Together Yet Apart: Multimodal Representation Learning for Personalised Visual Art Recommendation

Bereket A. Yilma
Luis A. Leiva
name.surname@uni.lu
University of Luxembourg
Luxembourg

ABSTRACT

With the advent of digital media, the availability of art content has greatly expanded, making it increasingly challenging for individuals to discover and curate works that align with their personal preferences and taste. The task of providing accurate and personalized Visual Art (VA) recommendations is thus a complex one, requiring a deep understanding of the intricate interplay of multiple modalities such as image, textual descriptions, or other metadata. In this paper, we study the nuances of modalities involved in the VA domain (image and text) and how they can be effectively harnessed to provide a truly personalized art experience to users. Particularly, we develop four fusion-based multimodal VA recommendation pipelines and conduct a large-scale user-centric evaluation. Our results indicate that early fusion (i.e, joint multimodal learning of visual and textual features) is preferred over a late fusion of ranked paintings from unimodal models (state-of-the-art baselines) but only if the latent representation space of the multimodal painting embeddings is entangled. Our findings open a new perspective for a better representation learning in the VA RecSys domain.

CCS CONCEPTS

• **Information systems** → **Personalization**; **Recommender systems**; • **Computing methodologies** → *Learning latent representations*; • **Applied computing** → *Media arts.*

KEYWORDS

Recommendation; Personalization; Artwork; User Experience; Machine Learning

ACM Reference Format:
Bereket A. Yilma and Luis A. Leiva. 2023. Together Yet Apart: Multimodal Representation Learning for Personalised Visual Art Recommendation. In *UMAP '23: Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization (UMAP '23), June 26–29, 2023, Limassol, Cyprus.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3565472.3592964

1 INTRODUCTION

Art is vast and diverse, with a wide range of styles, mediums, and forms. As a result, it can be challenging for individuals to discover new art that aligns with their personal preferences and interests. Art appreciation is also highly subjective, and what one person finds beautiful or inspiring may not be the same for another. Personalised recommender systems (RecSys) can help to address this challenge by suggesting artworks that are tailored to an individual's taste and preferences. However, creating effective personalised recommendations for Visual Art (VA) poses several challenges. In the VA domain, paintings are important items that bring together complex elements such as drawings, gestures, narration, composition, or abstraction [30]. The subjective nature of user's taste and the unique nature of their preferences, which are longstanding challenges in content personalization, are also salient issues in VA RecSys.

Furthermore, the kind of emotional and cognitive reflections paintings may trigger in users are also diverse, depending on their background, knowledge, and several other environmental factors [34]. Traditional RecSys often rely on collaborative filtering (CF), where the preferences of a group of users are used to make suggestions for a given user. However, such approaches may not be effective for VA recommendation, as the aesthetics of art can vary widely and may not be easily captured by a group of users. Additionally, the prevalence of the cold start problem in real applications of VA RecSys (i.e. museums and art galleries) where personalised recommendations are to be offered for new visitors, makes CF approaches impractical [48]. Hence, to enhance personalized VA recommendations, efficiently capturing latent semantic relationships of paintings is vital and yet remains an open research challenge.

The majority of previous work in VA RecSys often infers similarities and relationships among paintings from high-level features derived from the above-mentioned traditional metadata such as artist names, styles, materials, etc. However, these features may not be expressive enough to capture abstract concepts that are "hidden" in paintings and that could better adapt the recommendations to the subjective taste of the users. For this, a high-quality representation of the data is crucial [5]. Unfortunately, research on machine-generated data representation techniques for VA RecSys has been often overlooked, as prominent works have largely relied on manually curated metadata [27].

Over the past few years, representation learning techniques have gained attention in VA RecSys. For example, He et al. [15] were among the first ones to use latent visual features extracted using Deep Neural Networks (DNN) and also use pre-trained DNN models for VA RecSys. Messina et al. [33] showed that DNN-based visual features perform better than leveraging textual metadata, however they were focused on the artwork market, which is driven by transaction data rather than enhancing the users' quality of

UMAP '23, June 26–29, 2023, Limassol, Cyprus

ACM ISBN 978-1-4503-9932-6/23/06.
https://doi.org/10.1145/3565472.3592964

experience. Therefore, it is unclear if their findings would transfer to a more *user-centric* setting, which essentially entails investigating the actual relevance of recommendations to users.

Previous works argued that visual features tend to perform better than textual metadata [32, 46, 48] and hence they argued for not considering text-based information in VA RecSys. However, recent work [47] showed that both modalities significantly capture the complex semantics embedded in VA and their combination, using late fusion techniques, can even lead to a better performance. To this end, we set out to explore techniques that can jointly learn latent semantic representations of VA from different modalities for a personalized VA RecSys task. In particular, in this work we address the following key research questions:

RQ1: Can we jointly learn meaningful latent semantic representations from different modalities (textual descriptions and images) of VA to derive personalised recommendations?

RQ2: Does jointly learned features generate better recommendations than late-fused rankings?

To answer these questions, we study two multimodal VA representation learning techniques based on state-of-the-art approaches: Contrastive Language-Image Pre-training (CLIP) [37] and Bootstrapping Language-Image Pre-training (BLIP) [24]. Subsequently, we investigate whether jointly learned features (i.e, features that are learned from both modalities simultaneously) generate better recommendations than late-fused rankings (i.e., combining recommendations from each modality after they have been generated separately) of paintings.

For the latter, we study the combination of best performing models from previous work [47, 48]: Latent Dirichlet Allocation (LDA) [6] to learn text features and the popular Residual Neural Network (ResNet) [14] to learn visual features.

Finally, we conduct a large-scale user study that evaluated how accurate, diverse, novel, and serendipitous were the generated recommendations. In sum, this paper makes the following contributions:

- We develop and study four VA RecSys engines: two versions of late fusion combining LDA with ResNet and two early fusion engines (CLIP and BLIP).
- We conduct a large-scale user study ($N = 100$) to assess VA RecSys performance from a user-centric perspective.
- We contextualize our findings and provide guidance about how to design next-generation VA RecSys.

2 RELATED WORK

With the proliferation of online marketplaces, it has become easier than ever for artists to showcase and sell their work the web. However, with such a large volume of art available, it can be challenging for both artists and collectors to find and connect with each other. This is where VA RecSys come in. Nowadays, VA RecSys are becoming more and more prevalent in online platforms as well as in Cultural Heritage environments such as museums and art galleries [20]. The huge potential and benefit of personalized recommendations, in particular in the VA field, has been discussed by Esman [12], for which different approaches to VA RecSys has been proposed over the years. For example, Aroyo et al. [13] proposed a semantically-driven RecSys and semi-automatic generation of personalized museum visits guided by visitor models. Deladiennee et al. [10] and Kuflik et al. [22] introduced a graph-based semantic RecSys that relies on an ontological formalisation of knowledge about manipulated entities. However, there are several aspects that are challenging in VA RecSys.

