



Σχολή Θετικών Επιστημών & Τεχνολογίας
Μεταπτυχιακή Εξειδίκευση στα Πληροφορικά Συστήματα

**Δημιουργία Ρεαλιστικών Εικόνων με Χρήση
Προεκπαιδευμένων Μοντέλων Γενεσιουργού Τεχνητής
Νοημοσύνης από την Πλατφόρμα Hugging Face.**

Παρασκευή Σταματούκου

Επιτροπή Επίβλεψης Διπλωματικής Εργασίας

Επιβλέπων Καθηγητής:
Σωτήριος Κωτσιαντής

Συν-Επιβλέπων Καθηγητής:
Κωνσταντίνος Παξιμάδης

Αθήνα , Ιούνιος 2026

Η παρούσα εργασία αποτελεί πνευματική ιδιοκτησία της φοιτήτριας Παρασκευής Σταματούκου που την εκπόνησε. Στο πλαίσιο της πολιτικής ανοικτής πρόσβασης ο συγγραφέας/δημιουργός εκχωρεί στο ΕΑΠ, μη αποκλειστική άδεια χρήσης του δικαιώματος αναπαραγωγής, προσαρμογής, δημόσιου δανεισμού, παρουσίασης στο κοινό και ψηφιακής διάχυσής τους διεθνώς, σε ηλεκτρονική μορφή και σε οποιοδήποτε μέσο, για διδακτικούς και ερευνητικούς σκοπούς, άνευ ανταλλάγματος και για όλο το χρόνο διάρκειας των δικαιωμάτων πνευματικής ιδιοκτησίας. Η ανοικτή πρόσβαση στο πλήρες κείμενο για μελέτη και ανάγνωση δεν σημαίνει καθ' οιονδήποτε τρόπο παραχώρηση δικαιωμάτων διανοητικής ιδιοκτησίας του συγγραφέα/δημιουργού ούτε επιτρέπει την αναπαραγωγή, αναδημοσίευση, αντιγραφή, αποθήκευση, πώληση, εμπορική χρήση, μετάδοση, διανομή, έκδοση, εκτέλεση, «μεταφόρτωση» (downloading), «ανάρτηση» (uploading), μετάφραση, τροποποίηση με οποιονδήποτε τρόπο, τμηματικά ή περιληπτικά της εργασίας, χωρίς τη ρητή προηγούμενη έγγραφη συναίνεση του συγγραφέα/δημιουργού. Ο συγγραφέας/δημιουργός διατηρεί το σύνολο των ηθικών και περιουσιακών του δικαιωμάτων.

Περίληψη

Η ραγδαία εξέλιξη της Γενεσιουργού Τεχνητής Νοημοσύνης (Generative Artificial Intelligence) έχει οδηγήσει στην ανάπτυξη ισχυρών μοντέλων δημιουργίας εικόνων από κείμενο (text-to-image), με κυριότερους εκπροσώπους τα diffusion models. Τα μοντέλα αυτά αξιοποιούν τεχνικές βαθιάς μάθησης για τη δημιουργία ρεαλιστικών εικόνων υψηλής ποιότητας, παρουσιάζοντας σημαντικές εφαρμογές σε τομείς όπως η δημιουργική βιομηχανία, η εκπαίδευση και η ιατρική απεικόνιση.

Αντικείμενο της παρούσας διπλωματικής εργασίας είναι η συγκριτική αξιολόγηση προεκπαιδευμένων μοντέλων δημιουργίας εικόνων που είναι διαθέσιμα μέσω της πλατφόρμας Hugging Face, με έμφαση στα μοντέλα **Stable Diffusion v1.5** και **Stable Diffusion v2.1**. Παράλληλα, μελετώνται η αρχιτεκτονική των diffusion models, οι τεχνικές προσαρμογής (fine-tuning) και τα βασικά εργαλεία του οικοσυστήματος Hugging Face.

Στο πειραματικό μέρος εφαρμόστηκαν τεχνικές **Prompt Engineering**, δημιουργώντας πέντε διαφορετικά σύνολα προτροπών, με στόχο τη διερεύνηση της επίδρασής τους στην ποιότητα των παραγόμενων εικόνων. Για την αντικειμενική αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκαν οι μετρικές **CLIP Score**, **Fréchet Inception Distance (FID)** και **Aesthetic Score**, ενώ πραγματοποιήθηκε και στατιστική ανάλυση των αποτελεσμάτων μέσω **t-test** και **p-value**. Επιπλέον, πραγματοποιήθηκε ποιοτική αξιολόγηση των εικόνων με βάση τον ρεαλισμό, τη σημασιολογική συνέπεια ως προς το prompt και την οπτική τους ποιότητα.

Στο πλαίσιο της εργασίας πραγματοποιήθηκε επίσης fine-tuning του μοντέλου **Stable Diffusion v1.5** με την τεχνική **Low-Rank Adaptation (LoRA)**, χρησιμοποιώντας τη

βάση **COVID-19 Radiography Database**, με στόχο την παραγωγή συνθετικών ακτινογραφιών θώρακα. Η εκπαίδευση υλοποιήθηκε μέσω της βιβλιοθήκης **Diffusers** της Hugging Face σε περιβάλλον **Google Colab**, αξιοποιώντας κάρτα γραφικών **NVIDIA Tesla T4 GPU**.

Τα αποτελέσματα έδειξαν ότι τα μοντέλα Stable Diffusion v1.5 και Stable Diffusion v2.1 παρουσιάζουν παρόμοια απόδοση ως προς τη σημασιολογική αντιστοίχιση εικόνας-κειμένου, ενώ το Stable Diffusion v2.1 εμφανίζει ελαφρά βελτιωμένη συμπεριφορά σε πιο σύνθετες περιγραφές. Παράλληλα, η τεχνική **LoRA** απέδειξε ότι αποτελεί μία ιδιαίτερα αποδοτική λύση για την προσαρμογή προεκπαιδευμένων diffusion models σε εξειδικευμένα ιατρικά δεδομένα, απαιτώντας περιορισμένους υπολογιστικούς πόρους και μικρό χρόνο εκπαίδευσης. Συνολικά, η εργασία αναδεικνύει τις δυνατότητες των σύγχρονων diffusion models και των τεχνικών fine-tuning στην παραγωγή συνθετικών εικόνων υψηλής ποιότητας και συμβάλλει στην αξιολόγηση και αξιοποίησή τους σε εξειδικευμένα πεδία εφαρμογών.

Λέξεις-κλειδιά: Γενεσιουργός Τεχνητή Νοημοσύνη, Diffusion Models, Stable Diffusion, Hugging Face, Prompt Engineering, CLIP Score, Fréchet Inception Distance (FID), Aesthetic Score, LoRA, Fine-Tuning, Ιατρική Απεικόνιση.

Abstract

Recent advances in **Generative Artificial Intelligence (Generative AI)** have led to the rapid development of text-to-image diffusion models capable of generating high-quality and realistic images from natural language descriptions. These models have attracted significant research interest due to their wide range of applications in digital content creation, education, healthcare, and scientific research.

The objective of this thesis is to evaluate and compare pretrained text-to-image diffusion models available through the **Hugging Face** platform, with particular emphasis on **Stable Diffusion v1.5** and **Stable Diffusion v2.1**. The study also examines the architecture of diffusion models, fine-tuning techniques, and the software ecosystem supporting their development and deployment.

The experimental evaluation consisted of two main stages. First, various **prompt engineering** strategies were applied to investigate their impact on image generation quality. Five prompt categories were created, including baseline prompts and four prompt engineering techniques: **context tokens**, **emphasis**, **repetition**, and **word order**. The generated images were quantitatively evaluated using **CLIP Score**, **Fréchet Inception Distance (FID)**, and **Aesthetic Score**, while statistical analysis was performed using **t-tests** and **p-values**. Furthermore, a qualitative assessment was conducted to evaluate semantic consistency, realism, visual quality, and common generation artifacts.

The second stage involved **fine-tuning Stable Diffusion v1.5** using the **Low-Rank Adaptation (LoRA)** technique and the **COVID-19 Radiography Database** to generate synthetic chest X-ray images. The training process was implemented using the **Diffusers** library within the Hugging Face ecosystem in a **Google Colab** environment equipped with an **NVIDIA Tesla T4 GPU**.

The experimental results indicate that **Stable Diffusion v1.5** and **Stable Diffusion v2.1** exhibit comparable performance regarding text-image semantic alignment, while Stable Diffusion v2.1 demonstrates a slight advantage in understanding more complex prompts. The findings also confirm that **LoRA** provides an efficient and computationally lightweight fine-tuning approach, enabling the adaptation of pretrained diffusion models to specialized medical imaging datasets while maintaining high image quality.

Overall, this thesis demonstrates the effectiveness of modern diffusion models for text-to-image generation and highlights the practical value of prompt engineering, quantitative evaluation metrics, and parameter-efficient fine-tuning techniques for domain-specific image synthesis.

Keywords: Generative Artificial Intelligence, Diffusion Models, Stable Diffusion, Hugging Face, Prompt Engineering, CLIP Score, Fréchet Inception Distance (FID), Aesthetic Score, Low-Rank Adaptation (LoRA), Medical Image Generation.

Περιεχόμενα

Περίληψη.....	2
Abstract	4
Εισαγωγή	8
1.1 Σκοπός και Σημασία της Εργασίας	8
1.2 Ερευνητικά Ερωτήματα και Στόχοι	10
1.3 Δομή της Διπλωματικής	12
1.4 Επιστημονική Συνεισφορά της Εργασίας.....	15
Θεωρητικό Υπόβαθρο	17
2.1 Εισαγωγή στη Γενεσιουργό Τεχνητή Νοημοσύνη.....	17
2.2 Αρχές της Μετασχηματιστικής Μάθησης (Transformers)	18
2.3 Διάκριση μεταξύ Κειμένου-σε-Εικόνα και Εικόνα-σε-Εικόνα Μοντέλων	22
2.4 Τεχνικές Εκπαίδευσης και Fine-tuning Μοντέλων ΓΤΝ.....	25
2.4.1 DreamBooth	25
2.4.2 LORA - Low-Rank Adaptation	27
2.4.3 LoRA στο Stable Diffusion.....	28
2.4.4 Σύγκριση DreamBooth και LoRA	29
2.5 Αξιολόγηση Ποιότητας Γεννημένων Εικόνων (π.χ. FID, IS, CLIP score, aesthetic score)33	
Πλατφόρμα Hugging Face	37
3.1 Εισαγωγή και Αρχιτεκτονική της Πλατφόρμας	37
3.2 Model Hub και Προεκπαιδευμένα Μοντέλα	39
3.3 Hugging Face Datasets και Diffusers Library	41
3.4 Εργαλεία για Εικόνες: Diffusers, Transformers, Accelerate.....	43
Προεκπαιδευμένα Μοντέλα Γενεσιουργού ΤΝ για Εικόνες	45
4.1 Stable Diffusion	45
4.1.1 Αρχιτεκτονική Stable Diffusion.....	46
4.2 DALL·E.....	49
4.3 Άλλα μοντέλα	52
4.4 Συγκριτική παρουσίαση χαρακτηριστικών	55
4.5 Κριτική Ανάλυση των Σύγχρονων Μοντέλων Δημιουργίας Εικόνων	56
Πειραματικό Μέρος και Αποτελέσματα	60
5.1 Ορισμός Σεναρίων και Prompts	60
5.2 Πειραματική Διαδικασία	61
5.3 Ποιοτική και Ποσοτική Αξιολόγηση των Εικόνων	63
5.3.1 Prompt engineering - Clip score	63
5.3.2 FID και Clip score - Coco database.....	66
5.3.3 Aesthetic score - COCO dataset.....	70

5.3.4 Ποιοτική αξιολόγηση.....	73
5.4 Εκπαίδευση Stable Diffusion v1.5 με LoRA	89
5.4.1 Εκπαίδευση με απλά captions , rank =16 ,learning rate =1e-4.....	90
5.4.2 Εκπαίδευση με απλά captions , rank =8 , learning rate = 3e-5.....	93
5.5 Σύγκριση Μοντέλων και Προσεγγίσεων	97
Συμπεράσματα και Μελλοντικές Κατευθύνσεις.....	99
6.1 Κύρια Συμπεράσματα.....	99
6.2 Προτάσεις για Μελλοντική Έρευνα.....	102
ΒΙΒΛΙΟΓΡΑΦΙΑ	104
ΠΑΡΑΡΤΗΜΑΤΑ	110
Παράρτημα Α' : Κατηγορίες prompt και Clip score	110
Παράρτημα Β' : Scripts – prompt engineering.....	123
Παράρτημα Γ' : Script για FID –CLIP score από COCO dataset.....	128
Παράρτημα Δ' : Script - Aesthetic score.....	132

Εισαγωγή

1.1 Σκοπός και Σημασία της Εργασίας

Σκοπός της εργασίας

Σκοπός της παρούσας εργασίας είναι η συγκριτική αξιολόγηση προεκπαιδευμένων μοντέλων διάχυσης για τη δημιουργία εικόνων από κείμενο, καθώς και η διερεύνηση της επίδρασης τεχνικών Prompt Engineering στην ποιότητα των παραγόμενων εικόνων. Η αξιολόγηση πραγματοποιείται με συνδυασμό ποσοτικών μετρικών (CLIP Score, Fréchet Inception Distance (FID) και Aesthetic Score) και ποιοτικής αξιολόγησης, ενώ εξετάζεται επίσης η δυνατότητα προσαρμογής των μοντέλων σε εξειδικευμένο ιατρικό πεδίο μέσω fine-tuning με την τεχνική Low-Rank Adaptation (LoRA).

Σημασία της Εργασίας

Η ραγδαία ανάπτυξη των μοντέλων γενεσιουργού τεχνητής νοημοσύνης (Generative AI) έχει δημιουργήσει νέες δυνατότητες στη δημιουργία εικόνων από φυσική γλώσσα (text-to-image generation). Παρά τη μεγάλη πρόοδο των diffusion models, η αντικειμενική αξιολόγηση της ποιότητας των παραγόμενων εικόνων εξακολουθεί να αποτελεί πρόκληση, καθώς καμία μεμονωμένη μετρική δεν μπορεί να αποτυπώσει πλήρως όλες τις διαστάσεις της ποιότητας, όπως η σημασιολογική αντιστοίχιση με το κείμενο, ο φωτορεαλισμός και η αισθητική ποιότητα.

Η παρούσα εργασία συμβάλλει στη συστηματική αξιολόγηση των προεκπαιδευμένων μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1, συνδυάζοντας ποσοτικές μετρικές (CLIP Score, Fréchet Inception Distance και Aesthetic Score) με ποιοτική αξιολόγηση των παραγόμενων εικόνων. Παράλληλα, διερευνάται η επίδραση

διαφορετικών τεχνικών Prompt Engineering στη σημασιολογική ευθυγράμμιση μεταξύ του κειμένου και της παραγόμενης εικόνας, προσφέροντας χρήσιμα συμπεράσματα για τη βελτίωση της διαδικασίας δημιουργίας εικόνων.

Ιδιαίτερη σημασία έχει επίσης η διερεύνηση της δυνατότητας προσαρμογής ενός προεκπαιδευμένου μοντέλου μέσω της τεχνικής Low-Rank Adaptation (LoRA) σε ένα εξειδικευμένο πεδίο, όπως η παραγωγή συνθετικών ακτινογραφιών θώρακα. Η προσέγγιση αυτή αναδεικνύει τις δυνατότητες αξιοποίησης των diffusion models σε εφαρμογές της ιατρικής απεικόνισης, όπου η δημιουργία συνθετικών δεδομένων μπορεί να υποστηρίξει την έρευνα και την ανάπτυξη συστημάτων τεχνητής νοημοσύνης.

Συνολικά, η εργασία προτείνει ένα ολοκληρωμένο πλαίσιο αξιολόγησης μοντέλων δημιουργίας εικόνας, το οποίο συνδυάζει διαφορετικές τεχνικές Prompt Engineering, ποσοτικές μετρικές και ποιοτική αξιολόγηση. Το πλαίσιο αυτό μπορεί να αξιοποιηθεί τόσο για τη συγκριτική αξιολόγηση diffusion models όσο και για τη μελλοντική ανάπτυξη και βελτιστοποίηση μοντέλων δημιουργίας εικόνας σε γενικές και εξειδικευμένες εφαρμογές.

1.2 Ερευνητικά Ερωτήματα και Στόχοι

Ερευνητικά ερωτήματα

Η παρούσα εργασία διερευνά την απόδοση προεκπαιδευμένων μοντέλων διάχυσης (diffusion models) στη δημιουργία εικόνων από κείμενο (text-to-image generation), καθώς και τη δυνατότητα προσαρμογής τους σε εξειδικευμένες εφαρμογές. Για την επίτευξη του σκοπού αυτού διατυπώνονται τα ακόλουθα ερευνητικά ερωτήματα:

Για την επίτευξη του σκοπού αυτού διατυπώνονται τα ακόλουθα ερευνητικά ερωτήματα:

- **E1.** Υπάρχουν διαφορές στην απόδοση των μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1 ως προς τη δημιουργία εικόνων από κείμενο;
- **E2.** Σε ποιο βαθμό οι τεχνικές Prompt Engineering (context tokens, emphasis, repetition και αλλαγή της σειράς των λέξεων) επηρεάζουν τη σημασιολογική αντιστοίχιση μεταξύ του prompt και της παραγόμενης εικόνας;
- **E3.** Πώς αξιολογείται η ποιότητα των παραγόμενων εικόνων με τη συνδυαστική χρήση ποσοτικών μετρικών (CLIP Score, FID και Aesthetic Score) και ποιοτικής αξιολόγησης;
- **E4.** Είναι δυνατή η αποτελεσματική προσαρμογή ενός προεκπαιδευμένου μοντέλου Stable Diffusion μέσω της τεχνικής Low-Rank Adaptation (LoRA) για τη δημιουργία συνθετικών ακτινογραφιών θώρακα;

Στόχοι της Εργασίας

Με βάση τα παραπάνω ερωτήματα, οι βασικοί στόχοι της εργασίας είναι:

1. **Η συγκριτική αξιολόγηση** των μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1 ως προς την παραγωγή εικόνων από κείμενο.

2. **Η διερεύνηση της επίδρασης τεχνικών Prompt Engineering** (context tokens, emphasis, repetition και αλλαγή της σειράς των λέξεων) στη σημασιολογική αντιστοίχιση μεταξύ του prompt και της παραγόμενης εικόνας.
3. **Η ποσοτική αξιολόγηση** των παραγόμενων εικόνων με τη χρήση των μετρικών CLIP Score, Fréchet Inception Distance (FID) και Aesthetic Score.
4. **Η ποιοτική αξιολόγηση** των παραγόμενων εικόνων μέσω οπτικής σύγκρισης, με στόχο την αναγνώριση πλεονεκτημάτων και περιορισμών που δεν αποτυπώνονται πλήρως από τις ποσοτικές μετρικές.
5. **Η διερεύνηση της δυνατότητας προσαρμογής** ενός προεκπαιδευμένου μοντέλου Stable Diffusion σε εξειδικευμένο ιατρικό πεδίο μέσω fine-tuning με την τεχνική Low-Rank Adaptation (LoRA) για τη δημιουργία συνθετικών ακτινογραφιών θώρακα.
6. **Η διαμόρφωση ενός ολοκληρωμένου πλαισίου αξιολόγησης** μοντέλων δημιουργίας εικόνας, το οποίο συνδυάζει τεχνικές Prompt Engineering, ποσοτικές μετρικές και ποιοτική αξιολόγηση για την αποτίμηση της απόδοσης diffusion models.

1.3 Δομή της Διπλωματικής

Η παρούσα διπλωματική εργασία οργανώνεται σε έξι κεφάλαια, τα οποία καλύπτουν τόσο το θεωρητικό υπόβαθρο όσο και το πειραματικό μέρος της έρευνας.

Στο **Κεφάλαιο 1** παρουσιάζεται η εισαγωγή της εργασίας, το αντικείμενο της έρευνας, η σημασία της, οι στόχοι και τα ερευνητικά ερωτήματα που διερευνώνται, καθώς και η συνοπτική διάρθρωση της διπλωματικής.

Στο **Κεφάλαιο 2** παρουσιάζεται το θεωρητικό υπόβαθρο της εργασίας. Αρχικά γίνεται αναφορά στις βασικές έννοιες της Γενεσιουργού Τεχνητής Νοημοσύνης (Generative Artificial Intelligence) και στα μεγάλα γλωσσικά μοντέλα (Large Language Models). Στη συνέχεια περιγράφονται οι αρχές λειτουργίας των Transformers, των diffusion models. Παρουσιάζονται επίσης οι τεχνικές fine-tuning, όπως το DreamBooth και το Low-Rank Adaptation (LoRA).

Στο **Κεφάλαιο 3** παρουσιάζονται η **πλατφόρμα Hugging Face** και τα βασικά εργαλεία που χρησιμοποιήθηκαν στην εργασία, όπως οι βιβλιοθήκες **Diffusers**, **Transformers**, **Accelerate** και **Datasets**, καθώς και το **Hugging Face Model Hub**. Περιγράφονται οι δυνατότητες των εργαλείων αυτών για την ανάπτυξη, εκπαίδευση και αξιοποίηση μοντέλων γενεσιουργού τεχνητής νοημοσύνης, με ιδιαίτερη έμφαση στην εκπαίδευση και προσαρμογή μοντέλων Stable Diffusion μέσω της τεχνικής LoRA.

Στο **Κεφάλαιο 4** παρουσιάζονται τα σύγχρονα μοντέλα δημιουργίας εικόνων με έμφαση στο **Stable Diffusion**. Αρχικά αναλύεται η αρχιτεκτονική του Stable Diffusion καθώς και οι βασικοί μηχανισμοί λειτουργίας τους, όπως το Variational

Autoencoder (VAE), το U-Net, η διαδικασία διάχυσης (diffusion process) και ο μηχανισμός cross-attention. Στη συνέχεια παρουσιάζονται οι σημαντικότερες εκδόσεις του Stable Diffusion, καθώς και άλλα σύγχρονα μοντέλα δημιουργίας εικόνων, όπως τα **DALL·E**, **Imagen**, **FLUX** και άλλα diffusion models. Το κεφάλαιο ολοκληρώνεται με συγκριτική παρουσίαση των χαρακτηριστικών τους και κριτική ανάλυση των σύγχρονων προσεγγίσεων δημιουργίας εικόνων.

Στο **Κεφάλαιο 5** περιγράφεται η μεθοδολογία της έρευνας και η πειραματική διαδικασία. Παρουσιάζονται τα μοντέλα Stable Diffusion v1.5 και Stable Diffusion v2.1 που χρησιμοποιήθηκαν, οι τεχνικές Prompt Engineering που εφαρμόστηκαν, οι βάσεις δεδομένων COCO και COVID-19 Radiography Database, καθώς και το υπολογιστικό περιβάλλον υλοποίησης των πειραμάτων. Επιπλέον, περιγράφεται αναλυτικά η διαδικασία fine-tuning του Stable Diffusion v1.5 με την τεχνική LoRA.

Επίσης παρουσιάζονται τα αποτελέσματα των πειραμάτων. Αρχικά αναλύονται τα αποτελέσματα των τεχνικών Prompt Engineering μέσω του CLIP Score, ενώ στη συνέχεια παρουσιάζονται οι μετρικές FID και Aesthetic Score. Επιπλέον, πραγματοποιείται ποιοτική αξιολόγηση των παραγόμενων εικόνων και συγκρίνονται οι επιδόσεις των δύο μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1. Τέλος, παρουσιάζονται τα αποτελέσματα του fine-tuning με LoRA για την παραγωγή συνθετικών ακτινογραφιών θώρακα.

Στο **Κεφάλαιο 6** συνοψίζονται τα βασικά συμπεράσματα της έρευνας σε σχέση με τα ερευνητικά ερωτήματα και τους στόχους της εργασίας. Παράλληλα, συζητούνται οι περιορισμοί της μελέτης και διατυπώνονται προτάσεις για μελλοντική έρευνα, όπως η αξιολόγηση νεότερων diffusion models, η χρήση μεγαλύτερων και περισσότερο

εξειδικευμένων συνόλων δεδομένων και η διερεύνηση εναλλακτικών τεχνικών fine-tuning.

Τέλος, στο **Παράρτημα** παρατίθενται συμπληρωματικά στοιχεία της έρευνας, όπως τα πλήρη σύνολα των prompts, οι παραγόμενες εικόνες που χρησιμοποιήθηκαν στην ποιοτική αξιολόγηση, επιπλέον πίνακες και διαγράμματα, καθώς και αποσπάσματα του κώδικα και των πειραματικών υλοποιήσεων.

1.4 Επιστημονική Συνεισφορά της Εργασίας

Η παρούσα διπλωματική εργασία συμβάλλει στη μελέτη και αξιολόγηση σύγχρονων προεκπαιδευμένων μοντέλων γενεσιουργού τεχνητής νοημοσύνης για τη δημιουργία εικόνων μέσω της πλατφόρμας Hugging Face. Σε αντίθεση με εργασίες που περιορίζονται στη θεωρητική παρουσίαση των diffusion models, η παρούσα μελέτη συνδυάζει βιβλιογραφική ανάλυση, πειραματική αξιολόγηση και πρακτική προσαρμογή μοντέλων σε εξειδικευμένα δεδομένα.

Οι βασικές συνεισφορές της εργασίας συνοψίζονται ως εξής:

1. Παρουσιάζεται μια συστηματική επισκόπηση των κυριότερων προεκπαιδευμένων μοντέλων δημιουργίας εικόνων, με έμφαση στα Stable Diffusion, DALL·E και συναφή diffusion-based μοντέλα, αναδεικνύοντας τα χαρακτηριστικά, τις δυνατότητες και τους περιορισμούς τους.
2. Υλοποιείται πειραματικό πλαίσιο δημιουργίας εικόνων μέσω της βιβλιοθήκης Diffusers της πλατφόρμας Hugging Face, επιτρέποντας την αξιολόγηση διαφορετικών παραμέτρων παραγωγής και τεχνικών βελτιστοποίησης prompts.
3. Διερευνάται η επίδραση διαφορετικών διατυπώσεων prompts και παραμέτρων παραγωγής στην ποιότητα και πιστότητα των παραγόμενων εικόνων, αναδεικνύοντας κρίσιμους παράγοντες που επηρεάζουν το τελικό αποτέλεσμα.
4. Πραγματοποιείται προσαρμογή (fine-tuning) του Stable Diffusion μέσω της τεχνικής LoRA, αξιολογώντας τη δυνατότητα εξειδίκευσης του μοντέλου σε συγκεκριμένα οπτικά χαρακτηριστικά με περιορισμένες υπολογιστικές απαιτήσεις.

5. Παρουσιάζεται συγκριτική αξιολόγηση των παραγόμενων αποτελεσμάτων με χρήση ποιοτικών και ποσοτικών κριτηρίων, όπως η οπτική ποιότητα, η πιστότητα ως προς το κείμενο εισόδου και η υπολογιστική αποδοτικότητα.
6. Καταγράφονται τα πλεονεκτήματα και οι περιορισμοί των σύγχρονων προεκπαιδευμένων diffusion models σε πραγματικά σενάρια χρήσης, παρέχοντας κατευθύνσεις για μελλοντική έρευνα και εφαρμογές στον χώρο της δημιουργίας ψηφιακού περιεχομένου.

Θεωρητικό Υπόβαθρο

2.1 Εισαγωγή στη Γενεσιουργό Τεχνητή Νοημοσύνη

Η Γενεσιουργό Τεχνητή Νοημοσύνη (Generative AI) (GAI) δημιουργεί νέο περιεχόμενο, όπως κείμενο, εικόνες και ακόμη και ήχο, το οποίο γίνεται ολοένα και πιο δύσκολο να διακριθεί από έργο που έχει παραχθεί από ανθρώπους. Η GAI έγινε γνωστή παγκόσμια στις 30 Νοεμβρίου 2022, όταν η OpenAI κυκλοφόρησε το ChatGPT, ένα σύστημα συνομιλίας στο οποίο ο χρήστης μπορεί να δίνει ερωτήσεις και να παίρνει απαντήσεις.

Τα μοντέλα Γενεσιουργού Τεχνητής Νοημοσύνης, έχουν τεράστια δεδομένα και πολύπλοκες δομές και μπορούν να μάθουν να παράγουν νέο περιεχόμενο σε διαφορετικές μορφές, όπως κείμενο, βίντεο και ήχο, με βάση τα δεδομένα που τους δίνονται. Για παράδειγμα, η Γενεσιουργό Τεχνητή Νοημοσύνη (Generative AI) μπορεί να χρησιμοποιηθεί για τη δημιουργία ρεαλιστικών εικόνων από περιγραφές κειμένου, καθώς και για την παραγωγή βίντεο από γραπτές περιγραφές.

