Leveraging VLLMs for Visual Clustering: Image-to-text mapping shows increased semantic capabilities and interpretability*

Luigi Arminio[1], Matteo Magnani[2], Matias Piqueras[2], Luca Rossi[1], and Alexandra Segerberg[3]

[1] *IT University of Copenhagen, Denmark*
[2] *InfoLab, Dept. of Information Technology, Uppsala University, Sweden*
[3] *Dept. of Government, Uppsala University, Sweden*

This version: July 19, 2024

Abstract

Automated image categorization is vital for computational social science, particularly considering the rise of visual content on social media, as it helps the identification of emerging visual narratives in online debates. However, the methods currently used in the field to represent images numerically are unable to fully capture their connotative meaning and do not produce interpretable clusters. In response to these challenges, we evaluate an approach based on the automated generation of intermediate textual descriptions of the input images with respect to the connotative semantic validity of the generated clusters and their interpretability. We show that both aspects are improved over the currently typical clustering approach based on convolutional neural networks.

keywords: Image clustering, Vision Large Language Models, Semantic clustering, Connotation, Interpretability

1 Introduction

Automated image categorization represents a valuable analytical tool in the field of computational social science [1, 2, 3]. The visual turn in contemporary social media [4, 5, 6] has further increased the importance of computational analysis and the clustering of visual content for a wide range of research that leverage online social data [7, 8, 9, 10], including social media analysis [11, 12, 13, 14, 1].

*L.A., L.R., M.M., M.P. designed the study. L.A., M.M. implemented the clustering pipeline. L.R., A.S. performed the qualitative validation. All authors have revised and approved the manuscript.

In this context the automated annotation and categorization of visual content can be used to better understand a wide range of collective patterns including the proliferation of misinformation [15, 16] and the development of polarization dynamics [17, 18]. These and other potential applications have motivated the development of computational data analysis pipelines for visual content in social research [19, 20, 1, 21, 22].

Nevertheless, the state-of-the-art techniques used for image clustering still exhibit considerable limitations [23]. In particular, most traditional approaches tend to concentrate on object detection, succeeding in recognizing objects or patterns but potentially disregarding broader contextual and relational dynamics within pictures [24]. This implies that spatial relationships among visual elements and the external knowledge needed to interpret the image are neglected, making it difficult to go beyond its denotative representation [25, 26]. These limitations are more and more relevant considering the prominent role that internet memes [27, 28] and similar types of complex hybrid visual communication play in contemporary online social dynamics and political participation [29].

In response to these limitations, researchers have developed the concept of semantic clustering. Semantic clustering refers to the attempt to include as elements for the clustering both denotative aspects, such as the presence of specific objects, and connotative aspects, to capture the actual social meaning of what is represented [30]. The concept is exemplified in Figure 1. It has frequently been explored in the data science literature, and a variety of strategies to incorporate semantics into visual clustering have been proposed over time [31, 32, 33, 27]. Earlier studies focused on employing images' collateral textual information (such as captions or existing textual descriptions associated with a given image) along with low-level visual features in image clustering [31]. The reliance of this approach on the presence of textual annotations, though beneficial, remains a limitation, given the inconsistent availability of detailed textual metadata for all pictures and the fact that text and images may carry different meanings in online communication. Other researchers have attempted to derive semantic meanings from sets of visual items through low-level variables such as the color layout of the images and MPEG-7 color-structure descriptors [32]. However, while this approach was effective for preliminary screening tasks in image retrieval, it proved inadequate for achieving accurate semantic clustering [32]. Recent work attempts to integrate established methods such as Bag of Visual Words (BoVW) into the clustering pipeline, leading to an approach called Bag of Visual Phrases (BoVP), which effectively identifies dominant visual patterns in image collections while also connecting objects to their meanings and addressing their spatial relationships [33]. However, since it prioritizes dominant visual cues, this kind of approach might obscure elements that are less prominent but still significant for identifying visual narratives, particularly in the context of political analysis. For example, in the picture of a collective event like a political rally, approaches based on BoVP may not prioritize small and diverse visual elements such as symbols, party logos, or specific banners. This may in turn overlook significant differences between apparently similar vi-

Figure 1: Example of two images showing denotative difference but connotative similarity that should be clustered together in semantic clustering.

sual representations that actually refer to disparate socio-political subcultures, collective actors, or claims. An alternative approach is the Semantic-Enhanced Image Clustering (SIC) method [24], employing the CLIP neural network model, pre-trained on a large dataset of image-text pairs, to convert both pictures and their textual label into high-dimensional vectors, establishing a multi-modal embedding space [34]. In this case, these vectors are projected into a semantic space constructed from a refined list of nouns from the Wordnet lexical database [24, 35]. With this approach, SIC aims at categorising images based on their semantic similarity, harnessing CLIP's ability to model image-text relationships and the semantic variations captured by specific nouns in WordNet [24]. SIC methodology, leveraging CLIP's multimodal capabilities, may enable improvement in understanding subtle connections between visual elements within images [36]. However, this clustering approach still leans towards explicit elements characterizing the pictures, which might be inadequate when the goal shifts from mere clustering based on visual structures to categorizing images based on their implicit semantic meanings, symbolic representations and subtext that may instead be captured by LLMs' descriptive potential to capture latent contextual nuances within pictures [37, 38, 39]. In addition, the direct translation of images to feature vectors, whose individual elements are not directly associated to specific concepts as it is instead the case in word-based representations of text, hinders interpretability.

