ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

AN INVESTIGATION OF THE PERCEPTUAL CAPABILITIES OF CLIP IN THE ART DOMAIN

Relatore:

Prof. Andrea Asperti

Presentata da:

Leonardo Dessì

Sessione Ottobre 2025

Anno Accademico 2024/2025

Abstract

This thesis investigates the native perceptual capabilities of OpenAI's CLIP model in the domain of fine art, where its alignment with human interpretation is poorly understood. By treating CLIP as a fixed, pre-trained system without fine-tuning, this work isolates its intrinsic representational biases. The central objective is to assess how CLIP's representations encode artistic style and synthetic artifacts beyond semantic content. Using datasets of both human-made and AI-generated art, experiments evaluate image-text alignment, style recognition, and the model's correlation with human judgments.

The findings reveal that CLIP's perceptual framework is overwhelmingly dominated by semantic content. While robust in coarse semantic matching, the model struggles with fine-grained stylistic classification and exhibits poor generalization. Crucially, a significant "perceptual gap" is identified between CLIP's assessments and human evaluations of AI-generated art, stemming from the model's insensitivity to visual artifacts and compositional errors. These results underscore the limitations of using semantic similarity as a proxy for artistic fidelity and highlight the need for models that are more perceptually and culturally aligned for application in subjective domains.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent advances in artificial intelligence, and in particular the development of models that jointly integrate visual perception with linguistic understanding, have led to the emergence of Vision–Language Models (VLMs) [2,3]. Among these, OpenAI's Contrastive Language–Image Pre-training (CLIP) represents a foundational contribution, having introduced a paradigm in which natural language descriptions serve as supervision for visual representation learning [1]. Through its training objective, CLIP acquires transferable visual concepts by embedding images and texts into a shared feature space, where corresponding pairs are brought closer together and unrelated ones are separated [1, 4]. This contrastive mechanism [5] endows the model with the ability to generalize across tasks, enabling zero-shot transfer [6] without the need for domain-specific fine-tuning. The effectiveness of this approach has been repeatedly demonstrated, with CLIP achieving high performance in zero-shot image classification [7–9], retrieval, re-identification, and semantic search.

Beyond discriminative applications, CLIP has become integral to state-of-the-art generative pipelines. Its capacity to align textual and visual modalities allows it to act as a guidance mechanism [10, 11] in models such as GLIDE [12], DALL·E [13], and Stable Diffusion [14, 15], thereby shaping the production of synthetic imagery from textual prompts. This functionality has also enabled novel creative uses, including visual storytelling, interactive applications, and style transfer.

Despite this wide adoption, the perceptual alignment established by CLIP is not fully understood. While the model's ability to capture semantic content is well documented, its sensitivity to stylistic detail, historical context, and aesthetic coherence remains uncertain. The domain of art analysis is particularly well suited to reveal these limitations. Unlike everyday photographs or benchmark datasets, works of art embody a high degree of subjectivity, interpretative richness, and cultural specificity [16, 17]. Paintings, in particular, are not merely depictions of objects or scenes, but artifacts in which composition, medium, and style are as essential as content. For this reason, understanding art requires more than semantic labeling: it demands an appreciation of form, abstraction, symbolism, and historical situatedness.

CLIP, however, exhibits systematic difficulties in addressing these aspects. Models guided by CLIP in generative contexts often produce images with hyperrealistic surface details but anachronistic or stylistically inconsistent features [18]. These errors highlight a broader challenge: CLIP tends to prioritize large, easily recognizable objects in images and early tokens in text descriptions [8], thus favoring semantic cues over subtler visual patterns [2, 19]. As a consequence, brushwork, texture, and other fine-grained stylistic characteristics—which play a central role in art historical interpretation—are often neglected.

This thesis takes these challenges as its point of departure. Its central objective is to investigate the perceptual capabilities of CLIP in the artistic domain, with particular emphasis on whether the model encodes and differentiates not only semantic content but also stylistic traits, historical cues, and synthetic artifacts. To this end, CLIP's vision encoder is treated as a fixed perceptual system: no fine-tuning, prompt engineering, or adapters are employed. This methodological choice allows for an analysis of the inductive biases and representational priors intrinsic to the pretrained model, thereby isolating its "native" perceptual competence.

The evaluation is carried out across two complementary datasets: a collection of human-made artworks from the National Gallery of Art [20], and a set of AI-generated pastiches [18]. The dual perspective enables a systematic comparison between CLIP's representations of authentic artistic production and synthetic imagery, focusing on whether the model can detect stylistic irregularities, temporal inconsistencies, or characteristic artifacts introduced by generative processes. By examining interpretive dimensions such as semantic category, artistic style, historical period, and visual distortion, the thesis probes the extent to which CLIP's perceptual space aligns with human interpretations of art.

Ultimately, this inquiry aims to address a broader research question: can large-scale multimodal models like CLIP develop something akin to an "aesthetic sense"? If so, is this grounded in abstract perceptual structures, statistical regularities learned from training data, or biases arising from their design? By combining conceptual reflection with empirical evaluation, this thesis seeks to highlight both the potential and the limitations of CLIP in the context of art, while pointing to the need for models that are not only semantically accurate but also perceptually and culturally aligned.

TODO: Aggiungere scaletta tesi

Chapter 2

Foundations of Multimodal Learning and Visual Representation

2.1 The Multimodal Nature of Perception and Artificial Intelligence

Our experience of the world is inherently multimodal, involving the simultaneous perception of objects through sight, sound, texture, smell, and taste [21]. Consequently, a research problem or dataset is characterized as multimodal when it incorporates multiple such channels of information, referred to as modalities [21, 22]. A modality generally refers to the specific way in which something happens or is experienced [21]. For Artificial Intelligence to progress toward a comprehensive understanding of the surrounding world, it must be able to interpret and reason about these complex, multifaceted signals collectively [21].

Multimodal Machine Learning (MML) is the specialized and multidisciplinary field dedicated to building computational models capable of processing and relating information derived from disparate modalities. While the term modality often evokes sensory channels such as vision and touch, MML typically focuses on three primary channels in computational research: natural language (which can be written or spoken), visual signals (represented by images or videos), and vocal signals (encoding sounds and paralinguistic features) [21].

The central importance of MML stems from the recognition that multimodal data offers a richer, more comprehensive perspective on an entity or phenomenon than any single data source alone. Since multimodal data depict an object from different viewpoints, they often contain information that is complementary or supplementary in content [22]. For example, early investigations into speech recognition demonstrated that incorporating visual information, such as lip motion and mouth articulation, significantly enhanced overall performance [23], a concept motivated by observations like the McGurk effect in human speech perception [24]. By leveraging these combined informational sources, MML systems enable a wide range of applications, including audio visual speech recognition [23], image captioning [25–27], cross modal retrieval [28,29], visual question answering [30], and generative tasks like text to image synthesis [31,32]. In specialized domains, MML is vital for healthcare diagnostics, integrating complex data streams like medical imaging and genomic profiles [33–35].

A foundational challenge in MML is the heterogeneity gap [21,22]. Since modalities exist in unequal subspaces and feature vectors associated with similar semantics can be completely different across modalities, direct comparison or utilization by subsequent machine learning modules is hindered [22,33]. For instance, language is often represented symbolically, whereas visual and audio modalities are represented as signals [21]. The primary goal of multimodal representation learning is therefore to bridge this gap by learning how to represent diverse input signals in a unified, shared semantic subspace [22].

2.2 Core Technical Challenges in MML

The field of MML faces unique challenges due to the heterogeneity of the data [21]. Moving beyond traditional categorization methods like early and late fusion, research identifies five core technical challenges central to the field's advancement: representation, translation, alignment, fusion, and co-learning [21,22,36].

2.2.1 Representation

The challenge of representation concerns learning how to effectively summarize multimodal data in a way that exploits both the complementarity and the redundancy of the multiple modalities. A representation is commonly understood as a vector or tensor representation of an entity, such as an image, audio sample, or sentence. A multimodal representation combines information from multiple such entities. The difficulty lies in combining data from heterogeneous sources, managing varying levels of noise, and handling missing data [21].

MML models typically adopt one of two major strategies for representation [21,22]:

1. Joint Representations: This approach projects unimodal representations simultaneously into a single, common, shared semantic subspace. Joint representations are optimal when all modalities are consistently present during both training and inference, and they are frequently employed in tasks such as audio visual speech recognition and affective computing. The simplest instance of a joint representation is the concatenation

of individual modality features, often referred to as early fusion [21, 22].

2. Coordinated Representations: This approach maps each modality into a separate yet coordinated space. While the projection is independent for each modality, the resulting spaces are constrained to be coordinated, perhaps by minimizing cosine distance, maximizing correlation, or enforcing a partial order between them. This structure is advantageous for applications such as multimodal retrieval, translation, and zero shot learning, particularly where only one modality may be present during test time [21, 22].

2.2.2 Translation

Translation addresses the necessity of mapping or converting data from one modality to another. The objective is to generate the equivalent entity in a target modality given an input entity from a source modality. Examples include generating a descriptive sentence given an image (image captioning) or generating an image based on a textual description (text to image generation). This is challenging because the relationship between modalities is often ambiguous or subjective; for instance, many correct sentences can describe a single image, meaning a singular perfect translation may not exist [21].

Translation approaches are broadly categorized as example based (using a dictionary for translation) or generative (constructing a model that produces the translation). Generative approaches, particularly those built on end to end trained neural networks utilizing an encoder decoder architecture, are highly popular for multimodal translation. These models first encode the source modality into a vector representation and then use a decoder module to generate the target modality sequence or signal [21].

2.2.3 Alignment

The challenge of alignment focuses on identifying direct relationships and correspondences between the subcomponents of instances from two or more distinct modalities. For example, in visual language tasks, alignment involves matching specific words or phrases in a caption to corresponding regions within an image. Correct alignment is crucial for tasks like multimedia retrieval, allowing for complex searches such as finding video content based on textual descriptions. Alignment approaches must be robust enough to handle possible long range dependencies and ambiguities between the constituent elements [21].

Alignment can be categorized as explicit (seeking annotated correspondences, such as aligning steps in a recipe to a video showing the process [37]) or implicit (serving as an intermediate or latent step for another task, such as refining cross modal retrieval by aligning image regions and words [38]). Attention models are a major technique for implicit alignment, allowing models to align words in a question with subcomponents of an image or text source, which enhances accuracy and interpretability [21].

2.2.4 Fusion

Fusion is the process of integrating information from multiple modalities with the explicit goal of predicting an outcome measure, such as a classification label or a continuous value. MML systems pursue fusion for several key reasons [21]:

1. Robustness: Access to multiple modalities observing the same phenomenon yields more reliable predictions, particularly beneficial in areas like audio visual speech recognition.
2. Complementarity: Fusion can capture information that is not present in individual modalities alone.
3. Graceful Degradation: A multimodal system can continue to function even if data from one of the modalities is unavailable (e.g., recognizing emotion from visual signals when audio is missing).

Fusion often occurs late in the processing pipeline, interacting with the final prediction stages. The line between multimodal representation learning and fusion can become blurred when using deep neural networks, where representation learning and classification objectives are learned simultaneously. Fusion methods include model agnostic techniques (e.g., early, late, or hybrid combination strategies) and model based approaches like kernel methods, graphical models, and neural networks [21].

2.2.5 Co-learning

Co-learning is defined as aiding the modeling of a resource poor modality by leveraging knowledge acquired from a resource rich or cleaner modality. This challenge involves transferring knowledge among modalities, their representations, or their predictive models. Co-learning is particularly salient when one modality suffers from limited annotated data or noisy input [21].

This paradigm is exemplified by algorithms such as co-training, conceptual grounding, and zero shot learning (ZSL). Conceptual grounding, for example, involves learning semantic meanings based not only on language but also on sensorimotor experience and perceptual inputs like vision or sound [39, 40]. Crucially, Co-learning explores how knowledge learned from one modality can improve a computational model trained on a different modality [21]. This concept of knowledge transfer is central to the development of highly generalizable models, as will be discussed in the context of large scale pretraining.

2.3 Perceiving and Representing Visual Information

Before discussing the Contrastive Language Image Pretraining (CLIP) architecture, it is necessary to establish the modern computational mechanisms through which vision models interpret raw visual data. Visual data, such as images, are fundamentally structured as grid graphs in pixel space [36]. Historically, attempts to bridge the semantic gap—the gulf between simple visual features and rich user semantics—required explicit domain knowledge and detailed feature engineering. Early content based image retrieval relied on image processing methods to extract low level visual cues such as color, local geometry, and texture [41].

The paradigm shift brought about by deep learning revolutionized this process.

2.3.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) emerged as the dominant architecture for processing visual data, demonstrating powerful representation ability. CNNs, typified by models like ResNet [42], use successive layers of local convolutions and non linear activation functions to generate increasingly abstract representations of the input data [21].

A defining characteristic of CNNs is the inherent reliance on image specific inductive biases. Locality, two dimensional neighborhood structure, and translation equivariance are explicitly "baked into" every layer of the network [43,44]. While these biases significantly aid efficiency and performance on traditional image classification tasks like ImageNet [45], they restrict the architecture's generality. If the input modality changes (e.g., moving from a fixed resolution 2D image to a point cloud or audio spectrogram), the CNN architecture often requires re-design, as the spatial structure is fundamentally different [44].

2.3.2 Vision Transformers and Attention Mechanisms

More recently, the Transformer architecture [46], originally developed for sequence modeling in natural language processing (NLP), has been adapted for vision tasks. The key idea in the Vision Transformer (ViT) framework is to treat an image as a sequence of input elements, or "patches", analogous to tokens or words in a text sentence [43].

Transformers utilize self attention, a mechanism that allows the network to weigh the importance of different parts of the input relative to other parts, thereby enabling the capture of global relationships (non local patterns) [43,46]. Mathematically, the Transformer's self attention mechanism can be conceptualized from a geometrically topological perspective, modeling the input sequence as a fully connected graph. Each token embedding, regardless of modality, is treated as a node in this graph [36].

Unlike CNNs, the ViT architecture employs far fewer image specific inductive biases. The two dimensional neighborhood structure is utilized only minimally, primarily during the initial tokenization (cutting the image into patches) and when adjusting positional embeddings. Beyond this, the ViT must learn all spatial relationships among the patches from scratch. This makes the Transformer a modality agnostic pipeline [43]. The capacity of Transformers to handle diverse inputs flexibly, from pixels to audio waveforms or text sequences, without major architectural overhaul, is one of its notable advantages for MML [36].

The success of the Vision Transformer demonstrates that visual perception can be effectively achieved by models with minimal explicit visual domain knowledge, provided they are scaled appropriately [44]. This design philosophy paved the way for models that jointly leverage vast amounts of language and visual data, such as CLIP, whose visual encoder is often a Transformer variant [33].

2.4 Multimodal Representation Learning Frameworks

Multimodal representation learning structures how heterogeneous inputs are processed to achieve a unified understanding, serving as the essential backbone of any multimodal model [21]. The fundamental goal remains reducing the heterogeneity gap while preserving the modality specific semantics [22].

2.4.1 Joint Representation and Fusion

Joint representation models project unimodal inputs into a common space for subsequent fusion [22]. Neural networks are a highly popular method for constructing such joint representations. In a joint neural network model, each modality initially passes through several individual neural layers, followed by a hidden layer that projects the modalities into the shared joint space. The resultant joint multimodal representation can then be used directly for prediction. These models can be trained end to end, simultaneously learning both the representation of the data and the ability to perform a particular task, leading to a close relationship between representation learning and multimodal fusion [21].

Deep Graphical Models, such as multimodal Deep Boltzmann Machines (DBMs), also fall under the joint representation category, modeling the joint distribution over inputs (e.g., image and text) by fusing modalities in a unified latent space [22, 47].

2.4.2 Coordinated Representation

Coordinated representations are designed for scenarios where the modality representations remain separate but are structured to interact or coordinate knowledge [22]. This enables applications where certain modalities might be absent during inference [21]. The coordination typically involves forcing a relationship between the respective projection functions for each modality such that their outputs align in a shared, comparable space [21, 22].

Cross modal similarity methods learn coordinated representations by constraining similarity measurements. The learning objective is to ensure that the cross modal similarity distance for pairs describing the same semantics or object is minimized, while dissimilar pairs are maximized [22]. A representative application is the Deep Visual Semantic Embedding model (DeViSE), which coordinates visual features from a CNN with textual features from a word embedding model (like word2vec) [48]. The aim here is to transfer knowledge, allowing the model to improve visual representations by enforcing similarity between modalities during training, thereby capturing shared semantics [22].

2.4.3 Encoder Decoder Models and Translation

Encoder decoder models are fundamental to multimodal translation tasks, especially generation [49]. The process requires two main steps: an encoder converts the source modality into a latent vector, and a decoder uses this vector to construct the target modality [21, 49].

For visual language translation, such as image captioning, the image is encoded using networks like CNNs [50] or Vision Transformers [33], and the text is typically generated using sequence models like Recurrent Neural Networks (RNNs) [27, 51] or Transformer based decoders. The ability of this framework to generate novel samples of the target modality conditioned on the source modality is its significant advantage [22]. This framework also allows for explicit modeling of semantic consistency between modalities; for example, maximizing the likelihood of generating a correct sentence while minimizing the representation difference in a common subspace [52].

A particularly advanced application of this framework involves attention mechanisms. Attention allows the decoder to selectively and dynamically concentrate on salient parts of the source input during the prediction process. This ability to select prominent features enhances system performance and noise tolerance, leading to improved outcomes in tasks like image captioning [22].

2.5 Multimodal Scaling and The Path to CLIP

The increasing scale and complexity of MML models, particularly those leveraging Transformer architectures, have opened new frontiers in generalization and task transfer, forming the methodological context for foundational models like CLIP [1].

The emergence of Large Multimodal Models (LMMs), such as LLaVA and GPT-4o, since 2023 demonstrates the trend toward developing versatile systems capable of processing and generating diverse inputs and outputs like text, audio, and images. These systems often integrate a visual encoder with a powerful Large Language Model (LLM), connected via learned projection layers [3, 6].

This trend toward large scale models emphasizes the importance of transferability and zero shot capabilities [1]. This research investigating CLIP, which treats the model's vision encoder as a fixed perceptual system for artistic analysis, is situated precisely within this context of generalized, pretrained representation learning.

The key breakthrough exemplified by CLIP is the scalable pretraining task that uses natural language supervision for visual representation learning. Instead of being trained to predict a fixed, predetermined set of visual categories (as in older supervised computer vision systems), CLIP addresses the potentially easier proxy task of predicting which caption correctly pairs with which image [1].

This objective is achieved through a contrastive learning mechanism [5]. During pretraining on massive web scale datasets of image text pairs, CLIP jointly trains separate image and text encoders to maximize the cosine similarity of the embeddings belonging to the N real pairs in a batch, while simultaneously minimizing the similarity for the $N^2 - N$ incorrect pairings. The result is a shared multimodal embedding space where corresponding visual and linguistic concepts are brought closer together [1].

This contrastive mechanism endows CLIP with the power of zero shot transfer [6]. Because the visual concepts are learned and referenced through natural language, the model can classify or retrieve unseen visual concepts by generating a classifier directly from text prompts, eliminating the need for further domain specific training examples. This contrasts sharply with traditional supervised methods, which must infer concepts indirectly from labeled examples [1].

Understanding the foundational principles of multimodal learning and the functioning of vision models is essential for interpreting the results of this inquiry in Chapter 3, which will detail the CLIP architecture and its latent space.

Chapter 3

The CLIP Model

3.1 Introduction and Contextualization

The emergence of the CLIP model represents a fundamental methodological divergence from historical computer vision paradigms, establishing a new foundation for visual representation learning rooted in natural language supervision [1]. This section contextualizes CLIP's design rationale by first detailing the intrinsic limitations of conventional fixed-category classification systems, followed by an explanation of the pivotal shift toward leveraging web-scale linguistic data, which ultimately culminates in the introduction of the model's core contrastive objective.

3.1.1 Limitations of Fixed-Category Supervision

Historically, the dominant paradigm for developing state-of-the-art computer vision systems centered on training models to predict a predetermined, fixed set of discrete object categories [42, 43]. This approach typically involves matching image features, generated by a vision model (such as a ResNet or ViT) [42, 43], with a fixed set of randomly initialized weight vectors. These weight vectors are subsequently learned to represent visual concepts by minimizing the distance between the vector and images containing the corresponding category [53]. Prominent examples of this methodology include systems trained on large, manually curated datasets like ImageNet [50].

However, this conventional form of supervision presents inherent conceptual limitations, primarily restricting model generality and transferability [1]. Since classification labels are discretized, often converting rich textual descriptions into a simplistic scalar format, the rich semantics encapsulated within the original text are largely left unexploited [7, 54]. This constraint dictates that the visual recognition system operates solely on closed-set visual concepts, confining the model to a pre-defined taxonomy of categories [7].

Consequently, the utility and generalization capacity of these systems are severely limited: specifying any new visual concept beyond the initial closed set necessitates the acquisition and incorporation of additional labeled data [1]. The process of collecting and annotating large-scale, high-quality datasets for every specialized visual task is inherently resource-intensive, financially prohibitive, and often impracticable to scale across numerous distinct domains [9]. Training on such fixed sets leads to models focused on specific tasks, weakening the capacity to attain general visual representations and harming their transferability to novel, open-set applications [55]. The necessity of overcoming this dependence on labor-intensive annotation and predefined classification schemes motivated a fundamental shift in the training paradigm toward leveraging broader, more abstract sources of supervision [1].

