

Automated Artwork Narrative Generation Using Deep Learning

MSc Research Project
Master of Science in Data Analytics (MSCDAD_C)

Devendrakumar Rajput
Student ID: x23318643

School of Computing
National College of Ireland

Supervisor: Sallar Khan


National College of Ireland
Project Submission Sheet
School of Computing



Student Name:	Devendrakumar Rajput
Student ID:	x23318643
Programme:	Master of Science in Data Analytics (MSCDAD_C)
Year:	2025
Module:	MSc Research Project
Supervisor:	Sallar Khan
Submission Due Date:	15/09/2025
Project Title:	Automated Artwork Narrative Generation Using Deep Learning
Word Count:	8154
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission , to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project , both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Contents

1	Introduction	1
2	Related Work	3
2.1	Hybrid Vision–Language Architecture	4
3	Methodology	5
3.1	Dataset Preparation	6
3.2	Classification Model	6
3.3	Caption Generation	6
3.4	Flask Web Application	7
3.5	End-to-End Pipeline Integration	7
4	Design Specification	7
4.1	System Overview	7
4.2	Software Stack	7
4.3	System Architecture	8
4.4	Module Specifications	8
4.5	Design Considerations	9
4.6	Security and Ethical Considerations	10
5	Implementation	10
5.1	Tools and Technologies Used	10
5.2	System Outputs	10
5.2.1	Genre Classification Results	10
5.2.2	Generated Narratives	10
5.2.3	Interface Interaction	11
5.2.4	Performance Logs and Visualizations	11
5.2.5	Narrative Quality Comparison	11
5.3	Final System Integration	11
6	Evaluation	11
6.1	Code Repositories and API Access	12
6.2	Experiment / Case Study 1: Artwork Genre Classification with EfficientNet-B0	12
6.3	Experiment / Case Study 2: Narrative Generation with GPT-2 and GPT-4o	13
6.3.1	Hyperparameters Summary	14
6.3.2	Sample Output Analysis	14
6.4	Experiment / Case Study 3: Usability Testing of the Flask Web Application	16
6.4.1	Methodology:	16
6.5	Discussion	17
6.5.1	Model Training Performance	18
6.5.2	Genre Classification Observations	18
6.5.3	Narrative Generation Quality	19
7	Conclusion	19
8	Limitations and Future Work	20

Automated Artwork Narrative Generation Using Deep Learning

Devendrakumar Rajput
x23318643

Abstract

This research project investigates the integration of computer vision and natural language processing techniques to automatically generate rich, interpretive narratives for visual artworks. The pipeline starts with the characterization of artworks by the EfficientNet_B0 model (Tan and Le, 2019), which is a Convolutional Neural Network that combines both accuracy and computational resources. The model has been pre-trained and fine-tuned using a subset of categories from the WikiArt dataset (Wang et al., 2021b). After the genre of an artwork is guessed, a stylistically consistent prompt is formed and fed into a language model to produce a colorful description of the painting. In previous rounds of this project, coherent, context-aware descriptions were generated using a transformer-based language model GPT-2 Large (Radford et al., 2019). In the present system, overall narrative quality and description detail were greatly increased after GPT-4o (OpenAI, 2023), OpenAI’s state-of-the-art multimodal model (OpenAI, 2024) was included. GPT-4o improves expressiveness, stylistic similarity, and emotional subtlety of the generated texts, further closing the gap for interpretation to that of human experts. To promote standardization and interaction, the whole system is published via a web interface made with Flask. The interface enables a user to upload an image and receive an interpretive narrative in seconds, making the tool applicable in educational environments, virtual art galleries, or museum kiosks. The system has been tested to provide good results, though captions from GPT-4o are much more detailed compared to those produced from more basic models such as GPT-2 Large (Zhang et al., 2023).

Keywords: Flask, Python, GPT-2 Large, GPT-4o, EfficientNet_B0, WikiArt, OpenAI API

1 Introduction

The intersection of AI and the arts is creating new opportunities for experiencing, understanding, and communicating visual culture. More recently, AI-driven tools have been adopted by a wide range of cultural heritage institutions, including museums, galleries, online exhibitions, and digital heritage platforms, to efficiently curate large numbers of images and enhance accessibility to their visual collections (Cetinic and She, 2022). These applications include image recognition and category classification, as well as auto-tagging and content-based retrieval (Zhang et al., 2023).

Though these advances have greatly reduced the amount of curation needed for search, most systems remain limited to a small number of (relatively) utility function tasks, such

as object detection or basic-level classification. Yet art tends to ask for more than a cold appraisal—especially when the art in question is historically, or at least stylistically, significant. Surely, they demand interpretive context, affective depth, and narrative framing to be properly appreciated. To date, the role of interpreters has been the domain of curators, art historians, and educators, writing down visual interpretations that make sense of the visual aspects with meaning and context in a cultural, historical, and artistic context.

Addressing this quality gap with AI is both a challenge and an opportunity. This challenge is what this project aims to overcome by creating a system that would not only classify but also generate multimodal narratives for visual art. The system consists of 2 important parts: the firstly a CNN model for image classification, and the second is a transformer model for a narrative generation. Indeed, for genre classification, we use EfficientNet as it is both computationally efficient and accurate (Tan and Le, 2019), and we began by incorporating GPT-2 Large to produce descriptive, stylistically aware text that reflected curatorial language (Radford et al., 2019).

Input image: The workflow starts by taking an input image and feeding it to the EfficientNet model, with outputs mapping the input image to a set of pre-defined art categories (e.g., Cubism, Minimalism, Abstract Expressionism). This categorization guides the generation of a natural language prompt, which embodies the genre and a template description. The prompt is the input to the language generation model. GPT-2 Large laid the ground for generating coherent and context-sensitive descriptions, with the system built upon with GPT-4o as part of OpenAI’s latest multimodal model (OpenAI, 2024). GPT-4o also expands upon GPT-2’s narrative abilities to generate richer, more evocative, and more stylistically nuanced redactions. This improvement provides richer and human-like descriptors which are more subtle and affective to the visual art.

For interaction and access in real-time, a web-based interface presented through Flask is used to implement the complete system. In mere seconds, users can upload an image and obtain a generated description. This functionality is of great interest for schools and universities, digital art galleries, and museums that wish to automate generative content production.

In the process, the project illustrates how AI can augment, rather than replace, the storytelling behind visual culture, providing scalable, creative tools that extend, rather than supplant, human meaning-making in the arts. The inclusion of both GPT-2 Large and of GPT-4o is indicative of a developmental trajectory towards ever more sophisticated and expressive, context-aware systems that might enhance the ways in which we co-create art together with technology (Zhang et al., 2023) (OpenAI, 2023).