In this context, we propose to address these limitations and further develop the state of the art of connotative semantic and interpretable clustering of visual content for social research, by leveraging the visual capabilities of recent large multimodal models [40]. Large Language Models with visual capabilities (VLLMs) have been able to identify subtle nuances and contextual clues [41, 42] and recent extensive work shows the remarkable capabilities of contextual understanding offered by these models [43]. Therefore, we can use VLLMs to produce textual descriptions of the input images, specifically asking to focus on connotative elements. This then makes it possible to compute the clusters

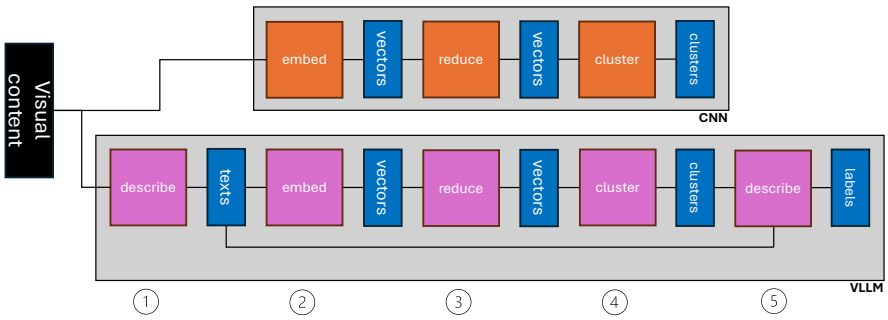


Figure 2: Pipeline

(or other summaries of the input images) working directly with the text. We compare the semantic quality of such clustering with the semantic quality of the clustering obtained using pretrained Convolutional Neural Networks (CNN) [1], that are commonly used in the recent literature on image clustering for the social sciences and that have shown state-of-the-art performance on the same data used in our experiments [1]. Differently from previous work, we assess semantic quality both with respect to the denotative and connotative similarity of images included in the same or different clusters.

In particular, we address the following research questions:

RQ1 Can a VLLM-based image clustering pipeline improve the connotative semantic validity of clusters compared with a CNN-based pipeline?

RQ2 Are the results obtained with a VLLM-based image clustering pipeline interpretable by the end-user?

2 Materials and methods

Figure 2 shows the two pipelines tested in this paper. The current mainstream approach relies on a CNN trained on an image recognition task to extract visual features (S2), followed by dimensionality reduction (S3) and clustering (S4) steps. Here we introduce Step 1 with the generation of textual descriptions using a VLLM (S1), followed by a textual embedding (S2). We also leverage the text produced in this phase to add an additional description step (S5) that increases the overall interpretability of the result. We note how the two parts of the pipeline that are different from the CNN-based pipeline are not in themselves new in the computer science literature [44]; we discuss the different angles explored by this study in the conclusion.

As a CNN we use VGG16, which has been shown to work well on the data used in this paper and so provides a strong baseline for comparison [1]. Other CNNs have also been tried, as reported in the appendix.

As a VLLM we use OpenAI's gpt-4-turbo. We note that some open source VLLMs also exist at the time of writing, e.g. LLaVA [45] and BLIP [46]. However the objective of this paper is not to compare alternative models to see what works best (which would also require a different experimental setup), but whether this type of models can improve connotative clustering. Therefore, we chose gpt-4-turbo as a state-of-the-art model providing fast executions if compared with open source models run over commodity hardware. As a prompt we used: "Describe the connotative meaning of this image, in one paragraph". We did not set model parameters (e.g., temperature), to test a default behavior, but manually checked that the generated texts provided good representations of the images instead. The only specific instruction we added ("in one paragraph") is used to generate texts that are not too short (so that there is space to provide a connotative description, mentioning what meaning the image conveys) and not too long (so that the model does not deviate from our instructions). We note that the model parameter limiting the number of tokens could not be used for this, as it resulted in truncated text.

To generate textual embeddings we used BERT (model: all-MiniLM-L6-v2), which is commonly used being open and having shown high accuracy in several studies. Given our research questions, we only need a good model (which we can evaluate downstream in the pipeline looking at the cluster results), as our objective is not to compare the accuracy of different text embedding models. We have however also tested the OpenAI text-embedding-3-small model, which has approximately the same accuracy of the large version of the same model while being significantly smaller (and thus cheaper and more energy-efficient). This does not lead to significant differences in our results; we report the details in the appendix.

Several dimensionality reduction methods exist, and we have no strong theoretical reasons to expect some to work better than others. Therefore, we tested one approach that preserves distances (UMAP) and a classical PCA, both for different parameter settings. For UMAP we tested 5 and 10 components, and 5, 15, and 30 nearest neighbors (for a total 6 settings), for PCA we tested 5, 10, and 20 components, and preserving 80% variability (4 settings).

As a clustering algorithm we used HDBSCAN, for three main reasons. First, it does not assume globular clusters or clusters of similar sizes, which are not guaranteed to be produced by UMAP. Second, it does not require to decide the number of clusters in advance. Third, it removes outliers, which is a reasonable choice for this study: our evaluation is focused on the generation of embeddings, and using a clustering method forcing all data points to belong to one cluster would make it difficult to discern whether a wrongly clustered image very different from the rest is a result of the embedding step, or if the embedding indeed clearly separated the image from the rest but the clustering method forced the image to be clustered anyway. HDBSCAN requires to specify the minimum cluster size. We argue that to be analytically valuable the minimum cluster size should be large enough, not to create too many clusters and not to create non-generalizing clusters. It should also not be too large, not to miss interesting clusters or to forcefully merge semantically diverse clusters. The specific value

depends on the research design, which implies that it should be optimized for specific empirical studies. Here we test 50, 100, and 200.

The data used in the experiments consists of 11873 images used in climate change communication [47].

Semantic and denotational validity

To evaluate if the VLLM-based approach creates higher-quality semantic clustering than the existing state of the art we adapt the cluster quality measure defined by [48]. In the original version, cluster quality measures the extent to which intracluster similarities outdistance intercluster similarities.