3.1.2 The Shift to Natural Language as Supervision

The adoption of natural language as a scalable source of supervision has been primarily driven by the fundamental advances achieved in the field of Natural Language Processing (NLP) [1]. Over preceding years, pre-training methods designed to learn directly from raw text—pioneered by systems such as BERT [56], T5 [57], and the GPT series [58], demonstrated remarkable success in revolutionizing NLP. These advancements established "text-to-text" as a standardized input-output interface, facilitating the creation of powerful, task-agnostic architectures capable of performing zero-shot transfer across diverse downstream linguistic tasks. Highly capable systems like GPT-3, for instance, became competitive with bespoke, task-specific models while requiring negligible amounts of training data specific to the target dataset [1].

Drawing analogous inspiration, researchers recognized that shifting visual learning toward natural language provided an opportunity to exploit a vastly broader and deeper source of supervision compared to constrained, manually curated datasets. The supervision available in web-scale text enables the exploration of open-set visual concepts, thereby leveraging the sheer accessibility of massive image-text pairs available across the internet, which fundamentally enhances the model's generalization power. CLIP, for example, utilized approximately 400 million (image, text) pairs collected from the web during its pre-training phase [1].

The core objective of this new paradigm is the learning of transferable visual representations by establishing a joint, conceptual understanding between raw text and images. CLIP achieves this by jointly training an image encoder and a text encoder to maximize the similarity between corresponding visual and textual representations within a shared feature space. This process enables images to be correctly classified by matching image features against a classifier synthesized directly from text, such as a textual description of the category. The foundational pre-training task is simplified to predicting which caption correctly pairs with which image in a batch [1]. This approach moves beyond optimization for a singular benchmark, allowing the learned representations to transfer effectively to a multitude of tasks [59].

3.1.3 Precursors to Contrastive Learning

The aspiration to computationally link the vision and language modalities precedes CLIP, with various prior works attempting to establish connections between visual concepts and their linguistic descriptions. Early investigative studies explored enhancing content-based image retrieval by training models to predict nouns and adjectives contained within accompanying image text [60]. Other methods demonstrated that mapping classifiers trained to predict words in captions to a weight space manifold could yield more data-efficient image representations [61]. A notable precursor, Visual N-Grams [62], was among the first studies to successfully apply zero-shot transfer methods to conventional image classification datasets using a pre-trained, task-agnostic model. Furthermore, methods like VirTex (2020) empirically demonstrated that image caption annotations provided sufficient supervision for representation learning, yielding feature quality comparable to or surpassing that of models trained on ImageNet, but utilizing significantly less data [63].

CLIP's innovation relative to this preceding work lies in its highly efficient and scalable approach, fundamentally altering the objective and the scale at which this learning occurs. Previous attempts often focused on detailed prediction tasks, such as predicting the exact words (as in image captioning) or relying on a bag-of-words (BoW) encoding. CLIP introduced the arguably simpler, yet highly effective, proxy task of merely predicting if a text as a whole correctly pairs with an image, rather than predicting the individual words within that text [1].

By keeping the dual-encoder architecture and swapping the predictive learning objective for a contrastive objective in the vein of the multi-class N-pair loss, CLIP achieved a measured 4x efficiency improvement in the rate of zero-shot transfer to ImageNet compared to the BoW baseline. The core of the method involves optimizing a symmetric cross-entropy loss that maximizes the cosine similarity between the N authentic (image, text) pairs in a batch while minimizing the cosine similarity for the remaining $N^2 - N$ incorrect pairings. This scalable methodology, referred to as Contrastive Language-Image Pre-training (CLIP), was successfully implemented from scratch at a massive scale of 400 million image-text pairs, marking its conceptual breakthrough and setting new standards for transferable visual models [1].

3.2 Data Scale and the WebImageText (WIT) Dataset

The remarkable performance of CLIP is inseparable from the WebImageText (WIT) dataset, a large-scale collection of 400 million image–text pairs specifically designed to overcome the shortcomings of earlier resources. Prior datasets such as MS-COCO and Visual Genome contained only around 100,000 photos, making them inadequate for web-scale learning. Even larger alternatives like YFCC100M or Conceptual Captions suffered from sparse metadata, low semantic richness, or overly simplistic captions, resulting in insufficient coverage and detail. These limitations highlighted the need for a dataset that could combine both massive scale and meaningful natural language supervision [1].

WIT addressed this by adopting an active collection strategy. Instead of passively scraping the web, its creators built a query-driven process starting from 500,000 Wikipedia-derived queries, later expanded with bi-grams. To avoid overrepresentation of specific concepts, they capped the number of samples per query at 20,000. This approach ensured not only enormous scale but also broader and more balanced coverage of visual concepts, providing the foundation required for CLIP's contrastive pre-training paradigm [1].

Nevertheless, the reliance on open-web data introduced unavoidable challenges. As with other large-scale datasets like LAION, WIT inevitably incorporates biases and problematic material, including explicit content, stereotypes, and demographic imbalances [1]. Analyses of CLIP and similar models show that these biases can shape embeddings in concerning ways—for example, NSFW associations with certain names or harmful misclassifications of people [64]. This tension highlights the core trade-off in constructing datasets like WIT: while their scale enables powerful generalization and robustness to distribution shifts, it comes at the cost of ethical control and heightened exposure to the biases present in internet-scale data [64].

3.3 The CLIP Architecture: Components and Design

The performance and remarkable transferability of the CLIP model are intrinsically linked to its modular, dual-encoder architecture. This design consists of two independently operating, modality-specific encoder networks—one for visual inputs and one for linguistic inputs—which are trained jointly to project their respective features into a common, shared vector space via a contrastive learning objective.

3.3.1 The Image Encoder Variants and Selection

For the image encoder, the CLIP framework explored two distinct families of highly performant deep learning architectures: the established ResNet (Deep residual learning for image recognition) family and the contemporary Vision Transformer (ViT) family [1].

The architectures studied within the ResNet family included the ResNet50 and the ResNet101 variants. These were often adapted using the EfficientNet-style model scaling approach, resulting in increasingly complex models denoted as RN50x4, RN50x16, and RN50x64, which utilized approximately 4x, 16x, and 64x the computational capacity of the base ResNet50 model, respectively. The ResNet models employed by CLIP incorporated several modifications from the original architecture, including the ResNetD improvements and antialiased rect-2 blur pooling. Crucially, the standard global average pooling layer in the conventional ResNet structure was replaced with an attention pooling mechanism. This attention pooling is implemented as a single layer of "transformer-style" multi-head QKV attention, where the query vector is conditioned on the global average-pooled representation of the input image [1].

For the second architectural family, CLIP leveraged the Vision Transformer (ViT). The ViT models explored included ViT-B/32, ViT-B/16, and the larger ViT-L/14 variants. These implementations closely followed the original ViT design. In Transformer-based architectures, feature extraction is typically performed using the learnable [CLS] token (class token), which is concatenated to the sequence of image patch embeddings. In the context of the ViT architecture utilized in CLIP, this approach is the mechanism by which the patch-based image representation yields the final class representation. Some configurations, such as ViT-B/16, consist of 12 transformer layers with a hidden size of 768 dimensions [1].

In terms of scaling, while prior computer vision research often scaled models by solely increasing width or depth, the ResNet image encoders in CLIP adapted a uniform scaling approach across width, depth, and resolution. In contrast, research found that Vision Transformers are about $3\times$ more compute efficient than the CLIP ResNets when trained on sufficiently large datasets. The final choice for the best-performing CLIP model configuration was the ViT-L/14 pre-trained at a higher 336×336 pixel resolution, denoted as ViT-L/14@336px. This selection was made because this specific model achieved the best overall performance and was deemed the most compute-efficient variant among those studied [1].

3.3.2 The Text Encoder

The text encoder employed in the CLIP architecture is a Transformer-based model. Its base size is structured as a 12-layer, 512-wide model with 8 attention heads. The architecture utilizes modifications detailed in earlier generative pre-training literature [1].

The input text processing involves converting raw text into a format suitable for the Transformer. Textual inputs are tokenized using a lower-cased byte pair encoding (BPE) representation, which has a vocabulary size of 49,152. The text sequence is bracketed by the special [SOS] (Start-of-Sequence) and [EOS] (End-of-Sequence) tokens. To ensure efficient batch processing and parallel computation, the text sequence is capped at a fixed maximum sequence length of 77 tokens [1].

The embeddings of the tokens are then passed through the Transformer layers. To derive the final sequence-level representation, the activations of the highest layer of the Transformer at the [EOS] token position are used. This resulting feature is subsequently layer normalized before being mapped to the final embedding space. Although masked self-attention was used in the text encoder, preserving the ability to potentially add language modeling as an auxiliary objective, this auxiliary function was reserved for future work [1].

3.3.3 Dimensionality and Latent Space Design

CLIP is fundamentally designed to align features from the image and text modalities within a shared multimodal embedding space. This alignment is critical because it allows visual and linguistic information to be compared using a similarity function, typically cosine distance [1].

The output feature representations produced by the image encoder (e.g., from the attention pooling layer of ResNet or the [CLS] token equivalent of ViT) and the text encoder (from the [EOS] token activation) are processed by a learned linear projection layer. The purpose of this linear layer is to map the encoder-specific feature representations into the joint multimodal space. CLIP simplifies this design compared to some preceding contrastive learning methods by removing the non-linear projection between the internal encoder representation and the contrastive embedding space, relying solely on a linear projection [1].

The final common dimensionality of the shared latent space (denoted d or D) varies depending on the specific model size. For instance, the smaller ResNet-50 (RN-50) model has an embedding space size of $d = 1024$, while the RN-50x4 model utilizes an embedding space size of $d = 640$. For the ViT-B/16 model, the image feature vector dimension (768) is reduced to 512 by a linear layer to match the text encoder output. Similarly, the smaller ViT-B/32 CLIP model fine-tuned for specialized tasks extracts image representations of 512-dimension [1].

Finally, for the purposes of calculating cosine similarity and facilitating comparison within the embedding space, the final embeddings produced by the encoders and projected into the shared space undergo L2 normalization. This step projects the features onto a unit hypersphere. The

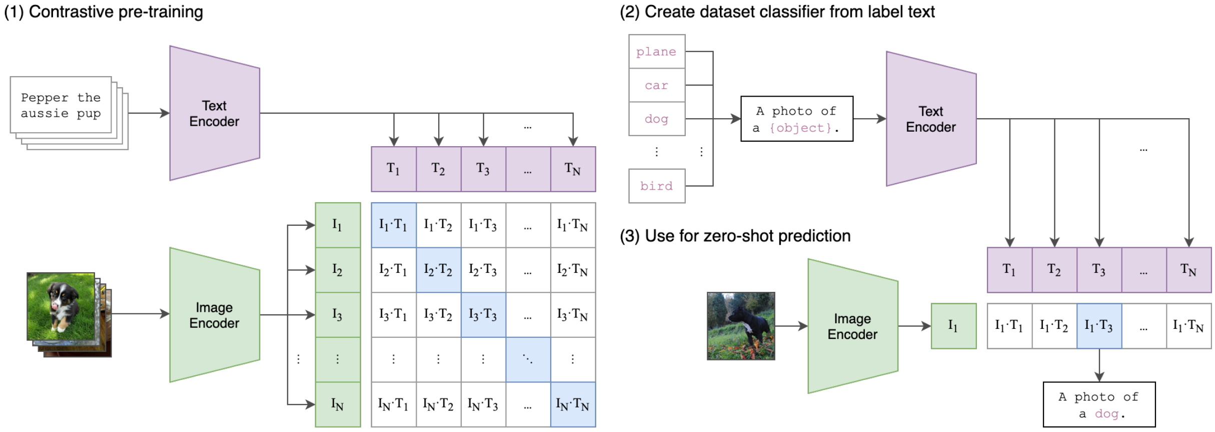

(1) Contrastive pre-training
Pepper the aussie pup
Text Encoder
Image Encoder
T1 T2 T3 ... TN
I1 I2 I3 ⋮ IN
I1·T1 I1·T2 I1·T3 ... I1·TN
I2·T1 I2·T2 I2·T3 ... I2·TN
I3·T1 I3·T2 I3·T3 ... I3·TN
IN·T1 IN·T2 IN·T3 ... IN·TN
(2) Create dataset classifier from label text
plane
car
dog
bird
A photo of a {object}.
Text Encoder
(3) Use for zero-shot prediction
Image Encoder
I1
I1·T1 I1·T2 I1·T3 ... I1·TN
A photo of a dog.

Figure 3.1: CLIP contrastative pre-training and its use for zero-shot prediction [1].

importance of this projection and normalization is paramount: the core contrastive objective relies on maximizing the cosine similarity (dot-product of L2-normalized features) between the real pairs and minimizing the similarity for the unmatched pairs across the batch [1].

3.4 The Contrastive Pre-training Objective and Loss Function

The efficacy of the CLIP model hinges upon a meticulously formulated and highly scalable training objective: the symmetric contrastive loss. This objective transcends the limitations of traditional fixed-category supervision by directly optimizing the alignment of feature representations across the visual and linguistic modalities in a shared embedding space.

3.4.1 The Batch Contrastive Learning Objective

The fundamental goal of CLIP's pre-training is to solve a proxy task: given a set of training examples, the model must predict which text description correctly corresponds to which image. This objective relies on batch contrastive learning, which operates on a batch of N randomly sampled (image, text) pairs $\{I_i, T_i\}_{i=1}^{N}$. The model is trained to maximize the similarity between the N authentic or "real" pairs while simultaneously minimizing the similarity for the remaining $N^2 - N$ incorrect pairings, referred to as in-batch negatives [1].

This process involves jointly training the image encoder (f_I) and the text encoder (f_T) to project their respective inputs into a normalized, joint multi-modal embedding space. Within this space, the affinity between any projected image feature $\nu_{I_i} = f_I(I_i)$ and any projected text feature $\nu_{T_j} = f_T(T_j)$ is quantified using the cosine similarity metric. The cosine similarity, mathematically equivalent to the normalized dot product of the L2-normalized feature vectors, is calculated for all possible combinations within the batch [1].

This computation yields an $N \times N$ similarity matrix, where the entry at position (i, j) represents the similarity score $S_{i,j} = \cos(\nu_{I_i}, \nu_{T_j})$ between the i-th image and the j-th text description. The diagonal elements of this matrix, where $i = j$, correspond precisely to the similarities of the positive (matched) pairs $\{I_i, T_i\}$ that occurred in the dataset. Conversely, all off-diagonal elements represent the similarities between mismatched (negative) pairs, which the learning objective endeavors to suppress. This core batch construction technique and learning paradigm were originally introduced as the multi-class N-pair loss objective [1].

3.4.2 The Symmetric Cross-Entropy Loss (NT-Xent)

The optimization of the contrastive objective utilizes a form of normalized temperature-scaled cross-entropy loss (NT-Xent). Specifically, CLIP employs a symmetric cross-entropy loss calculated over the pairwise similarity scores produced by the encoders. The dual nature of the loss ensures that the model learns alignment bi-directionally: predicting the correct text given an image, and predicting the correct image given a text.

Formally, given a batch of N image-text pairs, the total loss (L) is computed as the average of two distinct cross-entropy losses: the image-to-text loss (L_{I2T} or L_{image}) and the text-to-image loss (L_{T2I} or L_{text}).

The overall loss is defined as:

$$L = \frac{1}{2}(L_{\text{I2T}} + L_{\text{T2I}})$$

The image-to-text loss (L_{I2T}) calculates the probability that image I_i correctly matches its corresponding text T_i among all texts in the batch. This is achieved using the softmax function over the normalized similarity scores, maximizing the logarithm of the probability of the true positive pair:

$$L_{\text{I2T}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\cos(\nu_{I_i}, \nu_{T_i})/\tau)}{\sum_{j=1}^{N} \exp(\cos(\nu_{I_i}, \nu_{T_j})/\tau)}\right)$$

where $\nu_{I_i} = f_I(I_i)$ and $\nu_{T_j} = f_T(T_j)$ are the normalized feature vectors, and τ is the temperature parameter. The denominator represents the sum of exponentiated similarities between the anchor image ν_{I_i} and all N text captions in the batch.

Symmetrically, the text-to-image loss (L_{T2I}) calculates the probability that text T_i correctly matches its corresponding image I_i among all images in the batch:

$$L_{\text{T2I}} = -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{\exp(\cos(\nu_{I_i}, \nu_{T_i})/\tau)}{\sum_{j=1}^{N} \exp(\cos(\nu_{I_j}, \nu_{T_i})/\tau)}\right)$$

Here, the denominator sums the exponentiated similarities between the anchor text ν_{T_i} and all N images in the batch. By minimizing this symmetric objective, the model learns strong alignments that generalize across both directions of retrieval.

A central feature of CLIP's contrastive learning is the learnable temperature parameter (τ), which scales similarity scores in the softmax and adjusts how sharply the model distinguishes between positive and negative pairs. Unlike earlier methods that fixed τ as a manually tuned hyperparameter, CLIP optimizes it during training, typically starting from 0.07, so the model can dynamically adapt to diverse data scales and concepts. Training CLIP at web scale also required massive computational effort: the model was trained on 400 million image–text pairs using AdamW, a warmup–cosine learning rate schedule, and very large batch sizes, sometimes exceeding 30,000 samples. Runs were conducted on clusters of high-end GPUs such as NVIDIA A100s, with training costs on the order of 5×10^{22} FLOPs. To make this feasible, techniques like mixed-precision training and gradient checkpointing were used to reduce memory demands and speed up computation.

3.5 Zero-Shot Transfer, Emergent Capabilities, and Limitations

The capacity of CLIP models to perform zero-shot transfer represents the defining methodological departure from conventional computer vision systems. This mechanism, which leverages natural language to dynamically synthesize a classifier, provides CLIP with remarkable flexibility and generalization abilities [65]. Nevertheless, this scale of pre-training also imposes distinct limitations, particularly concerning fine-grained perception and robustness in highly specialized scenarios.

3.5.1 The Zero-Shot Classification Mechanism

The core functionality that enables CLIP to operate on unseen datasets without requiring any task-specific fine-tuning is the zero-shot classification mechanism. This capability is a direct consequence of the dual-encoder architecture and the contrastive training objective, which learns a joint embedding space where corresponding visual and linguistic concepts are highly aligned.

At the inference stage, the pre-trained CLIP model reuses its capability to predict the correct pairing of an image and a text snippet. To apply CLIP to a downstream classification task, such as identifying categories within ImageNet, the classification process is transformed into an image-to-text matching problem.

The technical explanation of classifier synthesis proceeds through the following steps:

1. **Feature Extraction** A query image I is processed by the image encoder, yielding a feature embedding $\nu_I \in \mathbb{R}^d$.

2. **Classifier Synthesis via Prompt Engineering** A classifier weight vector is dynamically synthesized for each of the K categories $\{C_1, \ldots, C_K\}$ present in the downstream dataset. This involves prompt engineering, where each category name is inserted into a pre-defined natural language template, resulting in a set of text strings T_i. A widely adopted hard prompt template for generic classification datasets, such as ImageNet, is simply "a photo of a $\{$class$\}$". For datasets requiring specialized context, task-relevant context is added, such as specifying the domain keyword in the template for fine-grained classification datasets (e.g., "a centered satellite photo of $\{$class$\}$" for EuroSAT). The use of prompts, whether manually crafted or automatically generated, is critical because the format of the prompts significantly affects the model's accuracy.

3. **Embedding the Prompts** Each category-specific text string T_i is fed into the text encoder, which generates a classification weight vector $\mathbf{w}_i$ (or text embedding $t_i \in \mathbb{R}^d$). The process of generating these text embeddings from the class name and prompt template is denoted as classifier weight generation.

4. **Classification via Cosine Similarity** The final prediction is achieved by measuring the cosine similarity (or dot product of L2-normalized features) between the image embedding ν_I and the synthesized text embeddings $\{\mathbf{w}_i\}_{i=1}^{K}$. The model predicts the class C_i corresponding to the text prompt $\mathbf{w}_i$ that exhibits the highest cosine similarity with the image feature ν_I. The prediction probability for class i is formulated mathematically as:

$$p(y = i|\mathbf{z}) = \frac{\exp(\text{sim}(\mathbf{z}, \mathbf{t}_i)/\tau)}{\sum_{j=1}^{K} \exp(\text{sim}(\mathbf{z}, \mathbf{t}_j)/\tau)}$$

   where $\mathbf{z}$ is the image embedding, $\mathbf{t}_i$ is the text embedding, $\text{sim}(\cdot, \cdot)$ is the cosine similarity, and τ is the learned temperature parameter of CLIP.

This approach allows the natural language text to directly communicate visual concepts, serving as a powerful alternative to inferring concepts indirectly from limited labeled examples.

3.5.2 Emergent Capabilities and Modality Alignment

A critical finding in the evaluation of CLIP is the emergence of powerful, transferable visual representations derived from its large-scale natural language supervision. CLIP's features demonstrate notable robustness to distribution shift and are often competitive with, and in some cases surpass, fully supervised baselines. For instance, the best CLIP model improved zero-shot accuracy on ImageNet from a proof-of-concept 11.5% to 76.2%, matching the performance of the original ResNet50 without using any of the 1.28 million crowd-labeled training examples. Across a 27-dataset evaluation suite, zero-shot CLIP either outperformed or was competitive with a fully supervised linear classifier fitted on ResNet50 features on 16 datasets, including ImageNet [1].