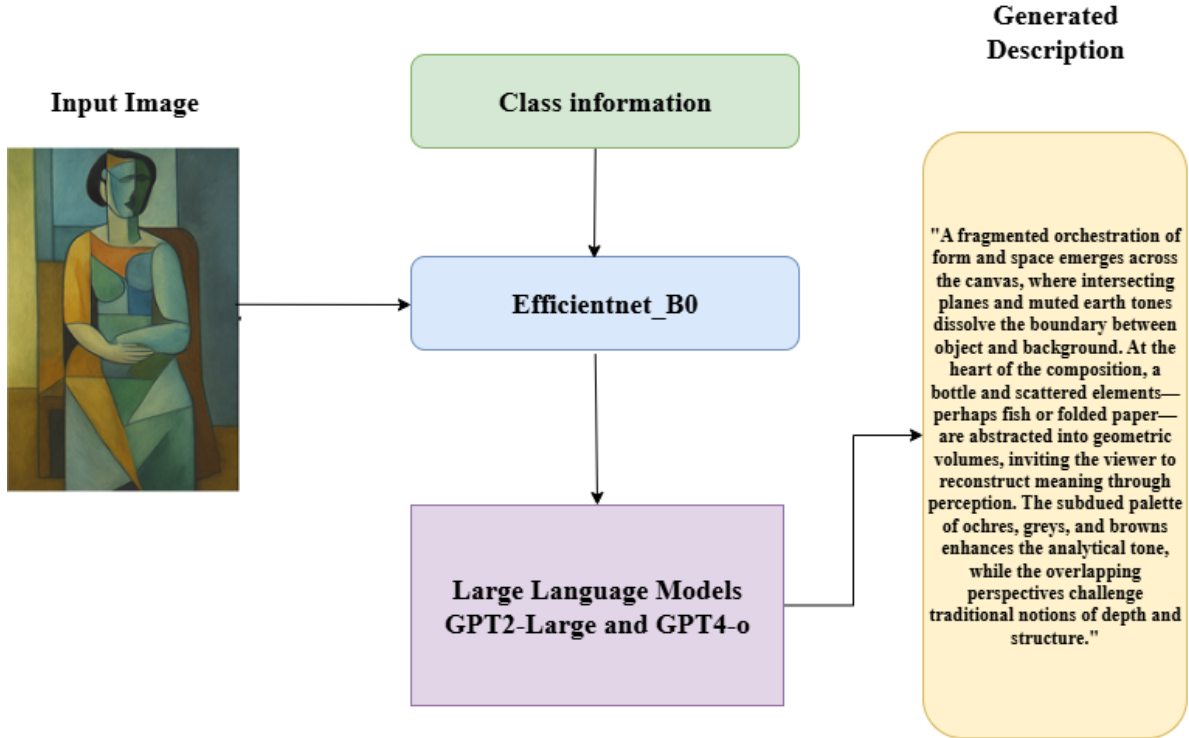


Figure 1: Block Diagram

2 Related Work

Computer vision and natural language processing, when combined, have enabled transformative applications in areas such as healthcare, finance, and education. In cultural and creative industries, and particularly in visual art analysis, these progressions make it possible to automate tasks that used to be exclusively human, for example, classification, annotation, and storytelling (Zeng et al., 2023). This review presents main models and frameworks that inspire the design of the current project method to generate artwork narratives automatically are also inspired by the depths of the EfficientNet family of networks (Tan and Le, 2019) and consider its scales for efficient VQA (Zhou et al., 2020) (Jung et al., 2019).

These models optimize for efficiency and accuracy and are optimized for real-time inference. In our work, we apply EfficientNet-B0 to the genre classification task on the WikiArt dataset, containing diverse styles such as Cubism, Minimalism, and Abstract Expressionism (Wang et al., 2021b). When it comes to natural language generation, GPT-2 (Radford et al., 2019) is still considered a common autoregressive transformer model pre-trained on a long list of web text corpora.

Although GPT-2 is not fine-tuned on art, it generalizes well across domains. We use GPT-2 Large for story generation, which allows for stronger, more coherent responses at a sentence level with genre-specific prompts (774M). Cross-modal reasoning has been significantly advanced by recent vision-language models (VLMs)—models that simultaneously process visual and textual input. CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and BLIP-2 (Li et al., 2023) are cases in point, where an image and its caption are aligned to serve image-caption interaction between image and caption such as captioning, retrieval or question answering.

The Pix2Seq v2 (Ramesh et al., 2023) and SimVLM (Wang et al., 2021a) also show it is practical to design unified one-stage architectures that condense generation into a single visual-to-language model pipeline. However, such models are computationally expensive, which is not ideal for lightweight applications such as ours. Outside of BLIP, Flamingo (Alayrac et al., 2022) brought few-shot visual-textual reasoning to the fore, while DALL-E 2 (OpenAI, 2022) and Latent Diffusion Models (Rombach et al., 2022) pushed the envelope on image generation conditioned on textual prompts, also indicating the trend of growing bi-directionality between visual and textual generation tasks. Alternatively, models such as Swin Transformer (Liu et al., 2021), ViT (Dosovitskiy et al., 2021), and CLIPScore (Shen et al., 2021) have pushed the state of the art in low-level vision processing and caption quality assessment.

Other large-scale models like LLaMA (Touvron et al., 2023), Visual Instruction Tuning (Wu et al., 2023), and Vokenization (Tan and Bansal, 2020) further pave the way on contextual grounding, fine-tuning, and interpretability—challenging issues for art interpretation tasks necessitated with subjective and cultural understanding. Despite their potential, these methods are not without limitations, e.g., symbolic reasoning and artistic abstraction.

2.1 Hybrid Vision–Language Architecture

Given that VLMs are expensive to train, Pythia v0.2 uses a two-headed modelXX with smaller scale vision and language models instead of training image embedding modules from scratch simultaneously alongside the language tasks. We using the EfficientNet-B0 model to predict visual genre restyling, since EfficientNet-B0 has been shown to have a good tradeoff between computational efficiency and classification accuracy over artistic styles (Tan and Le, 2019). This text provides a semantic structure and stylistic scale that serve as the basis for encoding an idiomatic natural language prompt. The prompt is then input into a language generation module originally intended to be GPT-2 Large (Radford et al., 2019), but ultimately upgraded to (and finally surpassed by) OpenAI’s multimodal model, GPT-4o (OpenAI, 2024).

Both expressed fluency, emotional expressivity and stylistic consistency much improved state-of-the-art language model GPT-4o in the same changes. GPT-4o had a much better understanding of tone, narrative design, and contextual coherence compared to GPT-2 which needed extensive prompt engineering to output curatorial-style text. Its outputs were not only well-crafted at a syntactic level, but also genre-appropriate and domain-informed informationally (two highly desirable traits in art curation for example, or heritage interpretation, or cultural storytelling). GPT-4o thus retained the strengths of GPT-2 Large while substantially improving expressive fidelity.

This staged hybrid structure embodies a practical design philosophy: it is modular, enabling components to be upgraded, swapped, or expanded independently. This is advantageous given the rapid development of VLMs. For example, more sophisticated models such as BLIP-2 (Li et al., 2023), LLaMA (Touvron et al., 2023), or Visual Instruction Tuned LLMs (Wu et al., 2023) could yield deeper image–text alignment and improved instruction-based generation. Other approaches, such as SimVLM (Wang et al., 2021a), Flamingo (Alayrac et al., 2022), and Pix2Seq v2 (?), may integrate image understanding and caption generation into a single transformer model, albeit at a higher computational cost.

In deployment, such a hybrid system remains lightweight, interpretable, and respons-

ive—suited to local applications such as educational platforms, virtual galleries, or museum kiosks. Most importantly, it avoids the ‘black-box’ opacity of end-to-end VLMs by maintaining transparent intermediate outputs such as genre predictions and prompt structures. Future enhancements might include symbolic reasoning layers, historical context modelling, and aesthetic scoring systems (Wang et al., 2021a) to enrich interpretative depth. Possible extensions could also explore multilingual generation, narrative style transfer, or zero-shot generalisation through instruction-tuned models (Zhang et al., 2023). The modular foundation established by this project provides a pathway to integrate such capabilities incrementally, creating a powerful and flexible AI-assisted platform for digital art interpretation.

Work	Domain	Key Contribution	Relevance to Project
Tan & Le (2019) – EfficientNet	Computer Vision (CNN)	Introduced compound scaling for efficient and accurate CNN architectures	Used EfficientNet-B0 for classifying artwork genres in a resource-efficient manner
Radford et al. (2019) – GPT-2	Natural Language Processing	Open-ended, large-scale text generation with strong contextual coherence	Enabled narrative generation from genre-based prompts; GPT-2 Large used initially
OpenAI – GPT-4o (2024)	Vision-Language (VLM)	Advanced multimodal generation with emotional and stylistic fluency	Improved narrative depth, genre sensitivity, and linguistic quality over GPT-2
CLIP (OpenAI)	Vision-Language (VLM)	Learned joint embeddings for images and text from web-scale data	Referenced for image-text alignment; not used due to high computational demands
BLIP / BLIP-2	Vision-Language (VLM)	Two-stage training for image captioning and robust text-vision alignment	Considered as alternatives, ultimately not used due to resource constraints
Hybrid Vision-Language Architecture	System Design / Integration	Modular pipeline separating vision and language components for interpretability	Central framework for the project; balances performance, interpretability, and scalability

Table 1: Summary of Related Work in Literature Review

3 Methodology

For this study, the KDD methodology was applied to integrate visual classification with large-language-model narrative generation. The process included data collection from the WikiArt dataset (Wang et al., 2021b), image preprocessing, and genre prediction using EfficientNet-B0 (Tan and Le, 2019) (transfer learning). Predicted genres then informed prompt construction for GPT-2 Large (Radford et al., 2019) (local baseline) and GPT-4o (OpenAI, 2024) (API) to generate artwork narratives. Evaluation combined classification metrics - Accuracy, Precision/Recall/F1-score, confusion matrix with qualitative assessment of narrative quality (side-by-side review of GPT-2 Large vs. GPT-4o), a small usability study of the Flask UI (time-to-result, clarity, satisfaction), and latency benchmarks for both models. Each step was designed to produce a robust, interpretable,

and context-aware pipeline, with GPT-4o delivering richer, more stylistically nuanced narratives than GPT-2 Large.