In this paper we split the similarity concept into *denotative* similarity, indicating the extent to which the images represent the same or similar visual elements, and *connotative* similarity, indicating the extent to which the two images refer to the same social meaning. In this way, for each clustering we can produce two scores assessing the extent to which they represent respectively denotative and connotative similarities.

We randomly drew 500 pairs of images and asked expert human coders (unaware if the images were sampled from the same cluster of from different clusters) to rate the denotative and connotative similarity of the images within each pair on a three-point scale: (1) unrelated, (2) loosely related, (3) closely related. In this paper we respectively assign numerical values 0, 1, and 2 to these three classes. Then, we compute the average score s_s for same-cluster pairs, and the average score s_d for different-cluster pairs. The final score is $s_s - s_d$, ranging from -2 to 2 and with expected value 0 for random assignments. The evaluation developed in three phases. First the coders rated denotative and connotative clustering quality for 250 pairs reaching an intercoder agreement measured with Krippendorff's α of .55 for the denotative scores and of .56 for the connotative scores. Then the coders discussed their differences during a consensus session and reached an agreement for all the scores. Then the coders individually assigned scores to the remaining 250 pairs and a new intercoder agreement was measured (α .81 for the denotative scores and .71 for the connotative scores). After that a new consensus session was held and consensus score was obtained for all the 500 pairs. We have also computed pair-counting F1-measure, which highlights the same trends in the results.

For this experiment, we only considered results where at least 50 pairs of manually labeled images were included in the same cluster and at least 50 pairs of manually labeled images were included in different clusters. This is not necessarily the case for all clusterings, because for 6 images we could not generate a text description and because some images are not included in any cluster but labeled as outliers. For example, the best clustering with respect to the connotative quality score labels 4027 images as outliers, leading to 82 same-cluster and 131 different-cluster pairs of images.

Interpretability

LLMs and more in general over-parametrized models have often been labeled as black boxes. However, we note that here we are interested in the interpretability of the clustering, which does not necessarily require to explain the models generating textual descriptions or embeddings, as argued in [49].

To assess the interpretability of the VLLM-based approach, we produced a description of each cluster by joining the descriptions of all images in the cluster and returning the k terms with the highest TF-IDF score. This is a common procedure in information retrieval, also available in commonly used text clustering software packages[1]. To reduce the number of terms and make the descriptions more compact, we removed terms not providing topic-specific meanings, such as *conveys* and *suggests*. Then we tested to what extent the descriptions corresponded to the clusters. To do so, we randomly selected sets of α images, where each set only contained images from the same cluster, and checked how often a human evaluator could correctly match a set with the explanation assigned by the algorithm. We also tested reliability, using three evaluators.

The reason for using $\alpha > 1$ is that we assume some images can be clustered by mistake. The fact that such images cannot be matched to the descriptions is not a sign of low interpretability, but a mistake in clustering that is already assessed as part of the semantic and denotational validity.

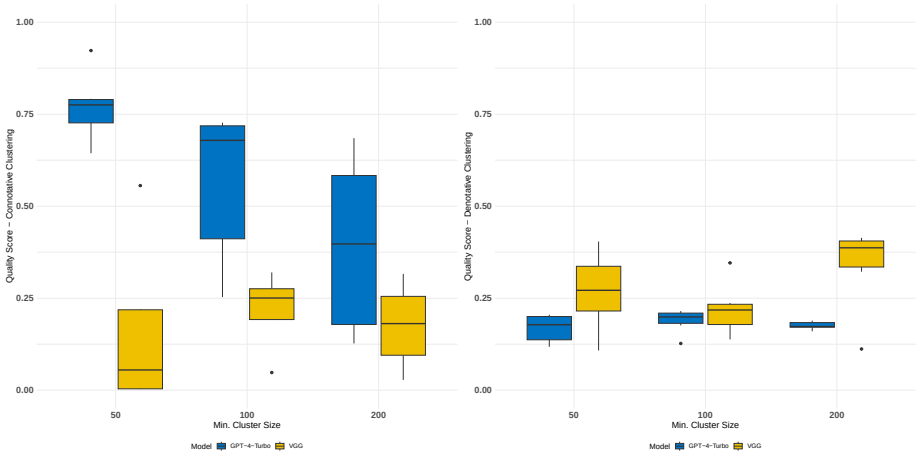
3 Results

3.1 Connotative and denotative clustering quality

Figure 3 shows a quality score for denotative (A) and connotative (B) clustering for the proposed VLLM-based strategy (using GPT-4-turbo as a specific model) and for the CNN-based clustering (using VGG16). This measure, defined in [48] for a generic similarity function and here differentiated into denotative and connotative similarities, is the difference between the average similarity of intra-cluster pairs of images and the average similarity of inter-cluster pairs of images, from a sample of pairs manually annotated (see the details in the methods section). Values range between -2 and +2, with 0 indicating a random clustering. Per each clustering strategy we show the scores per different values of the minimum cluster size parameter, with each boxplot summarizing the different scores obtained with multiple parameters for dimensionality reduction performed using UMAP [50].

The data show a higher quality of connotative clustering for the VLLM-based approach regardless of the minimum cluster size even if larger minimum cluster size degrades the performance. The quality of the connotative clustering obtained by the CNN-based approach is overall stable and it is not clearly affected by the minimum cluster size. The reason why the minimum cluster

[1] https://github.com/MaartenGr/BERTopic



(a) Connotative clustering quality

(b) Denotative clustering quality

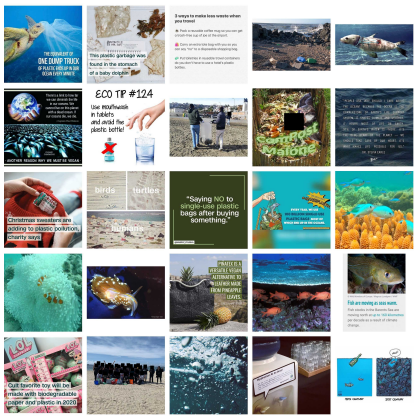
Figure 3: Clustering quality score for connotative (Figure 3a) and denotative (Figure 3b) clustering for VLLM (gpt-4-turbo) and CNN (VGG16). We visualize a boxplot to report the results for different strategies of dimensionality reduction.

size has such a profound impact on the connotative clustering quality of the VLLM-based approach seems to be that, forcing clusters to be larger, images that look similar but are connotatively diverse can be clustered together (e.g., images of street protests about two different issues).