Superior Robustness and Generalization

CLIP models display superior robustness compared to equivalent-accuracy supervised ImageNet models, substantially shrinking the "robustness gap" when confronted with natural distribution shifts across various image distributions. For instance, CLIP achieved 60.2% accuracy on the ImageNet Sketch Dataset, whereas the fully supervised ImageNet ResNet101 model only achieved 25.2% accuracy. This resilience stems from the exposure to a diverse array of signals within the extensive web-scale training dataset.

Emergent Task Learning

The contrastive pre-training task forces CLIP to learn a wide variety of tasks during pretraining simply to satisfy the objective of matching the correct image-text pairs. This task learning can then be leveraged for zero-shot transfer to many existing datasets. These emergent capabilities include generalizing across complex visual-linguistic tasks:

- **OCR (Text Recognition)** CLIP learns the ability to perform Optical Character Recognition (OCR). For example, CLIP's image embeddings are utilized in Scene Text Image Super-Resolution (STISR) systems to extract text features and guide the reconstruction process, confirming its proficiency in understanding textual content even in low-quality or noisy images [1].

- **Geo-localization** The model exhibits proficiency in tasks related to geographical location determination [1].

- **Action Recognition** CLIP has shown the ability to perform action recognition in videos [1].

The foundational alignment achieved by CLIP, which maps similar concepts expressed in images and text to similar feature representations (e.g., matching the image of a cat I_c with the text T_c "a photo of a cat") [59], makes it well-suited for subsequent applications such as image retrieval, composed image retrieval, image generation, and enhancing specialized systems like Zero-Shot Anomaly Segmentation (ZSAS) [59, 66].

Chapter 4

Related works

The emergence of Contrastive Language-Image Pre-training (CLIP) has fundamentally altered the landscape of multimodal artificial intelligence, offering a robust foundation for transferring knowledge across vision and language domains [2,67]. Trained on extensive web-scale collections of image-text pairs [64,68], CLIP learns to align features from separate visual and textual encoders into a unified embedding space through a contrastive learning objective [53, 59, 69]. This mechanism enables impressive zero-shot generalization and has led to its integration as a crucial component in contemporary generative models and vision-language pipelines [70]. However, CLIP's broad applicability has simultaneously necessitated a rigorous and critical evaluation of its intrinsic limits, leading to extensive research focusing on both model shortcomings and adaptation techniques.

4.1 CLIP: Limitations and Robustness

To understand the core capability of CLIP and, consequently, delineate the necessity of the current research, it is essential to first review the known failure modes and inherent biases of the pre-trained model when operating in its native, unmodified state [2].

4.1.1 Semantic Gaps and Biases

A primary area of inquiry concerns the depth of CLIP's semantic comprehension, specifically questioning whether it genuinely understands textual meaning beyond mere surface-level co-occurrence and statistical association [71]. Studies investigating CLIP's performance in complex, multi-object scenarios have revealed significant architectural biases within the model [8].

Notably, the image encoder exhibits a tendency to favor larger objects in a scene, while the text encoder prioritizes the first-mentioned objects within a prompt. These structural asymmetries lead to model instability, resulting in substantial performance drops in image-text matching tasks when captions are semantically equivalent but structurally varied—for instance, when object size or token order is manipulated. This behavior highlights a critical gap in understanding CLIP's inconsistency, which can manifest as systematic errors that betray inherent biases rather than robust conceptual understanding [8].

Furthermore, analyses suggest that CLIP tends to prioritize the encoding of semantic content, such as objects, scenes, and compositions, over more nuanced stylistic features, such as brushwork, color palette, or compositional structure. This bias is consistent with the model's large-scale training objective, which aims for broad semantic alignment rather than fine-grained aesthetic discernment. This limitation in capturing visual nuance becomes particularly acute in complex, subjective domains like visual art [2].

4.1.2 Adversarial Vulnerability and Internal Analysis

Despite demonstrating commendable resilience to challenging natural distributional shifts [64, 72], CLIP-based detectors are not impervious to malicious manipulation. Research has shown that these detectors are vulnerable to white-box adversarial attacks, a weakness they share with traditional Convolutional Neural Networks (CNNs) [73]. Interestingly, adversarial attacks do not easily transfer between CNN-based and CLIP-based methods, suggesting divergent internal feature representations are exploited by these attacks [73].

In addition to evaluating external vulnerability, model inversion techniques have been employed to understand the internal representations and knowledge embedded within CLIP [64]. In [64], traditional gradient ascent techniques [74–76] are applied to reconstruct images from CLIP embeddings. By generating images that best align with a given textual prompt (thus effectively inverting the model) researchers can gain insights into what the model "perceives". This approach has confirmed CLIP's capacity to seamlessly blend concepts (e.g., combining "panda mad scientist" and "sparkling chemicals"). Critically, inversion studies have also uncovered undesirable biases inherited from the web-scale training data, such as strong associations between certain celebrity names (particularly women) and sexually explicit content, a finding with significant implications for the use of CLIP embeddings in downstream generative models [64].

4.2 CLIP: Adaptation and Fine-Tuning

In response to CLIP's demonstrated limitations, a vast body of literature focuses on developing mechanisms to enhance its performance, improve generalization, and mitigate the modality gap—the persistent finding that image and text embeddings often occupy separated regions within the shared feature space [4]. These approaches primarily fall into three categories: feature modification via lightweight modules (adapters/projection), optimization of text inputs (prompt learning), and integration into larger multimodal systems (LMMs).

4.2.1 Adapters and Projection Methods

Standard fine-tuning of large models like CLIP is often slow and prone to overfitting, especially in few-shot scenarios. Adapters circumvent this by introducing minimal, lightweight modules—often implemented as bottleneck layers—which are inserted into the network and tuned, allowing the core pretrained weights to remain frozen [9]. Examples in this category include:

- CLIP-Adapter proposes an additional bottleneck layer to learn new features, utilizing a residual-style feature blending with the original CLIP features. This approach revived the "pretrain-finetuning" paradigm while requiring a small number of additional parameters [9].
- TIP-Adapter (Training-free CLIP-Adapter) aimed for higher efficiency by constructing the adapter using a key-value cache model derived from the few-shot training set, retrieving knowledge instead of requiring gradient updates. Its further variant, TIP-Adapter-F, achieves state-of-the-art results but reintroduces fine-tuning of the cache keys [77].
- Other adaptation methods, such as LIxP [65] and APE (Adaptive Prior rEfinement) [78], also focus on maximizing accuracy while minimizing computational overhead and the number of learnable parameters by refining CLIP's prior knowledge or selectively utilizing significant feature channels.

Projection-based methods specifically address the modality gap by projecting features into different subspaces to achieve better alignment. For instance, Selective Vision-Language Subspace Projection (SSP) proposes a training-free approach that utilizes local image features to construct unified visual and language subspaces, projecting the main features onto them to reduce distribution differences [4]. Similarly, networks like SgVA-CLIP focus on learning domain-specific visual features under the guidance of cross-modal knowledge [19].

4.2.2 Prompt Engineering and Tuning Methods

The success of CLIP relies heavily on the quality of its input prompts, making manual prompt engineering a critical yet cumbersome task that requires domain expertise, as minor wording changes can drastically impact performance [7, 9].

- Context Optimization (CoOp) proposed substituting cumbersome hard prompts with learnable continuous soft prompts [7,9]. By tuning these prompt vectors using few-shot data while keeping the main encoders fixed, CoOp significantly outperforms carefully engineered manual prompts [7]. However, a major limitation identified in CoOp is its tendency to overfit the base classes seen during training, resulting in poor generalization to novel or unseen classes [53].
- Conditional Context Optimization (CoCoOp) was introduced to overcome this generalization problem by extending CoOp to learn a lightweight meta-network that generates an input-conditional prompt vector for each image, making the prompt dynamic rather than static [53].
- Further advancements leverage Large Language Models (LLMs) themselves to generate more effective inputs. The Customized Prompts via Language models (CuPL) method uses an LLM (like GPT-3) to generate arbitrary numbers of descriptive, customized prompts (image-prompts) that contain discriminating visual details for fine-grained classification tasks. Using LLM-generated prompts improves zero-shot accuracy compared to hand-written prompts by enabling the model to focus on semantically important image regions [79].

4.2.3 Scaling to Large Multimodal Models (LMMs)

Modern research has scaled vision-language capabilities by integrating visual encoders, often based on frozen CLIP architectures, with powerful Large Language Models (LLMs) through learned projection layers (e.g., Q-Formers or linear projectors) [6]. These resulting multimodal LLMs (MLLMs) are designed for complex, conversational, and instruction-following tasks that go beyond mere classification or retrieval [3].

Prominent examples of these architectures include BLIP-2 [80], which uses a Q-Former to bridge the frozen image encoder and the LLM; InstructBLIP [81], which performs systematic vision-language instruction tuning; Qwen-VL [82]; and ShareGPT4V [83], which improved alignment using detailed, high-quality captions curated via GPT-4 Vision. These models represent the current frontier in multimodal reasoning, demonstrating advanced capabilities in visual analysis and generation [16].

4.3 Vision–Language Models in Art

The artistic domain presents a unique and demanding challenge for AI, requiring not only object recognition but also high-level interpretation of style, composition, and cultural context [16].

4.3.1 Datasets and Benchmarking Tasks

Early efforts applied computer vision techniques to art history tasks like style and author identification [17, 84]. These efforts have since been augmented by multimodal models:

- SemArt is a foundational multi-modal dataset comprising fine-art painting images, attributes, and artistic comments, designed for semantic art understanding and the Text2-Art retrieval challenge [17].

- CLIP-Art demonstrated an early adaptation of CLIP, fine-tuning the model using image-text pairs from the SemArt dataset to achieve improved performance in fine-grained art classification and retrieval tasks [85].

- AQUA (Art QUestion Answering) proposed a benchmark for answering visual and knowledge-based questions about paintings, highlighting that answering such questions often requires external knowledge beyond visual content [84]. The subsequent ArtQuest dataset was introduced to counter language biases that were found to be hidden in existing VQA datasets like AQUA [86].

- ArtBench provides a standardized and class-balanced dataset specifically for benchmarking the style classification performance and visual quality of generative models in reproducing artistic styles [87].

4.3.2 Advanced Art Analysis

More recent work leverages advanced LMMs for complex analysis tasks:

GalleryGPT addressed the need for deeper interpretation by introducing the task of generating comprehensive formal analysis for paintings, explicitly focusing on visual characteristics like color, composition, and light [16]. GalleryGPT employs a large dataset of paintings paired with LLM-generated formal analyses and uses a fine-tuned version of ShareGPT4V to push the model to perceive artistic skills rather than simply retrieving memorized knowledge (a phenomenon called "LLM-biased visual hallucination"). The GalleryGPT framework notably keeps the visual encoder frozen during the fine-tuning of the projection and LLM layers, an alignment in methodological spirit with the present work [16].

4.3.3 Current Approach: Style and AI-Generated Art

While sharing goals with works like GalleryGPT in investigating art perception, the current thesis presents a significant methodological distinction by incorporating the explicit comparison of human-made versus AI-generated artworks.

The investigation utilizes datasets of classical human-made art alongside synthetic images from collections like the AI-Pastiche dataset [18]. This critical duality allows the research to probe a fundamental question: To what extent can CLIP, as an unmodified perceptual system, recognize the subtle, systematic artifacts, inconsistencies, and aesthetic divergences introduced by generative processes that mimic historical styles?.

The existing literature on CLIP adaptation primarily aims at enhancing performance through the introduction of adapters such as CLIP-Adapter, TIP-Adapter, and APE, the development of prompt optimization methods like CoOp and CoCoOp, or the integration of CLIP into large multimodal models including BLIP-2, InstructBLIP, and ShareGPT4V. In contrast, the present work adopts a different perspective. Rather than modifying or fine-tuning the model, this thesis undertakes a fundamentally interpretative inquiry into CLIP's native capabilities. The vision encoder is treated as a fixed, pre-trained perceptual system, and the analysis seeks to uncover the inductive biases and visual priors that emerge from contrastive pre-training alone. This orientation is particularly important in the context of aesthetic evaluation, where the objective is not to optimize downstream performance but to investigate the representational limits of CLIP when confronted with nuanced artistic and stylistic judgments.

Existing studies indicate that while generative models can produce visually compelling outputs, they often fail to capture the deeper artistic principles, composition, and context of historical styles, often defaulting to surface-level details and hyperrealism [18]. This research focuses on detecting these critical shortcomings—such as failures in prompt adherence, compositional errors (e.g., missing figures), and stylistic artifacts, through the lens of CLIP's raw embeddings. Since prior analysis suggests that CLIP struggles with stylistic detail, period attribution, and identifying visual defects in synthetic imagery, this work aims to provide a structured critique of how CLIP's latent space represents complex aesthetic attributes and differentiates between authentic human creativity and statistical AI-driven imitation. By analyzing the visual features of style and artifacts in AI-generated art, this thesis clarifies the distinct space for research in critically evaluating the perceptual fidelity of foundational vision models like CLIP.

Chapter 5

Datasets and Experimental Methodology

This chapter outlines the rigorous methodological framework used to empirically investigate the perceptual capabilities and inherent biases of the CLIP model within the artistic domain. It begins by detailing the primary data sources: the National Gallery of Art Dataset (NGAD), which serves as a baseline of human-made art, and three collections of synthetic imagery: AI-Pastiche, AI-ArtBench, and AI-WikiArt. This strategic combination of real and AI-generated artworks facilitates a multifaceted assessment of CLIP's performance. Subsequently, the chapter describes the data preparation pipeline, including the generation of concise textual summaries from artwork descriptions to meet CLIP's input constraints. Finally, it details the experimental procedures for assessing image-text alignment, style recognition, and the model's ability to discern synthetic artifacts, thereby establishing the empirical foundation for the findings presented in the subsequent chapter.

5.1 Dataset

To construct a comprehensive evaluation, this research utilizes four distinct datasets, each selected to probe a specific aspect of CLIP's analytical capabilities. The National Gallery of Art Dataset (NGAD) provides a foundational corpus of authentic, human-created paintings. This is complemented by three specialized AI-generated datasets: the AI-Pastiche dataset is used for controlled evaluations of stylistic emulation and artifact detection; the AI-ArtBench dataset offers a large-scale benchmark for both AI art detection and style classification; and the AI-WikiArt dataset enables a robust, dual-domain assessment of artist attribution for real works versus their AI-generated counterparts. Together, these datasets create a holistic environment for testing the core hypotheses of this thesis.

5.1.1 National Gallery of Art Dataset (NGAD)

The National Gallery of Art Dataset (NGAD), utilized in this research, is a subset derived from the publicly available collection of over 130,000 pieces held by the National Gallery of Art (NGA) in Washington. The creation of this dataset involved selecting 4,436 artworks, focusing exclusively on paintings and drawings. Sculptures and modern artworks were deliberately excluded, as these pieces might pose greater challenges for analysis by models such as CLIP.

The NGAD is richly annotated, with each record comprising 11 key attributes. These attributes include:

- *"objectid"*: A unique identifier.
- *"title"*: the title of the artwork.
- *"period"*: the creation period.
- *"artist"*: the artist or attribution.
- *"link"*: a high-resolution image link provided via the IIIF protocol.

A defining feature of the dataset is the *"description"* attribute, which offers a detailed textual representation of each artwork. These descriptions are structured to convey visual elements, artistic techniques, and compositional aspects of the works.

The dataset covers a wide spectrum of historical periods and artistic styles, specifically reflecting a focus on European artistic traditions. The most heavily represented styles within the collection include Impressionist (693 artworks), Baroque (677 artworks), Realist (584 artworks) and Renaissance (535 artworks).

While core fields like *"objectid"*, *"title"*, *"period"*, *"artist"*, *"link"* and *"style"* are fully populated, the dataset does contain some missing values for other fields, specifically: *description* (3180 instances missing), ‘technique‘ (1708 instances missing),‘keyword‘ (906 instances missing), *"theme"* (227 instances missing) and *"school"* (66 instances missing).

5.1.2 AI-Pastiche Dataset

The AI-Pastiche dataset is a curated collection of 953 AI-generated paintings specifically engineered to mimic historical artistic styles. It originated from a larger research project dedicated to providing a critical assessment of the capabilities and limitations of contemporary generative models in accurately replicating artistic styles of the past [18].

The creation of the AI-Pastiche dataset involved comparing twelve modern generative models, comprising DallE3, StableDiffusion1.5, StableDiffusion3.5large, Flux1.1Pro, Flux1Schnell, Omnigen, Ideogram, Kolors1.5, FireflyImage3, LeonardoPheonix, MidjourneyV6.1 and Auto-Aestheticsv1. These models were evaluated using 73 meticulously crafted, uniform textual prompts that span a broad range of painting styles across five centuries [18].

Key characteristics and content details include:

- Scale and Curation: a total of over 20 images were generated per prompt across the selected models, from which the highest-quality samples were manually selected to form the final set of 953 images.
- Artistic Scope: The dataset covers a wide range of artistic periods, though a majority of the paintings emulate styles from the XIX-th and XX-th centuries. The most heavily represented artistic movements include Renaissance, Impressionism, Romanticism, and Baroque.
- Rich Annotation: The dataset is richly annotated with comprehensive metadata detailing the generation process. This metadata includes the specific generative model used, the full prompt text, the intended style and period, and a list of subject descriptors (e.g., "crowd", "landscape", or "soft tones").
- External Validation: The dataset creation included extensive user surveys designed to collect feedback on the perceived "authenticity" and prompt adherence of the AI-generated images.

The dataset is highly suitable for research into stylistic imitation, model benchmarking, user perception studies, and applications like deepfake detection and digital forensics.

5.1.3 AI-Artbench

The AI-ArtBench dataset is a novel, extensive collection of artworks specifically introduced for benchmarking artwork generation and evaluating models designed to detect and classify AI-generated art and style. It originated from the efforts of researchers at Carnegie Mellon University and the University of California, Berkeley, within the context of the ArtBrain project [88].

The AI-ArtBench dataset comprises over 185,015 artistic images across 10 distinctive artistic styles. A crucial characteristic of this dataset is its balance and comprehensiveness in representing both human and AI-generated content:

- Human-Created Art: It includes 60,000 human-drawn images, originally derived from the ArtBench-10 dataset. This set provides 6,000 images for each of the 10 artistic styles.

- AI-Generated Art: It contains 125,015 AI-generated paintings, which were synthesized using two prominent diffusion-based models: Latent Diffusion and Stable Diffusion. These models generated higher-quality images by conditioning the diffusion process using textual embeddings.
- Artistic Styles: The dataset is annotated across 10 distinct styles, including major movements like Baroque, Expressionism, Impressionism, Post-Impressionism, Realism, Renaissance, Romanticism, Surrealism, Art Nouveau, and Ukiyo-e.
- Resolution and Format: The human-drawn images are provided at 256×256 resolution. The synthetic images include versions in both 256×256 and higher 768×768 resolutions.
- Quality and Annotation: ArtBench-10, the human-art subset, is notably characterized as being class-balanced, high-quality, cleanly annotated, and standardized. This dataset avoids the typical long-tail class distributions often seen in previous art datasets [88].

The dataset is highly class-balanced within its splits, ensuring an unbiased evaluation when distinguishing between source models (AI vs. Human) and styles.

5.1.4 AI-WikiArt Dataset

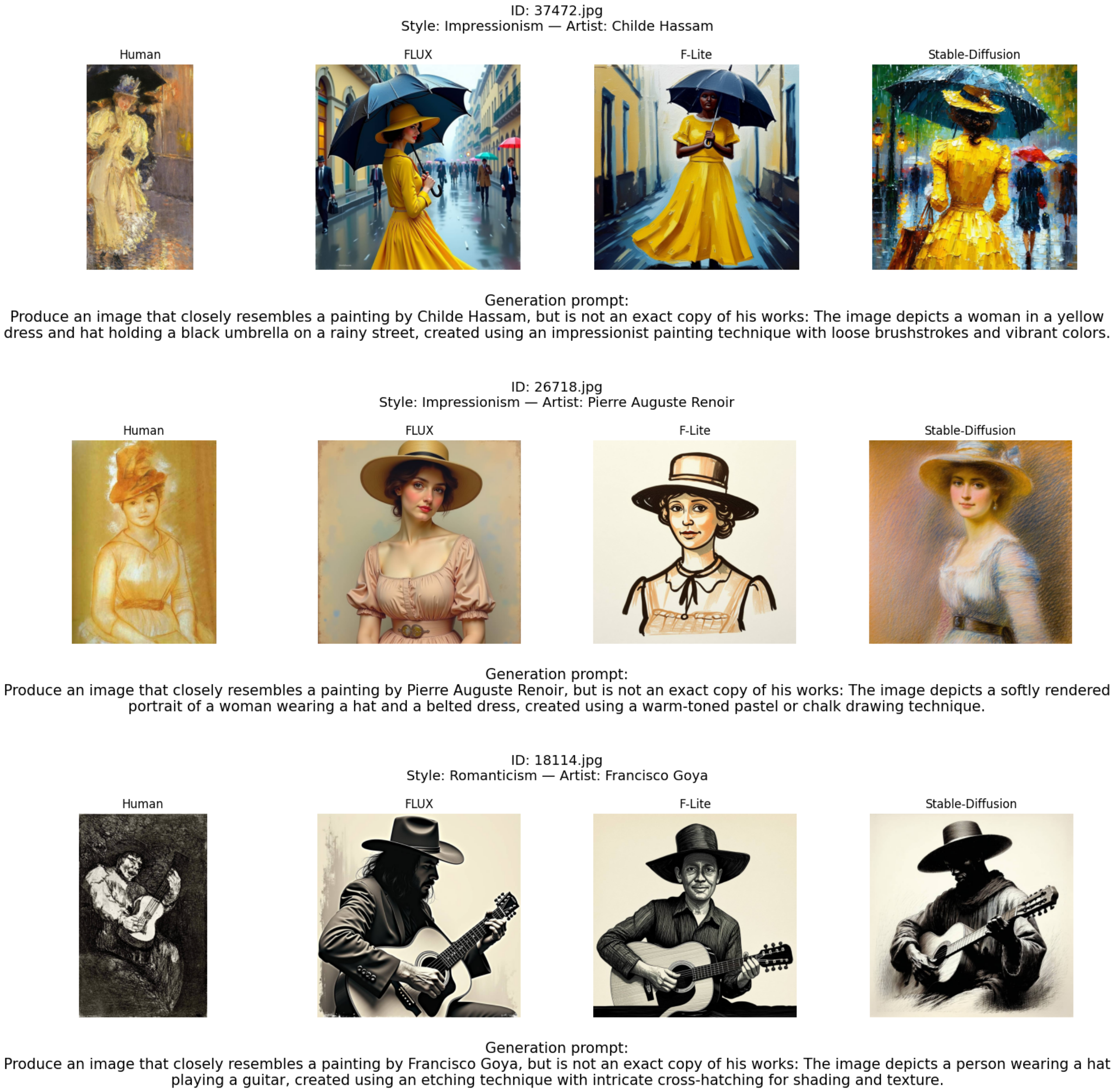
The AI-WikiArt dataset, is a comprehensive collection originally compiled to facilitate the systematic large-scale evaluation of Vision-Language Models (VLMs) in discerning artistic authorship and identifying synthetically generated content [89].