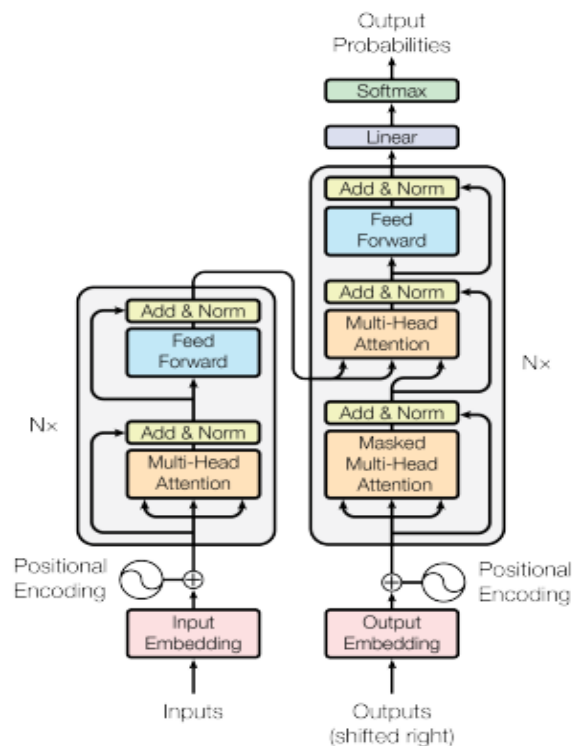
Η Γενεσιουργό Τεχνητή Νοημοσύνη (Generative AI), σηματοδοτήθηκε από την εισαγωγή των νευρωνικών δικτύων, ενός συγκεκριμένου τύπου μηχανικής μάθησης. Ένα νευρωνικό δίκτυο εκπαιδεύεται ώστε να αναγνωρίζει πρότυπα μέσα σε ένα σύνολο δεδομένων. Όταν το νευρωνικό δίκτυο εκπαιδευτεί, μπορεί να παίρνει αποφάσεις ή να πραγματοποιεί προβλέψεις χωρίς να είναι ρητά προγραμματισμένο για την εκτέλεση συγκεκριμένων εργασιών.

2.2 Αρχές της Μετασχηματιστικής Μάθησης (Transformers)

Η αρχιτεκτονική **Transformer** παρουσιάστηκε το 2017 από τους Vaswani et al. στο άρθρο *Attention Is All You Need* [5] και αποτέλεσε σημείο καμπής στην εξέλιξη της επεξεργασίας φυσικής γλώσσας (Natural Language Processing – NLP). Σε αντίθεση με προηγούμενες αρχιτεκτονικές, όπως τα Αναδρομικά Νευρωνικά Δίκτυα (RNNs) και τα Long Short-Term Memory (LSTM), οι Transformers βασίζονται αποκλειστικά στον μηχανισμό **attention**, επιτρέποντας την παράλληλη επεξεργασία ολόκληρων ακολουθιών και τη μοντελοποίηση μακροπρόθεσμων εξαρτήσεων μεταξύ των στοιχείων τους.

Η δυνατότητα αυτή οδήγησε σε σημαντική βελτίωση της απόδοσης σε πληθώρα εφαρμογών, όπως η μηχανική μετάφραση, η περίληψη κειμένου, η απάντηση σε ερωτήσεις, η παραγωγή φυσικής γλώσσας και, αργότερα, η γενεσιουργός τεχνητή νοημοσύνη.

Η βασική αρχιτεκτονική του Transformer ακολουθεί τη δομή **Encoder–Decoder**, όπως παρουσιάζεται στο Σχήμα 2.1. Ο **Encoder** μετατρέπει την είσοδο σε διανυσματικές σημασιολογικές αναπαραστάσεις (embeddings), ενώ ο **Decoder** χρησιμοποιεί τις αναπαραστάσεις αυτές για την παραγωγή της εξόδου. Κάθε επίπεδο του Encoder και του Decoder αποτελείται από μηχανισμούς **Multi-Head Attention**, πλήρως συνδεδεμένα δίκτυα προώθησης (Feed-Forward Networks), συνδέσεις υπολοίπου (Residual Connections) και κανονικοποίηση επιπέδου (Layer Normalization), επιτυγχάνοντας σταθερότερη και αποδοτικότερη εκπαίδευση.



Σχήμα 2.1. Βασική αρχιτεκτονική του Transformer, αποτελούμενη από τον Encoder και τον Decoder, όπως παρουσιάστηκε από τους Vaswani et al. (2017) στο άρθρο *Attention Is All You Need*. [5].

Μηχανισμός Attention

Η σημαντικότερη καινοτομία των Transformers είναι ο μηχανισμός **Attention**, ο οποίος επιτρέπει στο μοντέλο να εντοπίζει τα σημαντικότερα στοιχεία της εισόδου κατά την παραγωγή της εξόδου. Αντί να επεξεργάζεται τις λέξεις σειριακά, όπως τα RNNs, ο μηχανισμός attention υπολογίζει ταυτόχρονα τις σχέσεις μεταξύ όλων των στοιχείων της ακολουθίας, επιτρέποντας την αποτελεσματικότερη κατανόηση του συμφραζομένου.

Η λειτουργία του μηχανισμού προσοχής περιγράφεται από τη σχέση:

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

όπου:

- **Q (Query):** αναπαράσταση του ερωτήματος (query),
- **K (Key):** αναπαράσταση των χαρακτηριστικών που χρησιμοποιούνται για τον υπολογισμό της συνάφειας,
- **V (Value):** οι πληροφορίες που μεταφέρονται στην έξοδο,
- **d:** η διάσταση των διανυσμάτων (embedding dimension).

Self-Attention και Multi-Head Attention

Ο μηχανισμός **Self-Attention** επιτρέπει σε κάθε στοιχείο μιας ακολουθίας να λαμβάνει πληροφορία από όλα τα υπόλοιπα στοιχεία της ίδιας ακολουθίας, δημιουργώντας πλουσιότερες σημασιολογικές αναπαραστάσεις. Η επέκτασή του, γνωστή ως **Multi-Head Attention**, εκτελεί πολλαπλούς μηχανισμούς attention παράλληλα, επιτρέποντας στο μοντέλο να εξετάζει διαφορετικούς τύπους σχέσεων μεταξύ των δεδομένων. Οι έξοδοι όλων των μηχανισμών συνενώνονται και μετασχηματίζονται γραμμικά, οδηγώντας σε ακριβέστερες αναπαραστάσεις.

Feed-Forward Network, Residual Connections και Layer Normalization

Μετά τον μηχανισμό attention, κάθε επίπεδο του Transformer περιλαμβάνει ένα πλήρως συνδεδεμένο δίκτυο προώθησης (**Feed-Forward Network – FFN**), το οποίο εφαρμόζεται ανεξάρτητα σε κάθε θέση της ακολουθίας και ενισχύει τη μη γραμμική εκφραστικότητα του μοντέλου.

Παράλληλα, χρησιμοποιούνται **Residual Connections**, οι οποίες επιτρέπουν την ευκολότερη διάδοση της πληροφορίας στα βαθύτερα επίπεδα του δικτύου, καθώς και **Layer Normalization**, η οποία σταθεροποιεί την εκπαίδευση και επιταχύνει τη σύγκλιση του μοντέλου.

Η συνδυασμένη λειτουργία των παραπάνω μηχανισμών επιτρέπει στους Transformers να επεξεργάζονται αποτελεσματικά πολύ μεγάλες ακολουθίες δεδομένων και να εκπαιδεύονται παράλληλα σε μεγάλα σύνολα δεδομένων, γεγονός που τους κατέστησε τη βάση για την ανάπτυξη σύγχρονων μοντέλων, όπως τα **BERT**, **GPT**, **CLIP** και τα μοντέλα δημιουργίας εικόνων **Stable Diffusion**.

2.3 Διάκριση μεταξύ Κειμένου-σε-Εικόνα και Εικόνα-σε-Εικόνα

Μοντέλων

Η δημιουργία εικόνων με τεχνητή νοημοσύνη (AI Image Generation) αποτελεί έναν από τους ταχύτερα αναπτυσσόμενους τομείς της γενεσιουργού τεχνητής νοημοσύνης. Ανάλογα με τη μορφή της εισόδου που χρησιμοποιείται για τη δημιουργία της εικόνας, τα μοντέλα παραγωγής εικόνων διακρίνονται κυρίως σε δύο κατηγορίες:

Κείμενο-σε-Εικόνα (Text-to-Image) και **Εικόνα-σε-Εικόνα (Image-to-Image)**.

Μοντέλα Κειμένου-σε-Εικόνα (Text-to-Image)

Τα μοντέλα **Text-to-Image** δημιουργούν εικόνες ξεκινώντας αποκλειστικά από μία κειμενική περιγραφή (prompt). Κατά την εκπαίδευσή τους χρησιμοποιούν μεγάλα σύνολα δεδομένων που αποτελούνται από εικόνες και τις αντίστοιχες περιγραφές τους (captions), μαθαίνοντας τη σημασιολογική αντιστοίχιση μεταξύ της φυσικής γλώσσας και των οπτικών χαρακτηριστικών.

Κατά τη διαδικασία παραγωγής, το μοντέλο μετατρέπει το κείμενο σε σημασιολογικές αναπαραστάσεις (text embeddings) και δημιουργεί μία νέα εικόνα που είναι συνεπής με το περιεχόμενο της περιγραφής. Η ποιότητα της παραγόμενης εικόνας εξαρτάται σε μεγάλο βαθμό από τη σαφήνεια και την πληρότητα του prompt, γεγονός που έχει οδηγήσει στην ανάπτυξη τεχνικών **Prompt Engineering** για τη βελτιστοποίηση των αποτελεσμάτων.

Χαρακτηριστικά παραδείγματα μοντέλων Text-to-Image αποτελούν τα **DALL·E**, **Stable Diffusion**, **Imagen** και **FLUX**, τα οποία χρησιμοποιούνται ευρέως για τη δημιουργία φωτορεαλιστικών ή καλλιτεχνικών εικόνων από φυσική γλώσσα [11].

Μοντέλα Εικόνας-σε-Εικόνα (Image-to-Image)

Τα μοντέλα **Image-to-Image** δέχονται ως είσοδο μία ήδη υπάρχουσα εικόνα και παράγουν μία νέα εικόνα, διατηρώντας ή τροποποιώντας συγκεκριμένα χαρακτηριστικά της αρχικής. Σε αντίθεση με τα μοντέλα Text-to-Image, όπου η εικόνα δημιουργείται εξ ολοκλήρου από το κείμενο, στα Image-to-Image η αρχική εικόνα λειτουργεί ως οδηγός για την παραγωγή του τελικού αποτελέσματος.

Η κατηγορία αυτή περιλαμβάνει εφαρμογές όπως:

- αποκατάσταση εικόνας (image restoration),
- υπερδειγματοληψία (super-resolution),
- συμπλήρωση ελλειπόντων περιοχών (image inpainting),
- απομάκρυνση θορύβου (image denoising),
- μεταφορά καλλιτεχνικού ύφους (style transfer),
- χρωματισμό ασπρόμαυρων εικόνων (colorization),
- εκτίμηση βάθους (depth estimation),
- σημασιολογική ή στιγμιότυπη τμηματοποίηση (semantic και instance segmentation).

Η βασική αρχή λειτουργίας των μοντέλων Image-to-Image είναι η εκμάθηση της υπό συνθήκη κατανομής της εικόνας εξόδου δεδομένης της εικόνας εισόδου. Για τον σκοπό αυτό χρησιμοποιούνται σύγχρονα γενετικά μοντέλα, όπως τα **pix2pix**, **CycleGAN**, καθώς και diffusion models, όπως το **Stable Diffusion Img2Img** και το **ControlNet**, τα οποία επιτρέπουν την παραγωγή πολλαπλών ρεαλιστικών εκδοχών της ίδιας εικόνας [22], [24].

Τα μοντέλα Image-to-Image χρησιμοποιούνται σε ένα ευρύ φάσμα εφαρμογών, όπως η ιατρική απεικόνιση, η δορυφορική τηλεπισκόπηση, η βιομηχανική επιθεώρηση, η ψηφιακή αποκατάσταση εικόνων και η καλλιτεχνική δημιουργία.

Σύγκριση των δύο προσεγγίσεων

Η βασική διαφορά μεταξύ των δύο κατηγοριών έγκειται στη μορφή της εισόδου. Στα μοντέλα **Text-to-Image**, η εικόνα δημιουργείται αποκλειστικά από μία κειμενική περιγραφή, ενώ στα μοντέλα **Image-to-Image** η παραγωγή βασίζεται σε μία ήδη υπάρχουσα εικόνα, η οποία μετασχηματίζεται σύμφωνα με τις απαιτήσεις της εφαρμογής. Στην παρούσα εργασία αξιοποιούνται κυρίως μοντέλα **Text-to-Image**, και ειδικότερα οι εκδόσεις **Stable Diffusion v1.5** και **Stable Diffusion v2.1**, καθώς αντικείμενο της έρευνας αποτελεί η αξιολόγηση της παραγωγής εικόνων από κείμενο και η προσαρμογή των μοντέλων μέσω της τεχνικής **LoRA**.

2.4 Τεχνικές Εκπαίδευσης και Fine-tuning Μοντέλων ΓΤΝ

Στα μοντέλα διάχυσης (diffusion models), όπως το **Stable Diffusion**, το fine-tuning επιτρέπει την παραγωγή εικόνων που ανταποκρίνονται σε συγκεκριμένα αντικείμενα, πρόσωπα, καλλιτεχνικά στυλ ή επιστημονικούς τομείς, όπως η ιατρική απεικόνιση. Σε σύγκριση με την εκπαίδευση ενός μοντέλου από το μηδέν, οι τεχνικές fine-tuning απαιτούν σημαντικά λιγότερους υπολογιστικούς πόρους και μικρότερα σύνολα δεδομένων, καθιστώντας δυνατή την εκπαίδευση ακόμη και σε υπολογιστικά συστήματα με μία μόνο GPU.

Τα τελευταία χρόνια έχουν προταθεί αρκετές τεχνικές fine-tuning για τα diffusion models, με δημοφιλέστερες τις **DreamBooth** και **Low-Rank Adaptation (LoRA)**. Οι δύο μέθοδοι έχουν κοινό στόχο την προσαρμογή ενός προεκπαιδευμένου μοντέλου, αλλά διαφέρουν σημαντικά ως προς τον τρόπο εκπαίδευσης, τις απαιτήσεις σε υπολογιστικούς πόρους και το είδος των εφαρμογών στις οποίες υπερέχουν.

2.4.1 DreamBooth

Το **DreamBooth** προτάθηκε ως μία μέθοδος εξατομίκευσης προεκπαιδευμένων diffusion models μέσω της εκμάθησης νέων εννοιών (concepts). Βασική επιδίωξή του είναι να επιτρέπει στο μοντέλο να αναπαριστά με ακρίβεια ένα συγκεκριμένο αντικείμενο, πρόσωπο ή καλλιτεχνικό ύφος, διατηρώντας παράλληλα τις γενικές δυνατότητες του αρχικού μοντέλου.

Η βασική ιδέα του DreamBooth είναι η αντιστοίχιση ενός μοναδικού **identifier token** με ένα συγκεκριμένο αντικείμενο ή έννοια. Το token αυτό συνδυάζεται με μία γενική κατηγορία (**class prompt**), όπως *person*, *dog* ή *car*. Κατά τη διαδικασία fine-tuning,

το μοντέλο μαθαίνει να συσχετίζει το μοναδικό token με τα οπτικά χαρακτηριστικά των εικόνων εκπαίδευσης. Έτσι, όταν το token χρησιμοποιηθεί αργότερα σε ένα νέο prompt, το μοντέλο μπορεί να δημιουργήσει εικόνες που διατηρούν την ταυτότητα του αντικειμένου σε διαφορετικά περιβάλλοντα ή συνθέσεις.

Η διαδικασία εκπαίδευσης του DreamBooth βασίζεται σε δύο είδη prompts:

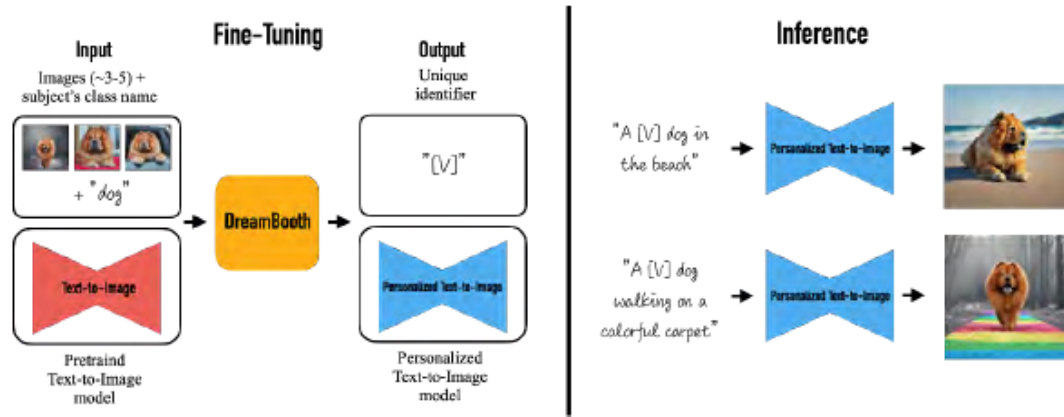
- **Class Prompt**, το οποίο περιγράφει τη γενική κατηγορία του αντικειμένου (π.χ. *photo of a dog* ή *photo of a person*).
- **Instance Prompt**, το οποίο περιλαμβάνει το μοναδικό identifier token που αντιπροσωπεύει το συγκεκριμένο αντικείμενο ή πρόσωπο.

Η εκπαίδευση πραγματοποιείται χρησιμοποιώντας συνήθως **10–20 εικόνες**, αριθμός αρκετός ώστε το μοντέλο να μάθει τα ιδιαίτερα χαρακτηριστικά του αντικειμένου χωρίς να απαιτείται μεγάλο σύνολο δεδομένων. Για τον λόγο αυτό, το DreamBooth χρησιμοποιείται ευρέως σε εφαρμογές **subject-driven generation**, όπου ζητούμενο είναι η πιστή αναπαράσταση ενός συγκεκριμένου αντικειμένου ή προσώπου.

Παρά τα ιδιαίτερα υψηλής ποιότητας αποτελέσματα που προσφέρει, το DreamBooth πραγματοποιεί fine-tuning μεγάλου μέρους των παραμέτρων του μοντέλου. Ως αποτέλεσμα, απαιτεί αυξημένη μνήμη GPU, μεγαλύτερο χρόνο εκπαίδευσης και περισσότερους υπολογιστικούς πόρους σε σχέση με νεότερες τεχνικές προσαρμογής.

Στην παρούσα εργασία το DreamBooth παρουσιάζεται ως μία από τις σημαντικότερες τεχνικές fine-tuning των diffusion models. Ωστόσο, για την εκπαίδευση του μοντέλου **Stable Diffusion v1.5** επιλέχθηκε η τεχνική **Low-Rank Adaptation (LoRA)**, καθώς επιτρέπει την αποτελεσματική προσαρμογή του μοντέλου με σημαντικά μικρότερες

απαιτήσεις σε υπολογιστικούς πόρους και χρόνο εκπαίδευσης, διατηρώντας παράλληλα υψηλή ποιότητα στα παραγόμενα αποτελέσματα.



Σχήμα 2.2. Απλοποιημένη διαδικασία λειτουργίας του DreamBooth [23].

2.4.2 LORA - Low-Rank Adaptation

Η **Low-Rank Adaptation (LoRA)** αποτελεί μία από τις σημαντικότερες και πλέον διαδεδομένες τεχνικές fine-tuning μεγάλων προεκπαιδευμένων μοντέλων (Large Pre-trained Models). Η μέθοδος προτάθηκε από τους Hu et al. (2021) με στόχο τη δραστική μείωση του υπολογιστικού κόστους που απαιτείται για την προσαρμογή ενός μοντέλου σε νέες εργασίες, χωρίς να απαιτείται η ενημέρωση όλων των παραμέτρων του.

Η βασική ιδέα της LoRA είναι ότι, αντί να τροποποιούνται απευθείας τα βάρη του προεκπαιδευμένου μοντέλου, εισάγονται μικροί πίνακες χαμηλής τάξης (**low-rank matrices**) οι οποίοι μαθαίνουν μόνο τις απαραίτητες προσαρμογές για τη νέα εργασία. Αν ο αρχικός πίνακας βαρών συμβολίζεται ως **W**, τότε κατά το συμβατικό

fine-tuning τα νέα βάρη εκφράζονται ως: $W' = W + \Delta W$, όπου το ΔW αντιστοιχεί στη μεταβολή που προκύπτει κατά την εκπαίδευση.

Στη LoRA, η μεταβολή αυτή δεν αποθηκεύεται ως πλήρης πίνακας, αλλά προσεγγίζεται από το γινόμενο δύο πινάκων χαμηλής τάξης:

$$\Delta W = BA$$

όπου οι πίνακες **A** και **B** έχουν πολύ μικρότερη διάσταση από τον αρχικό πίνακα βαρών. Με τον τρόπο αυτό μειώνεται σημαντικά ο αριθμός των εκπαιδευσιμων παραμέτρων, ενώ τα αρχικά βάρη του μοντέλου παραμένουν αμετάβλητα.

Η προσέγγιση αυτή επιτρέπει την προσαρμογή πολύ μεγάλων μοντέλων με ελάχιστες απαιτήσεις σε μνήμη GPU και χρόνο εκπαίδευσης. Σε πολλές περιπτώσεις εκπαιδεύονται μόνο λίγα εκατομμύρια παράμετροι αντί για εκατοντάδες εκατομμύρια ή ακόμη και δισεκατομμύρια, γεγονός που καθιστά τη LoRA ιδιαίτερα αποδοτική για εφαρμογές με περιορισμένους υπολογιστικούς πόρους.

2.4.3 LoRA στο Stable Diffusion

Στα μοντέλα **Stable Diffusion**, η LoRA εφαρμόζεται κυρίως στα επίπεδα **cross-attention** του δικτύου **U-Net**, τα οποία είναι υπεύθυνα για τη σύνδεση των κειμενικών αναπαραστάσεων (text embeddings) με τη διαδικασία παραγωγής της εικόνας. Αντί να τροποποιούνται όλα τα βάρη του U-Net, η LoRA εισάγει πρόσθετους πίνακες χαμηλής τάξης στα επίπεδα αυτά και εκπαιδεύει αποκλειστικά τις νέες παραμέτρους.

Μετά την ολοκλήρωση της εκπαίδευσης, οι πίνακες LoRA μπορούν να συνδυαστούν με το αρχικό μοντέλο χωρίς να απαιτείται επανεκπαίδευση ή αντικατάσταση των

βασικών βαρών. Έτσι, το μοντέλο αποκτά νέα γνώση διατηρώντας παράλληλα τις δυνατότητες του αρχικού προεκπαιδευμένου μοντέλου.

Η προσέγγιση αυτή προσφέρει σημαντικά πλεονεκτήματα:

- σημαντικά μικρότερες απαιτήσεις σε μνήμη GPU,
- ταχύτερη εκπαίδευση,
- μικρότερα αρχεία μοντέλων,
- δυνατότητα εύκολης αποθήκευσης και διαμοιρασμού των LoRA weights,
- δυνατότητα συνδυασμού πολλαπλών LoRA μοντέλων στο ίδιο βασικό μοντέλο.

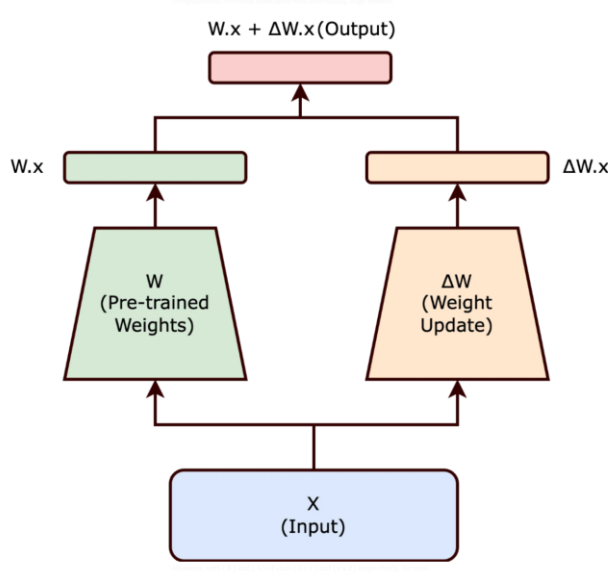
Για τον λόγο αυτό, η LoRA χρησιμοποιείται σήμερα ευρέως για την προσαρμογή μοντέλων Stable Diffusion σε εξειδικευμένα αντικείμενα, καλλιτεχνικά στυλ, χαρακτήρες και επιστημονικές εφαρμογές, όπως η παραγωγή συνθετικών ιατρικών εικόνων.

2.4.4 Σύγκριση DreamBooth και LoRA

Σε αντίθεση με το **DreamBooth**, το οποίο πραγματοποιεί fine-tuning μεγάλου μέρους των παραμέτρων του μοντέλου, η LoRA ενημερώνει μόνο ένα μικρό σύνολο πρόσθετων παραμέτρων. Ως αποτέλεσμα, απαιτεί σημαντικά λιγότερους υπολογιστικούς πόρους, μικρότερη μνήμη GPU και αισθητά μικρότερο χρόνο εκπαίδευσης, διατηρώντας παράλληλα υψηλή ποιότητα στις παραγόμενες εικόνες.

Τα χαρακτηριστικά αυτά καθιστούν τη LoRA ιδιαίτερα κατάλληλη για ερευνητικές εφαρμογές και περιβάλλοντα περιορισμένων υπολογιστικών πόρων, όπως το Google Colab.

Στην παρούσα εργασία, το fine-tuning του μοντέλου **Stable Diffusion v1.5** πραγματοποιήθηκε με την τεχνική **Low-Rank Adaptation (LoRA)**. Για την εκπαίδευση χρησιμοποιήθηκε η βάση **COVID-19 Radiography Database**, με στόχο την προσαρμογή του μοντέλου στην παραγωγή συνθετικών ακτινογραφιών θώρακα υψηλής ποιότητας. Η επιλογή της LoRA βασίστηκε στις χαμηλές απαιτήσεις της σε μνήμη και χρόνο εκπαίδευσης, καθώς και στη δυνατότητα αποτελεσματικής προσαρμογής του μοντέλου με την ενημέρωση μόνο ενός μικρού συνόλου παραμέτρων.



Σχήμα 2.3. Αναπαράσταση της τεχνικής Low-Rank Adaptation (LoRA) [21].

Στον παρακάτω πίνακα (Πίνακας 2.1) φαίνεται η σύγκριση των τεχνικών DreamBooth και LoRA.

Χαρακτηριστικό	DreamBooth	LoRA (Low-Rank Adaptation)
Τύπος προσαρμογής	Fine-tuning μεγάλου μέρους του μοντέλου	Fine-tuning μόνο πρόσθετων παραμέτρων χαμηλής τάξης
Εκπαιδύσιμες παράμετροι	Πολύ μεγάλος αριθμός	Μικρός αριθμός
Απαιτήσεις μνήμης GPU	Υψηλές	Χαμηλές
Χρόνος εκπαίδευσης	Μεγαλύτερος	Μικρότερος
Μέγεθος αποθηκευμένου μοντέλου	Μεγάλο	Πολύ μικρό (μερικά MB)
Υπολογιστικό κόστος	Υψηλό	Χαμηλό
Καταλληλότητα για Google Colab	Περιορισμένη	Πολύ υψηλή
Ποιότητα αποτελεσμάτων	Πολύ υψηλή	Πολύ υψηλή
Τυπικές εφαρμογές	Εξατομίκευση αντικειμένων ή προσώπων (subject-driven generation)	Fine-tuning μοντέλων για νέα σύνολα δεδομένων, καλλιτεχνικά στυλ και εξειδικευμένες εφαρμογές
Χρήση στην παρούσα εργασία	Δεν χρησιμοποιήθηκε	Fine-tuning του Stable Diffusion v1.5 με τη βάση COVID-19 Radiography Database

Πίνακας 2.1. Σύγκριση των τεχνικών DreamBooth και LoRA

Από τη σύγκριση των δύο τεχνικών προκύπτει ότι τόσο το DreamBooth όσο και η LoRA μπορούν να επιτύχουν υψηλής ποιότητας προσαρμογή προεκπαιδευμένων μοντέλων Stable Diffusion. Ωστόσο, η LoRA υπερτερεί ως προς τις απαιτήσεις σε μνήμη, τον χρόνο εκπαίδευσης και το μέγεθος των παραγόμενων μοντέλων, καθώς ενημερώνει μόνο ένα μικρό σύνολο παραμέτρων χωρίς να τροποποιεί τα αρχικά βάρη του μοντέλου. Για τον λόγο αυτό επιλέχθηκε στην παρούσα εργασία για το fine-tuning του **Stable Diffusion v1.5**, καθώς επέτρεψε την αποτελεσματική εκπαίδευση στο περιβάλλον **Google Colab** με χρήση κάρτας γραφικών **NVIDIA Tesla T4 (12 GB VRAM)** και την παραγωγή συνθετικών ακτινογραφιών θώρακα υψηλής ποιότητας.

2.5 Αξιολόγηση Ποιότητας Γεννημένων Εικόνων (π.χ. FID, IS, CLIP score, aesthetic score)

Η αξιολόγηση της ποιότητας των εικόνων που παράγονται από μοντέλα γενεσιουργού τεχνητής νοημοσύνης αποτελεί βασικό στάδιο της ερευνητικής διαδικασίας. Στόχος της αξιολόγησης είναι η εκτίμηση της ποιότητας, του ρεαλισμού και της σημασιολογικής συνέπειας των παραγόμενων εικόνων σε σχέση με το κείμενο ή τα πραγματικά δεδομένα αναφοράς.

Οι μέθοδοι αξιολόγησης διακρίνονται σε δύο βασικές κατηγορίες:

- **Ποσοτικές μέθοδοι (Quantitative Evaluation)**, οι οποίες βασίζονται σε αντικειμενικές αριθμητικές μετρικές.
- **Ποιοτικές μέθοδοι (Qualitative Evaluation)**, οι οποίες βασίζονται στην ανθρώπινη αξιολόγηση των παραγόμενων εικόνων.

Στην παρούσα εργασία χρησιμοποιούνται τόσο ποσοτικές όσο και ποιοτικές μέθοδοι, ώστε η αξιολόγηση των μοντέλων Stable Diffusion να είναι όσο το δυνατόν πιο ολοκληρωμένη.