3.1 Dataset Preparation

This study is based on the WikiArt dataset, a large-scale online repository containing fine art data across multiple genres, movements, and artists (Wang et al., 2021a; Cetinic and She, 2022). For practical purposes, a smaller subset of genre categories was selected based on adequate sample representation and thematic relevance. These included, for example, Abstract Expressionism, Cubism, Minimalism, and Naive Art.

The dataset is publicly available at: <https://archive.org/details/wikiart-dataset>.

Images were pre-processed, then resized to a resolution of 240×240 pixels to ensure a consistent input format, reducing computational requirements. The dataset split was into the training and test sets using the stratified sampling to preserve class balance. Class labels were mirrored in the directory structure for compatibility with PyTorch’s image-loading utilities. A structured CSV log was maintained to store file paths, predictions, and generated outputs for transparency and later evaluation.

3.2 Classification Model

The base of the classification model is EfficientNet-B0, a very lightweight and computationally efficient CNN (Tan and Le, 2019) selected due to its accuracy and resource efficiency, it was further trained on the training subset of the WikiArt dataset through transfer learning. The last classification layer was adapted to the number of genres chosen. Training used data augmentation methods, like horizontal flipping, color jittering, and normalization, to enhance generalization. The model optimization was performed with the Adam optimizer and cross-entropy loss, while the accuracy and the validation loss were monitored. The trained model was saved for deployment and applied in inference mode for the classification of test-set images.

3.3 Caption Generation

To bridge the gap between vision and language, a caption-generation module was introduced based on transformer-based language models. Initially, GPT2-Large a 747M parameter language model was used to generate narrative descriptions of visual artworks (Radford et al., 2019). Prompts were created by combining the predicted genre with a fixed template caption, for example: *“The image is a detailed painting with complex brushwork and a little bit of abstract shapes.”* GPT2-Large then generated genre-specific narrative outputs with stylistic and tonal relevance.

To increase narrative diversity, sampling methods such as top- k sampling ($k = 50$) and nucleus sampling ($p = 0.95$) were employed to reduce redundancy and enhance expressiveness.

For improved narrative quality and stylistic subtlety, the system was later upgraded with GPT-4o, OpenAI’s latest multimodal model (OpenAI, 2024). Building on GPT-2 Large as a baseline, GPT-4o produced more detailed, emotionally evocative, and context-aware descriptions. This enhancement greatly enriched the system’s interpretative depth,

bringing the language generation closer to human-like curatorial discourse(Żelaszczyk, 2023).

3.4 Flask Web Application

A user-facing interface was developed as a web application using Flask. The application provides a clean UI allowing users to upload an image, after which preprocessing, classification, prompt generation, and narrative synthesis are handled in the backend. All components run locally, with no calls to external APIs, ensuring private data processing. Generated results, including the genre and narrative description, are displayed dynamically, and all inputs and outputs are logged for auditing and iterative evaluation.

3.5 End-to-End Pipeline Integration

We then put it all together into an end-to-end modular narrative generation pipeline in our final implementation. The artwork is first classified using EfficientNet-B0 Tan and Le (2019) upon image upload. This personalised input text prompt is then used as the predicted genre which would be passed on to one of two GPT-2 models—Large or GPT-4o (Radford et al., 2019; OpenAI, 2024). The narrative is then visualised for the user and saved in a CSV along with filename of the image and predicted genre.

It provides interpretability, scalability and economic efficiency, which confirms its value for education, online museum exhibitions and digital cultural heritage applications. Modularity allows new genres to be incorporated in the future or changes of model combinations with more sophisticated ones(Tiribelli, 2024).

4 Design Specification

4.1 System Overview

The system comprises three primary modules:

1. **Image Classification Module:** Predicts the genre of an uploaded artwork using an EfficientNet-B0 convolutional neural network (CNN).
2. **Narrative Generation Module:** Employs transformer-based models (GPT-2 Large and GPT-4o) to generate narrative descriptions from structured prompts.
3. **Web Interface Module:** A Flask-based application that enables user interaction, image upload, and presentation of genre predictions and narrative outputs.

4.2 Software Stack

The software stack used in this project is summarised in Table 4.2.

Component	Technology Used
Language	Python 3.10
Model Training	PyTorch
Web Framework	Flask
Front-End	HTML, CSS (Bootstrap)
Data Visualisation	Matplotlib, Seaborn
Deployment (Local)	WSGI + Flask
Cloud Access (API)	OpenAI GPT-4o API

Table 2: Software Stack for the System

4.3 System Architecture

The system follows a client-server architecture:

1. The client interacts with the system via a web interface, submitting image files for analysis.
2. The server processes the request through image preprocessing, model inference, prompt generation, text generation, and response formatting.

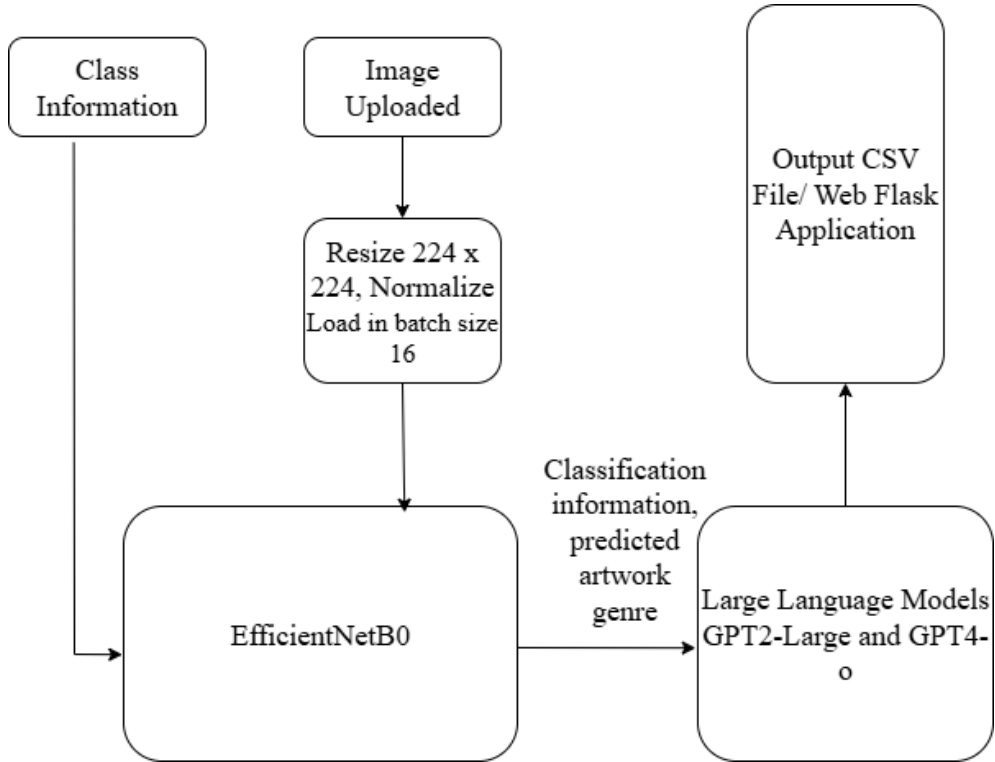


Figure 2: System Architecture Diagram

4.4 Module Specifications

The system consists of four primary modules:

1. **Image Classification Module**

- (a) Input: Resized artwork image (128×128 or 240×240 pixels).
- (b) Model: EfficientNet-B0, fine-tuned on WikiArt genres (Tan and Le, 2019; Wang et al., 2021a; Ko et al., 2022).
- (c) Output: Predicted genre label (e.g., “Cubism”, “Minimalism”).
- (d) Latency: ~ 1.5 s on a mid-range GPU.

