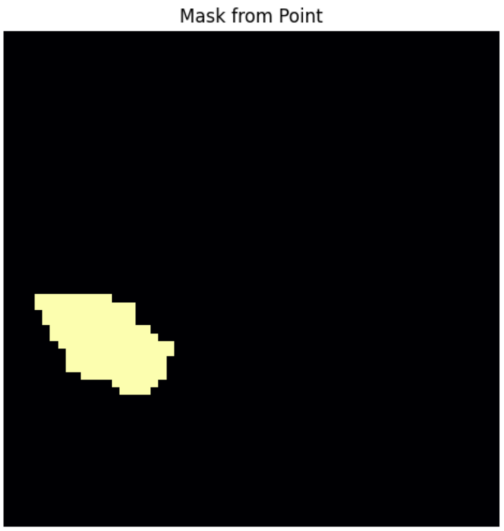


Figure 7.53. *Mask generated by Point-Prompted Pipeline*

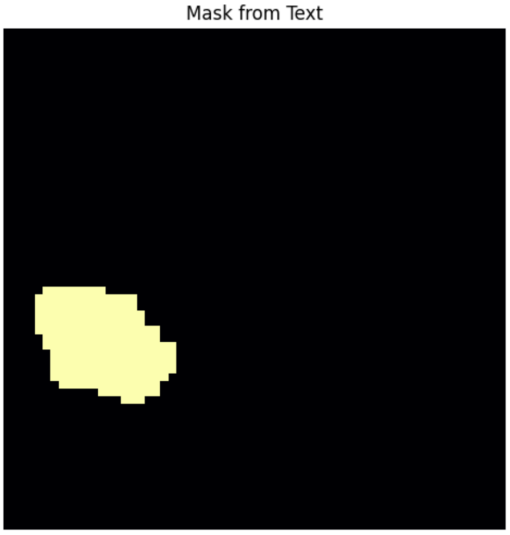


Figure 7.54. *Mask generated by Text-Prompted Pipeline*

7.10 Tumour - 4 - 4

	Point	Text	Intersection	Union
Dice	0.637	0.541	0.610	0.567
AUPRC	0.516	0.382	0.462	0.437
IoU	0.307	0.103	0.276	0.118
Precision	0.369	0.110	0.359	0.122
Recall	0.645	0.636	0.543	0.739
F1	0.470	0.187	0.432	0.211
Specificity	0.943	0.733	0.950	0.727
Hausdorff	11.200	19.293	11.258	19.241
ASSD	1.413	3.247	1.460	3.195

Table 7.10. *Evaluation Metrics for the "Tumour - 4 - 4" Configuration*

Figure 7.55. *Counterfactual generation*

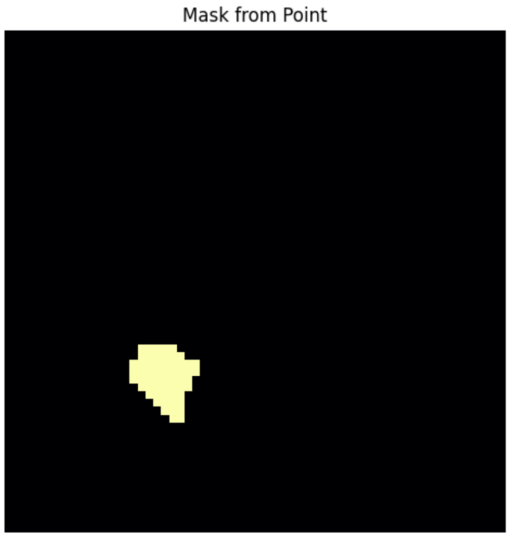


Figure 7.56. *Mask generated by Point-Prompted Pipeline*

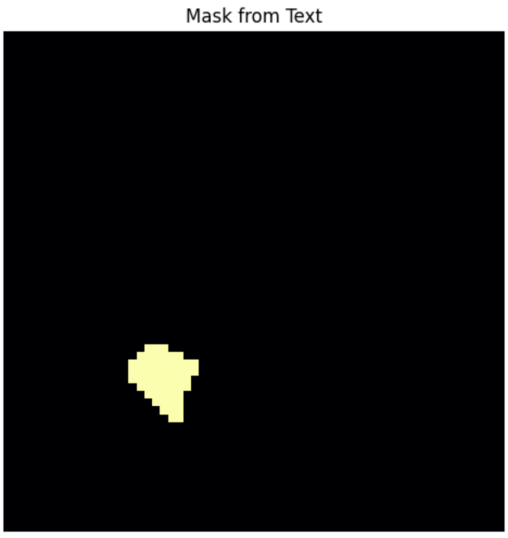


Figure 7.57. *Mask generated by Text-Prompted Pipeline*

Figure 7.58. *Counterfactual generation*

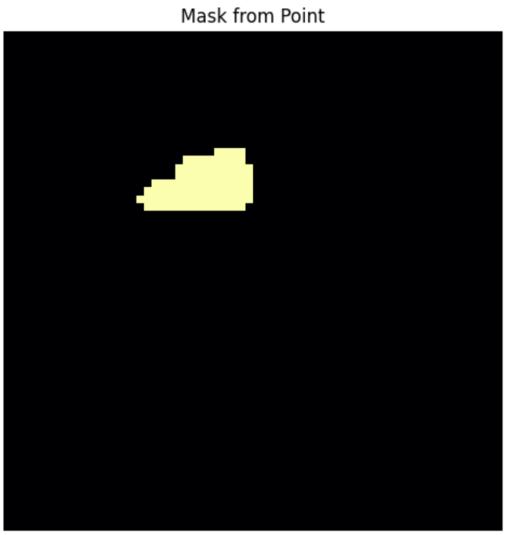


Figure 7.59. *Mask generated by Point-Prompted Pipeline*

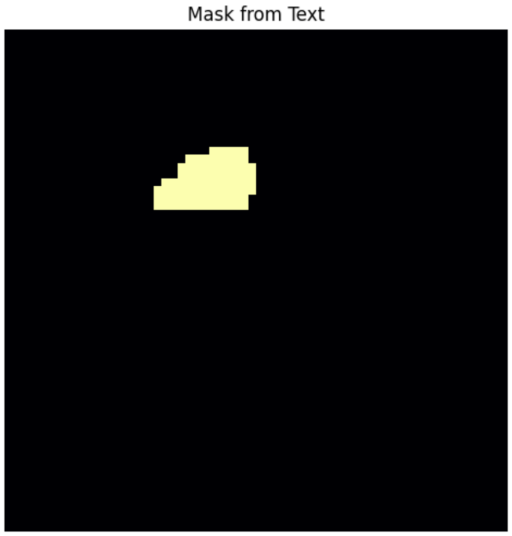


Figure 7.60. *Mask generated by Text-Prompted Pipeline*

7.11 Anomaly - 3 - 3

	Point	Text	Intersection	Union
Dice	0.806	0.781	0.776	0.811
AUPRC	0.761	0.707	0.831	0.698
IoU	0.606	0.527	0.676	0.494
Precision	0.747	0.615	0.941	0.544
Recall	0.762	0.787	0.705	0.844
F1	0.755	0.691	0.806	0.662
Specificity	0.987	0.974	0.998	0.963
Hausdorff	5.919	6.948	6.199	6.654
ASSD	0.450	0.685	0.528	0.610