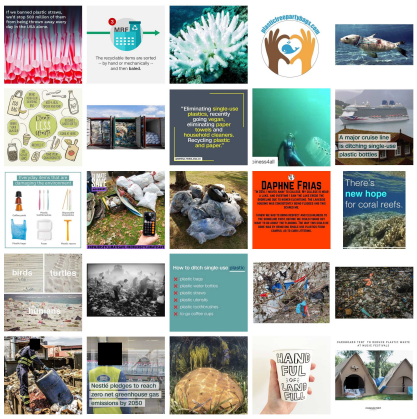
This can be clearly seen from comparing Figure 4a and 4b. The cluster represented by the random sample in Figure 4a has a very precise focus on the issue of plastics pollution in water, showing both sources and effect. The semantically comparable cluster showed in Figure 4b has a less well defined focus, mixing images referring to plastic pollution in the sea as well as on land together with images more generally connected with waste disposal and pollution.

Within this perspective, as it is often the case in clustering exercises, the right granularity of the concepts that should be clustered together is ultimately a decision for the researchers to make. The higher clustering quality of the VLLM-based approach is overall very clear peaking when the clustering algorithm allows smaller clusters.

Looking at the quality score for denotative clustering, we observe how the CNN-based approach is marginally better than the VLLM-based approach. This is unsurprising, since CNNs have proven capable of clustering images representing the same objects or similar pictures while the VLLM-based method can easily ignore shape similarity in favor of connotative semantic proximity. A good example of this that also exemplifies the limits of a denotative clustering, can be seen in Figure 5, where we compare a cluster obtained using the CNN-based approach (Figure 5a) with a cluster obtained with the VLLM-based one



(a) min. cluster size 50



(b) min. cluster size 200

Figure 4: An example of the impact of increasing minimum cluster size. Connotative semantic coherence of clustering degrades when moving from a minimum cluster size of 50 to a minimum cluster size of 200.

(a) CNN-based method (min.cl.size 50)

(b) VLLM-based method (min.cl.size 50)

Figure 5: A cluster obtained by the CNN-based approach (a) compared with a cluster obtained by the VLLM-based approach. The CNN-based approach favors shapes and object similarity while the VLLM-based approach ignores the visual difference between solar panels and wind turbines and produces a highly coherent cluster around the concept of renewable energy.

(Figure 5b). On the one side, it is clear how the CNN-based approach prioritizes shapes, objects and colors producing a cluster that, while centered around the visual representation of planet Earth, is semantically diverse (ranging from calls for veganism to climate data). On the other side, the VLLM-based approach clusters together the different shapes of wind turbines and solar panels with text-only messages calling for a transition to renewable energies. The resulting cluster is highly coherent from a connotative point of view at the expense of denotative coherence.

Building on these results, we can confidently answer the first research question. We demonstrate that a VLLM-based approach greatly improves the connotative semantic validity of visual clustering over a CNN-based approach from both a quantitative and a qualitative point of view. We also observe how adopting a VLLM-based approach results into a loss in quality of denotational clustering. This decline in denotational quality, albeit is small, is the side effect of introducing a connotative approach over a purely denotative one.

3.2 Interpretability

To assess interpretability, we compute a description of each of the 32 clusters generated by the best VLLM-based model, where each description is a list of terms with high TF-IDF score. This is a common approach from the field of information retrieval to summarize sets of texts, focusing on words that differentiate between them. As an example, the explanation for the cluster in Figure 5b is: *energy, renewable, solar, wind, turbines, sustainable, message* (see the appendix for the list of all descriptions). Then we had three human evaluators tasked to assign a random sample of sets of images (five sets for each cluster, three images per set, in total 160 sets) to the clusters only using the descriptions.

The average Cohen's Kappa between the three pairs of evaluators was .74. We then considered the majority label, that was available for 150 of the 160 samples (in the ten cases with three different labels, we just took the label of the first evaluator). As a result, the average precision and recall over all 32 clusters are, respectively, .83 and .83, with an overall accuracy of .83. As a reference, the expected precision and recall for a cluster assuming equal cluster sizes and random assignments is .03.

The values of precision and recall are not equally distributed across clusters (the detailed values for all clusters are provided in the appendix). Respectively 18 and 19 of the clusters obtain perfect precision and recall. Differences in cluster assignment are in some cases due to two clusters being semantically close. The following are three clear cases:

- One cluster about conferences and one whose description includes terms: *youth, formal*, and *speaking*.
- Two clusters about nature and water, with very similar terms (one having *life, boat*, the other *beauty, river*).

- One cluster about protests and one whose description includes terms: *youth, activism*, and *action*.
- One cluster about *climate change, nature, message* and one with arctic themes.

Merging these pairs of clusters with similar semantics, accuracy is increased to .87.

This evaluation clearly demonstrates how, exploiting the intermediate textual representations, the obtained clusters are not only semantically meaningful but also easily interpretable.

4 Conclusion

In this paper, we assessed the ability of a VLLM-based image clustering pipeline to produce semantically meaningful and interpretable clusters. We did so by defining denotative and connotative similarity functions, through manual labelling of randomly sampled pairs of images, and by generating textual descriptions of the clusters and testing the ability of human evaluators to match selected image samples with the right description, in both cases with the evaluators not being aware of the clusters produced by the algorithms. This procedure also allowed us to use these labels to evaluate and compare a large number of settings for the two pipelines.

