Event Recommendations through the Lens of Vision and Language Foundation Models

Haya Halimeh[1], Florian Freese[2] and Oliver Müller[1]

[1]Paderborn University, Paderborn, Germany

[2]University of Wuppertal, Wuppertal, Germany

Abstract

Recommender systems now span the entire customer journey. Amid the multitude of diversified experiences, immersing in cultural events has become a key aspect of tourism. Cultural events, however, suffer from fleeting lifecycles, evade exact replication, and invariably lie in the future. In addition, their low standardization makes harnessing historical data regarding event content or past patron evaluations intricate. The distinctive traits of events thereby compound the challenge of the cold-start dilemma in event recommenders. Content-based recommendations stand as a viable avenue to alleviate this issue, functioning even in scenarios where item-user information is scarce. Still, the effectiveness of content-based recommendations often hinges on the quality of the data representation they build upon. In this study, we explore an array of cutting-edge uni- and multimodal vision and language foundation models (VL-FMs) for this purpose. Next, we derive content-based recommendations through a straightforward clustering approach that groups akin events together, and evaluate the efficacy of the models through a series of online user experiments across three dimensions: similarity-based evaluation, comparison-based evaluation, and clustering assignment evaluation. Our experiments generated four major findings. First, we found that all VL-FMs consistently outperformed a naive baseline of recommending randomly drawn events. Second, unimodal text-based embeddings were surprisingly on par or in some cases even superior to multimodal embeddings. Third, multimodal embeddings yielded arguably more fine-grained and diverse clusters in comparison to their unimodal counterparts. Finally, we could confirm that cross event interest is indeed reliant on the perceived similarity of events, resonating with the notion of similarity in content-based recommendations. All in all, we believe that leveraging the potential of contemporary FMs for content-based event recommendations would help address the cold-start problem and propel this field of research forward in new and exciting ways.

Keywords

content-based event recommendation, clustering-based recommendation, cold-start problem, vision and language foundation models, domain adaptation

1. Introduction

Over the last years, the application of recommender systems in the tourism industry has expanded to cover more and more facets of the overall customer experience, ranging from recommending trips [1] over hotels [2] to restaurants [3]. Among these experiences, participating in local cultural events is a touristic activity that is gaining importance. By connecting travelers

RecSys Workshop on Recommenders in Tourism (RecTour 2023), September 19th, 2023, co-located with the 17th ACM Conference on Recommender Systems, Singapore

haya.halimeh@uni-paderborn.de (H. Halimeh); freese@uni-wuppertal.de (F. Freese); oliver.mueller@uni-paderborn.de (O. Müller)

with residents and local communities, these events enable cultural exchange and understanding.

Yet, the task of recommending events poses, arguably, greater challenges compared to recommending standardized mass services like transport or accommodation [4]. Events, by definition, have short life cycles, are never repeated in exactly the same way, and are always happening in the future [5]. These unique characteristics make it especially challenging to digitally represent events in their full breadth and depth. Unlike hotel rooms or flights, for example, we are lacking standard descriptors for representing events. In addition, due to their time-limited nature and low standardization, it is difficult to leverage historical data about the content of an event or past customer ratings as a basis for recommender systems. After all, an event recommendation is basically only reliable after the event took place, but must be created before it happens [6]. In addition, compared to global e-commerce or streaming platforms, travel and cultural platforms typically have only information about very few purchases or ratings of their customers, making collaborative filtering strategies difficult to implement, as these platforms are often limited to specific services, geographies, or genres and customers are switching between many platforms. In sum, the unique attributes of events amplify the well-known cold-start problem of recommender systems.

In scenarios characterized by limited or constrained data about item-user relationships (e.g., purchases, ratings), the system's ability to proficiently represent the content of these items is vital. Yet, as described above, the effective modeling of event contents hinges on encoding them into semantically meaningful and expressive representations. Recent advances in Deep Representation Learning (DRL), notably through foundation models (FMs) pretrained on massive amounts of broad data and adaptable to a wide range of specific tasks [7], offer a promising avenue to achieve this objective. In fact, since the groundbreaking introduction of BERT in 2018 [8], there has been a remarkable upsurge in the widespread adoption of large-scale FMs in various contexts. Owning to their capability to automatically extract rich representations from raw data, the consensus within the AI community now gravitates towards embracing FMs as the fundamental framework for training machine learning models on downstream tasks, moving away from the conventional practice of building models from scratch. Large-scale FMs were initially introduced for natural language processing (NLP) and later extended to include applications in computer vision (CV). More recently, the growing prominence of FMs within these two fields has led to increased research attention towards amalgamating both modalities. Multimodal FMs emerged as a natural outcome for this trend, specifically vision and language models (VL-FMs), which can handle both text and visual data simultaneously.

Considering that the intangible nature of events begs for multimodal content descriptions, it seems promising to evaluate the capacity of VL-FMs for integrating multimedia data into content-based event recommendations [9]. Previous research in this avenue has predominantly focused on designing modality-specific features based on event textual content [5, 10, 11, 12]. As a result, there remains an unaddressed gap in incorporating multimodal content and leveraging images as supplementary signals within the representation learning process.

In light of the above, we set out to explore the potential of harnessing both uni- and multimodal VL-FMs to learn informative representations of cultural events. The resultant event embeddings formed, in turn, the foundational cornerstone for content-based recommendations. We derived these content-based recommendations through a simple clustering approach that groups the events into semantically related cluster.

We conducted our computational experiments on an event dataset sourced from the event-based platform Meetup.com[1]. The dataset comprises $10,658$ distinct cultural events from the ten largest cities in the US, each accompanied by its respective descriptions and corresponding images. To evaluate the usefulness of different VL-FMs for generating content-based recommendations on this dataset, we conducted a series of online user experiments, in which we presented users with recommendations based on different VL-FMs and asked them to evaluate these recommendations with regards to three dimensions: similarity-based evaluation, comparison-based evaluation, and clustering assignment evaluation.

Our experiments generated four major findings. First, we found that all VL-FMs consistently outperform a naive baseline of recommending randomly drawn events. Second, surprisingly we found that unimodal text-based embeddings matched or in some cases even outperformed multimodal embeddings in terms of perceived similarity. Third, multimodal embeddings yielded arguably more fine-grained and diverse clusters in comparison to their unimodal counterparts. Finally, we could confirm that cross-event interest is reliant on the perceived similarity of events.

Our contributions can be summarized as follows: First, to the best of our knowledge, we present a novel exploratory method for grasping the impact of FMs on event recommendation systems. Second, to tackle the cold-start problem, we provide a straightforward clustering-based approach aimed at automatically identifying semantically relevant events based on their multimedia content. Third, we conduct a sequence of user experiments to confirm the results across three different dimensions. In doing so, we assess the efficacy of modern unimodal and multimodal FM techniques for cultural events representation.

The remainder of the paper is organized as follows: The next section offers an overview of related work on using embeddings for content-based event recommendations. The subsequent section imparts foundational knowledge regarding the technical background. Following that, section four delves into the approaches for generating uni- and multimodal embeddings and introduces representative FMs for each one. In section five, we outline the general approach, including the clustering framework, data and user experiments. Section six elaborates on and discusses the results, while section seven wraps up the study by addressing its limitations, implications, and providing a perspective on potential future research directions.

2. Events in the Recommender Landscape

For decades, there has been a growing interest in incorporating richer information beyond numerical ratings to promote recommendations. Over time, the field has evolved significantly with the development of various frameworks and methodologies [13, 14]. Acquiring a semantic representation of the recommended items is a pivotal aspect in the functionality of most of these frameworks, whether they are content-based or collaborative in nature.

Several works considered content-based event recommendations. Authors in [10] used event-related discussions to estimate the future popularity of events, while authors in [15] sought to enrich culture event metadata with open linked data. The latter enabled adding semantic knowledge structure into their recommendation methods. In [11], topic modeling techniques and Gibbs Sampling method were used to generate topic distributions based on the content of the

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

events and then map it to user features. In [16], the authors investigated event recommendation within the framework of the Douban network. They introduced a model that considers semantics and context by making use of content information analysis and social relations. Other authors introduced a hybrid approach that combines content-based and collaborative methods to provide recommendations for academic events to users [12].