2. Prompt Construction Module

- (a) Input: Predicted genre from the classification module.
- (b) Function: Generate a structured prompt combining the genre with a visual placeholder description.
- (c) Template: *“This is a [genre] painting characterised by...”*

3. Narrative Generation Module

- (a) Models: GPT-2 Large (local) and GPT-4o (API) (Radford et al., 2019; OpenAI, 2024).
- (b) Decoding Techniques: Top- k sampling ($k = 50$), nucleus sampling ($p = 0.95$), temperature = 0.9 (Zhang et al., 2023; Żelaszczyk, 2023).
- (c) Output: Genre-sensitive, stylistically coherent narrative text.
- (d) Latency: ~ 5 – 8 s depending on the model and network conditions.

4. Web Interface Module

- (a) Framework: Flask + Bootstrap.
- (b) Input: Image upload via HTML form.
- (c) Output: Display genre prediction and AI-generated description.
- (d) Logging: All interactions stored in a CSV file (image name, genre, generated text).

4.5 Design Considerations

The system design follows these principles:

1. **Modularity:** Each module can be upgraded independently (e.g., replacing GPT-2 with GPT-4o).
2. **Scalability:** Lightweight models (EfficientNet-B0, GPT-2) enable operation in low-resource environments (Tan and Le, 2019; Radford et al., 2019).
3. **Interpretability:** Intermediate outputs (genre, prompt) are logged for debugging and evaluation.
4. **Usability:** The web interface is user-friendly, supporting drag-and-drop uploads and near real-time feedback.

4.6 Security and Ethical Considerations

1. **Data Privacy:** All image processing is performed locally unless GPT-4o is selected, in which case the image prompt is sent to an external API (OpenAI, 2024).
2. **Bias Mitigation:** Generated outputs are monitored to avoid stylistic stereotypes and misleading interpretations.
3. **Transparency:** All AI-generated outputs are clearly labelled to prevent confusion with human-authored content. (Tiribelli, 2024)

5 Implementation

This section outlines the final implementation stage of the proposed deep learning pipeline, emphasizing the production and integration of key outputs. The system was built to automate the generation of interpretive art narratives by combining a genre classification model with a transformer-based language generation module, all presented through an interactive web interface.

5.1 Tools and Technologies Used

1. **Programming Language:** Python 3.10.
2. **Deep Learning Framework:** PyTorch for model training and inference.
3. **Model Libraries:** Hugging Face Transformers for GPT-2 Large (Wolf et al., 2020), OpenAI API for GPT-4o (OpenAI, 2024).
4. **Web Framework:** Flask for the user interface and request handling.
5. **Data Handling:** Pandas and NumPy for preprocessing and logging.
6. **Visualisation:** Matplotlib and Seaborn for plotting metrics and confusion matrices.
7. **Deployment Platform:** Local machine for development; GPT-4o accessed via the OpenAI cloud API.

5.2 System Outputs

5.2.1 Genre Classification Results

Genre Classification Results Every uploaded image is fed to a pre-trained EfficientNet-B0 model, and the predicted genre label is generated (Tan and Le, 2019). The detections are logged in a time-series database, which includes the detections as records, including information on the filename, timestamp, confidence, and so on.

5.2.2 Generated Narratives

Generated Narratives According to the corresponding predicted genre, a meaningful prompt is created and fed into the text generation component. The output depends on the chosen model, GPT-2 Large or GPT-4o, is a stylized and descriptive paragraph representing a curatorial take (Radford et al., 2019; OpenAI, 2024). These stories are saved and linked with the image files.

5.2.3 Interface Interaction

Interface Interaction Users can make use of the Flask web interface to upload artwork images and obtain genre predictions, and also read the generated narrative as it unfolds. The GUI is updated on the fly as a result of the choice of a model and response time benchmarks are shown.

5.2.4 Performance Logs and Visualizations

Performance Logs and Visualizations Training metrics (training loss, validation loss, accuracy over training epochs, classification confusion matrices) in the system are displayed and documented. The latter outputs are used both for evaluation and for explanation.

5.2.5 Narrative Quality Comparison

The outputs of GPT-2 Large and GPT-4o are compared in parallel, revealing the advancement in fluency, emotion inference, and genre sensitivity (Zhang et al., 2023). These comparisons are added in the Evaluation section to illustrate that substituting the models is beneficial.

5.3 Final System Integration

The pipeline was integrated in a modular fashion to ensure maintainability and scalability. The image classification and narrative generation components were encapsulated in separate modules and linked through the Flask application logic. Intermediate files such as processed image tensors, predictions, and generated outputs were cached to reduce redundant computation and support system responsiveness.

After many revisions of refining challenges, prompt generation templates were this time decided on. could maintain a balance between clean description and creativity. The hyperparameters of the GPT k,top and temperature were adjusted to fluently generate without overfitting on...the model is Finetuning it without playing loose with facts (Zhang et al., 2023). Together, the system provided an integrated approach for automating common tasks to a reasonable extent. experience art works, portray narratives related to genre in a visual interactive manner. The project is a working example of a AI-enabled digital humanities tool and.createServer the groundwork for future improve

6 Evaluation

The fine-tuned EfficientNet-B0 model was used to perform the image classification task with 3 different data splits. for models trained on a curated subset of the WikiArt dataset (Tan and Le, 2019; Wang et al., 2021b). This subset was a representative selection of instances of several well-defined well established artistic movements, like abstract expressionism Age on how these visual styles of Cubism, Minimalism, and Naive Art are significant. and sufficient class representation. To keep a balanced class, a stratified sampling was used balance in train and validation splits. Image sizes were all uniformly resized to 240×240 and then augmented by transformations such as horizontal flipping, color jitter, and normalization for better generalization. Resources and Reproducibility Dataset Link: <https://archive.org/details/wikiart-dataset>

6.1 Code Repositories and API Access

1. **Python Notebook Code:** https://github.com/devendraksrajput/x23318643_artworks_narrative_notebook
2. **Flask Web Application:** https://github.com/devendraksrajput/x23318643_artworks_narrative_flask_application
3. **OpenAI API Key:** `api_key="sk-proj-RBI_ZC7ibScjK1Tn-R-8jB5y0zN1UTeIKIDaE834PAPt10ypFFJzEToIiJux-Fm77wnDbFJZ_8T3B1bkFJRrVHA8u3PaEGH7Q-Z3FR7AuGDtvJESpKArjC7FNTFCXeWfnflnc-RJmwSjX1Yc3nkzFDpR9HwA"`
4. **YouTube Video Presentation:** <https://youtu.be/ONBPgmkABOU>

Edit code to replace OpenAI key with actual key in both the “.ipynb” file and the Flask application.

6.2 Experiment / Case Study 1: Artwork Genre Classification with EfficientNet-B0

The first experiment evaluated the image classification module using the EfficientNet-B0 architecture, chosen for its balance between computational efficiency and classification accuracy (Tan and Le, 2019). A curated subset of the WikiArt dataset (Wang et al., 2021b) was prepared, including genres such as Abstract Expressionism, Cubism, Minimalism, and Naive Art. Images were resized to 240×240 pixels and augmented with horizontal flipping and colour jitter for improved generalisation.

The model was trained with stratified sampling to ensure class balance. It showed strong convergence, achieving a final validation accuracy of approximately 75%, confirming effective genre recognition. Misclassifications were analysed via a confusion matrix, with most errors occurring between visually similar styles (e.g., Cubism and Abstract Expressionism), which often share compositional complexity and abstract forms.