Table 7.11. *Evaluation Metrics for the "Anomaly - 3 - 3" Configuration*

Figure 7.61. *Counterfactual generation*

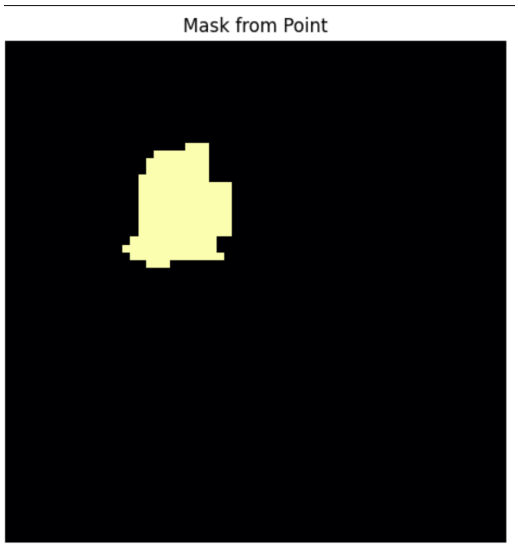


Figure 7.62. *Mask generated by Point-Prompted Pipeline*

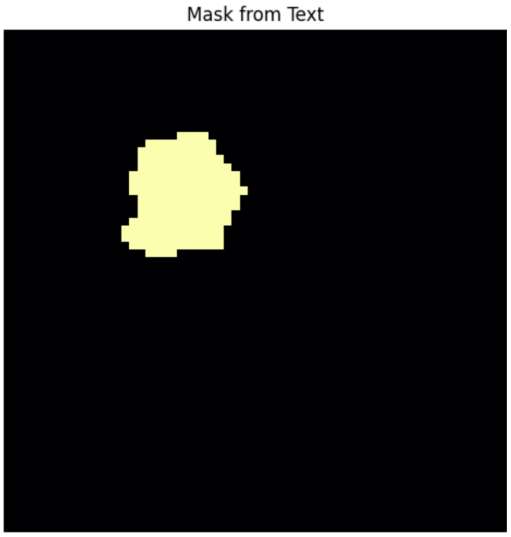


Figure 7.63. *Mask generated by Text-Prompted Pipeline*

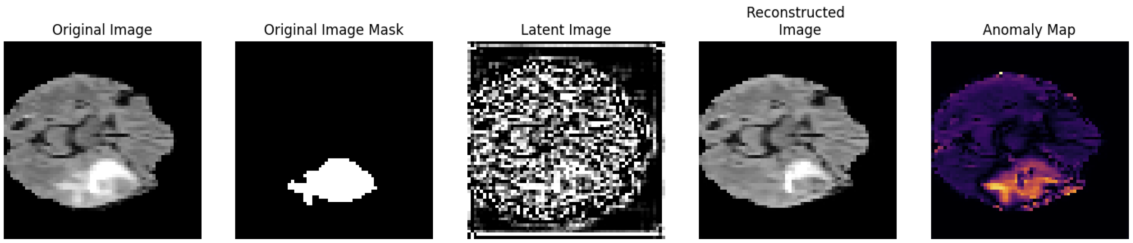


Figure 7.64. *Counterfactual generation*

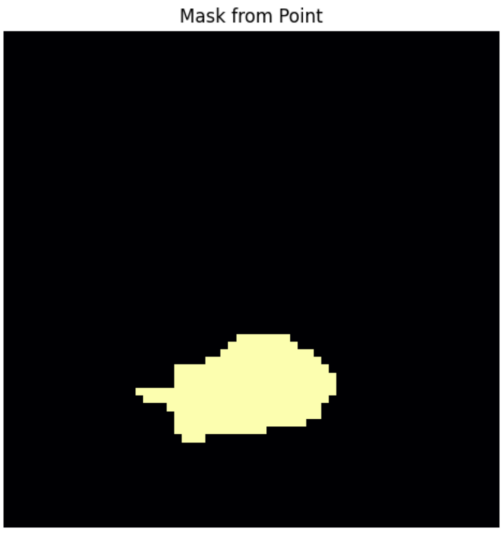


Figure 7.65. *Mask generated by Point-Prompted Pipeline*

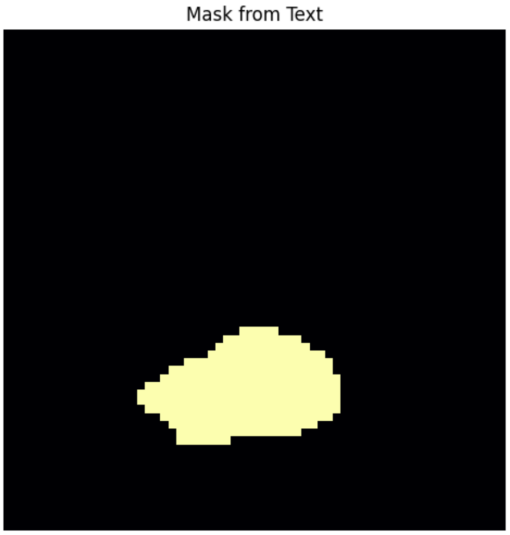


Figure 7.66. *Mask generated by Text-Prompted Pipeline*

7.12 Anomaly - 2 - 2

	Point	Text	Intersection	Union
Dice	0.799	0.800	0.816	0.783
AUPRC	0.743	0.754	0.849	0.713
IoU	0.580	0.586	0.717	0.502
Precision	0.683	0.656	0.929	0.540
Recall	0.793	0.845	0.758	0.880
F1	0.734	0.739	0.835	0.669
Specificity	0.981	0.977	0.997	0.961
Hausdorff	6.670	5.954	5.545	7.057
ASSD	0.531	0.534	0.416	0.649

