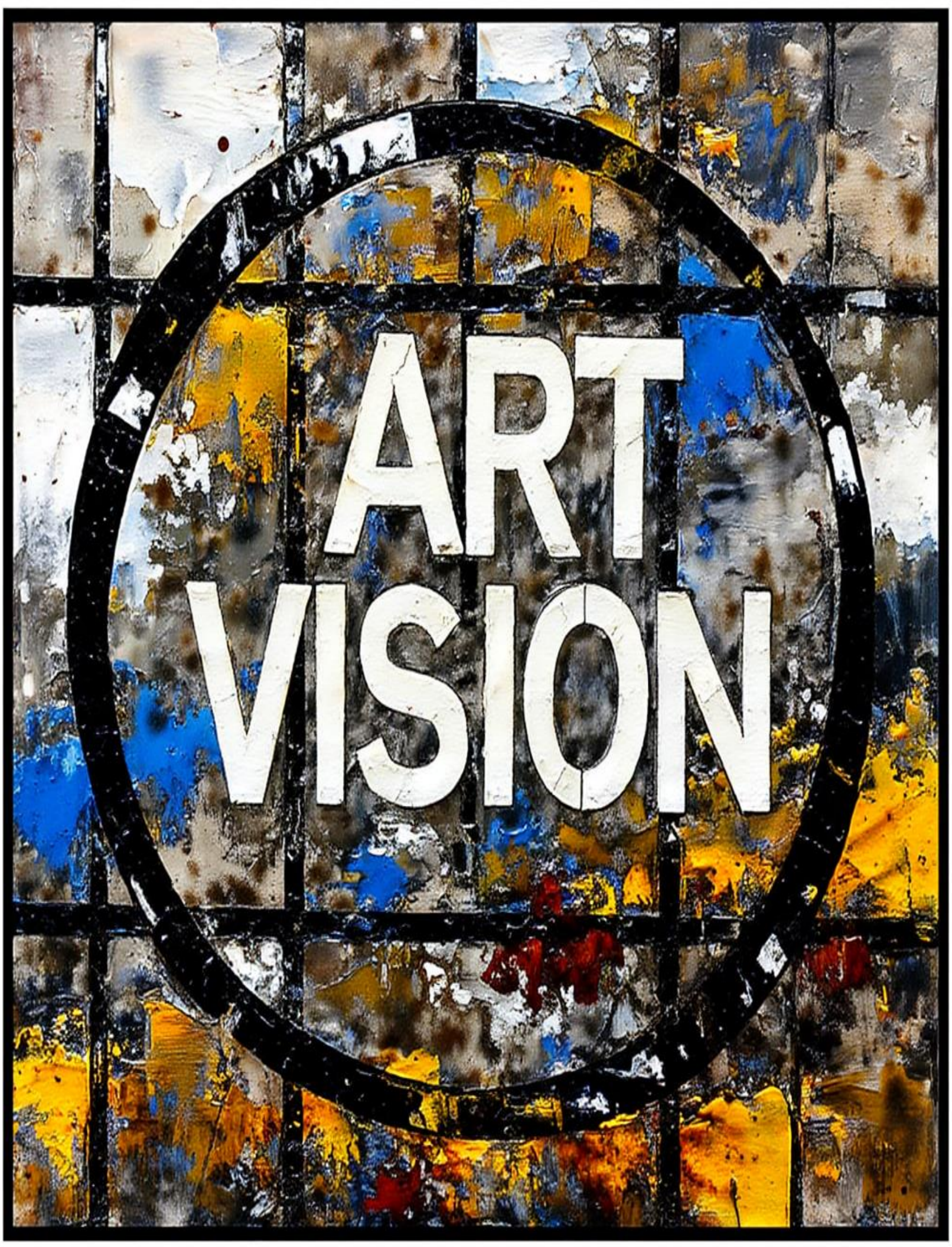

ART
VISION

Table of Contents

1.Introduction:

- **Project Name:** “Art-Vision: Art Critique and Creation Tool”
- **Project Scope:** Web Application
- **Project Idea Overview:**

  Art-Vision is an AI-powered tool designed to offer two core functionalities:

  - **Art Creation**: Generates original artwork based on user-defined description, styles, or parameters, providing creative inspiration art pieces.
  - **Art Critique**: Analyzes an artist's work and provides a critique of various elements, such as color palettes, composition, texture, and style, which can help artists improve or refine their works.

  Art-Vision leverages advanced AI techniques in computer vision and natural language processing to both create and critique artworks, making it a versatile tool in the artistic space. As it aims to provide an intuitive and efficient platform for Artisans, designers, and creatives seeking automated insights and AI-generated visuals.

  With a focus on performance, security, and usability, the project is designed to support a growing user base while maintaining compliance with industry standards and best practices.

---

2. Features List *(*Arranged by importance*)*:

1. **Generate Art from Text Prompt** *(Mandatory)*:
   - Users can input a description and select a specific art style (e.g., sketching, realism, or animation). The system then generates original artwork based on the provided description and the chosen style.
2. **Analyze and Critique Submitted Artwork** *(Mandatory)*:
   - Users can submit an artwork image, and the system generates a critique text of the submitted artworks based on:
     1. **Emotional State**
     2. **Color Harmony**
     3. **Composition**
     4. **Light/Shadow**
     5. **Originality with Art School**
     6. **Art Style**
3. **Enhance Submitted Art Based on Color Harmony** *(Optional)*:
   - Users can submit an artwork image, and the system improves their artwork’s color palette generating more harmonious artwork.

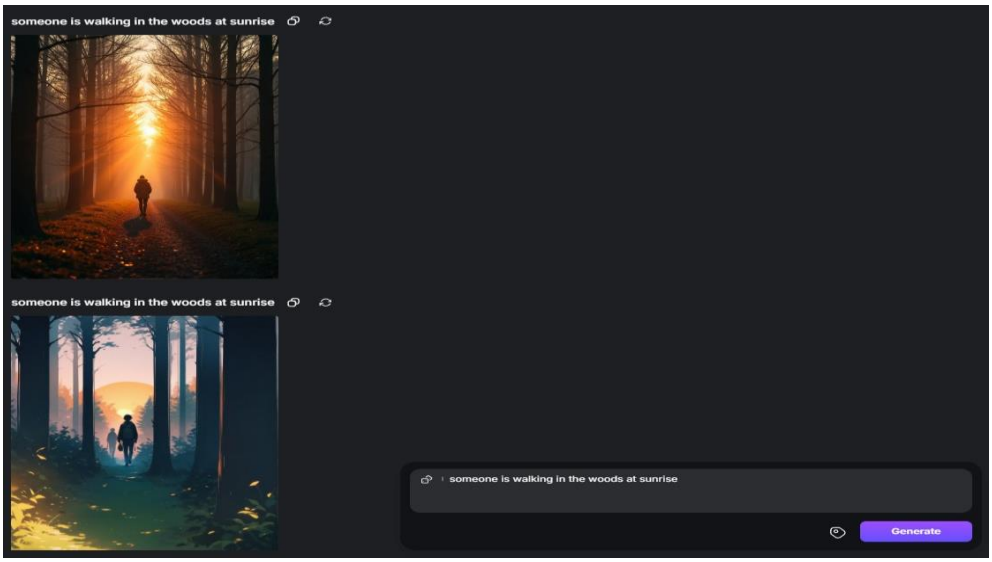
4. **Convert Doodles to Real Art** *(Optional)*:
    - Users can submit simple doodle sketches, and the system will convert these into more refined, realistic artwork.

---

3. Existing Reference Applications:

- **Imagine-Art: https://www.imagine.art/dashboard**
    - Imagine-Art enables users to generate AI-powered artwork based on specific inputs, such as text descriptions or style preferences. It excels in generating high-quality and visually appealing artworks using state-of-the-art AI models.
    - *Advantages*: The platform offers a wide range of artistic styles and generates high-resolution, detailed art from user input.
    - *What Art-Vision Improves*: Art-Vision combines the ability to generate art with a critique feature that helps artists understand and improve their work, making it a more comprehensive tool for artists.

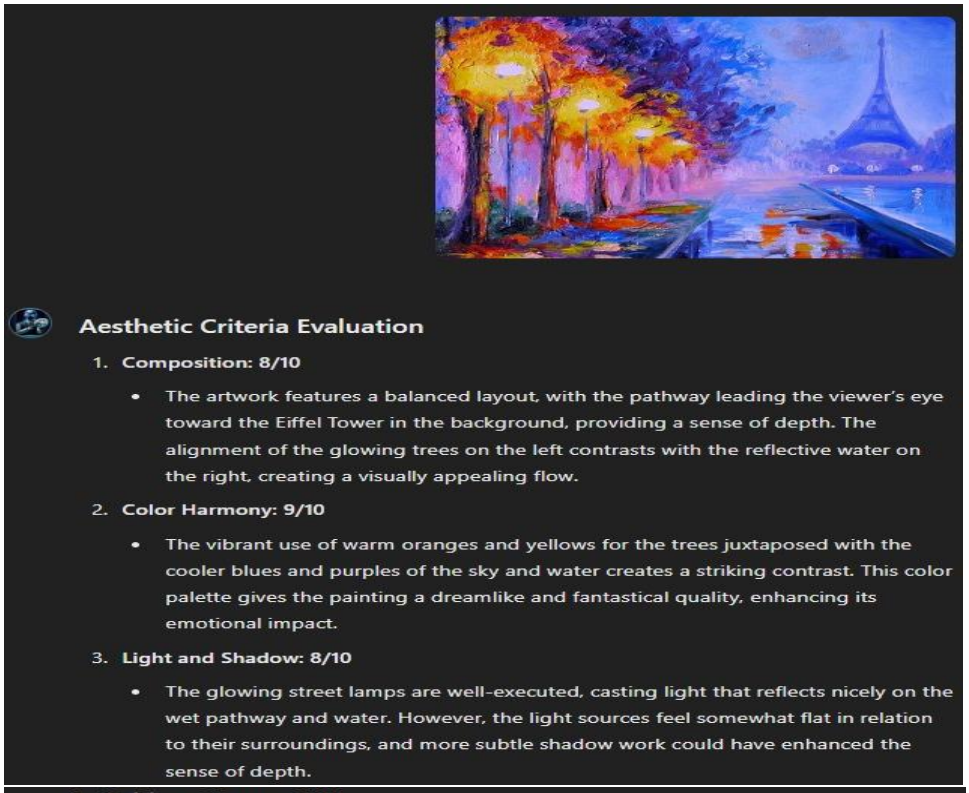
    e.g of **Imagine-Art** website:



- **Art-Alyzer: https://theresanaiforthat.com/gpt/artalyzer/**
    - Art-Alyzer is an AI tool that provides users with a critique of their artwork. It focuses on evaluating aspects such as composition, use of color, and balance, offering artists feedback on how to improve their work.
    - *Advantages*: It provides a deep and professional critique for improving various visual elements.
    - *What Art-Vision Improves*: Art-Vision offers more detailed critique options, such as emotional state analysis and originality with respect to art schools.