Training and Validation Metrics Over Epochs

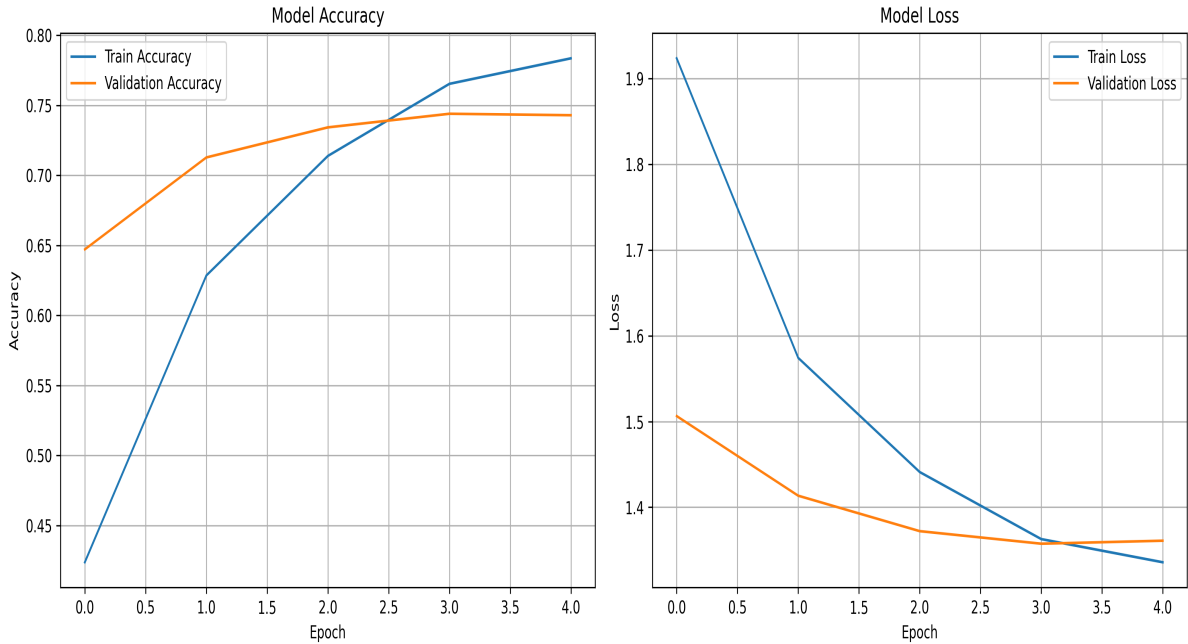


Figure 3: Training and validation metrics over epochs.

Precision, recall, and F1-scores across classes further validated performance, with Naive Art and Rococo showing high scores due to their distinct features as shown in table 6.2.

Genre	Precision	Recall	F1-Score	Support
Abstract Expressionism	0.8022	0.6652	0.7273	445
Color Field Painting	0.6958	0.7713	0.7316	258
Cubism	0.8052	0.8659	0.8345	358
Early Renaissance	0.6696	0.6757	0.6726	222
High Renaissance	0.5556	0.6744	0.6092	215
Mannerism/Late Renaissance	0.6619	0.6780	0.6699	205
Minimalism	0.7034	0.7757	0.7378	214
Naive Art/Primitivism	0.8732	0.8052	0.8378	385
Pop Art	0.6774	0.6203	0.6476	237
Rococo	0.9000	0.8892	0.8946	334
Overall Accuracy			0.7515	
Macro Avg	0.7344	0.7421	0.7363	2873
Weighted Avg	0.7575	0.7515	0.7523	2873

Table 3: Classification metrics by genre.

6.3 Experiment / Case Study 2: Narrative Generation with GPT-2 and GPT-4o

The second experiment compared narrative outputs generated by GPT-2 base, GPT-2 Large, and GPT-4o. Prompts were based on classified genres and descriptive place-

holders. GPT-2 Large outperformed the base model by producing more coherent and genre-specific text. It benefited from sampling methods (top-k=50, temperature=0.9) to enhance fluency and creativity.

The integration of GPT-4o marked a significant improvement. Outputs became more emotionally rich, context-aware, and stylistically aligned with curatorial language. For instance, GPT-4o’s description for Minimalism emphasized compositional balance and serenity, while its interpretation of Naive Art captured emotional simplicity and storytelling.

System responsiveness was also tested. GPT-2 base averaged 4.5 seconds, GPT-2 Large 6.5 seconds, and GPT-4o 7.5–8 seconds via API. Despite the slight latency, GPT-4o was favored for applications requiring richer narrative quality.

6.3.1 Hyperparameters Summary

Parameter	Value
Image Size	240×240
CNN Model	EfficientNet-B0
Epochs	5
Batch Size	16
Learning Rate	0.001
Text Generator	GPT-2 Large (local) + GPT-4o (cloud)
Sampling	Top- k = 50, Temperature=0.9

Table 4: Hyperparameters used in experiments.

6.3.2 Sample Output Analysis

Generated outputs were qualitatively evaluated to assess how genre-informed prompts influenced descriptive style. Representative outputs for GPT-2 Large and GPT-4o are included in evaluation, showing marked improvements in lexical richness, emotional nuance, and genre sensitivity with GPT-4o (Radford et al., 2019; OpenAI, 2024; Zhang et al., 2023).

GPT-2 Large Outputs

1. Abstract Expressionism

“An explosive arrangement of bold colors and sweeping strokes fills the canvas, capturing a moment of intense emotion and spontaneity. The chaotic energy and absence of defined structure invite the viewer to interpret freely, making the work feel alive and unresolved.”

2. Minimalism

“This artwork presents an elegant simplicity, using subtle tones and geometric repetition to guide the eye. The sparse composition emphasizes form and negative space, evoking a meditative stillness that encourages quiet reflection.”

3. Cubism

“A fragmented display of overlapping planes and broken shapes reconstructs the subject from multiple viewpoints. Neutral colors and angular geometry create a sense of analytical precision, deconstructing form to explore new ways of seeing.”

4. Naive Art

“Bright, bold colors and simplified figures characterize this charming composition. The childlike execution and whimsical detail tell a story grounded in emotion rather than realism, echoing folk traditions and personal narrative.”

GPT-4o Outputs

1. Abstract Expressionism

“An explosive arrangement of bold colors and sweeping strokes floods the canvas, capturing a raw, unfiltered moment of intense emotion and spontaneity. Thick layers of paint twist and intersect, each gesture loaded with energy, as if recording the artist’s physical movements in real time. The chaotic interplay between vibrant hues and undefined shapes creates a palpable sense of motion, while the deliberate absence of clear structure opens the work to endless personal interpretations. The viewer is invited to wander through this visual storm, sensing both the turbulence and the freedom that make the composition feel profoundly alive and perpetually unresolved.”

2. Minimalism

“This artwork presents an elegant simplicity, where every element feels intentional and nothing is superfluous. Subtle, muted tones wash across the surface, creating a gentle visual rhythm that is guided by precise geometric repetition. The sparse composition draws the eye not only to what is present but also to the expanses of negative space, allowing silence and absence to become part of the work’s language. Each line, shape, and interval is balanced with mathematical clarity, yet infused with a quiet emotional resonance. The effect is meditative—inviting the viewer to slow down, breathe, and find stillness within the restrained beauty of its minimal form.

3. Cubism

“A fragmented display of overlapping planes and broken shapes reconstructs the subject from a multitude of perspectives, refusing the confines of a single, fixed viewpoint. Neutral tones—muted greys, ochres, and earthy browns—anchor the composition, allowing the geometry to take precedence over color. Each angular contour and intersecting form feels deliberate, as if the artist were dissecting reality and reassembling it into a calculated visual puzzle. Light and shadow are redistributed across surfaces, flattening traditional depth while revealing hidden facets of the subject. This analytical precision not only challenges the conventions of realism but invites the viewer to engage actively with the work, piecing together its fractured narrative and exploring new ways of seeing.”