The AI-WikiArt dataset is notable for its substantial scale and dual composition of authentic and machine-generated art.

1. Real Paintings Subset: The dataset contains 39,530 real paintings. Only images that had identifiable artists in the original metadata were retained, leading to the exclusion of works labeled as "unknown". This curated set covers a wide diversity of art based on annotations for 128 artists, spanning 10 genres, and representing 27 distinct styles. The core metadata provided for each image includes the artist, genre, and style.
2. AI-Generated Counterparts: A defining characteristic is the creation of a matched set of AI-generated counterparts designed to mimic the historical paintings. The process for generating these images involved a two-step, automated pipeline:
    - Caption Extraction: Descriptive captions for the original 39,530 real WikiArt images were extracted using the language model GPT4.1-mini.
    - Image Generation: These captions were subsequently used as textual prompts to

generate imitation images using three powerful text-to-image models: Stable Diffusion, Flux, and F-Lite.

The prompt used during generation was meticulously structured to request the AI models to "Produce an image that closely resembles a painting by [correct painter], but is not an exact copy of his works". This methodological choice ensured that the generated dataset remained homogeneous with the real dataset in terms of image content and type, allowing for fair and meaningful comparative experiments. Some examples of paintings in AI-WikiArt are in Figure 5.1.


ID: 37472.jpg
Style: Impressionism — Artist: Childe Hassam
Human
FLUX
F-Lite
Stable-Diffusion

Generation prompt:
Produce an image that closely resembles a painting by Childe Hassam, but is not an exact copy of his works: The image depicts a woman in a yellow dress and hat holding a black umbrella on a rainy street, created using an impressionist painting technique with loose brushstrokes and vibrant colors.

ID: 26718.jpg
Style: Impressionism — Artist: Pierre Auguste Renoir
Human
FLUX
F-Lite
Stable-Diffusion

Generation prompt:
Produce an image that closely resembles a painting by Pierre Auguste Renoir, but is not an exact copy of his works: The image depicts a softly rendered portrait of a woman wearing a hat and a belted dress, created using a warm-toned pastel or chalk drawing technique.

Generation prompt:
Produce an image that closely resembles a painting by Francisco Goya, but is not an exact copy of his works: The image depicts a person wearing a hat playing a guitar, created using an etching technique with intricate cross-hatching for shading and texture.

Figure 5.1: Example of images present in AI-WikiArt Dataset

5.2 Experimental Methodology

This section details the experimental framework designed to empirically investigate CLIP's ability to extract high-level semantic and stylistic information from artworks. Our investigation is guided by a central principle: to assess the native perceptual abilities of the pre-trained CLIP model, treating it as a fixed perceptual system analogous to the human sensory apparatus. In contrast to the extensive body of research focused on enhancing performance through adaptation, our approach deliberately avoids any modification to the core model. Consequently, no fine-tuning, adapters, or prompt optimization techniques were employed. This methodological choice allows for a direct analysis of the inductive biases and visual priors encoded in CLIP's pretrained form, isolating the representational structures learned from contrastive pre-training alone.

To achieve this, we designed a series of experiments to evaluate CLIP across multiple interpretive dimensions. These experiments assess its capacity for fundamental image-text alignment, its more nuanced understanding of artistic style, and its sensitivity to the visual characteristics of AI-generated art by comparing its judgments on prompt adherence and synthetic artifacts against human evaluations. Collectively, this framework provides a structured critique of how CLIP's latent space represents complex aesthetic attributes and to what extent its computational perception aligns with human interpretation.

5.2.1 Experiment 1: Image-Text Alignment

The first experiment was designed to quantitatively assess CLIP's core capability: its ability to align visual representations of artworks with their corresponding linguistic descriptions. This task serves as a baseline for evaluating the model's cross-modal understanding in the artistic domain, probing the extent to which the shared embedding space captures meaningful semantic correspondences between images and text. The experiment was systematically conducted across both human-made and AI-generated datasets to investigate the robustness of this alignment.

The underlying methodology for this experiment hinges on leveraging CLIP's dual-encoder architecture to project images and texts into a common, high-dimensional feature space, where their semantic similarity can be measured. Formally, given an image I_i and a text snippet T_j, their respective embeddings, $v_i \in \mathbb{R}^D$ and $t_j \in \mathbb{R}^D$, are generated by the image encoder f_I and the text encoder f_T:

$$v_i = f_I(I_i) \tag{5.1}$$

$$t_j = f_T(T_j) \tag{5.2}$$

The degree of alignment between the image and the text is then quantified by the cosine similarity between their L2-normalized embedding vectors. The similarity score $S_{i,j}$ is computed as:

$$S_{i,j} = \frac{v_i \cdot t_j}{\|v_i\| \|t_j\|} \tag{5.3}$$

This score, which ranges from -1 to 1, provides a quantitative measure of the semantic proximity between the visual and textual content in CLIP's latent space.

This general procedure was specifically adapted to the unique characteristics of each dataset used in this thesis:

- **National Gallery of Art Dataset (NGAD):** The experiment was performed on a curated subset of 1,521 artworks for which detailed curatorial descriptions were available. For each image in this subset, we computed its cosine similarity against the textual summaries of all 1,521 descriptions. This setup framed the task as an **information retrieval** problem. The performance was evaluated using the **Recall@k** metric, which measures the percentage of images for which the correct corresponding summary is ranked within the top k most similar texts.
- **AI-Pastiche Dataset:** The analysis was conducted by calculating the alignment between each generated image and the summary of the textual prompt used to create it. Given the comparatively small and repetitive set of unique prompts (73 in total), this task was treated as a multi-class classification problem. The primary evaluation metric was **accuracy**, defined as the proportion of images for which the corresponding prompt summary achieved the highest similarity score among all possible prompts.
- **AI-WikiArt Dataset:** For this dataset, the experiment was focused exclusively on the subset of AI-generated images. The alignment was calculated between each synthetic artwork and the caption that was used as its generative prompt. In this case, there are 39,471 prompts, each of which is used to generate 3 images with the 3 different generative models, for a total of 118,590 images so it was evaluated as an information retrieval task using the Recall@k metric.

A key methodological challenge had to be addressed before the alignment could be computed. The CLIP architecture imposes a strict constraint on its text encoder, which can only process a maximum sequence length of 77 tokens. This limitation was frequently exceeded by the detailed, lengthy curatorial descriptions found in the NGAD, as well as by the elaborate prompts used to generate images in the AI-Pastiche and AI-WikiArt datasets. A naive truncation of these texts was deemed unacceptable, as it would lead to an arbitrary loss of potentially crucial semantic and stylistic information. To overcome this, we implemented a systematic preprocessing step. We generated concise summaries of each description or prompt using ChatGPT 4o-mini, explicitly instructing the language model to preserve the core elements—such as subject, style, and period—within the condensed text. This procedure ensured that the textual inputs were not only compliant with the model's technical constraints but also semantically faithful to the original source. In Appendix A.1 and A.2 there is an example of a summary of NGAD image description summary and a summary of an AI-Pastiche image generation prompt

To ensure a comprehensive analysis, this entire experimental procedure was replicated across multiple pre-trained CLIP model variants, encompassing both ResNet-based and Vision Transformer-based architectures. This allowed for a comparative study of how model capacity, architectural design, and input resolution affect image-text alignment performance within the specialized domain of fine art.

5.2.2 Experiment 2: Artistic Style Recognition

Moving beyond the direct alignment of semantic content, the second experiment was designed to probe a more abstract and challenging dimension of CLIP's perceptual capabilities: its ability to recognize and classify artistic styles. This task is significantly more complex than matching an image to a descriptive summary, as it requires the model to identify subtle, distributed visual features (such as brushwork, color palette, composition, and texture) and map them to high-level, culturally-defined concepts. The primary objective was to quantify CLIP's native performance on this task and to investigate whether a simple supervised mapping could improve upon the baseline zero-shot approach by better aligning the visual and textual embedding spaces.

The investigation began with a standard zero-shot classification setup. For each of the C distinct artistic styles present in the datasets, a standardized textual prompt was constructed following the template: "an artwork in [style] style". These prompts were then encoded using CLIP's text encoder to produce a set of style-representative textual embeddings $\{t_c\}_{c=1}^{C}$. Concurrently, all images in the test set were processed by the image encoder to obtain their corresponding visual embeddings $\{v_i\}_{i=1}^{N}$. The classification of an image v_i was determined by finding the style prompt t_c that yielded the highest cosine similarity:

$$\hat{y}_i = \arg \max_{c \in \{1,\ldots,C\}} \frac{v_i \cdot t_c}{||v_i|| ||t_c||} \tag{5.4}$$

For the NGAD dataset, performance was primarily assessed using Accuracy, Precision, Recall, F1 and Recall@k metrics. For the AI-Pastiche dataset, where the labels correspond to the intended style, the same procedure was applied. For the AI-Wikiart and AI-ArtBench datasets, the experiment was first conducted on the entire datasets and then on their respective human and AI parts to test how well CLIP can distinguish between styles in human and AI works. The metrics used are the same as those for NGAD and AI-Pastiche.

However, as anticipated and confirmed by preliminary results, this direct zero-shot approach yields modest performance. This aligns with a well-documented limitation of CLIP: the model exhibits a persistent gap between its strong grasp of semantic content and its weaker sensitivity to visual nuance. Its pre-training objective prioritizes the association between objects and their textual labels, often causing it to conflate artworks that share a similar subject matter but differ stylistically. This semantic dominance necessitates more sophisticated methods to explicitly guide the model's focus toward stylistic features.

Context: Advanced Adaptation and Alignment Methods

The challenge of fine-grained classification in CLIP has motivated the development of numerous adaptation techniques designed to enhance its performance. These methods serve as an important theoretical context for our own approach.

One prominent example is Adaptive Prior rEfinement (APE) [78], a method designed to improve CLIP's efficiency and accuracy in few-shot learning for downstream vision tasks. Its key actions are:

- **Refinement of Prior Knowledge:** APE refines CLIP's pretrained knowledge by analyzing its visual representations, starting from the observation that not all extracted visual features are equally meaningful along the channel dimension for specific downstream tasks.
- **Prior Refinement Module:** The core of the method is a module that adaptively selects the most significant feature channels based on two main criteria:
  - *Inter-Class Similarity:* The module seeks to minimize similarity between different classes, thereby identifying the most discriminative channels for classification.
  - *Intra-Class Variance:* It aims to eliminate feature channels that remain nearly constant across various categories, as these provide little to no discriminative information.
- **Efficiency Improvement:** By maximizing inter-class disparity on the few-shot training data, APE effectively discards redundant information. This not only improves accuracy but also reduces the cache size required for inference, ensuring high computational efficiency.

In its training-free variant, APE explores trilateral affinities among the test image, a cache model, and textual representations for robust recognition. The trainable version (APE-T) introduces a lightweight module of category residuals.

Another advanced approach, Selective Vision-Language Subspace Projection (SSP) [4], directly addresses the feature alignment problem in CLIP. The SSP method focuses on reducing modality gaps, which occur when text and image features corresponding to the same concept lie far apart in the unified feature space. The main characteristics of SSP are:

1. **Subspace Projection:** SSP is a training-free method that leverages local image features as a bridge to improve the alignment between global text-image pairs.
2. **Vision Projector:** It uses local image feature regions that are most similar to the global image features to construct a unified "vision subspace". The global image features are then projected into this space for better alignment.
3. **Language Projector:** For each class, it uses local image features that exhibit high semantic correlation with the corresponding text features to construct a class-specific "linguistic subspace". The text features are then projected into their respective linguistic subspaces.
4. **Matrix-Based Computations:** As a training-free method, SSP involves only matrix computations, such as Singular Value Decomposition (SVD), and can be integrated into existing frameworks like APE to further improve classification performance.

Proposed Method: Style Classification via Cross-Modal Alignment

While complex methods like SSP offer a sophisticated, training-free approach to the modality gap, and techniques like APE focus on optimizing performance in few-shot scenarios, our research takes a different methodological path. We adopt simpler, supervised methods (primarily a direct Linear Probe and our proposed Cross-Modal Alignment) not with the primary goal of outperforming these specialized techniques, but to use them as diagnostic tools. Our objective is to investigate the fundamental structure of CLIP's latent space and to determine to what extent stylistic information is linearly separable and alignable. To contextualize our findings, we did, however, benchmark our methods against APE to assess their performance relative to a well-established adaptation technique. Our primary proposed method, Cross-Modal Alignment via logistic regression, is detailed below.

This experiment aims to evaluate CLIP's capability to recognize and align artworks with their corresponding artistic styles, leveraging both its image and text embedding spaces. Our objective is to examine whether the visual features extracted by CLIP can be effectively mapped into the semantic space induced by textual style descriptions, thereby enabling accurate cross-modal retrieval and classification.

We consider a dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where each image x_i is annotated with a style label $y_i \in \{1, \ldots, C\}$. Two sets of embeddings are extracted:

- **Visual embeddings** $v_i \in \mathbb{R}^D$, obtained by encoding each image with CLIP's image encoder.
- **Textual embeddings** $t_c \in \mathbb{R}^D$, computed for each style c by encoding the prompt "an artwork in {style} style".

Our key finding was that classification performance improved substantially when using unnormalized embeddings, suggesting that the magnitude of the feature vectors carries important discriminative information for style, which is discarded during L2 normalization. Consequently, all subsequent steps were performed on unnormalized embeddings.

To bridge the modality gap, we first characterized the structure of each space independently by training separate multiclass logistic regression classifiers. A classifier was trained on the visual embeddings to learn a weight matrix $W_{\text{img}} \in \mathbb{R}^{C \times D}$, where each row vector w_c^{img} corresponds to the learned decision boundary for style c. An analogous classifier was trained on the textual embeddings to yield a weight matrix $W_{\text{txt}} \in \mathbb{R}^{C \times D}$.

These learned weight matrices implicitly capture the modality-specific features most relevant for style discrimination. We leveraged them to construct a linear transformation matrix $T \in \mathbb{R}^{D \times D}$ designed to project visual embeddings into the semantic space of the textual classifier:

$$T = W_{\text{txt}}^{\top} W_{\text{img}} \tag{5.5}$$

This transformation is derived by solving the ordinary least squares problem to find the optimal mapping from the visual classifier's weights to the textual classifier's weights. The transformed representation of a visual embedding v_i is then given by $v_i' = Tv_i$, which is intended to be semantically aligned with its corresponding textual style embedding.

Finally, the similarity between a transformed visual embedding v_i' and a textual style embedding t_c is computed as their dot product:

$$S_{c,i} = t_c^{\top} v_i' = t_c^{\top} T v_i$$

This produces a similarity matrix $S \in \mathbb{R}^{C \times N}$. Performance was assessed in both classification and retrieval scenarios, using accuracy, precision, recall and macro-averaged F1-score for the former, and Recall@k for the latter. This approach not only provides a robust measure of

CLIP's style recognition accuracy but also offers insights into the degree of linear separability and semantic alignment between its visual and textual embedding spaces for the complex task of style perception.

5.2.3 Experiment 3: Adherence Analysis (on AI-Pastiche)

While the previous experiments quantified CLIP's ability to perform classification tasks on discrete labels (retrieving a description or identifying a style), this third experiment delves into a more nuanced and subjective problem: evaluating the model's capacity to assess the quality of alignment between a generative prompt and its resulting image. The primary objective is to determine whether CLIP's quantitative measure of similarity correlates with qualitative human judgments of prompt adherence. This is a critical test, as any discrepancy would highlight a divergence between the model's learned feature space and the perceptual cues that humans deem important for semantic and stylistic consistency.

The experiment was conducted exclusively on the AI-Pastiche dataset, which is uniquely suited for this analysis due to its associated human evaluation data. As detailed in [18], an extensive adherence survey was conducted where human participants were shown a set of images generated by different models from the same textual prompt. In a comparative setting, they were asked to rate each image as "good", "neutral", or "bad" based on its perceived adherence to the prompt. By averaging the responses from multiple participants, the authors of the dataset derived a continuous *adherence score* for each image, which serves as a ground-truth measure of perceived stylistic and semantic alignment.

To investigate whether CLIP captures similar perceptual cues, we designed a parallel analysis using the model's native similarity scores. The procedure was as follows:

1. For each of the unique textual prompts P_j in the AI-Pastiche dataset, we first computed its corresponding text embedding t_j using CLIP's text encoder.

2. For each image I_i generated from prompt P_j, we computed its image embedding v_i using CLIP's image encoder.

3. The cosine similarity $S_{ij} = \frac{v_i \cdot t_j}{\|v_i\| \|t_j\|}$ was calculated between the image embedding and its corresponding prompt embedding. This produced a raw "CLIP adherence score" for each image.

4. To enable a direct comparison with the human-derived scores, the vector of CLIP's similarity scores for a given prompt was normalized and scaled.

Finally, to quantify the degree of alignment between machine and human perception, we computed the cosine similarity between the vector of CLIP's normalized similarity scores and the vector of human adherence scores. This cosine similarity serves as a synthetic measure of how well CLIP's assessments align with human perception. A high positive cosine similarity would suggest that the features CLIP deems important for similarity are congruent with those valued by human observers, while a low cosine similarity would indicate a significant perceptual gap.

This entire procedure was systematically repeated across all available CLIP model variants to assess whether architectural differences influence the model's ability to emulate human-like judgment of generative quality.

5.2.4 Experiment 4: Perception of Artifacts and Deformations (on AI-Pastiche)

The final experiment addresses a critical aspect of evaluating AI-generated art: the presence of visual artifacts. Human assessment of prompt adherence is not solely based on semantic and stylistic correspondence; it is also heavily influenced by the overall technical quality of the image, specifically the absence of the characteristic distortions, inconsistencies, and anatomical errors common in generative models [18]. This experiment was designed to investigate whether CLIP's internal representations are sensitive to such structural flaws. The core hypothesis is that CLIP, having been pre-trained to prioritize semantic content, may be largely "blind" to low-level artifacts that do not significantly alter the high-level interpretation of a scene.

This investigation leverages the human-annotated survey data from the AI-Pastiche dataset, which specifically targeted the presence of visible artifacts. In the survey, defects were categorized as:

- *Major*: Clearly visible or frequent errors (e.g., severe anatomical mistakes).
- *Minor*: Less critical imperfections (e.g., extra fingers, small distortions).
- *None*: No apparent visual defects.

From these evaluations, each image in the dataset was assigned a continuous *defect score*, providing a ground-truth measure of its perceived technical quality.

Our analysis proceeded in two stages. First, we conducted a direct test to determine if artifact-related information could be linearly decoded from CLIP's visual embeddings. We attempted to predict the human-annotated defect score (d_i) using the CLIP image embedding (v_i) as input to a linear regression model. This initial approach yielded a coefficient of determination (R^2) close to zero, providing strong evidence that CLIP's global feature representations do not explicitly or linearly encode information about such visual flaws.

Second, to further explore the impact of this perceptual "blindness", we designed an experiment to test whether augmenting CLIP's similarity score with external artifact information could improve its alignment with overall human adherence judgments. We introduced a linear combination model that integrates CLIP's perceptual similarity score with the human-annotated defect score. Given the normalized perceptual similarity $\hat{s}_i$ for image i (from Experiment 3) and its defect score d_i, we compute a new predicted adherence score $\tilde{y}_i$ as:

$$\tilde{y}_i = a \cdot \hat{s}_i + b \cdot d_i + c \tag{5.6}$$

where a and b are learned coefficients representing the weights of CLIP's similarity and the defect information, respectively, and c is a bias term. The optimal coefficients are estimated by minimizing the squared error between the model's prediction $\tilde{y}_i$ and the true human-annotated adherence score y_i, using the ordinary least squares method:

$$\min_{a,b,c} \sum_{i=1}^{n} \left(y_i - (a \cdot \hat{s}_i + b \cdot d_i + c)\right)^2 \tag{5.7}$$

Finally, the resulting amended score vector $\tilde{\mathbf{y}}$ is compared to the human adherence vector $\mathbf{y}$ by computing their correlation. An increase in correlation compared to the results from Experiment 3 would provide indirect but compelling evidence that CLIP's evaluation of adherence is missing a critical component, sensitivity to artifacts, that is central to human perception. This outcome would underscore a fundamental limitation in using CLIP as a standalone evaluator for the quality of generated art.