Table 7.12. *Evaluation Metrics for the "Anomaly - 2 - 2" Configuration*

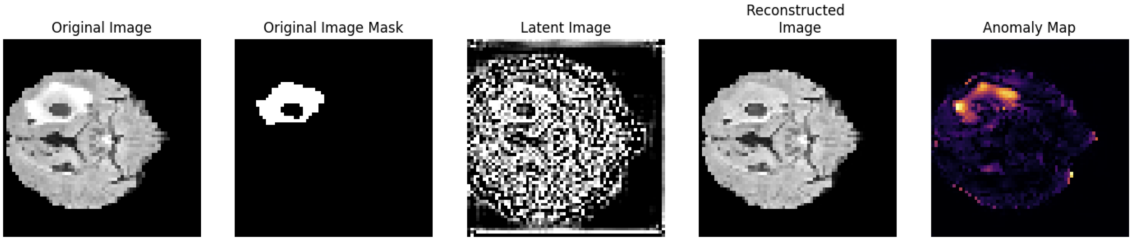


Figure 7.67. *Counterfactual generation*

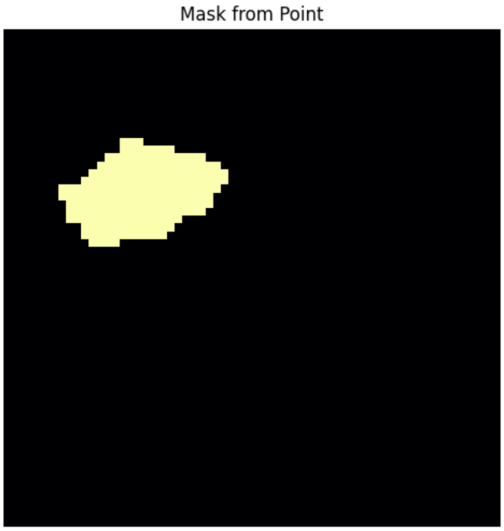


Figure 7.68. *Mask generated by Point-Prompted Pipeline*

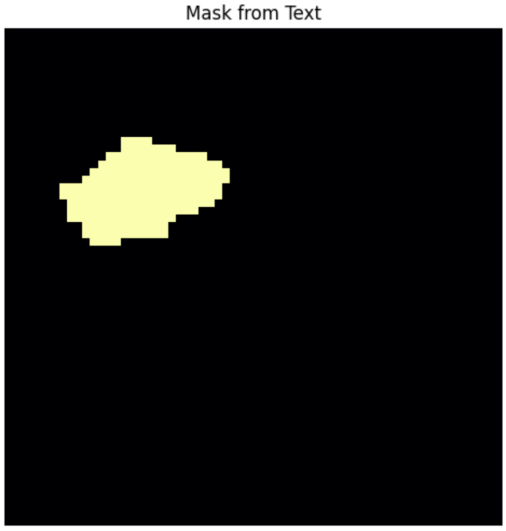


Figure 7.69. *Mask generated by Text-Prompted Pipeline*

Figure 7.70. *Counterfactual generation*

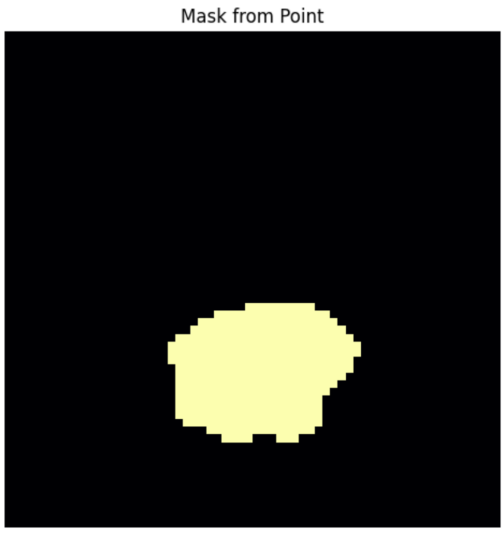


Figure 7.71. *Mask generated by Point-Prompted Pipeline*

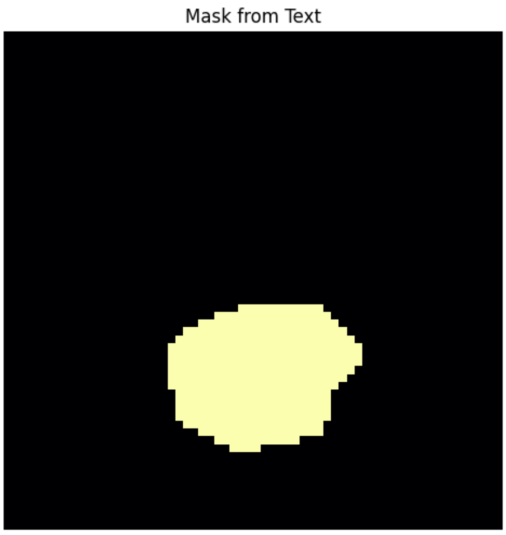


Figure 7.72. *Mask generated by Text-Prompted Pipeline*

7.13 Anomaly - 1 - 1

	Point	Text	Intersection	Union
Dice	0.774	0.685	0.780	0.680
AUPRC	0.734	0.624	0.756	0.620
IoU	0.566	0.329	0.599	0.321
Precision	0.662	0.342	0.709	0.332
Recall	0.795	0.901	0.793	0.902
F1	0.723	0.496	0.749	0.486
Specificity	0.979	0.910	0.984	0.907
Hausdorff	7.177	10.826	6.843	11.160
ASSD	0.657	1.553	0.582	1.628

Table 7.13. *Evaluation Metrics for the "Anomaly - 1 - 1" Configuration*

Figure 7.73. *Counterfactual generation*

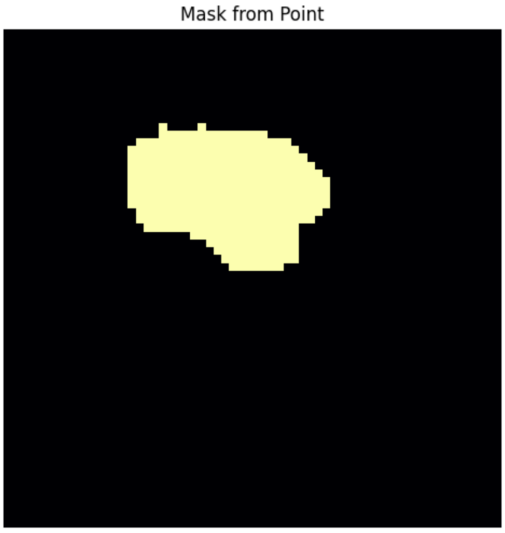


Figure 7.74. *Mask generated by Point-Prompted Pipeline*

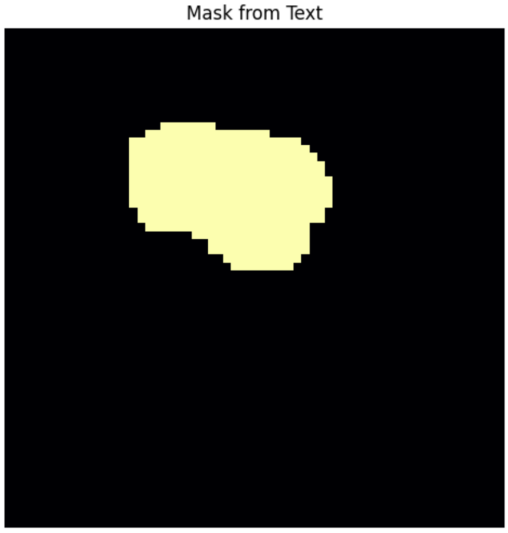


Figure 7.75. *Mask generated by Text-Prompted Pipeline*

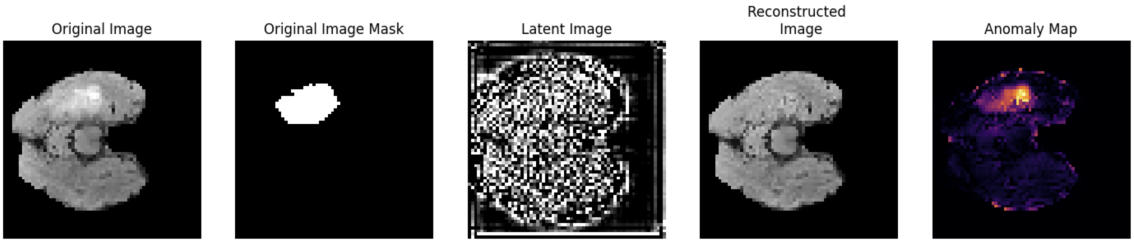


Figure 7.76. *Counterfactual generation*

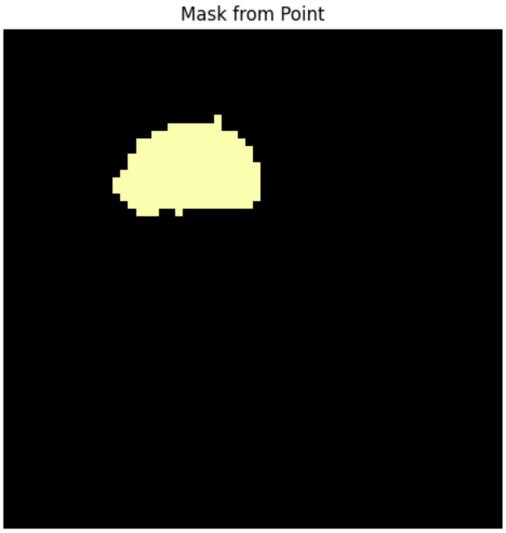


Figure 7.77. *Mask generated by Point-Prompted Pipeline*

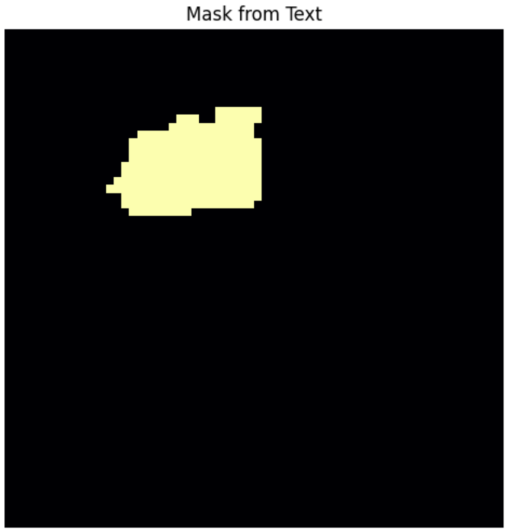


Figure 7.78. *Mask generated by Text-Prompted Pipeline*

7.14 Smaller U-Net

We experiment with a smaller architecture for our diffusion model's U-Net. In particular, our encoder now processes the images through three, instead of four, stages with constant channel widths: 64, 64 and 64 channels. Each stage now comprises of one, instead of two, residual blocks. We provide the results of the hyperparameters configuration of "lesion-3-3" below.