We note how the two parts of the pipeline that are different from the CNN-based pipeline have already been used in the computer science literature. [44] recently tested both the impact of clustering text descriptions and interpretability based on TF-IDF keywords. The usage of TF-IDF keywords within text clustering is also a common task, implemented in popular libraries such as BERTopic[2]. However, our research questions focus on new aspects that are fundamental for application to social research. The evaluation in [44] assumes that methods based on deep neural networks can map semantically similar images together, and that a text-based approach can outperform one not based on text. However, while not explicitly discussing this, the experiments in [44] focus on denotative clustering, e.g., showing that a text-based approach is better at identifying and explaining clusters about objects depicted in the images (bridges, churches, towers, etc.). In this paper we differentiate between denotative and connotative semantics, implying that the latter is what is often important in social research, and show how the different approaches target one or the other. We also study the concepts of semantics and interpretability in the context of social research, which is visible in our validation based on online communication images and alignment with human-defined and reliable notions of connotative and denotative similarity (instead of, e.g., benchmark datasets used to train classifiers, such as ImageNet and CIFAR), and also in the design of the interpretability tests.

[2]https://github.com/MaartenGr/BERTopic

The generation of intermediate image descriptions not only improves semantic clustering and interpretability, which is the focus of the experimental part of this paper, but also opens several new directions. First, based on the evidence that a VLLM-based approach can provide a good representation of semantic similarity, one could consider using topic modelling instead of an embeddig+clustering approach, e.g. LDA and its variations [51]. Directly testing topic modelling would not have been methodologically appropriate for this paper, where our objective was to perform a comparison with CNN-based pipelines. To enable this comparison, we only replaced the CNN-based embedding step. In this way, both pipelines would produce the same type of intermediate output (embeddings), which could then be processed without changing the rest of the pipeline. However, the fact that VLLMs enable other pipelines should also be considered an advantage of using this general approach, and something to be tested in future work.

Compared with typical topic models, the partitioning clustering approach used in this paper is limited when the same image would fit multiple clusters (mixed membership). While most of the images used in this paper do fit a single cluster of the ones generated by the VLLM-based method, there are some that would fit two or even more, e.g., an image with statistics about global warming which could fit both the cluster on awareness of global warming and the cluster on data and percentages. We note how these have made the validation of interpretability harder, because images whose algorithmic cluster was correct could also be correctly assigned to a different cluster by the human evaluator, but still counted as a mistake in the validation measure.

Another interesting direction following from this study concerns prompt engineering. In this paper we focus on semantic clustering and interpretability, but one additional potential advantage of a VLLM-based method is flexibility, as the same approach can in principle be used to generate descriptions highlighting different types of visual cues, e.g., emotional and aesthetic. [44] suggests how prompting can also be used to inject external knowledge. This requires a different type of experimental setting. We note that in our experiments we also tested alternative prompts (without prompt engineering) obtaining no conclusive evidence, and thus suggesting a separate and focused experimental study as future work.

When LLMs are used, something that should be checked is the possible bias introduced in the analysis, both with respect to possible mistakes but also with respect to possible preferences assigned to one instead of a different cluster. This requires a specific analysis, but we anecdotally note that the VLLM-based approach produced a cluster about culture, showing people wearing non-western clothes. While this interpretation was aligned with the one of the human evaluators, and thus did not result in lower validity scores, it is contextual and could also well have not been aligned if using other models or evaluators. It is also interesting to observe, again anecdotally, the prevalent arctic themes in a cluster supposedly about nature and climate change.

In our experiments we did not post-process the outliers produced by the clustering method, for example assigning (part of) them to the clusters. This

would have introduced an additional variable, making it more difficult to control the experiments and not contributing to answering our research questions. However, in studies where it is important to cluster everything that can be clustered, for example to compute a more precise measure of prevalence of the different visual themes, such a post-processing should be considered.

As a last important consideration, we should point out how our experiments show relative improvements (with respect to CNN-based approaches) and also good absolute performances, but the results are still not perfect in two regards. First, while semantically more meaningful (assuming that we are often interested in connotative semantics), the produced clusters are still based on some algorithmic choices that decrease the researcher's control over the used semantic similarity. One thing is to see that the clusters are meaningful, another thing is to produce the clusters that the researcher would have liked to produce in the absence of computational constraints. Second, there are still several mistakes in the clusters, for example animals incorrectly identified as polar bears by the VLLM or minor visual elements that gain semantic centrality (e.g., an image of an activist eating candy from a bag representing penguins ended up being clustered together with pictures showing the consequences of climate change in the polar regions). Both issues suggest that, while in this paper we have focused on a specific part of the image clustering pipeline, depending on the research question we expect many applications of this pipeline to be followed by (or intertwined with) interactive steps. Merging clusters, for example based on semantic similarity as emerged in our experiments, or the removal of wrong samples, are examples of these steps, which also require a thorough validation.

Acknowledgments

This research has been funded by Independent Research Fund Denmark (DFF) grant 0257-00007B and Swedish Research Council for Health, Working Life and Welfare (FORTE) grant 2021-01646 as part of the CHANSE (Collaboration of Humanities and Social Sciences in Europe) initiative, Swedish Research Council grant 2021-02769, by the AI4Research initiative at Uppsala University, and by the National Academic Infrastructure for Supercomputing in Sweden (NAISS).

References

[1] Han Zhang and Yilang Peng. Image clustering: An unsupervised approach to categorize visual data in social science research. *Sociological Methods & Research*, 2022.

[2] Nora Webb Williams, Andreu Casas, and John D Wilkerson. *Images as data for social science research: An introduction to convolutional neural nets for image classification*. Cambridge University Press, 2020.