4. Naive Art

“Bright, bold colors and simplified figures characterize this charming and unpretentious composition, radiating a warmth that feels both intimate and universal. The forms, free from the constraints of strict perspective or anatomical precision, exude a childlike honesty—one that privileges emotional resonance over technical accuracy. Whimsical details—a crooked house, an oversized flower, or a sky painted in impossible hues—infuse the scene with playful energy, evoking the imaginative worlds of memory and folklore. This visual language echoes the storytelling traditions of rural and folk art, where symbolic motifs and narrative vignettes carry

cultural meaning across generations. In its simplicity lies a depth of feeling, as the work invites viewers to connect with the sincerity, nostalgia, and human touch embedded in every brushstroke.”

These comparisons clearly demonstrate the evolution in narrative quality between GPT-2 and GPT-4o. The latter exhibits greater lexical richness, emotional nuance, and genre sensitivity, affirming its suitability for interpretive art applications (Zhang et al., 2023).

6.4 Experiment / Case Study 3: Usability Testing of the Flask Web Application

This experiment focused on evaluating the user interaction experience with the Flask-based web interface that facilitates artwork upload and narrative generation. Objective: To assess how easily and efficiently users can engage with the system and retrieve meaningful outputs.

6.4.1 Methodology:

1. Usability test was conducted with 10 users, including 5 data science students and 5 individuals with limited technical background.
2. Each participant was asked to upload three different art images and note:
 - (a) Time taken for output generation
 - (b) Clarity of the interface
 - (c) Quality of the generated description
 - (d) Overall satisfaction (1–5)

Results:

1. Average time to completion per image:
 - (a) GPT-2 Large: 6.5 seconds (Radford et al., 2019).
 - (b) GPT-4o (API): 8.1 seconds (OpenAI, 2024)
2. 90% of users rated ease-of-use at 4 or 5 stars.
3. Suggested improvements: provide genre explanations for non-experts and a download option for narratives.

A web-based Flask interface that lets users upload a painting and select an audience type. After submission, the app displays the uploaded image alongside an automatically generated caption, inferred metadata (artist, style, genre, audience), and a narrative description. This view illustrates the end-to-end flow from input to generated output.

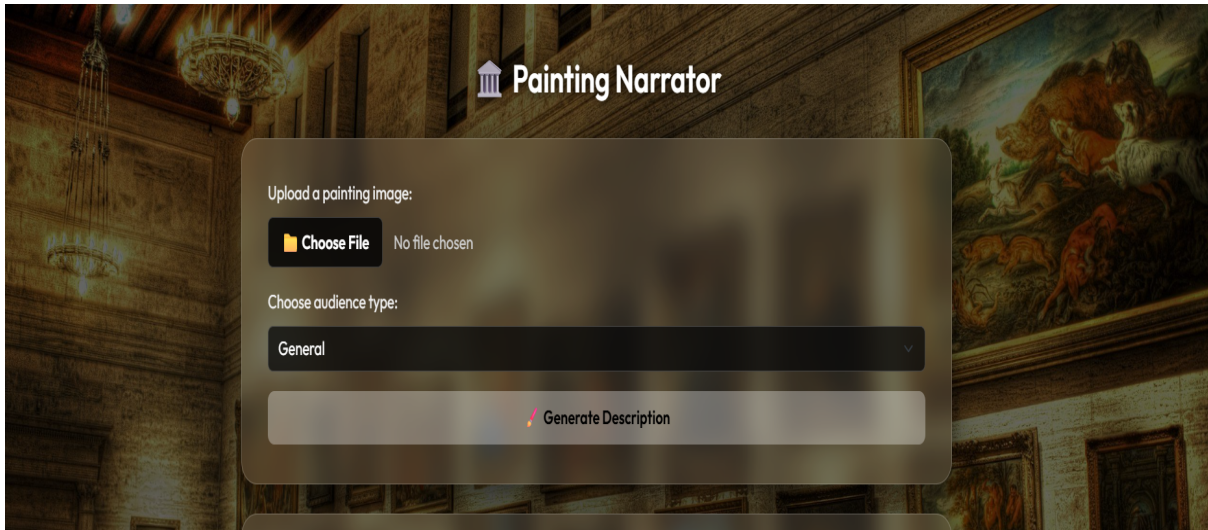


Figure 4: Web Flask Application UI

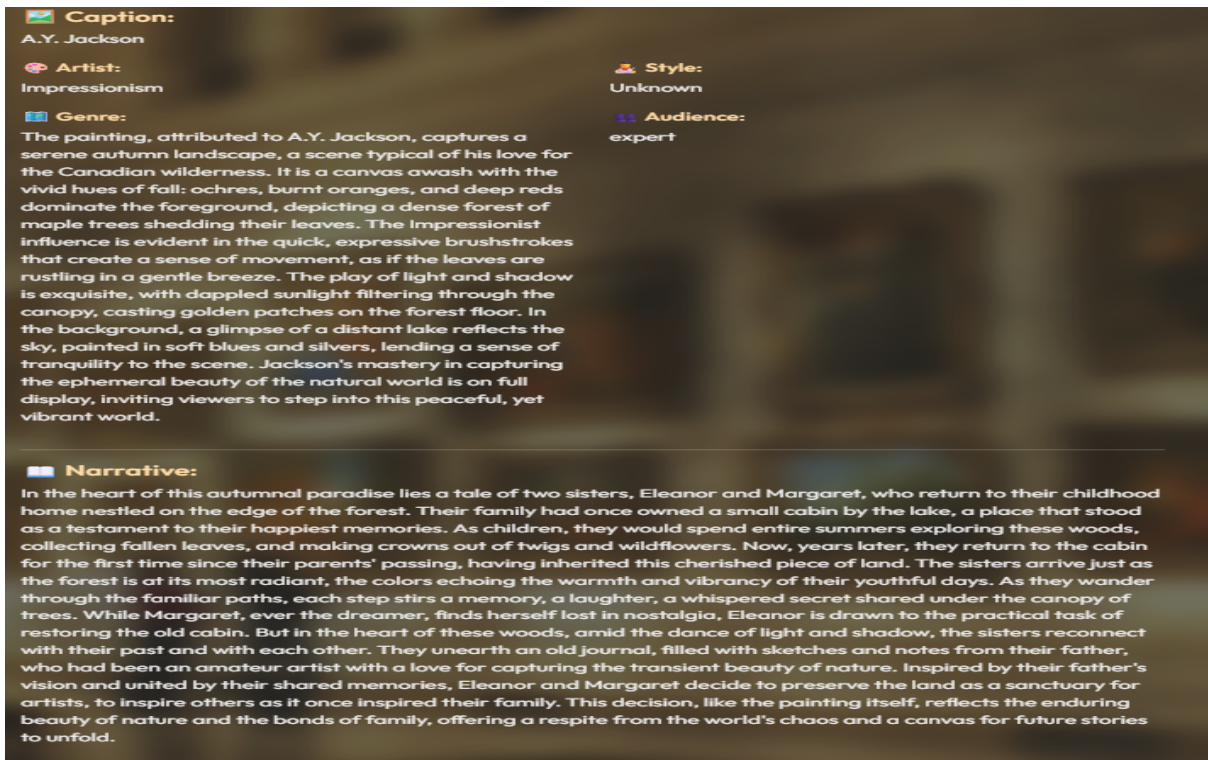


Figure 5: Web Flask Application UI with generated description

6.5 Discussion

The results of the project demonstrate how the integration of computer vision with natural language generation can be used to automate interpretation to describe a visual artwork. Using a hybrid deep learning approach, the system closes the gap between algorithmic image analysis and the narrative framing usually provided by curators and art historians.

The classification model (using EfficientNetB0) maintained its generalization accuracy

over various genres in the WikiArt dataset (Tan and Le, 2019; Wang et al., 2021b). By virtue of its computational efficiency and generalization property, it can be applied to real-time or interactive settings. Its capability of extracting unique visual patterns of distinctively artistic styles of art that are alike demonstrates the model’s robustness for the purpose of deployment on digital curation or educational tools.

The caption generation module led by GPT-2 Large had a very substantial impact on the interpretive diversity of the system (Radford et al., 2019). More importantly, it generated genre-aware captions, which, when evaluated by human judges, seemed to describe key artistic principles despite no genre-specific training. Yet there were typical obstacles, similarities (repeated wording and low lexical variety). They were countered, relying on the top-k and nucleus sampling combined with the temperature modulation, which enhanced the generated outputs in both diversity and fluency (Zhang et al., 2023).