Primarily, because paintings are both high-dimensional and semantically complex, we need a computationally efficient way of modelling both their content and their context. This essentially calls for efficient data representation techniques that are capable of capturing the complex semantics embedded in paintings. To this end, He et al. [15] proposed a visually, socially, and temporally-aware model for artistic recommendation. This was among the first works that utilized the power of DNNs to exploit latent representations for VA recommendation. Their work primarily builds upon two methods, factorized personalized Markov chains [39] and visual Bayesian personalized ranking [16]. Although, the method is only applicable under collaborative filtering scenario.

Subsequently, Messina et al. [31–33] explored content-based artwork recommendation using images, keywords, and transaction data from the UGallery online artwork store.[1] Their work suggested that automatically computed visual features perform better than manually-engineered visual features extracted from images (i.e, texture, sharpness, brightness, etc.). Their work also indicated that a hybrid approach combining visual features and textual keyword attributes such as artist, title, style, etc., yields a further performance improvement. However, their hybrid approach was based on computing a score as a convex linear combination of the scores of individual methods (visual similarity and keyword similarity). Particularly, they did not explore feature learning approaches which are more scalable and generalizable. Recent work by Yilma et al. [48] proposed a VA recommendation approach that leveraged topic modeling techniques from textual descriptions of paintings and performed a comparative study against visual features automatically extracted using DNNs. Their study demonstrated the potential of learning features from text-based data, especially when it comes to explaining the recommendations to the user. However, they did not study the combination of text-based and image-based RecSys engines. A follow-up work by Yilma and Leiva [47] explored VA RecSys through reciprocal rank fusion to combine recommendations generated from engines independently trained on image and text. The results from this study indicated that a combination of both modalities performs better.

In sum, a number of VA Recsys strategies have been proposed over the years, but given that (i) user preferences are highly subjective and (ii) visual artwork is particularly complex to grasp, VA recommendation remains a rather challenging task. It demands a more accurate representation of not only VA content but also user profiles such as modelling temporal and social dynamics in terms of users' tendency to interact with content more or less consistently, as well as their preferences towards individual artists, styles, colors, etc. However, these are rarely available or not directly accessible in practice, making the so called cold-start problem a prevalent issue in VA RecSys. Thus, research effort in uncovering latent semantics of visual art is still considered a worthwhile endeavour, especially

[1] https://www.ugallery.com/

with regards to evaluating the quality of the recommendations from a user-centric perspective.

3 LEARNING LATENT REPRESENTATIONS OF PAINTINGS

Figures 1 to 3 summarize the VA RecSys approaches we have studied in this work. In the following we provide the essential information to understand the backbone models in each case.

3.1 Feature learning from text-based representations of paintings

Latent Dirichlet Allocation (LDA) [6] has demonstrated superiority over several other models in capturing hidden semantic structures in document modeling. It has been applied in several text-based RecSys tasks such as scientific paper recommendation [1], personalized hashtag recommendation [49], and online course recommendation [2]. Recent work by Devlin et al. [11] developed Bidirectional Encoder Representations from Transformers (BERT) and set a new state-of-the-art performance on sentence-pair related tasks like semantic textual similarity and question answering. However, LDA based representations have shown superiority over BERT on a VA Recsys task [47]. Hence, in this work we adopt LDA to learn painting feature representations from their associated textual metadata, where each painting is represented by a document containing detailed annotations such as title, format, or a curated description; see Figure 4 for an example. A detailed discussion on LDA topic modeling can be found in [6] and [18].

Once the LDA model is trained over the entire text dataset, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is produced where each entry $\mathbf{A}_{ij}$ is the cosine similarity measure between document embeddings. This similarity matrix therefore captures the latent topic distribution over all documents, which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks, as explained in the next section.

3.2 Feature learning from image-based representations of paintings

Visual feature extraction is critical to have a discriminative representation of images [29], and it is widely used in several tasks such as object detection, classification, or segmentation [40]. Traditional approaches to feature extraction include Harris Corner Detection [8], or the more advanced version Shi-Tomasi Corner Detector [3]. Other approaches have been proposed, such as SURF [28] or BRIEF [7], but they have been superseded by recent advances in Deep Learning, in particular in Convolutional Neural Networks (CNN). Today, image feature extraction techniques are mostly based on pre-trained CNN architectures such as AlexNet [21], GoogLeNet [43], and VGG [42]. The winner of the 2015 ImageNet challenge, ResNet, proposed by He et al. [14] introduced the use of residual layers to train very deep CNNs, setting a world record of more than 100 layers. ResNet-50 is the 50-layer version of this architecture, trained on more than a million images from the ImageNet database.[2] Thus, it has learned rich feature representations for a wide range of images and has shown superiority over other pre-trained models as a feature extractor [4, 17, 23]. We use the ResNet-50 model pre-trained on ImageNet to extract latent visual features (image embeddings) from paintings. By passing each painting image through the network, a convolutional feature map (i.e., a feature vector representation) is obtained.

Once we extract all image features from the entire dataset, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ is produced where each entry $\mathbf{A}_{ij}$ is the cosine similarity measure between all image embeddings. This similarity matrix therefore captures the latent visual distribution over all images, which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks, as explained in the next section.

3.3 Joint feature learning from visual and textual representations of paintings

3.3.1 Contrastive Language-Image Pre-training (CLIP). CLIP is a technique for pre-training a language model on a large dataset of images and their associated text [37]. Unlike traditional models that use an image encoder (e.g., a CNN) and a classifier (e.g., a fully-connected network), CLIP jointly trains an image encoder and a text encoder to encourage a close embedding space between the ones that form a pair via contrastive learning. The model predicts the correct image given a text prompt, and vice versa. During pre-training, the network learns to generate a shared embedding space for both image and text inputs, where similar images and captions are close to each other in the embedding space. The idea is that by pre-training on this task, the model will learn to understand the relationship between language and visual concepts, which can then be fine-tuned for a variety of natural language understanding and image understanding tasks.

Particularly, given a batch of M <image,text> pairs, CLIP is trained to predict which of the $M \times M$ possible image-text pairings across a batch actually occurred. Thus, the model learns a multimodal embedding space by jointly training an image encoder, either a ResNet or a Vision Transformer (ViT), and a text encoder (a Transformer such as BERT) to maximize the cosine similarity of the image and text embeddings of the M real pairs in the batch, while minimizing the cosine similarity of the embeddings of the M^2 - M incorrect pairings. Hence, a symmetric cross-entropy loss is optimized over these similarity scores. In this work, we leveraged CLIP as a feature extractor to learn a joint embedding space for images of paintings and their corresponding textual descriptions in order to uncover the latent semantic relationships between paintings embedded within the two modalities. We use ResNet-50 as our image encoder and BERT as our text encoder in our CLIP architecture. Once features from both modalities are extracted, a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ can be produced where each entry $\mathbf{A}_{ij}$ is the cosine similarity measure between the joint painting embeddings which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks. Figure 2 illustrates our multimodal approach to learn latent semantic representations of paintings with CLIP.