[3] Yilang Peng and Yingdan Lu. Computational visual analysis in political

communication. In *Research Handbook on Visual Politics*, pages 42–54. Edward Elgar Publishing, 2023.

[4] Tama Leaver, Tim Highfield, and Crystal Abidin. *Instagram: Visual social media cultures*. John Wiley & Sons, 2020.

[5] Tim Highfield and Tama Leaver. Instagrammatics and digital methods: Studying visual social media, from selfies and gifs to memes and emoji. *Communication research and practice*, 2(1):47–62, 2016.

[6] Nicholas Carah. Curators of databases: Circulating images, managing attention and making value on social media. *Media International Australia*, 150(1):137–142, 2014.

[7] Wim Bernasco, Evelien M. Hoeben, Dennis Koelma, Lasse Suonperä Liebst, Josephine Thomas, Joska Appelman, Cees GM Snoek, and Marie Rosenkrantz Lindegaard. Promise into practice: Application of computer vision in empirical research on social distancing. *Sociological Methods & Research*, 52(3):1239–1287, 2023.

[8] Jungseock Joo and Zachary C Steinert-Threlkeld. Image as data: Automated content analysis for visual presentations of political actors and events. *Computational Communication Research*, 4(1), 2022.

[9] Fan Zhang, Lun Wu, Di Zhu, and Yu Liu. Social sensing from street-level imagery: A case study in learning spatio-temporal urban mobility patterns. *ISPRS journal of photogrammetry and remote sensing*, 153:48–58, 2019.

[10] Lijia Deng, Qinghua Zhou, Shuihua Wang, Juan Manuel Górriz, and Yudong Zhang. Deep learning in crowd counting: A survey. *CAAI Transactions on Intelligence Technology*, 2023.

[11] Donghyeon Won, Zachary C Steinert-Threlkeld, and Jungseock Joo. Protest activity detection and perceived violence estimation from social media images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 786–794, 2017.

[12] Hung Nguyen, Van Nguyen, Thin Nguyen, Mark E Larsen, Bridianne O'Dea, Duc Thanh Nguyen, Trung Le, Dinh Phung, Svetha Venkatesh, and Helen Christensen. Jointly predicting affective and mental health scores using deep neural networks of visual cues on the web. In *Web Information Systems Engineering–WISE 2018: 19th International Conference, Dubai, United Arab Emirates, November 12-15, 2018, Proceedings, Part II 19*, pages 100–110. Springer, 2018.

[13] Naina Said, Kashif Ahmad, Michael Riegler, Konstantin Pogorelov, Laiq Hassan, Nasir Ahmad, and Nicola Conci. Natural disasters detection in social media and satellite imagery: a survey. *Multimedia Tools and Applications*, 78:31267–31302, 2019.

[14] Conor Lynch, Christian O'Leary, Gary Smith, Rose Bain, Jacqueline Kehoe, Alex Vakaloudis, and Richrd Linger. A review of open-source machine learning algorithms for twitter text sentiment analysis and image classification. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2020.

[15] Ambica Ghai, Pradeep Kumar, and Samrat Gupta. A deep-learning-based image forgery detection framework for controlling the spread of misinformation. *Information Technology & People*, 37(2):966–997, 2024.

[16] Anastasia Giachanou, Guobiao Zhang, and Paolo Rosso. Multimodal multi-image fake news detection. In *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*, pages 647–654. IEEE, 2020.

[17] Amogh Joshi and Cody Buntain. Examining similar and ideologically correlated imagery in online political communication. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 774–786, 2024.

[18] Angelina Mooseder, Cornelia Brantner, Rodrigo Zamith, and Jürgen Pfeffer. (social) media logics and visualizing climate change: 10 years of# climatechange images on twitter. *Social Media + Society*, 9(1), 2023.

[19] Firoj Alam, Muhammad Imran, and Ferda Ofli. Image4act: Online social media image processing for disaster response. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 601–604, 2017.

[20] Matteo Magnani and Alexandra Segerberg. On the conditions for integrating deep learning into the study of visual politics. In *13th ACM Web Science Conference*, 2021.

[21] Michelle Torres and Francisco Cantú. Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data. *Political Analysis*, 30(1):113–131, January 2022.

[22] Michelle Torres. A framework for the unsupervised and semi-supervised analysis of visual frames. *Political Analysis*, 32(2):199–220, 2024.

[23] Ying Yu, Chunping Wang, Qiang Fu, Renke Kou, Fuyu Huang, Boxiong Yang, Tingting Yang, and Mingliang Gao. Techniques and challenges of image segmentation: A review. *Electronics*, 12(5):1199, 2023.

[24] Shaotian Cai, Liping Qiu, Xiaojun Chen, Qin Zhang, and Longteng Chen. Semantic-enhanced image clustering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 6869–6878, 2023.

[25] Xiaofei He, Deng Cai, Haifeng Liu, and Jiawei Han. Image clustering with tensor representation. In *Proceedings of the 13th annual ACM international conference on Multimedia*, pages 132–140, 2005.

[26] Robert Laurenson and Clark F Olson. Adding color information to spatially-enhanced, bag-of-visual-words models. In *Advances in Visual Computing: 16th International Symposium, ISVC 2021, Virtual Event, October 4-6, 2021, Proceedings, Part II*, pages 263–275. Springer, 2021.

[27] Bradley E Wiggins and G Bret Bowers. Memes as genre: A structurational analysis of the memescape. *New media & society*, 17(11):1886–1906, 2015.

[28] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. Understanding visual memes: An empirical analysis of text superimposed on memes shared on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 153–164, 2020.

[29] Mette Mortensen and Christina Neumayer. The playful politics of memes. In *The Playful Politics of Memes*, pages 1–11. Routledge, 2023.