Furthermore, Art-Vision not only critiques artwork but also includes an art creation system, allowing users to generate artworks from text prompts.

e.g of **Art-Alyzer** website*:*



Aesthetic Criteria Evaluation

1. **Composition: 8/10**
   - The artwork features a balanced layout, with the pathway leading the viewer's eye toward the Eiffel Tower in the background, providing a sense of depth. The alignment of the glowing trees on the left contrasts with the reflective water on the right, creating a visually appealing flow.
2. **Color Harmony: 9/10**
   - The vibrant use of warm oranges and yellows for the trees juxtaposed with the cooler blues and purples of the sky and water creates a striking contrast. This color palette gives the painting a dreamlike and fantastical quality, enhancing its emotional impact.
3. **Light and Shadow: 8/10**
   - The glowing street lamps are well-executed, casting light that reflects nicely on the wet pathway and water. However, the light sources feel somewhat flat in relation to their surroundings, and more subtle shadow work could have enhanced the sense of depth.
4. **Subject Matter: 7/10**
   - The scene is a romanticized view of Paris, with the Eiffel Tower as a clear cultural symbol. While beautiful, the subject feels a bit conventional, relying on iconic imagery without pushing boundaries in terms of narrative or meaning.
5. **Texture and Detail: 7/10**
   - The thick, expressive brushstrokes lend the painting a lively texture, particularly in the foliage and water reflections. However, some areas, like the background and sky, feel less refined, which slightly detracts from the overall polish.
6. **Originality: 6/10**
   - While the artist's use of color is vibrant and distinctive, the subject matter (a Paris street scene with the Eiffel Tower) is fairly common in art. More experimentation with form or concept could have added uniqueness.

Overall Score: 45/60

This painting effectively captures a vivid and enchanting atmosphere, using bold colors and expressive strokes to evoke a romanticized vision of Paris. While the execution is beautiful, it leans on familiar imagery, and further innovation in subject matter or light handling could elevate it even further.

- **Art-Vision's Unique Contributions:**

Art-Vision distinguishes itself from existing tools and research by offering a comprehensive, dual-functionality system. It goes beyond merely generating art from text prompts by integrating a critique system that provides actionable feedback on various artistic elements. Furthermore, it offers optional enhancement features and the ability to convert simple doodles into more sophisticated artworks, making it a versatile tool for artists at all levels.

In comparison to existing applications, Art-Vision's uniqueness lies in:

- The combination of **art generation** and **detailed critique**.
- A wider range of art styles, from sketching to realistic to anime.
- Advanced critique systems that analyze not just aesthetics, but emotional tone, light and shadow usage, and originality compared to established art schools.
- Enhancement and doodle-to-art conversion options, which provide more versatility to users.

4.Requirements:

- **Functional Requirements:**

  1-User Authentication & Authorization

  - Users can register and log in securely.
  - Role-based access control (admin, standard user).

  2-Image Upload & Management

  - Users can upload images for critique or creation.
  - Ability to view, delete, and manage uploaded images.
  - Support for multiple image formats (JPEG, PNG, etc.).

  3-AI Image Critique

  - Users can submit images for AI-based critique.
  - AI model provides feedback based on predefined criteria (e.g., composition, lighting, style).
  - Display critique results with detailed analysis.

  4-AI Image Creation

  - Users can generate images based on input parameters (e.g., style, content).
  - Provide options to refine generated images.
  - Display and download created images.

  5-API for AI Model Integration

  - Backend communicates with the AI model to send and receive image data.
  - Handle AI model requests asynchronously for performance optimization.

6-Result History & Reporting

- Users can view their critique and creation history.
- Generate downloadable reports summarizing feedback and generated images.

7-Notifications System

- Notify users when their image critique or creation is complete.
- Email notifications for important actions (e.g., new critiques, report downloads).

8-Search & Filter Capabilities

- Users can search and filter their image critiques and created images.
- Filters based on date, tags, critique score, etc.

9-User Feedback Collection

- Allow Users to rate AI critique results and provide feedback.

10-Admin Dashboard

- Monitor usage statistics and system performance.
- Manage users and their uploaded/generated content.

- **Non-Functional Requirements:**

1-Performance

- System should handle concurrent user requests efficiently.
- AI model response time should be optimized for smooth user experience.

2-Scalability

- Support for growing user base and large image uploads.
- Cloud-based storage and processing for scalability.

3-Security

- Secure authentication and authorization (OAuth, JWT).
- Data encryption (images and user data).
- Protection against common vulnerabilities (SQL injection, XSS, CSRF).

4-Availability

- Ensure system uptime of 99.9% with proper failover mechanisms.
- Implement monitoring and alerting.

5-Maintainability

- Clean and modular codebase for easy updates.
- Comprehensive logging for issue diagnosis.

6-Usability

- Intuitive user interface and smooth workflow.
- Support for responsive design on different devices.

7-Compliance

- Adherence to data privacy regulations (GDPR, CCPA).
- Secure data storage and retrieval policies.

8-Interoperability

- RESTful API support for integrations with other applications.
- Compatible with third-party AI services if needed.

---

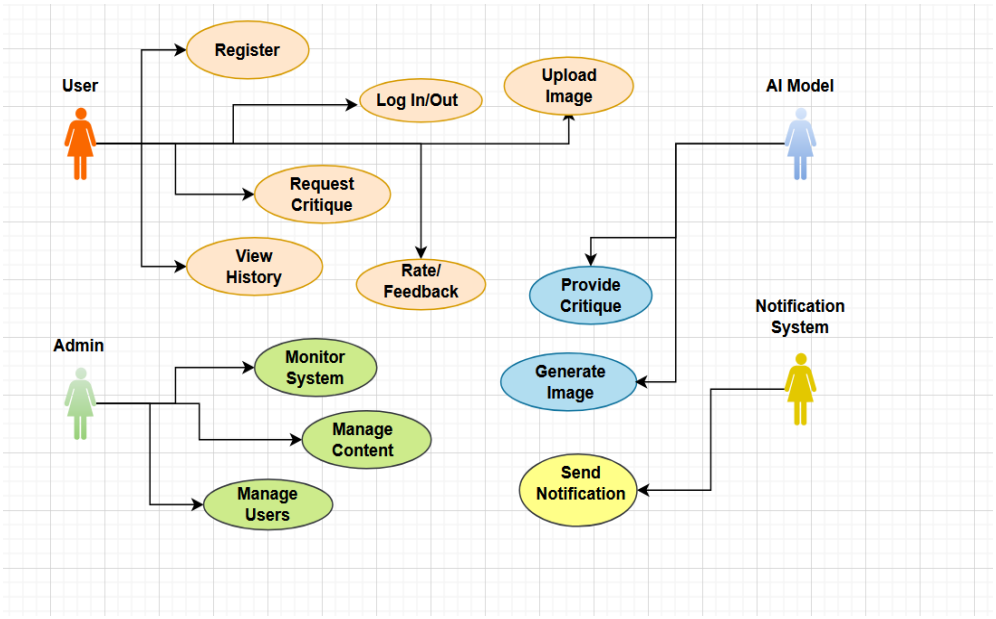
5. UML Diagrams:

- **Use Case (Textual Representation):**

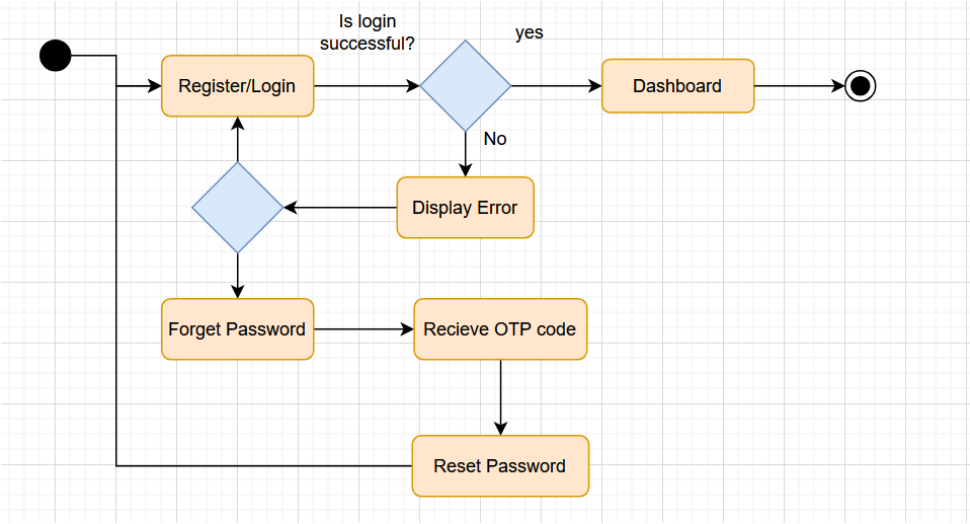
```
+-------------------+       +-------------------+       +-------------------+
|      User         |       |      Admin        |       |    AI Model       |
+-------------------+       +-------------------+       +-------------------+
| Register          |       | Monitor System    |       | Provide Critique  |
| Log In/Out        |       | Manage Users      |       | Generate Image    |
| Upload Image      |       | Manage Content    |       +-------------------+
| Request Critique  |       +-------------------+
| Generate Image    |
| View History      |
| Rate/Feedback     |
+-------------------+

+-------------------+
| Notification Sys. |
+-------------------+
| Send Notification |
```

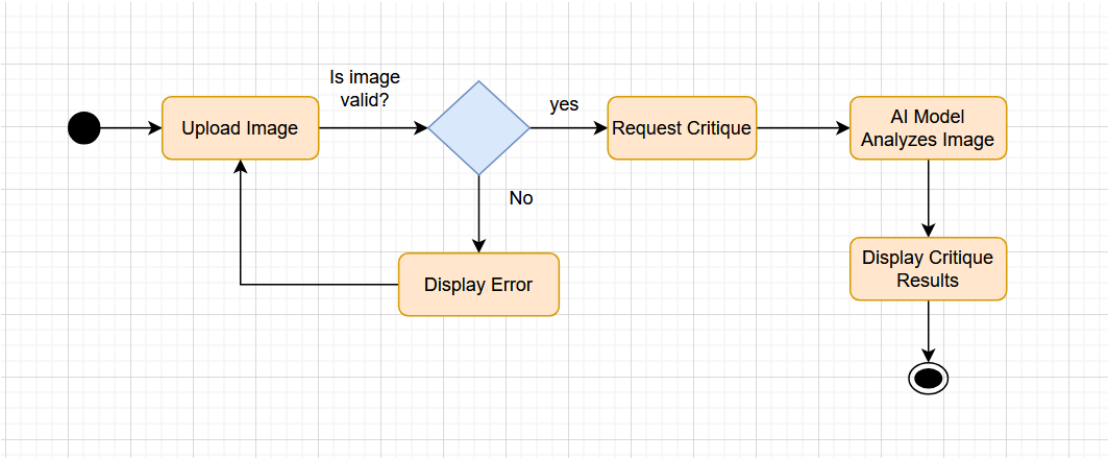
- **Use Case (Visual Representation):**



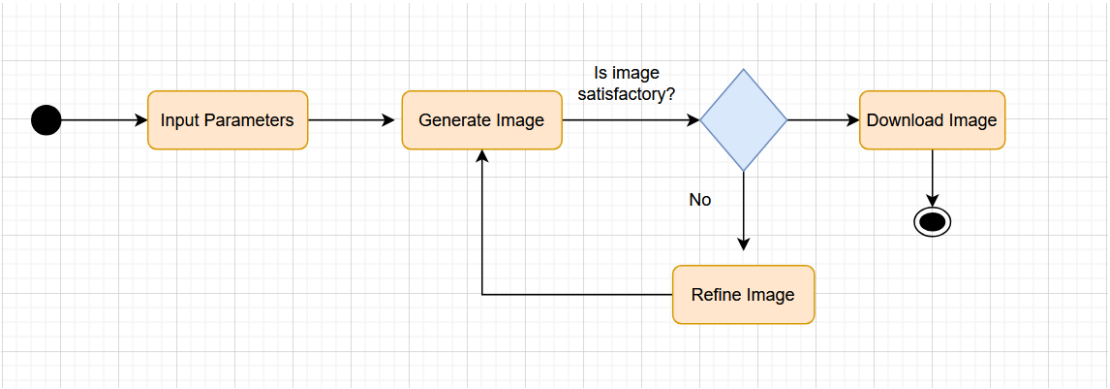
- **Activity diagrams:**
  - **User Authentication Flow:**