Chapter 6

Experimental Results and Analysis

Following the methodological framework established in Chapter 5, this chapter presents and critically analyzes the empirical results obtained from the series of experiments designed to probe CLIP's perceptual capabilities in the artistic domain. The objective is to translate the experimental procedures into quantitative data and qualitative insights, thereby forming a comprehensive assessment of the model's strengths, weaknesses, and intrinsic biases when confronted with human-made and AI-generated art.

The analysis is structured to mirror the sequence of the experiments, progressing from a foundational assessment of cross-modal alignment to more nuanced evaluations of stylistic perception and human-centric judgments. We begin by examining the results of the image-text alignment task, which serves as a baseline for CLIP's core semantic understanding. Subsequently, we delve into the more challenging task of artistic style recognition, comparing the model's performance across different datasets and evaluation strategies. The final sections are dedicated to a direct comparison between CLIP's computational assessments and human perception, focusing on the subjective evaluation of prompt adherence and the model's sensitivity (or lack thereof) to synthetic artifacts in AI-generated images. Collectively, these results provide a data-driven critique of how CLIP "sees" art, revealing the extent to which its computational perception aligns with, and diverges from, human interpretation.

6.1 Image-Text Alignment Results (Experiment 1)

6.1.1 Quantitative Analysis on NGAD

The initial experiment provides a foundational assessment of CLIP's core cross-modal alignment capability within the domain of human-created art. The results for the image-to-summary retrieval task on the NGAD are presented in Table 6.1. This task was framed as an information retrieval problem, where each of the 1,521 artworks was matched against the entire corpus of 1,521 textual summaries, with performance measured using the Recall@k metric.

Model	recall@1	recall@5	recall@10
RN50	0.663	0.915	0.966
RN101	0.693	0.926	0.966
RN50x4	0.741	0.946	0.978
RN50x16	0.791	0.964	0.988
RN50x64	**0.828**	0.970	0.990
ViT-B/32	0.678	0.925	0.970
ViT-B/16	0.709	0.928	0.969
ViT-L/14	0.794	0.972	0.989
ViT-L/14@336px	0.814	**0.974**	**0.991**

Table 6.1: Summary–image alignment performance on the NGAD subset. The table shows Recall@k scores for various CLIP model architectures, demonstrating the model's ability to retrieve the correct textual summary for a given artwork.

The results reveal a clear and consistent trend: performance in aligning images with their textual descriptions improves directly with the scale and capacity of the CLIP models. This holds true for both the ResNet and Vision Transformer (ViT) architectural families.

Within the ResNet-based models, there is a steady improvement in recall scores as the model size increases. The baseline RN50 achieves a Recall@1 of 66.3%, which progressively climbs to a robust 82.8% for the RN50x64 model, the top performer in this family. This indicates that the increased computational capacity allows the model to capture more nuanced visual features that are essential for distinguishing between artworks with potentially similar subjects or compositions.

A similar pattern is observed for the ViT-based models. The larger architectures and, notably, higher input resolutions, lead to superior alignment. The "ViT-L/14@336px" model, which processes images at a higher resolution, achieves the best overall performance among the transformer variants and is the top performer at higher recall thresholds, with a Recall@5 of 97.4% and a Recall@10 of 99.1%. This suggests that for fine-grained association tasks, such as matching a detailed artwork to a concise summary, increased input resolution provides critical visual information that enhances the model's discriminative power.

Overall, the high recall scores across the board confirm that CLIP's fundamental ability to map visual and linguistic concepts to a shared semantic space is highly effective, even in the specialized and complex domain of fine art. The best models, "RN50x64" and "ViT-L/14@336px", demonstrate a remarkable capacity to correctly identify the specific textual description for an artwork out of over 1,500 possibilities more than 80% of the time on the first attempt. This strong baseline performance in capturing semantic content serves as a crucial point of comparison for the more abstract and challenging task of style recognition, which will be examined in the following section.

6.1.2 Quantitative Analysis on AI-Pastiche and AI-WikiArt

To complement the analysis on human-created art, the image-text alignment experiment was extended to two datasets composed of AI-generated images: AI-Pastiche and AI-WikiArt. This phase of the investigation was designed to assess CLIP's ability to form semantic correspondences when dealing with synthetic visual data, providing insight into both the model's robustness and the output quality of the generative systems themselves.

Performance on AI-Pastiche

The AI-Pastiche dataset, with its structure of multiple images generated from a limited set of 73 distinct prompts, framed the alignment task as a multi-class classification problem. The results, presented in Table 6.2, show a consistently high level of accuracy across all evaluated CLIP model variants.

Model	**RN50**	**RN101**	**RN50x4**	**RN50x16**	**RN50x64**	**ViT-B/32**	**ViT-B/16**	**ViT-L/14**	**ViT-L/14@336px**
Accuracy	0.866	0.887	0.891	0.893	**0.896**	0.881	0.880	**0.896**	**0.896**

Table 6.2: Accuracy of different CLIP models in matching generated images with their corresponding summarized prompts in the AI-Pastiche dataset.

Performance is strong across the board, with every model exceeding 86% accuracy. The top-performing models ("RN50x64", "ViT-L/14", and "ViT-L/14@336px"), all achieve an impressive accuracy of 89.6%. This task is inherently simpler than the retrieval task on NGAD, as the model only needs to distinguish between 73 relatively distinct semantic concepts. Nevertheless, these results confirm CLIP's robust capability to capture visual-semantic correspondences even in synthetically generated image-text pairs.

From a generative standpoint, the high accuracy scores indicate that the underlying image generation models were largely successful in producing outputs that are semantically aligned with the subject matter described in the prompts. The mean cosine similarity between each generated image and its corresponding prompt summary is 0.278, with a standard deviation of 0.344. However, this finding comes with a critical caveat: while cosine similarity proves effective for associating an image with its intended prompt, its reliability as a standalone metric for evaluating the overall *quality* of the generation is uncertain. A comprehensive assessment of quality requires evaluating not only the semantic correspondence but also the stylistic fidelity and technical execution, including the absence of visual artifacts. This limitation will be addressed in subsequent experiments.

Performance on AI-WikiArt

The AI-WikiArt dataset presented a different and more demanding challenge. The experiment was focused exclusively on its large AI-generated subset, which contains 118,590 images generated from 39,471 unique prompts, where each prompt was used to create three distinct images via three different generative models. This structure framed the task as a large-scale information retrieval problem, where for each image, the model had to retrieve the correct generative prompt from the entire corpus of nearly 40,000 candidates. The results for this task are presented in Table 6.3.

	Accuracy	**Precision**	**Recall**	**F1**	**Recall@1**	**Recall@5**	**Recall@10**
AI-WikiArt (AI)	0.4258	0.4719	0.4262	0.4057	0.4258	0.6565	0.7416

Table 6.3: Prompt-image alignment performance for the AI-generated subset of the WikiArt dataset, measured with the "ViT-L/14@336px" architecture.

The performance on the AI-WikiArt dataset is noticeably lower than that observed on NGAD, with a Recall@1 of just 42.58%. This disparity is attributable to two main factors that significantly increase the complexity of the task. First is the sheer scale of the retrieval space: discriminating the correct prompt from a pool of nearly 40,000 candidates is substantially more difficult than from the 1,521 candidates in the NGAD experiment. Second is the nature of the textual data itself. The prompts in AI-WikiArt, having been generated by an LLM to be descriptive, are far more varied, complex, and nuanced than the concise, structured summaries used for the NGAD experiment. This creates a much denser and more semantically complex space for retrieval, where subtle differences in phrasing can lead to lower similarity scores and incorrect rankings. The results underscore how sensitive retrieval performance is to both the scale and the nature of the text corpus, even when the underlying visual-semantic alignment capability of the model remains constant.

6.1.3 Qualitative Error Analysis

While the quantitative results demonstrate a strong overall performance in matching images to their correct textual descriptions, a qualitative analysis of the model's failures is essential for understanding the nature and limitations of its semantic understanding. These errors reveal that while CLIP successfully captures high-level thematic content, it often falters when confronted with finer-grained semantic details, stylistic nuances, and specific contextual information.

A representative example of such a misclassification on the NGAD dataset, committed by the "ViT-L/14@336px" model, is presented in Figure 6.1. The model incorrectly associated a Baroque painting by Canaletto with a summary describing a Rococo work by Jean-Honoré Fragonard.

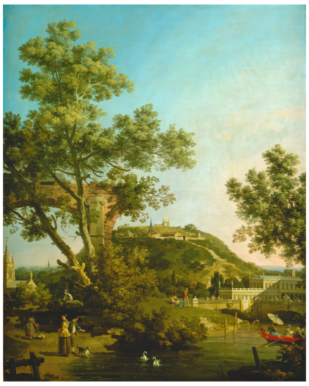


Figure 6.1: An example of a summary misclassification on the NGAD dataset. The model incorrectly associates Canaletto's Baroque landscape with the description of a Rococo scene by Fragonard.

Actual Artwork: A Baroque painting by Canaletto.

> *In this Baroque painting by Canaletto (1751-1775), sunlight illuminates a vibrant landscape featuring an arched stone ruin, a bridge, and a river. Scattered figures in colorful attire engage with nature, while buildings and churches rise on a distant hill. The scene is rich with detail, showcasing Venetian life and architecture.*

Incorrectly Predicted Summary:

> *In this Rococo painting by Jean Honoré Fragonard (1751-1775), a lively scene unfolds in a lush park where light-skinned figures enjoy leisure by a river. A couple in elegant attire sits nearby, while boys engage in playful horse-riding games. Tall trees frame the idyllic landscape, enhancing the theme of amusement.*

The incorrect predicted summary description corresponds to artwork in the Figure 6.2.

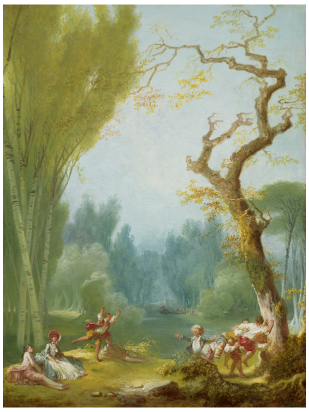


Figure 6.2: Actual artwork that corresponds to the Jean Honoré Fragonard artwork description

At a coarse semantic level, the confusion is understandable. Both the image and the predicted summary describe an 18th-century European landscape painting featuring figures near water and architectural or natural elements. CLIP correctly identifies the general scene composition. However, a closer inspection reveals critical failures in comprehension:

- **Content Mismatch:** The model overlooks the specific, defining objects in the painting. The prominent architectural features—an arched stone ruin and a bridge—are central to Canaletto's work, but they are entirely absent from the predicted summary, which instead describes a generic "lush park" and "tall trees".
- **Stylistic and Contextual Mismatch:** The model confuses two distinct artistic styles and cultural contexts. It misattributes a Venetian *veduta*, characteristic of the Baroque period's detailed cityscapes, to a French Rococo *fête galante*, which typically depicts idyllic and leisurely aristocratic scenes. This error foreshadows the difficulties in fine-grained style recognition that will be explored in the next section.
- **Atmospheric Mismatch:** The actual summary correctly captures the scene's focus on "Venetian life and architecture", while the predicted summary evokes a different mood of "amusement" and "playful games", which is not the primary theme of the artwork.

This example is emblematic of a broader pattern in CLIP's errors. It demonstrates that the model's alignment is predominantly driven by high-level object and scene recognition. While it can successfully identify broad categories like "landscape with figures", it struggles to differentiate based on the specific attributes of those objects, their artistic rendering, or their historical and cultural significance. This highlights a fundamental gap between coarse semantic matching and true, nuanced visual comprehension.

6.2 Artistic Style Recognition Results

Having established CLIP's robust performance in aligning images with their direct semantic descriptions, this section transitions to a more abstract and challenging dimension of its perceptual capabilities: the recognition of artistic style. This task moves beyond the identification of objects and scenes, probing the model's sensitivity to the subtle, distributed visual features (such as brushwork, color palette, composition, and texture) that constitute an artistic style. As foreshadowed by the qualitative errors in the previous section and consistent with the literature on CLIP's semantic biases, this high-level classification is expected to be a significantly more difficult task for a model pre-trained to prioritize content over form.

The primary objective of this section is to quantitatively evaluate CLIP's native competence in style recognition and to investigate the degree to which this information is encoded within its visual feature space. To this end, we present a multi-faceted analysis. We begin by establishing a baseline with the standard zero-shot classification approach. We then explore two supervised methods, a linear probe and a cross-modal alignment, to determine whether stylistic information is linearly separable and generalizable, a critical test of true stylistic understanding versus dataset-specific pattern recognition.

6.2.1 Baseline Zero-Shot Performance

The baseline for assessing CLIP's intrinsic capability for artistic style recognition was established via the standard zero-shot classification paradigm. In this approach, image embeddings are directly compared against text embeddings synthesized from prompts formatted as "an artwork in [style] style". The results from this experiment, utilizing the "ViT-L/14@336px" architecture across all evaluated datasets, are systematically presented in Table 6.4.

Dataset	Accuracy	Precision	Recall	F1	Recall@1	Recall@3	Recall@5
NGAD	0.3006	0.3578	0.3173	0.2303	0.3006	0.6080	0.7189
AI-Pastiche	0.4974	0.4499	0.4454	0.3830	0.4974	0.7696	0.8429
AI-WikiArt (Human + AI)	0.3281	0.3478	0.4808	0.2894	0.3281	0.6189	0.7607
AI-WikiArt (Human)	0.3664	0.4059	0.4630	0.3210	0.3664	0.5850	0.7189
AI-WikiArt (AI)	0.1490	0.2622	0.3050	0.1806	0.1490	0.3822	0.5454
AI-ArtBench (Human + AI)	0.6378	0.7096	0.6380	0.6367	0.6378	0.8830	0.9554
AI-ArtBench (Human)	0.5516	0.5727	0.5516	0.5156	0.5516	0.8336	0.9370
AI-ArtBench (AI)	0.6754	0.7750	0.6764	0.6841	0.6754	0.9053	0.9657

Table 6.4: Results of Artistic Style Recognition with zero-shot approach with CLIP model using "ViT-L/14@336px" architecture

The empirical data confirm that direct zero-shot style classification constitutes a formidable challenge for the CLIP model. On the NGAD, a corpus of human-created artworks, the model achieves a modest top-1 accuracy of 30.06%. This figure highlights the model's inherent difficulty in distinguishing between nuanced and often inter-related artistic styles based solely on its pre-trained representational space. Nevertheless, the Recall@3 (60.80%) and Recall@5 (71.89%) metrics are informative, revealing that while the primary prediction is often inaccurate, the correct stylistic category is frequently present within the top-k predictions. This indicates that while the latent space encodes some measure of stylistic affinity, it lacks the requisite discriminative granularity for precise classification.

6.2.2 Performance with Linear Probing

To further isolate and evaluate the richness of the visual feature space, a linear probing experiment was conducted. This standard evaluation protocol involves training a linear classifier (logistic regression) directly on the frozen visual embeddings, thereby measuring how linearly separable the style classes are within the image modality alone, without reference to the text encoder's semantic space.

Intra-Dataset Performance

To establish a baseline for supervised performance, each dataset was individually split into an 80% training set and a 20% testing set. A linear probe (logistic regression) was then trained on the embeddings of the training split and evaluated on the corresponding test split.

A critical preliminary finding was the significant impact of embedding normalization. The contrastive pre-training objective of CLIP utilizes L2-normalized embeddings for its cosine similarity calculations. However, our experiments revealed that the magnitude of the feature vectors, which is discarded during normalization, contains substantial discriminative information pertinent to artistic style. The results, presented in Tables 6.5 (normalized) and 6.6 (unnormalized), demonstrate this effect.

	Accuracy	**Precision**	**Recall**	**F1**
NGAD	0.7074	0.7418	0.7074	0.7122
AI-Pastiche	0.8482	0.8709	0.8482	0.8433
AI-WikiArt (Human + AI)	0.5380	0.5917	0.5380	0.5522
AI-WikiArt (Human)	0.7239	0.7408	0.7239	0.7278
AI-WikiArt (AI)	0.4884	0.5521	0.4884	0.5052
AI-ArtBench (Human + AI)	0.8635	0.8630	0.8635	0.8631
AI-ArtBench (Human)	0.7523	0.7514	0.7523	0.7518
AI-ArtBench (AI)	0.9471	0.9472	0.9471	0.9471

Table 6.5: Results of Artistic Style Recognition with linear probe method using CLIP "ViT-L/14@336px" architecture with normalized embedding (trained and tested on 80/20 splits of the same dataset).

	Accuracy	Precision	Recall	F1
NGAD	0.7771	0.7825	0.7771	0.7775
AI-Pastiche	0.9581	0.9692	0.9581	0.9602
AI-WikiArt (Human + AI)	0.5509	0.5971	0.5509	0.5635
AI-WikiArt (Human)	0.7528	0.7593	0.7528	0.7543
AI-WikiArt (AI)	0.5056	0.5608	0.5056	0.5206
AI-ArtBench (Human + AI)	0.8718	0.8713	0.8718	0.8715
AI-ArtBench (Human)	0.7495	0.7489	0.7495	0.7491
AI-ArtBench (AI)	0.9533	0.9534	0.9533	0.9533

Table 6.6: Results of Artistic Style Recognition with linear probe method using CLIP "ViT-L/14@336px" architecture with not normalized embedding (trained and tested on 80/20 splits of the same dataset).

The use of unnormalized embeddings proved critical. As shown by comparing the two tables, forgoing normalization consistently and significantly improved classification accuracy. For instance, on the NGAD, accuracy increased from 70.74% to 77.71%. This reinforces the conclusion that vector magnitude in CLIP's feature space is a key carrier of stylistic information. All subsequent analysis is therefore based on the results from unnormalized embeddings in Table 6.6.

The high accuracy achieved with this method, especially the exceptional 95.81% on AI-Pastiche, was initially promising. However, this particular result is misleading. Given the small size of the AI-Pastiche dataset and the fact that many images were generated from identical prompts, the random 80/20 split resulted in near-duplicate images appearing in both the training and testing sets. This means the model likely learned to recognize superficial patterns specific to those prompts rather than generalizable stylistic features.

Cross-Dataset Generalization Test

The high intra-dataset scores suggested potential overfitting. To test for true stylistic generalization, we conducted a more rigorous cross-dataset evaluation. We trained a single logistic regression classifier exclusively on the AI-WikiArt (Human) training data and tested its performance on the entirety of the other datasets, mapping styles between them. The results, shown in Table 6.7, confirm our hypothesis: performance declined significantly across all datasets, revealing that the model had learned dataset-specific patterns rather than abstract, transferable style concepts.

Dataset	Accuracy	Precision	Recall	F1
AI-WikiArt (Human) (test split)	0.7654	0.7719	0.7654	0.7670
NGAD	0.4729	0.5784	0.4729	0.4831
AI-Pastiche	0.3452	0.6018	0.3452	0.3682
AI-ArtBench (Human+AI)	0.5267	0.6410	0.5267	0.5498
AI-ArtBench (Human)	0.6024	0.6924	0.6024	0.6068
AI-ArtBench (AI)	0.4905	0.6032	0.4905	0.5093
AI-WikiArt (AI)	0.3088	0.4110	0.3088	0.3087

Table 6.7: Results of Artistic Style Recognition with linear probe method trained on AI-WikiArt (Human) and tested on other datasets.

The accuracy on AI-Pastiche plummeted from 95.81% to 34.52%, reinforcing the conclusion that the earlier result was due to pattern recognition of near-duplicates. The performance on the AI-ArtBench (AI) subset is also revealing. While it dropped from 95.33% to 49.05%, this score is still anomalously high. This is attributable to a "shared inductive bias": the generative models used for AI-ArtBench (Latent and Standard Diffusion) are themselves conditioned on CLIP. They were prompted with simple text ("A painting in <art style> art style"), with variation introduced only by changing the seed. This process creates outputs optimized for CLIP's perceptual space, where images of the same style are visually very similar and cluster tightly together. As shown in Figure 6.3, an image from this dataset and its nearest neighbor in CLIP's embedding space can be nearly identical. Consequently, the high accuracy on AI-ArtBench (AI) is considered negligible as it merely exposes CLIP's bias rather than demonstrating genuine style recognition.

6.2.3 Performance with Supervised Cross-Modal Alignment

Intra-Dataset Performance

Next, we investigated whether the modality gap for artistic style could be bridged by learning an explicit mapping from the visual to the textual space. We employed a supervised cross-modal alignment method (detailed in Chapter 5), which learns a linear transformation (T) to project CLIP's unnormalized visual embeddings into the semantic space of the textual style descriptors. The model was trained and tested on an 80/20 split of each dataset.