	Point	Text	Intersection	Union
Dice	0.757	0.778	0.758	0.777
AUPRC	0.778	0.776	0.807	0.766
IoU	0.611	0.624	0.626	0.611
Precision	0.855	0.787	0.944	0.736
Recall	0.681	0.751	0.650	0.783
F1	0.759	0.769	0.770	0.759
Specificity	0.993	0.988	0.998	0.982
Hausdorff	6.838	7.489	7.205	6.971
ASSD	0.414	1.325	0.420	1.535

Table 7.14. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration of the smaller U-Net*

7.15 Removing the Diffusion Model

In this section, we present the results obtained by directly applying the point-prompted and text-prompted segmentation pipelines to the original MRI images from the Medical Segmentation Decathlon Task 1 (Brain Tumor) dataset, without generating counterfactual images.

	Point	Text	Intersection	Union
Dice	0.337	0.164	0.303	0.218
AUPRC	0.195	0.056	0.125	0.124
IoU	0.114	0.022	0.083	0.052
Precision	0.145	0.049	0.178	0.072
Recall	0.353	0.285	0.364	0.324
F1	0.315	0.223	0.444	0.356
Specificity	0.514	0.422	0.463	0.355
Hausdorff	28.436	39.323	30.147	0.347
ASSD	4.147	7.348	4.736	6.548

Table 7.15. *Evaluation Metrics for the "Lesion - 3 - 3" Configuration without the usage of the Counterfactual*

7.16 Failure Cases

Despite the overall robustness of our framework, there are instances where it failed to correctly segment the lesions in the MRI images. These failure cases primarily arise due to the presence of very small and non-obvious lesions, which pose significant challenges for accurate detection and segmentation. In this section, we present several images where our model struggled to perform effectively.

The main reasons for the failure of our segmentation pipeline include:

Very Small Lesions: Lesions that are too small are often difficult to detect and segment accurately. The small size leads to insufficient contrast and prominence in the images, making it challenging for the model to differentiate them from the surrounding tissue.

Poor Quality of Original MRI Images: In some cases, the original MRI images may be of poor quality, with low resolution or high noise levels. This can hinder the model's ability to accurately identify and segment the lesions.

Below are examples of images where our model failed to correctly segment the lesions. Each set of images includes the original MRI, the original image mask, the latent image, the reconstructed image, and the anomaly map. These examples illustrate the challenges faced by our model in accurately detecting very small and subtle lesions.