Some studies tackled the cold-start problem in event-based social networks (EBSNs). Authors in [5], for example, suggest exploiting contextual signals, such as social, location based, and temporal signals, to enhance the recommendation quality. For the content based signals, they use basic TF-IDF analysis on the event descriptions. Authors in [6] developed a collective Bayesian Poisson factorization model that integrates location, organizer, user relationships, and event textual content to infer content topics and mitigate the cold-start local event recommendations. In [17], information about event venue, event popularity, temporal influence and geographical distance are used to create group event recommendations.

While these studies collectively exploit content based signals, they frequently rely on user features or contextual input and overlook scenarios where such information might be unavailable. Furthermore, most of these works typically start from scratch, without exploring the possibilities offered by large-scale state-of-the-art FMs for efficiently encoding event data. They also overlook including images as valuable signals in the recommendation process.

Motivated by these factors, we shift our focus towards modeling event data by leveraging accessible and advanced uni- and multimodal VL-FMs. By doing so, we (i) intend to experiment with integrating image information into event embeddings and (ii) probe into the potential of these models for enhancing content-based event recommendations against cold-start scenarios.

3. Deep Representation Learning

The key to enabling an intelligent system to comprehend the world around us lies in its ability to identify and separate the fundamental explanatory factors that are hidden within the low-level sensory data it observes. Representation Learning (RL) is a subarea of machine learning (ML) that aims to accomplish precisely that. At its core, RL learns a set of meaningful features from a given collection of data, making it simpler to derive valuable information when performing different ML tasks [18, 19, 20]. Deep RL seeks to achieve this by relying on sophisticated neural architectures such as FM models.

More recently, the growing popularity of unimodal FMs has sparked increased research interest in the integration of multiple modalities together. Indeed, at the most basic level, it is theoretically insufficient to imitate a whole range of human perceptions and understanding through only one modality [21]. For instance, describing a concept that is grounded in visual representation – such as shape constancy [22]– solely through non-visual means can be difficult.

The overarching goal for multimodal FMs thus becomes to capture the joint distribution of multiple modalities and to learn a shared representation space that reduces the semantic heterogeneity gap between the modalities while preserving the modality-specific semantics [23]. As mentioned in the introduction, a representative of this class of models are VL-FMs.

Pre-training VL-FMs typically involves three main steps: (i) encoding images and text into latent representations, (ii) designing a high-performing architecture for modeling interactions between modalities, and (iii) devising effective pre-training tasks [24].

The main architectural distinction lies in the interaction step, and there are several strategies to generating multimodal embeddings. However, the literature lacks agreement on the optimal design. For cultural event data, we find it intriguing to experiment with a single-stream fusion encoder approach, a dual encoder approach, and a two-step approach.

In the single-stream fusion encoder approach, the text embeddings and image features are concatenated together before feeding them into a transformer-based encoder to model the vision and language (VL) interaction [24, 25].

In the dual encoder approach, the text and image modalities are encoded separately through two single-modal encoders. As next, the embeddings are projected through a shallow interaction module to the same semantic space to compute VL similarity scores [24].

Lastly, in the two-step approach, one modality is translated into the other and the multimodal embeddings are generated using an unimodal FM based on the combined resulting set.

As is widely recognized, pre-trained FM models amass broad knowledge through extensive pre-training on numerous source tasks with abundant labeled and unlabeled data [7]. Research, however, suggests that domain adaptation could enhance the model's performance, even in cases where the source and target domains closely align [26]. Putting this into perspective, in our experiments, we prioritize adapting the pre-trained models to the target domain of event data, before proceeding to the recommendation task.

4. Vision and Language Foundation Models

We experiment with a separate unimodal FM for each individual modality– i.e. a language model and a vision model. For the multimodal VL-FMs, we evaluate the three approaches introduced in the previous section 3 using corresponding VL-FMs. Throughout this study, we denote embeddings generated from text only as text-based embeddings and those generated solely from vision as vision-based embeddings. An overall overview of the all approaches is provided in Figure 1. The forthcoming subsections delve deeper into the architectures, pre-training paradigms, and adaptation processes of the models for each approach.

4.1. Text-Based Embeddings

To create unimodal embeddings that rely exclusively on text but are adjusted to suit the event data domain, we employ the methodology described in [27]. We opt for this method, as it was recognized by the authors as the most effective approach in their investigation of domain adaptation for dense retrievals. We outline the steps visually in Figure 2.

The method combines two domain adaptation strategies, namely, Transformer-based Denoising AutoEncoder (TSDAE) [28] and Generative Pseudo Labeling (GPL) [27]. TSDAE is a denoising autoencoder-based architecture that focuses on reconstructing the original input sentences from a corrupted version of itself without accessing all contextualized word embeddings. GPL on the other hand combines a query generator and pseudo labeling. The approach involves creating queries for unlabeled sentences in the target domain, followed by pairing them with sentence passages. Subsequently, the resulting (query, passage) pairs are pseudo-labeled using a

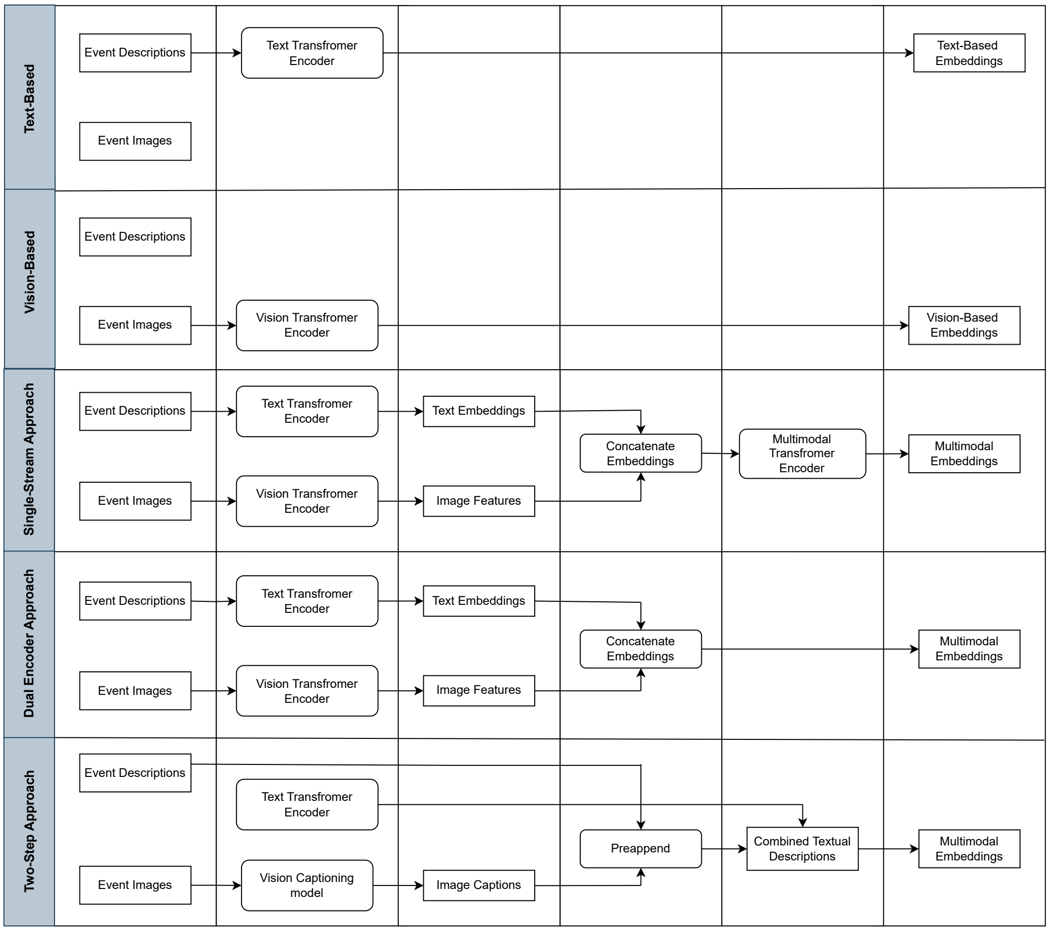


Figure 1: The approaches for the generation of vision-based, text-based, and multimodal vision & language embeddings.