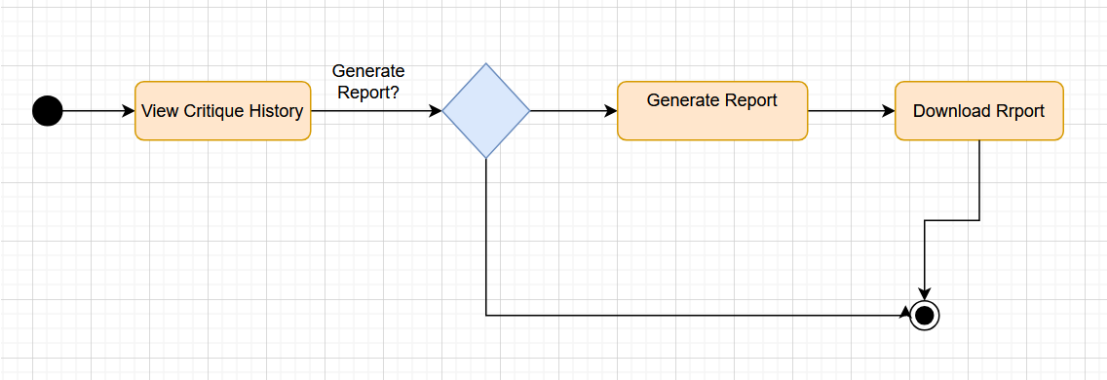
  - **Artwork Critique Flow:**



  - **Artwork Creation Flow:**



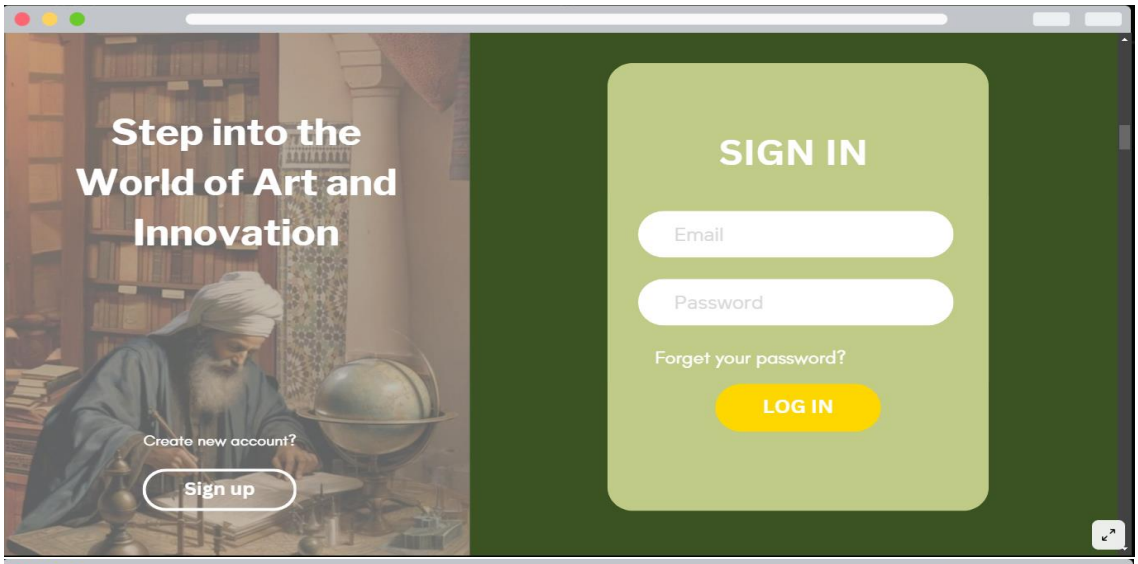
  - **History & Reporting Flow:**

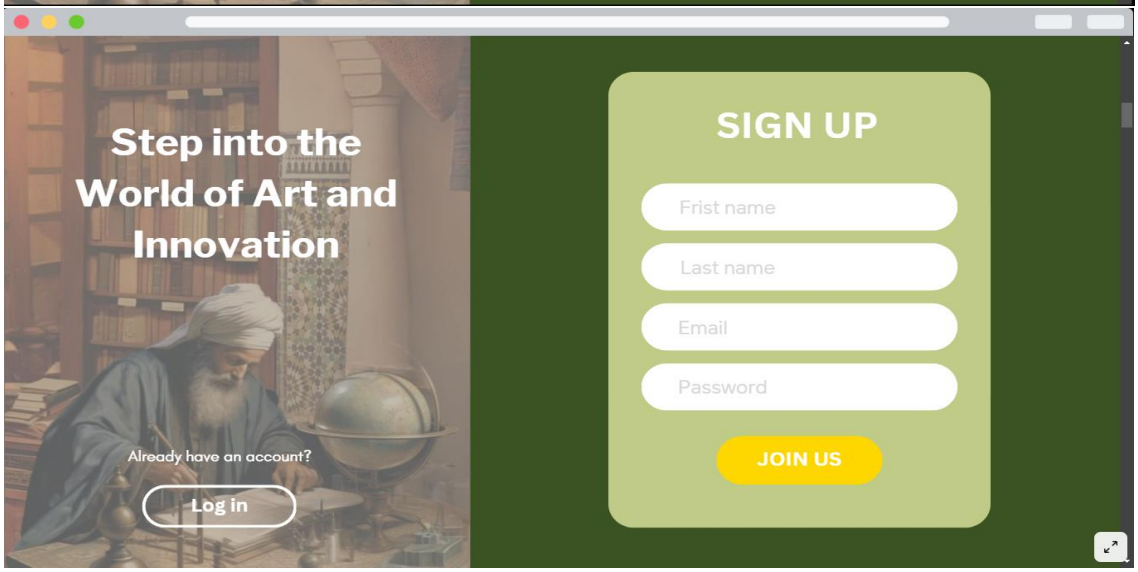


---

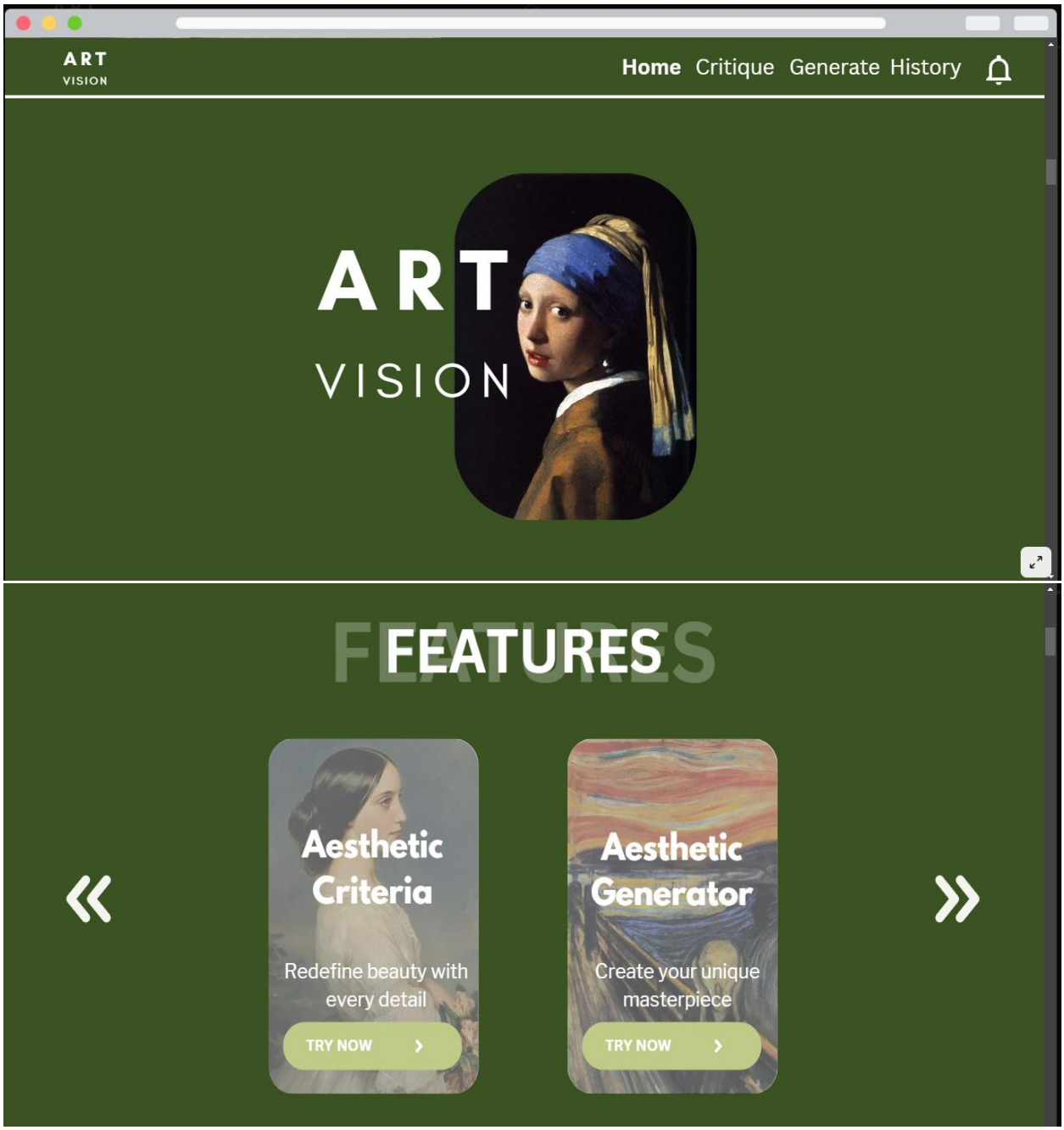
6. User Interface (Canva Link):

- Sign up / Sign In:

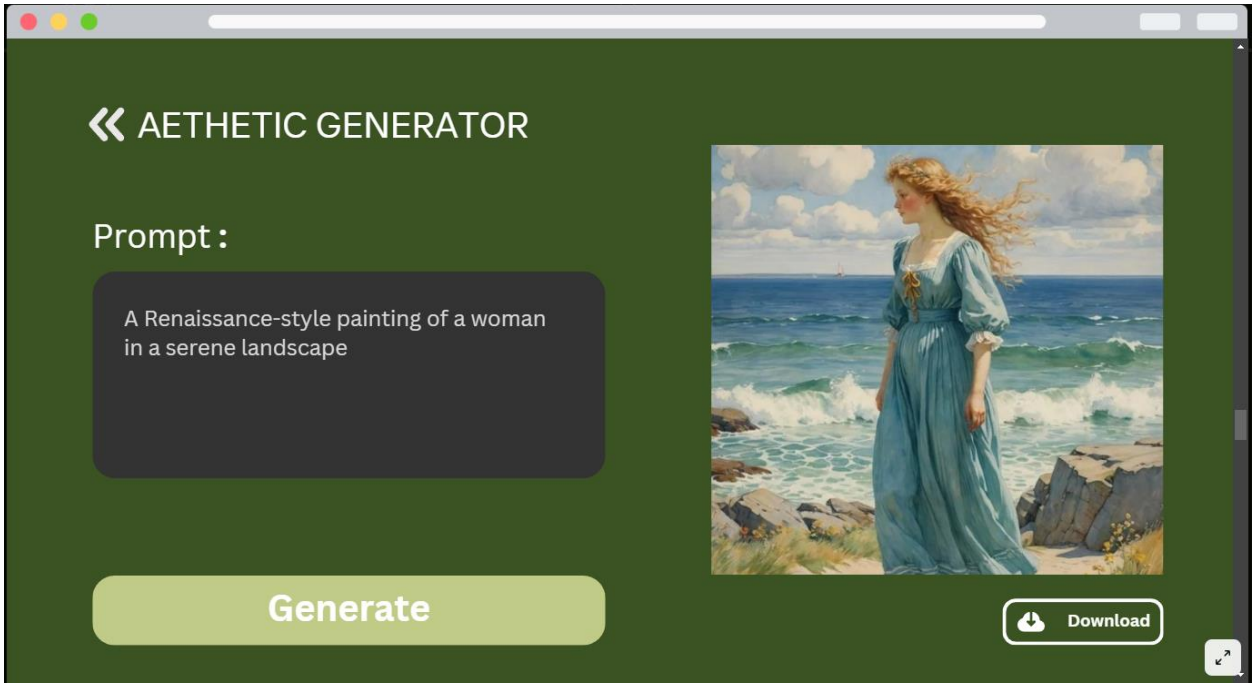




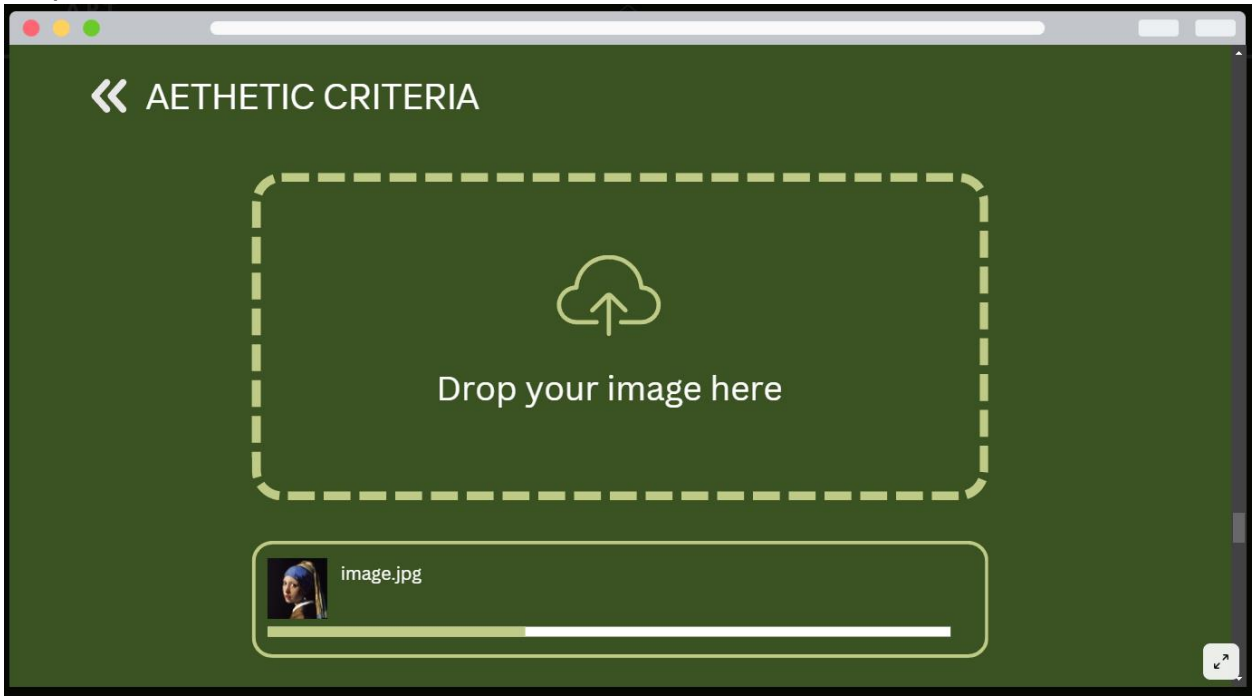
- Landing Page:



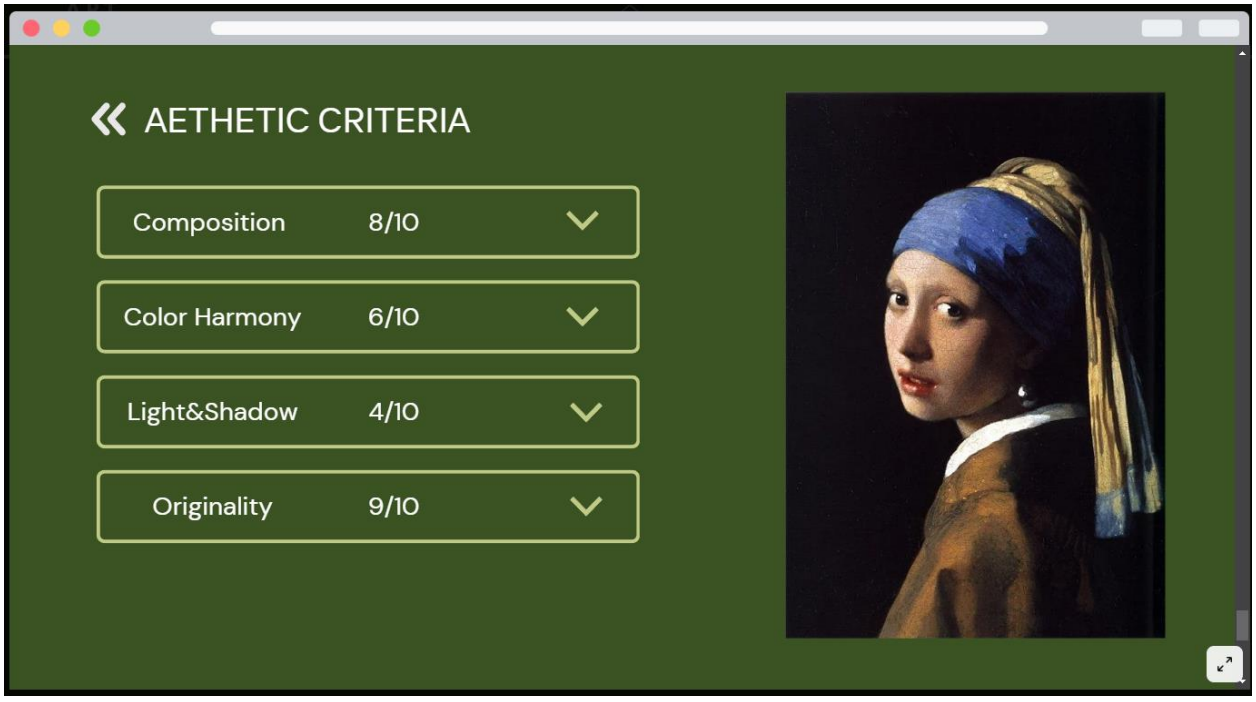
- Generate Artwork:



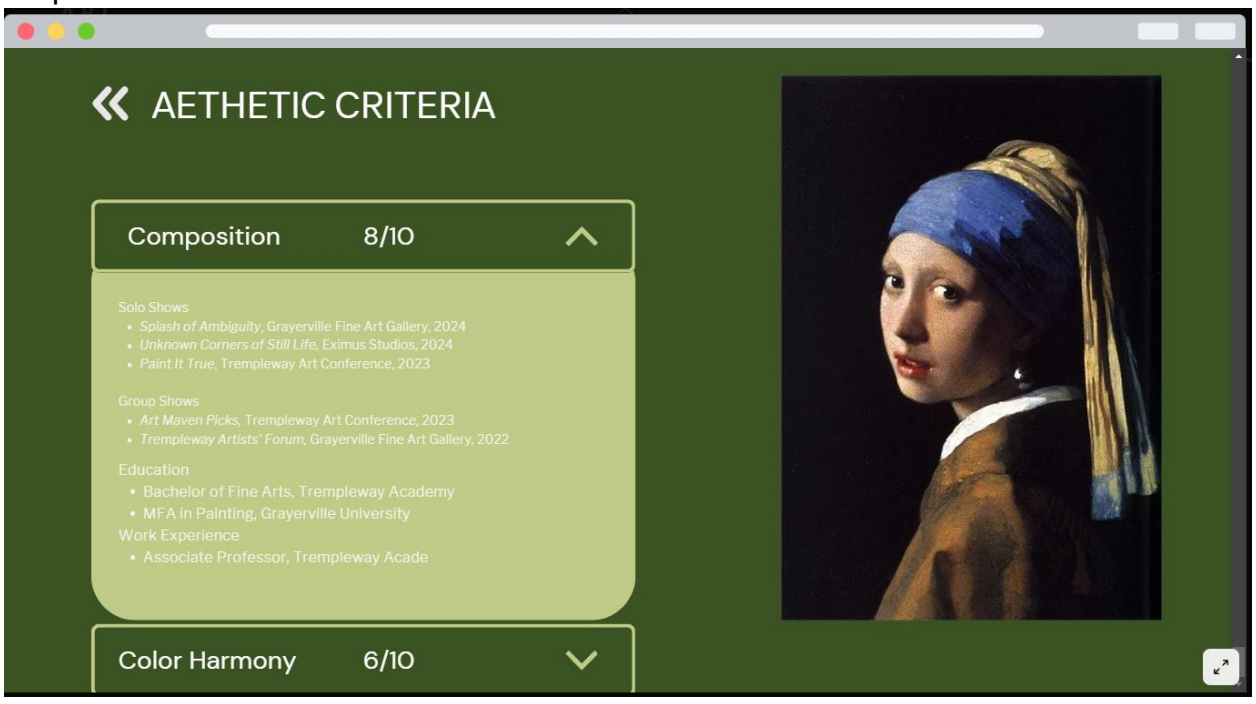
- Critique Artwork:
  Step-1:



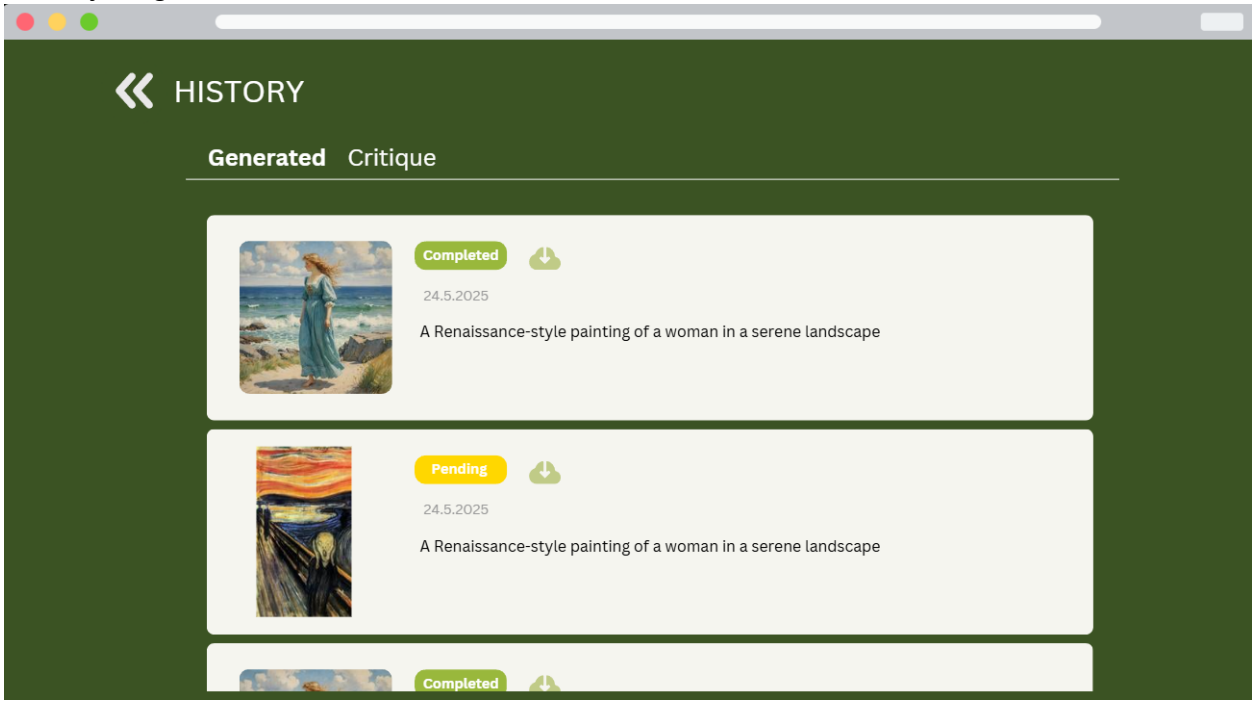
Step-2:


AETHETIC CRITERIA
Composition 8/10
Color Harmony 6/10
Light&Shadow 4/10
Originality 9/10

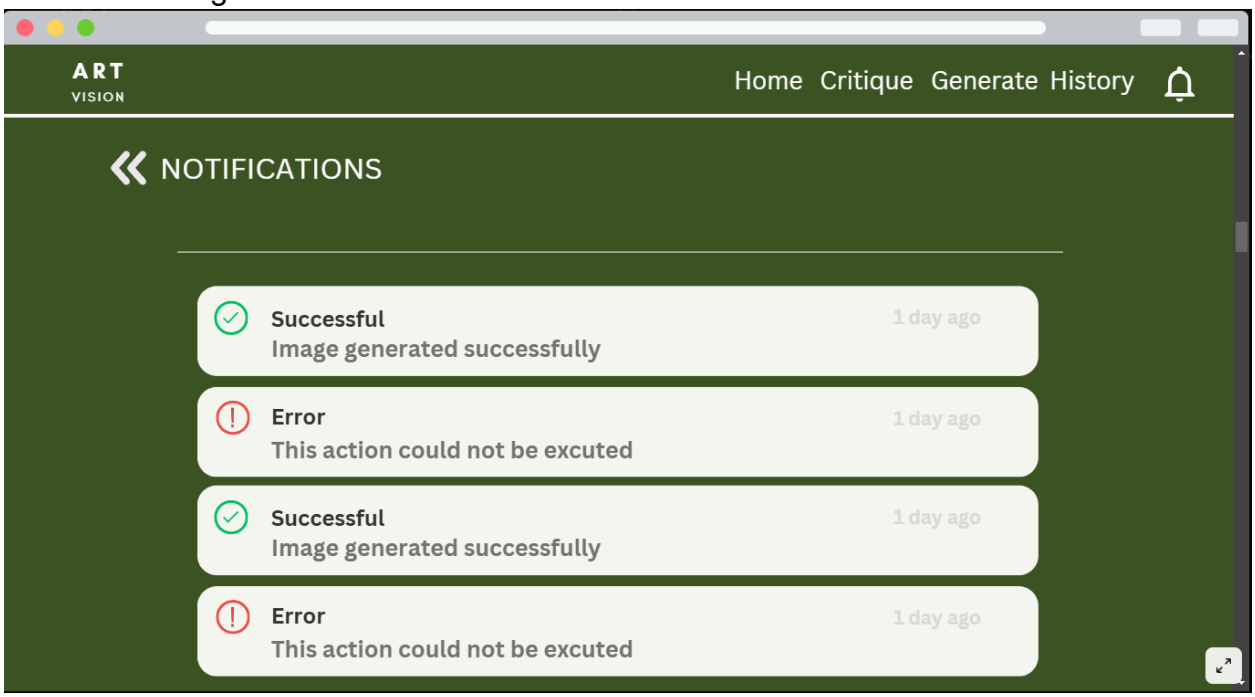
Step-3:


AETHETIC CRITERIA
Composition 8/10
Solo Shows
Splash of Ambiguity, Grayerville Fine Art Gallery, 2024
Unknown Corners of Still Life, Eximus Studios, 2024
Paint It True, Trempleway Art Conference, 2023
Group Shows
Art Maven Picks, Trempleway Art Conference, 2023
Trempleway Artists' Forum, Grayerville Fine Art Gallery, 2022
Education
Bachelor of Fine Arts, Trempleway Academy
MFA in Painting, Grayerville University
Work Experience
Associate Professor, Trempleway Acade
Color Harmony 6/10

- History Page:



- Notification Page:



7. Existing Research Papers:

- **The following table overview art critique papers:**

Paper Link	Publish year	Author	Dataset	Accuracy	pros	cons
APDDv2: Aesthetics of Paintings and Drawings Dataset with Artist Labeled Scores and Comments	2024	Xin Jin, Qianqian Qiao, Yi Lu, Huaye Wang , Heng Huang, Shan Gao, Jianfei Liu, and Rui Li	Dataset encompassing 24 distinct artistic categories and 10 aesthetic attributes. Comprising 10,023 painting images, 85,191 Scoring labels, and 6,249 linguistic comments. Dataset's link.	The ArtCLIP model outperformed other models (AANSPS and SAAN) in aesthetic evaluation metrics. **Total Aesthetic Score (TAS)**: ACC (Accuracy) = **0.89**.	**APDDv2** is the first multimodal dataset featuring detailed aesthetic language commentary, with over 10,000 images . Learns rich aesthetic concepts from categorized multimodal databases Combines text and image embeddings for better multimodal understanding	**Dataset Limitations:** -Focused on expert opinions, excluding broader public aesthetic perspectives - Underrepresented art genres. -May require refinement to adapt to subjective aesthetic judgments across different cultures
Towards Artistic Image Aesthetics Assessment: a Large-scale Dataset and a New Method	2023	Ran Yi, Haoyuan Tian , Zhihao Gu , Yu-Kun Lai , Paul L. Rosin	BAID dataset consists of 60,337 artistic images covering various art forms annotated with scores. Dataset's link.	SAAN model outperforms existing IAA methods according to quantitative comparisons. **Total Aesthetic Score (TAS)**: (Accuracy) = 76.80%	**BAID** is the largest artistic image aesthetic assessment dataset and far exceeds existing IAA and AIAA datasets in quantity and quality of artworks.	**SAAN** model involves a complex neural network architecture (e.g., combining style-specific and generic aesthetic features) which increases computational costs and makes the model harder to deploy

ArtEmis: Affective Language for Visual Art.	2021	Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J. Guibas.	Dataset contains 455K emotion attributions and explanations on 80K artworks from WikiArt dataset. Dataset's link	Quantitative accuracy not directly applicable due to task subjectivity	**Emotion-Centric**: ArtEmis focuses on subjective, emotional, and abstract descriptions **Linguistic Diversity**: The dataset captures a wide range of vocabulary and linguistic structures	**Limited Scope of Emotions**: The dataset uses a predefined set of eight emotions plus a "something-else" category
Explain Me the Painting: Multi-Topic	2021	Zechen Bai, Yuta Nakashima and Noa Garcia	Dataset contains a 17,249 images of the SemArt dataset annotated with 33,543 sentences about the content, form and context. Dataset's link.	The proposed framework with Parallel decoder model appears to perform the best overall based on these metrics: GM:77.6 S-T:30.9 EA:92.6	**Multi-Topic Descriptions**: The framework effectively generates descriptions covering artistic style, composition, historical context, and artist background, offering a holistic understanding of the artwork.	**Complexity of Implementation**: Integrating multiple topics with external knowledge and ensuring coherence in generated text requires sophisticated modeling and computational resources.
Photo Aesthetics Ranking Network with Attributes and Content Adaptation.	2016	Shu Kong, Xiaohui She, Zhe Lin, Radomir Mech and Charless Fowlkes.	Dataset contains aesthetic scores and meaningful attributes assigned to each image. Dataset's link	Model achieves a classification accuracy of **77.3%.**	**State-of-the-Art Performance**: Achieved superior classification accuracy on the AVA dataset and a high correlation score for aesthetic ranking	**Complexity**: The multi-branch network design and content-adaptive features add complexity to training and inference
Large-scale Classification of Fine-Art Paintings	2015	Babak Saleh, Ahmed Elgammal	Dataset containing 81,444 pieces of visual art labeled with artist, style and genre.	Model achieves a classification accuracy of 45.97 %	**Comprehensive Feature Analysis**: evaluates various visual features like SIFT, HOG, and CNN-based features for	**Potential Overfitting**: focused metric learning may overfit to specific datasets or styles, reducing generalization across less

			Dataset's link		effective painting representation.	common art forms or unseen styles.
AVA: A Large-Scale Database for Aesthetic Visual Analysis.	2012	Naila Murray and Luca Marchesotti	AVA dataset contains over 250,000 images with aesthetic scores for each image. Dataset's link	achieved classification accuracies superior to **90%** on this dataset.	**Rich Annotations**: includes multiple ratings per image, alongside photographic attributes,offering a nuanced resource for aesthetic models	**No Contextual Metadata**: lacks contextual metadata such as the purpose of the image or the photographer's intent, limiting the interpretability of the ratings
Pandora: Description of a Painting Database for Art Movement Recognition with Baselines and Perspectives	2016	Corneliu Florea, Razvan Condorovici , Constantin Vertan, Raluca Boia, Laura Florea and Ruxandra Vrânceanu	Dataset contains 18,038 images ,18 art movements Classes and is designed for studying the recognition of art movements. Dataset's link.	The best performance achieved was **54.7%** using a combination of pyramidal Local Binary Patterns and Color Structure Descriptor with a Support Vector Machine (SVM) classifier	Using a **fixed evaluation protocol** with 4-fold cross-validation, reducing overfitting issues present in smaller datasets. Introduction of a significantly larger and more diverse art database covering wide time periods and diverse styles.	**Lower accuracy** compared to smaller, less diverse datasets. Challenges in separating closely related styles. Database size deemed **too small for effective deep learning**, as CNN.
DeepArt: Learning Joint Representations of Visual Arts.	2017	Hui Mao, Ming Cheung, and James She.	Dataset contains 500K images, each with 10 labels: artist, genre, art movement, event, historical figure and description. Dataset's link	Quantitative accuracy not directly applicable due to task subjectivity	**Dual Feature Representation**: The model integrates both content and style of visual art through a unified framework using VGG-16 for content extraction and Gram matrices for style profiling.	**Computational Intensity**: Using deep learning architectures like VGG-16 and triplet-based ranking requires significant computational resources.