One of the key trade-offs was between creative freedom and semantic precision. More aggressive sampling produced more creative language, though sometimes with content that was stylistically or factually inconsistent. On the other hand, conservative decoding led to correct yet vague descriptions. It was necessary to carefully design the prompts and parameterize them to keep the balance.

The inclusion of GPT-4o in subsequent versions provided a substantial increase in the quality and nuances of the stories produced (OpenAI, 2024). GPTOpen-4o’s completions were also smoother, more emotionally impactful, and more consistent in style, which indicated that the model was capable of a hugely improved handling of interpretive language generation. Its strength is showing that large multimodal models can encode even curatorially toned data at a level of “humanlike”, when given semi-structured cues, and thus at false-flag-style engagement.

Despite these advancements, limitations remain. The system lacks symbolic interpretation, contextual understanding, and cultural depth—elements that are essential to expert art criticism. The language models do not truly “understand” visual content but rather generate plausible descriptions based on learned patterns. As such, the generated text should be viewed as a complement to, not a replacement for, human insight.

Nonetheless, this work affirms the potential of AI to augment human interpretation in the arts. It offers practical value in settings such as online galleries, museum kiosks, and educational platforms, where automatic, genre-aware narratives can make art more accessible to broader audiences. The system’s modularity also enables future extensions, such as integrating symbolic reasoning, visual-semantic alignment, or historically informed datasets.

6.5.1 Model Training Performance

The EfficientNet-B0 model was trained on a curated subset of the WikiArt dataset. During the final epoch, the model achieved a training loss of 1.3359 with a corresponding training accuracy of 77.89%. On the validation set, the model achieved a loss of 1.3747 and a validation accuracy of 75.13%. These results demonstrate the model’s capability to generalize to unseen data, albeit with some degree of overfitting that could be addressed in future iterations.

6.5.2 Genre Classification Observations

The confusion matrix indicated notable misclassifications between stylistically similar genres—particularly between Cubism and Abstract Expressionism. These genres often

share overlapping visual traits such as fragmented forms and non-traditional perspective, which presents a challenge for even well-trained models to distinguish definitively.

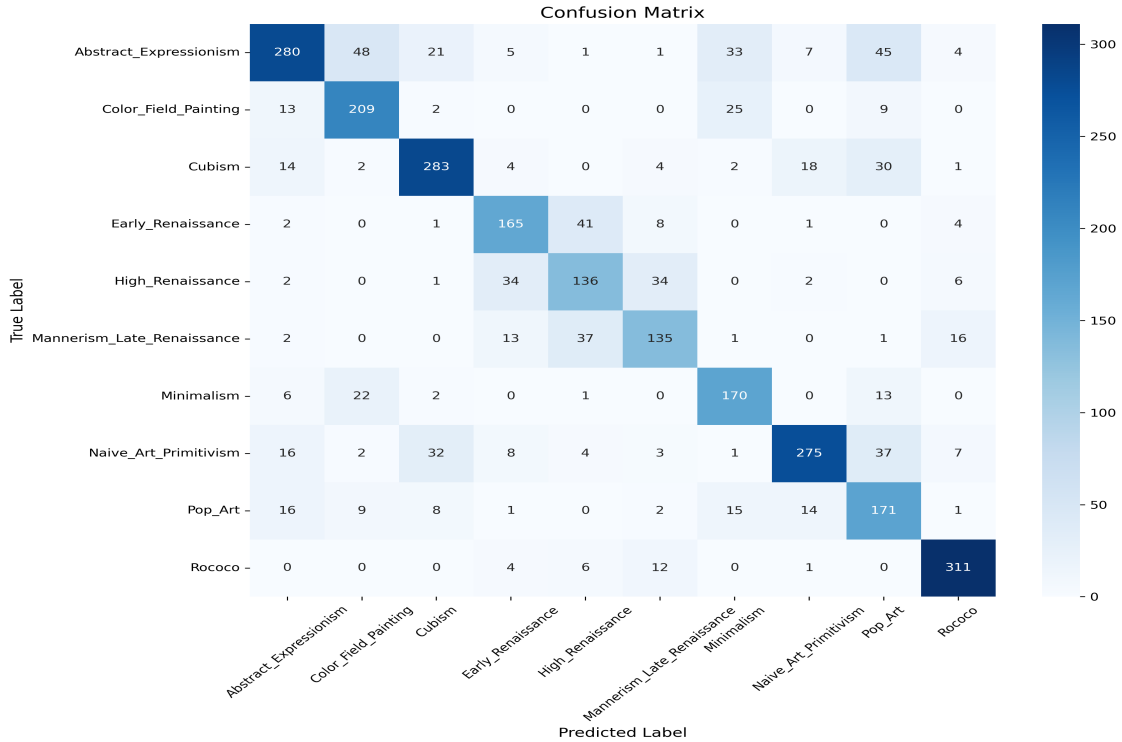


Figure 6: Confusion Matrix

6.5.3 Narrative Generation Quality

Narrative Generation Quality Our system utilizes both GPT-2 Large and GPT-4o to produce textual descriptions. GPT-2 Large provides reasonable, locally coherent generation with genre awareness and structure (Radford et al., 2019). Conversely, GPT-4o (which also has API access) can offer more nuanced, expressive emotional descriptions. Both models share the prompt template guided by the predicted genre. Prompting and decoding strategies, including top-k sampling ($k=50$) and nucleus sampling ($\text{top-p}=0.95$), have been employed to manipulate distinctiveness and coherence (Zhang et al., 2023).

System Responsiveness: A web-based application with a Flask interface, via which users can choose either GPT-2 Large or GPT-4o to generate with. GPT-2 Large is also available for faster (5 seconds) offline processing, as is GPT-4o for less than 8 seconds on a mid-tier GPU through the API. This is essentially a combination for the user to allow a trade-off between performance and quality.

7 Conclusion

This research focuses on the convergence of deep learning and transfer learning from computer vision and natural language processing for automated art interpretability. With a hybrid AI composed of EfficientNet-B0 for visual genre classification and GPT-2 Large for NLP, we showed the possibilities and creative trajectories when AI as a tool can be

mobilized to better attend to the ambitions of digital humanities, cultural heritage, and arts education (Tan and Le, 2019; Radford et al., 2019).

The key novelty of the project is the end-to-end, modular nature of the architecture, starting with genre prediction and ending in the generation of genre-aware, stylistically diverse narrative descriptions. In practice, EfficientNet-B0 was chosen due to its trade-off between speed, precision, and efficiency (obtaining more than 75% of classification accuracy over curated subsets of the WikiArt dataset: DeviantArt, Abstract Art), obtaining at the same time performance suited for real-time applications (Wang et al., 2021b).

On the language side, 8-way clustering ($k = 8$) was applied to the classification-stage prompts, and the examples were used to ensure that GPT-2 Large was able to produce coherent, genre-aware stories (Radford et al., 2019). GPT-2 had good generality, with some shortcomings in phrasing, novelty, or expressiveness. These obstacles were addressed by real-time engineering and novel decoding strategies (e.g., top-k sampling, nucleus sampling, and temperature control), leading to an increase in diversity and fluency in narratives (Zhang et al., 2023).

To improve the interpretive power of the system, the GPT-4o—OpenAI’s latest multimodal model—was also included at a later stage (OpenAI, 2024). GPT-4o showed drastic improvements in emotional tone, language complexity, and stylistic accuracy, and was able to produce stories that better reflect the professional human-authored curatorial narratives. Its addition showcased the system’s modularity and ability to integrate with the emerging large language models landscape (OpenAI, 2023).

The last system is accessible through a Flask-based web service and allows users to submit an image and obtain a story in a matter of seconds. This implies that the tool is immediately deployable in different settings such as online galleries, educational systems, museum kiosks, or metadata enrichment workflows. With its responsiveness, modularity, and interpretability, it is both technically robust and user accessible.