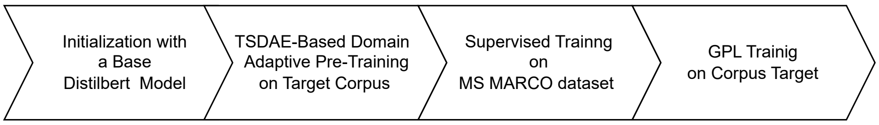

Initialization with a Base Distilbert Model
TSDAE-Based Domain Adaptive Pre-Training on Target Corpus
Supervised Trainng on MS MARCO dataset
GPL Trainig on Corpus Target

Figure 2: The pipeline for generating text-based embeddings, as proposed by [27].

cross-encoder. The model is then trained on these synthetically generated labels, utilizing the MarginMSE loss as introduced in [29].

Our implementation heavily relies on the publicly available GitHub repositories[2,3] of the two methods and adapts largely the same training arguments.

[2] Last visited 10.08.2023: https://github.com/UKPLab/gpl

[3] Last visited 10.08.2023: https://github.com/UKPLab/sentence-transformers/

4.2. Vision-Based Embeddings

Until recently, CV applications were dominated by convolutional neural networks (CNNs), while transformer architectures were completely absent from the field. In 2020, this architectural gap was bridged with the introduction of ViT -i.e. the first Vision Transformer Model by authors in [30]. ViT operates by partitioning an image into smaller patches and arranging them linearly as sequence elements, akin tokens in NLP. It then encodes the patches through linear projection and incorporates positional embeddings. The resultant set is further processed through a series of transformer blocks with a supervised training objective, commonly applied for image classification tasks.

In 2022, the Masked Autoencoder (MAE) for CV was proposed [31]. Following a standard self-supervised pre-training paradigm, MAE is designed to reconstruct partially corrupted patches in input images using a ViT initialized encoder and a lightweight decoder.

The unsupervised nature of autoencoders renders MAE as a viable tool for pre-training based domain adaptation on in-domain data. As such, we utilized the Huggingface library[4] and continued training the (facebook/vitmae-base) model for further 30 epochs and with the same default hyper-parameters as set in the original implementation. Finally, during inference, we extract akin to authors in [32] the image features from the penultimate layer pf the model.

4.3. Single-Stream Fusion Encoder Approach

As a representative for the single-stream architecture, we selected the Vision-and-Language Transformer (ViLT) [33]. ViLT is a self-supervised single-stream VL architecture that deviates from other VL-FMs through its minimal convolution-free pipeline. In particular, instead of relying on heavy convolutional networks (like Faster R-CNN [34]) ViLT employs ViT to encode pixel-level inputs. ViT is also utilized to initialize the interaction transformer in the model.

The image and text embeddings are concatenated into a single sequence, that undergoes multiple interaction block layer updates up until the final contextualized sequence. The pooled multimodal representation is then obtained by linear projection upon the first index of this sequence. The model is pre-trained on Image text matching (ITM) and masked language modeling (MLM) [8] objectives and targets cross-modal and multi-modal vision-and language tasks.

Since ViLT is pre-trained on smaller public paired datasets, we opted for domain adapting a base ViLT model on the event data following the same methodology as in 4.2. For the implementation, we leveraged the publicly available ViLT repository[5] and trained (vilt_200k_mlm_itm.ckpt) with the default settings for further 3 epochs.

4.4. Dual Encoder Approach

Contrastive Language-Image Pre-training (CLIP) [32] is a state-of-the-art deep learning model designed for vision-and-language understanding using a form of supervised contrastive learning.

CLIP undergoes pre-training on an immense private image-and-text dataset, comprising 400 million (image, text) pairs and sourced from publicly available web data. In line with other

[4] Last visited 10.08.2023: https://github.com/huggingface/transformers

[5] Last visited 10.08.2023: https://github.com/dandelin/vilt

VL-FMs in this class [35, 36], CLIP employs two standard transformer-based encoders to embed image-and-text pairs individually. By training the encoders to pull together associated text-image pairs while pulling apart mismatching ones, the model is compelled to learn a joint vector space that adeptly captures the intricate connections between the images and their associated texts. CLIP's representation learning capabilities were demonstrated through a standard linear probing protocol[6].

Taken the above into account, there appears to be no need for any further training. Instead, we apply the CLIP model[7] directly to the event texts and images to obtain their respective embeddings. As per the authors' approach, we extract the features before the linear projection layer, and fuse them into a unified vector representation through simple concatenation.

4.5. Two-Step Approach

Up until now, we have covered two VL-FMs, each with a different architecture. However, another viable approach is to convert one modality into the other. Once both modalities are represented in the same modality form, multimodal representations can be obtained by applying an unimodal FM on the combined information. By adopting such a strategy, one can benefit from the strengths of unimodal models in handling singular modalities, while still achieving multimodal VL understanding.

Converting an item from one modality into another would require creating an adequate representation of the same item in the other modality. For this purpose, VL generation techniques, such as visual captioning (VC) or text-to-image synthesis [37, 38], could be used.

Here, we employ VC to automatically generate textual captions for the event images. In particular, we implement the generative image-to-text transformer (GiT) proposed in [39]. Generative models typically contain complex architectures. In contrast, GiT adopts a straightforward encoder-decoder structure and the common language modeling (LM) loss while still maintaining state-of-the-art performance.

The obtained multimodal representations are achieved by applying the domain-adapted language model as discussed in subsection 4.1 on the combined set of event captions and event descriptions.

5. Approach

5.1. Clustering-Based Framework

As already discussed, content-based recommendations can mitigate the cold-start problem in event recommenders by delivering suggestions solely grounded in event content. The idea behind clustering-based recommendations on the other hand builds upon similarity within clusters to provide alike suggestions. Events within the same cluster are assumed to have shared characteristics, making it likely that users showing interest in a specific event will tend to also show interest in others from the same cluster. Alternately, cluster-based recommendations can serve to diversify suggestions by ensuring that events from various clusters are suggested,

[6] For detailed experiments and results, please refer to the supplementary material provided by the authors.

[7] Last visited 10.08.2023: https://github.com/huggingface/transformers

thereby expanding the spectrum of choices available to users and enhancing overall exposure to different events.

This study is primarily concerned with exploring whether contemporary FMs are suitable candidates for event content representations. As such, our emphasis lies in evaluating a select set of unimodal and multimodal VL-FMs for event representation within a simple clustering framework designed for content-based recommendations.

Embeddings generated by large-scale FMs are typically high-dimensional. However, as the number of dimensions in data increases, the proximity of a random data point to its nearest neighbor and its farthest neighbor approximate each other. In spaces with high dimensionality, the notion of spatial locality therefore loses its clarity [40, 41]. This complexity adds difficulty to the clustering task and renders applying an appropriate dimension reduction technique necessary [42, 43].

One of the available techniques is Uniform Manifold Approximation and Projection (UMAP) [44]. Over the last few years, UMAP has acquired recognition as a promising alternative to PCA [45] and t-SNE [46] for its ability to balance the preservation of the local and global structure of high-dimensional data when projected none-linearly onto lower dimensions. Recent research even shows that applying UMAP improves the performance of several clustering algorithms, both in terms of accuracy and computation time [47].

We proceeded by clustering the embeddings with Hierarchical Density-Based Spatial Clustering (HDBSCAN) [48]. HDBSCAN is a density-based clustering technique known for its ability to identify clusters of arbitrary densities, its resilience to outliers and noisy data, and its minimal reliance on prior knowledge or data assumptions.