Ποσοτικές μέθοδοι αξιολόγησης

Οι ποσοτικές μέθοδοι χρησιμοποιούν αριθμητικές μετρικές για την αντικειμενική αξιολόγηση διαφορετικών χαρακτηριστικών των παραγόμενων εικόνων. Οι σημαντικότερες μετρικές που χρησιμοποιούνται στη βιβλιογραφία είναι οι **Inception Score (IS)**, **Fréchet Inception Distance (FID)**, **CLIP Score** και **Aesthetic Score**.

Inception Score (IS)

Το **Inception Score (IS)** προτάθηκε το 2016 ως μία από τις πρώτες μετρικές αξιολόγησης μοντέλων παραγωγής εικόνας. Η μετρική βασίζεται στο προεκπαιδευμένο νευρωνικό δίκτυο **Inception v3**, το οποίο έχει εκπαιδευτεί στο σύνολο δεδομένων **ImageNet**.

Για κάθε παραγόμενη εικόνα, το Inception v3 υπολογίζει μία κατανομή πιθανοτήτων πάνω στις διαθέσιμες κατηγορίες αντικειμένων. Ένα υψηλό Inception Score υποδηλώνει ότι οι εικόνες είναι ταυτόχρονα ευδιάκριτες ως προς το περιεχόμενό τους και παρουσιάζουν μεγάλη ποικιλία μεταξύ τους.

Παρότι το IS αποτέλεσε σημαντικό βήμα στην αξιολόγηση γενετικών μοντέλων, παρουσιάζει περιορισμούς, καθώς δεν συγκρίνει τις παραγόμενες εικόνες με πραγματικές εικόνες αναφοράς και δεν αξιολογεί τη σημασιολογική αντιστοίχιση με το κείμενο.

Fréchet Inception Distance (FID)

Η **Fréchet Inception Distance (FID)** αποτελεί σήμερα μία από τις πιο διαδεδομένες μετρικές αξιολόγησης γενετικών μοντέλων εικόνας. Η μετρική συγκρίνει τις κατανομές χαρακτηριστικών πραγματικών και παραγόμενων εικόνων, χρησιμοποιώντας τα χαρακτηριστικά που εξάγονται από το προεκπαιδευμένο δίκτυο **Inception v3**.

Η σύγκριση πραγματοποιείται μέσω της απόστασης Fréchet μεταξύ δύο πολυδιάστατων κανονικών κατανομών. Μικρότερες τιμές FID υποδηλώνουν ότι οι

παραγόμενες εικόνες προσεγγίζουν περισσότερο την κατανομή των πραγματικών εικόνων και συνεπώς θεωρούνται υψηλότερης ποιότητας.

CLIP Score

Το **CLIP (Contrastive Language–Image Pre-training)** είναι ένα πολυτροπικό μοντέλο της OpenAI, το οποίο εκπαιδεύεται ταυτόχρονα σε εικόνες και στις αντίστοιχες κειμενικές περιγραφές τους. Μέσω της κοινής εκπαίδευσης ενός **Image Encoder** και ενός **Text Encoder**, το μοντέλο μαθαίνει έναν κοινό σημασιολογικό χώρο αναπαράστασης εικόνων και κειμένου.

Το **CLIP Score** αξιοποιεί τις αναπαραστάσεις αυτές για να μετρήσει τον βαθμό σημασιολογικής ευθυγράμμισης μεταξύ μιας παραγόμενης εικόνας και του αντίστοιχου prompt. Η μετρική βασίζεται στην **cosine similarity** μεταξύ των embeddings της εικόνας και του κειμένου. Υψηλότερες τιμές CLIP Score υποδηλώνουν μεγαλύτερη σημασιολογική συνάφεια μεταξύ εικόνας και κειμένου.

Aesthetic Score

Το **Aesthetic Score** αποτελεί μία συμπληρωματική μετρική αξιολόγησης, η οποία εκτιμά την αισθητική ποιότητα των παραγόμενων εικόνων. Σε αντίθεση με το FID και το CLIP Score, δεν αξιολογεί ούτε την ομοιότητα με πραγματικές εικόνες ούτε τη σημασιολογική αντιστοίχιση με το κείμενο, αλλά προβλέπει πόσο αισθητικά ευχάριστη θεωρείται μία εικόνα.

Η αξιολόγηση πραγματοποιείται μέσω ενός προεκπαιδευμένου μοντέλου που βασίζεται στα χαρακτηριστικά του **CLIP ViT-L/14** και έχει εκπαιδευτεί πάνω σε

ανθρώπινες αξιολογήσεις αισθητικής. Η τελική βαθμολογία κυμαίνεται από **1 έως 10**, όπου μεγαλύτερες τιμές αντιστοιχούν σε υψηλότερη αισθητική ποιότητα.

Ποιοτικές μέθοδοι αξιολόγησης

Οι ποιοτικές μέθοδοι βασίζονται στην ανθρώπινη αξιολόγηση των παραγόμενων εικόνων. Η αξιολόγηση πραγματοποιείται μέσω οπτικής επιθεώρησης (visual inspection), σύγκρισης ζευγών εικόνων (pairwise comparison) ή κατάταξης προτίμησης (preference ranking), εξετάζοντας χαρακτηριστικά όπως ο ρεαλισμός, η αισθητική ποιότητα, η συνέπεια με το prompt και η παρουσία ανεπιθύμητων τεχνουργημάτων.

Παρότι οι ποιοτικές μέθοδοι είναι περισσότερο υποκειμενικές και απαιτούν σημαντικό χρόνο αξιολόγησης, αποτελούν απαραίτητο συμπλήρωμα των ποσοτικών μετρικών, καθώς μπορούν να αναδείξουν χαρακτηριστικά που δεν αποτυπώνονται πλήρως μέσω αριθμητικών δεικτών.

Στην παρούσα εργασία η τελική αξιολόγηση των μοντέλων πραγματοποιήθηκε συνδυάζοντας τις μετρικές **FID**, **CLIP Score** και **Aesthetic Score** με ποιοτική αξιολόγηση των παραγόμενων εικόνων, ώστε να ληφθούν υπόψη τόσο η αντικειμενική όσο και η αντιληπτική ποιότητα των αποτελεσμάτων.

Πλατφόρμα Hugging Face

3.1 Εισαγωγή και Αρχιτεκτονική της Πλατφόρμας

Η **Hugging Face** αποτελεί σήμερα μια από τις πιο δημοφιλείς και ολοκληρωμένες πλατφόρμες για ανάπτυξη και εκπαίδευση μοντέλων τεχνητής νοημοσύνης, καλύπτοντας εφαρμογές φυσικής γλώσσας, όρασης, ήχου και multimodal δεδομένων. Η πλατφόρμα παρέχει ένα πλήρες οικοσύστημα εργαλείων, βιβλιοθηκών και προεκπαιδευμένων μοντέλων, επιτρέποντας την εύκολη πρόσβαση σε state-of-the-art μοντέλα και την γρήγορη ανάπτυξη πειραμάτων με υψηλή αναπαραγωγιμότητα.

Η αρχιτεκτονική της πλατφόρμας Hugging Face βασίζεται σε τρία κύρια στοιχεία:

1. **Transformers:** Βιβλιοθήκη που παρέχει προεκπαιδευμένα μοντέλα για φυσική γλώσσα και multimodal εφαρμογές. Στα πλαίσια της δημιουργίας εικόνων από κείμενο, τα Transformers χρησιμοποιούνται για την παραγωγή textual embeddings και την καθοδήγηση των diffusion models, εξασφαλίζοντας σημασιολογική ευθυγράμμιση μεταξύ κειμένου και εικόνας (Wolf et al., 2020).
2. **Diffusers:** Βιβλιοθήκη για diffusion-based μοντέλα όπως το **Stable Diffusion**, προσφέροντας pipelines για **text-to-image**, **image-to-image** και fine-tuning με τεχνικές όπως **DreamBooth** και **LoRA**. Τα Diffusers επιτρέπουν εύκολη διαχείριση μοντέλων, παραγωγή εικόνων υψηλής πιστότητας και αξιολόγηση πειραμάτων, ενώ υποστηρίζουν και custom datasets (Hugging Face, 2025).

3. **Accelerate:** Βιβλιοθήκη που αυτοματοποιεί την εκπαίδευση και inference σε διαφορετικές υπολογιστικές διαμορφώσεις, υποστηρίζοντας multi-GPU setups, TPUs και μεικτή ακρίβεια (mixed precision). Αυτό μειώνει τον χρόνο εκπαίδευσης και βελτιώνει την αποδοτικότητα των πειραμάτων, ενώ διευκολύνει την αναπαραγωγικότητα (Hugging Face, 2025).

Η συνδυαστική χρήση των εργαλείων αυτών καθιστά την πλατφόρμα Hugging Face μια ισχυρή και επεκτάσιμη λύση για πειραματική αξιολόγηση και ανάπτυξη μοντέλων τεχνητής νοημοσύνης, ιδιαίτερα στον τομέα της παραγωγής εικόνων από κείμενο και εικόνα.

3.2 Model Hub και Προεκπαιδευμένα Μοντέλα

Η πλατφόρμα **Hugging Face Model Hub** αποτελεί ένα από τα μεγαλύτερα αποθετήρια προεκπαιδευμένων μοντέλων μηχανικής μάθησης και τεχνητής νοημοσύνης, παρέχοντας πρόσβαση σε χιλιάδες μοντέλα για επεξεργασία φυσικής γλώσσας (Natural Language Processing), υπολογιστική όραση (Computer Vision), παραγωγή εικόνων και πολυτροπικές (multimodal) εφαρμογές (Wolf et al., 2020; Hugging Face, 2025). Εκτός από τα ίδια τα μοντέλα, η πλατφόρμα προσφέρει έτοιμα σύνολα δεδομένων (Datasets), βιβλιοθήκες λογισμικού και παραδείγματα κώδικα, διευκολύνοντας σημαντικά την ανάπτυξη εφαρμογών τεχνητής νοημοσύνης.

Η χρήση προεκπαιδευμένων μοντέλων, όπως τα **GPT**, **CLIP** και **Stable Diffusion**, επιτρέπει την αξιοποίηση γνώσης που έχει αποκτηθεί από εκπαίδευση σε πολύ μεγάλα σύνολα δεδομένων, μειώνοντας σημαντικά τον χρόνο και τους υπολογιστικούς πόρους που απαιτούνται για την ανάπτυξη νέων εφαρμογών. Αντί της εκπαίδευσης ενός μοντέλου από την αρχή (training from scratch), οι ερευνητές μπορούν να χρησιμοποιήσουν υπάρχοντα μοντέλα και να τα προσαρμόσουν στις απαιτήσεις μιας συγκεκριμένης εφαρμογής μέσω τεχνικών fine-tuning, όπως οι **DreamBooth** και **Low-Rank Adaptation (LoRA)** (Rombach et al., 2022; Radford et al., 2021).

Ένα ακόμη σημαντικό χαρακτηριστικό του Hugging Face Model Hub είναι η υποστήριξη της **αναπαραγωγιμότητας (reproducibility)** και της **συνεργατικής έρευνας (collaborative research)**. Κάθε μοντέλο συνοδεύεται από τεχνική τεκμηρίωση (Model Card), πληροφορίες σχετικά με τα δεδομένα εκπαίδευσης, τις προτεινόμενες χρήσεις, τους περιορισμούς και την άδεια χρήσης, επιτρέποντας στους

ερευνητές να αξιολογούν την καταλληλότητα του μοντέλου για την εκάστοτε εφαρμογή. Παράλληλα, η εύκολη διαμοίραση μοντέλων, συνόλων δεδομένων και πειραμάτων διευκολύνει την επαλήθευση των αποτελεσμάτων και την περαιτέρω εξέλιξη της έρευνας (Hugging Face, 2025).

Η επιλογή της πλατφόρμας Hugging Face βασίστηκε στην ευρεία αποδοχή της από την ερευνητική κοινότητα, στην υποστήριξη σύγχρονων diffusion models και στη δυνατότητα εύκολης αναπαραγωγής των πειραμάτων, στοιχεία που συνέβαλαν στην αξιοπιστία και την επαναληψιμότητα της πειραματικής διαδικασίας.

3.3 Hugging Face Datasets και Diffusers Library

Το οικοσύστημα της **Hugging Face** παρέχει ένα ολοκληρωμένο σύνολο εργαλείων για την ανάπτυξη, εκπαίδευση και αξιολόγηση μοντέλων τεχνητής νοημοσύνης. Μεταξύ των σημαντικότερων βιβλιοθηκών του περιλαμβάνονται οι **Diffusers** και **Datasets**, οι οποίες χρησιμοποιούνται ευρέως σε εφαρμογές γενεσιουργού τεχνητής νοημοσύνης και ειδικότερα στην παραγωγή εικόνων από κείμενο (text-to-image generation).

Η βιβλιοθήκη **Diffusers** της Hugging Face [27] αποτελεί μία από τις πλέον διαδεδομένες βιβλιοθήκες ανοιχτού κώδικα για την υλοποίηση και αξιοποίηση μοντέλων διάχυσης (diffusion models). Παρέχει έτοιμα pipelines για παραγωγή εικόνων από κείμενο (text-to-image), μετασχηματισμό εικόνας σε εικόνα (image-to-image), inpainting και άλλες εφαρμογές, ενώ υποστηρίζει σύγχρονες αρχιτεκτονικές όπως τα **Stable Diffusion**, **ControlNet** και άλλα diffusion models.

Ένα από τα σημαντικότερα πλεονεκτήματα της βιβλιοθήκης είναι η διάθεση επίσημων σεναρίων εκπαίδευσης (training scripts), τα οποία συντηρούνται από την κοινότητα της Hugging Face και ενημερώνονται διαρκώς σύμφωνα με τις τελευταίες εξελίξεις της βιβλιοθήκης. Τα σενάρια αυτά είναι αυτοτελή, εύκολα προσαρμόσιμα και συνοδεύονται από εκτενή τεκμηρίωση, επιτρέποντας τόσο σε αρχάριους όσο και σε έμπειρους ερευνητές να υλοποιήσουν διαδικασίες fine-tuning χωρίς να απαιτείται ανάπτυξη κώδικα από την αρχή.

Παράλληλα, η βιβλιοθήκη **Datasets** της Hugging Face αποτελεί ένα ευέλικτο εργαλείο για τη διαχείριση και την προεπεξεργασία συνόλων δεδομένων μηχανικής μάθησης. Υποστηρίζει δεδομένα κειμένου, εικόνας, ήχου και πολυτροπικών

εφαρμογών, παρέχοντας ενιαίο τρόπο πρόσβασης τόσο σε δημόσια όσο και σε τοπικά σύνολα δεδομένων.

Η βιβλιοθήκη προσφέρει δυνατότητες όπως:

- φόρτωση συνόλων δεδομένων με μία μόνο γραμμή κώδικα,
- αποδοτική διαχείριση πολύ μεγάλων συνόλων δεδομένων μέσω streaming,
- μετασχηματισμό, φιλτράρισμα και προεπεξεργασία των δεδομένων,
- αυτόματη ενσωμάτωση με βιβλιοθήκες όπως οι PyTorch, TensorFlow και JAX,
- συμβατότητα με τα pipelines εκπαίδευσης της βιβλιοθήκης Diffusers.

Στον χώρο της δημιουργίας εικόνων από κείμενο, η βιβλιοθήκη Datasets διευκολύνει τη δημιουργία και διαχείριση προσαρμοσμένων συνόλων δεδομένων που αποτελούνται από εικόνες και τις αντίστοιχες περιγραφές τους (captions), επιτρέποντας την εύκολη επαναχρησιμοποίηση των δεδομένων και την αναπαραγωγή των πειραμάτων.

3.4 Εργαλεία για Εικόνες: Diffusers, Transformers, Accelerate

Η ανάπτυξη και η εκπαίδευση σύγχρονων μοντέλων δημιουργίας εικόνων βασίζονται σε εξειδικευμένες βιβλιοθήκες λογισμικού, οι οποίες υποστηρίζουν την υλοποίηση, την εκπαίδευση, τη βελτιστοποίηση και την αξιολόγηση μοντέλων βαθιάς μάθησης. Στο οικοσύστημα της Hugging Face, οι βιβλιοθήκες **Diffusers** [27], **Transformers** [25] και **Accelerate** [28] αποτελούν τον βασικό πυρήνα για την ανάπτυξη εφαρμογών παραγωγής εικόνων από κείμενο (text-to-image generation), προσφέροντας ένα ενιαίο και πλήρως αναπαραγωγίμο περιβάλλον ανάπτυξης.

Η βιβλιοθήκη **Diffusers** παρέχει ολοκληρωμένη υποστήριξη για μοντέλα διάχυσης (diffusion models), όπως τα **Stable Diffusion**, καθώς και έτοιμα pipelines για εργασίες παραγωγής εικόνων από κείμενο (text-to-image), μετασχηματισμού εικόνας σε εικόνα (image-to-image), inpainting και άλλες εφαρμογές γενεσιουργού τεχνητής νοημοσύνης. Παράλληλα, περιλαμβάνει επίσημα σενάρια εκπαίδευσης (training scripts) και υποστηρίζει σύγχρονες τεχνικές fine-tuning, όπως **DreamBooth** και **Low-Rank Adaptation (LoRA)**, επιτρέποντας την αποδοτική προσαρμογή προεκπαιδευμένων μοντέλων σε εξειδικευμένα σύνολα δεδομένων με μειωμένες απαιτήσεις σε υπολογιστικούς πόρους.

Η βιβλιοθήκη **Transformers** αποτελεί το βασικό εργαλείο για την επεξεργασία της φυσικής γλώσσας και τη δημιουργία των σημασιολογικών αναπαραστάσεων (embeddings) που χρησιμοποιούνται από τα diffusion models. Στα μοντέλα δημιουργίας εικόνων αξιοποιούνται προεκπαιδευμένοι **CLIP text encoders**, οι οποίοι μετατρέπουν τα κείμενα (prompts) σε διανυσματικές αναπαραστάσεις υψηλής διάστασης. Οι αναπαραστάσεις αυτές χρησιμοποιούνται από το μοντέλο διάχυσης για

την καθοδήγηση της διαδικασίας παραγωγής εικόνων, εξασφαλίζοντας τη σημασιολογική αντιστοίχιση μεταξύ της φυσικής γλώσσας και του παραγόμενου οπτικού περιεχομένου.

Η βιβλιοθήκη **Accelerate** συμπληρώνει το οικοσύστημα της Hugging Face, παρέχοντας εργαλεία για την αποδοτική εκπαίδευση μεγάλων μοντέλων σε διαφορετικές υπολογιστικές υποδομές. Υποστηρίζει αυτόματη διαχείριση των διαθέσιμων υπολογιστικών συσκευών (CPU και GPU), εκπαίδευση σε πολλαπλές GPU, τεχνικές μικτής ακρίβειας (mixed precision), gradient accumulation και άλλες μεθόδους βελτιστοποίησης της χρήσης μνήμης. Με τον τρόπο αυτό μειώνεται σημαντικά ο χρόνος εκπαίδευσης και διευκολύνεται η αναπαραγωγή των πειραμάτων ακόμη και σε περιβάλλοντα με περιορισμένους υπολογιστικούς πόρους.

Προεκπαιδευμένα Μοντέλα Γενεσιουργού ΤΝ για Εικόνες

4.1 Stable Diffusion

Το **Stable Diffusion** αποτελεί ένα από τα πλέον διαδεδομένα μοντέλα γενεσιουργού τεχνητής νοημοσύνης (Generative Artificial Intelligence) για τη δημιουργία εικόνων από κείμενο (text-to-image generation). Αναπτύχθηκε το 2022 από τη **Stability AI** σε συνεργασία με τις **CompVis**, **Runway ML** και **LAION** και βασίζεται στην αρχιτεκτονική των **Latent Diffusion Models (LDMs)** (Rombach et al., 2022).

Σε αντίθεση με τα κλασικά diffusion models, τα οποία πραγματοποιούν τη διαδικασία διάχυσης απευθείας στον χώρο των εικονοστοιχείων (pixel space), το Stable Diffusion εκτελεί τη διαδικασία αυτή σε έναν **λανθάνοντα χώρο (latent space)** μειωμένων διαστάσεων. Η προσέγγιση αυτή μειώνει σημαντικά τις απαιτήσεις σε υπολογιστική ισχύ και μνήμη, διατηρώντας παράλληλα υψηλή ποιότητα στις παραγόμενες εικόνες. Ως αποτέλεσμα, το μοντέλο μπορεί να εκτελεστεί ακόμη και σε συστήματα με μία μόνο GPU, γεγονός που συνέβαλε καθοριστικά στη διάδοσή του τόσο στην ερευνητική κοινότητα όσο και στη βιομηχανία.

Ένα ακόμη σημαντικό χαρακτηριστικό του Stable Diffusion είναι η ανοικτή διάθεση του προεκπαιδευμένου μοντέλου και του πηγαίου κώδικα, γεγονός που διευκόλυνε την ανάπτυξη πλήθους εφαρμογών, εργαλείων και τεχνικών fine-tuning από την κοινότητα ανοικτού λογισμικού.

4.1.1 Αρχιτεκτονική Stable Diffusion

Η αρχιτεκτονική του Stable Diffusion αποτελείται από τέσσερα βασικά υποσυστήματα, τα οποία συνεργάζονται για την παραγωγή εικόνων από κείμενο (Σχήμα 1).

α) Variational Autoencoder (VAE)

Το **Variational Autoencoder (VAE)** χρησιμοποιείται για τη μετατροπή των εικόνων από τον χώρο των εικονοστοιχείων σε έναν λανθάνοντα χώρο μικρότερων διαστάσεων. Για παράδειγμα, μία εικόνα ανάλυσης 512×512 συμπιέζεται σε μία λανθάνουσα αναπαράσταση περίπου $64 \times 64 \times 4$. Μετά την ολοκλήρωση της διαδικασίας δημιουργίας, το VAE αποκωδικοποιεί τη λανθάνουσα αναπαράσταση και παράγει την τελική εικόνα υψηλής ανάλυσης.

β) Διαδικασία Διάχυσης (Diffusion Process)

Η διαδικασία διάχυσης αποτελείται από δύο διαδοχικά στάδια. Στο στάδιο της **forward diffusion** προστίθεται σταδιακά Γκαουσιανός θόρυβος στη λανθάνουσα αναπαράσταση μέχρι αυτή να μετατραπεί σε τυχαίο θόρυβο. Στη συνέχεια, κατά τη **reverse diffusion**, το μοντέλο μαθαίνει να αφαιρεί προοδευτικά τον θόρυβο, ανακατασκευάζοντας μία νέα εικόνα που καθοδηγείται από το δοθέν κείμενο (prompt).

γ) U-Net

Ο βασικός πυρήνας του Stable Diffusion είναι ένα συνελικτικό νευρωνικό δίκτυο τύπου **U-Net**, το οποίο προβλέπει τον θόρυβο που πρέπει να αφαιρεθεί σε κάθε βήμα

της αντίστροφης διάχυσης. Μέσω επαναλαμβανόμενων βημάτων αποθορυβοποίησης, το U-Net δημιουργεί σταδιακά τη λανθάνουσα αναπαράσταση της τελικής εικόνας.

δ) Text Conditioning και Cross-Attention

Η παραγωγή εικόνων καθοδηγείται από κειμενικές προτροπές (**prompts**). Αρχικά, το κείμενο επεξεργάζεται από έναν προεκπαιδευμένο **CLIP Text Encoder**, ο οποίος το μετατρέπει σε σημασιολογικές αναπαραστάσεις (**text embeddings**). Οι αναπαραστάσεις αυτές ενσωματώνονται στο U-Net μέσω του μηχανισμού **cross-attention**, επιτρέποντας στο μοντέλο να συσχετίζει τα χαρακτηριστικά του κειμένου με τα οπτικά χαρακτηριστικά της εικόνας κατά τη διαδικασία αποθορυβοποίησης.

Η λειτουργία του μηχανισμού **Scaled Dot-Product Attention** περιγράφεται από τη σχέση:

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

όπου:

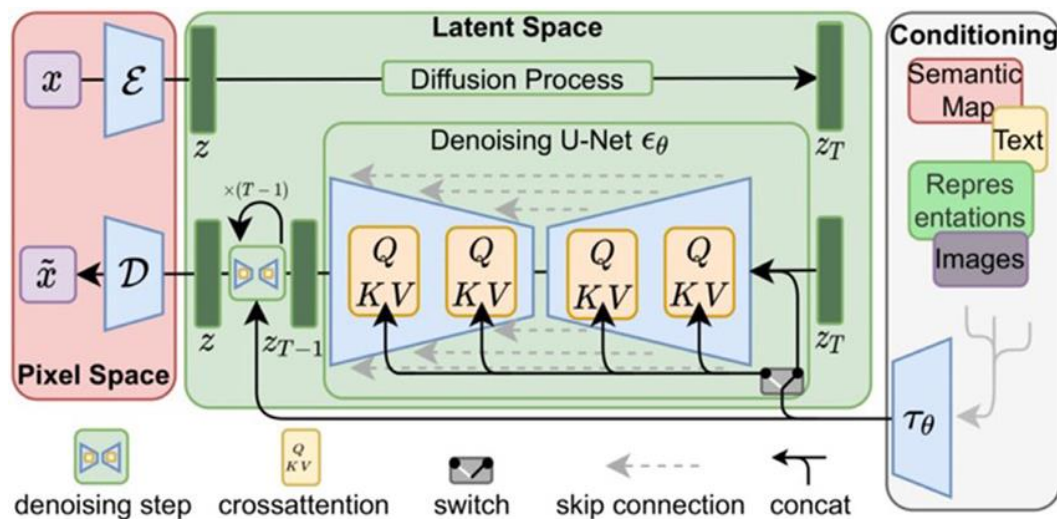
- **Q (Query)**: αναπαράσταση του ερωτήματος (query),
- **K (Key)**: αναπαράσταση των χαρακτηριστικών που χρησιμοποιούνται για τον υπολογισμό της συνάφειας,
- **V (Value)**: οι πληροφορίες που μεταφέρονται στην έξοδο,
- **d**: η διάσταση των διανυσμάτων (embedding dimension).

Αρχικά υπολογίζεται το εσωτερικό γινόμενο μεταξύ των διανυσμάτων **Q** και **K**, το οποίο εκφράζει τον βαθμό συσχέτισης μεταξύ τους. Η τιμή αυτή διαιρείται με τον παράγοντα $\sqrt{d}$, ώστε να αποφεύγονται υπερβολικά μεγάλες τιμές κατά την

εκπαίδευση και να βελτιώνεται η αριθμητική σταθερότητα. Στη συνέχεια εφαρμόζεται η συνάρτηση **softmax**, η οποία μετατρέπει τις τιμές σε βάρη προσοχής (attention weights). Τέλος, τα βάρη αυτά χρησιμοποιούνται για τον υπολογισμό ενός σταθμισμένου αθροίσματος των διανυσμάτων $\mathbf{V}$, το οποίο αποτελεί την έξοδο του μηχανισμού attention.

Στο Stable Diffusion η παραπάνω λειτουργία υλοποιείται μέσω του **cross-attention**, επιτρέποντας στο U-Net να ενσωματώνει τις σημασιολογικές πληροφορίες των text embeddings σε κάθε βήμα της διαδικασίας reverse diffusion. Με τον τρόπο αυτό, το παραγόμενο οπτικό περιεχόμενο ευθυγραμμίζεται με τη σημασιολογία του prompt, βελτιώνοντας τη συνάφεια μεταξύ κειμένου και εικόνας.

Η συνεργασία των παραπάνω υποσυστημάτων επιτρέπει στο Stable Diffusion να παράγει εικόνες υψηλής ποιότητας με σχετικά μικρές απαιτήσεις σε υπολογιστικούς πόρους, γεγονός που το καθιστά ιδιαίτερα κατάλληλο τόσο για ερευνητικές εφαρμογές όσο και για πρακτικές εφαρμογές δημιουργίας περιεχομένου.



Σχήμα 1: Αρχιτεκτονική του Stable Diffusion βασισμένη στα Latent Diffusion Models (Rombach et al., 2022)

Ένα από τα σημαντικότερα πλεονεκτήματα του Stable Diffusion είναι η δυνατότητα **fine-tuning**, δηλαδή η προσαρμογή ενός προεκπαιδευμένου μοντέλου σε εξειδικευμένα σύνολα δεδομένων. Η διαδικασία αυτή επιτρέπει στο μοντέλο να μάθει νέα οπτικά χαρακτηριστικά και να παράγει εικόνες προσαρμοσμένες σε συγκεκριμένους τομείς εφαρμογής, χωρίς να απαιτείται εκπαίδευση από την αρχή.

4.2 DALL·E

Το **DALL-E** [12] αναπτύχθηκε από την OpenAI και κυκλοφόρησε για πρώτη φορά τον Ιανουάριο του 2021, δημιουργώντας ψηφιακές εικόνες από περιγραφές φυσικής γλώσσας, που ονομάζονται «prompts». Πρόκειται για μια έκδοση του GPT-3 [13] με 12 δισεκατομμύρια παραμέτρους, εκπαιδευμένη σε 250 εκατομμύρια ζεύγη εικόνας-κειμένου που συλλέχθηκαν από το διαδίκτυο, προκειμένου να παράγει εικόνες από περιγραφές κειμένου, χρησιμοποιώντας ένα σύνολο δεδομένων ζευγών κειμένου-εικόνας.