3.3.2 Bootstrapping Language-Image Pre-training (BLIP). BLIP is also a method for pre-training a neural network that combines language and image modalities [24]. However, unlike CLIP where a model is pre-trained by learning to distinguish between the correct and wrong image-text pairs, BLIP pre-trains a model by learning

[2] http://www.image-net.org

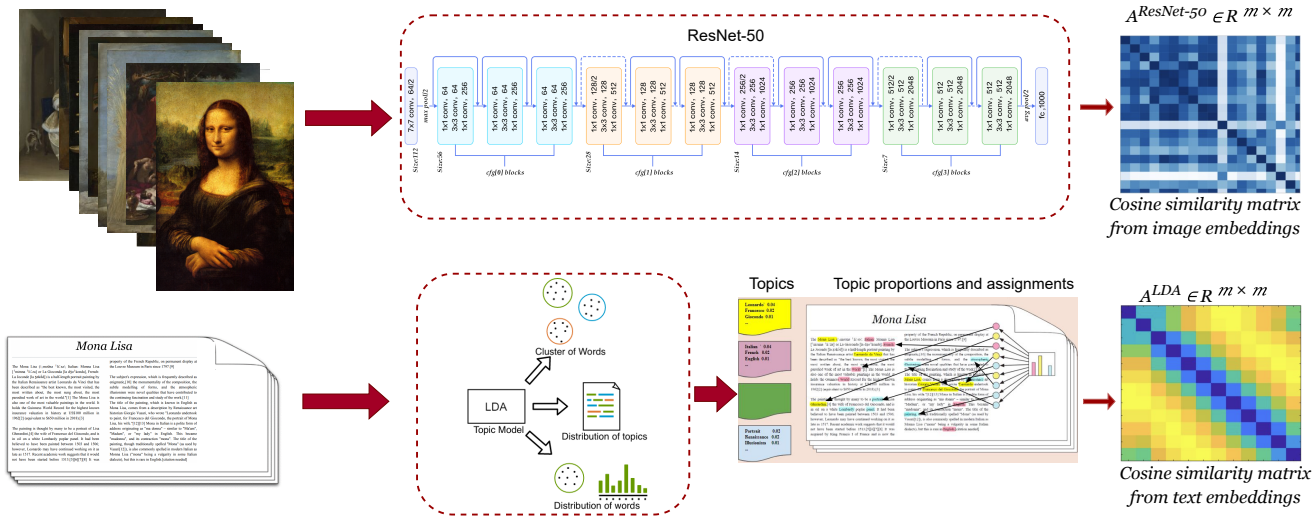


Figure 1: Overview of our unimodal approaches to learn latent semantic representations of paintings: Image-based (top) and text-based (bottom).

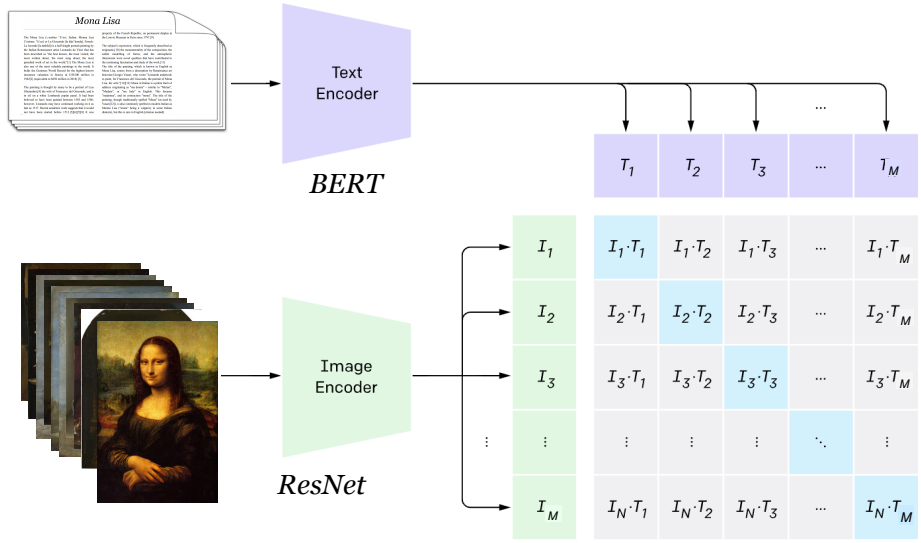


Figure 2: Overview of our multimodal approach with CLIP to learn latent semantic representations of paintings.

to predict an image or a text given the other modality, and makes sure the model is able to learn from a large dataset with the goal of bootstrapping the model's understanding of language and image relationships. In order to pre-train a unified vision-language model with both understanding and generation capabilities, BLIP introduces multimodal mixture of encoder-decoder, a multi-task model which can operate in one of the following three functionalities. (1) *Unimodal encoders*, which separately encode image and text. The image encoder is a ViT and the text encoder is BERT. A `[CLS]` token is appended to the beginning of the text input to summarize the sentence. (2) *Image-grounded text encoder*, which injects visual information by inserting a cross-attention layer between the self-attention layer and the feed forward network for each transformer block of the text encoder. A special `[Encode]` token is appended to the text, and its output embedding is used as the multimodal representation of the image-text pair. (3) *Image-grounded text decoder*, which replaces the bi-directional self-attention layers in the

text encoder with causal self-attention layers. A special [Decode] token is used to signal the beginning of a sequence.

During pre-training, BLIP jointly optimizes three objectives namely Image-Text Contrastive Loss (ITC), Image-Text Matching Loss (ITM), and Language Modeling Loss (LM). ITC activates the unimodal encoder with the aim of aligning the feature space of the visual and text transformers by encouraging positive image-text pairs to have similar representations in contrast to the negative pairs. ITM activates the image-grounded text encoder for a binary classification task, where the model is asked to predict whether an image-text pair is positive (matched) or negative (unmatched) given their multimodal feature. LM activates the image-grounded text decoder, which aims to generate textual descriptions conditioned on the images. By jointly optimizing these three objectives, BLIP allows to maximize the similarity between image and text representation. For the task of painting representation learning, we leveraged the pre-trained BLIP as a multimodal feature extractor.

Unlike CLIP, BLIP has an ITM head which is known to perform better at computing image-text similarity [24]. The ITM head uses cross-attention to fuse image and text features, which can capture finer-grained similarity compared to the simple cosine similarity function (used by the ITC loss and CLIP). Thus, we compute ITM scores for every painting by first extracting multimodal features and passing them to the ITM head which generates a probability matching scores for the (image, text) pairs. Once this is done a matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ can be produced where each entry $\mathbf{A}_{ij}$ is the the probability matching score between the joint painting embeddings which is then leveraged to compute semantic similarities of paintings for VA RecSys tasks. Figure 3 illustrates our multimodal approach to learn latent semantic representations of paintings with BLIP.