In the last step, we draw on approaches proposed by [49, 50] and transfer the same setup to the multimodal case to pull representative exemplars for each cluster. In detail, we harness the Maximal Marginal Relevance (MMR) method [51] to select exemplars that are most relevant to each representative embedding while still being sufficiently dissimilar to each other.

It is well established that the efficacy of clustering algorithms is contingent upon the inherent structure of the data. By the same token, we argue that the quality of representations can be assessed by evaluating the quality of the resultant clusters. That is, provided that the experimental setting and clustering procedure remain consistent across all models. Figure 3 illustrates the clustering framework visually.

5.2. Data Acquisition and Preprocessing

We fill in the footsteps of previous research about event recommendations [5, 11, 52] and create the dataset using events from the platform Meetup.com. Meetup.com is an online event-social networking platform (EBSN) that facilitates online and face-to-face meetings, known as Meetups. The platform is built around a web-based structure of groups where individuals with similar interests can collaborate, plan, create, comment, share, and promote cultural events.

We used the Meetup GraphQL API in November 2022 to perform a comprehensive crawl of all publicly available activity on the platform from the ten largest cities located in the USA[8], namely, New York, Los Angeles, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San

[8] Last visited 25.04.2023: https://worldpopulationreview.com/us-cities

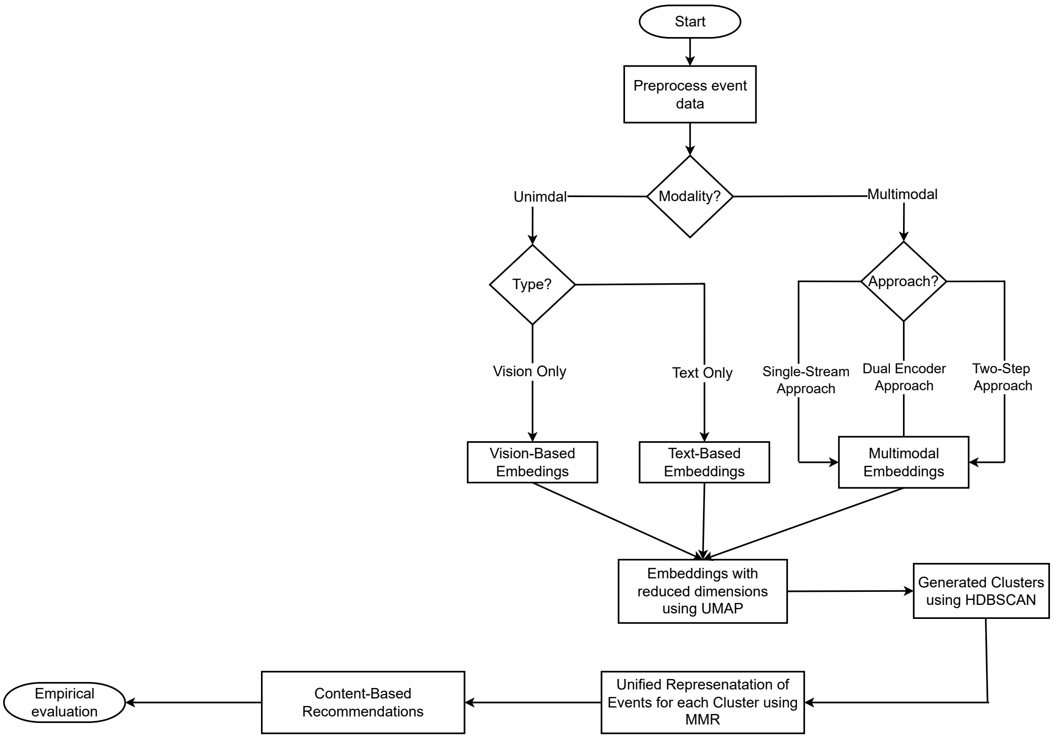


Figure 3: The clustering framework for content-based event recommendations.

Diego, Dallas, and San Jose, California. The resulting collection served as the training dataset, on which we trained all FMs models. In April 2023, we performed another crawl to create a second independent set for testing purposes. This way, we emulate real-world conditions by delivering recommendations for events that are scheduled to occur in the future.

We selected these particular cities for the analysis based on two primary reasons: Firstly, as the largest cities in the United States, they are considered to be among the most popular and vibrant locations, which often corresponds with a high level of event activity. Secondly, these cities are spread across different states, which provides a level of cultural diversity[9]. In total, the collection comprises $13,685$ distinct events, with $10,658$ in the training set and $3,027$ in the testing set. For each event, we obtained (i) the event image and (ii) the event description and its title. The events do not have true labels, thus, our setting is unsupervised.

Prior to the analysis, we carried out several standard NLP preprocessing steps such as excluding non-English texts and removing special characters HTML tags and hashtags. Additionally, we applied regex [53] filtering to exclude any irrelevant information such as Zoom invitations, email addresses, phone numbers, and URLs from the text data. Further preprocessing steps are handled by the respective models' tokenizers.

Regarding the images, we applied data augmentation including resizing, cropping, and normalization techniques for the vision-based and VL-FMs separately as outlined in their

[9] A similar approach was applied in [5].

original research papers.

5.3. Quantitative Empirical Evaluation

One significant limitation of clustering research is the lack of validation methods for unlabeled data. Many studies in this area of research attempt to solve the clustering problem on datasets with known true labels. However, in real-life situations, having prior knowledge of the ground truth is often the exception rather than the norm. There are few external evaluation methods available that can be applied in such settings, like the silhouette coefficient [54] and the density-based clustering validation metric (DBCV) [55]. While these metrics may offer valid alternatives to measure the clustering performance, they have their own limitations and are consequently not applicable in many actual situations. The Silhouette Coefficient, for example, assumes that the clusters are convex and well-separated, which is not always the case in density-based clustering. On the other hand, although developed for density-based clustering, DBCV suffers from increasing run times as the number of data points increases.

These reasons motivated us to validate the results by justifying them with human perception through a series of online user experiments. We leveraged the crowdsourcing service Prolific[10] to recruit 300 human evaluators, all of which passed an attention check. We presented users with content-based recommendations founded on the VL-FMs introduced in this study and prompted them to evaluate these recommendations along three dimensions: similarity-based evaluation, comparison-based evaluation, and clustering assignment evaluation.

All in all, the recommendations were based on one of the following types: (i) Text-based embeddings. (ii) Vision-based embeddings. (iii) VL-based embeddings obtained through single-stream fusion encoder approach. (iv) VL-based embeddings obtained through the dual encoder approach. (v) VL-based embeddings obtained through the two-step approach. As a baseline, we match and recommend events randomly.

In the first set of experiments, participants were presented with two events displayed side by side. For each event, participants viewed the corresponding image, title, and description. Both events were sampled randomly from the same cluster and had the same embedding type. The events were randomly placed on either the left or right side and were not presented in any particular order.

As previously mentioned, clustering-based recommendations rely on the assumption of similarities among events within a cluster and consider other events within the same cluster as relevant. We attempt to examine this assumption by modeling participants' perceived similarity and cross event interest between two such events. We refer to Figure 5 in Appendix for an illustrated example.

In the second set, participants were shown one event and two clusters, with one cluster being the correct cluster to which the event belonged. The events and clusters were randomly sampled and based on the same embedding type. We created the cluster visualizations by drawing representative exemplars from each cluster (as explained in subsection 5.1), and then consolidating them into one single visual representation. To evaluate the representations quality, we gauge the agreement between the clustering assignments and the participants' perceived

[10] https://internal-api.prolific.co/

memberships by asking them to indicate the cluster to which they believed the event belonged. We refer to Figure 8 in Appendix for an illustrated example.

In the last set, we asked the respondents to choose the form of modality (text,image or both) that they believe had the most significant impact on their judgment and to motivate their choice. Here, we attempted to model the participants' preferences and perceptions regarding the influence of different modalities on their decision-making process.