[30] Gholamhosein Sheikholeslami, Wendy Chang, and Aidong Zhang. Semantic clustering and querying on heterogeneous features for visual data. In *Proceedings of the sixth ACM international conference on Multimedia*, pages 3–12, 1998.

[31] Meng Zhu and Atta Badii. Semantic-associative visual content labelling and retrieval: A multimodal approach. *Signal processing: Image communication*, 22(6):569–582, 2007.

[32] Raquel E Patiño-Escarcina and Jose Alfredo Ferreira Costa. The semantic clustering of images and its relation with low level color features. In *2008 IEEE International Conference on Semantic Computing*, pages 74–79. IEEE, 2008.

[33] Achref Ouni, Thierry Chateau, Eric Royer, Marc Chevaldonné, and Michel Dhome. An efficient ir approach based semantic segmentation. *Multimedia Tools and Applications*, 82(7):10145–10163, 2023.

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

[35] C Fellbaum. Wordnet: An on-line lexical database, 1998.

[36] Yi Zhu, Zhaoqing Zhu, Bingqian Lin, Xiaodan Liang, Feng Zhao, and Jianzhuang Liu. Relclip: Adapting language-image pretraining for visual relationship detection via relational contrastive learning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4800–4810, 2022.

[37] Karsten Roth, Jae Myung Kim, A Koepke, Oriol Vinyals, Cordelia Schmid, and Zeynep Akata. Waffling around for performance: Visual classification with random words and broad concepts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15746–15757, 2023.

[38] Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo. Gpt-4v (ision) as a social media analysis engine. *arXiv preprint arXiv:2311.07547*, 2023.

[39] Xinlu Zhang, Yujie Lu, Weizhi Wang, An Yan, Jun Yan, Lianke Qin, Heng Wang, Xifeng Yan, William Yang Wang, and Linda Ruth Petzold. Gpt-4v (ision) as a generalist evaluator for vision-language tasks. *arXiv preprint arXiv:2311.01361*, 2023.

[40] Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256. IEEE, 2023.

[41] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[42] Roman Johannes Gertz, Thomas Dratsch, Alexander Christian Bunck, Simon Lennartz, Andra-Iza Iuga, Martin Gunnar Hellmich, Thorsten Persigehl, Lenhard Pennig, Carsten Herbert Gietzen, Philipp Fervers, et al. Potential of gpt-4 for detecting errors in radiology reports: Implications for reporting accuracy. *Radiology*, 311(1):e232714, 2024.

[43] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[44] Andreas Stephan, Lukas Miklautz, Kevin Sidak, Jan Philip Wahle, Bela Gipp, Claudia Plant, and Benjamin Roth. Text-Guided Image Clustering. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2960–2976, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[45] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022.

[47] Han Zhang and Yilang Peng. Data for: Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research, 2022.

[48] Justin Grimmer and Gary King. General purpose computer-assisted clustering and conceptualization. *Proceedings of the National Academy of Sciences*, 108(7):2643–2650, 2011.

[49] Justin Grimmer, Margaret E. Roberts, and Brandon M. Stewart. Machine Learning for Social Science: An Agnostic Approach. *Annual Review of Political Science*, 24(Volume 24, 2021), May 2021.

[50] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[51] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

A Alternative pipeline settings

We tested a variation of the parameters used in our main pipeline. To embed the text, in addition to Bert we also used text-embedding-3-small, provided by OpenAI, and a TF-IDF vectorization with a hyperparemeter for the number of terms kept for each image (10, 20, and 30). To embed the images directly we used three CNNs: VGG16 (which is the one for which we reported results in the main paper), xception, and resnet50.

We also tested K-means as an alternative clustering algorithms with the same range of parameters and dimensionality reduction strategies. The distribution of quality scores for these alternatives is reported in Figure 6.

The violin plots show the distributions varying the parameters for clustering and dimensionality reduction. For KMeans, we set the number of clusters to 7 (as for the best denotative clustering obtained using HDBSCAN), 15 (as an intermediate value), and 32 (as for the best connotative clustering obtained using HDBSCAN). For dimensionality reduction, as in the main paper, we preserved 5, 10, and 20 components, and the components explaining 80% of the variance, for PCA, and 5 and 10 components for UMAP. For UMAP we also used 5, 15, and 30 nearest neighbors.

B Interpretability

Table 1 shows the top TF-IDF terms for each of the 32 clusters obtained by the pipeline with the best combination of parameters (UMAP: 10 components, 15 neighbors, HDBSCAN: 50 min cluster size), which is the one used to test interpretability.

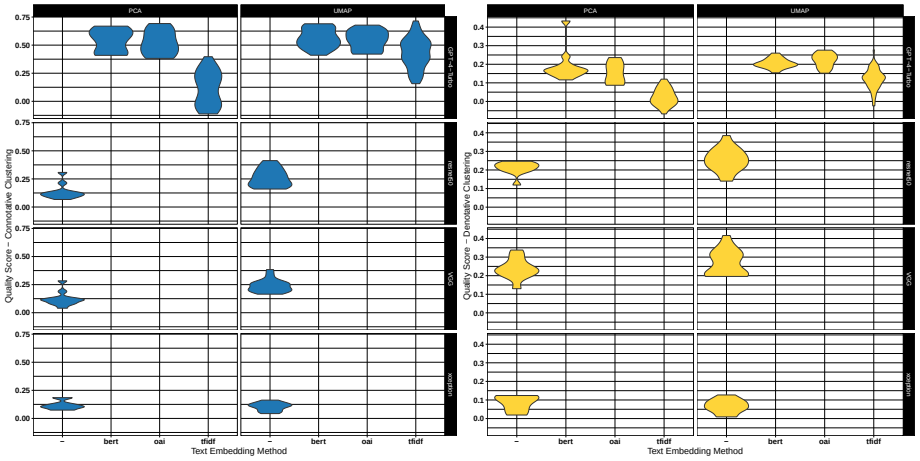
Table 2 shows the cluster assignment by the three human evaluators (consensus) and by the pipeline, including 5 image sets per cluster in the validation data with $\alpha = 3$. Table 3 shows the corresponding precision and recall for each cluster.