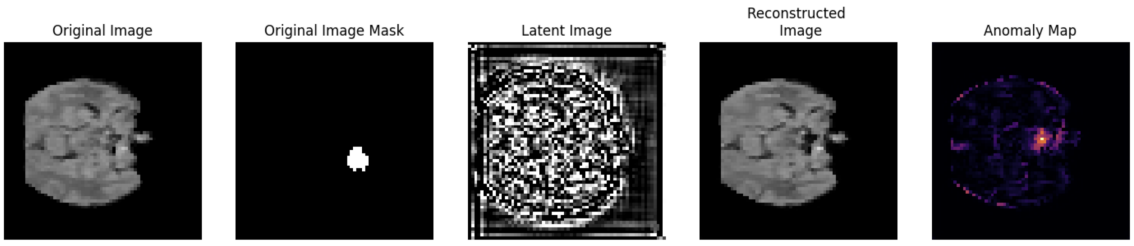


Figure 7.79. *Counterfactual generation*

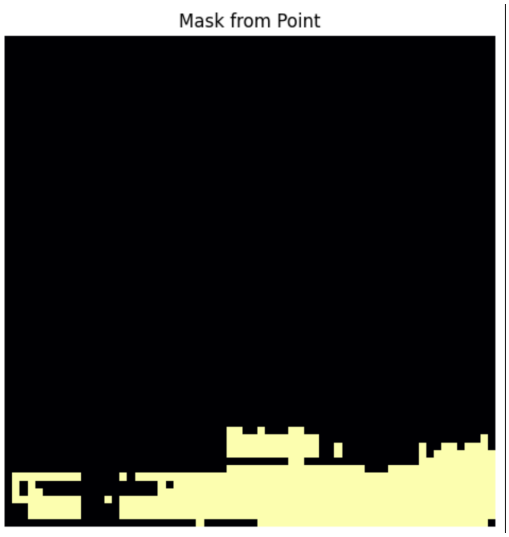


Figure 7.80. *Mask generated by Point-Prompted Pipeline*

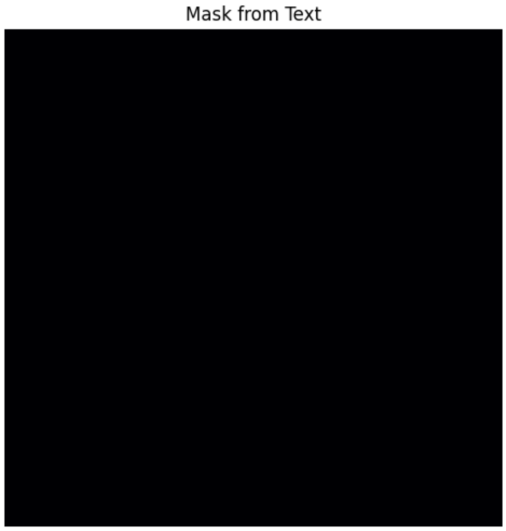


Figure 7.81. *Mask generated by Text-Prompted Pipeline*

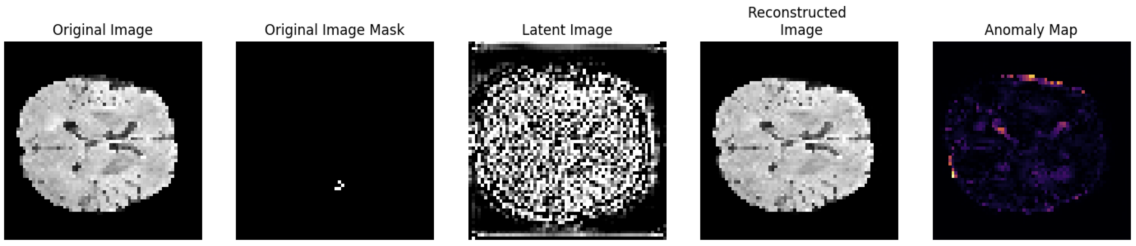


Figure 7.82. *Counterfactual generation*

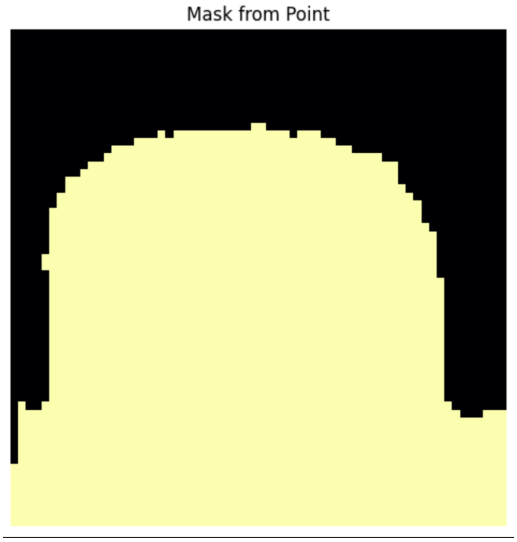


Figure 7.83. *Mask generated by Point-Prompted Pipeline*

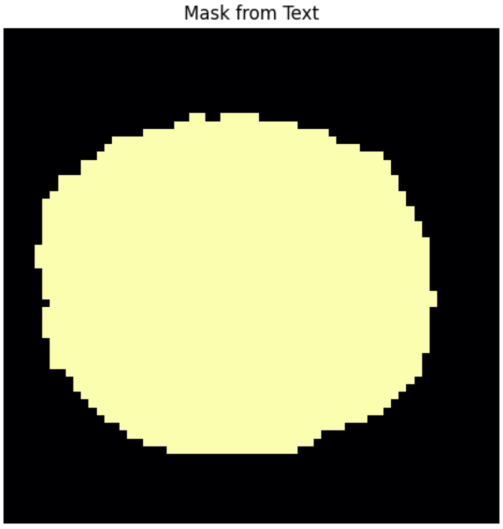


Figure 7.84. *Mask generated by Text-Prompted Pipeline*

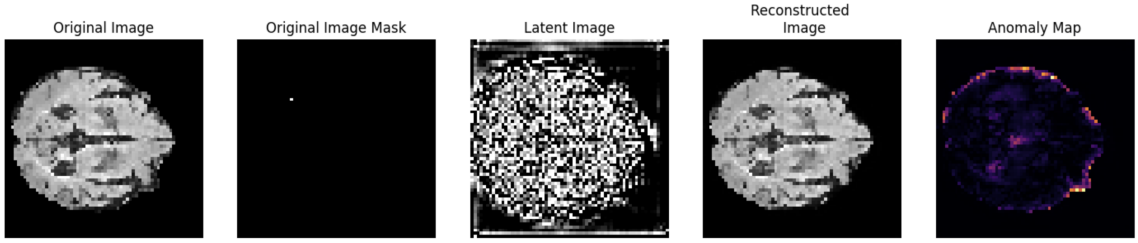


Figure 7.85. *Counterfactual generation*

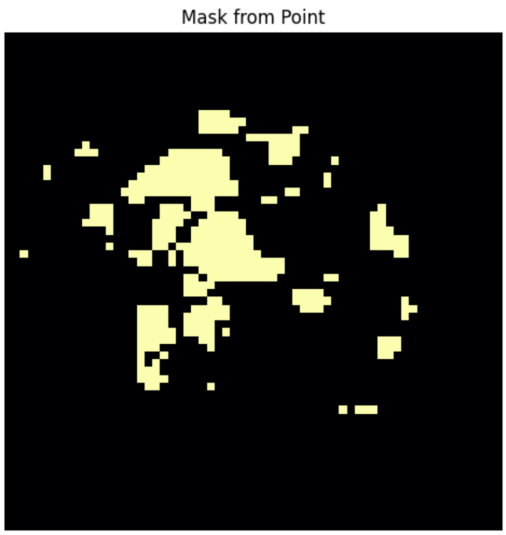


Figure 7.86. *Mask generated by Point-Prompted Pipeline*

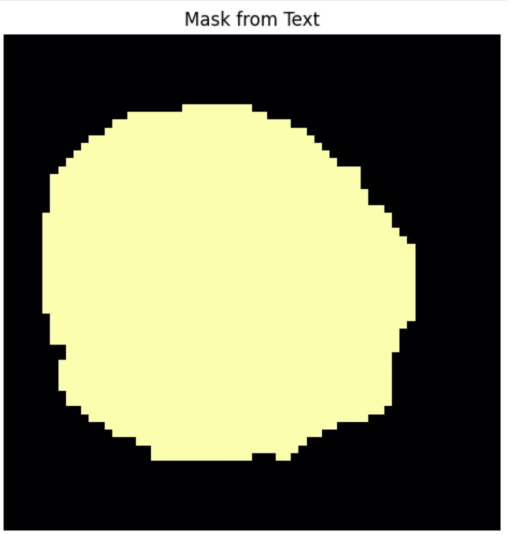


Figure 7.87. *Mask generated by Text-Prompted Pipeline*

7.17 Interpretation of Results and Analysis

In this section, we present a comprehensive analysis of the experimental results obtained using our segmentation framework on the Brain Tumor task of the Medical Segmentation Decathlon challenge. The analysis is based on various hyperparameter configurations, including different text prompts, the number of preprocessing steps, and the application of the SAM-GroundingDINO pipeline.

Our experiments demonstrated that both the ***"lesion"*** and ***"tumour"*** text prompts were equally effective in guiding the segmentation process. These prompts consistently yielded high performance across all evaluation metrics. The text prompt ***"anomaly"***, while still effective, produced slightly lower results compared to "lesion" and "tumor." This suggests that text prompts closely related to the specific nature of the target (i.e., lesion or tumor) are more effective in enhancing the segmentation accuracy.

We observed a clear trend where increasing the number of preprocessing steps in both the point-prompted and text-prompted pipelines led to improved results. This trend held true up to a certain point; specifically, the quality of the segmentation improved as the number of preprocessing steps increased up to three. However, when the number of preprocessing steps reached four, the quality of the results began to decline. This decline is attributed to the excessive loss of information caused by the repeated multiplication process, which ultimately hinders the model's ability to accurately segment the lesions. Conversely, the configuration with zero preprocessing steps (0-0) yielded the worst results, underscoring the importance of appropriate preprocessing in enhancing segmentation performance.

Our framework demonstrated robustness across different configurations, maintaining high performance regardless of the specific text prompt used, as long as it pertained to the general category of "lesion" or "tumor." Both the point-prompted and text-prompted segmentation pipelines produced comparable results, indicating that either method can be effectively used in practice. Similarly, the intersection and union methods, which combine the results from the point and text pipelines, also showed similar quality, further validating the robustness of our approach.

As another part of our ablation study, we experimented with a smaller diffusion model configuration, which resulted in slightly worse performance compared to the larger model but still produced good results overall. This indicates that while the diffusion model size does impact performance, the framework is still effective with smaller models.

Furthermore, we evaluated the performance of the SAM-GroundingDINO pipeline applied directly to the original MRI images, bypassing the counterfactual generation step. In this scenario, we observed significantly worse results, highlighting the critical role of the diffusion model and the counterfactual generation process in enhancing segmentation accuracy.

Chapter 8

Comparative Analysis with Existing Approaches in Brain Tumor Segmentation

8.1 Challenges in Direct Comparison with Existing Work

In this chapter, we provide a comparative analysis of the results obtained from our framework with those reported in various other studies that have employed diffusion models for brain tumor segmentation and localization. While this comparison is aimed at placing our work in the context of the existing literature, it is important to note several challenges that prevent a direct, one-to-one comparison.

One major limitation is the lack of publicly available benchmarks from other works. Many studies in this area do not explicitly specify key details, such as the specific subset of the BraTS dataset they used or the precise architecture of the models they employed, including critical variations in U-Net configurations. These differences in datasets and architectural designs can significantly affect segmentation performance, making it difficult to draw direct comparisons across all evaluation metrics.

Furthermore, the training protocols, hyperparameter settings, and preprocessing steps are often not consistently reported in these papers, adding another layer of complexity to the comparison. For instance, variations in the number of preprocessing steps, the type of text prompts used, and the inclusion of attention mechanisms in U-Net architectures are critical factors that can influence the results but are not always fully detailed in the related work.

Despite these challenges, we aim to provide a high-level comparison of the overall performance trends by focusing on the commonly reported metrics such as Dice score, IoU, precision, recall, and specific segmentation performance metrics. This qualitative comparison will highlight the strengths of our framework and its alignment with or improvement over existing approaches in the field of brain tumor segmentation using diffusion models.

In the following sections, we present the performance of our framework alongside the results reported in several key papers, with careful consideration of the aforementioned constraints. While a direct benchmark comparison is not possible, this analysis provides valuable insights into the broader impact of diffusion models on brain tumor segmentation tasks.