Furthermore, to ensure that the participants' attention is not diverted to or biased by secondary information such as the varying length of descriptions, we constrained the length of the description for the sampled events to one standard deviation around the mean length of all deceptions. We reconstructed the events by parsing the raw HTML data and converting it into a structured HTML document. Before this process, we took measures to protect privacy by removing sensitive personal data and anonymizing images containing individuals using Gaussian blurring. The remaining settings, including image resolution and font size, were left unaltered. In the last step, we integrated the reconstructed events into windows of equal sizes and blended them into the experiment's user interface.

6. Results and Discussion

In this section, we detail the findings of the study. We commence by delivering quantitative measures and further support these findings with a concise qualitative analysis.

6.1. Similarity-Based Evaluation

For the similarity-based evaluation (corresponding to the first set of experiments), we carried out a series of simple linear regressions in which we regressed the variable of interest against the type of embeddings used. Table 1 displays the regression results of the perceived similarity degree across the reference groups, random, vision-based and text-based, respectively.

First and foremost, the results indicate that recommendations based on event representations, as encoded by the FMs, performed on average significantly better than a naive random baseline. The two-step approach emerged, notably, as the most superior among all models.

The results also suggest, to our surprise, that text-based embeddings led to recommendations that were either comparable or in some cases even superior in the perceived similarity degree to their multimodal counterparts. This strongly suggests that text-based clusters seem to satisfy the assumption of semantic similarity and demonstrate comparable traits.

Additionally, when pitted against vision-based embeddings, all models once again achieved superior performance with the two-step approach surpassing all others. Comparing the random baseline with vision-based embeddings however yielded insignificance. This implies that there was minimal distinction in the perceived similarity between recommendations based solely on visual similarities and a randomly selected sample.

To explore this in more detail, we looked into cases where the similarities between events were rated high by the study participants and compared them to those rated low[11].

[11] In all examples, we retrieved similar images under Creative Commons licenses due to copyright reasons from https://commons.wikimedia.org (last visited on 10.09.2023)

Vision-based clusters can be interpreted along two dimensions: content and context. The content dimension reflects the entities that are visible in the images, and it is the most intuitive dimension. In contrast, the context dimension reflects the circumstances in which the content occurs [56]. We notice homogeneous clusters on the content dimension, yet recommendations derived from them may not be reliable as the context is mostly unclear.

Further, the instances where participants rated the events as highly similar were relatively limited. These cases featured events that seem to have images matching across the content and the context dimensions. As an example for such a case we provide Figure 6, in which the event images of a sample recommendation are shown. Both images depict a book and both events were about literary gatherings. On the other hand, instances rated as dissimilar by the participants appeared to feature events with images that matched only in terms of their content. Figure 6 in the Appendix visually displays such an example. While both images share similar attributes, there is an absence of contextual similarity between both events. To elaborate, the event depicted on the left pertains to travel and tourist excursions, whereas the event on the right is centered around meditation journeys. These observations could potentially therefore reaffirm the notion that vision models encode images primarily based on content, lacking the incorporation of sufficient semantic signals within them.

Furthermore, and perhaps most importantly, we conducted a regression analysis to examine the relationship between cross-event interest as a dependent and embeddings type as an independent variable. The rationale behind this is that the perceived similarity may not always convey a complete picture. For instance, two events with the same image color are likely to be perceived as similar but the two events may still be completely different.

Table 2 demonstrates that cross interest is significantly higher to the text-based and two-step approach, thereby confirming that instances with higher perceived similarity indeed correspond to higher cross interest levels.

We also found that all multimodal embeddings resulted in clusters that are more varied and detailed in comparison to those achieved by their unimodal counterparts, all the while maintaining comparable perceived similarity and cross-event interest as the text-based ones. This observation could hint that the VL-FMs provided richer and more ample event representations, as Figure 7 in Appendix shows.

6.2. Clustering Assignment Evaluation

In the second set of experiments, we attempted to assess the clustering quality by measuring the agreement degree between human judgment and clustering assignment.

As can be seen in Table 5 in Appendix, the agreement metric appears to vary across the different embeddings' types. In particular, recommendations founded on embeddings created by the VL two-step approach showed the highest agreement frequencies. To test for statistical significance, we opted for a simple logistic regression, modeling the variable agreement as the dependent variable and the type of embedding as the independent variable. The two-step approach was set as the reference level.

The coefficient estimates in Table 3 suggest that clusters formed using the two-step approach tend to exhibit notably higher agreement probabilities in comparison to other VL and text-based approaches, although not when compared to the vision-based approach.

Table 1
Similarity-based evaluation: Perceived similarity between events.

	Dependent variable: Perceived Similarity		
	(1)	(2)	(3)
Intercept	2.100***	2.380***	4.340***
	(0.348)	(0.348)	(0.348)
Treatment(reference=Random)			
C(Vision-Based)	0.280		
	(0.493)		
C(Text-Based)	2.240***		
	(0.493)		
C(VL Single-Stream Approach)	1.120**		
	(0.493)		
C(VL Dual Encoder Approach)	1.840***		
	(0.493)		
C(VL Two-Step Approach)	2.340***		
	(0.493)		
Treatment(reference=Vision-Based)			
C(Random)		-2.280	
		(0.493)	
C(Text-Based)		1.960***	
		(0.493)	
C(VL Single-Stream Approach)		0.840*	
		(0.493)	
C(VL Dual Encoder Approach)		1.560***	
		(0.493)	
C(VL Two-Step Approach)		2.060***	
		(0.493)	
Treatment(reference=Text-Based)			
C(Random)			-2.240***
			(0.493)
C(Vision-Based)			-1.960***
			(0.493)
C(VL Single-Stream Approach)			-1.120**
			(0.493)
C(VL Dual Encoder Approach)			-0.400
			(0.493)
C(VL Two-Step Approach)			0.100
			(0.493)
Observations	300	300	300
R^2	0.123	0.123	0.123
Adjusted R^2	0.108	0.108	0.108
Residual Std. Error	2.464 (df=294)	2.464 (df=294)	2.464 (df=294)
F Statistic	8.266*** (df=5; 294)	8.266*** (df=5; 294)	8.266*** (df=5; 294)
Note:			*p<0.1; **p<0.05; ***p<0.01

This observation aligns with the principles of Gestalt [57, 58], and particularly resonates with the concept of similarity, which refers to our capacity to perceive identical visual elements as cohesive units. Items sharing resemblances in terms of shapes and colors for instance are likely perceived to fall within the same category.

6.3. Comparison-Based Evaluation

A Chi Square test of independence revealed that among participants, the form of modality that they believe had the most significant impact on their judgment and embeddings type were significantly associated, $\chi^2(10) = 29.36, p < 0.001$. Post hoc comparisons in Table 4 for each pair, with FDR correction applied, revealed that participants who were shown random events were more likely to answer based on event descriptions. In comparison, statistical similarity was observed among all other cases.

Table 2
Similarity-based evaluation: Cross event interest.

	Dependent variable: Cross Event Interest
	(1)
Intercept	3.360***
	(0.355)
Treatment(reference=Random)	
C(Vision-Based Approach)	0.420
	(0.502)
C(Text-Based Approach)	1.260**
	(0.502)
C(VL Single-Stream Approach)	0.660
	(0.502)
C(VL Dual Encoder Approach)	0.320
	(0.502)
C(VL Two-Step Approach)	1.100**
	(0.502)
Observations	300
R^2	0.030
Adjusted R^2	0.014
Residual Std. Error	2.511 (df=294)
F Statistic	1.833 (df=5; 294)
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01

Table 3
Clustering assignment evaluation: Agreement between clustering assignment and human judgment.