4 PERSONALIZED RECOMMENDATION OF PAINTINGS

We study approaches that can learn features from both textual and visual information of paintings, either independently or jointly. On the one hand, we use LDA for learning text-based representations, as it has shown superiority on VA RecSys [47]. On the other hand, since visual features have been extensively explored in VA RecSys [15, 19, 31, 48], we study ResNet-50 for learning image-based representations, which it is considered the state of the art in prior work [19, 48]. Finally, for joint representation learning, we study CLIP and BLIP, presented in Section 3.3, which have not been considered before in the domain of VA RecSys.

4.1 Problem formulation

Let $\mathcal{P} = \{p_1, p_2, \ldots, p_m\}$ be a set of paintings with their associated embeddings $\mathcal{P}^* = \{\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_m\}$ according to CLIP or BLIP, and $\mathcal{P}^u = \{p_1^u, p_2^u, \ldots, p_n^u\}$ be the set of paintings a user u has rated, where $\mathcal{P}^u \subset \mathcal{P}$ and $\omega^u = \{\omega_1^u, \omega_2^u, \ldots, \omega_n^u\}$ are the normalised ratings[3] that u gave to a small set of paintings $\mathcal{P}^u$. Once the dataset embeddings (latent feature vectors) are learned using the models (LDA, ResNet, CLIP or BLIP) we compute the similarity matrix for all the paintings $\mathbf{A}$ as discussed in Section 3. Then, the predicted score $S^u(p_i)$ the user would give to each painting in the collection $\mathcal{P}$ is calculated based on the weighted average distance between the rated paintings and all other paintings:

$$S^u(p_i) = \frac{1}{n} \sum_{j=1}^{n} \left(\omega_j^u \cdot \mathbf{A}_{ij} \right) \tag{1}$$

where $\mathbf{A}_{ij} = d(\boldsymbol{p}_i, \boldsymbol{p}_j)$ is the similarity between embeddings of paintings p_i and p_j in the computed similarity matrix. The summation in Equation 1 is taken over all user's rated paintings $n = |\mathcal{P}^u|$. Once the scoring procedure is complete, the paintings are sorted and the r most similar paintings constitute a ranked recommendation list. In sum, the VA RecSys task consists of recommending the most similar paintings to a user based on a small set of paintings rated before, i.e., the elicited preferences.

4.2 Early vs. Late fusion

Traditionally, late fusion (i.e. at post-hoc) has been preferred over early fusion (i.e. at the feature level) in previous work because the fused models can be independent from each other, so each can use their own features, numbers of dimensions, etc. [38]. However, recent advances in multimodal representation learning techniques have shown promising results in several downstream tasks [25]. Thus, we explore whether that jointly learned features can be more effective at capturing the complex relationships between the different modalities and producing better recommendations.

In order to answer our research questions, we study four VA RecSys engines: the first two engines are combinations of text and image representations of paintings by fusing independently trained LDA and ResNet rankings, which we refer to as "late fusion engines". The other two engines are based on jointly learned embeddings of text and image representations of paintings using CLIP and BLIP, which we refer to as "early fusion engines". For the late fusion engines, we adopted the reciprocal rank fusion strategy [9] for combining rankings in information retrieval systems, as it is simple to use and has been proved effective for VA RecSys tasks [47].

We consider two approaches to late fusion: "partial late fusion" (or Late-partial) and "total late fusion" (or Late-total). In partial late fusion we fuse the individual image and text-based rankings (of r paintings each) and keep the top-ranked r paintings. In total late fusion we rank first the whole collection of paintings for each modality (image and text representations), then fuse the individual rankings, and finally pick the top r paintings. This later approach tries to promote more painting agreements, thereby encouraging more consistent fused rankings.

5 DATASET

The dataset used in our study contains 2,368 paintings from The National Gallery, London.[4] This curated set of paintings belongs to the CrossCult Knowledge Base.[5] Each painting image is accompanied by a set of text-based metadata, which makes this dataset suitable for testing the proposed feature learning approaches. A sample data point is shown in Figure 4. For image-based representation learning, the actual images of paintings are used, whereas

[3] In our application, to be described later, users elicit their preferences in a 5-point scale rating (higher is better), thus we transform those values into weights $\omega_i^u \in [0, 1]$ for every painting p_i^u the user has rated.

[4] https://www.nationalgallery.org.uk/

[5] https://www.crosscult.lu/

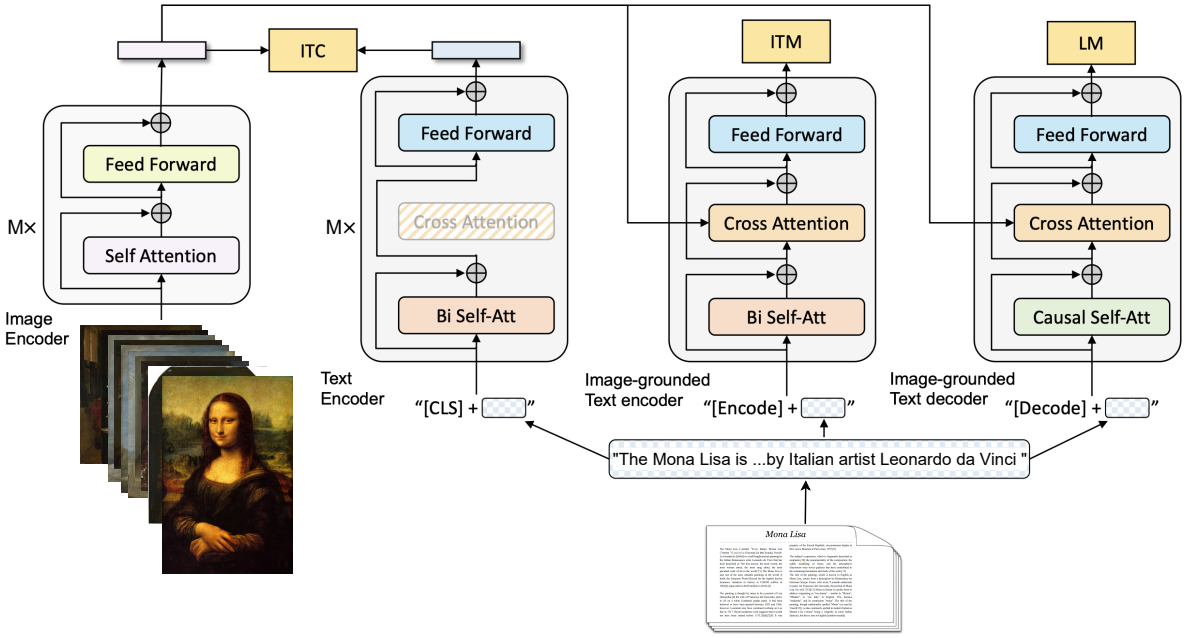


Figure 3: Overview of our multimodal approach with BLIP to learn latent semantic representations of paintings.

for text-based representation learning, we use all available painting attributes, such as artist name, painting title, technique used, etc. as well as descriptions provided by museum curators. These descriptions carry complementary information about the paintings, such as stories and narratives, that can be exploited to better capture the painting semantics.