Art Critique Models:

- **ArtCLIP:**

- The model is pre-trained on the **DPC2022** dataset (contains 510K images with over 5 million comments and 350K aesthetic scores).
- The pre-training phase focuses on attribute-aware learning, where the model learns to associate images with their aesthetic attributes.
- After pre-training, the model is fine-tuned on the **APDDv2** dataset.
- During fine-tuning, the model's image encoder extracts feature vectors, which are then passed through an MLP (Multi-Layer Perceptron) network for score regression.

Architecture:

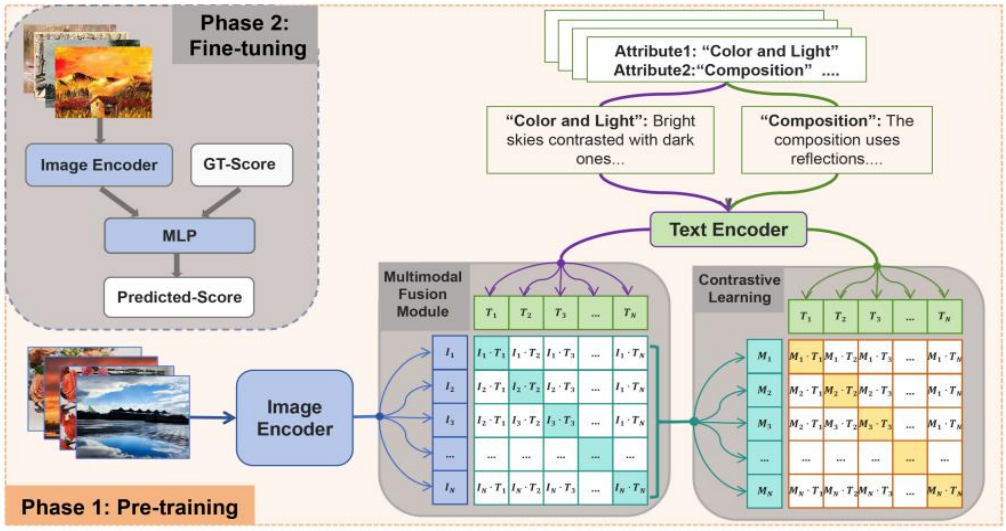


Figure 9: ArtCLIP samples two comments from different aesthetic attribute perspectives and further introduces a multimodal fusion module to integrate text and image embeddings, thereby generating multimodal embeddings for contrastive learning.

- **SAAN(Style-specific Art Assessment Network):**
  - A **VGG-19** model pretrained on **ImageNet dataset** is used to extract style features from the input image. These features represent the artistic style of the image, such as brush strokes, texture, and color distribution
  - A **ResNet-50** model pretrained with self-supervised learning is used to extract aesthetic Features
  - Style and aesthetic features are combined using an **AdaIN** layer, the output of AdaIN is a **style-specific aesthetic feature**, which combines the aesthetic qualities of the image with its artistic style
  - A **ResNet-50** model is pretrained on the **BAID dataset** with distortion classification and intensity estimation tasks used to extract universal aesthetic features.
  - A **Non-local blocks** are designed To capture the spatial relationships and global context of the artwork and then the fused features are passed through an MLP to predict the aesthetic score.

Architecture:

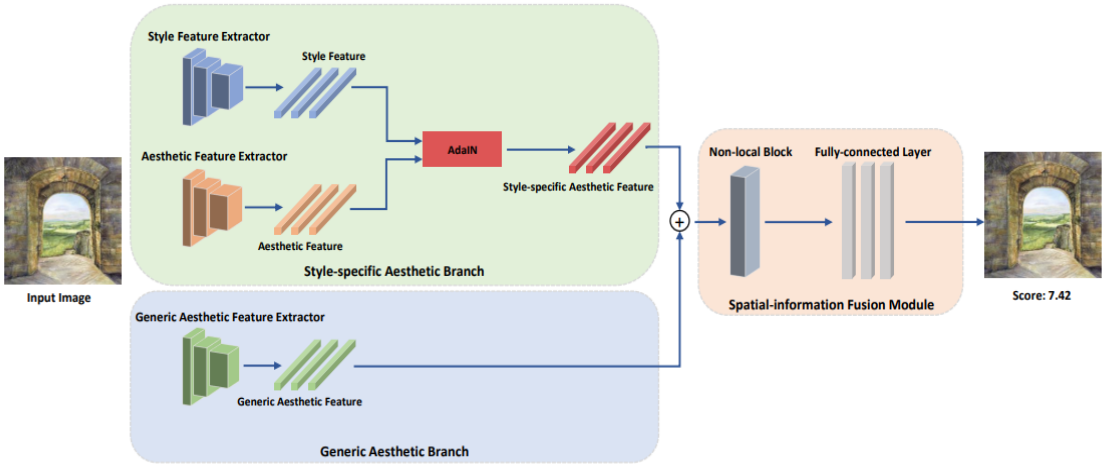
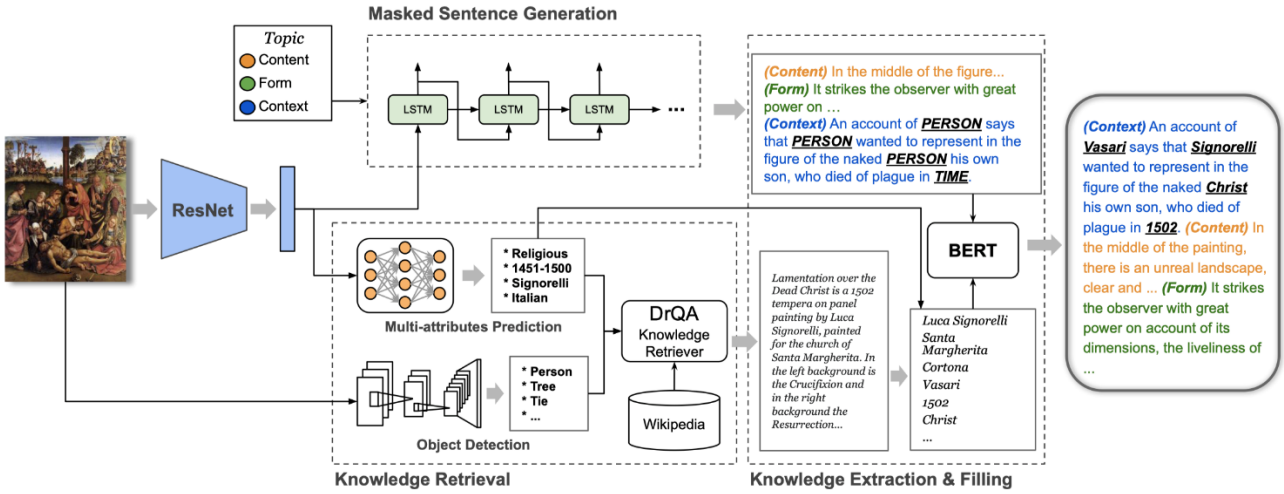


Figure 4. Overall architecture of the proposed SAAN. SAAN consists of three modules: 1) a style-specific branch to extract style-specific features; 2) a generic aesthetic branch to extract generic aesthetic features; and 3) a spatial information fusion module that fuses the spatial information using a non-local block and considers the composition of artwork during the assessment. See Sec. 4 for further details.

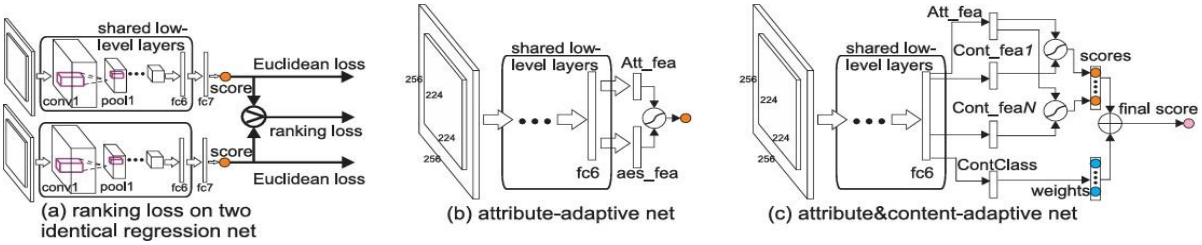
- **Multi-topic and knowledgeable art description framework:**
  - A pre-trained **ResNet** on ImageNet dataset is used to extract the visual features from the input artwork.
  - Then a topic decoder is used to generate multi-topic masked sentences that describe the painting from multiple aspects.
  - The average-pooling vector is computed as the global visual feature for multi-attribute prediction.
  - object detector is employed to detect visual concepts.
  - The predicted attributes and the detected objects are combined to form a query for **DrQA**
  - **DrQA** is used to retrieve relevant articles from an external knowledge base (e.g., Wikipedia) based on the query. The top-5 articles with the highest similarity scores are selected.
  - Use Stanford **CoreNLP** to extract named entities from the top-5 articles
  - Combine the extracted named entities and artistic attributes to form a set of candidate words G.
  - Train a **BERT**-based model to fill the masked concepts in the generated sentences with appropriate words from G.
  - The model is trained using a sequence-to-sequence approach, where the input sequence includes the masked description and the candidate words, and the output is the filled description.