	Dependent variable: Agreement	
Treatment(reference=VL Two-Step Approach)		
C(Vision-Based)		-0.926
		(0.638)
C(Text-Based)		-1.953***
		(0.597)
C(VL Single-Stream Approach)		-1.689***
		(0.603)
C(VL Dual Encoder Approach)		-1.290 **
		(0.618)
Intercept	1.516***	2.442***
	(0.368)	(0.521)
Observations	250	250
Note:	*p$<$0.1; **p$<$0.05; ***p$<$0.01	

The results suggest therewith that both the textual and visual stimuli were perceived as equally important regardless of the embedding type.

7. Conclusion

This work has showcased the potential of harnessing contemporary uni- and multimodal FMs as a means for event data representation. We contend that (i) these models present themselves as effective tools for appropriately capturing event content, and that (ii) content-based recommendations grounded on these models can effectively address the cold-start problem in event recommendations.

We started off by building a collection of events from the event-based platform Meetup.

Table 4
Comparison-based evaluation: Post hoc pairwise pearson χ^2 comparisons with FDR correction.

	Comparison	p.Chisq	p.adj.Chisq	Cramer.V
1	VL Dual Encoder Approach : VL Two-Step Approach	0.25	0.41	0.17
2	VL Dual Encoder Approach : Vision-Based	0.24	0.41	0.17
3	VL Dual Encoder Approach : Random	0.03	**0.09***	**0.26**
4	VL Dual Encoder Approach: Text-Based	0.57	0.70	0.11
5	VL Dual Encoder Approach : VL Single-Stream Approach	0.43	0.59	0.13
6	VL Two-Step Approach : Vision-Based	0.66	0.71	0.09
7	VL Two-Step Approach : Random	0.00	**0.00*****	**0.42**
8	VL Two-Step Approach : Text-Based	0.35	0.52	0.14
9	Two-Step : VL Single-Stream Approach	0.61	0.70	0.10
10	Vision-Based : Random	0.00	**0.00*****	**0.41**
11	Vision-Based : Text-Based	0.14	0.35	0.20
12	Vision-Based : VL Single-Stream Approach	0.91	0.91	0.04
13	Random : Text-Based	0.00	**0.02****	**0.33**
14	Random : VL Single-Stream Approach	0.00	**0.01****	**0.37**
15	Text-Based : VL Single-Stream Approach	0.22	0.41	0.17

Subsequently, we transformed these into informative vector representations using a selected set of FMs. The resultant representations were arranged into clusters, bringing together events with common characteristics. The clusters served in turn as the corner stone for generating content-based recommendations. we then executed a series of user experiments to ascertain the quality of these recommendations with regards to three dimensions: similarity-based evaluation, comparison-based evaluation, and clustering assignment evaluation. The findings of this study can be summarized to : (i) Recommendations generated through the use of FMs outperformed a naive baseline that randomly drew events, with the VL two-step approach emerging as the top performer. (ii) Surprisingly, text-based embeddings performed on par or even surpassed multimodal VL embeddings in some cases. (iii) All multimodal VL-FMs yielded more diverse clusters, each with different scopes, while still demonstrating comparable perceived similarity to the text-based clusters. (v) The agreement between cluster assignments and human judgment revealed that the two-step approach was statistically superior to all other models except for vision-based embeddings, aligning with the principles of Gestalt theory. (iv) The results indeed verified that cross interest is associated with perceived similarity, confirming that similar content-based recommendations are more likely to be regarded as relevant.

Nonetheless, the empirical results reported herein should be considered in light of some limitations. First, the baseline is limited to a naive random approach. Additional standard content-based extraction methods could be considered to substantiate the superior performance of FM models. Second, the dataset was sourced from Meetup during the months of November and April of 2023, limiting the scope of our data to the type of events commonly promoted on the platform during these months. In order to offer a more thorough representation of the broader spectrum of events, further experiments spanning extended time frames are imperative.

In conclusion, our experiments indicate that leveraging the capabilities of modern FMs for content-based event recommendations could effectively address the cold-start problem. It would be interesting to test these models in a real-world recommendation setting with post-recommendation human feedback. We view our study as a step in this direction and hope that our results can support the development and application of future technologies in this field.

References

[1] W. Wörndl, A. Hefele, D. Herzog, Recommending a sequence of interesting places for tourist trips, Information Technology & Tourism 17 (2017) 31–54.

[2] I. Partalas, A. Morvan, A. SADEGHIAN, S. MINAEE, X. LI, B. COWAN, D. Z. WANG, Hotel2vec: Learning hotel embeddings from user click sessions with side information (2021).

[3] Y. Zhang, Q. Ai, X. Chen, W. B. Croft, Joint representation learning for top-n recommendation with heterogeneous information sources, in: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, 2017, pp. 1449–1458.

[4] Y. Du, X. Meng, Y. Zhang, Cvtm: A content-venue-aware topic model for group event recommendation, IEEE Transactions on Knowledge and Data Engineering 32 (2019) 1290–1303.

[5] A. Q. Macedo, L. B. Marinho, R. L. Santos, Context-aware event recommendation in event-based social networks, in: Proceedings of the 9th ACM Conference on Recommender Systems, 2015, pp. 123–130.

[6] W. Zhang, J. Wang, A collective bayesian poisson factorization model for cold-start local event recommendation, in: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, 2015, pp. 1455–1464.

[7] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al., On the opportunities and risks of foundation models (2023).

[8] J. D. M.-W. C. Kenton, L. K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[9] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system: A survey and new perspectives, ACM computing surveys (CSUR) 52 (2019) 1–38.

[10] S. Madisetty, Event recommendation using social media, in: 2019 IEEE 35th International Conference on Data Engineering (ICDE), IEEE, 2019, pp. 2106–2110.

[11] T. Trinh, D. Wu, R. Wang, J. Z. Huang, An effective content-based event recommendation model, Multimedia Tools and Applications 80 (2021) 16599–16618.

[12] E. Minkov, B. Charrow, J. Ledlie, S. Teller, T. Jaakkola, Collaborative future event recommendation, in: Proceedings of the 19th ACM international conference on Information and knowledge management, 2010, pp. 819–828.

[13] U. Javed, K. Shaukat, I. A. Hameed, F. Iqbal, T. M. Alam, S. Luo, A review of content-based and context-based recommendation systems, International Journal of Emerging Technologies in Learning (iJET) 16 (2021) 274–306.

[14] M. M. Afsar, T. Crump, B. Far, Reinforcement learning based recommender systems: A survey, ACM Computing Surveys 55 (2022) 1–38.

[15] T. De Pessemier, S. Coppens, E. Mannens, S. Dooms, L. Martens, K. Geebelen, An event distribution platform for recommending cultural activities, in: 7th International Confer-

ence on Web Information Systems and Technologies (WEBIST-2011), Ghent University, Department of Information technology, 2011, pp. 231–236.

[16] M. Xu, S. Liu, Semantic-enhanced and context-aware hybrid collaborative filtering for event recommendation in event-based social networks, IEEE Access 7 (2019) 17493–17502.

[17] Y. Jhamb, Y. Fang, A dual-perspective latent factor model for group-aware social event recommendation, Information Processing & Management 53 (2017) 559–576.

[18] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE transactions on pattern analysis and machine intelligence 35 (2013) 1798–1828.

[19] P. Mohan, B. C. Brilley, N. K. AA, Multimodal representation learning: Cross-modality and shared representation, in: 2022 International Conference on Industry 4.0 Technology (I4Tech), IEEE, 2022, pp. 1–5.

[20] G. E. Hinton, R. Zemel, Autoencoders, minimum description length and helmholtz free energy, Advances in neural information processing systems 6 (1993).

[21] C. Zhang, Z. Yang, X. He, L. Deng, Multimodal intelligence: Representation learning, information fusion, and applications, IEEE Journal of Selected Topics in Signal Processing 14 (2020) 478–493.

[22] P. Thompson, Margaret thatcher: A new illusion, Perception 9 (1980) 483–484.

[23] W. Guo, J. Wang, S. Wang, Deep multimodal representation learning: A survey, IEEE Access 7 (2019) 63373–63394.