5.1 Story groups

The dataset provides 8 curated stories (categories) linked to a few of the paintings, namely: *'Women's lives', 'Contemporary style and fashion', 'Water, Monsters and Demons', 'Migration: Journeys and exile', 'Death', 'Battles and Commanders', and 'Warfare'*. Figure 5 shows a 2D projection map of the story groups in the dataset using the non-linear projection t-SNE algorithm [44]. We can see that the majority of the paintings belong to the 'uncategorized' class. These story groups are meant to provide context to a selected group of paintings, according to the museum experts who created the dataset. We can observe from the latent space projections that the story groups are scattered across the entire dataset, suggesting that museum curators considered them to be representative examples of the collection. The map projection also surfaces the complex latent semantic relationships among the paintings.

5.2 Preprocessing

In order to learn visual features (i.e, extract convolutional feature maps) with the pre-trained ResNet-50 model and ViT, we used the actual images of paintings.[6] In order to learn textual features, the painting metadata were pre-processed: text fields concatenation, removal of punctuation symbols and stop-words, lowercasing, and lemmatization. For the purpose of unimodal semantic representation learning using LDA, a "topic coherence" analysis was conducted. Topic coherence is a commonly used technique to evaluate topic models and select the optimal number of topics that yield a meaningful representation [18]. Ideally, a good model should generate coherent topics; i.e the higher the coherence score the better the model is [35]. Following a coherence analysis the number of topics in our implementation were set to 10. For more details on LDA topic coherence we refer the reader to [18] and [35].

[6]All paintings are available under a Creative Commons (CC) license.

5.3 Modality gap

Multimodal models map inputs from different data modalities (e.g. image and text) into a shared latent representation space [24, 37]. A recent work by Liang et al. [26] discovered an intriguing geometric phenomenon of the representation space, where embeddings from different modalities are located in two completely separate regions of the embedding space. A systematic analysis demonstrated that this gap is caused by a combination of model initialization and contrastive learning optimization [26]. Their work also showed that varying the modality gap has a significant impact in the model's downstream performance. However, they did not evaluate potential impacts of the modality gap in other models such as BLIP, which is a multimodal architecture that optimizes not only contrastive loss (ITC) but also other objectives jointly (ITM and LM). In Figure 6 we can see that CLIP embeddings adhere to the reported phenomenon of the modality gap [26] but BLIP embeddings are rather entangled (i.e., a reduced modality gap).

Postulate 1. *A small modality gap has a positive impact on downstream performance [26], therefore we expect BLIP to achieve higher user-centred performance than CLIP in VA RecSys tasks.*

painting_id	000-0349-0000
title	Master of the Clarisse (possibly Rinaldo da Siena)
artist	Rinaldo da Siena
publication_date	13th_century
size_format	Portrait
size	Very Small
technique	-
description	This painting was probably influenced by a Byzantine icon. It shows the Virgin and Child, together with images which represent the Redemption of Man by Christ: the Annunciation and the Crucifixion. On either side is the Last Judgement, with trumpeting angels calling people from their tombs. It may be missing a gable with the Blessing Redeemer at the top and possibly wings on either side.

Figure 4: Sample painting and associated metadata from the National Gallery dataset.

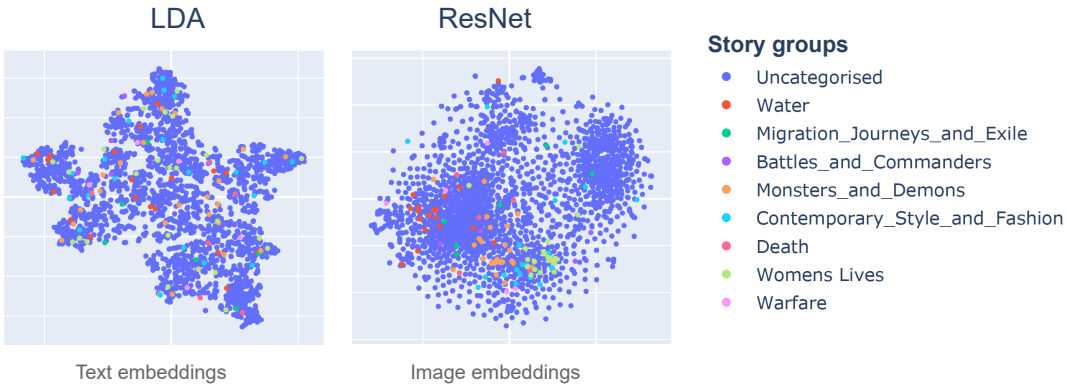


Figure 5: Unimodal Latent space projection (t-SNE projection) of the curated story groups.

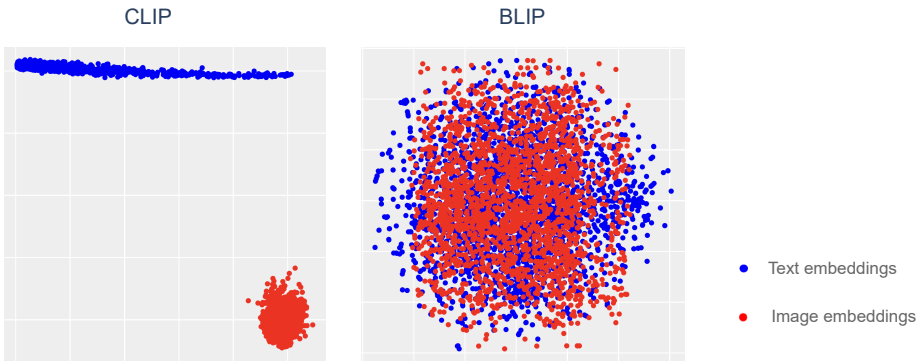


Figure 6: Modality gap in multimodal representation learning, in this case corresponding to CLIP and BLIP embeddings (t-SNE projection) of paintings.

6 EVALUATION

The main goal of our evaluation was to understand the user's perception towards the quality of our studied RecSys engines, and ultimately to assess which feature learning approach best captures the semantic relatedness of paintings. We conducted a large-scale user study, discussed below, that was approved by the Ethics Review Panel of the University of Luxembourg with ID 22-031.

Apparatus. We created a web application that first collected preference elicitation ratings from users and then showed a set of VA recommendations based on their elicited preferences. Then, participants were provided with one set of recommendations from each VA RecSys engine at a time (see Figure 7 left). Since participants could use any device (desktop computer, laptop, or mobile) to complete the study, we used a responsive design. By clicking or tapping on any image, in both the elicitation and rating screens, a modal window displays an enlarged version of the image (see Figure 7 right).