8.2 Comparison with Related Work

In this section, we compare the results of our framework with three relevant studies that employed diffusion models for brain tumor segmentation using counterfactual images. These papers share a similar conceptual foundation: utilizing a diffusion model to generate counterfactual images and calculating their segmentation metrics based on the difference between the original and generated images. However, several key differences in the implementation and pipeline design should be noted.

The primary distinction lies in the specific details of the diffusion model used in each work. The hyperparameters of the diffusion models, including the number of time steps, noise schedules, and model architecture, vary across the studies. These differences in model design and counterfactual generation contribute to variations in the results. Furthermore, while all studies focus on leveraging the difference image for tumor segmentation, we extend this approach by introducing a preprocessing step that enhances the difference image before applying the segmentation models.

Our pipeline also distinguishes itself by incorporating advanced segmentation techniques, specifically SAM (Segment Anything Model) and GroundingDINO, which provide a more structured and flexible method for segmentation compared to traditional approaches. These additions allow for improved lesion localization and segmentation by refining the input difference images and utilizing both point- and text-prompted pipelines. As a result, our framework not only addresses some of the limitations observed in prior works but also introduces a more comprehensive approach to segmenting brain tumors in MRI images.

8.2.1 Comparison with Wolleb et al.'s Method for Medical Anomaly Detection Using DDPMs

The first paper, written by Wolleb et al. [50], presents a method for medical anomaly detection using Denoising Diffusion Probabilistic Models (DDPMs) to generate counterfactual images, where only the pathological regions are altered, and the rest of the image remains unchanged. The difference between the original image and the generated "healthy" image forms the anomaly map, which is used for segmentation.

Comparison with Our Approach:

- **U-Net Architecture**: Their U-Net employs 128 channels in the first layer and uses one attention head at a resolution of 16. In contrast, our U-Net is more scalable, with four stages featuring increasing channel widths (64, 128, 256, and 512 channels), and includes residual blocks at each stage. Additionally, our model integrates attention at three levels (second, third, and fourth) with cross-attention mechanisms, which provide enhanced feature focus and improve lesion localization.
- **Diffusion Steps:** Their model utilizes 1,000 diffusion steps, while our setup is optimized to use fewer steps, improving computational efficiency. Despite fewer diffusion steps, our advanced preprocessing and segmentation models (SAM and GroundingDINO) allow us to maintain high segmentation accuracy, whereas their method directly relies on classifier-guided denoising.
- **Classifier vs. Segmentation Models:** While their method uses a classifier to guide the diffusion process toward generating healthy images, our framework incorporates advanced segmentation models after the difference image is processed. Our approach includes a preprocessing step that enhances the quality of the difference image and uses sophisticated segmentation techniques, resulting in more accurate and refined lesion segmentation.
- **Dataset Size:** Their dataset consists of 16,205 slices for training and 1,787 slices for testing, while our dataset is based on 388 brain MRIs for training, 96 for validation, and 251 for testing. While they focus on individual slices, our approach benefits from working on full MRI scans, which provide better spatial context for segmentation tasks.
- **Training and Batch Size:** They train their model for 50,000 iterations with a batch size of 10, while our framework uses 4,000 iterations with a batch size of 32. Our structured training approach, featuring dynamic learning rates, gradient scaling, and conditioning dropout, allows us to achieve efficient convergence with fewer iterations.

Results:

The results from the first paper demonstrate how the two key hyperparameters, the classifier gradient scale $\boldsymbol{s}$ and the noise level $\boldsymbol{L}$, impact the performance of the diffusion model on the BRATS2020 dataset. The **Dice score** and **AUROC** were used as evaluation metrics. As seen in the figures, the Dice score reaches its peak value of around 0.7 when L=500 and s is close to 100. The AUROC score remains high across different values of s, with a maximum of approximately 0.98 at both L=500 and L=250. However, when the noise level L is set too high (e.g., L=750), the performance degrades significantly, with the Dice score dropping below 0.4 and the AUROC falling below 0.95. Similarly, increasing the gradient scale beyond a certain threshold introduces artifacts, particularly at the edges

of the brain, leading to a decrease in the Dice score. The results emphasize that careful tuning of both L and s is critical to achieve optimal segmentation performance.

8.2.2 Comparison with Sanchez et al.'s Method for Medical Anomaly Detection Using DDPMs

The second paper, written by Sanchez et al. [40], presents a method for generating counterfactual images to localize lesions in brain MRIs using a diffusion model with implicit guidance and attention conditioning. The core difference between their approach and ours lies in the specific techniques used to generate the counterfactual and handle the resulting difference image.

Comparison with Our Approach:

- **Attention Mechanisms:** While both approaches rely on U-Net architectures, the second paper incorporates conditional attention mechanisms throughout the network, which guide the model using conditioning information during the decoding process. Our method, in contrast, uses attention at specific levels (second, third, and fourth) but includes cross-attention to integrate additional external context, such as class labels, ensuring more precise segmentation by focusing attention on relevant areas in the brain MRI.
- **Dynamic Normalization:** One of the unique aspects of their approach is the use of dynamic normalization during inference. This prevents pixel saturation in the latent space, ensuring that the model retains reconstruction quality over iterative denoising steps. Our framework does not rely on dynamic normalization. Instead, we preprocess the difference image through multiple multiplication steps with the original image to enhance lesion visibility and reduce noise, offering an alternative pathway to maintaining quality in the reconstructed image.
- **Diffusion Process:** Their method applies implicit guidance in the diffusion process, combining conditional and unconditional predictions to generate the counterfactual. This is achieved by randomly dropping conditioning during training to prevent overfitting. Our approach, however, focuses on a simpler but more controlled conditioning technique in which external information is explicitly integrated. The counterfactual generation in our model is further enhanced by advanced segmentation tools such as SAM and GroundingDINO, which are used after preprocessing to provide refined segmentation results.
- **Dataset Size and Image Resolution:** The second paper uses a larger training dataset from the BraTS 2021 challenge, with 938 MRIs for training, compared to our 388. They downsample images to a 64×64 resolution for training, while evaluating at 128×128 for better comparison. Our approach uses a consistent image resolution across training and evaluation. This difference in dataset size and resolution could explain some of the variations in segmentation quality, as larger datasets and higher resolutions typically improve model performance.

Results:

The results from the second paper demonstrate strong performance in tumor detection and localization, achieving an AUPRC of 82.8 $\pm$ 0.4 and a Dice score of 76.2 $\pm$ 0.3 when trained on the full dataset. The model incorporates implicit guidance, attention conditioning, and dynamic normalization, which significantly contribute to these results. Specifically, the inclusion of implicit guidance improves the Dice score to 52.0, while adding attention conditioning boosts it further to 74.3. The final improvement comes from dynamic normalization, bringing the Dice score to 76.2, which represents the optimal performance for their framework. These results indicate the effectiveness of combining diffusion models with advanced guidance and normalization techniques for brain tumor segmentation.

8.2.3 Comparison with Fontanella et al.'s Method for Medical Anomaly Detection Using DDPMs

The third paper, written by Fontanella et al. [11], introduces a method called "Dif-fuse" that combines a DDPM trained on healthy samples with saliency maps, leveraging counterfactual examples to detect and localize anomalies. Their approach involves generating a noised version of a diseased image using inverse DDIM sampling, then smoothing the saliency maps to generate masks that localize the pathological regions. The pathological regions are edited using DDPM sampling, while the healthy parts of the image are preserved through DDIM sampling. A coherent final image is achieved by mixing the DDPM and DDIM components at each sampling step, effectively preventing structural changes in healthy area

Comparison with Our Approach: Compared to our approach, the third paper also uses a DDPM for counterfactual generation, but it adds a saliency map mechanism for refining the localization of pathological regions. Our method differs in that we employ SAM and GroundingDINO for segmentation, focusing on preprocessing the difference image, while "Dif-fuse" uses ACAT for generating saliency maps. Moreover, while we use four metrics for evaluation (e.g., Dice, AUPRC), this paper highlights the importance of thresholding saliency maps and noise levels in their anomaly detection process, refining performance by optimizing these parameters.