[24] Y. Du, Z. Liu, J. Li, W. X. Zhao, A survey of vision-language pre-trained models, arXiv preprint arXiv:2202.10936 (2022).

[25] P. Xu, X. Zhu, D. A. Clifton, Multimodal learning with transformers: A survey, IEEE Transactions on Pattern Analysis and Machine Intelligence (2023).

[26] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8342–8360.

[27] K. Wang, N. Thakur, N. Reimers, I. Gurevych, Gpl: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2022, pp. 2345–2360.

[28] K. Wang, N. Reimers, I. Gurevych, Tsdae: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 671–688.

[29] S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, A. Hanbury, Improving efficient neural ranking models with cross-architecture knowledge distillation, arXiv preprint arXiv:2010.02666 (2020).

[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[31] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, R. Girshick, Masked autoencoders are scalable vision learners, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 16000–16009.

[32] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell,

P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.

[33] W. Kim, B. Son, I. Kim, Vilt: Vision-and-language transformer without convolution or region supervision, in: International Conference on Machine Learning, PMLR, 2021, pp. 5583–5594.

[34] R. Girshick, Fast r-cnn, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.

[35] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, T. Duerig, Scaling up visual and vision-language representation learning with noisy text supervision, in: International conference on machine learning, PMLR, 2021, pp. 4904–4916.

[36] K.-H. Lee, X. Chen, G. Hua, H. Hu, X. He, Stacked cross attention for image-text matching, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 201–216.

[37] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, Information Fusion 77 (2022) 149–171.

[38] W. Chai, G. Wang, Deep vision multimodal learning: Methodology, benchmark, and trend, Applied Sciences 12 (2022) 6588.

[39] J. Wang, Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, L. Wang, Git: A generative image-to-text transformer for vision and language, arXiv preprint arXiv:2205.14100 (2022).

[40] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, When is "nearest neighbor" meaningful?, in: Database Theory—ICDT'99: 7th International Conference Jerusalem, Israel, January 10–12, 1999 Proceedings 7, Springer, 1999, pp. 217–235.

[41] C. C. Aggarwal, A. Hinneburg, D. A. Keim, On the surprising behavior of distance metrics in high dimensional space, in: Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8, Springer, 2001, pp. 420–434.

[42] M. Steinbach, L. Ertöz, V. Kumar, The challenges of clustering high dimensional data, New directions in statistical physics: econophysics, bioinformatics, and pattern recognition (2004) 273–309.

[43] D. Pandove, S. Goel, R. Rani, Systematic review of clustering high-dimensional and large datasets, ACM Transactions on Knowledge Discovery from Data (TKDD) 12 (2018) 1–68.

[44] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[45] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (2010) 433–459.

[46] L. Van der Maaten, G. Hinton, Visualizing data using t-sne., Journal of machine learning research 9 (2008).

[47] M. Allaoui, M. L. Kherfi, A. Cheriet, Considerably improving clustering algorithms using umap dimensionality reduction technique: a comparative study, in: Image and Signal Processing: 9th International Conference, ICISP 2020, Marrakesh, Morocco, June 4–6, 2020, Proceedings 9, Springer, 2020, pp. 317–325.

[48] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering., J. Open Source Softw. 2 (2017) 205.

[49] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[50] M. Grootendorst, Concept, https://github.com/MaartenGr/Concept, 2022. GitHub repository.

[51] A. Z. Broder, On the resemblance and containment of documents, in: Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171), IEEE, 1997, pp. 21–29.

[52] X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, J. Han, Event-based social networks: linking the online and offline social worlds, in: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, 2012, pp. 1032–1040.

[53] A. V. Aho, Algorithms for finding patterns in strings, handbook of theoretical computer science (vol. a): algorithms and complexity, 1991.

[54] P. J. Rousseeuw, Silhouettes: a graphical aid to the interpretation and validation of cluster analysis, Journal of computational and applied mathematics 20 (1987) 53–65.

[55] D. Moulavi, P. A. Jaskowiak, R. J. Campello, A. Zimek, J. Sander, Density-based clustering validation, in: Proceedings of the 2014 SIAM international conference on data mining, SIAM, 2014, pp. 839–847.

[56] J. Klostermann, A. Plumeyer, D. Böger, R. Decker, Extracting brand information from social networks: Integrating image, text, and social tagging data, International Journal of Research in Marketing 35 (2018) 538–556.

[57] K. Koffka, Principles of Gestalt psychology, volume 44, Routledge, 2013.

[58] M. Wertheimer, Laws of organization in perceptual forms. (1938).

A. Appendix

Table 5
Clustering assignment evaluation: Absolute and relative frequencies for agreement.

	Vision-Based	Text-Based	Single-Stream App.	Dual Encoder App.	Two-Step App.
0	0.18 (9)	0.38 (19)	0.32 (16)	0.24 (12)	0.08 (4)
1	0.82 (41)	0.62 (31)	0.68 (34)	0.76 (38)	**0.92 (46)**

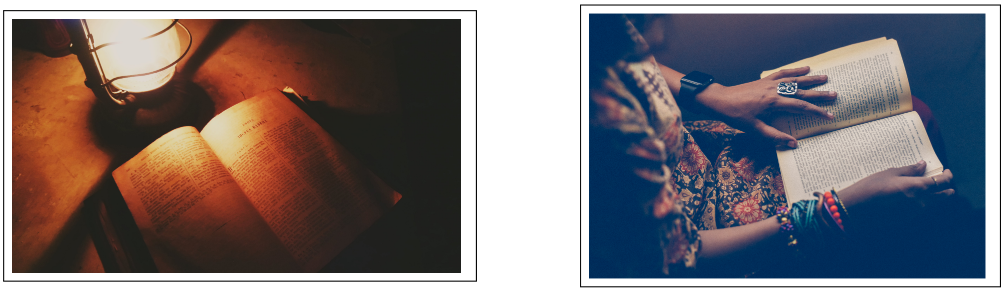
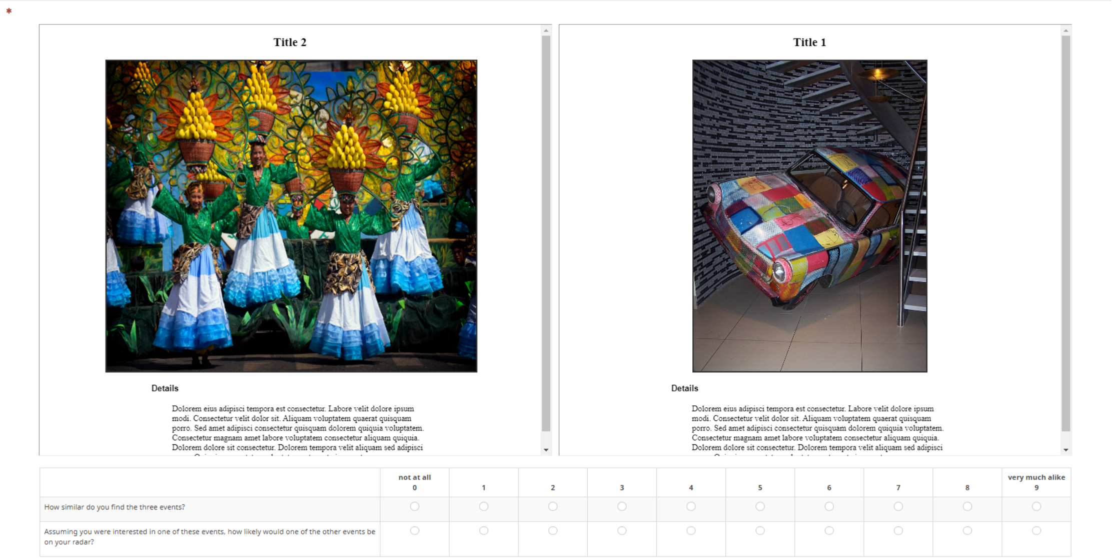


Figure 4: Example of a good recommendation based on vision-based embeddings. Images are sourced from https://commons.wikimedia.org (last visited on 10.09.2023).