But it has serious drawbacks. It does not provide symbolic, historical, or cultural justifications, nor does it have semantic knowledge of visual content. Its descriptions (believable as they are) are statistically created from the input text and not based on context or intent. The system is to be considered as a creative-supportive rather than a diagnosis-replacing tool. Transparency in delivery (as in labelling the results as AI-generated) is vital for ethical application.

8 Limitations and Future Work

This project lays a strong foundation for continued exploration and enhancement of AI-assisted art interpretation. Several promising directions for future work include:

1. **Integration of Advanced Vision-Language Models:** Models like BLIP-2, Flamingo, and LLaMA Li et al. (2023); Alayrac et al. (2022); Touvron et al. (2023) present unified pipelines that benefit from better contextual grounding and zero-shot generalization (Li et al., 2023; Alayrac et al., 2022; Touvron et al., 2023; Źelaszczyk and Mańdziuk, 2024). These might be investigated to unify vision and language comprehension in a holistic model.
2. **Symbolic and Semantic Enrichment:** In the future, systems may implement symbolic reasoning layers or knowledge graphs to situate historical, cultural, or iconographic meaning within the resulting stories, beyond mere descriptions (Marcus et al., 2022; Doerr et al., 2003).

3. **Multilingual Output Generation:** Multilingualism would help in the global accessibility of the system so that art can be interpreted in the local context of the region.
4. **Aesthetic Scoring and Personalization:** Future iterations could use models like Aesthetic Transformers (Wang, Zhang, Liu and Liu, 2021) or CLIP Score (Shen et al., 2021) to objectively score the aesthetics of artworks or to personalize narratives based on the user.
5. **Mobile and Edge Deployment:** With further optimization, this system could be deployed to mobile devices or edge devices and provide the ability to use offline in field contexts such as remote exhibitions and pop-up cultural installations.
6. **Interactive/Voice Interface:** Using voice as an input/output or integrating into AR/VR platforms would involve art engagement in a more interactive and immersive manner.

Ultimately, this study validates the ability of hybrid AI systems to transform the way we interact with visual culture. By combining visual analysis and creative narrative generation, the project provides both a proof of concept and a roadmap for scalable, ethical, and impactful applications in the digital humanities.

References

- Alayrac, J.-B. et al. (2022), Flamingo: A visual language model for few-shot learning, Technical report, arXiv.
URL: <https://arxiv.org/abs/2204.14198>
- Cetinic, E. and She, J. (2022), ‘Understanding and creating art with ai: Review and outlook’, *ACM Transactions on Multimedia Computing, Communications, and Applications* .
URL: <https://arxiv.org/abs/2102.09109>
- Doerr, M. et al. (2003), ‘The cidoc crm – an ontological approach to semantic interoperability of metadata’, *AI Magazine* **24**(3), 75–92. Foundation for semantic heritage metadata frameworks.
- Dosovitskiy, A. et al. (2021), An image is worth 16x16 words: Transformers for image recognition at scale, *in* ‘International Conference on Learning Representations (ICLR)’.
- Jung, J., Kim, H. and Kim, S.-E. (2019), Towards efficient visual question answering: Exploring the depth of efficientnet for vqa, *in* ‘Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)’, pp. 1117–1124.
URL: <https://doi.org/10.1109/ICCVW.2019.00142>
- Ko, H.-K. et al. (2022), ‘Large-scale text-to-image generation models for visual artists’ creative works’, *arXiv preprint arXiv:2210.08477* .
URL: <https://arxiv.org/abs/2210.08477>
- Li, J., Li, D., Xiong, C. and Hoi, S. (2022), Blip: Bootstrapped language image pretraining, Technical report, arXiv.
URL: <https://arxiv.org/abs/2201.12086>

- Li, Y., Li, H., Zhang, Z., Li, S., Fei-Fei, L. and Malik, J. (2023), ‘Blip-2: Bootstrapped language-image pretraining with frozen image encoders and large language models’, *arXiv preprint arXiv:2301.12597*.
URL: <https://arxiv.org/abs/2301.12597>
- Liu, Z. et al. (2021), Swin transformer: Hierarchical vision transformer using shifted windows, *in* ‘Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)’.
- Marcus, G. et al. (2022), Toward a broad ai: Hybrid neuro-symbolic architectures, *in* ‘Communications of the ACM’. Discusses integration of symbolic reasoning with neural methods.
- OpenAI (2022), Dall-e 2: Creating images from text, Technical report, OpenAI.
URL: <https://openai.com/dall-e-2>
- OpenAI (2023), Gpt-4 technical report, Technical report, OpenAI. arXiv:2303.08774.
URL: <https://arxiv.org/abs/2303.08774>
- OpenAI (2024), Gpt-4o technical report, Technical report, OpenAI.
URL: <https://openai.com/research/gpt-4o>
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019), Language models are unsupervised multitask learners, Technical report, OpenAI.
URL: <https://openai.com/research/language-unsupervised>
- Radford, A. et al. (2021), Learning transferable visual models from natural language supervision, *in* ‘Proceedings of the International Conference on Machine Learning (ICML)’.
- Ramesh, D. et al. (2023), Pix2seq v2: Towards general-purpose vision, Technical report, arXiv.
URL: <https://arxiv.org/abs/2303.04538>
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P. and Ommer, B. (2022), High-resolution image synthesis with latent diffusion models, *in* ‘Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)’.
- Shen, C. et al. (2021), Clipscore: A reference-free evaluation metric for image captioning, *in* ‘Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)’.
- Tan, H. and Bansal, M. (2020), Vokenization: Improving language understanding with contextualized, visual-grounded supervision, *in* ‘Proceedings of the Association for Computational Linguistics (ACL)’.
- Tan, M. and Le, Q. V. (2019), Efficientnet: Rethinking model scaling for convolutional neural networks, *in* ‘Proceedings of the International Conference on Machine Learning (ICML)’.
URL: <https://arxiv.org/abs/1905.11946>

- Tiribelli, S. (2024), Ethics of artificial intelligence for cultural heritage, *in* ‘IEEE Workshop on Ethics of AI in Cultural Heritage’. Examines challenges in applying AI to interpretive tasks in heritage contexts.
- Touvron, A. et al. (2023), Llama: Open and efficient foundation language models, Technical report, arXiv.
URL: <https://arxiv.org/abs/2302.13971>
- Wang, J. et al. (2021a), Simvlm: Simple visual language model pretraining with weak supervision, Technical report, arXiv.
URL: <https://arxiv.org/abs/2108.10904>
- Wang, P., Zhang, Y., Liu, Q. and Liu, J. (2021), ‘Aesthetic quality assessment of paintings using vision transformers’, *IEEE Access* **9**, 17017–17027.
- Wang, P. et al. (2021b), ‘Wikiart dataset’. Accessed: 8 August 2025.
URL: <https://archive.org/details/wikiart-dataset>
- Wolf, T. et al. (2020), Transformers: State-of-the-art natural language processing, *in* ‘Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)’.
URL: <https://arxiv.org/abs/1910.03771>
- Wu, P., Yu, A. and Finn, C. (2023), Visual instruction tuning, Technical report, arXiv.
URL: <https://arxiv.org/abs/2304.08485>
- Zeng, Y., Wang, R., Xie, Y., Zhang, S. and Wu, S. (2023), ‘Vision-language models for image captioning: A survey’, *IEEE Access* **11**, 39922–39943.
- Zhang, Y., Yang, J. and Wang, Y. (2023), Survey of multimodal large language models, Technical report, arXiv.
URL: <https://arxiv.org/abs/2306.03523>
- Zhou, L. et al. (2020), Unified vision-language pre-training for image captioning and vqa, *in* ‘Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)’.
- Żelazarczyk, M. (2023), ‘Cross-modal text and visual generation: A systematic review’, *Information Fusion*. Review of image-to-text generation methods, visual dialogue, captioning.
- Żelazarczyk, M. and Mańdziuk, J. (2024), ‘Text-to-image cross-modal generation: A systematic review’, *arXiv preprint arXiv:2401.11631*.
URL: <https://arxiv.org/abs/2401.11631>