The results, shown in Table 6.8, initially suggested this was a highly effective approach, substantially improving performance over the zero-shot baseline. Accuracy on the NGAD dataset surged to 74.29%, and the model achieved even higher scores on cleaner datasets, reaching 93.72% on AI-Pastiche and 92.18% on the AI-generated subset of AI-ArtBench. The success of this method on intra-dataset splits indicates that the modality gap is not chaotic but structurally coherent within a specific dataset. The fact that a linear transformation can bridge the

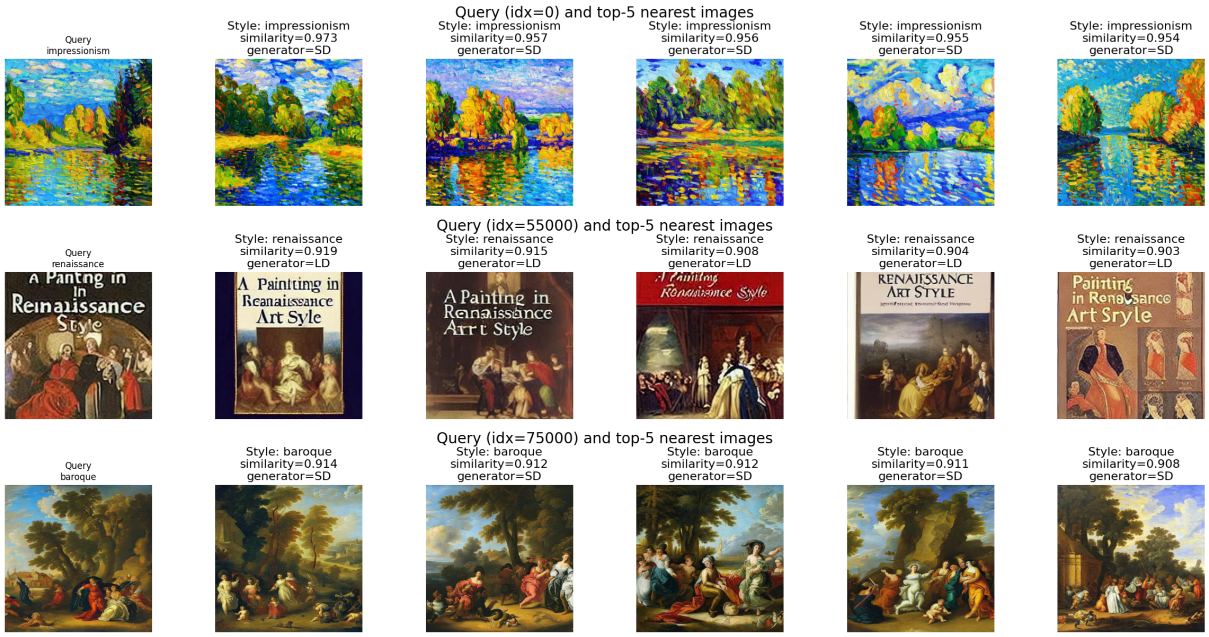


Figure 6.3: Example of AI-ArtBench generated images and their nearest neighbor, demonstrating the visual similarity that leads to inflated performance metrics.

spaces suggests that the geometric relationships between style concepts are analogous across both modalities, just oriented differently. However, these high scores must be interpreted with caution, as they likely reflect the same overfitting to dataset-specific patterns observed in the linear probing experiment.

Dataset	Accuracy	Precision	Recall	F1	Recall@1	Recall@3	Recall@5
NGAD	0.7429	0.6517	0.6517	0.6304	0.7429	0.9554	0.9863
AI-Pastiche	0.9372	0.9372	0.9665	0.9450	0.9372	1.0000	1.0000
AI-WikiArt (Human + AI)	0.3116	0.4455	0.4838	0.3784	0.3116	0.5576	0.6939
AI-WikiArt (Human)	0.6505	0.6306	0.7164	0.6230	0.6505	0.9089	0.9652
AI-WikiArt (AI)	0.2961	0.4244	0.4556	0.3627	0.2961	0.5333	0.6762
AI-ArtBench (Human + AI)	0.8304	0.8516	0.8302	0.8331	0.8304	0.9691	0.9905
AI-ArtBench (Human)	0.6809	0.7362	0.6809	0.6763	0.6809	0.9170	0.9755
AI-ArtBench (AI)	0.9218	0.9348	0.9222	0.9208	0.9218	0.9966	0.9998

Table 6.8: Results of Artistic Style Recognition with cross-alignment method using CLIP "ViT-L/14@336px" architecture with not normalized embedding

Cross-Dataset Generalization Test

To verify this suspicion, we conducted the same cross-dataset experiment with the cross-modal alignment method. The alignment transformation was trained only on the AI-WikiArt (Human) dataset and then tested on all other datasets. The results are presented in Table 6.9.

Dataset	Accuracy	Precision	Recall	F1	Recall@1	Recall@3	Recall@5
AI-WikiArt (Human)	0.6486	0.6111	0.6915	0.5943	0.6486	0.9091	0.9694
NGAD	0.4499	0.2840	0.2291	0.2283	0.4499	0.7523	0.8678
AI-Pastiche	0.3998	0.2767	0.2049	0.2153	0.3998	0.7153	0.8683
AI-ArtBench (Human + AI)	0.4625	0.2693	0.1980	0.2149	0.4625	0.8225	0.9327
AI-ArtBench (Human)	0.5202	0.2803	0.2229	0.2289	0.5202	0.8149	0.9111
AI-ArtBench (AI)	0.4350	0.2752	0.1865	0.2034	0.4350	0.8261	0.9431
AI-WikiArt (AI)	0.2779	0.2899	0.3325	0.2638	0.2779	0.5650	0.7142

Table 6.9: Cross-dataset results of the Supervised Cross-Modal Alignment method trained on AI-WikiArt (Human).

As with the linear probe, the cross-dataset performance dropped dramatically. The accuracy on NGAD fell from 74.29% to 44.99%, and on AI-Pastiche from 93.72% to 39.98%. This confirms that the supervised alignment, while powerful, also overfits to the source dataset's specific patterns and biases. It learns a mapping that is effective for a particular data distribution but fails to capture the abstract essence of artistic style required for generalization. Therefore, the high intra-dataset performance validates that stylistic information is robustly encoded, but this information can only be "unlocked" in a way that is highly sensitive to the training domain, failing to bridge the modality gap in a universally applicable manner.

6.2.4 Comparative Analysis: Human-Created vs. AI-Generated Art

The cross-dataset generalization results reveal a clear pattern: the linear probe trained on human art performs consistently better on other datasets of human art than on those containing AI-generated images. For instance, the accuracy on AI-ArtBench (Human) is 60.24%, whereas it drops to 49.05% on the corresponding AI-generated subset. This suggests that while the model learns patterns from human art, these patterns are less applicable to synthetic images, which may lack the coherent stylistic features of their human counterparts.

The particularly poor performance on AI-WikiArt (AI) (30.88%) and AI-Pastiche (34.52%) reinforces this. The generative processes used for these datasets seem to produce images that are stylistically ambiguous, making them difficult to classify even for a model trained on a diverse set of human art. Conversely, the relatively higher performance on AI-ArtBench (AI) suggests a "shared inductive bias", where the generative process, which involves CLIP, creates images that are easier for a CLIP-based classifier to understand, even if that understanding does not generalize well.

6.2.5 Benchmarking Against Few-Shot Adaptation (APE)

To provide a robust benchmark against the state-of-the-art in model adaptation, we evaluated the Adaptive Prior Refinement (APE) method, a prominent technique for efficient few-shot learning. This experiment serves a crucial contextualizing purpose: by comparing our simpler supervised methods against a more complex, specialized adaptation technique, we can better situate our findings within the broader literature and validate the effectiveness of our chosen approach. The test was conducted on the NGAD dataset, simulating data-scarce scenarios by randomly sampling a small number of training examples ("shots") per class. The results are summarized in Table 6.10.

Shots	Accuracy	Recall@1	Recall@2	Recall	Recall@4	Recall@5
1	37.21	37.21	53.20	64.16	71.23	78.31
2	38.13	38.13	59.13	73.29	80.82	84.93
4	50.23	50.23	66.44	79.00	86.76	90.18
8	41.10	41.10	64.16	74.43	83.79	88.36
16	53.42	53.42	74.43	84.93	87.90	91.55

Table 6.10: APE model performance on the NGAD test dataset.

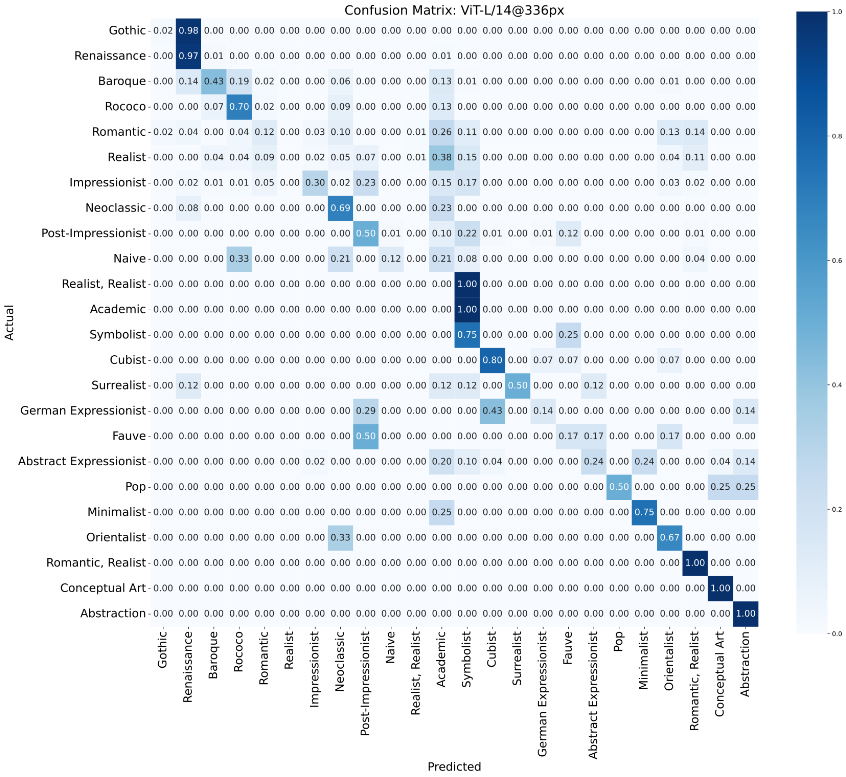
APE's performance exhibits a clear dependency on the number of training shots, with accuracy improving from 37.21% in the 1-shot setting to a peak of 53.42% with 16 shots. However, even in its optimal configuration, APE's performance remains significantly below that of our supervised methods trained on the full dataset. Both the cross-modal alignment (74.29% accuracy) and the linear probe (77.71% accuracy) on non-normalized embeddings are considerably more effective.

This outcome is highly instructive. The sub-optimal performance of APE serves to strengthen our central arguments. First, it demonstrates that for a fine-grained task like style recognition, which exhibits high intra-class variance, a straightforward supervised classifier that leverages the entire dataset is more powerful than a specialized few-shot method. Second, it suggests that the stylistic information within CLIP's embeddings is distributed across the entire feature space rather than being concentrated in a few highly discriminative channels that APE is designed to select. This reinforces the validity of our approach, which utilizes the full, unnormalized embedding space, and confirms that in a context where sufficient data is available, a simple linear model can outperform more complex adaptation techniques.

6.2.6 Qualitative Analysis of Zero-Shot Errors

To better characterize the nature of model errors, Figure 6.4 shows the normalized confusion matrix for the best-performing model. While the recall scores might appear modest at first glance, the matrix reveals a more nuanced picture: many errors occur between adjacent or stylistically related categories. For example, Impressionism is often confused with Post-Impressionism, Abstract Expressionism with Minimalism, and Baroque with Rococo.

Interestingly, the Academic style functions as a sort of fallback category, absorbing misclassifications from Baroque, Neoclassicism, Romanticism, and Realism. This pattern suggests that CLIP associates "academic" with a broad set of classical or representational visual features, particularly when more specific stylistic cues are ambiguous or absent.



	Gothic	Renaissance	Baroque	Rococo	Romantic	Realist	Impressionist	Neoclassic	Post-Impressionist	Naive	Realist, Realist	Academic	Symbolist	Cubist	Surrealist	German Expressionist	Fauve	Abstract Expressionist	Pop	Minimalist	Orientalist	Romantic, Realist	Conceptual Art	Abstraction
Gothic	0.02	0.98	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Renaissance	0.00	0.97	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Baroque	0.00	0.14	0.43	0.19	0.02	0.00	0.00	0.06	0.00	0.00	0.00	0.13	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.00
Rococo	0.00	0.00	0.07	0.70	0.02	0.00	0.00	0.09	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Romantic	0.02	0.04	0.00	0.04	0.12	0.00	0.03	0.10	0.00	0.00	0.01	0.26	0.11	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.14	0.00	0.00
Realist	0.00	0.00	0.04	0.04	0.09	0.00	0.02	0.05	0.07	0.00	0.01	0.38	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.11	0.00	0.00
Impressionist	0.00	0.02	0.01	0.01	0.05	0.00	0.30	0.02	0.23	0.00	0.00	0.15	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.02	0.00	0.00
Neoclassic	0.00	0.08	0.00	0.00	0.00	0.00	0.00	0.69	0.00	0.00	0.00	0.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Post-Impressionist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.01	0.00	0.10	0.22	0.01	0.00	0.01	0.12	0.00	0.00	0.00	0.00	0.01	0.00	0.00
Naive	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.21	0.00	0.12	0.00	0.21	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00
Realist, Realist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Academic	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Symbolist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Cubist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.80	0.00	0.07	0.07	0.00	0.00	0.00	0.07	0.00	0.00	0.00
Surrealist	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.12	0.12	0.00	0.50	0.00	0.00	0.12	0.00	0.00	0.00	0.00	0.00	0.00
German Expressionist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.43	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.14
Fauve	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.17	0.00	0.00	0.00
Abstract Expressionist	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.20	0.10	0.04	0.00	0.00	0.00	0.24	0.00	0.24	0.00	0.00	0.04	0.14
Pop	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.25	0.25
Minimalist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.75	0.00	0.00	0.00	0.00
Orientalist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00
Romantic, Realist	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Conceptual Art	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
Abstraction	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

Figure 6.4: Normalized confusion matrix of the best CLIP model (ViT-L/14@336px) on NGAD style classification.

The same phenomenon can be observed in Figure 6.5, which compares the distributions of true and predicted styles across the dataset. This analysis highlights both dataset imbalances and the most frequent error patterns in model predictions.

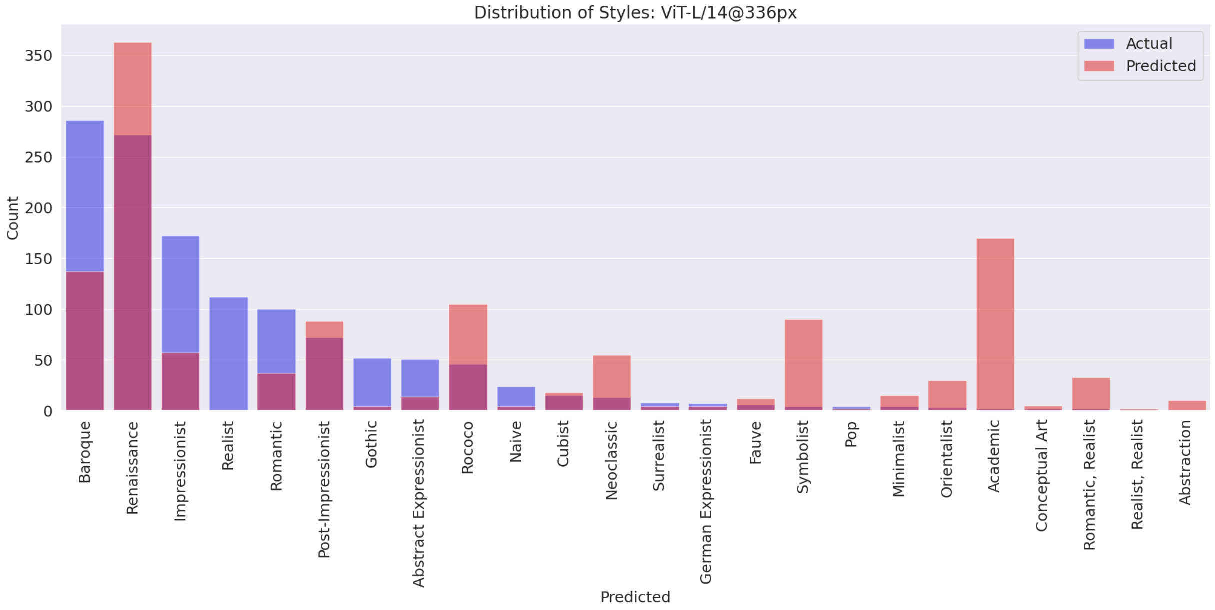
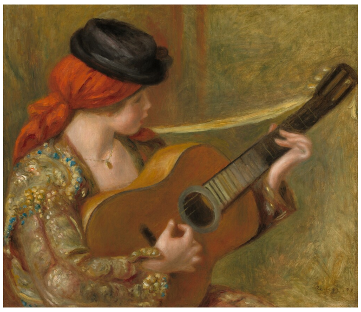


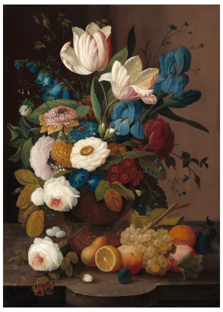
Figure 6.5: Comparison between actual and predicted distribution of styles in NGAD.

These results indicate that while CLIP exhibits some sensitivity to artistic style, its internal representations are not yet optimized for fine-grained distinctions requiring a nuanced understanding of artistic conventions or visual grammar. Further refinement, potentially through domain-specific supervision, may be required for such objectives.

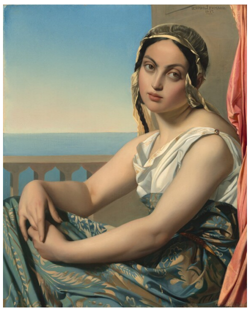
A qualitative inspection of misclassified examples further illustrates these limitations. Figure 6.6 presents three representative failure cases, each annotated with the ground-truth label and the incorrect prediction from the best-performing model.



True: Impressionist
Pred: Renaissance



True: Realist
Pred: Rococo



True: Orientalist
Pred: Neoclassic

Figure 6.6: Examples of misclassifications in style recognition.

In many of these cases, the predicted style may still appear visually plausible, underscoring the inherent subjectivity of the task and the challenge of disentangling stylistic cues from overlapping visual features.

When extending the analysis to AI-Pastiche, the evaluation compares the style predicted by CLIP for generated artworks with the style specified in the generation prompt. Notably, the classification accuracy achieved by CLIP in this context is comparable to that observed in NGAD, despite the visual inspection suggesting that the generated images often fail to convincingly reproduce historical styles.

This unexpectedly high agreement may be partly explained by a shared inductive bias: many generative models use CLIP during training or inference—as a scoring function, conditioning mechanism, or similarity guide—leading to predictions that align with the prompt without necessarily reflecting genuine stylistic fidelity.

As in the NGAD analysis, a confusion matrix (Figure 6.7) and style distribution comparison (Figure 6.8) were generated to further investigate model behavior on AI-Pastiche.

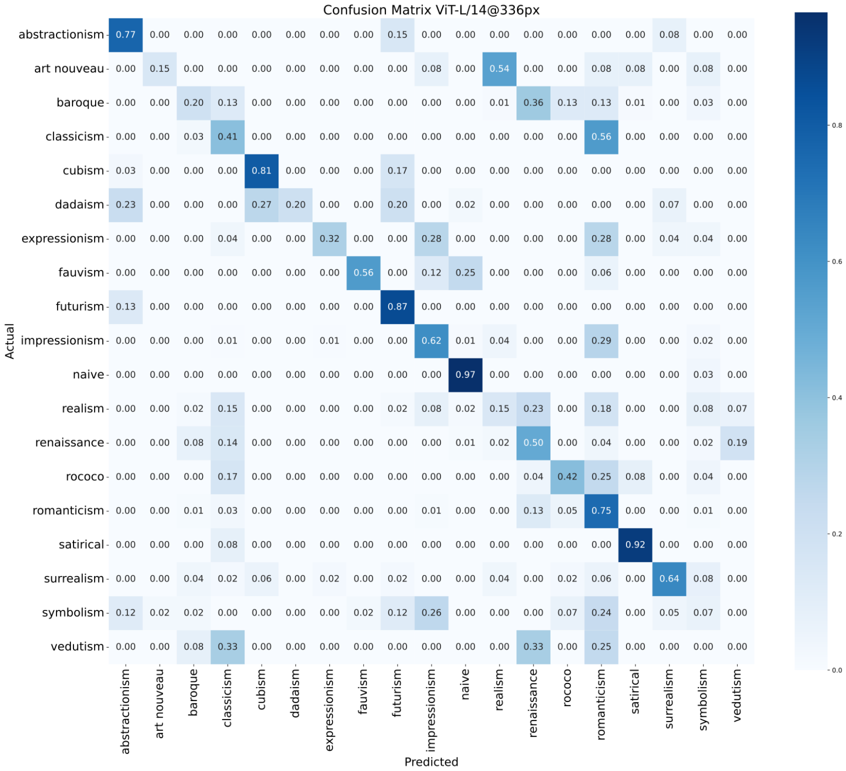


Figure 6.7: Normalized confusion matrix for style prediction in AI-Pastiche (best CLIP model).