Διαθέτει ένα ευρύ φάσμα δυνατοτήτων, όπως η δημιουργία ανθρωπομορφισμένων εκδόσεων ζώων και αντικειμένων, ο συνδυασμός ασύνδετων εννοιών με εύλογους τρόπους, η απόδοση κειμένου, καθώς και η εφαρμογή μετασχηματισμών σε υπάρχουσες εικόνες.

Το **DALL-E** χρησιμοποιεί ένα νευρωνικό δίκτυο τύπου Transformer για να επιτρέψει στο μοντέλο να δημιουργεί και να κατανοεί συνδέσεις μεταξύ διαφορετικών εννοιών. Πρόκειται για ένα γλωσσικό μοντέλο τύπου Transformer. Λαμβάνει τόσο το κείμενο όσο και την εικόνα ως μία ενιαία ροή δεδομένων που

περιέχει έως 1280 tokens (256 για το κείμενο και 1024 για την εικόνα) και εκπαιδεύεται χρησιμοποιώντας τη μέθοδο της μέγιστης πιθανοφάνειας (maximum likelihood) ώστε να παράγει όλα τα tokens, το ένα μετά το άλλο.

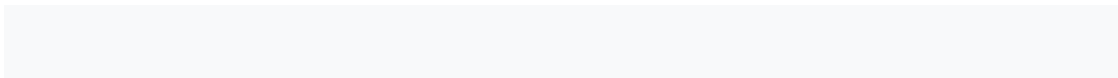
Αυτή η διαδικασία εκπαίδευσης επιτρέπει στο DALL-E όχι μόνο να δημιουργεί μια εικόνα από το μηδέν, αλλά και να αναδημιουργεί οποιαδήποτε ορθογώνια περιοχή μιας υπάρχουσας εικόνας που εκτείνεται έως την κάτω-δεξιά γωνία, με τρόπο συμβατό με την κειμενική εντολή (text prompt).

Το DALL-E μπορεί να δημιουργεί εικόνες σε πολλαπλά στυλ, συμπεριλαμβανομένων ρεαλιστικών φωτογραφιών, ζωγραφικών έργων και emojis. Μπορεί να «χειρίζεται και να αναδιατάσσει» αντικείμενα στις εικόνες του και να τοποθετεί σωστά στοιχεία σχεδίασης σε νέες συνθέσεις χωρίς ρητή οδηγία.

Το **DALL·E mini** αναπτύχθηκε από τον **Boris Dayma** και την κοινότητα του **Hugging Face** [20]. Το έργο δημιουργήθηκε ως μια ελαφριά, προσβάσιμη αναπαραγωγή της βασικής ιδέας του DALL·E της OpenAI, με στόχο να επιτρέψει την ευρεία πειραματική χρήση μοντέλων μετατροπής κειμένου σε εικόνα χωρίς τις υψηλές υπολογιστικές απαιτήσεις των εμπορικών συστημάτων.

Το DALL·E mini παρότι επέτρεψε την ευρεία χρήση και συνέβαλε στη διάδοση των μοντέλων μετατροπής κειμένου σε εικόνα, παρουσιάζει σημαντικούς περιορισμούς ως προς την ποιότητα της παραγόμενης εικόνας, την ευκρίνεια των οπτικών χαρακτηριστικών και τη σημασιολογική ευθυγράμμιση με το κείμενο εισόδου. Συγκεκριμένα, το DALL·E mini συχνά παράγει εικόνες χαμηλής ανάλυσης, με οπτικά τεχνουργήματα, ασυνέπειες στη δομή των αντικειμένων και περιορισμένη ικανότητα απόδοσης λεπτομερειών, ιδιαίτερα σε πολύπλοκες σκηνές ή στην

παραγωγή ρεαλιστικών ανθρώπινων προσώπων. Σε σύγκριση με πιο σύγχρονες αρχιτεκτονικές μοντέλων διάχυσης, όπως το DALL·E 3 ή το Stable Diffusion, το DALL·E mini υστερεί σημαντικά σε επίπεδο οπτικού ρεαλισμού και ακρίβειας, γεγονός που το καθιστά λιγότερο κατάλληλο για εφαρμογές που απαιτούν υψηλή πιστότητα και ρεαλιστική αναπαράσταση.



4.3 Άλλα μοντέλα

Τα τελευταία χρόνια έχουν παρουσιαστεί νέα μοντέλα δημιουργίας εικόνων που βελτιώνουν σημαντικά την ποιότητα των παραγόμενων αποτελεσμάτων, την κατανόηση των λεκτικών περιγραφών και την αποδοτικότητα της εκπαίδευσης. Αν και η παρούσα εργασία επικεντρώνεται στα Stable Diffusion 1.5 και 2.1, αξίζει να αναφερθούν συνοπτικά ορισμένες από τις σημαντικότερες πρόσφατες εξελίξεις.

Το **Stable Diffusion XL (SDXL)** [38] αποτελεί σημαντική εξέλιξη της οικογένειας Stable Diffusion και έχει σχεδιαστεί για την παραγωγή εικόνων υψηλότερης ποιότητας και ανάλυσης. Σε σύγκριση με τα Stable Diffusion 1.5 και 2.1, προσφέρει βελτιωμένη απόδοση στην απόδοση λεπτομερειών, στη σύνθεση σύνθετων σκηνών και στη φωτορεαλιστική αναπαράσταση αντικειμένων. Παράλληλα, διατηρεί το πλεονέκτημα του ανοικτού κώδικα και της δυνατότητας εφαρμογής τεχνικών fine-tuning, όπως οι LoRA και DreamBooth. Ωστόσο, οι αυξημένες απαιτήσεις σε μνήμη GPU και υπολογιστικούς πόρους αποτελούν σημαντικό περιορισμό για την εκπαίδευση και αξιοποίησή του.

Το **DALL·E 3** [15] αποτελεί ένα από τα πλέον προηγμένα εμπορικά μοντέλα δημιουργίας εικόνων. Βασικό του πλεονέκτημα είναι η ιδιαίτερα αποτελεσματική κατανόηση σύνθετων prompts και η δυνατότητα παραγωγής εικόνων που ανταποκρίνονται με μεγάλη ακρίβεια στις λεκτικές περιγραφές του χρήστη. Επιπλέον, παρουσιάζει βελτιωμένη ικανότητα ενσωμάτωσης κειμένου μέσα στις εικόνες. Παρόλα αυτά, πρόκειται για κλειστό σύστημα, χωρίς δυνατότητα τοπικής εκπαίδευσης ή προσαρμογής σε εξειδικευμένα σύνολα δεδομένων.

Το **Imagen** [33] της Google ανήκει επίσης στην κατηγορία των σύγχρονων μοντέλων diffusion και είναι γνωστό για την υψηλή φωτορεαλιστικότητα και την εξαιρετική ποιότητα των παραγόμενων εικόνων. Χρησιμοποιεί προηγμένες τεχνικές κατανόησης φυσικής γλώσσας και επιτυγχάνει υψηλή αντιστοίχιση μεταξύ prompt και παραγόμενου αποτελέσματος. Ωστόσο, όπως και το DALL·E 3, παραμένει κλειστό μοντέλο με περιορισμένη πρόσβαση για την ερευνητική κοινότητα.

Stable Diffusion 3 (SD3) [11]. Το Stable Diffusion 3 αποτελεί εξέλιξη της οικογένειας Stable Diffusion και εισάγει αρχιτεκτονική βασισμένη σε Diffusion Transformers (DiT). Σε σύγκριση με τις προηγούμενες εκδόσεις παρουσιάζει σημαντικά βελτιωμένη κατανόηση σύνθετων prompts, καλύτερη απόδοση στη δημιουργία κειμένου μέσα στις εικόνες και υψηλότερη συνολική ποιότητα. Παράλληλα διατηρεί τη φιλοσοφία του ανοικτού οικοσυστήματος που επιτρέπει την προσαρμογή και το fine-tuning.

Flux [42]. Το Flux είναι ένα σύγχρονο ανοικτό μοντέλο δημιουργίας εικόνων που αναπτύχθηκε από την Black Forest Labs. Βασίζεται επίσης σε αρχιτεκτονικές τύπου Transformer και έχει προσελκύσει σημαντικό ενδιαφέρον λόγω της υψηλής πιστότητας στην ερμηνεία των prompts και της ανταγωνιστικής ποιότητας εικόνας. Ωστόσο, οι απαιτήσεις του σε υπολογιστικούς πόρους είναι σημαντικά υψηλότερες σε σχέση με τα Stable Diffusion 1.5 και 2.1.

Imagen 3 [39]. Το Imagen 3 αποτελεί τη νεότερη γενιά του μοντέλου Imagen της Google. Διακρίνεται για την εξαιρετική φωτορεαλιστικότητα, την υψηλή αισθητική ποιότητα και την ιδιαίτερα καλή κατανόηση φυσικής γλώσσας. Παρόλα αυτά, παραμένει κλειστό εμπορικό σύστημα και δεν παρέχει τη δυνατότητα τοπικής εκπαίδευσης ή προσαρμογής από ερευνητές και προγραμματιστές.

Midjourney [41]. Το Midjourney είναι ένα από τα δημοφιλέστερα εμπορικά συστήματα δημιουργίας εικόνων. Είναι γνωστό για την παραγωγή εικόνων υψηλής αισθητικής και καλλιτεχνικής ποιότητας, ιδιαίτερα σε δημιουργικές εφαρμογές. Ωστόσο, η κλειστή φύση του μοντέλου περιορίζει τη δυνατότητα ελέγχου, παραμετροποίησης και επιστημονικής αναπαραγωγής των αποτελεσμάτων.

AuraFlow [40]. Το AuraFlow αποτελεί μια νεότερη ανοικτού κώδικα προσέγγιση που βασίζεται σε αρχιτεκτονικές diffusion και transformers. Στόχος του είναι η παροχή ανταγωνιστικών επιδόσεων σε σχέση με εμπορικά μοντέλα, διατηρώντας παράλληλα τη διαφάνεια και την προσβασιμότητα του ανοικτού λογισμικού. Αν και βρίσκεται ακόμη σε σχετικά πρώιμο στάδιο ανάπτυξης, παρουσιάζει ιδιαίτερο ερευνητικό ενδιαφέρον ως εναλλακτική λύση ανοικτού κώδικα.

Η εμφάνιση των παραπάνω μοντέλων δείχνει ότι η έρευνα στον τομέα της δημιουργίας εικόνων εξελίσσεται με ταχύ ρυθμό. Παρότι τα Stable Diffusion 1.5 και 2.1 εξακολουθούν να χρησιμοποιούνται ευρέως λόγω της ευκολίας προσαρμογής και των χαμηλότερων απαιτήσεων σε υπολογιστικούς πόρους, τα νεότερα μοντέλα προσφέρουν βελτιωμένη ποιότητα εικόνας και αποτελούν σημαντικές κατευθύνσεις για μελλοντική έρευνα.

.

4.4 Συγκριτική παρουσίαση χαρακτηριστικών

Στον παρακάτω πίνακα φαίνεται η Παρουσίαση Σύγχρονων Μοντέλων Δημιουργίας Εικόνων.

Μοντέλο	Κατηγορία	Ανοιχτό / Κλειστό	Fine-Tuning	Πλεονεκτήματα	Μειονεκτήματα
Stable Diffusion 1.5	Latent Diffusion	Ανοιχτό	Ναι (LoRA, DreamBooth)	Χαμηλές απαιτήσεις GPU, μεγάλη κοινότητα χρηστών	Χαμηλότερη ποιότητα από νεότερα μοντέλα
Stable Diffusion 2.1	Latent Diffusion	Ανοιχτό	Ναι (LoRA, DreamBooth, Textual Inversion)	Βελτιωμένη ποιότητα εικόνας σε σχέση με το SD 1.5, καλύτερη κατανόηση prompts, υποστήριξη υψηλότερης ανάλυσης (768×768), ανοιχτός κώδικας	Μικρότερη κοινότητα και λιγότερα διαθέσιμα μοντέλα/πρόσθετα σε σχέση με το SD 1.5, υπολείπεται του SDXL σε λεπτομέρεια και φωτορεαλισμό
Stable Diffusion XL (SDXL)	Latent Diffusion	Ανοιχτό	Ναι	Υψηλότερη ανάλυση και λεπτομέρεια	Αυξημένες απαιτήσεις μνήμης
DALL·E 3	Transformer + Diffusion	Κλειστό	Όχι	Πολύ καλή κατανόηση prompts	Μη διαθέσιμο για τοπική εκπαίδευση
Imagen	Diffusion	Κλειστό	Όχι	Υψηλή φωτορεαλιστικότητα	Περιορισμένη πρόσβαση
Midjourney	Proprietary	Κλειστό	Όχι	Εξαιρετική καλλιτεχνική ποιότητα	Έλλειψη ελέγχου και παραμετροποίησης
Flux	Transformer-based Diffusion	Ανοιχτό	Ναι	Πολύ καλή πιστότητα στο prompt	Υψηλές απαιτήσεις υπολογιστικών πόρων

4.5 Κριτική Ανάλυση των Σύγχρονων Μοντέλων Δημιουργίας Εικόνων

Τα μοντέλα δημιουργίας εικόνων παρουσιάζουν σημαντικές διαφορές ως προς την ποιότητα των παραγόμενων αποτελεσμάτων, τις υπολογιστικές απαιτήσεις και τη δυνατότητα προσαρμογής τους σε εξειδικευμένα δεδομένα. Για τον λόγο αυτό, η επιλογή ενός μοντέλου εξαρτάται από τις ανάγκες της εφαρμογής και τους διαθέσιμους υπολογιστικούς πόρους.

Το Stable Diffusion 1.5 αποτελεί μία από τις πιο δημοφιλείς ανοικτού κώδικα λύσεις, κυρίως λόγω των χαμηλών απαιτήσεων σε υπολογιστικούς πόρους και της εκτεταμένης υποστήριξης από την κοινότητα. Επιπλέον, υποστηρίζει αποτελεσματικά τεχνικές προσαρμογής όπως οι LoRA και DreamBooth. Ωστόσο, η ποιότητα των παραγόμενων εικόνων και η πιστότητα στα σύνθετα prompts υστερούν σε σχέση με νεότερα μοντέλα.

Το Stable Diffusion 2.1 αποτελεί εξέλιξη του Stable Diffusion 1.5 και εισάγει βελτιώσεις στην ποιότητα των παραγόμενων εικόνων, στην κατανόηση των λεκτικών περιγραφών και στην υποστήριξη υψηλότερων αναλύσεων. Διατηρεί το βασικό πλεονέκτημα των ανοικτών μοντέλων, δηλαδή τη δυνατότητα προσαρμογής μέσω τεχνικών όπως LoRA, DreamBooth και Textual Inversion, γεγονός που το καθιστά ιδιαίτερα χρήσιμο για ερευνητικές εφαρμογές και εξειδικευμένα σύνολα δεδομένων.

Σε σύγκριση με το Stable Diffusion 1.5, προσφέρει πιο συνεκτικές συνθέσεις και βελτιωμένη απόδοση σε σύνθετα prompts. Ωστόσο, η μετάβαση από την έκδοση 1.5 στην 2.1 δεν συνοδεύτηκε από αντίστοιχη αύξηση της δημοτικότητας στην ερευνητική και αναπτυξιακή κοινότητα, με αποτέλεσμα να υπάρχουν λιγότερα διαθέσιμα προεκπαιδευμένα checkpoints και εργαλεία υποστήριξης.

Επιπλέον, παρότι υπερτερεί του SD 1.5 σε ποιότητα εικόνας, υστερεί έναντι του Stable Diffusion XL (SDXL) ως προς τη λεπτομέρεια, τη φωτορεαλιστικότητα και τη συνολική ποιότητα των παραγόμενων αποτελεσμάτων.

Επομένως, το Stable Diffusion 2.1 μπορεί να θεωρηθεί μια ενδιάμεση λύση μεταξύ του SD 1.5 και του SDXL, συνδυάζοντας σχετικά χαμηλές απαιτήσεις υπολογιστικών πόρων με βελτιωμένη ποιότητα εικόνας, χωρίς όμως να επιτυγχάνει τις επιδόσεις των νεότερων μοντέλων.

Το Stable Diffusion XL (SDXL) βελτιώνει σημαντικά την ποιότητα εικόνας, την ανάλυση και την απόδοση σε σύνθετες περιγραφές κειμένου. Παρά τα πλεονεκτήματά του, απαιτεί σημαντικά περισσότερη μνήμη GPU και μεγαλύτερο χρόνο εκπαίδευσης και παραγωγής εικόνων, γεγονός που περιορίζει τη χρήση του σε συστήματα περιορισμένων πόρων.

Το DALL·E 3 διακρίνεται για την εξαιρετική κατανόηση φυσικής γλώσσας και την ικανότητά του να ακολουθεί με ακρίβεια σύνθετες λεκτικές οδηγίες. Ωστόσο, πρόκειται για κλειστό εμπορικό σύστημα, γεγονός που δεν επιτρέπει την εκπαίδευση ή προσαρμογή του σε εξειδικευμένα σύνολα δεδομένων.

Το Imagen επιτυγχάνει ιδιαίτερα υψηλό επίπεδο φωτορεαλισμού και θεωρείται από τα κορυφαία μοντέλα ως προς την οπτική ποιότητα. Παρόλα αυτά, η περιορισμένη πρόσβαση στο μοντέλο δυσχεραίνει την αξιοποίησή του στην ακαδημαϊκή έρευνα και την αναπαραγωγή πειραμάτων.

Το Midjourney παράγει εικόνες υψηλής αισθητικής και καλλιτεχνικής ποιότητας, γεγονός που το καθιστά ιδιαίτερα δημοφιλές σε δημιουργικές εφαρμογές. Εντούτοις, η κλειστή φύση του συστήματος και οι περιορισμένες δυνατότητες

παραμετροποίησης δεν επιτρέπουν ουσιαστική έρευνα πάνω στην εκπαίδευση και προσαρμογή του μοντέλου.

Τέλος, το Flux αποτελεί μία από τις νεότερες ανοικτές προσεγγίσεις και εμφανίζει εξαιρετική πιστότητα στην απόδοση των prompts, προσεγγίζοντας σε αρκετές περιπτώσεις τις επιδόσεις εμπορικών λύσεων. Το βασικό του μειονέκτημα είναι οι αυξημένες απαιτήσεις σε υπολογιστικούς πόρους, οι οποίες καθιστούν δυσκολότερη την εκπαίδευση και αξιοποίησή του σε συνηθισμένο εξοπλισμό.

Συνολικά, τα ανοικτά μοντέλα, όπως τα Stable Diffusion και Flux, υπερτερούν ως προς τη δυνατότητα προσαρμογής και πειραματισμού, ενώ τα κλειστά μοντέλα, όπως τα DALL·E 3, Imagen και Midjourney, προσφέρουν συχνά ανώτερη ποιότητα εικόνας και ευχρηστία. Η επιλογή μεταξύ των δύο κατηγοριών αποτελεί συμβιβασμό μεταξύ ελέγχου, παραμετροποίησης, κόστους και ποιότητας αποτελεσμάτων.

Για την υλοποίηση των πειραμάτων επιλέχθηκαν τα μοντέλα **Stable Diffusion 1.5** και **Stable Diffusion 2.1**, καθώς αποτελούν δύο από τις πλέον διαδεδομένες ανοικτού κώδικα υλοποιήσεις μοντέλων δημιουργίας εικόνων. Η επιλογή τους βασίστηκε τόσο στη δυνατότητα εφαρμογής τεχνικών προσαρμογής (fine-tuning), όπως οι LoRA και DreamBooth, όσο και στις σχετικά περιορισμένες απαιτήσεις τους σε υπολογιστικούς πόρους σε σύγκριση με νεότερα μοντέλα, όπως το SDXL ή το Flux.

Το Stable Diffusion 1.5 επιλέχθηκε λόγω της ευρείας χρήσης του στην ερευνητική κοινότητα, της μεγάλης διαθεσιμότητας εργαλείων και προεκπαιδευμένων μοντέλων, καθώς και της εκτεταμένης τεκμηρίωσης που υποστηρίζει την αναπαραγωγή των πειραμάτων. Παρά το γεγονός ότι η ποιότητα των παραγόμενων εικόνων υπολείπεται

των νεότερων μοντέλων, εξακολουθεί να αποτελεί σημείο αναφοράς για τη μελέτη τεχνικών προσαρμογής.

Το Stable Diffusion 2.1 επιλέχθηκε ως νεότερη έκδοση της ίδιας αρχιτεκτονικής, προσφέροντας βελτιωμένη κατανόηση των prompts, καλύτερη συνοχή στη σύνθεση των εικόνων και υποστήριξη υψηλότερης ανάλυσης. Η σύγκρισή του με το Stable Diffusion 1.5 επιτρέπει τη διερεύνηση του κατά πόσο οι βελτιώσεις της νεότερης έκδοσης επηρεάζουν την αποτελεσματικότητα των τεχνικών fine-tuning και την ποιότητα των παραγόμενων αποτελεσμάτων.

Η χρήση και των δύο μοντέλων επιτρέπει μια πιο ολοκληρωμένη αξιολόγηση, καθώς παρέχει τη δυνατότητα σύγκρισης διαφορετικών εκδόσεων της ίδιας οικογένειας μοντέλων υπό κοινές πειραματικές συνθήκες. Με τον τρόπο αυτό μπορούν να εξαχθούν ασφαλέστερα συμπεράσματα σχετικά με την επίδραση των τεχνικών προσαρμογής και την απόδοση των μοντέλων δημιουργίας εικόνων σε εξειδικευμένα σύνολα δεδομένων.

Πειραματικό Μέρος και Αποτελέσματα

5.1 Ορισμός Σεναρίων και Prompts

Η παρούσα εργασία ακολουθεί πειραματική και συγκριτική μεθοδολογική προσέγγιση με στόχο την αξιολόγηση προεκπαιδευμένων diffusion models για παραγωγή εικόνας από κείμενο (text-to-image). Τα πειράματα πραγματοποιήθηκαν χρησιμοποιώντας μοντέλα διαθέσιμα μέσω της πλατφόρμας Hugging Face και συγκεκριμένα τις εκδόσεις του Stable Diffusion 1.5 και 2.1.

Χρησιμοποιήσαμε απλά prompts και τα χωρίσαμε σε 5 κατηγορίες (πίνακας 1). Στην συνέχεια εφαρμόσαμε τεχνικές prompt engineering και από το βασικό σύνολο prompts δημιουργήσαμε 4 σύνολα με τις παρακάτω τεχνικές: context tokens , emphasis (έμφαση λέξεων), repetition (επαναληπτικές εντολές), order (αλλαγή σειράς στα πλαίσια του περιεχομένου). Για κάθε σύνολο υπολογίσαμε το Clip score.

Απλά prompts

Απλά αντικείμενα	A cup on a table with a book beside it A car parked in front of a small house A lamp standing next to a sofa A chair under a window A plate with food in the center of a table
Ζώα	A brown dog running on grass A white cat sitting on a sofa A small yellow bird sitting on a tree branch A horse standing in a field A fish swimming underwater
Άνθρωποι	A woman wearing a black coat, standing on a city street A man smiling while holding a cup of coffee A child playing with a ball in a park A person sitting at a desk using a laptop A man walking down a street at night
Σκηνές Περιβάλλον	A quiet beach at sunset A snowy mountain landscape with pine trees A city street during heavy rain A small village surrounded by hills A forest path in autumn

Χωρικές οδηγίες	A cup on a table with a book beside it A car parked in front of a small house A lamp standing next to a sofa A chair under a window A plate with food in the center of a table
-----------------	--

Πίνακας 1: Απλά prompts

5.2 Πειραματική Διαδικασία

Στο πλαίσιο της πειραματικής αξιολόγησης πραγματοποιήθηκαν πειράματα στα μοντέλα διάχυσης (diffusion models) **Stable Diffusion v1.5** και **Stable Diffusion v2.1**.

Αρχικά εφαρμόστηκαν τεχνικές **prompt engineering**. Για κάθε prompt παράχθηκαν τέσσερις εικόνες με διαφορετικά **seeds**, ώστε να υπολογιστούν η μέση τιμή και η τυπική απόκλιση του **CLIP Score**. Κατά την παραγωγή των εικόνων χρησιμοποιήθηκαν οι παράμετροι **guidance scale = 7.5** και **50 inference steps**. Οι εικόνες που προέκυψαν από τις τεχνικές prompt engineering χρησιμοποιήθηκαν επίσης για την ποιοτική αξιολόγηση των δύο μοντέλων. Τα αποτελέσματα της αξιολόγησης παρουσιάζονται στο **Παράρτημα Α΄**.

Στη συνέχεια χρησιμοποιήθηκε η βάση δεδομένων **COCO (Common Objects in Context)** για τον υπολογισμό των μετρικών **FID (Fréchet Inception Distance)** και **CLIP Score**. Για τον σκοπό αυτό χρησιμοποιήθηκαν 1.000 πραγματικές εικόνες του συνόλου δεδομένων μαζί με τις αντίστοιχες λεζάντες (captions).

Η ποσοτική αξιολόγηση ολοκληρώθηκε με τον υπολογισμό του **Aesthetic Score (αισθητική βαθμολογία)**. Για τον υπολογισμό του Aesthetic Score χρησιμοποιήθηκαν 1.000 captions από τη βάση δεδομένων COCO. Με βάση τα captions δημιουργήθηκαν

συνθετικές εικόνες από τα δύο μοντέλα Stable Diffusion, οι οποίες στη συνέχεια αξιολογήθηκαν με το προεκπαιδευμένο μοντέλο **Improved Aesthetic Predictor (CLIP ViT-L/14)**, ώστε να υπολογιστεί η αισθητική τους βαθμολογία.

Τέλος, πραγματοποιήθηκε **fine-tuning** του μοντέλου **Stable Diffusion v1.5** με την τεχνική **Low-Rank Adaptation (LoRA)** για τη δημιουργία συνθετικών ακτινογραφιών θώρακα. Για την εκπαίδευση χρησιμοποιήθηκε η βάση δεδομένων **COVID-19 Radiography Database**.

Όλα τα πειράματα υλοποιήθηκαν με χρήση **Jupyter Notebooks**, της βιβλιοθήκης **PyTorch** και του περιβάλλοντος **Google Colab**. Ως υπολογιστικό υλικό χρησιμοποιήθηκε κάρτα γραφικών **NVIDIA Tesla T4 GPU** με μνήμη **12 GB VRAM**.

5.3 Ποιοτική και Ποσοτική Αξιολόγηση των Εικόνων

5.3.1 Prompt engineering - Clip score

Στο συγκεκριμένο πείραμα συγκρίνουμε τα μοντέλα **Stable Diffusion v1.5** και **Stable Diffusion v2.1** με χρήση της τεχνικής prompt engineering. Χρησιμοποιήσαμε 5 κατηγορίες prompts. Τα απλά prompts (P0) αποτελούνται από 5 υποκατηγορίες «Απλά αντικείμενα», «Ζώα», «Άνθρωποι», «Σκηνές - Περιβάλλον» και «Χωρικές οδηγίες» με 5 prompts ανά υποκατηγορία. Άρα συνολικά έχουμε 25 prompts ανά κατηγορία.

Από τα απλά prompts δημιουργήσαμε 4 σύνολα με τις παρακάτω τεχνικές: context tokens (P1), emphasis (έμφαση λέξεων) (P2), repetition (επαναληπτικές εντολές) (P3), order (αλλαγή σειράς στα πλαίσια του περιεχομένου) (P4).

Για κάθε σύνολο και υποκατηγορία υπολογίστηκαν μέσοι όροι του CLIP score και η τυπική απόκλιση. Τα αποτελέσματα φαίνονται στον παρακάτω πίνακα (Πίνακας 2).

CLIP Score		
	MONTEAO	
P0-Base	Stable Diffusion 1.5	Stable Diffusion 2.1
Απλά αντικείμενα	0,306±0,022	0,298±0,029
Ζώα	0,323±0,028	0,323±0,028
Άνθρωποι	0,309±0,030	0,311±0,034
Σκηνές Περιβάλλον	0,300±0,014	0,300±0,009
Χωρικές οδηγίες	0,294±0,017	0,299±0,014
CLIP Score		
	MONTEAO	
P1-Context	Stable Diffusion 1.5	Stable Diffusion 2.1
Απλά αντικείμενα	0,305±0,020	0,317±0,007

Ζώα	0,324±0,019	0,327±0,033
Άνθρωποι	0,320±0,037	0,317±0,039
Σκηνές Περιβάλλον	0,306±0,015	0,323±0,012
Χωρικές οδηγίες	0,316±0,023	0,326±0,015

CLIP Score		
	MONTELO	
P2-Emphasis	Stable Diffusion 1.5	Stable Diffusion 2.1
Απλά αντικείμενα	0,297±0,030	0,295±0,031
Ζώα	0,314±0,020	0,315±0,023
Άνθρωποι	0,308±0,024	0,314±0,030
Σκηνές Περιβάλλον	0,302±0,022	0,310±0,018
Χωρικές οδηγίες	0,297±0,019	0,310±0,011

CLIP Score		
	MONTELO	
P3-Repetition	Stable Diffusion 1.5	Stable Diffusion 2.1
Απλά αντικείμενα	0,298±0,029	0,299±0,028
Ζώα	0,318±0,023	0,317±0,029
Άνθρωποι	0,313±0,027	0,306±0,032
Σκηνές Περιβάλλον	0,303±0,011	0,304±0,016
Χωρικές οδηγίες	0,298±0,017	0,306±0,009

CLIP Score		
	MONTELO	
P4-Order	Stable Diffusion 1.5	Stable Diffusion 2.1
Απλά αντικείμενα	0,301±0,018	0,315±0,024
Ζώα	0,327±0,024	0,327±0,030

Άνθρωποι	0,327±0,038	0,317±0,041
Σκηνές Περιβάλλον	0,310±0,010	0,316±0,014
Χωρικές οδηγίες	0,310±0,015	0,308±0,015

Πίνακας 2: μέσοι όροι clip score και τυπική απόκλιση

Τα συνολικά αποτελέσματα από τα πειράματα δίνονται στο *Παράρτημα Α'* και τα scripts που χρησιμοποιήθηκαν δίνονται στο *Παράρτημα Β'*. Επίσης οι εικόνες που παράχθηκαν με τις τεχνικές prompt engineering χρησιμοποιήθηκαν στην παράγραφο 5.3.4 για την ποιοτική αξιολόγηση των δύο μοντέλων.