- **Photo Aesthetics Ranking Network:**
  - **Base Network (AlexNet Backbone):** The backbone of all the models is AlexNet, which serves as the feature extractor to extract meaningful image features. This involves:
    - Shared low-level convolutional layers (Conv1 to Conv5).
    - Fully connected layers (Fc6 and Fc7) for feature representation.
  - **Regression Network (Baseline Model)**
    - The first architecture replaces the softmax layer of AlexNet with a regression network.
    - This network predicts an aesthetic score using a Euclidean loss function to minimize the difference between predicted scores and ground-truth ratings.
  - **Ranking Loss Network**
    - In this variant, a pairwise ranking loss is added in addition to the regression loss.
    - Key components: Two identical regression networks (sharing weights).
  - **Sampling strategies:**
    - Within-rater: Compares images rated by the same person to capture individual preferences.
    - Cross-rater: Compares images rated by different raters for general aesthetics.
    - Combining both strategies improves ranking performance.
    - This variation is often called Reg+Rank in the paper.
  - **Attribute-Adaptive Network:**
    - Adds an attribute prediction branch to the regression network.
    - Key components: Attribute features (Att_fea): Extracted from the fc6 layer and passed to the attribute branch.
    - This variation is referred to as Reg+Rank+Att
  - **Content-Adaptive Branches:**
    - Groups images into semantic categories (content clusters) using unsupervised K-means clustering (k=10).

- Cluster information is integrated into the model to refine predictions.
- The content-specific model variation is called Reg+Rank+Cont.

- **Final Unified Model:**
  - The most comprehensive version combines regression, ranking loss, attribute prediction, and content adaptation.
  - This is the Reg+Rank+Att+Cont model.
- **AADB Dataset:**
  - Contains 10,000 images with:
    - Aesthetic ratings.
    - Attribute ratings (e.g., light, color, composition).
    - Anonymized rater identities for within-rater comparisons.
  - Dataset split:
    - Training: 8,500 images.
    - Validation: 500 images.
    - Testing: 1,000 images.
  - Content Groups:
    - Unsupervised K-means clustering groups training images into 10 clusters using features from AlexNet's fc7 layer.
    - Softmax is applied to transform distances between a test image and cluster centroids into prediction weights.
  - Use in the Paper: Key dataset for training the model with attribute supervision and ranking loss.
- **AVA Dataset:**
  - Contains ~250,000 images, each rated on a scale of 1–10 by ~200 users.
  - Dataset split:
    - Training: 230,000 images.
    - Testing: 20,000 images.
  - Images are divided into low (score ≤ 5) and high (score > 5) aesthetic categories for binary classification tasks.
  - Use in the Paper:
    - Demonstrates the model's generalization ability to a larger and more challenging dataset.
    - Validates the approach's effectiveness for binary classification and ranking on professional photos.



- **LeNet-5 Model and Feature Extraction Techniques in Pandora Paper:**
  - **Texture Feature Extractors:**
    - These techniques focus on identifying patterns in the texture. Texture features are useful for recognizing patterns like brush strokes, texture styles, and other surface-level properties of the painting.

- **Histogram of Oriented Gradients (HOG)**:
  - **How it works**: HOG computes the gradients (changes in pixel intensity) of an image and accumulates them into a histogram. The gradient is taken in different orientations, and the histogram helps describe the spatial distribution of these gradients in a region of the image.
  - **Usage in painting analysis**: It's useful for detecting the structure of edges in the painting, helping to distinguish between different artistic styles.
- **Pyramidal HOG (pHOG)**:
  - **How it works**: This is an extension of HOG applied to multiple levels of a Gaussian pyramid (an image representation that captures different scales of detail). At each level, the HOG descriptor is computed, which helps capture texture at different scales.
  - **Usage**: This provides a more detailed description of the texture and structure in the image at varying resolutions, useful for recognizing more subtle or large-scale features in paintings.
- **Color HOG**:
  - **How it works**: This variant applies the HOG method separately to each color channel of an image (RGB). This helps capture the texture information in the context of specific color distributions.
  - **Usage**: Useful when color plays a significant role in texture or style, distinguishing paintings with unique color patterns.
- **Local Binary Pattern (LBP)**:
  - **How it works**: LBP creates a histogram of patterns formed by comparing the intensity of each pixel in a local neighborhood (typically a 3x3 pixel area) and encoding it as a binary value. It generates a set of quantized patterns based on local image textures.
  - **Usage**: LBP is often used to identify repeating structures, such as regular brush strokes or textured surfaces, and has been applied to painting analysis.
- **Pyramidal LBP (pLBP)**:
  - **How it works**: Similar to pHOG, pLBP computes the LBP descriptor over multiple levels of a Gaussian pyramid, which allows it to capture texture features at different scales.
  - **Usage**: This improves the detection of textural features that span a wide range of scales in the painting.
- **Local Invariant Order Pattern (LIOP)**:
  - **How it works**: LIOP focuses on the order of pixel intensities in a local neighborhood. After sorting the intensities in increasing order, it creates a pattern based on the relative ordering of these values.
  - **Usage**: Useful for capturing the relative arrangement of intensities within a region, which can be helpful for distinguishing certain artistic techniques or brushwork.
- **Edge Histogram Descriptor (EHD)**:
  - **How it works**: This descriptor is part of the MPEG-7 standard, which divides the image into regions and computes the distribution of four basic

gradient orientations within each region. It helps describe the edge structure of the image.

- **Usage**: EHD is useful for identifying patterns in edges and shapes within the painting.

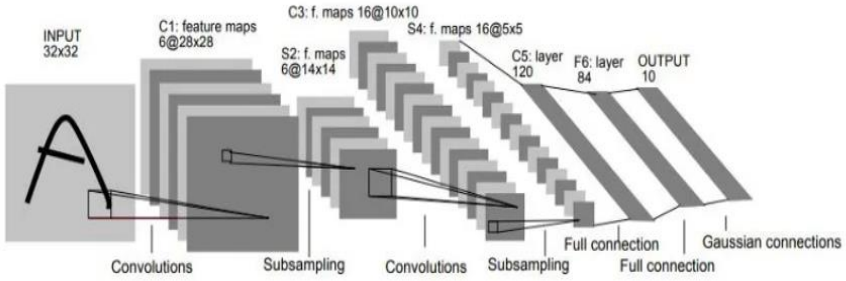
- **Spatial Envelope (GIST)**:
  - **How it works**: The GIST descriptor captures the overall spatial layout of the image, which can describe global properties like the composition of the painting and its general structure (e.g., the arrangement of objects, large shapes, and space).
  - **Usage**: GIST is often used to **characterize the global composition or shape** of a scene in a painting.

- **Color Descriptors:**
  - These techniques focus on capturing the color properties of the painting, which are often distinctive between different artistic movements, artists, or periods.
- **Discriminative Color Names (DCN)**:
  - **How it works**: DCN represents dominant colors in an image using an information-oriented approach. It extracts key color names (like "red," "blue," "green") and their prevalence in the image.
  - **Usage**: By analyzing the distribution of color names, this technique helps identify the overall color scheme of the painting, which can reveal its style or painter.

- **Color Structure Descriptor (CSD)**:
  - **How it works**: CSD generalizes the color histogram by considering the spatial relationships of colors. It captures the distribution of quantized colors and their spatial coherence in the painting.
  - **Usage**: CSD is particularly useful for differentiating between different art movements, as various styles tend to use color differently (e.g., impressionism vs. realism).
- **Features of LeNet-5:**
  - **Every convolutional layer includes three parts: convolution, pooling, and nonlinear activation functions**
  - **Using convolution to extract spatial features (Convolution was called receptive fields originally)**
  - **The average pooling layer is used for subsampling.**
  - **'tanh' is used as the activation function**
  - **Using Multi-Layered Perceptron or Fully Connected Layers as the last classifier**
  - **The sparse connection between layers reduces the complexity of computation**

Architecture:

The LeNet-5 CNN architecture has seven layers. Three convolutional layers, two subsampling layers, and two fully linked layers make up the layer composition.

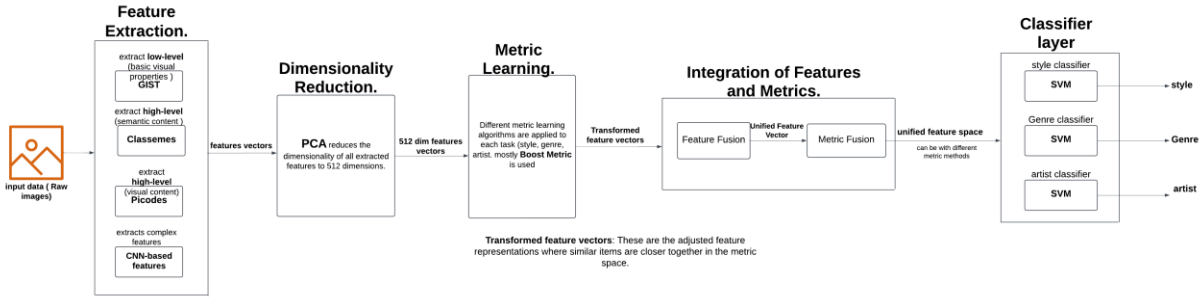


LeNet-5 Architecture

- **Large-scale Classification of Fine-Art Painting Paper:**
  - **The model is designed to:**
    - **Predict the Style**: Identify the artistic movement or period (e.g., Impressionism, High Renaissance).
    - **Predict the Genre**: Determine the subject matter of the painting (e.g., Portrait, Landscape, Still Life).
    - **Predict the Artist**: Attribute the painting to a specific artist (e.g., Leonardo da Vinci, Vincent van Gogh).
  - **The dataset, "Wikiart Paintings," consists of:**

    81,449 images from **1,119 artists** spanning **27 styles** and **45 genres**. Data splits for the tasks are as follows:

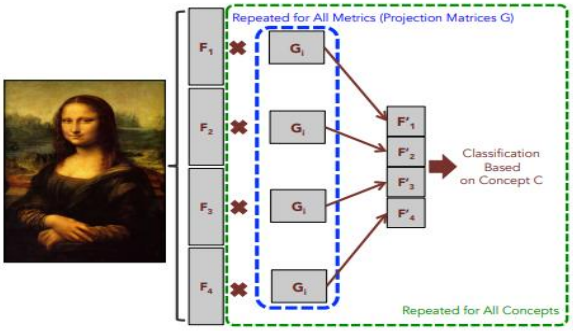
    - **Style Classification**: 27 styles, each with at least 1,500 paintings.
    - **Genre Classification**: 10 genres, each with at least 1,500 paintings.
    - **Artist Classification**: 23 artists, each with at least 500 paintings.
  - **Workflow:**



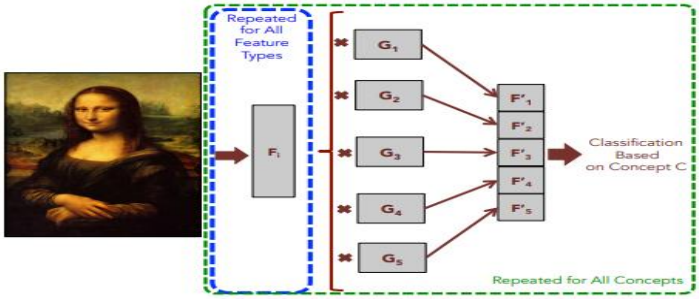
  - **Feature Extraction:**
    - The features are divided into **low-level** (basic visual properties) and **high-level**(semantic content (e.g., objects or concepts).
    - Low-Level Features: **GIST** descriptors.

- High-Level Features: **Classemes**, **Picodes**, **CNN-based features**.
- CNNs have four convolutional layers followed by three fully connected layers, and use the last layer of a pre-trained CNN as another feature vector.