*

Title 2

Details

Dolorem eius adipisci tempora est consectetur. Labore velit dolore ipsum modi. Consectetur velit dolor sit. Aliquam voluptatem quaerat quisquam porro. Sed amet adipisci consectetur quisquam dolorem quiquia voluptatem. Consectetur magnam amet labore voluptatem consectetur aliquam quiquia. Dolorem dolore sit consectetur. Dolorem tempora velit aliquam sed adipisci

Title 1

Details

Dolorem eius adipisci tempora est consectetur. Labore velit dolore ipsum modi. Consectetur velit dolor sit. Aliquam voluptatem quaerat quisquam porro. Sed amet adipisci consectetur quisquam dolorem quiquia voluptatem. Consectetur magnam amet labore voluptatem consectetur aliquam quiquia. Dolorem dolore sit consectetur. Dolorem tempora velit aliquam sed adipisci

	not at all 0	1	2	3	4	5	6	7	8	very much alike 9
How similar do you find the three events?	○	○	○	○	○	○	○	○	○	○
Assuming you were interested in one of these events, how likely would one of the other events be on your radar?	○	○	○	○	○	○	○	○	○	○

Figure 5: An example for the first set of experiments. Images are sourced from https://commons.wikimedia.org (last visited on 10.09.2023).

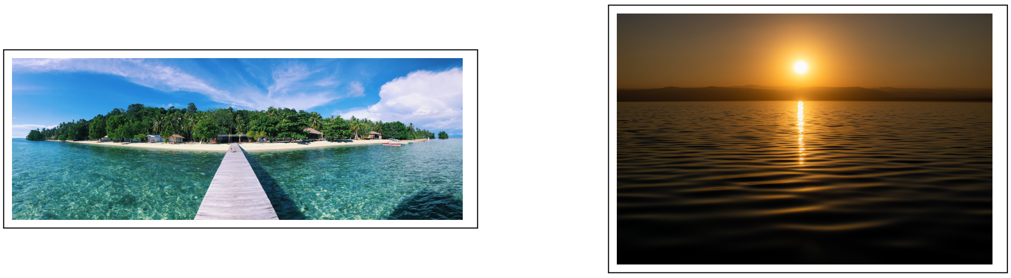


Figure 6: Example of a Poor recommendation based on vision-based embeddings. Images are sourced from https://commons.wikimedia.org (last visited on 10.09.2023).

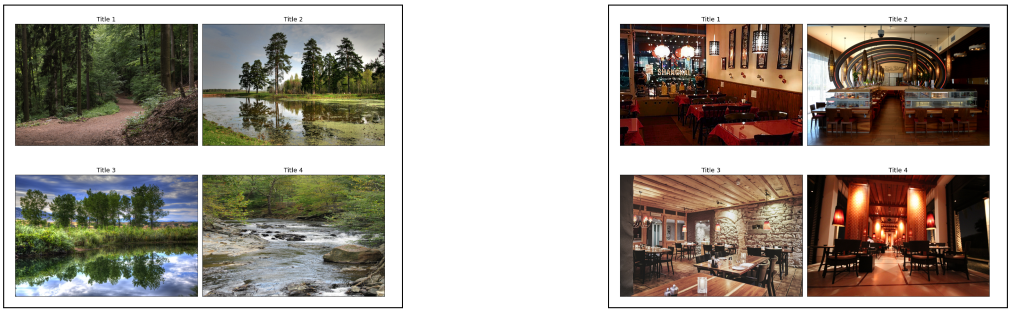
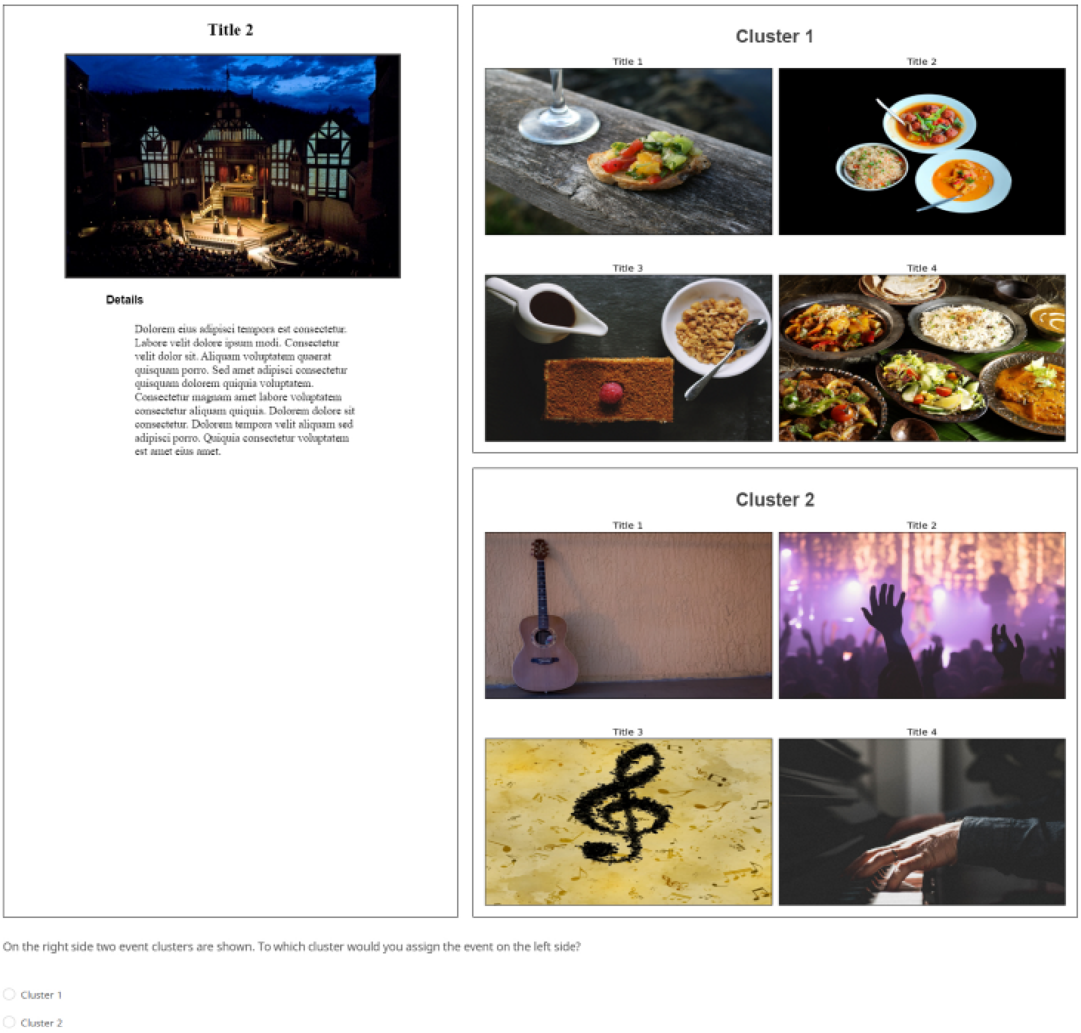


Figure 7: Examples of Recommendations Based on VL-Based Embeddings. Images are sourced from https://commons.wikimedia.org (last visited on 10.09.2023).



Details

Dolorem eius adipisci tempora est consectetur. Labore velit dolore ipsum modi. Consectetur velit dolor sit. Aliquam voluptatem quaerat quisquam porro. Sed amet adipisci consectetur quisquam dolorem quiquia voluptatem. Consectetur magnam amet labore voluptatem consectetur aliquam quiquia. Dolorem dolore sit consectetur. Dolorem tempora velit aliquam sed adipisci porro. Quiquia consectetur voluptatem est amet eius amet.

On the right side two event clusters are shown. To which cluster would you assign the event on the left side?

○ Cluster 1

○ Cluster 2

Figure 8: An example for the second set of experiments. Images are sourced from https://commons.wikimedia.org (last visited on 10.09.2023).