Από τον Πίνακα 2 παρατηρείται ότι οι διαφορές στο CLIP Score μεταξύ των μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1 είναι γενικά μικρές. Επιπλέον, λαμβάνοντας υπόψη και τις αντίστοιχες τυπικές αποκλίσεις, δεν προκύπτει σαφής υπεροχή κάποιου από τα δύο μοντέλα ως προς τη σημασιολογική ευθυγράμμιση εικόνας – κειμένου. Το συμπέρασμα αυτό ενισχύεται από τα αποτελέσματα της επόμενης ενότητας, όπου ο υπολογισμός των μετρικών FID και CLIP Score με χρήση της βάσης δεδομένων COCO οδηγεί επίσης σε πολύ παρόμοιες τιμές CLIP Score για τα δύο μοντέλα.

5.3.2 FID και Clip score - Coco database

Για την ποσοτική αξιολόγηση των μοντέλων χρησιμοποιήθηκε η βάση δεδομένων **COCO (Common Objects in Context)**. Επιλέχθηκαν **1.000 εικόνες** από το σύνολο **COCO val2017** και για κάθε εικόνα χρησιμοποιήθηκε μία αντίστοιχη λεζάντα (caption). Όλες οι πραγματικές εικόνες μετατράπηκαν σε ανάλυση **512×512 pixels**, ώστε να είναι συμβατές με την ανάλυση εξόδου των μοντέλων. Το script που χρησιμοποιήθηκε δίνεται στο *Παράρτημα Γ'*.

Με βάση τα ίδια captions δημιουργήθηκαν δύο σύνολα συνθετικών εικόνων, ένα με το **Stable Diffusion v1.5** και ένα με το **Stable Diffusion v2.1**. Για τη διασφάλιση συγκρίσιμων αποτελεσμάτων χρησιμοποιήθηκε το ίδιο **seed (42)** και για τα δύο μοντέλα. Η παραγωγή των εικόνων πραγματοποιήθηκε με **30 inference steps** και **guidance scale = 10**.

Η αξιολόγηση πραγματοποιήθηκε με δύο συμπληρωματικές μετρικές. Ο **Fréchet Inception Distance (FID)** χρησιμοποιήθηκε για τη μέτρηση της ομοιότητας μεταξύ των πραγματικών εικόνων της βάσης COCO και των εικόνων που παρήχθησαν από κάθε μοντέλο. Παράλληλα, υπολογίστηκε και ο **CLIP Score**, ο οποίος εκτιμά τον βαθμό σημασιολογικής αντιστοίχισης μεταξύ του caption και της παραγόμενης εικόνας.

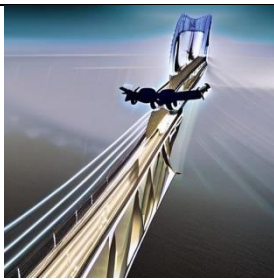
Τα αποτελέσματα παρουσιάζονται στον Πίνακα 3

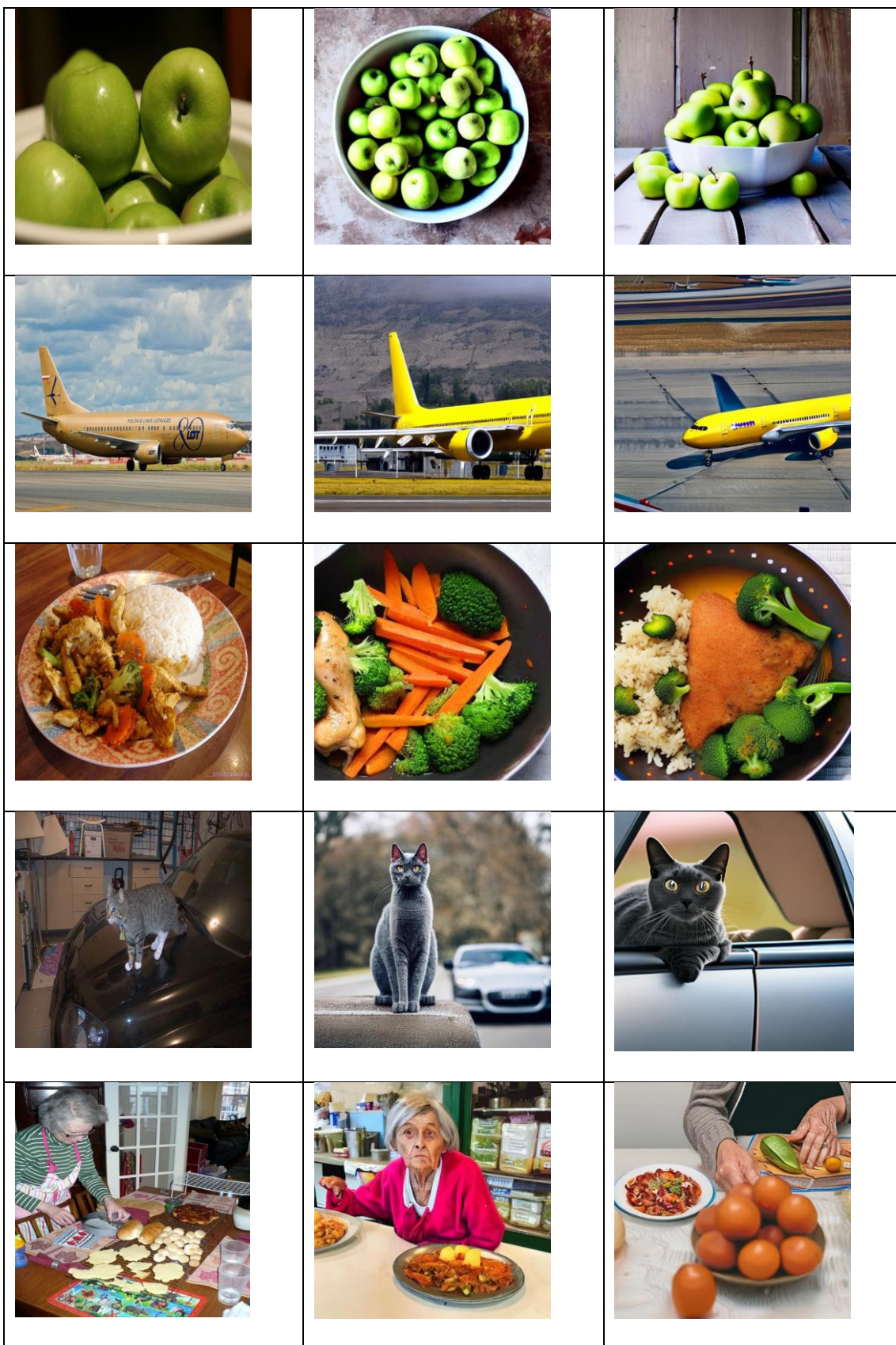
Model	FID	CLIP
SD1.5	62.40	30.984
SD2.1	70.86	29.490

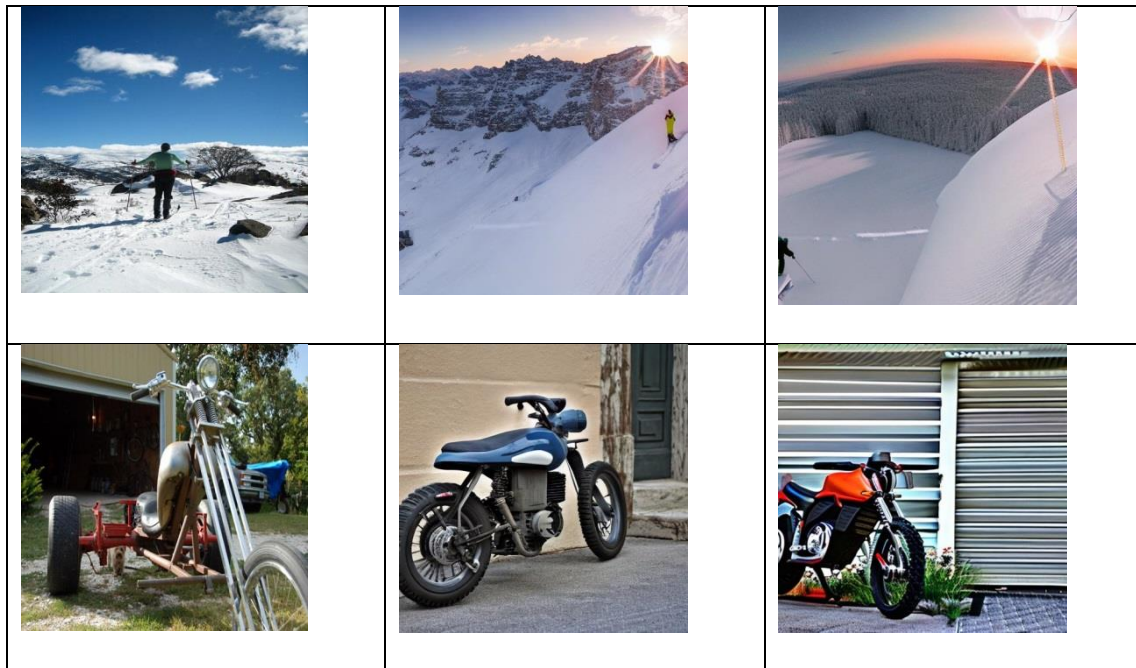
. Πίνακας 3

Από τα αποτελέσματα παρατηρείται ότι το **Stable Diffusion v1.5** επιτυγχάνει μικρότερη τιμή **FID** σε σχέση με το **Stable Diffusion v2.1**, γεγονός που υποδηλώνει ότι οι παραγόμενες εικόνες του εμφανίζουν μεγαλύτερη ομοιότητα με τις πραγματικές εικόνες της βάσης COCO. Παράλληλα, το μοντέλο Stable Diffusion v1.5 παρουσιάζει ελαφρώς υψηλότερο **CLIP Score**, γεγονός που δείχνει καλύτερη σημασιολογική αντιστοίχιση μεταξύ της παραγόμενης εικόνας και του αντίστοιχου caption. Ωστόσο, οι διαφορές στις τιμές του CLIP Score είναι μικρές, γεγονός που συμφωνεί με τα αποτελέσματα της προηγούμενης ενότητας, όπου επίσης δεν παρατηρήθηκε σαφής υπεροχή κάποιου από τα δύο μοντέλα ως προς τη σημασιολογική ευθυγράμμιση εικόνας–κειμένου.

Στον παρακάτω πίνακα (Πίνακας 4) έχουμε παραθέσει ενδεικτικά μερικές εικόνες που παράχθηκαν από τα δύο μοντέλα καθώς και τις αντίστοιχες πραγματικές εικόνες του dataset.

Real Image	SD 1.5	SD 2.1
		





Πίνακας 4: Πραγματικές εικόνες και generated images (SD1.5 – SD2.1)

5.3.3 Aesthetic score - COCO dataset

Για την αξιολόγηση της αισθητικής ποιότητας των παραγόμενων εικόνων χρησιμοποιήθηκε η μετρική **Aesthetic Score**. Επιλέχθηκαν **1.000 captions** από τη βάση δεδομένων **COCO**, από τα οποία δημιουργήθηκαν αντίστοιχες εικόνες με τα μοντέλα **Stable Diffusion v1.5** και **Stable Diffusion v2.1**. Το script που χρησιμοποιήθηκε δίνεται στο *Παράρτημα Δ'*.

Η αξιολόγηση πραγματοποιήθηκε με το προεκπαιδευμένο μοντέλο **Improved Aesthetic Predictor**, το οποίο βασίζεται στα χαρακτηριστικά του **CLIP ViT-L/14** και έχει εκπαιδευτεί να προβλέπει ανθρώπινες αισθητικές αξιολογήσεις σε κλίμακα **1–10**. Για κάθε εικόνα υπολογίστηκε μία αισθητική βαθμολογία και στη συνέχεια εξήχθη ο μέσος όρος για κάθε μοντέλο.

Τα αποτελέσματα παρουσιάζονται στον Πίνακα 5.

Μοντέλο	Mean Aesthetic Score
Stable Diffusion v1.5	5.251
Stable Diffusion v2.1	5.148

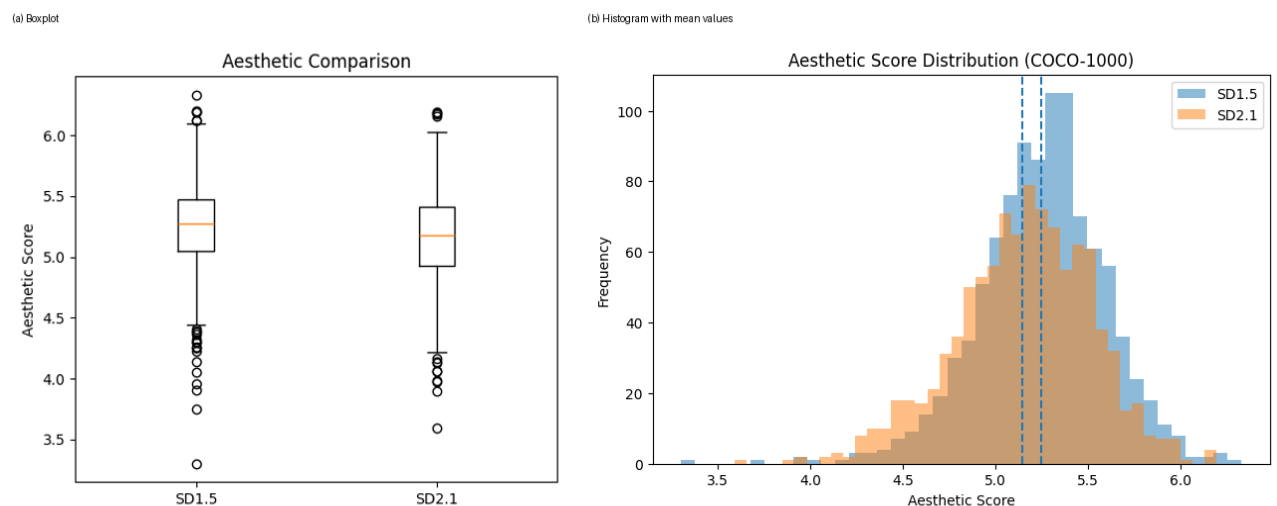
Πίνακας 5. Μέση αισθητική βαθμολογία (Aesthetic Score).

Από τα αποτελέσματα παρατηρείται ότι και τα δύο μοντέλα επιτυγχάνουν μέσες αισθητικές βαθμολογίες περίπου ίσες με **5**, γεγονός που υποδηλώνει εικόνες **μέτριας έως καλής αισθητικής ποιότητας**. Το **Stable Diffusion v1.5** παρουσιάζει ελαφρώς υψηλότερη μέση βαθμολογία (**5.25**) σε σχέση με το **Stable Diffusion v2.1** (**5.15**). Η διαφορά αυτή είναι μικρή (περίπου **0.10 μονάδες**) και επομένως, από πρακτική

άποψη, τα δύο μοντέλα εμφανίζουν παρόμοια αισθητική ποιότητα στις εικόνες που παράγουν.

Οι σχετικά μέτριες τιμές του Aesthetic Score θεωρούνται αναμενόμενες, καθώς χρησιμοποιήθηκαν γενικά περιγραφικά captions από τη βάση COCO και όχι prompts ειδικά σχεδιασμένα για τη δημιουργία εικόνων υψηλής καλλιτεχνικής ή αισθητικής ποιότητας

Παρότι ο στατιστικός έλεγχος έδειξε στατιστικά σημαντική διαφορά μεταξύ των δύο μοντέλων ($t = 6.40$, $p < 0.001$), η διαφορά στη μέση αισθητική βαθμολογία είναι μικρή (≈ 0.10 μονάδες). Επομένως, από πρακτική άποψη, τα δύο μοντέλα παρουσιάζουν παρόμοια αισθητική απόδοση. Για μεγαλύτερη ευκρίνεια παραθέτουμε τα διαγράμματα boxplot και Histogram με τις μέσες τιμές στο σχήμα 5.1



Σχήμα 5.1. Σύγκριση του Aesthetic Score για τα μοντέλα Stable Diffusion v1.5 και Stable Diffusion v2.1. **(α)** Boxplot της κατανομής των αισθητικών βαθμολογιών. **(β)** Ιστόγραμμα των βαθμολογιών με τις αντίστοιχες μέσες τιμές των δύο μοντέλων.

Το Σχήμα 5.1 παρουσιάζει την κατανομή των αισθητικών βαθμολογιών των δύο μοντέλων. Από το boxplot (Σχήμα 5.1α) παρατηρείται ότι οι διάμεσοι, η διασπορά

και το ενδοτεταρτημοριακό εύρος είναι ιδιαίτερα παρόμοια, ενώ και τα δύο μοντέλα εμφανίζουν περιορισμένο αριθμό ακραίων τιμών. Το ιστόγραμμα (Σχήμα 5.1β) δείχνει μεγάλη επικάλυψη των δύο κατανομών, με το Stable Diffusion v1.5 να εμφανίζει ελαφρώς υψηλότερη μέση αισθητική βαθμολογία από το Stable Diffusion v2.1

5.3.4 Ποιοτική αξιολόγηση

Παρότι οι μετρικές **FID**, **CLIP Score** και **Aesthetic Score** παρέχουν αντικειμενικά ποσοτικά κριτήρια αξιολόγησης, δεν μπορούν να αποτυπώσουν πλήρως την ποιότητα μιας εικόνας. Για παράδειγμα, το FID αξιολογεί την ομοιότητα της κατανομής των παραγόμενων εικόνων με τις πραγματικές εικόνες, χωρίς όμως να λαμβάνει υπόψη τη σημασιολογική ορθότητα κάθε εικόνας. Αντίστοιχα, το CLIP Score εκτιμά τη σημασιολογική συνάφεια εικόνας και κειμένου, χωρίς να αξιολογεί λεπτομέρειες όπως η σωστή απεικόνιση αντικειμένων ή η συνολική αισθητική ποιότητα. Για τον λόγο αυτό πραγματοποιήθηκε και ποιοτική αξιολόγηση των εικόνων, με οπτική σύγκριση των παραγόμενων εικόνων ως προς το αντίστοιχο prompt .

Η ποιοτική αξιολόγηση βασίστηκε στις εικόνες που δημιουργήθηκαν κατά τα πειράματα Prompt Engineering (P0–P4) και οργάνωθηκε στις πέντε κατηγορίες prompts: απλά αντικείμενα, ζώα, άνθρωποι, σκηνές/περιβάλλον και χωρικές οδηγίες.

5.3.4.1 P0 –Base prompts

Με τα βασικά prompts τα δύο μοντέλα παρουσιάζουν παρόμοια συνολική απόδοση. Το Stable Diffusion v1.5 αποδίδει ελαφρώς καλύτερα στα απλά αντικείμενα, ενώ το Stable Diffusion v2.1 εμφανίζει μικρή υπεροχή στις κατηγορίες άνθρωποι και χωρικές οδηγίες. Οι διαφορές παραμένουν περιορισμένες και συμφωνούν με τα αποτελέσματα του CLIP Score.

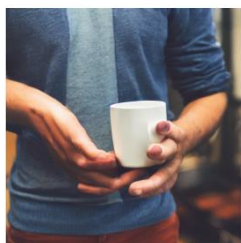
SD 1.5 απλά αντικείμενα			
A wooden chair CLIP score: 0.310 ± 0.003			
			
A Chinese ceramic jar CLIP score: 0.336 ± 0.006			
			
SD 2.1 απλά αντικείμενα			
A wooden chair CLIP score: 0.312 ± 0.009			
			

A Chinese ceramic jar
CLIP score: 0.332 ± 0.010



SD 1.5 'Άνθρωποι

A man smiling while holding a cup of coffee
CLIP score: 0.324 ± 0.024



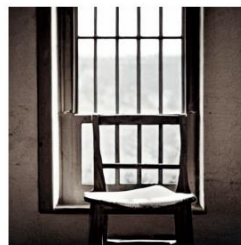
SD 2.1 'Άνθρωποι

A man smiling while holding a cup of coffee
CLIP score: 0.339 ± 0.009



SD 1.5 Χωρικές οδηγίες

A chair under a window
CLIP score: 0.299 ± 0.011



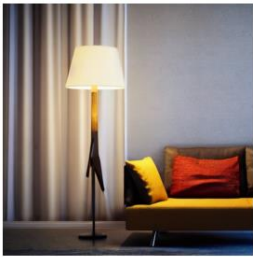
SD 2.1 Χωρικές οδηγίες

A chair under a window
CLIP score: 0.310 $\pm$ 0.007



5.3.4.2 P1 –Context

Η προσθήκη context tokens, δηλαδή επιπλέον περιγραφικών στοιχείων που εμπλουτίζουν το αρχικό prompt και παρέχουν περισσότερα συμφραζόμενα, φαίνεται να ευνοεί περισσότερο το Stable Diffusion v2.1. Το μοντέλο αποδίδει πληρέστερες σκηνές και εμφανίζει καλύτερη κατανόηση σύνθετων περιγραφών, ιδιαίτερα στις κατηγορίες σκηνές/περιβάλλον και χωρικές οδηγίες.

SD 1.5 Σκηνές / Περιβάλλον -Χωρικές οδηγίες			
A snowy mountain landscape with pine trees, wide landscape CLIP score: 0.296 ± 0.012			
			
A lamp standing next to a sofa, indoor living room, realistic lighting CLIP score: 0.340 ± 0.009			
			
A cup on a table with a book beside it, indoor scene, realistic lighting CLIP score: 0.319 ± 0.027			
			
SD 2.1 Σκηνές / Περιβάλλον -Χωρικές οδηγίες			

A snowy mountain landscape with pine trees, wide landscape
CLIP score: 0.338 ± 0.009



A lamp standing next to a sofa, indoor scene, realistic lighting
CLIP score: 0.332 ± 0.016

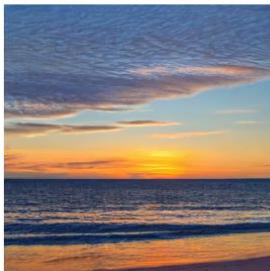
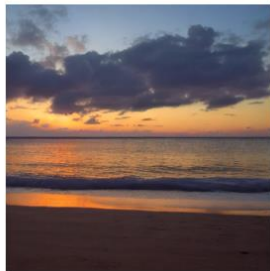
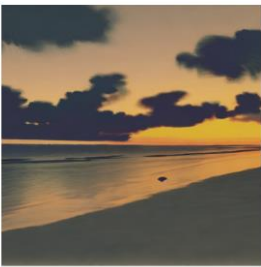
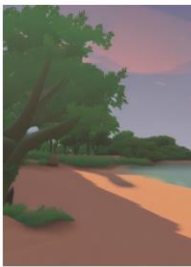


A cup on a table with a book beside it, realistic photography, natural lighting
CLIP score: 0.302 ± 0.018



5.3.4.3 P2 -Emphasis

Η χρήση έμφασης σε επιλεγμένες λέξεις δεν διαφοροποιεί σημαντικά την απόδοση των δύο μοντέλων στα απλά prompts. Ωστόσο, στις πιο σύνθετες περιγραφές το Stable Diffusion v2.1 παρουσιάζει ελαφρώς καλύτερη συνέπεια ως προς τη σημασιολογική απόδοση της εικόνας.

SD 1.5 Emphasis			
(A quiet beach at sunset:1.4) CLIP score: 0.283 ± 0.006			
			
(A car parked in front of a small house:1.4) CLIP score: 0.277 ± 0.023			
			
SD 2.1 Emphasis			
(A quiet beach at sunset:1.4) CLIP score: 0.304 ± 0.010			
			
(A car parked in front of a small house:1.4) CLIP score: 0.323 ± 0.007			
			

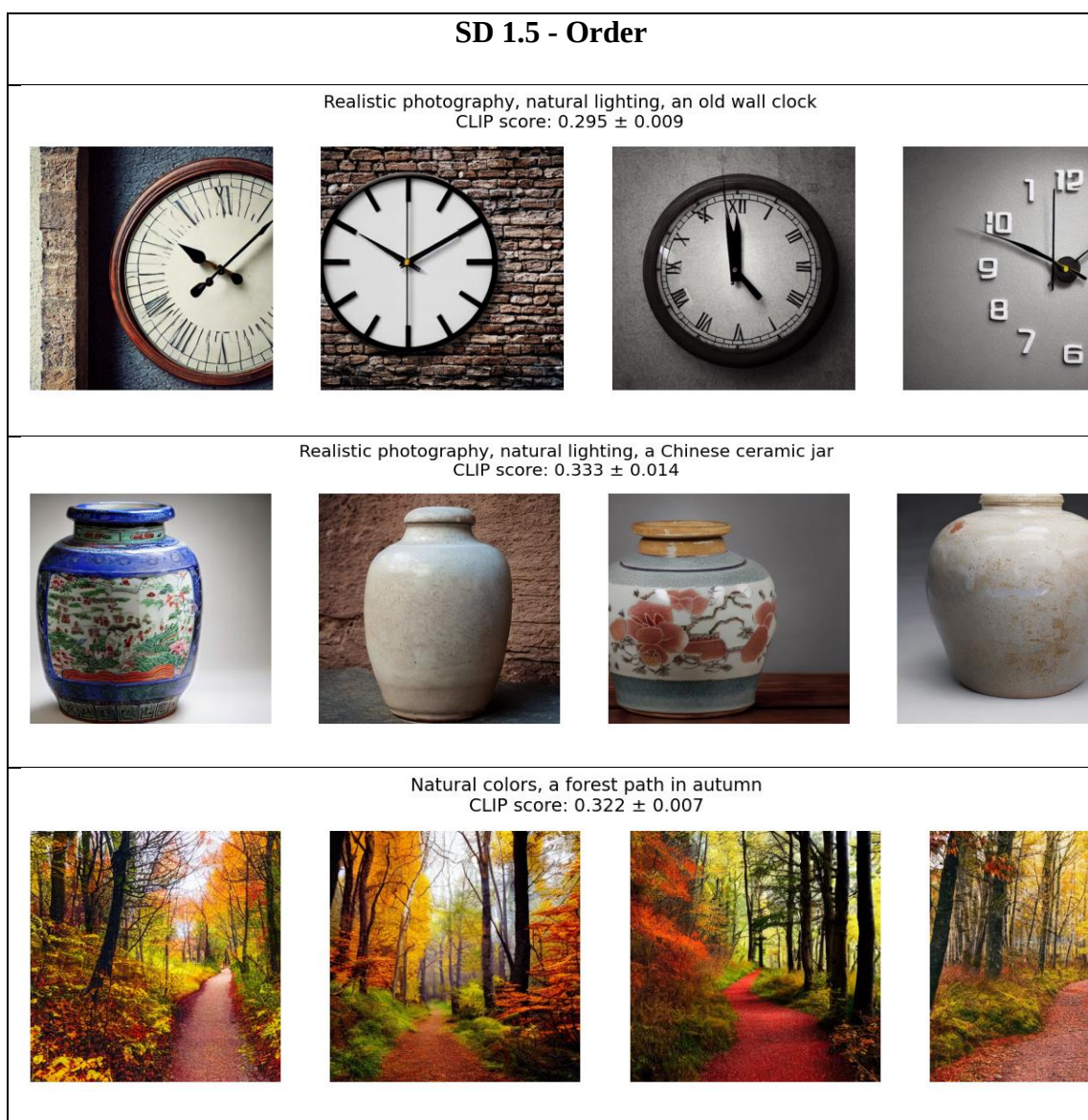
5.3.4.4 P3 -Repetition

Η επανάληψη βασικών όρων του prompt δεν οδηγεί σε ουσιαστική βελτίωση της ποιότητας των εικόνων. Τα δύο μοντέλα εμφανίζουν παρόμοια συμπεριφορά, με το Stable Diffusion v2.1 να υπερέχει στις χωρικές οδηγίες και το Stable Diffusion v1.5 να παρουσιάζει ελαφρώς καλύτερα αποτελέσματα στην κατηγορία άνθρωποι.

SD 1.5 Χωρικές οδηγίες - Άνθρωποι				
A chair under a window, chair under window CLIP score: 0.297 ± 0.012				
				
A man smiling while holding a cup of coffee, man smiling CLIP score: 0.343 ± 0.009				
				
SD 2.1 Χωρικές οδηγίες - Άνθρωποι				
A chair under a window, chair under window CLIP score: 0.311 ± 0.002				
				
A man smiling while holding a cup of coffee, man smiling CLIP score: 0.337 ± 0.011				
				

5.3.4.5 P4 -Order

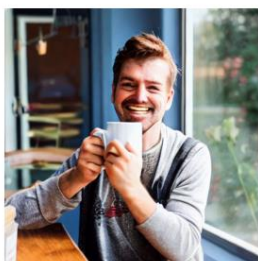
Η αλλαγή της σειράς των λέξεων στο prompt δεν επηρεάζει σημαντικά τη συνολική απόδοση των μοντέλων. Το Stable Diffusion v2.1 εμφανίζει καλύτερα αποτελέσματα στις κατηγορίες απλά αντικείμενα και σκηνές/περιβάλλον, ενώ το Stable Diffusion v1.5 παρουσιάζει μικρή υπεροχή στην κατηγορία άνθρωποι. Στις κατηγορίες ζώα και χωρικές οδηγίες δεν παρατηρούνται αξιοσημείωτες διαφορές.