Participants. We recruited a large sample of $N = 100$ participants via the Prolific crowdsourcing platform.[7] We enforced the following screening criteria for any participant to be eligible: being fluent in the English language, art is listed among their interests/hobbies, minimum approval rate of 100% in previous crowdsourcing studies in the platform, and registration date before January 2022 (i.e., participants had been active for at least one year in the platform).

Our recruited participants (60 female, 40 male) were aged 28.55 years (SD=8.1) and could complete the study only once. Most of them had South African (20%), Portuguese (18%), or Italian (10%) nationality. The study took 6 min on average to complete (SD=3.9) and participants were paid an equivalent hourly wage of $10/h.

Design. Participants were exposed to all VA engines twice (within-subjects design), to serve as an attention check and also to ensure a consistent intra-rater agreement. Five users had a low intra-rater agreement[8] and thus were excluded from analysis. Our dependent variables are widely accepted proxies of recommendation quality [36]: **Accuracy** ("*The recommendations match my personal preferences and interests*"); **Diversity** ("*The recommended paintings are diverse*"); **Novelty** ("*The recommender helped me discover paintings I did not know before*"); **Serendipity** ("*I found surprisingly interesting paintings*").

Procedure. Participants accessed our web application and entered their demographics information (age, gender) on a welcome screen. There, they were informed about the purpose of the study and the data collection policy.

Then, participants advanced to the preference elicitation screen, where they were shown one painting at random from each of the nine curated story groups. They rated each painting in a 5-point numerical scale (5 is better, i.e. the user likes the painting the most). Finally, users advanced to the RecSys assessment screen, where they were shown a set of nine painting recommendations drawn from each VA RecSys engine. Note that each user initially rated nine paintings (one from each story group) but recommendations may come from only one or a few story groups, depending on their elicited preferences. Each set of paintings was rated in a 5-point Likert scale (1: Strongly disagree, ..., 5: Strongly agree) for each of the considered dependent variables.

[7] https://www.prolific.co/

[8] The difference between ratings for the same ranking was at least 2 points for at least one of the four dependent variables.

6.1 Results

We investigated whether there is any difference between any of the four VA RecSys engines, for which we use a linear mixed-effects (LME) model where each dependent variable is explained by each engine. Participants are considered random effects.

An LME model is appropriate here because the dependent variables are discrete and have a natural order. In addition, LME models are quite robust to violations of several distributional assumptions [41].

We fit the LME models (one per dependent variable) and compute the estimated marginal means for specified factors. We then run pairwise comparisons (also known as *contrasts* in LME parlance) with Bonferroni-Holm correction to guard against multiple comparisons.

Figure 8 shows the distributions of user ratings for each of the dependent variables considered. We report below statistical significance (p-values) and effect sizes (r) to better gauge the differences between our four VA RecSys engines.

Accuracy analysis. Differences between BLIP and CLIP were statistically significant ($p < .0001, r = 0.28$). Differences between BLIP and both late fusion engines were also statistically significant ($p = .0001, r = 0.24$). All other comparisons were not found to be statistically significant. Effect sizes suggest a moderate practical importance of the results.

Diversity analysis. Differences between BLIP and CLIP were statistically significant ($p < .0001, r = 0.26$). Differences between BLIP and the partial late fusion engine were statistically significant ($p = .043, r = 0.13$). Differences between CLIP and both late fusion engines were also statistically significant ($p = .043, r = 0.14$). No statistically significant differences were found between both late fusion engines. Effect sizes suggest a small practical importance.

Novelty analysis. No statistically significant differences between BLIP and CLIP were found. Differences between BLIP and both late fusion engines were also statistically significant ($p < .01, r = 0.22$). Differences between CLIP and the partial late fusion engine were statistically significant ($p = .015, r = 0.16$). No statistically significant differences were found between both late fusion engines. Effect sizes suggest a small to moderate practical importance.

Serendipity analysis. Differences between BLIP and CLIP were statistically significant ($p < .0001, r = 0.29$). Differences between BLIP and both late fusion engines were also statistically significant ($p < .001, r = 0.27$). All other comparisons were not found to be statistically significant. Effect sizes suggest a moderate practical importance.

Ranking overlap analysis. We conducted an additional analysis that checked whether our participants did receive truly personalized recommendations. For this, we computed the Intersection over Union (IoU) and Rank-Biased Overlap (RBO), which are widely used measures in information retrieval to assess ranking quality [45].

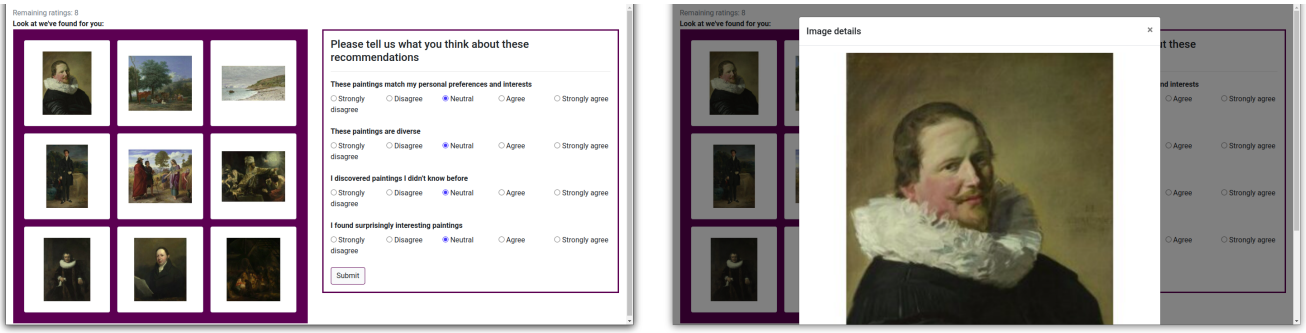
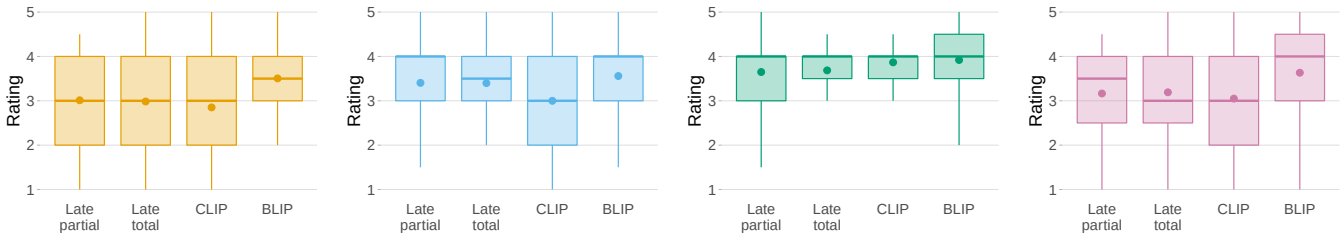


Figure 7: Screenshots of our web application for evaluation, in this case in a desktop browser.