In terms of dataset, both approaches use the BraTS2021 dataset, but the "Dif-fuse" paper also tests on the IST-3 dataset, which we do not. Regarding architecture, "Dif-fuse" uses a U-Net with 128 channels and attention heads at specific resolutions (8, 16, 32), whereas our U-Net model uses a wider range of channels (64-512) with attention layers integrated at different levels to enhance feature focus. Lastly, both methods address the issue of excessive noise and artifacts introduced by over-processing, but we implement it by carefully choosing the number of preprocessing steps, while "Dif-fuse" adjusts the noise levels and thresholds for binarizing saliency maps.

Their results indicate optimal performance with 500 noising steps and thresholding at the 90th percentile, showcasing their best Dice score when artifacts are minimized and pathological regions are sufficiently removed.

Results:

The results from the third paper, "Dif-fuse," indicate that their approach outperforms other weakly-supervised methods, both those employing GANs and diffusion models. Specifically, the anomaly maps generated by "Dif-fuse" achieve a Dice score of 0.7056, surpassing other models like f-Ano GAN (0.5407), classifier-guided diffusion models (0.6534), and classifier-free diffusion models (0.6393). Additionally, their ablation study on the saliency maps from ACAT reveals a Dice score of 0.5753, highlighting that integrating their diffusion model with the saliency map process significantly enhances performance.

In comparison to our approach, while "Dif-fuse" achieves a commendable Dice score of 0.7056, our framework, which integrates SAM, GroundingDINO, and multiple preprocessing steps, achieves higher Dice scores in certain configurations. This reinforces the effectiveness of our counterfactual generation pipeline when combined with advanced segmentation techniques like SAM. Additionally, "Dif-fuse" focuses heavily on optimizing saliency maps and noising steps, while we emphasize a robust preprocessing of difference images and multiple segmentation pipelines to achieve more flexible lesion localization across various test cases.

Chapter 9

Epilogue

In this thesis, we have developed and evaluated a robust framework for the segmentation and localization of brain tumors in MRI images. Our approach integrates a diffusion model to generate healthy counterfactual images, followed by segmentation using the SAM (Segment Anything Model) and GroundingDINO models. We explored various hyperparameter configurations and conducted an ablation study to understand the significance of each component in our pipeline.

Our key findings include the following:

Effectiveness of Text Prompts: Our experiments demonstrated that text prompts such as "lesion" and "tumor" are equally effective, consistently yielding high performance across all evaluation metrics. The "anomaly" prompt, while still effective, resulted in slightly lower performance. This highlights the importance of choosing text prompts that are closely related to the specific nature of the target lesions.

Impact of Preprocessing Steps: Increasing the number of preprocessing steps generally improved the segmentation quality, with optimal results observed at three steps. Beyond this point, performance began to decline due to excessive information loss caused by repeated multiplications. The worst results were observed with no preprocessing, underscoring the necessity of adequate preprocessing.

Robustness of the Framework: Our framework proved to be robust across different configurations, maintaining high performance irrespective of the specific text prompt used, provided it was relevant to the target lesions. Both point-prompted and text-prompted pipelines performed comparably well, as did the intersection and union methods combining their results.

Ablation Study: Our ablation study confirmed the critical role of the diffusion model

and counterfactual generation in achieving high segmentation accuracy. A smaller diffusion model configuration resulted in slightly lower but still acceptable performance. However, removing the diffusion model entirely and applying the SAM-GroundingDINO pipeline directly to the original MRI images led to significantly worse results.

Overall, this thesis contributes to the field of medical image segmentation by demonstrating the efficacy of integrating generative models with advanced segmentation techniques. The proposed framework enhances the accuracy of lesion localization and segmentation, making it a valuable tool for assisting radiologists and medical professionals.

Future work could explore several avenues to further improve the framework:

- **Enhanced Preprocessing Techniques:** Investigating alternative preprocessing methods that preserve more information could yield better segmentation results.
- **Extended Hyperparameter Tuning:** A broader exploration of hyperparameters may uncover configurations that further optimize performance.
- **Real-time Application and Clinical Trials:** Testing the framework in real-time clinical settings would provide insights into its practical utility and potential areas for improvement.

In conclusion, our framework represents a significant advancement in the automatic segmentation of brain tumors in MRI images. By leveraging the strengths of diffusion models and state-of-the-art segmentation techniques, we have developed a robust, flexible, and effective tool that holds promise for aiding medical diagnosis and treatment planning.

Bibliography

[1] URL: https://pareto.ai/blog/diffusion-models.

[2] Brian D.O. ANDERSON. *REVERSE-TIME DIFFUSION EQUATION MODELS*. URL: https://core.ac.uk/download/pdf/82826666.pdf. Department of Electrical Engineering, The University of Newcastle, N.S. W. 2308, Australia - 1980.

[3] Michela Antonelli et al. *The Medical Segmentation Decathlon*. URL: https://arxiv.org/pdf/2106.05735. arXiv:2106.05735v1 [eess.IV] 10 Jun 2021.

[4] M. Jorge Cardoso et al. *MONAI: An open-source framework for deep learning in healthcare*. URL: https://arxiv.org/pdf/2211.02701. arXiv:2211.02701v1 [cs.LG] 4 Nov 2022.

[5] Ziyi Chang, George Koulieris, and Hubert P. H. Shum. *On the Design Fundamentals of Diffusion Models: A Survey*. URL: https://arxiv.org/pdf/2306.04542. arXiv:2306.04542v3 [cs.LG] 19 Oct 2023.

[6] Patrick Ferdinand Christa et al. *Automatic Liver and Tumor Segmentation of CT and MRI Volumes Using Cascaded Fully Convolutional Neural Networks*. URL: https://arxiv.org/pdf/1702.05970.

[7] Prafulla Dhariwa and Alex Nichol. *Diffusion Models Beat GANs on Image Synthesis*. URL: https://arxiv.org/pdf/2105.05233.pdf. arXiv:2105.05233v4 [cs.LG] 1 Jun 2021.

[8] Francisco Javier Díaz-Pernas et al. *A Deep Learning Approach for Brain Tumor Classification and Segmentation Using a Multiscale Convolutional Neural Network*. URL: https://arxiv.org/pdf/2402.05975.

[9] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. *DENSITY ESTIMATION USING REAL NVP*. URL: https://arxiv.org/pdf/1605.08803. arXiv:1605.08803v3 [cs.LG] 27 Feb 2017.

[10] Jan Egger et al. *Medical deep learning—A systematic meta-review*. URL: https://arxiv.org/pdf/2010.14881.

[11] Alessandro Fontanella et al. *Diffusion Models for Counterfactual Generation and Anomaly Detection in Brain Images*. URL: https://arxiv.org/pdf/2308.02062. arXiv:2308.02062v1 [eess.IV] 3 Aug 2023.

[12] Ian J. Goodfellow et al. *Generative Adversarial Nets*. URL: https://arxiv.org/pdf/1406.2661.

[13] Agrim Gupta, Piotr Dollar, and Ross Girshick. *LVIS: A Dataset for Large Vocabulary Instance Segmentation*. URL: https://arxiv.org/pdf/1908.03195. arXiv:1908.03195v2 [cs.CV] 15 Sep 2019.

[14] Ayan Gupta et al. *Brain Tumor Segmentation from MRI Images using Deep Learning Techniques*. URL: https://arxiv.org/pdf/2305.00257.