Realistic photo, a woman wearing a black coat, standing on a city street
CLIP score: 0.340 ± 0.020



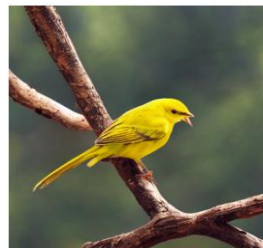
Natural lighting, a man smiling while holding a cup of coffee
CLIP score: 0.347 ± 0.014



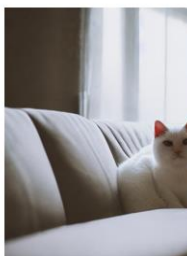
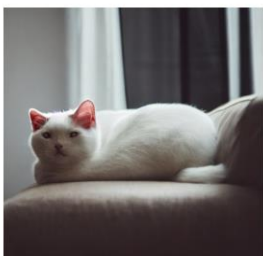
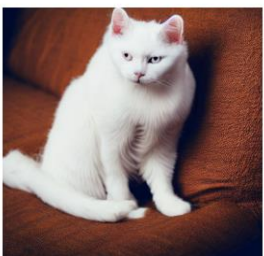
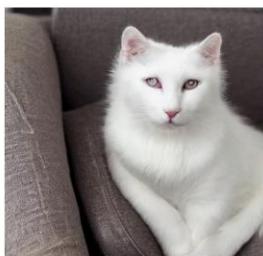
Outdoor photography, a child playing with a ball in a park
CLIP score: 0.355 ± 0.007



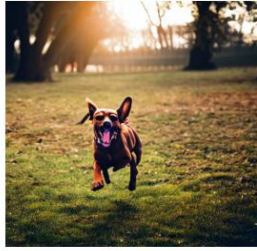
Wildlife photography, a small yellow bird sitting on a tree branch
CLIP score: 0.349 ± 0.003



Indoor photography, soft lighting, a white cat sitting on a sofa
CLIP score: 0.351 ± 0.007



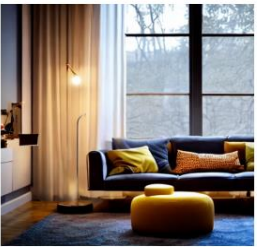
Outdoor photography, natural light, a brown dog running on grass
CLIP score: 0.322 ± 0.006



Indoor scene, realistic lighting, a cup on a table with a book beside it
CLIP score: 0.315 ± 0.012



Indoor living room, realistic lighting, a lamp standing next to a sofa
CLIP score: 0.311 ± 0.009



Indoor scene, natural light, a chair under a window
CLIP score: 0.331 ± 0.011



SD 2.1 - Order

Realistic photography, natural lighting, an old wall clock
CLIP score: 0.323 $\pm$ 0.006



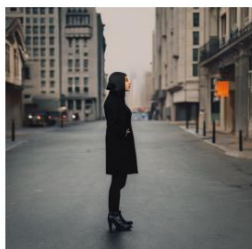
Realistic photography, natural lighting, a Chinese ceramic jar
CLIP score: 0.349 $\pm$ 0.005



Natural colors, a forest path in autumn
CLIP score: 0.333 $\pm$ 0.004



Realistic photo, a woman wearing a black coat, standing on a city street
CLIP score: 0.324 $\pm$ 0.024



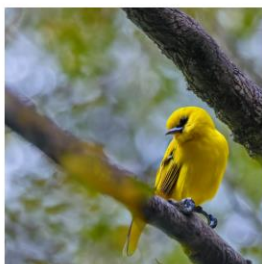
Natural lighting, a man smiling while holding a cup of coffee
CLIP score: 0.332 $\pm$ 0.014



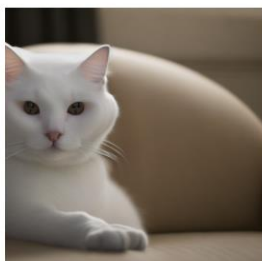
Outdoor photography, a child playing with a ball in a park
CLIP score: 0.352 ± 0.006



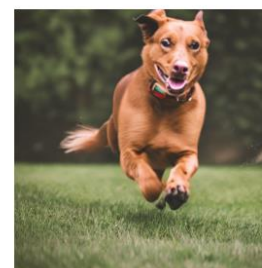
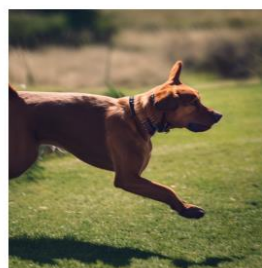
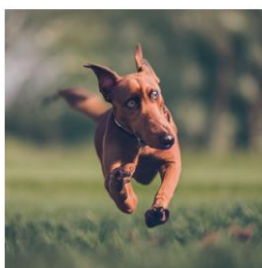
Wildlife photography, a small yellow bird sitting on a tree branch
CLIP score: 0.351 ± 0.011



Indoor photography, soft lighting, a white cat sitting on a sofa
CLIP score: 0.352 ± 0.013



Outdoor photography, natural light, a brown dog running on grass
CLIP score: 0.341 ± 0.007



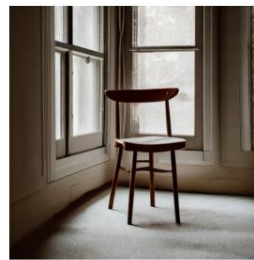
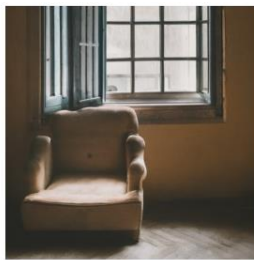
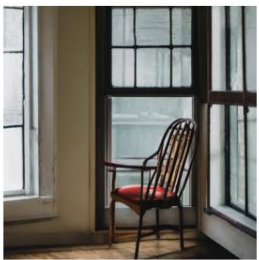
Indoor scene, realistic lighting, a cup on a table with a book beside it
CLIP score: 0.309 ± 0.035



Indoor living room, realistic lighting, a lamp standing next to a sofa
CLIP score: 0.315 ± 0.004



Indoor scene, natural light, a chair under a window
CLIP score: 0.320 ± 0.006

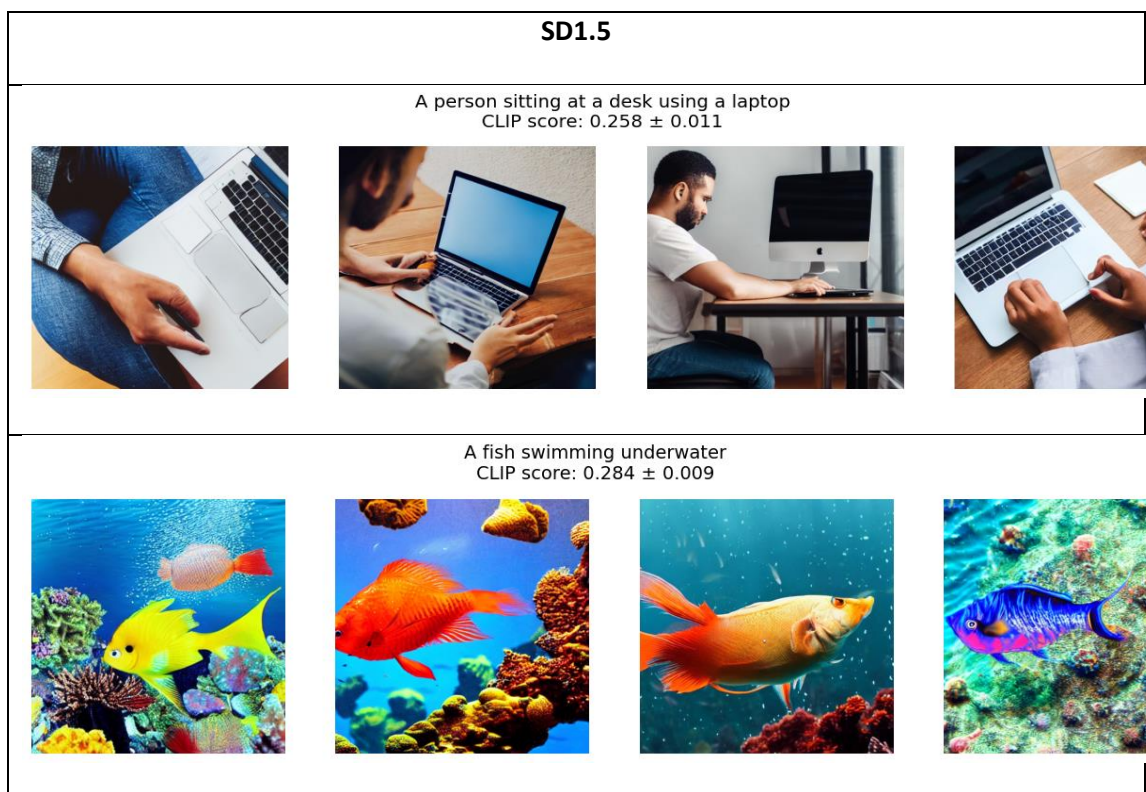


5.3.4.6 Γενικές Παρατηρήσεις

Πέρα από τη σύγκριση των δύο μοντέλων στις διαφορετικές τεχνικές prompt engineering, από την οπτική αξιολόγηση των παραγόμενων εικόνων προέκυψαν ορισμένα κοινά χαρακτηριστικά και περιορισμοί.

Πρώτον, τόσο το **Stable Diffusion v1.5** όσο και το **Stable Diffusion v2.1** παρουσιάζουν δυσκολία στην πιστή αναπαράσταση των ανθρώπινων χεριών, ιδιαίτερα στην απεικόνιση των δακτύλων. Σε αρκετές περιπτώσεις παρατηρήθηκαν λανθασμένος αριθμός δακτύλων, αφύσικες αναλογίες ή παραμορφώσεις.

Δεύτερον, σε ορισμένα prompts οι παραγόμενες εικόνες εμφάνισαν χαρακτηριστικά που προσεγγίζουν περισσότερο εικονογράφηση ή καρτουνίστικη αισθητική παρά φωτορεαλιστική απεικόνιση.

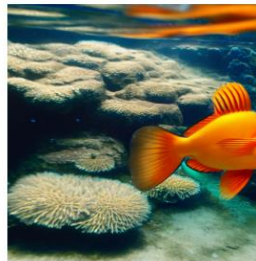


SD 2.1

A person sitting at a desk using a laptop
CLIP score: 0.253 ± 0.006



A fish swimming underwater
CLIP score: 0.285 ± 0.008



5.4 Εκπαίδευση Stable Diffusion v1.5 με LoRA

Εκπαιδεύσαμε το μοντέλο Stable Diffusion v1.5 με την τεχνική fine tuning Low-Rank Adaptation (LoRA) για την παραγωγή ακτινογραφιών θώρακα. Για την εκπαίδευση χρησιμοποιήθηκε η βάση [covid19-radiography-database](https://huggingface.co/datasets/covid19-radiography-database) και το script `train_text_to_image_lora.py` από το [Hugging face](https://huggingface.co).

Για την εκπαίδευση δημιουργήθηκε ένα dataset με τέσσερις κατηγορίες normal, covid-19, Viral Pneumonia, lung opacity με 250 εικόνες ακτινογραφιών θώρακα ανά κατηγορία. Έγιναν πειράματα με απλά captions π.χ. "<xray> chest x-ray, covid-19".

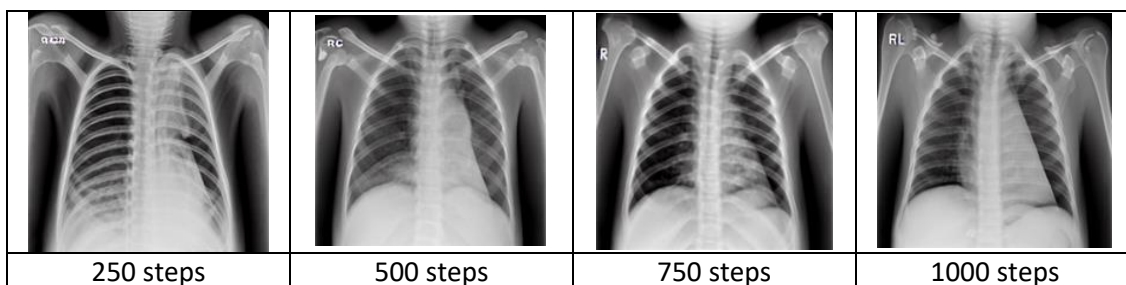
Για την εξαγωγή FID score χρησιμοποιήθηκαν 200 ακτινογραφίες ανά κατηγορία από την βάση δεδομένων και συγκρίθηκαν με 200 ακτινογραφίες ανά κατηγορία που παράχθηκαν από τον εκπαιδευμένο μοντέλο. Επίσης υπολογίστηκε το CLIP score για prompts με χρήση του tag και χωρίς χρήση του tag.

Έγιναν πειράματα με rank=32,16,8 και learning rate = 1e-4, 3e-5. Τα καλύτερα αποτελέσματα προέκυψαν για rank 8 και learning rate 3e-5.

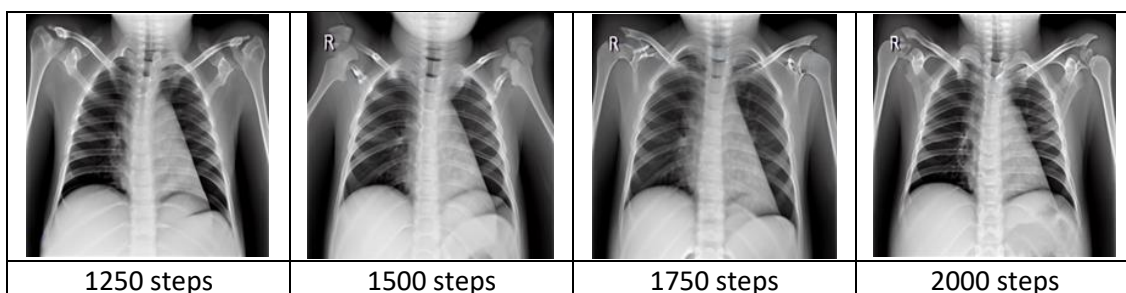
Κατά την εκπαίδευση χρησιμοποιήθηκε σύνδεση με το [Weights & Biases](https://wandb.ai) (wandb) για παρατήρηση των ενδιάμεσων αποτελεσμάτων και προσδιορισμό του βέλτιστου σημείου εκπαίδευσης.

5.4.1 Εκπαίδευση με απλά captions , rank =16 ,learning rate =1e-4

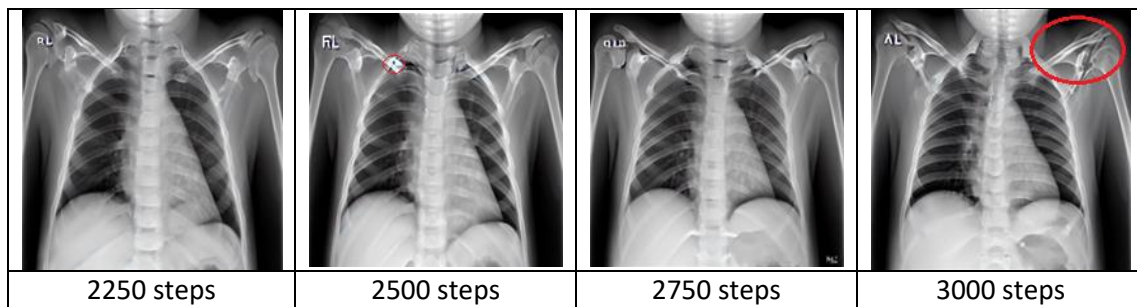
Παρακάτω δίνονται εικόνες από διαδοχικά στάδια εκπαίδευσης. Όλες οι φωτογραφίες παράχθηκαν από το ίδιο validation prompt "< xray> chest xray, frontal view, detailed lungs, medical imaging, high contrast".



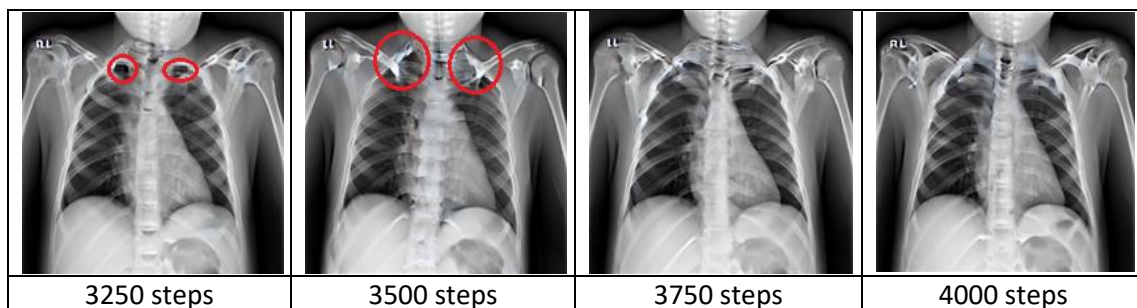
Παρατηρούμε ότι από τα πρώτα στάδια έχουμε εικόνες που έχουν χαρακτηριστικά ακτινογραφίας. Αυτό είναι αναμενόμενο αφού το μοντέλο είναι εκπαιδευμένο και γνωρίζει τα γενικά χαρακτηριστικά του ανθρώπινου σώματος και την μορφή μιας ακτινογραφίας. Επίσης παρατηρούμε ότι διαδοχικά η δομή του σκελετού (θώρακας, οστά ωμοπλάτης, κλειδώσεις) βελτιώνεται και οι πνεύμονες αρχίζουν να σχηματίζονται.



Στα πρώτα στάδια εκπαίδευσης 250 - 1000 steps παρατηρήσαμε γρήγορη βελτίωση ενώ στα στάδια 1250 – 2000 steps παρατηρούμε διόρθωση λεπτομερειών όπως πιο σωστή εμφάνιση των οστών –σκελετού.

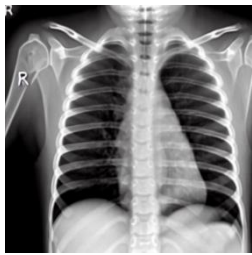
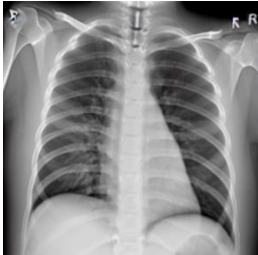


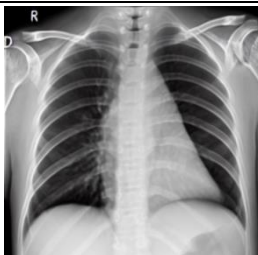
Στα επόμενα βήματα της εκπαίδευσης το μοντέλο αρχίζει να παράγει παραμορφώσεις και αρχίζουν να εμφανίζονται μπλε αποχρώσεις.

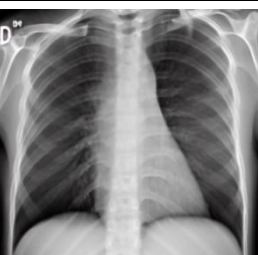
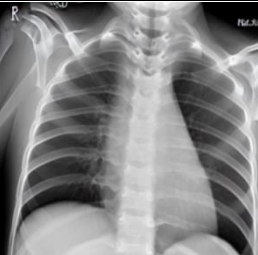


Στα 3250 steps αρχίζουν να παρουσιάζονται τεχνητά χαρακτηριστικά (artifacts) που γίνονται πιο έντονα μετά τα 3500 steps. Επομένως μετά τα 2000 steps το μοντέλο δεν βελτιώνεται και αρχίζει να παρουσιάζει σημάδια υπερεκπαίδευσης. Τα καλύτερα αποτελέσματα με βάση τις παραγόμενες εικόνες φαίνεται να είναι μεταξύ 1500 με 2000 steps.

Διαλέξαμε checkpoints στα 1400, 1600 και 1800 steps για να ελέγξουμε τις εξόδους με βάση prompts.

			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1400 steps			

			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1600 steps			

			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1800 steps			

Από τις εικόνες που παράχθηκαν από το μοντέλο για 1400 ,1600 και 1800 steps πιο έντονη διαφοροποίηση ανάλογα με το prompt παρατηρούμε στα 1400 steps. Επίσης έχουμε χαρακτηριστικά ακτινογραφίας και σωστή ανατομία. Στο παρακάτω πίνακα δίνουμε το FID και CLIP score για το μοντέλο των 1400 steps που επιλέξαμε.

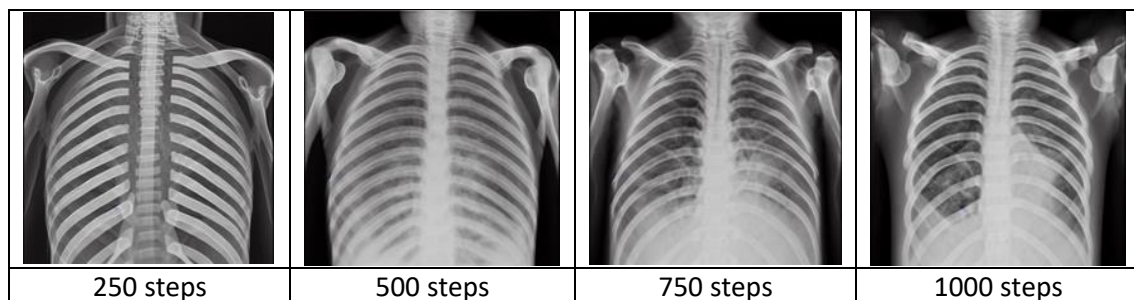
	FID score	CLIP score with tag	CLIP score without tag
Normal	151	32	31
Covid-19	271	34	34
Lung-Opacity	244	33	33
Viral Pneumonia	189	34	34

Από τα FID scores παρατηρούμε ότι σε όλες τις κατηγορίες έχουμε υψηλές τιμές που δείχνουν ότι το μοντέλο δεν έμαθε τις διαφορετικές περιπτώσεις. Τα Clip scores δείχνουν καλή ευθυγράμμιση με το κείμενο. Επίσης από τα clips scores παρατηρούμε ότι το μοναδικό tag δεν δίνει καμία διαφορά εφόσον το prompt αναφέρει ρητά την φράση x-ray.

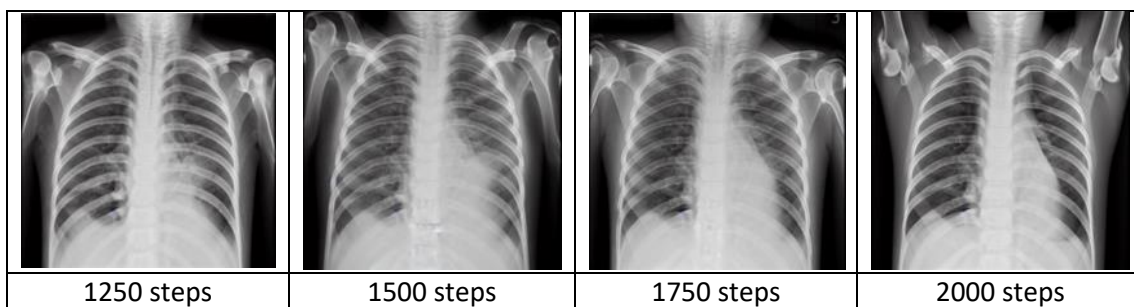
Για να πετύχουμε καλύτερα αποτελέσματα αλλάξαμε την παράμετρο rank σε 8 ώστε να μην έχουμε μεγάλη απομνημόνευση λεπτομερειών και την παράμετρο learning_rate σε $3e-5$ ώστε να έχουμε πιο αργή μάθηση.

5.4.2 Εκπαίδευση με απλά captions , rank =8 , learning rate = $3e-5$

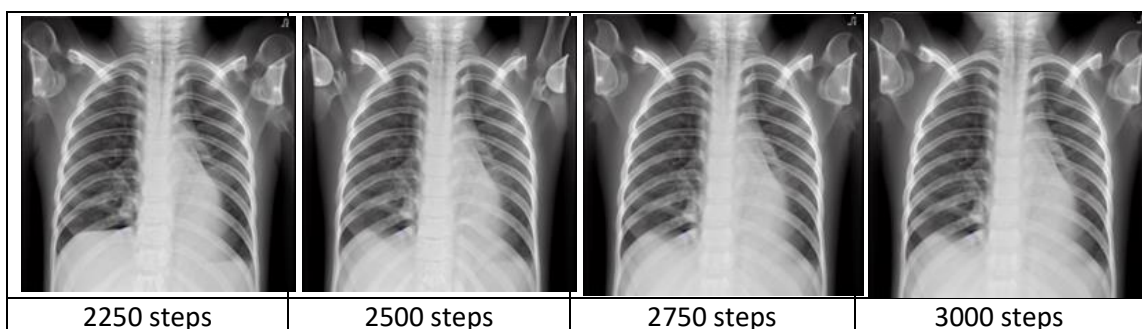
Παρακάτω δίνονται εικόνες από διάφορα στάδια της εκπαίδευσης.



Παρατηρούμε ότι στα 250 steps η εικόνα που έχουμε με rank 8 είναι πολύ μικρότερης λεπτομέρειας σε σχέση με την αντίστοιχη εικόνα με rank 16. Αυτό εξηγεί γιατί το μοντέλο με rank 16 έφτασε πολύ γρήγορα σε στάδια υπερκπαίδευσης χωρίς να γενικεύσει και να ξεχωρίζει τις διάφορες κατηγορίες.



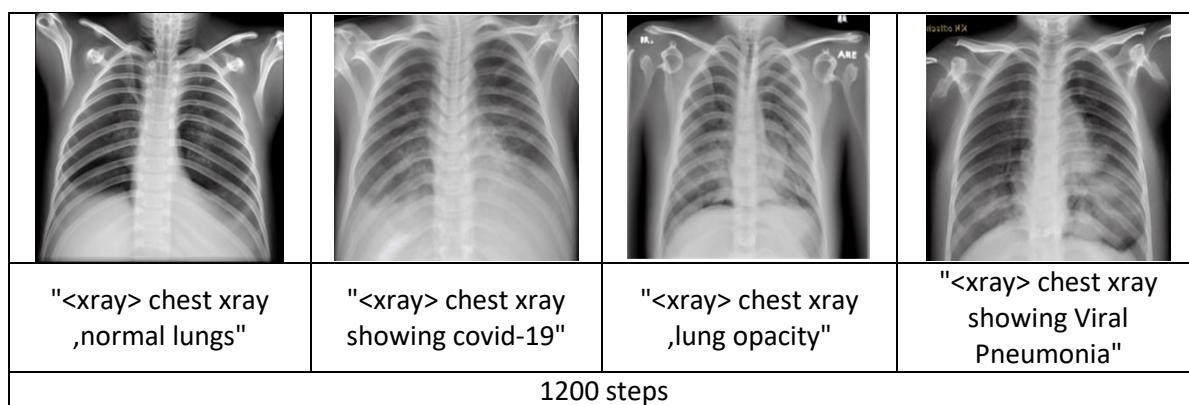
Στα επόμενα στάδια 1250 -2000 steps παρατηρούμε μικρές βελτιώσεις



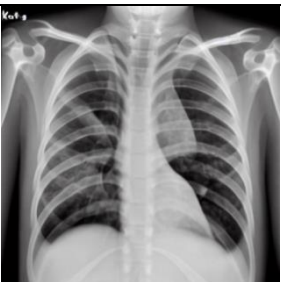
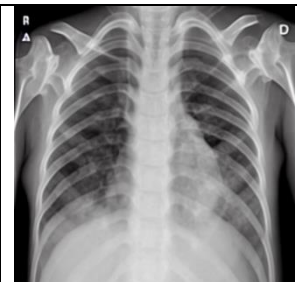
Στα τελευταία στάδια της εκπαίδευσης δεν παρατηρούμε περαιτέρω βελτίωση.

Από τις παραπάνω εικόνες παρατηρούμε ότι μεταξύ 1250 -2000 steps έχουμε τις καλύτερες εικόνες αφού στα τελευταία στάδια δεν παρατηρούμε αλλαγές.

Επιλέξαμε τα checkpoint στα 1200,1400,1600 και 1800 steps για να συγκρίνουμε τις εξόδους του μοντέλου σε απλά prompts για τις τέσσερις κατηγορίες.



			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1400 steps			

			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1600 steps			

			
"<xray> chest xray ,normal lungs"	"<xray> chest xray showing covid-19"	"<xray> chest xray ,lung opacity"	"<xray> chest xray showing Viral Pneumonia"
1800 steps			

Από τα παραπάνω αποτελέσματα καλύτερη διαφοροποίηση με τα prompts έχουμε στα 1400 steps. Για το μοντέλο στα 1400 steps υπολογίσαμε FID και CLIP scores:

Model 1400 steps

	FID score	CLIP score with tag	CLIP score without tag
Normal	102	32	31
Covid-19	179	34	34
Lung-Opacity	157	34	33
Viral Pneumonia	113	35	34

Παρατηρούμε μεγάλη βελτίωση (περίπου 35%) στα FID scores σε σχέση με το προηγούμενο πείραμα. Τα scores για τις κατηγορίες normal και viral pneumonia είναι αρκετά καλά ενώ για τις κατηγορίες Covid και Lung opacity δείχνουν ότι το μοντέλο εμφανίζει δυσκολία στην μάθηση αυτών των κατηγοριών. Ειδικά για την περίπτωση του Covid-19 αυτό είναι αναμενόμενο αφού δεν είναι εύκολο να αναγνωριστεί η νόσος με μια απλή ακτινογραφία.