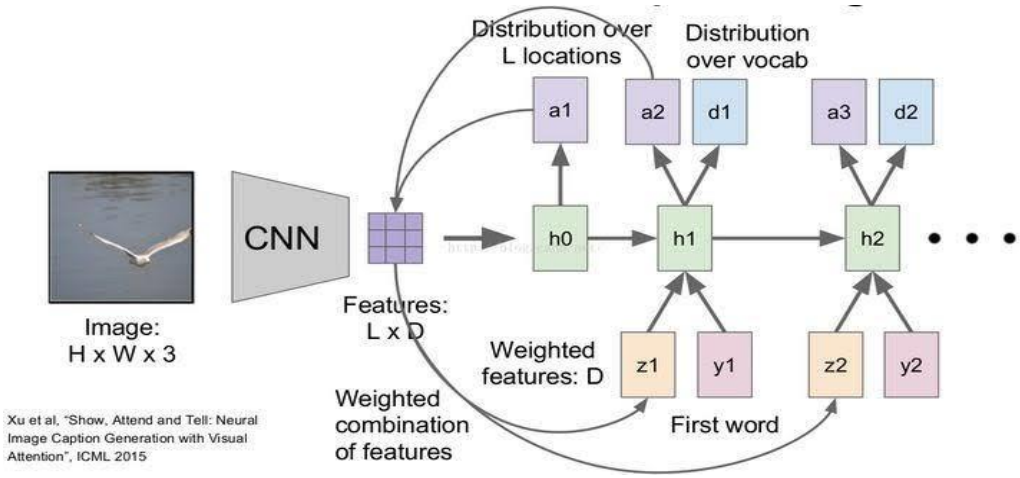
- **Dimensionality Reduction:**
  - The paper uses **Principal Component Analysis (PCA)** to reduce the dimensionality of all features to a uniform size of **512 dimensions** for computational efficiency and faster metric learning.
- **Metric Learning:**
  - Different metric learning approaches are applied to adjust the extracted features for each prediction task (style, genre, artist):
    - **Boost Metric**: always gives the best or the second best results for all classification tasks.
    - **ITML (Information-Theoretic Metric Learning).**
    - **LMNN (Large-Margin Nearest Neighbors)**
    - **NCA (Neighborhood Component Analysis)** and **MLKR (Metric Learning for Kernel Regression)** approaches are performing worse than other metrics.
- **Integration of Features and Metrics:**
  - **Feature Fusion:** Combines multiple feature types (e.g., GIST, Classemes, CNN) into a single feature vector.



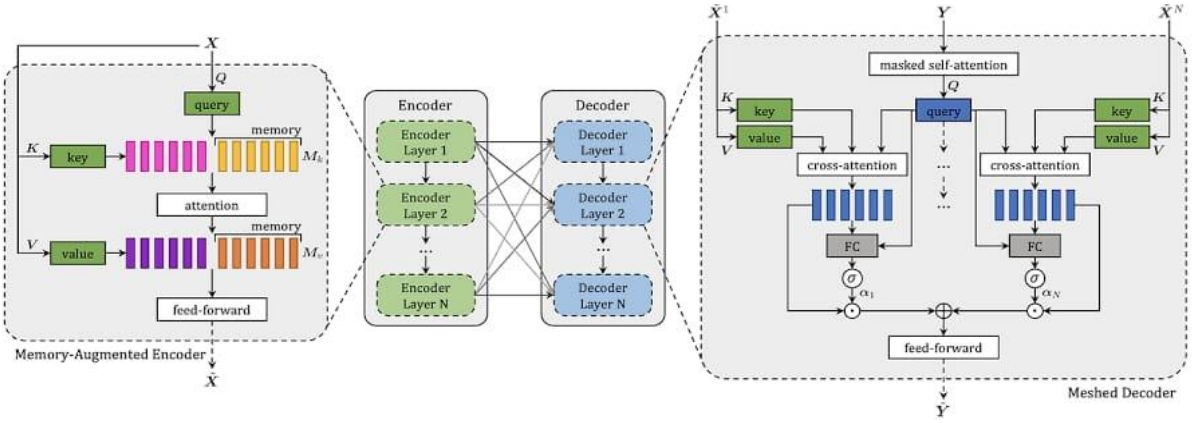
  - **Metric Fusion:** Combines multiple task-specific projections (e.g., different metric learning results) into a unified feature representation.
  - results in a composite feature vector that is richer and more informative, which is then used as input for training the classifiers.



        - **After the fusion process, Support Vector Machines (SVMs) are trained on the fused feature vector for the respective task (style, genre, artist).**
    - **ArtEmis Dataset :**
        - **The ArtEmis dataset is built on top of the publicly available WikiArt1 dataset, It's consisting of 454,684 paintings, emotional explanatory utterances and emotional responses + OLA is dataset of 5,000 objective captions (used in ANP model)**
    - **ANP + KNN Models**
        - ANP:
            - Created without fully leveraging the ArtEmis dataset
            - This baseline generates emotionally expressive captions in a simpler way by modifying captions from a different dataset
            - This baseline is simpler and less sophisticated than a model directly trained with ArtEmis
        - KNN :
            - This is a non-neural, nearest-neighbor approach that uses ArtEmis captions in a straightforward way
            - Since this approach doesn't generate captions but rather reuses existing ones, its limitations highlight the benefits of learning-based methods.
        - **Basic ArtEmis speakers + Emotion grounded speakers**
            - **Emotion grounded speakers :**
                - An emotion classifier is trained separately to predict the emotion label for an image Cemotion|image .
                - This means the speaker can generate captions that emphasize a specific emotion, even if the emotion is not directly tied to the image's visual characteristics.
            - Use a pretrained CNN (ResNet-32) to extract features and Add a fully connected layer for emotion classification.
    - **SAT model:**
      This is a classic image captioning architecture that combines:
      Image Encoder: Converts the image into a feature representation (e.g., using a convolutional neural network like ResNet or VGG).
      Attention Mechanism: Focuses on specific parts of the image while generating each word of the caption.
      LSTM (Long Short-Term Memory): A type of recurrent neural network (RNN) that generates captions word by word, considering both the current image features and previously generated words.
        - **Emotion grounded**
          Pass the emotion label (from the classifier or dataset) through a fully connected layer to obtain an emotion embedding.
          LSTM Decoder: Input the image features, attention-weighted features, and emotion embedding into the LSTM at every time step.



- **M2 model :**
  has three components:
  Image Encoder: Encodes visual features from paintings.Use Faster R-CNN for object-level features (bounding boxes and corresponding embeddings). Optionally, use ResNet for global image features.
  Meshed-Memory Mechanism: A transformer-based architecture integrates image features and previous text context.The "meshed memory" allows multi-level interactions between the image and generated text.
  Text Decoder: A transformer-based decoder generates the emotional description word-by-word based on the image features and previously generated words.
  - **Emotion grounded**
    Convert emotion labels into embeddings or one-hot vectors.
    Add the encoded emotion features as an input to the meshed-memory layers.
    Inject emotion features into the decoder at each step.



- **Text-to-art generation papers:**

Paper Link	Publish year	Author	Dataset	Accuracy	pros	cons
"Parti: Pathways Autoregressive Text-to-Image"	2022	Google AI	LAION-400M MS-COCO Localized Narratives PartiPrompts (P2)	people prefer 20B: 63.2% for image realism/quality 75.9% for image-text match	• High-quality outputs • adaptable to complex prompts • scalable • robust reconstruction.	• Computationally expensive • slower generation • vulnerable to dataset biases
"Artistic Image Generation Using GANs"	2018	Saikat Basu, Xiaodan Liang, and others	WikiArt CIFAR-10	Demonstrated improved performance in generating high-quality artwork and natural images compared to prior models	• Supports conditional image synthesis • better stability in training • enhanced diversity in outputs	• Limited resolution of generated images • challenges in scaling to very large datasets
Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding	2022	Aiden N. et al.	LAION-400M ms-coco	Outperforms GLIDE, DALL-E 2, and Make-A-Scene in image realism and text alignment	• Generates highly realistic images from detailed text descriptions. • Strong text understanding for complex prompts. • Versatile applications in creative industries.	• High computational cost. • Reliant on dataset quality, with potential biases.
CLIPDraw: Exploring Text-to-Drawing Synthesisthroug	2021	Kevin Frans L.B. Soros Olaf Witkowski	this dataset are not publicly disclosed	Not mentioned	• Innovative Use of CLIP better	• Innovative Use of CLIP better • Computational Intensity

h Language-Image Encoders					• Abstraction-Friendly Approach • Optimization-Based Framework	• Limited Dataset Transparency • Evaluation Challenges
VQGAN-CLIP: Open Domain Image Generationand Editing with Natural Language Guidance	2022	Katherine Crowson Stella Biderman Daniel Kornis Dashiell Stander Eric Hallahan Louis Castricato Edward Raff	ImageNet	Not mentioned	• Intuitive Interaction • High-Quality Outputs • Combines Strengths of Two Models	• Computational Costs • Ambiguity in Prompts
CogView: Mastering Text-to-Image Generation viaTransformers	2021	Ming Ding, Zhuoyi Yang Wenyi Hong Wendi Zheng Chang Zhou Da Yin Junyang Lin Xu Zou Zhou Shao Hongxia Yang Jie Tang	MS-COCO	CogView sometimes outperforms DALL-E on tasks requiring detailed semantic understanding but may be slightly weaker in fine-grained visual details or style transfer.	• High-Quality Text-to-Image Generation • Transformer-Based Innovation • Use of Discrete VAE (dVAE) • Competitive with DALL-E	• Computational Resource Requirements • Resolution Limitation • Limited Generalization for Rare Concepts
Hierarchical Text-ConditionalImage Generation with CLIP Latents	2022	Aditya Ramesh□Prafulla Dhariwal□Alex Nichol□Casey Chu□Mark Chen	this dataset are not publicly disclosed	- highlights qualitative results and reports less than 1% mean-squared error in reconstructing image embeddings -The results demonstrate superior image quality compared to earlier models like **GLIDE** and **DALL·E 1**	• High Photorealism • Improved Diversity • Zero-shot Capabilities • Semantic Control	• Complexity • Dependence on Data Quality • Biases in Outputs
https://ar5iv.labs.arxiv.org/html/2409.11340	2024	**Xi Li**	X2I Dataset ("Anything to Image")	specific quantitative accuracy metrics	• Unified Framework	• Resource Intensive

		Linchao Bao **Sheng Guo** **Xin Zhao**		are not explicitly stated. However, the use of high-resolution training stages and progressive optimization highlights the model's superior aesthetic quality and adaptability	• Multimodal Inputs • HighQuality Outputs • Scalable Dataset	• Quality Dependence • Complex Implementation
How to Read Paintings: Semantic Art.	2018	Noa Garcia and George Vogiatzis	SemArt dataset contains 21,384 samples that provides artistic comments along with fine-art paintings and their attributes Dataset's link	human accuracy is high, reaching **88.9%.**	**Multimodal Approach:** leverages both textual and visual modalities **Human Evaluation Comparison**: comparing model outputs with human evaluations gives credibility to the results.	**Focus on Retrieval**: The work focuses heavily on retrieval tasks rather than exploring other aspects of semantic understanding.