[15] Qisheng He et al. *Modality-Agnostic Learning for Medical Image Segmentation Using Multi-modality Self-distillation*. URL: https://arxiv.org/pdf/2306.03730. arXiv:2306.03730v1 [eess.IV] 6 Jun 2023.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. *Denoising Diffusion Probabilistic Models*. URL: https://arxiv.org/pdf/2006.11239.pdf. arXiv:2006.11239v2 [cs.LG] 16 Dec 2020.

[17] Jonathan Ho and Tim Salimans. *CLASSIFIER-FREE DIFFUSION GUIDANCE*. URL: https://arxiv.org/pdf/2207.12598.pdf. arXiv:2207.12598v1 [cs.LG] 26 Jul 2022.

[18] Daniel P. Huttenlocher, Gregory A. Klanderman, and William J. Rucklidge. *Comparing Images Using the Hausdorff Distance*. URL: https://people.eecs.berkeley.edu/~malik/cs294/Huttenlocher93.pdf.

[19] Fabian Isensee et al. *nnU-Net: Self-adapting Framework for U-Net-Based Medical Image Segmentationn*. URL: https://arxiv.org/pdf/1809.10486. arXiv:1809.10486v1 [cs.CV] 27 Sep 2018.

[20] Muhammad Waseem Khan. *A Survey: Image Segmentation Techniques*. URL: https://www.ijfcc.org/papers/274-B317.pdf.

[21] Diederik P. Kingma and Max Welling. *Auto-Encoding Variational Bayes*. URL: https://arxiv.org/pdf/1312.6114. arXiv:1312.6114v11 [stat.ML] 10 Dec 2022.

[22] Alexander Kirillov1 et al. *Segment Anything*. URL: https://arxiv.org/pdf/2304.02643. arXiv:2304.02643v1 [cs.CV] 5 Apr 2023.

[23] Jianning Li et al. *Open-Source Skull Reconstruction with MONAI*. URL: https://arxiv.org/pdf/2211.14051. arXiv:2211.14051v2 [eess.IV] 15 Jun 2023.

[24] Tsung-Yi Lin et al. *Microsoft COCO: Common Objects in Context*. URL: https://arxiv.org/pdf/1405.0312. arXiv:1405.0312v3 [cs.CV] 21 Feb 2015.

[25] Shilong Liu et al. *Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection*. URL: https://arxiv.org/pdf/2303.05499. arXiv:2303.05499v4 [cs.CV] 20 Mar 2023.

[26] Jonathan Long, Evan Shelhamer, and Trevor Darrell. *Fully Convolutional Networks for Semantic Segmentation*. URL: https://arxiv.org/pdf/1411.4038. arXiv:1411.4038v2 [cs.CV] 8 Mar 2015.

[27] D.G. Lowe. *Object recognition from local scale-invariant features*. URL: https://ieeexplore.ieee.org/document/790410/authors#authors. Print ISBN:0-7695-0164-8.

[28] Richard Meyes et al. *Ablation Studies in Artificial Neural Networks*. URL: https://arxiv.org/pdf/1901.08644. arXiv:1901.08644v2 [cs.NE] 18 Feb 2019.

[29] Alex Nichol and Prafulla Dhariwal. *Improved Denoising Diffusion Probabilistic Models*. URL: https://arxiv.org/pdf/2102.09672.pdf. arXiv:2102.09672v1 [cs.LG] 18 Feb 2021.

[30] Alex Nichol et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. URL: https://arxiv.org/pdf/2112.10741. arXiv:2112.10741v3 [cs.CV] 8 Mar 2022.

[31] Sanguk Park and Minyoung Chung. *Cardiac Segmentation on CT Images through Shape-Aware Contour Attentions*. URL: https://arxiv.org/pdf/2105.13153.

[32] Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. *TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning*. URL: https://arxiv.org/pdf/2003.04696. arXiv:2003.04696v5 [eess.IV] 5 Aug 2021.

[33] Francesco Piccialli et al. *A survey on deep learning in medicine: Why, how and when?* URL: https://www.sciencedirect.com/science/article/abs/pii/S1566253520303651.

[34] Walter H. L. Pinaya et al. *Generative AI for Medical Imaging: extending the MONAI Framework*. URL: https://arxiv.org/pdf/2307.15208. arXiv:2307.15208v1 [eess.IV] 27 Jul 2023.

[35] Marta B.M. Ranzini et al. *MONAIFBS: MONAI-BASED FETAL BRAIN MRI DEEP LEARNING SEGMENTATION*. URL: https://arxiv.org/pdf/2103.13314. arXiv:2103.13314v1 [eess.IV] 21 Mar 2021.

[36] Shaoqing Ren et al. *Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks*. URL: https://arxiv.org/pdf/1506.01497. arXiv:1506.01497v3 [cs.CV] 6 Jan 2016.

[37] Robin Rombach et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. URL: https://arxiv.org/pdf/2112.10752.pdf. arXiv:2112.10752v2 [cs.CV] 13 Apr 2022.

[38] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. URL: https://arxiv.org/pdf/1505.04597. arXiv:1505.04597v1 [cs.CV] 18 May 2015.

[39] Pedro Sanchez and Sotirios A. Tsaftaris. *Diffusion Causal Models for Counterfactual Estimation*. URL: https://arxiv.org/pdf/2202.10166. arXiv:2202.10166v1 [cs.LG] 21 Feb 2022.

[40] Pedro Sanchez et al. *What is Healthy? Generative Counterfactual Diffusion for Lesion Localization*. URL: https://arxiv.org/pdf/2207.12268. arXiv:2207.12268v1 [cs.CV] 25 Jul 2022.

[41] Jascha Sohl-Dickstein et al. *Deep Unsupervised Learning using Nonequilibrium Thermodynamics*. URL: https://arxiv.org/pdf/1503.03585.pdf. arXiv:1503.03585v8.

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. *DENOISING DIFFUSION IMPLICIT MODELS*. URL: https://arxiv.org/pdf/2010.02502.pdf. arXiv:2010.02502v4 [cs.LG] 5 Oct 2022.

[43] Yang Song and Stefano Ermon. *Generative Modeling by Estimating Gradients of the Data Distribution*. URL: https://arxiv.org/pdf/1907.05600.pdf. arXiv:1907.05600v3 [cs.LG] 10 Oct 2020.

[44] Yang Song et al. *SCORE-BASED GENERATIVE MODELING THROUGH STOCHASTIC DIFFERENTIAL EQUATIONS*. URL: https://arxiv.org/pdf/2011.13456.pdf. arXiv:2011.13456v2 [cs.LG] 10 Feb 2021.

[45] Hai Siong Tan, Kuancheng Wang, and Rafe Mcbeth. *Exploring UMAP in hybrid models of entropy-based and representativeness sampling for active learning in biomedical segmentation*. URL: https://arxiv.org/pdf/2312.10361. arXiv:2312.10361v2 [cs.CV] 27 May 2024.

[46] Naofumi Tomita et al. *Automatic Post-Stroke Lesion Segmentation on MR Images using 3D Residual Convolutional Neural Network*. URL: https://arxiv.org/pdf/1911.11209.

[47] Ashish Vaswani et al. *Attention Is All You Need*. URL: https://arxiv.org/pdf/1706.03762. arXiv:1706.03762v7 [cs.CL] 2 Aug 2023.

[48] Khushboo Verma, Satwant Kumar, and David Paydarfar. *Automatic Segmentation and Quantitative Assessment of Stroke Lesions on MR Images*. URL: https://www.mdpi.com/2075-4418/12/9/2055.

[49] Khushboo Verma, Satwant Kumar, and David Paydarfar. *Automatic Segmentation and Quantitative Assessment of Stroke Lesions on MR Images*. URL: https://www.mdpi.com/2075-4418/12/9/2055.

[50] Julia Wolleb et al. *Diffusion Models for Medical Anomaly Detection*. URL: https://arxiv.org/pdf/2203.04306. arXiv:2203.04306v2 [eess.IV] 5 Oct 2022.