Από τα πειράματα φαίνεται ότι τουλάχιστον για μικρά datasets σαν αυτό που χρησιμοποιήσαμε μικρότερο rank και learning rate μπορούν να δώσουν σημαντικά καλύτερα αποτελέσματα.

5.5 Σύγκριση Μοντέλων και Προσεγγίσεων

Η πειραματική διαδικασία βασίστηκε στη συγκριτική αξιολόγηση των μοντέλων **Stable Diffusion v1.5** και **Stable Diffusion v2.1**, καθώς και στη μελέτη διαφορετικών προσεγγίσεων για τη βελτίωση και αξιολόγηση της δημιουργίας εικόνων από κείμενο.

Αρχικά, τα δύο μοντέλα συγκρίθηκαν ως προς την ικανότητά τους να παράγουν εικόνες που αντιστοιχούν σημασιολογικά στα δοθέντα prompts. Για τον σκοπό αυτό χρησιμοποιήθηκαν πέντε κατηγορίες prompts και εφαρμόστηκαν τέσσερις τεχνικές Prompt Engineering (context tokens, emphasis, repetition και αλλαγή της σειράς των λέξεων), ώστε να διερευνηθεί η επίδρασή τους στην απόδοση των μοντέλων.

Η σύγκριση πραγματοποιήθηκε τόσο με ποσοτικές όσο και με ποιοτικές μεθόδους αξιολόγησης. Η ποσοτική αξιολόγηση περιλάμβανε τον υπολογισμό των μετρικών **CLIP Score**, **Fréchet Inception Distance (FID)** και **Aesthetic Score**, οι οποίες εξετάζουν διαφορετικές διαστάσεις της ποιότητας των παραγόμενων εικόνων. Παράλληλα, πραγματοποιήθηκε ποιοτική αξιολόγηση μέσω οπτικής σύγκρισης των εικόνων, με στόχο την αναγνώριση χαρακτηριστικών και περιορισμών που δεν αποτυπώνονται πλήρως από τις αριθμητικές μετρικές.

Τέλος, διερευνήθηκε μία διαφορετική προσέγγιση αξιοποίησης των diffusion models μέσω **fine-tuning** του Stable Diffusion v1.5 με την τεχνική **Low-Rank Adaptation (LoRA)**. Η προσέγγιση αυτή εφαρμόστηκε σε εξειδικευμένο σύνολο ιατρικών εικόνων (COVID-19 Radiography Database), προκειμένου να αξιολογηθεί η δυνατότητα προσαρμογής ενός προεκπαιδευμένου μοντέλου σε εφαρμογές δημιουργίας συνθετικών ακτινογραφιών θώρακα.

Με τον τρόπο αυτό η πειραματική διαδικασία επιτρέπει τη σύγκριση τόσο των δύο μοντέλων Stable Diffusion όσο και των διαφορετικών προσεγγίσεων αξιολόγησης και προσαρμογής τους, παρέχοντας μια ολοκληρωμένη εικόνα των δυνατοτήτων και των περιορισμών τους.

Για τη συστηματική σύγκριση των δύο μοντέλων σχεδιάστηκαν τέσσερις διακριτές πειραματικές διαδικασίες, καθεμία από τις οποίες αξιολογεί διαφορετική πτυχή της απόδοσής τους. Συγκεκριμένα, εξετάστηκε η επίδραση των τεχνικών Prompt Engineering, η ποιότητα των παραγόμενων εικόνων με χρήση των μετρικών CLIP Score, FID και Aesthetic Score, καθώς και η δυνατότητα προσαρμογής του μοντέλου μέσω της τεχνικής LoRA. Ο Πίνακας 5.5.1 συνοψίζει τα πειράματα που πραγματοποιήθηκαν, τα μοντέλα που συγκρίνονται και τα κριτήρια αξιολόγησης που χρησιμοποιήθηκαν σε κάθε περίπτωση.

Πείραμα	Σύγκριση	Κριτήρια αξιολόγησης
Prompt Engineering	Stable Diffusion v1.5 – Stable Diffusion v2.1	CLIP Score, τυπική απόκλιση και ποιοτική αξιολόγηση
COCO Evaluation	Stable Diffusion v1.5 – Stable Diffusion v2.1	FID και CLIP Score
Aesthetic Evaluation (Στατιστική αξιολόγηση)	Stable Diffusion v1.5 – Stable Diffusion v2.1	Aesthetic Score, t-test και p-value
Fine-tuning (LoRA)	Stable Diffusion v1.5 πριν και μετά το LoRA	CLIP Score, FID και ποιοτική αξιολόγηση

Πίνακας 5.5.1. Σύνοψη των πειραμάτων και των κριτηρίων αξιολόγησης.

Συμπεράσματα και Μελλοντικές Κατευθύνσεις

6.1 Κύρια Συμπεράσματα

Η παρούσα εργασία είχε ως στόχο τη συγκριτική αξιολόγηση των προεκπαιδευμένων μοντέλων **Stable Diffusion v1.5** και **Stable Diffusion v2.1** στη δημιουργία εικόνων από κείμενο, καθώς και τη διερεύνηση της επίδρασης τεχνικών Prompt Engineering και της δυνατότητας προσαρμογής των μοντέλων μέσω της τεχνικής **Low-Rank Adaptation (LoRA)**.

Τα αποτελέσματα έδειξαν ότι τα δύο μοντέλα παρουσιάζουν συνολικά παρόμοια απόδοση ως προς τη σημασιολογική αντιστοίχιση μεταξύ εικόνας και κειμένου. Οι διαφορές που παρατηρήθηκαν στο **CLIP Score** ήταν γενικά μικρές, ενώ στις περισσότερες περιπτώσεις δεν προέκυψε σαφής υπεροχή κάποιου από τα δύο μοντέλα.

Η αξιολόγηση των τεχνικών Prompt Engineering έδειξε ότι η διατύπωση του prompt επηρεάζει την ποιότητα των παραγόμενων εικόνων. Ιδιαίτερα, η χρήση **context tokens** βελτίωσε την απόδοση του Stable Diffusion v2.1 σε σύνθετες περιγραφές, όπως σκηνές, περιβάλλοντα και χωρικές οδηγίες. Αντίθετα, οι τεχνικές **emphasis**, **repetition** και **αλλαγής της σειράς των λέξεων (order)** δεν οδήγησαν σε σημαντικές και σταθερές βελτιώσεις, γεγονός που υποδηλώνει ότι η αποτελεσματικότητά τους εξαρτάται από το περιεχόμενο του prompt.

Η αξιολόγηση με χρήση της βάσης δεδομένων **COCO** έδειξε ότι το **Stable Diffusion v1.5** πέτυχε χαμηλότερη τιμή **Fréchet Inception Distance (FID)** και ελαφρώς υψηλότερο **CLIP Score**, γεγονός που υποδηλώνει καλύτερη συνολική ομοιότητα με

τις πραγματικές εικόνες και ελαφρώς καλύτερη σημασιολογική αντιστοίχιση. Ωστόσο, οι διαφορές παρέμειναν περιορισμένες και δεν μεταβάλλουν ουσιαστικά τη συνολική εικόνα της απόδοσης των δύο μοντέλων.

Η αξιολόγηση της αισθητικής ποιότητας μέσω του **Aesthetic Score** έδειξε ότι και τα δύο μοντέλα παράγουν εικόνες μέτριας έως καλής αισθητικής ποιότητας, με το Stable Diffusion v1.5 να παρουσιάζει ελαφρώς υψηλότερη μέση βαθμολογία. Παρότι ο στατιστικός έλεγχος (t-test) έδειξε στατιστικά σημαντική διαφορά μεταξύ των δύο μοντέλων, η πραγματική διαφορά στη μέση αισθητική βαθμολογία ήταν μικρή, γεγονός που υποδηλώνει παρόμοια αισθητική απόδοση στην πράξη.

Η ποιοτική αξιολόγηση επιβεβαίωσε τα αποτελέσματα των ποσοτικών μετρικών. Τα δύο μοντέλα παρήγαγαν εικόνες υψηλής ποιότητας στις περισσότερες περιπτώσεις, ενώ εμφάνισαν κοινούς περιορισμούς, όπως δυσκολία στην πιστή απεικόνιση των ανθρώπινων χεριών και περιστασιακή παραγωγή εικόνων με χαρακτηριστικά εικονογράφησης αντί πλήρους φωτορεαλισμού.

Τέλος, το **fine-tuning** του Stable Diffusion v1.5 με την τεχνική **LoRA** απέδειξε ότι ένα προεκπαιδευμένο diffusion model μπορεί να προσαρμοστεί αποτελεσματικά σε εξειδικευμένο πεδίο, όπως η δημιουργία συνθετικών ακτινογραφιών θώρακα. Τα αποτελέσματα έδειξαν ότι ακόμη και με σχετικά περιορισμένο σύνολο δεδομένων είναι δυνατή η παραγωγή εικόνων που προσεγγίζουν ικανοποιητικά τα χαρακτηριστικά των πραγματικών ιατρικών εικόνων.

Συνολικά, η εργασία έδειξε ότι η αξιολόγηση των diffusion models δεν μπορεί να βασίζεται σε μία μόνο μετρική. Ο συνδυασμός **Prompt Engineering**, **CLIP Score**, **FID**, **Aesthetic Score** και **ποιοτικής αξιολόγησης** παρέχει μια πιο ολοκληρωμένη

και αξιόπιστη εικόνα των δυνατοτήτων και των περιορισμών των μοντέλων δημιουργίας εικόνας. Παράλληλα, αναδείχθηκε η ευελιξία των προεκπαιδευμένων diffusion models και η δυνατότητά τους να προσαρμόζονται σε εξειδικευμένες εφαρμογές μέσω τεχνικών αποδοτικής εκπαίδευσης όπως η LoRA.

6.2 Προτάσεις για Μελλοντική Έρευνα

Η παρούσα εργασία επικεντρώθηκε στη συγκριτική αξιολόγηση των μοντέλων Stable Diffusion v1.5 και Stable Diffusion v2.1, καθώς και στη διερεύνηση της επίδρασης τεχνικών Prompt Engineering και της προσαρμογής μέσω της τεχνικής LoRA. Παρότι τα αποτελέσματα ήταν ιδιαίτερα ικανοποιητικά, αναδεικνύονται αρκετές κατευθύνσεις για μελλοντική έρευνα.

Αρχικά, ενδιαφέρον παρουσιάζει η επέκταση της συγκριτικής αξιολόγησης σε νεότερα και πιο σύγχρονα μοντέλα δημιουργίας εικόνων, όπως το **Stable Diffusion 3**, το **FLUX**, το **Imagen** και άλλα σύγχρονα diffusion models, προκειμένου να διερευνηθεί η εξέλιξη της ποιότητας των παραγόμενων εικόνων και η αποτελεσματικότητα των τεχνικών Prompt Engineering.

Επιπλέον, θα μπορούσε να εξεταστεί η επίδραση περισσότερων τεχνικών Prompt Engineering, όπως η χρήση **negative prompts**, η αξιοποίηση παραδειγμάτων (**few-shot prompting**), η αλυσιδωτή διαμόρφωση προτροπών (**prompt chaining**) ή η αυτόματη βελτιστοποίηση prompts με χρήση μεγάλων γλωσσικών μοντέλων (Large Language Models).

Μία ακόμη ενδιαφέρουσα κατεύθυνση αφορά τη χρήση επιπρόσθετων μετρικών αξιολόγησης, καθώς και τη διεξαγωγή εκτενέστερων μελετών με ανθρώπινους αξιολογητές. Ο συνδυασμός αντικειμενικών μετρικών και ανθρώπινης αξιολόγησης μπορεί να οδηγήσει σε πιο ολοκληρωμένη αποτίμηση της ποιότητας των παραγόμενων εικόνων.

Στον τομέα της ιατρικής απεικόνισης, η μελλοντική έρευνα μπορεί να επεκταθεί στη χρήση μεγαλύτερων και πιο ποικίλων συνόλων δεδομένων, καθώς και στην εφαρμογή τεχνικών fine-tuning σε διαφορετικές ιατρικές απεικονιστικές εξετάσεις, όπως αξονικές ή μαγνητικές τομογραφίες. Επιπλέον, μπορεί να διερευνηθεί η αξιοποίηση των συνθετικών εικόνων ως μέσο ενίσχυσης (data augmentation) για την εκπαίδευση μοντέλων ταξινόμησης και διάγνωσης.

Επίσης, ιδιαίτερο ενδιαφέρον παρουσιάζει η μελέτη της υπολογιστικής αποδοτικότητας των diffusion models, τόσο ως προς τον χρόνο παραγωγής εικόνων όσο και ως προς τις απαιτήσεις σε υπολογιστικούς πόρους, καθώς και η αξιολόγηση της δυνατότητας ανάπτυξής τους σε περιβάλλοντα περιορισμένων υπολογιστικών δυνατοτήτων.

Τέλος, μία σημαντική κατεύθυνση για μελλοντική έρευνα είναι η ανάπτυξη ενός ολοκληρωμένου πλαισίου αξιολόγησης diffusion models, το οποίο θα συνδυάζει ποσοτικές μετρικές, στατιστικούς ελέγχους, ποιοτική αξιολόγηση και τεχνικές Prompt Engineering. Ένα τέτοιο πλαίσιο θα μπορούσε να αποτελέσει σημείο αναφοράς για τη συστηματική και αντικειμενική σύγκριση διαφορετικών μοντέλων δημιουργίας εικόνας και να συμβάλει στη βελτίωση της αξιοπιστίας των σχετικών ερευνητικών αποτελεσμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. “Diffusion models: A comprehensive survey of methods and applications”, 2024. URL: <https://arxiv.org/abs/2209.00796>
2. Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. “Deep unsupervised learning using nonequilibrium thermodynamics”, 2015. URL: <https://arxiv.org/abs/1503-03585>
3. Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”, 2020. URL: <https://arxiv.org/abs/2006.11239>
4. Ajay Bandi, Pydi Venkata Satya Ramesh Adapa, and Yudu Eswar Vinay Pratap Kumar Kuchi. The power of generative ai: A review of requirements, models, input-output formats, evaluation metrics, and challenges. *Future Internet*, 15(8), 2023. URL: <https://www.mdpi.com/1999-5903/15/8/260>
5. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. “Attention is all you need”, 2023. URL: <https://arxiv.org/abs/1706.03762>
6. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. “Neural machine translation by jointly learning to align and translate”, 2016. URL: <https://arxiv.org/abs/1409.0473>
7. Ian Goodfellow, Yoshua Bengio, and Aaron Courville. “*Deep Learning*. MIT Press”, 2016. URL: <https://www.deeplearningbook.org/>.

8. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. “Deep residual learning for image recognition”, 2015. URL: <https://arxiv.org/abs/1512.03385>
9. Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. “Layer normalization”, 2016. URL: <https://arxiv.org/abs/1607.06450>
10. Bommasani et al. “On the opportunities and risks of foundation models”, 2022. URL: <https://arxiv.org/abs/2108.07258>
11. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. “High resolution image synthesis with latent diffusion models”, 2022. URL: <https://arxiv.org/abs/2112.10752>
12. Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. “Zero-shot text-to-image generation”, 2021. URL: <https://arxiv.org/abs/2102.12092>
13. B. Brown et al. “ Language models are few-shot learners”, 2020. URL: <https://arxiv.org/abs/2005.14165>
14. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. “Hierarchical text - conditional image generation with clip latents”, 2022. URL: <https://arxiv.org/abs/2204.06125>
15. OpenAI. (2023). *DALL-E 3 system card*. <https://openai.com/index/dall-e-3-system-card/>
16. Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. “Improved techniques for training gans” , 2016. URL: <https://arxiv.org/abs/1606.03498>
17. Shane Barratt and Rishi Sharma. “A note on the inception score”, 2018. URL: <https://arxiv.org/abs/1801.01973>

18. Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. “Gans trained by a two time-scale update rule converge to a local nash equilibrium” , 2018. URL: <https://arxiv.org/abs/1706.08500>
19. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. “Learning transferable visual models from natural language supervision” , 2021. URL: <https://arxiv.org/abs/2103.00020>
20. Dayma, B., Patil, S., Cuenca, P., Saifullah, K., Abraham, T., Lê, P., Melas, L., & Ghosh, R. (2021). *DALL·E Mini*. Hugging Face. <https://huggingface.co/dalle-mini/dalle-mini>
21. Federated learning and conditional diffusion model for privacy-preserving short-term voltage stability assessment in multi-region power systems. Chong Ma, Hanjie Wang, Jiting Cao. <https://doi.org/10.1016/j.egyr.2025.05.087>
22. Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, Mohammad Norouzi. “Palette: Image-to-Image Diffusion Models”, 2022. URL: <https://arxiv.org/abs/2111.05826>
23. Rahman, A., Al-Mamoon, F., Saquib, M. N., Bahar, F. B., & Fabia, M. T. (2024). *Implementation of diffusion model in realistic face generation* [Πτυχιακή εργασία, BRAC University]. BRAC University Institutional Repository. <https://dspace.bracu.ac.bd/xmlui/handle/10361/22780>
24. Sagar Saxena, Mohammad Nayeem Teli. “Comparison and Analysis of Image-to-Image Generative Adversarial Networks: A Survey”, 2022. URL: <https://arxiv.org/abs/2112.12625>

25. Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, In So Kweon, Junmo Kim. “Text-to-image Diffusion Models in Generative AI: A Survey” , 2024. URL: <https://arxiv.org/abs/2303.07909>
26. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, Alexander M. Rush. “HuggingFace's Transformers: State-of-the-art Natural Language Processing”, 2020. URL: <https://arxiv.org/abs/1910.03771>
27. von Platen, P., Patil, S., Lozhkov, A., Cuenca, P., Lambert, N., Rasul, K., Davaadorj, M., Nair, D., Paul, S., Berman, W., Xu, Y., Liu, S., & Wolf, T. (2022). *Diffusers: State-of-the-art diffusion models* [Computer software]. GitHub. <https://github.com/huggingface/diffusers>
28. Gugger, S., Debut, L., Wolf, T., Schmid, P., Mueller, Z., Mangrulkar, S., & Bossan, B. (2022). *Accelerate: Training and inference at scale made simple, efficient and adaptable* [Computer software]. GitHub. <https://github.com/huggingface/accelerate>
29. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever. “Learning Transferable Visual Models From Natural Language Supervision”. 2021. URL: <https://arxiv.org/abs/2103.00020>
30. CompVis. (2022). *Stable diffusion: A latent text-to-image diffusion model* [Αποθετήριο κώδικα]. GitHub. <https://github.com/CompVis/stable-diffusion>

31. Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical textconditional image generation with clip latents. 2022. URL: <https://arxiv.org/abs/2204.06125>
32. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, Lawrence Zitnick, and Piotr Dollár. “Microsoft COCO: Common Objects in Context”, 2015. URL: <https://arxiv.org/abs/1405.0312>
33. Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, Mohammad Norouzi. “Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding”, 2022. URL: <https://arxiv.org/abs/2205.11487>
34. Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer. “High-Resolution Image Synthesis with Latent Diffusion Models”, 2021. URL: <https://arxiv.org/abs/2112.10752>
35. Alex Nichol, Prafulla Dhariwal. “Improved Denoising Diffusion Probabilistic Models”, 2021. URL: <https://arxiv.org/abs/2102.09672>
36. Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, Jenia Jitsev. “LAION-5B: An open large-scale dataset for training next generation image-text models”, 2022. URL: <https://arxiv.org/abs/2210.08402>

37. Lopez, J. (2022, December 7). *Stable Diffusion v2.1 and DreamStudio updates 7-Dec 22*. Stability AI. <https://stability.ai/news-updates/stablediffusion2-1-release7-dec-2022>
38. Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, Robin Rombach. “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis”, 2023. URL: <https://arxiv.org/abs/2307.01952>
39. Imagen 3 Team. (2024). Imagen 3 [Technical report]. Google DeepMind. https://storage.googleapis.com/deepmind-media/imagen/imagen_3_tech_report_update_dec2024_v3.pdf
40. fal. (n.d.). AuraFlow v0.3 [Model card]. Hugging Face. <https://huggingface.co/fal/AuraFlow-v0.3>
41. Midjourney. (n.d.). Version. Midjourney Documentation. <https://docs.midjourney.com/hc/en-us/articles/32199405667853-Version>
42. Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, Luke Smith. “FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space”, 2025. URL: <https://arxiv.org/abs/2506.15742>
43. Ojonugwa Oluwafemi Ejiga Peter, Md Mahmudur Rahman, Fahmi Khalifa “Advancing AI-Powered Medical Image Synthesis: Insights from MedVQA-GI Challenge Using CLIP, Fine-Tuned Stable Diffusion, and Dream-Booth + LoRA” , 2025. URL: <https://arxiv.org/abs/2502.20667>

ΠΑΡΑΡΤΗΜΑΤΑ

Παράρτημα Α' : Κατηγορίες prompt και Clip score

BASIKES- ΚΑΤΗΓΟΡΙΕΣ ΡΟ	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair	0,310	0,003
2	A Chinese ceramic jar	0,336	0,006
3	An old wall clock	0,313	0,019
4	A hardcover book	0,297	0,018
5	A small house	0,276	0,010
	ΜΕΣΟΣ ΟΡΟΣ	0,306	0,011

Ζώα			
1	A brown dog running on grass	0,316	0,002
2	A white cat sitting on a sofa	0,358	0,009
3	A small yellow bird sitting on a tree branch	0,339	0,005
4	A horse standing in a field	0,318	0,003
5	A fish swimming underwater	0,284	0,009
	ΜΕΣΟΣ ΟΡΟΣ	0,323	0,006

Άνθρωποι			
1	A woman wearing a black coat, standing on a city street	0,309	0,022
2	A man smiling while holding a cup of coffee	0,324	0,024
3	A child playing with a ball in a park	0,332	0,007
4	A person sitting at a desk using a laptop	0,258	0,011
5	A man walking down a street at night	0,322	0,013
	ΜΕΣΟΣ ΟΡΟΣ	0,309	0,015

Σκηνές Περιβάλλον			
1	A quiet beach at sunset	0,279	0,005
2	A snowy mountain landscape with pine trees	0,309	0,005
3	A city street during heavy rain	0,310	0,010
4	A small village surrounded by hills	0,292	0,009
5	A forest path in autumn	0,308	0,009
	ΜΕΣΟΣ ΟΡΟΣ	0,300	0,008

Χωρικές			
----------------	--	--	--

οδηγίες			
1	A cup on a table with a book beside it	0,269	0,019
2	A car parked in front of a small house	0,296	0,018
3	A lamp standing next to a sofa	0,317	0,005
4	A chair under a window	0,299	0,011
5	A plate with food in the center of a table	0,290	0,014
	ΜΕΣΟΣ ΟΡΟΣ	0,294	0,014

P1-Context	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair, realistic photography, natural lighting	0,315	0,013
2	A Chinese ceramic jar, realistic photography, natural lighting	0,324	0,014
3	An old wall clock, realistic photography, natural lighting	0,317	0,018
4	A hardcover book, realistic photography, natural lighting	0,291	0,017
5	A small house, realistic photography, natural lighting	0,279	0,018
	ΜΕΣΟΣ ΟΡΟΣ	0,305	0,016

Ζώα			
1	A brown dog running on grass, outdoor photography, natural light	0,330	0,001
2	A white cat sitting on a sofa, indoor photography, soft lighting	0,335	0,011
3	A small yellow bird sitting on a tree branch, wildlife photography	0,343	0,007
4	A horse standing in a field, natural landscape	0,319	0,010
5	A fish swimming underwater, underwater photography	0,293	0,010
	ΜΕΣΟΣ ΟΡΟΣ	0,324	0,008
Άνθρωποι			
1	A woman wearing a black coat, standing on a city street, realistic photo	0,317	0,026
2	A man smiling while holding a cup of coffee, natural lighting	0,346	0,017
3	A child playing with a ball in a park, outdoor photography	0,353	0,016
4	A person sitting at a desk using a laptop, indoor scene	0,259	0,007
5	A man walking down a street at night, low light photography	0,327	0,011

	ΜΕΣΟΣ ΟΡΟΣ	0,320	0,016
Σκηνές Περιβάλλον			
1	A quiet beach at sunset, wide landscape	0,286	0,007
2	A snowy mountain landscape with pine trees, wide landscape	0,296	0,012
3	A city street during heavy rain, urban photography	0,323	0,011
4	A small village surrounded by hills, landscape photography	0,308	0,007
5	A forest path in autumn, natural colors	0,315	0,005
	ΜΕΣΟΣ ΟΡΟΣ	0,306	0,009
Χωρικές οδηγίες			
1	A cup on a table with a book beside it, indoor scene, realistic lighting	0,319	0,027
2	A car parked in front of a small house, outdoor scene, daylight	0,301	0,010
3	A lamp standing next to a sofa, indoor living room, realistic lighting	0,340	0,009
4	A chair under a window, indoor scene, natural light	0,335	0,019
5	A plate with food in the center of a table, indoor scene, realistic photography	0,286	0,013
	ΜΕΣΟΣ ΟΡΟΣ	0,316	0,016

P2-Emphasis	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	(A wooden chair:1.4)	0,306	0,004
2	(A Chinese ceramic jar:1.4)	0,334	0,012
3	(An old wall clock:1.4)	0,315	0,012
4	(A hardcover book:1.4)	0,264	0,013
5	(A small house:1.4)	0,268	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,297	0,010

Ζώα			
1	(A brown dog running on grass:1.4)	0,308	0,008
2	(A white cat sitting on a sofa:1.4)	0,324	0,014
3	(A small yellow bird sitting on a tree branch:1.4)	0,339	0,008
4	(A horse standing in a field:1.4)	0,315	0,006
5	(A fish swimming underwater:1.4)	0,285	0,009
	ΜΕΣΟΣ ΟΡΟΣ	0,314	0,009
Ανθρωποι			

1	(A woman wearing a black coat, standing on a city street:1.4)	0,307	0,017
2	(A man smiling while holding a cup of coffee:1.4)	0,333	0,010
3	(A child playing with a ball in a park:1.4)	0,313	0,012
4	(A person sitting at a desk using a laptop:1.4)	0,268	0,013
5	(A man walking down a street at night:1.4)	0,318	0,008
	ΜΕΣΟΣ ΟΡΟΣ	0,308	0,012
Σκηνές Περιβάλλον			
1	(A quiet beach at sunset:1.4)	0,283	0,006
2	(A snowy mountain landscape with pine trees:1.4)	0,328	0,011
3	(A city street during heavy rain:1.4)	0,321	0,008
4	(A small village surrounded by hills:1.4)	0,279	0,008
5	(A forest path in autumn:1.4)	0,298	0,004
	ΜΕΣΟΣ ΟΡΟΣ	0,302	0,007
Χωρικές οδηγίες			
1	(A cup on a table with a book beside it:1.4)	0,280	0,020
2	(A car parked in front of a small house:1.4)	0,277	0,023
3	(A lamp standing next to a sofa:1.4)	0,314	0,002
4	(A chair under a window:1.4)	0,319	0,004
5	(A plate with food in the center of a table:1.4)	0,295	0,017
	ΜΕΣΟΣ ΟΡΟΣ	0,297	0,013

P3- Repetition	PROMPTS	CLIP SCOR E	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair, wooden chair	0,314	0,008
2	A Chinese ceramic jar, Chinese ceramic jar	0,331	0,012
3	An old wall clock, old wall clock	0,311	0,009
4	A hardcover book, hardcover book	0,267	0,019
5	A small house, small house	0,267	0,002
	ΜΕΣΟΣ ΟΡΟΣ	0,298	0,010

Ζώα			
1	A brown dog running on grass, brown dog running	0,315	0,009
2	A white cat sitting on a sofa, white cat sitting	0,336	0,009

3	A small yellow bird sitting on a tree branch, yellow bird	0,346	0,010
4	A horse standing in a field, horse standing	0,308	0,004
5	A fish swimming underwater, fish swimming	0,287	0,004
	ΜΕΣΟΣ ΟΡΟΣ	0,318	0,007
Ανθρωποι			
1	A woman wearing a black coat, standing on a city street, woman wearing a black coat	0,304	0,017
2	A man smiling while holding a cup of coffee, man smiling	0,343	0,009
3	A child playing with a ball in a park, child playing	0,325	0,009
4	A person sitting at a desk using a laptop, person using a laptop	0,271	0,011
5	A man walking down a street at night, man walking	0,320	0,005
	ΜΕΣΟΣ ΟΡΟΣ	0,313	0,010
Σκηνές Περιβάλλον			
1	A quiet beach at sunset, quiet beach	0,286	0,007
2	A snowy mountain landscape with pine trees, snowy mountains	0,306	0,003
3	A city street during heavy rain, heavy rain	0,312	0,008
4	A small village surrounded by hills, small village	0,297	0,011
5	A forest path in autumn, forest path	0,314	0,005
	ΜΕΣΟΣ ΟΡΟΣ	0,303	0,007
Χωρικές οδηγίες			
1	A cup on a table with a book beside it, cup on table, book beside	0,273	0,016
2	A car parked in front of a small house, car in front of house	0,309	0,011
3	A lamp standing next to a sofa, lamp next to sofa	0,317	0,009
4	A chair under a window, chair under window	0,297	0,012
5	A plate with food in the center of a table, plate in center	0,294	0,013
	ΜΕΣΟΣ ΟΡΟΣ	0,298	0,012

P4 - Order	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	Realistic photography, natural lighting, a wooden chair	0,293	0,034
2	Realistic photography, natural lighting, a Chinese	0,333	0,014

	ceramic jar		
3	Realistic photography, natural lighting, an old wall clock	0,295	0,009
4	Realistic photography, natural lighting, a hardcover book	0,294	0,022
5	Realistic photography, natural lighting, a small house	0,291	0,010
	ΜΕΣΟΣ ΟΡΟΣ	0,301	0,018