(a) Accuracy ratings **(b) Diversity ratings** **(c) Novelty ratings** **(d) Serendipity ratings**

Figure 8: Distribution of user ratings. Dots denote mean values.

RBO and IoU were calculated in a pairwise manner among all users exposed to the same engine. Tables 1 and 2 present the results of this analysis. As shown in the tables, there is little overlap in the rankings produced by each engine. This analysis thus indicates that each participant indeed was shown a personalized set of recommendations and that each VA Recsys engine produced very different rankings.

7 DISCUSSION

To the best of our knowledge, our work is the first to exploit jointly learned multimodal approaches in VA RecSys, and subjected to a rigorous user evaluation. Particularly, leveraging joint optimization objectives such as BLIP set our work apart from previous studies, which have seemingly limited their scope to benchmarking with publicly available datasets, without delving deeper into user evaluations. Hence, we believe that our findings will inspire researchers to further explore multimodal approaches and open a new perspective for a better representation learning that serves not only the domain of VA RecSys but also several downstream tasks such as multimodal classification of paintings or prediction of user ratings towards artwork, for example.

Our results indicate that early fusion is preferred over late fusion but only if the latent space of the painting embeddings is entangled, which is the one provided by BLIP. Our results also show that both versions of late fusion engines were preferred over CLIP (which ensures the modality gap). This further complements our previous knowledge about reduced downstream performance of multimodal models in varying the modality gap [26]. This phenomenon was reported to be caused by model initialization and contrastive learning optimization. While this phenomenon was observed across multiple scenarios and benchmarks on publicly available datasets, there were no concrete recommendations regarding how to modify the gap and shift the embeddings to improve downstream performance.

Following on this discussion, multimodal architectures that depart from solely optimizing contrastive loss, such as BLIP, have not been studied in VA RecSys before. Hence, our work represents a significant contribution by extending BLIP to a real world application and conducting a large-scale user evaluation to assess its downstream performance. From our analysis, we believe that the multimodal representation learning technique employed in BLIP, which jointly optimizes three objective functions, allows to enhance model's understanding of text and image relationships, maximizing the similarity between these representations. Notably, it helps to overcome the modality gap and provides better recommendations, which validates our Postulate 1.

Based on our results, we can also conclude that instead of only contrastive learning, leveraging a late-total fusion approach (which ranks the whole database before using reciprocal rank fusion) is beneficial to generate better recommendations, as it promotes finding more common items. However, we can conclude that the best strategy for a personalized VA RecSys task is leveraging BLIP, which allows to jointly learn meaningful latent semantic representations from image and text modalities. Hence, we can provide affirmative answers to our initial research questions.

Table 1: Within-ranking overlap results, showing Mean ± SD of IoU and RBO measures.

	Late-partial	Late-total	CLIP	BLIP
IoU	0.09 ± 0.15	0.09 ± 0.15	0.09 ± 0.15	0.08 ± 0.17
RBO	0.10 ± 0.16	0.10 ± 0.16	0.10 ± 0.17	0.09 ± 0.16

Table 2: Between-ranking overlap results, showing Mean ± SD of IoU and RBO measures.

	Late-partial	Late-total	CLIP	BLIP
Late-partial		0.14 ± 0.14	0.01 ± 0.02	0.00 ± 0.01
Late-total	0.08 ± 0.11		0.00 ± 0.01	0.00 ± 0.00
CLIP	0.01 ± 0.01	0.00 ± 0.01		0.00 ± 0.01
BLIP	0.00 ± 0.01	0.00 ± 0.01	0.00 ± 0.01	

7.1 Limitations and future work

We have analysed a rich dataset with joint image and text modalities, which is a good representative of the VA domain. However, in practice it is difficult to get access to such datasets from other museums and galleries, therefore it would be interesting in the future to study the performance of our multimodal approaches on other art collections.

We also acknowledge that participants in a crowdsourcing study might not be intrinsically motivated, as they are rewarded with monetary incentives to take part in the study. However, to account for such a potential bias, among other strict screening criteria, we selected a large pool of participants, enforced that art would be listed among their interests/hobbies, and ensured that all their previous crowdsourcing studies were successfully approved. Furthermore, participants were considered as a random effect in our statistical analysis. Nonetheless, as a follow-up of this work, conducting a museum study with in-situ visitors may be beneficial to have a more accurate insight into the representation power of our models.

8 CONCLUSION

We have presented a novel approach to personalized VA recommendation grounded on multimodal representation learning. We explored different fusion strategies for image and text modalities by developing four different RecSys engines. Two engines were combinations of independently trained LDA and ResNet rankings, which we refer to as "late fusion engines", whereas the other two engines were based on jointly learned embeddings of images and textual descriptions of paintings using CLIP and BLIP, which we refer to as "early fusion engines". We then conducted a large-scale user study to evaluate the performance of each engine. Our results indicate that BLIP produces the best recommendation for users, followed by both late fusion approaches. Our data, code, and models are available as open-source software at https://github.com/Bekyilma/MRL_VA_RecSys.

ACKNOWLEDGMENTS

This work was supported by the Horizon 2020 FET program of the European Union through the ERA-NET Cofund funding grant CHIST-ERA-20-BCI-001 and the European Innovation Council Pathfinder program (SYMBIOTIK project, grant 101071147).

REFERENCES

[1] Maha Amami, Gabriella Pasi, Fabio Stella, and Rim Faiz. 2016. An lda-based approach to scientific paper recommendation. In *International conference on applications of natural language to information systems*. Springer, 200–210.
[2] Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe, and José Ochoa Luna. 2014. Online Courses Recommendation based on LDA.. In *SIMBig*. Citeseer, 42–48.
[3] Monika Bansal, Munish Kumar, Manish Kumar, and Krishan Kumar. 2021. An efficient technique for object recognition using Shi-Tomasi corner detection algorithm. *Soft Computing* 25, 6 (2021), 4423–4432.
[4] Catarina Barata, M Emre Celebi, and Jorge S Marques. 2018. A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE journal of biomedical and health informatics* 23, 3 (2018), 1096–1109.
[5] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35, 8 (2013), 1798–1828.
[6] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.
[7] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. Brief: Binary robust independent elementary features. In *European conference on computer vision*. Springer, 778–792.
[8] Jie Chen, Li-hui Zou, Juan Zhang, and Li-hua Dou. 2009. The Comparison and Application of Corner Detection Algorithms. *Journal of multimedia* 4, 6 (2009).
[9] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114
[10] Louis Deladiennee and Yannick Naudet. 2017. A graph-based semantic recommender system for a reflective and personalised museum visit: Extended abstract. In *2017 12th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*. 88–89. https://doi.org/10.1109/SMAP.2017.8022674
[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[12] Abigail R Esman. 2012. The World's Strongest Economy? The Global Art Market.
[13] Willem Robert van Hage, Natalia Stash, Yiwen Wang, and Lora Aroyo. 2010. Finding your way through the Rijksmuseum with an adaptive mobile museum guide. In *Extended semantic web conference*. Springer, 46–59.
[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
[15] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. 2016. Vista: a visually, socially, and temporally-aware model for artistic recommendation. In *Proceedings of the 10th ACM Conference on Recommender Systems*. 309–316.
[16] Ruining He and Julian McAuley. 2016. VBPR: visual bayesian personalized ranking from implicit feedback. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 30.
[17] A Victor Ikechukwu, S Murali, R Deepu, and RC Shivamurthy. 2021. ResNet-50 vs VGG-19 vs training from scratch: a comparative analysis of the segmentation and classification of Pneumonia from chest X-ray images. *Global Transitions Proceedings* 2, 2 (2021), 375–381.
[18] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. 2019. Latent Dirichlet Allocation (LDA) and Topic modeling:

models, applications, a survey. *Multimedia Tools and Applications* 78, 11 (2019), 15169–15211.

[19] Mehmet Oguz Kelek, Nurullah Calik, and Tulay Yildirim. 2019. Painter classification over the novel art painting data set via the latest deep neural networks. *Procedia Computer Science* 154 (2019), 369–376.

[20] Kalliopi Kontiza, Olga Loboda, Louis Deladiennee, Sylvain Castagnos, and Yannick Naudet. 2018. A museum app to trigger users' reflection. In *International Workshop on Mobile Access to Cultural Heritage (MobileCH2018)*.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2017. Imagenet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[22] Tsvi Kuflik, Einat Minkov, and Keren Kahanov. 2014. Graph-based Recommendation in the Museum. In *DMRS*. Citeseer, 46–48.

[23] Bin Li and Dimas Lima. 2021. Facial expression recognition via ResNet-50. *International Journal of Cognitive Computing in Engineering* 2 (2021), 57–64.

[24] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086* (2022).

[25] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.

[26] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. 2022. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *arXiv preprint arXiv:2203.02053* (2022).

[27] Ioanna Lykourentzou, Xavier Claude, Yannick Naudet, Eric Tobias, Angeliki Antoniou, George Lepouras, and Costas Vasilakis. 2013. Improving museum visitors' Quality of Experience through intelligent recommendations: A visiting style-based approach. In *Workshop Proceedings of the 9th International Conference on intelligent environments*. IOS Press, 507–518.

[28] Elmar Mair, Gregory D Hager, Darius Burschka, Michael Suppa, and Gerhard Hirzinger. 2010. Adaptive and generic corner detection based on the accelerated segment test. In *European conference on Computer vision*. Springer, 183–196.

[29] Karim Malik, Colin Robertson, Steven A Roberts, Tarmo K Remmel, and Jed A Long. 2022. Computer vision models for comparing spatial patterns: understanding spatial scale. *International Journal of Geographical Information Science* (2022), 1–35.

[30] R. Mayer and S. Sheehan. 1991. *The Artist's Handbook of Materials and Techniques*. Viking. https://books.google.lu/books?id=tQ9nYreyGwEC

[31] Pablo Messina, Manuel Cartagena, Patricio Cerda, Felipe del Rio, and Denis Parra. 2020. CuratorNet: Visually-aware Recommendation of Art Images. (2020).

[32] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2017. Exploring Content-based Artwork Recommendation with Metadata and Visual Features. *arXiv preprint arXiv:1706.05786* (2017).

[33] Pablo Messina, Vicente Dominguez, Denis Parra, Christoph Trattner, and Alvaro Soto. 2019. Content-based artwork recommendation: integrating painting metadata with neural and manually-engineered visual features. *User Modeling and User-Adapted Interaction* 29, 2 (2019), 251–290.

[34] Yannick Naudet, Angeliki Antoniou, Ioanna Lykourentzou, Eric Tobias, Jenny Rompa, and George Lepouras. 2015. Museum personalization based on gaming and cognitive styles: the BLUE experiment. *International Journal of Virtual Communities and Social Networking (IJVCSN)* 7, 2 (2015), 1–30.

[35] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*. Association for Computational Linguistics, 100–108.

[36] Pearl Pu, Li Chen, and Rong Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 157–164.

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*. PMLR, 8748–8763.

[38] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep Multimodal Learning: A Survey on Recent Advances and Trends. *IEEE Signal Process. Mag.* 34, 6 (2017), 96–108.

[39] Steffen Rendle, Christoph Freudenthaler, and Lars Schmidt-Thieme. 2010. Factorizing personalized markov chains for next-basket recommendation. In *Proceedings of the 19th international conference on World wide web*. 811–820.

[40] Ayodeji Olalekan Salau and Shruti Jain. 2019. Feature extraction: a survey of the types, techniques, applications. In *2019 International Conference on Signal Processing and Communication (ICSC)*. IEEE, 158–164.

[41] Holger Schielzeth, Niels J. Dingemanse, Shinichi Nakagawa, David F. Westneat, Hassen Allegue, Céline Teplitsky, Denis Réale, Ned A. Dochtermann, László Zsolt Garamszegi, and Yimen G. Araya-Ajoy. 2020. Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* 11, 9 (2020), 1141–1152.

[42] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[44] L.J.P. van der Maaten and G.E. Hinton. 2008. Visualizing Data Using t-SNE. *J. Mach. Learn. Res.* (2008), 2579–2605.

[45] William Webber, Alistair Moffat, and Justin Zobel. 2010. A Similarity Measure for Indefinite Rankings. 28, 4, Article 20 (2010), 38 pages. https://doi.org/10.1145/1852102.1852106

[46] Bereket Abera Yilma, Najib Aghenda, Marcelo Romero, Yannick Naudet, and Hervé Panetto. 2020. Personalised visual art recommendation by learning latent semantic representations. In *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization (SMA*. IEEE, 1–6.

[47] Bereket A. Yilma and Luis A. Leiva. 2023. The Elements of Visual Art Recommendation: Learning Latent Semantic Representations of Paintings. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 24, 17 pages. https://doi.org/10.1145/3544548.3581477

[48] Bereket Abera Yilma, Yannick Naudet, and Hervé Panetto. 2021. Personalisation in Cyber-Physical-Social Systems: A Multi-Stakeholder Aware Recommendation and Guidance. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (Utrecht, Netherlands) *(UMAP '21)*. Association for Computing Machinery, New York, NY, USA, 251–255. https://doi.org/10.1145/3450613.3456847

[49] Feng Zhao, Yajun Zhu, Hai Jin, and Laurence T Yang. 2016. A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. *Future Generation Computer Systems* 65 (2016), 196–206.