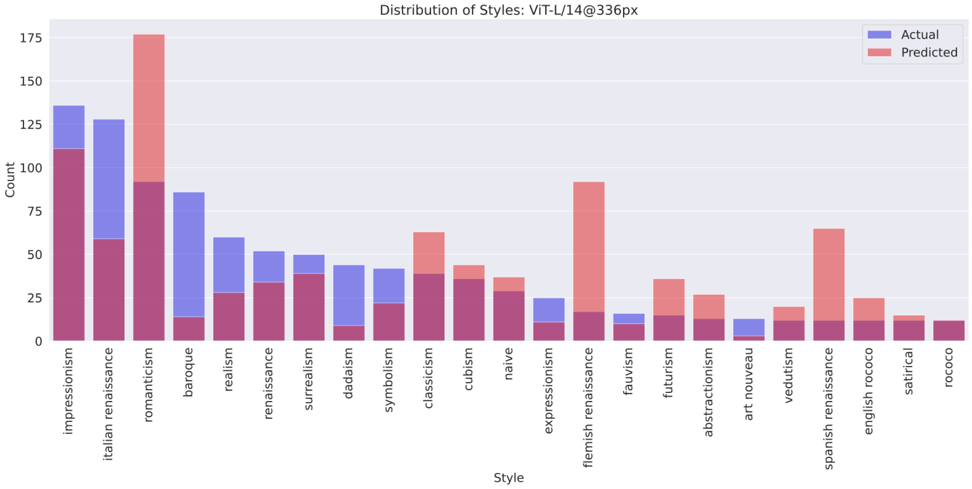


Figure 6.8: Comparison between actual and predicted style distributions in AI-Pastiche.

The distribution analysis reveals that CLIP achieves relatively strong performance for frequently prompted styles such as Renaissance, Impressionism, Surrealism, and Cubism, while Romanticism, Dadaism, and Classicism are more prone to misclassification.

6.3 Correlation with Human Judgment: Adherence and Artifacts

While the preceding experiments focused on CLIP's ability to perform objective classification and retrieval tasks, the following analyses shift to a more subjective and challenging dimension: assessing the degree to which the model's internal similarity metrics align with qualitative human perception. By leveraging the human evaluation data from the AI-Pastiche dataset, these final two experiments directly probe the "perceptual gap" between machine and human judgment. The core objective is to determine whether the features CLIP deems salient for evaluating prompt adherence and image quality are the same as those prioritized by human observers.

6.3.1 Adherence to Generative Prompts

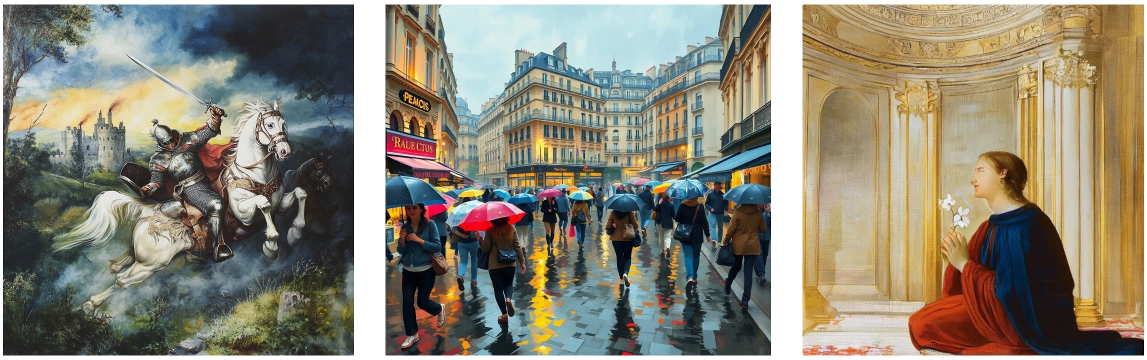
This experiment sought to answer a fundamental question: does a higher similarity score from CLIP correspond to a higher rating of prompt adherence from a human evaluator? To quantify the alignment between machine and human perception, we computed the cosine similarity between the vector of CLIP's normalized similarity scores and the vector of human adherence scores from the AI-Pastiche dataset. The results are presented in Table 6.11.

Model	Alignment with Human Judgment	Alignment with Defect Integration
RN50	0.406	0.478
RN101	0.398	0.462
RN50x4	0.379	0.448
RN50x16	0.413	0.489
RN50x64	0.428	0.484
ViT-B/32	0.411	0.481
ViT-B/16	0.383	0.458
ViT-L/14	0.425	0.482
ViT-L/14@336px	**0.437**	**0.497**

Table 6.11: Cosine similarity between the vector of CLIP's adherence scores and the vector of average human judgments. The second column shows the improved alignment when human-annotated defect scores are integrated into the model.

Across all models, the cosine similarity between the vector of CLIP scores and the vector of human scores hovers around a moderate value of approximately 0.4, with the best-performing model, "ViT-L/14@336px", reaching a similarity of 0.437. While this positive value indicates some degree of alignment, it must be contextualized. As a benchmark, the average inter-annotator correlation between different human evaluators on the same task is approximately 0.7. The substantial gap between CLIP's alignment score and the high level of human consensus provides quantitative evidence that the model's criteria for judging adherence diverge significantly from those of human observers.

A qualitative analysis of the discrepancies, illustrated in Figure 6.9, reveals the nature of this gap. CLIP's evaluation is heavily biased towards semantic content, often overlooking compositional or stylistic errors that are immediately apparent to humans. For example, Figure 6.9(a) was intended to represent a "fight between two knights on horseback", but was penalized by human evaluators for showing only one knight. CLIP, however, assigned it a high similarity score, as its feature space successfully identified the presence of "knight", "horse", and a "fight" scene, but was insensitive to the compositional error of the missing subject. Similarly, Figure 6.9(b) was criticized by humans for its lack of adherence to the requested Impressionist style, while Figure 6.9(c) was downgraded for missing one of the two figures described in the prompt. These examples highlight a critical limitation: CLIP's perception is atomistic, verifying the presence of semantic concepts without necessarily understanding their relational structure or completeness as specified in the prompt.



(a) Midjourney (b) Auto-Aesthetics (c) Omnigen

Figure 6.9: Examples of images in the AI-Pastiche dataset not aligning well with their prompts, for content or style.

6.3.2 Perception of Artifacts and Deformations

The final experiment investigated CLIP's sensitivity to a key factor in human aesthetic judgment: the presence of visual artifacts and technical flaws. The analysis proceeded in two stages. First, we attempted to directly predict the human-annotated defect scores using a linear regression model trained on the frozen CLIP image embeddings. This attempt failed conclusively, yielding a coefficient of determination (R^2) close to zero. This result provides strong evidence that information related to visual artifacts is not linearly encoded in CLIP's feature space; the model appears "blind" to these low-level structural flaws.

In the second stage, we tested the hypothesis that this "blindness" contributes to the perceptual gap observed in the adherence task. We integrated the human-annotated defect scores into a simple linear model alongside CLIP's similarity score to predict the final human adherence rating. The results, shown in the second column of Table 6.11, demonstrate a consistent and significant improvement in the cosine similarity for all models. For the best-performing model, "ViT-L/14@336px", the alignment score increased from 0.437 to 0.497.

This improvement is highly significant. It demonstrates that a substantial portion of the variance in human judgment that CLIP fails to capture is attributable to the perception of technical quality and artifacts. The fact that adding this external information brings the machine's evaluation closer to the human baseline is indirect but compelling proof that CLIP's assessment of adherence is fundamentally incomplete. It confirms that human perception is a composite judgment, weighing semantic content, stylistic consistency, and technical execution, whereas CLIP's perception is overwhelmingly dominated by the first of these alone.

6.4 Discussion: Semantic Relationships in the Latent Space

To synthesize the preceding results and investigate the internal organization of the latent space learned by CLIP, we employed the UMAP algorithm [90] to generate a three-dimensional projection of the high-dimensional embeddings. These embeddings included both image representations and textual prompts describing artistic styles (e.g., "an artwork in [style] style"). The resulting visualization (Figure 6.10) reveals a clear separation between textual encodings (the small cluster on the left) and image encodings (the large cluster on the right). For the images, we also distinguish correctly classified samples, shown as green bullets, from misclassified ones, shown as red crosses.

The substantial entanglement of the two image classes suggests a dominance of non-stylistic features in the embeddings. However, the chaotic pattern could also be a consequence of the aggressive dimensionality reduction and may not accurately reflect the semantic structure present in the original high-dimensional space.

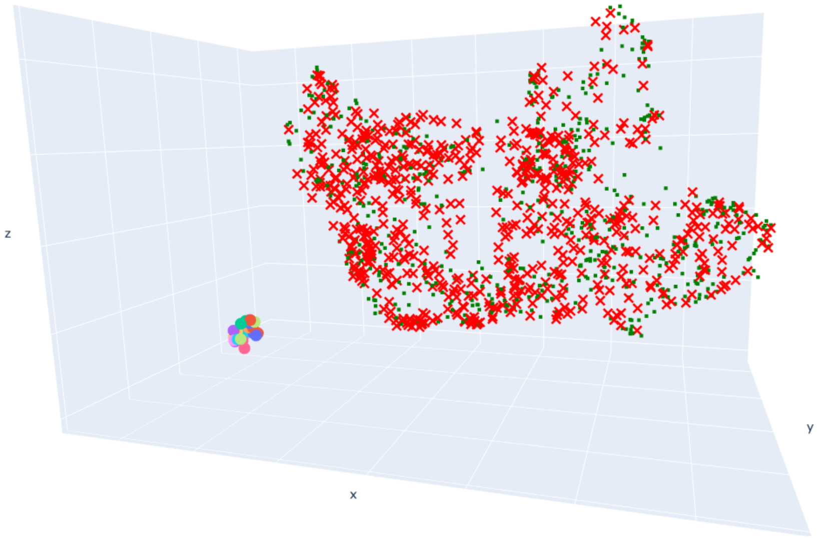


Figure 6.10: Three-dimensional projection of the textual embeddings of artistic styles and the visual embeddings of correctly classified (green) and misclassified (red) images using UMAP.

To gain a deeper understanding of the structure of CLIP's latent space, we conducted a detailed analysis based on nearest neighbors. Specifically, we focused on image pairs in which one image was correctly classified while the other was not. By identifying such pairs, we were able to isolate semantically coherent examples where classification performance diverged. This setup offered valuable insight into how subject matter similarity and stylistic cues interact within CLIP's representation space, and where the model may struggle to disentangle the two.

Representative examples are shown in Figure 6.11, where the misclassified image is displayed on the left, and its nearest correctly classified neighbor appears on the right. For each image, we report the cosine similarity scores with respect to both the true and the predicted style prompts, computed in both the original latent space and its lower-dimensional projection. This comparison allows us to examine how proximity in the embedding space relates to classification outcomes and whether stylistic distinctions are preserved through dimensionality reduction.

These findings underscore a key limitation of CLIP: it tends to encode and prioritize semantic content (like objects, scenes, and compositions) over stylistic features such as brushwork, color palette, or compositional structure. This leads to frequent misclassifications when artworks differ in style but share similar subject matter. This bias is consistent with CLIP's training objectives, which favor semantic alignment based on large-scale image-text data, where captions often emphasize content over formal style. As a result, CLIP's latent space tends to conflate stylistically diverse images with similar semantics. While we explored projections to mitigate this issue, the results were unsatisfactory. Techniques like light adapters could potentially fine-tune CLIP's vision encoder; such modifications fall outside the scope of our current research aims.

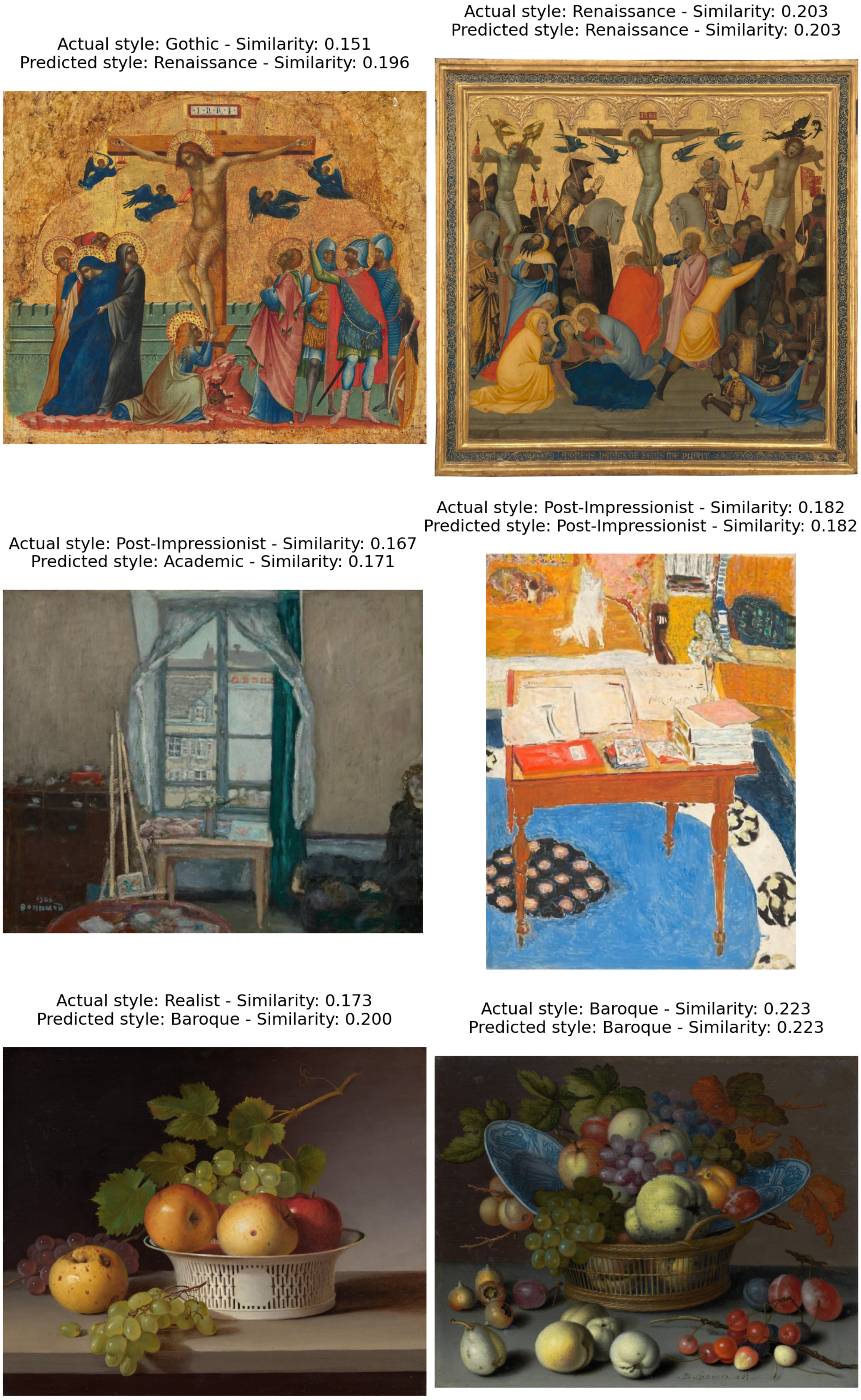

Actual style: Gothic - Similarity: 0.151
Predicted style: Renaissance - Similarity: 0.196
Actual style: Renaissance - Similarity: 0.203
Predicted style: Renaissance - Similarity: 0.203
Actual style: Post-Impressionist - Similarity: 0.167
Predicted style: Academic - Similarity: 0.171
Actual style: Post-Impressionist - Similarity: 0.182
Predicted style: Post-Impressionist - Similarity: 0.182
Actual style: Realist - Similarity: 0.173
Predicted style: Baroque - Similarity: 0.200
Actual style: Baroque - Similarity: 0.223
Predicted style: Baroque - Similarity: 0.223

Figure 6.11: Visual comparison between a misclassified image (left) and a correctly classified image (right). For each artwork, the actual and predicted artistic styles are shown, along with their respective similarity scores to the image in the latent CLIP space.

Chapter 7

Conclusion

This thesis has conducted a multifaceted investigation into the perceptual capabilities of the CLIP model, specifically examining its proficiency in the complex domain of fine art. By treating the model as a fixed, pre-trained perceptual system, this research has sought to uncover the intrinsic biases and representational priorities that emerge from its contrastive learning objective. The central inquiry has revolved around a critical question: to what extent does CLIP's computational perception align with human interpretation, particularly concerning the nuanced dimensions of artistic style, semantic content, and the technical quality of AI-generated imagery?

7.1 Summary of Core Findings

The empirical evidence presented in this thesis converges on a central conclusion: CLIP's perceptual framework is overwhelmingly dominated by semantic content, often at the expense of stylistic nuance and technical fidelity. This core finding is substantiated by a series of structured experiments that collectively delineate the model's strengths and limitations.

First, the research confirms that CLIP possesses a robust capability for high-level semantic understanding. In the image-text alignment tasks, the model demonstrated a strong ability to match artworks with their corresponding textual descriptions, underscoring the effectiveness of its shared embedding space in capturing broad thematic and compositional elements. However, a qualitative analysis of its errors revealed a recurring pattern: the model's success is predicated on coarse object and scene recognition, while it frequently overlooks finer-grained details, specific attributes, and the relational structure of elements within a composition.

Second, when confronted with the more abstract challenge of artistic style recognition, CLIP's native zero-shot performance proved to be modest. While the model exhibits a rudimentary sensitivity to stylistic affinity, often grouping related artistic movements, it lacks the discriminative power for precise classification. The investigation into supervised methods, including linear probing and cross-modal alignment, further illuminated this limitation. Although these methods substantially improved intra-dataset performance, confirming that stylistic information is linearly separable within the feature space, this success did not translate to robust cross-dataset generalization. This critical failure indicates that the model learns to recognize dataset-specific patterns and superficial visual cues rather than abstract, transferable concepts of artistic style.

Third, and perhaps most significantly, the direct comparison with human judgments on AI-generated art exposed a substantial perceptual gap. The moderate correlation between CLIP's similarity scores and human adherence ratings provides quantitative evidence that the model's criteria for evaluating prompt alignment diverge from human perception. This divergence is largely attributable to CLIP's insensitivity to visual artifacts and compositional errors—flaws that are immediately salient to human observers but are not explicitly encoded in the model's feature space. The finding that integrating human-annotated defect scores significantly improves the alignment with human judgment serves as compelling evidence that CLIP's assessment is fundamentally incomplete, prioritizing semantic correspondence while neglecting the technical execution that is integral to human aesthetic evaluation.

7.2 Broader Implications and Contributions

The findings of this thesis carry several broader implications for the field of multimodal AI and its application in culturally sensitive domains. The methodological decision to analyze CLIP as an unmodified perceptual system offers a crucial counterpoint to the prevailing focus on performance optimization through adaptation. By isolating the model's native capabilities, this research provides a clearer understanding of the foundational biases that may be inherited by downstream applications, including the large multimodal models that build upon CLIP-like encoders.

Furthermore, the explicit comparison between human-made and AI-generated art addresses a timely and critical issue. As generative models become increasingly sophisticated, the need for reliable evaluation metrics that align with human perception is paramount. This thesis demonstrates that while CLIP can serve as a useful tool for assessing semantic consistency, its limitations in perceiving stylistic integrity and technical quality make it an unreliable arbiter of overall artistic fidelity.

Ultimately, this work contributes to a growing body of critical inquiry into the nature of "understanding" in large-scale AI models. It highlights the distinction between statistical pattern recognition and genuine, nuanced comprehension, suggesting that the path toward more culturally and perceptually aligned AI will require not only greater scale but also more sophisticated training objectives that move beyond simple semantic matching.

7.3 Future Research Directions

The insights and limitations identified in this thesis open several promising avenues for future research. The following directions represent logical next steps to build upon the findings of this work:

- **Detecting AI-Generated Art:** One of the most direct and pressing future tasks is to investigate whether CLIP's feature space can be leveraged to explicitly distinguish between human-made and AI-generated artworks. While this thesis has shown that stylistic and artifact-related information is not linearly separable in a straightforward manner, it is possible that more complex, non-linear classifiers trained on CLIP's embeddings could learn to identify the subtle statistical "fingerprints" of synthetic media. Such an investigation would have significant practical implications for digital forensics and content authenticity.
- **Extending to Other Multimodal Models:** The focus of this thesis has been exclusively on the CLIP architecture. A valuable extension would be to apply the same methodological framework to other prominent Vision-Language Models, as well as to enhanced versions of CLIP that can accommodate longer and more descriptive textual inputs (i.e., more than 77 tokens). A comparative analysis across different architectures could reveal whether the semantic bias observed in CLIP is a general characteristic of current contrastive learning paradigms or a specific artifact of its design. This would provide a more comprehensive overview of the state-of-the-art and could inform the development of models with a more balanced perceptual understanding.

In conclusion, while CLIP represents a landmark achievement in multimodal learning, its application to the nuanced and subjective world of art reveals the profound challenges that lie ahead. The journey toward creating AI systems that can not only see but also perceive with the depth and sensitivity of human vision is still in its early stages. It is hoped that the critical analysis presented in this thesis will contribute to a deeper understanding of both the potential and the limitations of these powerful technologies, guiding future efforts to create models that are not only more intelligent but also more culturally and aesthetically aware.