Ζώα			
1	Outdoor photography, natural light, a brown dog running on grass	0,322	0,006
2	Indoor photography, soft lighting, a white cat sitting on a sofa	0,351	0,007
3	Wildlife photography, a small yellow bird sitting on a tree branch	0,349	0,003
4	Natural landscape, a horse standing in a field	0,316	0,012
5	Underwater photography, a fish swimming underwater	0,295	0,005
	ΜΕΣΟΣ ΟΡΟΣ	0,327	0,007

Άνθρωποι			
1	Realistic photo, a woman wearing a black coat, standing on a city street	0,340	0,020
2	Natural lighting, a man smiling while holding a cup of coffee	0,347	0,014
3	Outdoor photography, a child playing with a ball in a park	0,355	0,007
4	Indoor scene, a person sitting at a desk using a laptop	0,260	0,023
5	Low light photography, a man walking down a street at night	0,335	0,011
	ΜΕΣΟΣ ΟΡΟΣ	0,327	0,015

Σκηνές Περιβάλλον			
1	Wide landscape, a quiet beach at sunset	0,298	0,009
2	Wide landscape, a snowy mountain landscape with pine trees	0,307	0,027
3	Urban photography, a city street during heavy rain	0,318	0,008
4	Landscape photography, a small village surrounded by hills	0,306	0,008
5	Natural colors, a forest path in autumn	0,322	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,310	0,012

Χωρικές οδηγίες			
1	Indoor scene, realistic lighting, a cup on a table with a book beside it	0,315	0,012

2	Outdoor scene, daylight, a car parked in front of a small house	0,301	0,026
3	Indoor living room, realistic lighting, a lamp standing next to a sofa	0,311	0,009
4	Indoor scene, natural light, a chair under a window	0,331	0,011
5	Indoor scene, realistic photography, a plate with food in the center of a table	0,291	0,014
ΜΕΣΟΣ ΟΡΟΣ		0,310	0,014

Stable diffusion 2.1

BASIKES- ΚΑΤΗΓΟΡΙΕΣ ΡΟ	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair	0,312	0,009
2	A Chinese ceramic jar	0,332	0,010
3	An old wall clock	0,310	0,010
4	A hardcover book	0,257	0,009
5	A small house	0,281	0,004
ΜΕΣΟΣ ΟΡΟΣ		0,298	0,008

Ζώα			
1	A brown dog running on grass	0,329	0,008
2	A white cat sitting on a sofa	0,351	0,006
3	A small yellow bird sitting on a tree branch	0,345	0,010
4	A horse standing in a field	0,304	0,008
5	A fish swimming underwater	0,285	0,008
ΜΕΣΟΣ ΟΡΟΣ		0,323	0,008
Άνθρωποι			
1	A woman wearing a black coat, standing on a city street	0,310	0,014
2	A man smiling while holding a cup of coffee	0,339	0,009
3	A child playing with a ball in a park	0,328	0,009
4	A person sitting at a desk using a laptop	0,253	0,006
5	A man walking down a street at night	0,325	0,006
ΜΕΣΟΣ ΟΡΟΣ		0,311	0,009
Σκηνές Περιβάλλον			
1	A quiet beach at sunset	0,295	0,009

2	A snowy mountain landscape with pine trees	0,297	0,020
3	A city street during heavy rain	0,300	0,006
4	A small village surrounded by hills	0,292	0,007
5	A forest path in autumn	0,316	0,006
	ΜΕΣΟΣ ΟΡΟΣ	0,300	0,010
Χωρικές οδηγίες			
1	A cup on a table with a book beside it	0,288	0,011
2	A car parked in front of a small house	0,285	0,023
3	A lamp standing next to a sofa	0,318	0,019
4	A chair under a window	0,310	0,007
5	A plate with food in the center of a table	0,294	0,015
	ΜΕΣΟΣ ΟΡΟΣ	0,299	0,015

P1-Context	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair, realistic photography, natural lighting	0,323	0,007
2	A Chinese ceramic jar, realistic photography, natural lighting	0,325	0,004
3	An old wall clock, realistic photography, natural lighting	0,316	0,015
4	A hardcover book, realistic photography, natural lighting	0,314	0,008
5	A small house, realistic photography, natural lighting	0,307	0,008
	ΜΕΣΟΣ ΟΡΟΣ	0,317	0,008

Ζώα			
1	A brown dog running on grass, outdoor photography, natural light	0,347	0,005
2	A white cat sitting on a sofa, indoor photography, soft lighting	0,359	0,015
3	A small yellow bird sitting on a tree branch, wildlife photography	0,349	0,009
4	A horse standing in a field, natural landscape	0,291	0,013
5	A fish swimming underwater, underwater photography	0,292	0,009
	ΜΕΣΟΣ ΟΡΟΣ	0,327	0,010
Άνθρωποι			
1	A woman wearing a black coat, standing on a city street, realistic photo	0,317	0,019
2	A man smiling while holding a cup of coffee,	0,341	0,009

	natural lighting		
3	A child playing with a ball in a park, outdoor photography	0,349	0,008
4	A person sitting at a desk using a laptop, indoor scene	0,250	0,002
5	A man walking down a street at night, low light photography	0,327	0,008
	ΜΕΣΟΣ ΟΡΟΣ	0,317	0,009
Σκηνές Περιβάλλον			
1	A quiet beach at sunset, wide landscape	0,330	0,020
2	A snowy mountain landscape with pine trees, wide landscape	0,338	0,009
3	A city street during heavy rain, urban photography	0,314	0,009
4	A small village surrounded by hills, landscape photography	0,309	0,006
5	A forest path in autumn, natural colors	0,323	0,010
	ΜΕΣΟΣ ΟΡΟΣ	0,323	0,011
Χωρικές οδηγίες			
1	A cup on a table with a book beside it, indoor scene, realistic lighting	0,302	0,018
2	A car parked in front of a small house, outdoor scene, daylight	0,342	0,024
3	A lamp standing next to a sofa, indoor living room, realistic lighting	0,332	0,016
4	A chair under a window, indoor scene, natural light	0,331	0,007
5	A plate with food in the center of a table, indoor scene, realistic photography	0,322	0,012
	ΜΕΣΟΣ ΟΡΟΣ	0,326	0,015

P2-Emphasis	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	(A wooden chair:1.4)	0,309	0,005
2	(A Chinese ceramic jar:1.4)	0,318	0,006
3	(An old wall clock:1.4)	0,309	0,007
4	(A hardcover book:1.4)	0,241	0,016
5	(A small house:1.4)	0,300	0,013
	ΜΕΣΟΣ ΟΡΟΣ	0,295	0,009

Ζώα			
1	(A brown dog running on grass:1.4)	0,324	0,003
2	(A white cat sitting on a sofa:1.4)	0,314	0,008

3	(A small yellow bird sitting on a tree branch:1.4)	0,347	0,007
4	(A horse standing in a field:1.4)	0,308	0,007
5	(A fish swimming underwater:1.4)	0,285	0,009
	ΜΕΣΟΣ ΟΡΟΣ	0,315	0,007
Άνθρωποι			
1	(A woman wearing a black coat, standing on a city street:1.4)	0,332	0,007
2	(A man smiling while holding a cup of coffee:1.4)	0,326	0,022
3	(A child playing with a ball in a park:1.4)	0,336	0,010
4	(A person sitting at a desk using a laptop:1.4)	0,262	0,006
5	(A man walking down a street at night:1.4)	0,314	0,011
	ΜΕΣΟΣ ΟΡΟΣ	0,314	0,011
Σκηνές Περιβάλλον			
1	(A quiet beach at sunset:1.4)	0,304	0,010
2	(A snowy mountain landscape with pine trees:1.4)	0,339	0,014
3	(A city street during heavy rain:1.4)	0,311	0,009
4	(A small village surrounded by hills:1.4)	0,290	0,011
5	(A forest path in autumn:1.4)	0,305	0,013
	ΜΕΣΟΣ ΟΡΟΣ	0,310	0,011
Χωρικές οδηγίες			
1	(A cup on a table with a book beside it:1.4)	0,295	0,014
2	(A car parked in front of a small house:1.4)	0,323	0,007
3	(A lamp standing next to a sofa:1.4)	0,301	0,010
4	(A chair under a window:1.4)	0,314	0,013
5	(A plate with food in the center of a table:1.4)	0,314	0,008
	ΜΕΣΟΣ ΟΡΟΣ	0,310	0,010

P3- Repetition	PROMPTS	CLIP SCOR E	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	A wooden chair, wooden chair	0,320	0,004
2	A Chinese ceramic jar, Chinese ceramic jar	0,324	0,005
3	An old wall clock, old wall clock	0,313	0,017
4	A hardcover book, hardcover book	0,259	0,003
5	A small house, small house	0,281	0,012

	ΜΕΣΟΣ ΟΡΟΣ	0,299	0,008
Ζώα			
1	A brown dog running on grass, brown dog running	0,323	0,005
2	A white cat sitting on a sofa, white cat sitting	0,336	0,008
3	A small yellow bird sitting on a tree branch, yellow bird	0,352	0,018
4	A horse standing in a field, horse standing	0,294	0,007
5	A fish swimming underwater, fish swimming	0,283	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,317	0,009
Άνθρωποι			
1	A woman wearing a black coat, standing on a city street, woman wearing a black coat	0,293	0,015
2	A man smiling while holding a cup of coffee, man smiling	0,337	0,011
3	A child playing with a ball in a park, child playing	0,322	0,010
4	A person sitting at a desk using a laptop, person using a laptop	0,256	0,004
5	A man walking down a street at night, man walking	0,323	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,306	0,009
Σκηνές Περιβάλλον			
1	A quiet beach at sunset, quiet beach	0,289	0,007
2	A snowy mountain landscape with pine trees, snowy mountains	0,296	0,025
3	A city street during heavy rain, heavy rain	0,298	0,003
4	A small village surrounded by hills, small village	0,310	0,009
5	A forest path in autumn, forest path	0,329	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,304	0,010
Χωρικές οδηγίες			
1	A cup on a table with a book beside it, cup on table, book beside	0,300	0,017
2	A car parked in front of a small house, car in front of house	0,319	0,007
3	A lamp standing next to a sofa, lamp next to sofa	0,298	0,019
4	A chair under a window, chair under window	0,311	0,002
5	A plate with food in the center of a table, plate in center	0,300	0,008
	ΜΕΣΟΣ ΟΡΟΣ	0,306	0,010

P4 - Order	PROMPTS	CLIP SCORE	ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ
Απλά αντικείμενα			
1	Realistic photography, natural lighting, a wooden chair	0,317	0,011
2	Realistic photography, natural lighting, a Chinese ceramic jar	0,349	0,005
3	Realistic photography, natural lighting, an old wall clock	0,323	0,006
4	Realistic photography, natural lighting, a hardcover book	0,290	0,028
5	Realistic photography, natural lighting, a small house	0,296	0,018
	ΜΕΣΟΣ ΟΡΟΣ	0,315	0,013

Ζώα			
1	Outdoor photography, natural light, a brown dog running on grass	0,341	0,007
2	Indoor photography, soft lighting, a white cat sitting on a sofa	0,352	0,013
3	Wildlife photography, a small yellow bird sitting on a tree branch	0,351	0,011
4	Natural landscape, a horse standing in a field	0,296	0,008
5	Underwater photography, a fish swimming underwater	0,293	0,007
	ΜΕΣΟΣ ΟΡΟΣ	0,327	0,009
Άνθρωποι			
1	Realistic photo, a woman wearing a black coat, standing on a city street	0,324	0,024
2	Natural lighting, a man smiling while holding a cup of coffee	0,332	0,014
3	Outdoor photography, a child playing with a ball in a park	0,352	0,006
4	Indoor scene, a person sitting at a desk using a laptop	0,246	0,008
5	Low light photography, a man walking down a street at night	0,333	0,006
	ΜΕΣΟΣ ΟΡΟΣ	0,317	0,012
Σκηνές Περιβάλλον			
1	Wide landscape, a quiet beach at sunset	0,296	0,008
2	Wide landscape, a snowy mountain landscape with pine trees	0,326	0,005
3	Urban photography, a city street during heavy rain	0,313	0,005
4	Landscape photography, a small village surrounded by hills	0,310	0,002

5	Natural colors, a forest path in autumn	0,333	0,004
	ΜΕΣΟΣ ΟΡΟΣ	0,316	0,005
Χωρικές οδηγίες			
1	Indoor scene, realistic lighting, a cup on a table with a book beside it	0,309	0,035
2	Outdoor scene, daylight, a car parked in front of a small house	0,315	0,013
3	Indoor living room, realistic lighting, a lamp standing next to a sofa	0,315	0,004
4	Indoor scene, natural light, a chair under a window	0,320	0,006
5	Indoor scene, realistic photography, a plate with food in the center of a table	0,282	0,004
	ΜΕΣΟΣ ΟΡΟΣ	0,308	0,012

Παράρτημα Β' : Scripts – prompt engineering

Κώδικας SD1.5

```
!pip install -q diffusers transformers accelerate torch
torchvision pillow tqdm
!pip install -q git+https://github.com/openai/CLIP.git

from huggingface_hub import login
login(token="your token")

from huggingface_hub import whoami
whoami()

from diffusers import StableDiffusionPipeline
import torch

#torch.cuda.empty_cache()
device = "cuda" if torch.cuda.is_available() else "cpu"
model_id = "sd-legacy/stable-diffusion-v1-5"

pipe = StableDiffusionPipeline.from_pretrained(
    model_id,
    torch_dtype=torch.float16
)

pipe = pipe.to(device)

import clip
from PIL import Image

clip_model, preprocess = clip.load("ViT-B/32", device=device)
clip_model.eval()

import numpy as np
from tqdm import tqdm

prompts = [

]

results = {}
images_dict = {} # dictionary για να κρατάμε εικόνες

for p in prompts:
```

```
scores = []
imgs = []
for i in range(4):
    gen = torch.Generator(device).manual_seed(100 + i)

    img = pipe(
        p,
        num_inference_steps=50, #
        guidance_scale=7.5,
        generator=gen
    ).images[0]

    imgs.append(img) # κρατάμε την εικόνα

# CLIP score
img_input = preprocess(img).unsqueeze(0).to(device)
txt_input = clip.tokenize([p]).to(device)

with torch.no_grad():
    img_feat = clip_model.encode_image(img_input)
    txt_feat = clip_model.encode_text(txt_input)

    img_feat /= img_feat.norm(dim=-1, keepdim=True)
    txt_feat /= txt_feat.norm(dim=-1, keepdim=True)

    score = (img_feat @ txt_feat.T).item()
    scores.append(score)

results[p] = {
    "mean_clip": np.mean(scores),
    "std_clip": np.std(scores)
}
images_dict[p] = imgs # αποθηκεύουμε όλες τις εικόνες για το
prompt

import matplotlib.pyplot as plt

# Εμφάνιση εικόνων σε grid ανά prompt
def show_images_grid(images_dict, results_dict, n_cols=4):
    for prompt, imgs in images_dict.items():
        n_images = len(imgs)
        n_rows = 1 # ορίζοντο grid
        fig, axes = plt.subplots(n_rows, n_cols,
            figsize=(4*n_cols, 4*n_rows))

        # Αν έχουμε μόνο μια εικόνα, axes δεν είναι λίστα
        if n_images == 1:
```

```
axes = [axes]

for i, ax in enumerate(axes):
    if i < n_images:
        ax.imshow(imgs[i])
        ax.axis("off")
    else:
        ax.axis("off") # κενά axes

mean_clip = results_dict[prompt]["mean_clip"]
std_clip = results_dict[prompt]["std_clip"]
plt.suptitle(f"{prompt}\nCLIP score: {mean_clip:.3f} ± {std_clip:.3f}", fontsize=14)
plt.show()

# Κλήση της συνάρτησης
show_images_grid(images_dict, results)
```

Κώδικας SD2.1

```
!pip install -q diffusers transformers accelerate torch
torchvision pillow tqdm
!pip install -q git+https://github.com/openai/CLIP.git

from huggingface_hub import login
login(token=" your token ")

from huggingface_hub import whoami
whoami()

import torch
from diffusers import StableDiffusionPipeline

device = "cuda" if torch.cuda.is_available() else "cpu"

model_id = "sd2-community/stable-diffusion-2-1"

pipe = StableDiffusionPipeline.from_pretrained(
    model_id,
    torch_dtype=torch.float16
).to(device)

pipe.enable_attention_slicing()
```

```
import clip
from PIL import Image

clip_model, preprocess = clip.load("ViT-B/32", device=device)
clip_model.eval()

import numpy as np
from tqdm import tqdm

prompts = [ ]

results = {}
images_dict = {} # dictionary για να κρατάμε εικόνες

for p in prompts:
    scores = []
    imgs = []
    for i in range(4):
        gen = torch.Generator(device).manual_seed(100 + i)

        img = pipe(
            p,
            num_inference_steps=50,
            guidance_scale=7.5,
            generator=gen
        ).images[0]

        imgs.append(img) # κρατάμε την εικόνα

    # CLIP score
    img_input = preprocess(img).unsqueeze(0).to(device)
    txt_input = clip.tokenize([p]).to(device)

    with torch.no_grad():
        img_feat = clip_model.encode_image(img_input)
        txt_feat = clip_model.encode_text(txt_input)

        img_feat /= img_feat.norm(dim=-1, keepdim=True)
        txt_feat /= txt_feat.norm(dim=-1, keepdim=True)

        score = (img_feat @ txt_feat.T).item()
        scores.append(score)

    results[p] = {
        "mean_clip": np.mean(scores),
        "std_clip": np.std(scores)
    }
```

```
images_dict[p] = imgs # αποθηκεύουμε όλες τις εικόνες για το
prompt

import matplotlib.pyplot as plt

# Εμφάνιση εικόνων σε grid ανά prompt
def show_images_grid(images_dict, results_dict, n_cols=4):
    for prompt, imgs in images_dict.items():
        n_images = len(imgs)
        n_rows = 1 # οριζόντιο grid
        fig, axes = plt.subplots(n_rows, n_cols,
figsize=(4*n_cols, 4*n_rows))

        # Αν έχουμε μόνο μια εικόνα, axes δεν είναι λίστα
        if n_images == 1:
            axes = [axes]

        for i, ax in enumerate(axes):
            if i < n_images:
                ax.imshow(imgs[i])
                ax.axis("off")
            else:
                ax.axis("off") # κενά axes

        mean_clip = results_dict[prompt]["mean_clip"]
        std_clip = results_dict[prompt]["std_clip"]
        plt.suptitle(f"{prompt}\nCLIP score: {mean_clip:.3f} ±
{std_clip:.3f}", fontsize=14)
        plt.show()

# Κλήση της συνάρτησης
show_images_grid(images_dict, results)
```

Παράρτημα Γ' : Script για FID -CLIP score από COCO dataset

FID SCORE SD15 – SD21

```
!pip install diffusers transformers accelerate torch torchvision
pytorch-fid pycocotools ftfy regex tqdm -quiet

!wget -q http://images.cocodataset.org/zips/val2017.zip
!unzip -q val2017.zip

!wget -q
http://images.cocodataset.org/annotations/annotations_trainval201
7.zip
!unzip -q annotations_trainval2017.zip

import json, os, shutil
from PIL import Image

N = 1000 # βάλτε 5000 για serious experiment

with open('annotations/captions_val2017.json') as f:
    data = json.load(f)

annotations = data['annotations']

# Κρατάμε ένα caption ανά image_id
seen = set()
selected = []

for ann in annotations:
    if ann['image_id'] not in seen:
        selected.append(ann)
        seen.add(ann['image_id'])
    if len(selected) == N:
        break

os.makedirs("real", exist_ok=True)

prompts = []
image_ids = []

for ann in selected:
    img_name = f"{ann['image_id']:012d}.jpg"
    src = os.path.join("val2017", img_name)

    img = Image.open(src).convert("RGB")
    img = img.resize((512,512))
```

```
img.save(os.path.join("real", img_name))

prompts.append(ann['caption'])
image_ids.append(img_name)

print("Matched samples:", len(prompts))

from huggingface_hub import login
login(token="your token")

import torch
from diffusers import StableDiffusionPipeline
import os

device = "cuda"

pipe15 = StableDiffusionPipeline.from_pretrained(
    "sd-legacy/stable-diffusion-v1-5",
    torch_dtype=torch.float16
).to(device)

pipe15.safety_checker = None
os.makedirs("sd15", exist_ok=True)

generator = torch.Generator(device=device).manual_seed(42)

for i, prompt in enumerate(prompts):
    image = pipe15(
        prompt,
        num_inference_steps=30,
        guidance_scale=7.5, #7.5
        height=512,
        width=512,
        generator=generator
    ).images[0]

    image.save(f"sd15/{image_ids[i]}")

    if i % 100 == 0:
        print("SD1.5:", i)

pipe21 = StableDiffusionPipeline.from_pretrained(
    "stabilityai/stable-diffusion-2-1-base ", # #sd2-
    community/stable-diffusion-2-1
    torch_dtype=torch.float16
).to(device)

pipe21.safety_checker = None
```

```
os.makedirs("sd21", exist_ok=True)

generator = torch.Generator(device=device).manual_seed(42)

for i, prompt in enumerate(prompts):
    image = pipe21(
        prompt,
        num_inference_steps=30,
        guidance_scale=7.5, #7.5
        height=512,
        width=512,
        generator=generator
    ).images[0]

    image.save(f"sd21/{image_ids[i]}")

    if i % 100 == 0:
        print("SD2.1:", i)

from pytorch_fid import fid_score

device = torch.device("cuda")

fid_15 = fid_score.calculate_fid_given_paths(
    ["real", "sd15"],
    batch_size=32,
    device=device,
    dims=2048
)

fid_21 = fid_score.calculate_fid_given_paths(
    ["real", "sd21"],
    batch_size=32,
    device=device,
    dims=2048
)

print("FID SD1.5:", fid_15)
print("FID SD2.1:", fid_21)

from transformers import CLIPProcessor, CLIPModel
import torch
from PIL import Image
from tqdm import tqdm

model = CLIPModel.from_pretrained("openai/clip-vit-base-
patch32").to(device)
```

```
processor = CLIPProcessor.from_pretrained("openai/clip-vit-base-
patch32")

def compute_clip_score(folder):
    scores = []
    for i, prompt in tqdm(list(enumerate(prompts))):
        img =
Image.open(f"{folder}/{image_ids[i]}").convert("RGB")
        inputs = processor(text=[prompt], images=img,
return_tensors="pt", padding=True).to(device)
        outputs = model(**inputs)
        logits = outputs.logits_per_image
        scores.append(logits.item())
    return sum(scores)/len(scores)

clip15 = compute_clip_score("sd15")
clip21 = compute_clip_score("sd21")

print("CLIP Score SD1.5:", clip15)
print("CLIP Score SD2.1:", clip21)

print("\n===== FINAL RESULTS =====")
print(f"SD1.5 | FID: {fid_15:.2f} | CLIP: {clip15:.3f}")
print(f"SD2.1 | FID: {fid_21:.2f} | CLIP: {clip21:.3f}")
```

Παράρτημα Δ' : Script - Aesthetic score

```
!pip install diffusers transformers accelerate torch torchvision  
tqdm pillow pandas matplotlib  
!pip install open_clip_torch
```

```
!wget https://github.com/christophschuhmann/improved-aesthetic-  
predictor/raw/main/sac+logos+ava1-l14-linearMSE.pth
```

```
import torch  
import torch.nn as nn  
import open_clip  
import numpy as np  
import os  
import json  
import random  
import pandas as pd  
  
from tqdm import tqdm  
from PIL import Image  
from diffusers import StableDiffusionPipeline
```

```
device = "cuda" if torch.cuda.is_available() else "cpu"  
print(device)
```

```
class AestheticPredictor(nn.Module):  
    def __init__(self, input_size=768):  
        super().__init__()  
        self.layers = nn.Sequential(  
            nn.Linear(input_size, 1024),  
            nn.Dropout(0.2),  
            nn.Linear(1024, 128),  
            nn.Dropout(0.2),  
            nn.Linear(128, 64),  
            nn.Dropout(0.1),  
            nn.Linear(64, 16),  
            nn.Linear(16, 1)  
        )
```

```
    def forward(self, x):  
        return self.layers(x)
```

```
clip_model, _, preprocess =  
open_clip.create_model_and_transforms(  
    "ViT-L-14",  
    pretrained="openai"  
)
```

```
clip_model = clip_model.to(device)
clip_model.eval()

aesthetic_model = AestheticPredictor(768).to(device)

state_dict = torch.load(
    "sac+logos+ava1-114-linearMSE.pth",
    map_location=device
)

aesthetic_model.load_state_dict(state_dict)
aesthetic_model.eval()

!wget
http://images.cocodataset.org/annotations/annotations_trainval201
7.zip
!unzip -q annotations_trainval2017.zip

with open("annotations/captions_val2017.json") as f:
    coco=json.load(f)

captions=[x["caption"] for x in coco["annotations"]]

random.seed(42)
prompts=random.sample(captions,1000)

print("Prompts:",len(prompts))

pipe15 = StableDiffusionPipeline.from_pretrained(
    "sd-legacy/stable-diffusion-v1-5",
    torch_dtype=torch.float16
).to(device)

pipe21 = StableDiffusionPipeline.from_pretrained(
    "sd2-community/stable-diffusion-2-1",
    torch_dtype=torch.float16
).to(device)

pipe15.safety_checker=None
pipe21.safety_checker=None

def generate(pipe, folder):

    os.makedirs(folder,exist_ok=True)
    generator=torch.Generator(device).manual_seed(42)

    paths=[]
```

```
for i, prompt in enumerate(tqdm(prompts)) :

    image=pipe(
        prompt,
        height=512,
        width=512,
        guidance_scale=7.5,
        num_inference_steps=30,
        generator=generator
    ).images[0]

    path=f"{folder}/{i}.png"
    image.save(path)
    paths.append(path)

return paths

d15_paths = generate(pipe15, "sd15")
sd21_paths = generate(pipe21, "sd21")

def compute_aesthetic(paths):

    scores=[]

    with torch.no_grad():

        for p in tqdm(paths):

            img=preprocess(
                Image.open(p).convert("RGB")
            ).unsqueeze(0).to(device)

            feat=clip_model.encode_image(img)
            feat/=feat.norm(dim=-1, keepdim=True)

            score=aesthetic_model(feat)
            scores.append(score.item())

    return np.mean(scores), scores

mean15,scores15=compute_aesthetic(sd15_paths)
mean21,scores21=compute_aesthetic(sd21_paths)

print("\n==== AESTHETIC RESULTS =====")
print("SD1.5:",mean15)
print("SD2.1:",mean21)
```

```
from google.colab import drive
drive.mount('/content/drive')

df=pd.DataFrame({
    "prompt":prompts,
    "sd15_aesthetic":scores15,
    "sd21_aesthetic":scores21
})

df.to_csv("aesthetic_results.csv",index=False)

import pandas as pd

df = pd.read_csv("aesthetic_results.csv")

df.head()

import matplotlib.pyplot as plt

plt.figure(figsize=(8,5))

plt.hist(df["sd15_aesthetic"],
        bins=40,
        alpha=0.5,
        label="SD1.5")

plt.hist(df["sd21_aesthetic"],
        bins=40,
        alpha=0.5,
        label="SD2.1")

plt.xlabel("Aesthetic Score")
plt.ylabel("Frequency")
plt.title("Aesthetic Score Distribution")
plt.legend()

plt.show()

plt.figure(figsize=(6,5))

plt.boxplot(
    [df["sd15_aesthetic"],
     df["sd21_aesthetic"]],
    labels=["SD1.5", "SD2.1"]
)

plt.ylabel("Aesthetic Score")
plt.title("Aesthetic Comparison")
```

```
plt.show()

means = [
    df["sd15_aesthetic"].mean(),
    df["sd21_aesthetic"].mean()
]

plt.figure()

plt.bar(["SD1.5", "SD2.1"], means)

plt.ylabel("Mean Aesthetic Score")
plt.title("Average Aesthetic Quality")

plt.show()

from scipy.stats import ttest_ind

t, p = ttest_ind(
    df["sd15_aesthetic"],
    df["sd21_aesthetic"]
)

print("t-stat:", t)
print("p-value:", p)

plt.figure(figsize=(8, 5))

plt.hist(df["sd15_aesthetic"], bins=40, alpha=0.5, label="SD1.5")
plt.hist(df["sd21_aesthetic"], bins=40, alpha=0.5, label="SD2.1")

plt.axvline(df["sd15_aesthetic"].mean(),
            linestyle="--")

plt.axvline(df["sd21_aesthetic"].mean(),
            linestyle="--")

plt.xlabel("Aesthetic Score")
plt.ylabel("Frequency")
plt.title("Aesthetic Score Distribution (COCO-1000)")
plt.legend()

plt.show()
```

Υπεύθυνη Δήλωση Συγγραφέα:

Δηλώνω ρητά ότι, σύμφωνα με το άρθρο 8 του Ν.1599/1986, η παρούσα εργασία αποτελεί αποκλειστικά προϊόν προσωπικής μου εργασίας, δεν προσβάλλει κάθε μορφής δικαιώματα διανοητικής ιδιοκτησίας, προσωπικότητας και προσωπικών δεδομένων τρίτων, δεν περιέχει έργα/εισφορές τρίτων για τα οποία απαιτείται άδεια των δημιουργών/δικαιούχων και δεν είναι προϊόν μερικής ή ολικής αντιγραφής, οι πηγές δε που χρησιμοποιήθηκαν περιορίζονται στις βιβλιογραφικές αναφορές και μόνον και πληρούν τους κανόνες της επιστημονικής παράθεσης.