Bibliography

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, PMLR, July 2021. ISSN: 2640-3498.

[2] A. Asperti, L. Dessì, M. C. Tonetti, and N. Wu, "Does CLIP perceive art the same way we do?," May 2025. arXiv:2505.05229 [cs].

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 34892–34916, Dec. 2023.

[4] X. Zhu, B. Zhu, Y. Tan, S. Wang, Y. Hao, and H. Zhang, "Selective Vision-Language Subspace Projection for Few-shot CLIP," in *Proceedings of the 32nd ACM International Conference on Multimedia*, (Melbourne VIC Australia), pp. 3848–3857, ACM, Oct. 2024.

[5] Y. Tian, D. Su, S. Lauria, and X. Liu, "Recent advances on loss functions in deep learning for computer vision," *Neurocomputing*, vol. 497, pp. 129–158, Aug. 2022.

[6] W. Tu, W. Deng, and T. Gedeon, "Toward a Holistic Evaluation of Robustness in CLIP Models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, pp. 8280–8296, Sept. 2025.

[7] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to Prompt for Vision-Language Models," *International Journal of Computer Vision*, vol. 130, pp. 2337–2348, Sept. 2022.

[8] R. Abbasi, A. Nazari, A. Sefid, M. Banayeeanzade, M. H. Rohban, and M. S. Baghshah, "CLIP Under the Microscope: A Fine-Grained Analysis of Multi-Object Representation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9308–9317, June 2025.

[9] P. Gao, S. Geng, R. Zhang, T. Ma, R. Fang, Y. Zhang, H. Li, and Y. Qiao, "CLIP-Adapter: Better Vision-Language Models with Feature Adapters," *International Journal of Computer Vision*, vol. 132, pp. 581–595, Feb. 2024.

[10] X. Liu, D. H. Park, S. Azadi, G. Zhang, A. Chopikyan, Y. Hu, H. Shi, A. Rohrbach, and T. Darrell, "More Control for Free! Image Synthesis with Semantic Diffusion Guidance," in *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 289–299, Jan. 2023. ISSN: 2642-9381.

[11] A. Taghipour, M. Ghahremani, M. Bennamoun, A. Miri Rekavandi, Z. Li, H. Laga, and F. Boussaid, "Faster Image2Video Generation: A Closer Look at CLIP Image Embedding's Impact on Spatio-Temporal Cross-Attentions," *IEEE Access*, vol. 13, pp. 141313–141327, 2025.

[12] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models," in *Proceedings of the 39th International Conference on Machine Learning*, pp. 16784–16804, PMLR, June 2022. ISSN: 2640-3498.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," Apr. 2022. arXiv:2204.06125 [cs].

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10674–10685, June 2022. ISSN: 2575-7075.

[15] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, and R. Rombach, "Scaling Rectified Flow Transformers for High-Resolution Image Synthesis," in *Forty-first International Conference on Machine Learning*, June 2024.

[16] Y. Bin, W. Shi, Y. Ding, Z. Hu, Z. Wang, Y. Yang, S.-K. Ng, and H. T. Shen, "GalleryGPT: Analyzing Paintings with Large Multimodal Models," in *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 7734–7743, Oct. 2024. arXiv:2408.00491 [cs].

[17] N. Garcia and G. Vogiatzis, "How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval," in *Computer Vision – ECCV 2018 Workshops* (L. Leal-Taixé and S. Roth, eds.), (Cham), pp. 676–691, Springer International Publishing, 2019.

[18] A. Asperti, F. George, T. Marras, R. C. Stricescu, and F. Zanotti, "A Critical Assessment of Modern Generative Models' Ability to Replicate Artistic Styles," Feb. 2025. arXiv:2502.15856 [cs].

[19] F. Peng, X. Yang, L. Xiao, Y. Wang, and C. Xu, "SgVA-CLIP: Semantic-Guided Visual Adapting of Vision-Language Models for Few-Shot Image Classification," *IEEE Transactions on Multimedia*, vol. 26, pp. 3469–3480, 2024.

[20] "NationalGalleryOfArt/opendata," Sept. 2025. original-date: 2021-03-23T14:24:49Z.

[21] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, pp. 423–443, Feb. 2019.

[22] W. Guo, J. Wang, and S. Wang, "Deep Multimodal Representation Learning: A Survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[23] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Communications Magazine*, vol. 27, no. 11, pp. 65–71, 2002.

[24] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[25] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: Data, models and evaluation metrics," *Journal of Artificial Intelligence Research*, vol. 47, pp. 853–899, 2013.

[26] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *Proceedings of the 32nd International Conference on Machine Learning* (F. Bach and D. Blei, eds.), vol. 37 of *Proceedings of Machine Learning Research*, (Lille, France), pp. 2048–2057, PMLR, 07–09 Jul 2015.

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and Tell: A Neural Image Caption Generator," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3156–3164, 2015.

[28] F. Feng, X. Wang, and R. Li, "Cross-modal retrieval with correspondence autoencoder," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 7–16, 2014.

[29] M. Hendriksen, M. Bleeker, S. Vakulenko, N. van Noord, E. Kuiper, and M. de Rijke, "Extending clip for category-to-image retrieval in e-commerce," 2022.

[30] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2425–2433, 2015.

[31] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, pp. 1060–1069, Pmlr, 2016.

[32] Y. Shi, B. Paige, P. Torr, *et al.*, “Variational mixture-of-experts autoencoders for multimodal deep generative models,” *Advances in neural information processing systems*, vol. 32, 2019.

[33] M. Hendriksen, S. Vakulenko, E. Kuiper, and M. de Rijke, “Scene-Centric vs. Object-Centric Image-Text Cross-Modal Retrieval: A Reproducibility Study,” in *Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, (Berlin, Heidelberg), pp. 68–85, Springer-Verlag, Apr. 2023.

[34] X. Wang, J. Zhao, E. Marostica, W. Yuan, J. Jin, J. Zhang, R. Li, H. Tang, K. Wang, Y. Li, *et al.*, “A pathology foundation model for cancer diagnosis and prognosis prediction,” *Nature*, vol. 634, no. 8035, pp. 970–978, 2024.

[35] R. J. Chen, M. Y. Lu, D. F. Williamson, T. Y. Chen, J. Lipkova, Z. Noor, M. Shaban, M. Shady, M. Williams, B. Joo, *et al.*, “Pan-cancer integrative histology-genomic analysis via multimodal deep learning,” *Cancer cell*, vol. 40, no. 8, pp. 865–878, 2022.

[36] P. Xu, X. Zhu, and D. A. Clifton, “Multimodal Learning with Transformers: A Survey,” May 2023. arXiv:2206.06488 [cs].

[37] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, “What’s cookin’? interpreting cooking videos using text, speech and vision,” in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (R. Mihalcea, J. Chai, and A. Sarkar, eds.), (Denver, Colorado), pp. 143–152, Association for Computational Linguistics, May–June 2015.

[38] A. Karpathy, A. Joulin, and L. Fei-Fei, “Deep fragment embeddings for bidirectional image sentence mapping,” *Advances in neural information processing systems*, vol. 27, 2014.

[39] L. W. Barsalou, “Grounded cognition,” *Annu. Rev. Psychol.*, vol. 59, no. 1, pp. 617–645, 2008.

[40] M. M. Louwerse, “Symbol interdependency in symbolic and embodied cognition,” *Topics in Cognitive Science*, vol. 3, no. 2, pp. 273–302, 2011.

[41] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-based image retrieval at the end of the early years,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, Dec. 2000.

[42] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, OpenReview.net, 2021.

[44] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General Perception with Iterative Attention," in *Proceedings of the 38th International Conference on Machine Learning*, pp. 4651–4664, PMLR, July 2021. ISSN: 2640-3498.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, Ieee, 2009.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, (Red Hook, NY, USA), pp. 6000–6010, Curran Associates Inc., Dec. 2017.

[47] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal Neural Language Models," in *Proceedings of the 31st International Conference on Machine Learning*, pp. 595–603, PMLR, June 2014. ISSN: 1938-7228.

[48] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov, "Devise: A deep visual-semantic embedding model," *Advances in neural information processing systems*, vol. 26, 2013.

[49] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[50] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[51] J. Mao, W. Xu, Y. Yang, J. Wang, Z. Huang, and A. Yuille, "Deep captioning with multimodal recurrent neural networks (m-rnn)," *arXiv preprint arXiv:1412.6632*, 2014.

[52] L. Gao, Z. Guo, H. Zhang, X. Xu, and H. T. Shen, "Video captioning with attention-based lstm and semantic consistency," *IEEE Transactions on Multimedia*, vol. 19, no. 9, pp. 2045–2055, 2017.

[53] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional Prompt Learning for Vision-Language Models," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16795–16804, June 2022. ISSN: 2575-7075.

[54] T. Yu, Z. Lu, X. Jin, Z. Chen, and X. Wang, "Task Residual for Tuning Vision-Language Models," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10899–10909, June 2023. ISSN: 2575-7075.

[55] Z. Guo, R. Zhang, L. Qiu, X. Ma, X. Miao, X. He, and B. Cui, "CALIP: Zero-Shot Enhancement of CLIP with Parameter-free Attention," Dec. 2022. arXiv:2209.14169 [cs].

[56] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[57] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of machine learning research*, vol. 21, no. 140, pp. 1–67, 2020.

[58] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[59] A. Baldrati, M. Bertini, T. Uricchio, and A. d. Bimbo, "Composed Image Retrieval using Contrastive Learning and Task-oriented CLIP-based Features," Aug. 2023. arXiv:2308.11485 [cs].

[60] A. Quattoni, M. Collins, and T. Darrell, "Learning visual representations using images with captions," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8, IEEE, 2007.

[61] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," *Advances in neural information processing systems*, vol. 25, 2012.

[62] A. Li, A. Jabri, A. Joulin, and L. Van Der Maaten, "Learning visual n-grams from web data," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4183–4192, 2017.

[63] D. Ruta, A. Gilbert, P. Aggarwal, N. Marri, A. Kale, J. Briggs, C. Speed, H. Jin, B. Faieta, A. Filipkowski, Z. Lin, and J. Collomosse, "StyleBabel: Artistic Style Tagging and Captioning," Mar. 2022. arXiv:2203.05321 [cs].

[64] H. Kazemi, A. Chegini, J. Geiping, S. Feizi, and T. Goldstein, "What do we learn from inverting CLIP models?," Mar. 2024. arXiv:2403.02580 [cs].

[65] K. Roth, Z. Akata, D. Damen, I. Balažević, and O. J. Hénaff, "Context-Aware Multimodal Pretraining," Nov. 2024. arXiv:2411.15099 [cs].

[66] S. Li, J. Cao, P. Ye, Y. Ding, C. Tu, and T. Chen, "ClipSAM: CLIP and SAM collaboration for zero-shot anomaly segmentation," *Neurocomputing*, vol. 618, p. 129122, Feb. 2025.

[67] H. Yang, N. Wang, H. Li, L. Wang, and Z. Wang, "Application of CLIP for efficient zero-shot learning," *Multimedia Systems*, vol. 30, p. 219, July 2024.

[68] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," Nov. 2021. arXiv:2111.02114 [cs].

[69] Lviv Polytechnic National Universit, V. Lytvyn, R. Peleshchak, Lviv Polytechnic National Universit, I. Rishnyak, Lviv Polytechnic National Universit, B. Kopach, Lviv Polytechnic National Universit, Y. Gal, and Drohobych Ivan Franko State Pedagogical University, "Detection of Similarity Between Images Based on Contrastive Language-Image Pre-Training Neural Network," in *Proceedings of the 8th International Conference on Computational Linguistics and Intelligent Systems. Volume I: Machine Learning Workshop*, CoLInS, Apr. 2024.

[70] S. Shen, L. H. Li, H. Tan, M. Bansal, A. Rohrbach, K.-W. Chang, Z. Yao, and K. Keutzer, "How Much Can CLIP Benefit Vision-and-Language Tasks?," in *International Conference on Learning Representations*, Oct. 2021.

[71] X. Pan, T. Ye, D. Han, S. Song, and G. Huang, "Contrastive Language-Image Pre-Training with Knowledge Graphs," *Advances in Neural Information Processing Systems*, vol. 35, pp. 22895–22910, Dec. 2022.

[72] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo, D. Song, J. Steinhardt, and J. Gilmer, "The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8320–8329, Oct. 2021. ISSN: 2380-7504.

[73] V. De Rosa, F. Guillaro, G. Poggi, D. Cozzolino, and L. Verdoliva, "Exploring the Adversarial Robustness of CLIP for AI-generated Image Detection," in *2024 IEEE International*

Workshop on Information Forensics and Security (WIFS), pp. 1–6, Dec. 2024. ISSN: 2157-4774.

[74] A. Mahendran and A. Vedaldi, "Understanding deep image representations by inverting them," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5188–5196, June 2015. ISSN: 1063-6919.

[75] H. Yin, P. Molchanov, J. M. Alvarez, Z. Li, A. Mallya, D. Hoiem, N. K. Jha, and J. Kautz, "Dreaming to Distill: Data-Free Knowledge Transfer via DeepInversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8715–8724, 2020.

[76] A. Dosovitskiy and T. Brox, "Inverting Visual Representations with Convolutional Networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4829–4837, June 2016. ISSN: 1063-6919.

[77] R. Zhang, W. Zhang, R. Fang, P. Gao, K. Li, J. Dai, Y. Qiao, and H. Li, "Tip-Adapter: Training-Free Adaption of CLIP for Few-Shot Classification," in *Computer Vision – ECCV 2022* (S. Avidan, G. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, eds.), vol. 13695, pp. 493–510, Cham: Springer Nature Switzerland, 2022. Series Title: Lecture Notes in Computer Science.

[78] X. Zhu, R. Zhang, B. He, A. Zhou, D. Wang, B. Zhao, and P. Gao, "Not All Features Matter: Enhancing Few-shot CLIP with Adaptive Prior Refinement," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2605–2615, Oct. 2023. ISSN: 2380-7504.

[79] S. Pratt, I. Covert, R. Liu, and A. Farhadi, "What does a platypus look like? Generating customized prompts for zero-shot image classification," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, (Paris, France), pp. 15645–15655, IEEE, Oct. 2023.

[80] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models," in *Proceedings of the 40th International Conference on Machine Learning*, pp. 19730–19742, PMLR, July 2023. ISSN: 2640-3498.

[81] W. Dai, J. Li, D. Li, A. Tiong, J. Zhao, W. Wang, B. Li, P. N. Fung, and S. Hoi, "InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 49250–49267, Dec. 2023.

[82] S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang, H. Zhong, Y. Zhu, M. Yang, Z. Li, J. Wan, P. Wang, W. Ding, Z. Fu, Y. Xu, J. Ye, X. Zhang,

T. Xie, Z. Cheng, H. Zhang, Z. Yang, H. Xu, and J. Lin, "Qwen2.5-VL Technical Report," Feb. 2025. arXiv:2502.13923 [cs].

[83] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "ShareGPT4V: Improving Large Multi-modal Models with Better Captions," in *Computer Vision – ECCV 2024* (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.), (Cham), pp. 370–387, Springer Nature Switzerland, 2025.

[84] N. Garcia, C. Ye, Z. Liu, Q. Hu, M. Otani, C. Chu, Y. Nakashima, and T. Mitamura, "A Dataset and Baselines for Visual Question Answering on Art," in *Computer Vision – ECCV 2020 Workshops* (A. Bartoli and A. Fusiello, eds.), (Cham), pp. 92–108, Springer International Publishing, 2020.

[85] M. V. Conde and K. Turgutlu, "CLIP-Art: Contrastive Pre-training for Fine-Grained Art Classification," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, (Nashville, TN, USA), pp. 3951–3955, IEEE, June 2021.

[86] T. Bleidt, S. Eslami, and G. de Melo, "ArtQuest: Countering Hidden Language Biases in ArtVQA," in *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 7311–7320, Jan. 2024. ISSN: 2642-9381.

[87] P. Liao, X. Li, X. Liu, and K. Keutzer, "The ArtBench Dataset: Benchmarking Generative Models with Artworks," June 2022. arXiv:2206.11404 [cs].

[88] R. S. R. Silva, A. Lotfi, I. K. Ihianle, G. Shahtahmassebi, and J. J. Bird, "ArtBrain: An Explainable end-to-end Toolkit for Classification and Attribution of AI-Generated Art and Style," Dec. 2024. arXiv:2412.01512 [cs].

[89] T. Fu, J. Conde, G. Martínez, P. Reviriego, E. Merino-Gómez, and F. Moral, "Artificial intelligence and misinformation in art: Can vision language models judge the hand or the machine behind the canvas?," *arXiv preprint arXiv:2508.01408*, 2025.

[90] L. McInnes, J. Healy, N. Saul, and L. Großberger, "UMAP: Uniform Manifold Approximation and Projection," *Journal of Open Source Software*, vol. 3, p. 861, Sept. 2018.

Appendix A

A.1 Example of a summary of an artwork from the NGAD

To illustrate the methodology adopted in our study, we selected a painting from the National Gallery of Art, Van Gogh's self-portrait. Since zero-shot CLIP accepts textual descriptions of no more than 77 tokens, we employed GPT-4o mini to generate a summary of the description of the painting, restricted to a maximum of 300 characters. The summarization process aimed to preserve essential information related to the subject and the stylistic characteristics of the artwork. This is the actual prompt used to generate the summaries:

Listing A.1: Prompt used to generate summaries

```
Your goal is to summarize the following painting descriptions in 300 characters. You will be
provided the description of a painting, its subject, its style, and its period, and you will
output a JSON object containing the following information:
{
    summary: string // at most 300 characters summary of the painting based on the painting
        description.
}
The summary must retain information about the subject, style, and period.
```

In Figure A.1 we present the painting, the official description provided by the National Gallery, and the summarized version produced using GPT-4o mini.

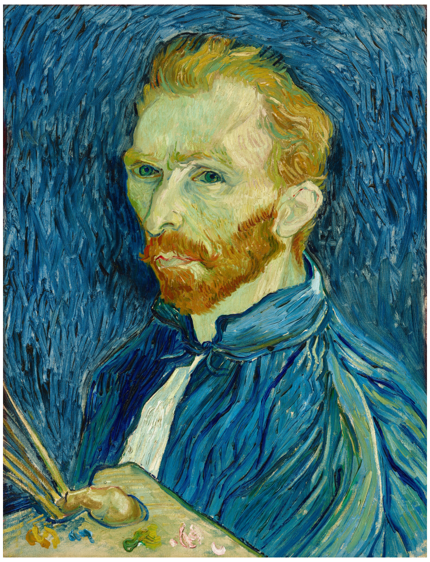


Figure A.1: Example of an artwork from the National Gallery.

National Gallery Description:

> *Shown from the chest up, a man with short, orange hair and green-tinted, pale skin looks at us, wearing a vivid blue painter's smock in this vertical portrait painting. His smock and the background are painted with long, mostly parallel strokes of cobalt, azure, and lapis blue. His shoulders are angled to our left, and he looks at us from the corners of his blue eyes. He has a long, slightly bumped nose, and his lips are closed within a full, rust-orange beard. He holds a palette and paintbrushes in his left hand, in the lower left corner of the canvas. The background is painted with long brushstrokes that follow the contours of his head and torso to create an aura-like effect.*

Summarized version of the description:

> *This Post-Impressionist portrait by Vincent van Gogh (1876-1900) depicts a man with orange hair and pale skin in a blue painter's smock. He gazes at the viewer with blue eyes, holding a palette and brushes. The background features long, parallel strokes of blue, enhancing the aura around him.*

A.2 Example of a summary of an AI-generated artwork from the AI-Pastiche Dataset

As an example of the AI-Pastiche dataset, we chose an Impressionist painting generated by generative models, depicting a bridge landscape. As observed in the artwork, the generative model adhered to the given prompt, producing an image strongly reminiscent of Monet's Japanese bridge. For the CLIP processing, the same procedure was applied to the description, limiting it to 300 characters of text, as done for the National Gallery descriptions.

In Figure A.2, we present the generated painting, the prompt provided to the model, and the corresponding summary.

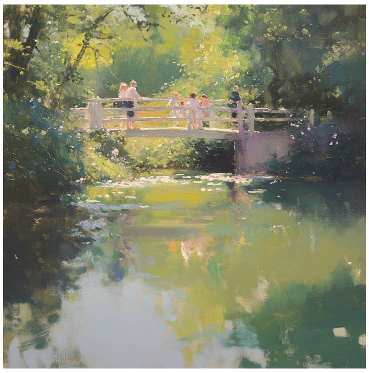


Figure A.2: Example of an artwork from the AI-Pastiche.

Prompt for the generated image:

> *Create an Impressionist-style painting depicting a serene outdoor scene, such as a sunlit garden, a riverside, or a city park. The image should focus on capturing the play of natural light and atmosphere, with soft, loose brushstrokes and a pastel-like color palette of light blues, greens, pinks, and yellows. The figures and landscape should appear slightly blurred, as if seen from a distance, giving a sense of movement and fleeting moments. Include reflections in water, dappled sunlight, and subtle shifts in color to evoke a peaceful, idyllic mood, typical of Impressionist art.*

Summarized version of the prompt:

> *This Impressionist painting from the XIX century features a serene landscape with soft tones, depicting a sunlit garden or riverside. It captures natural light and atmosphere through loose brushstrokes and a pastel palette. Figures and scenery appear slightly blurred, evoking movement and tranquility, with reflections in water.*

.