CL	Description
0	bear:0.38, polar:0.24, environmental:0.21, ice:0.21, climate:0.19, wildlife:0.19, arctic:0.18, change:0.12, bears:0.11, penguins:0.11
1	environmental:0.34, plastic:0.27, waste:0.19, marine:0.16, natural:0.15, pollution:0.15, human:0.13, message:0.13, impact:0.13, coral:0.11
2	ipcc:0.41, climate:0.29, change:0.23, intergovernmental:0.21, global:0.17, environmental:0.16, panel:0.13, conference:0.12, addressing:0.11, formal:0.11
3	environmental:0.21, human:0.2, natural:0.18, wildfire:0.18, smoke:0.17, wildfires:0.16, disasters:0.15, koala:0.14, destruction:0.14, urgency:0.13
4	energy:0.31, renewable:0.27, solar:0.22, wind:0.2, environmental:0.16, turbines:0.16, sustainable:0.13, message:0.12, sustainability:0.1, clean:0.1
5	coal:0.34, environmental:0.26, pollution:0.19, energy:0.19, industrial:0.13, message:0.11, emitting:0.1, text:0.1, public:0.1, smoke:0.09
6	satellite:0.24, earth:0.24, human:0.23, space:0.2, orbiting:0.2, technological:0.18, exploration:0.17, technology:0.15, advancement:0.13, scientific:0.13
7	amazon:0.34, bolsonaro:0.31, environmental:0.29, rainforest:0.22, action:0.13, jair:0.13, policies:0.12, president:0.12, activism:0.12, political:0.12
8	environmental:0.3, vegan:0.29, dietary:0.21, choices:0.2, animal:0.19, message:0.19, meat:0.17, veganism:0.16, change:0.13, climate:0.13
9	climate:0.26, temperatures:0.25, map:0.21, temperature:0.18, environmental:0.17, global:0.14, average:0.14, change:0.13, areas:0.13, warming:0.11
10	meme:0.3, climate:0.26, change:0.25, environmental:0.21, humor:0.16, humorously:0.13, issues:0.13, character:0.13, uses:0.13, humorous:0.1
11	environmental:0.32, human:0.3, natural:0.27, industrial:0.15, impact:0.13, nature:0.12, stark:0.11, conservation:0.1, deforestation:0.1, degradation:0.09
12	earth:0.38, planet:0.25, space:0.18, home:0.15, universe:0.12, fragility:0.12, responsibility:0.11, awe:0.11, life:0.11, stewardship:0.11
13	climate:0.33, environmental:0.25, change:0.23, action:0.2, message:0.16, urgency:0.14, global:0.14, 24:0.11, planet:0.11, reality:0.11
14	political:0.2, majority:0.19, percentage:0.15, opinion:0.15, data:0.15, party:0.14, republican:0.13, design:0.13, use:0.12, pie:0.12
15	warming:0.36, global:0.27, climate:0.27, change:0.19, environmental:0.16, americans:0.14, awareness:0.13, issue:0.12, concern:0.12, urgency:0.11
16	resilience:0.25, human:0.19, disaster:0.19, natural:0.17, scene:0.15, disasters:0.14, debris:0.13, devastation:0.12, destruction:0.12, adversity:0.12
17	urban:0.32, city:0.29, life:0.18, human:0.12, bustling:0.11, scene:0.11, nature:0.11, buildings:0.1, sky:0.1, cityscape:0.09
18	life:0.21, nature:0.19, pastoral:0.16, rural:0.15, scene:0.14, animals:0.14, cattle:0.14, animal:0.13, cow:0.13, themes:0.13
19	environmental:0.32, climate:0.28, activism:0.26, change:0.2, youth:0.2, action:0.16, young:0.15, message:0.14, urgency:0.14, issues:0.13
20	expression:0.16, woman:0.16, personal:0.15, overall:0.15, young:0.14, moment:0.14, casual:0.14, gaze:0.13, background:0.12, possibly:0.1
21	child:0.31, children:0.21, resilience:0.21, childhood:0.17, life:0.14, innocence:0.13, environment:0.12, young:0.12, scene:0.11, themes:0.11
22	nature:0.22, natural:0.18, scene:0.18, water:0.17, human:0.17, life:0.16, boat:0.16, solitude:0.13, sea:0.13, serene:0.12
23	youth:0.27, young:0.24, engagement:0.18, empowerment:0.17, formal:0.16, involvement:0.14, speaking:0.14, political:0.13, woman:0.13, civic:0.13
24	activism:0.24, protest:0.19, public:0.18, cause:0.16, demonstration:0.14, social:0.13, change:0.13, message:0.13, political:0.12, individuals:0.12
25	labor:0.3, agricultural:0.27, work:0.18, hard:0.16, resilience:0.14, farming:0.14, workers:0.13, rural:0.13, perseverance:0.12, nature:0.11
26	cultural:0.31, traditional:0.2, resilience:0.19, life:0.19, attire:0.15, woman:0.14, heritage:0.14, tradition:0.11, community:0.11, identity:0.1
27	river:0.33, natural:0.27, nature:0.19, beauty:0.17, water:0.17, life:0.17, human:0.15, landscape:0.15, meandering:0.12, winding:0.1
28	friendship:0.21, camaraderie:0.19, smiles:0.19, group:0.18, individuals:0.16, casual:0.14, shared:0.14, possibly:0.14, moment:0.14, event:0.12
29	formal:0.23, conference:0.22, professional:0.19, academic:0.16, engagement:0.15, speaker:0.14, seminar:0.14, intellectual:0.14, event:0.13, likely:0.13
30	nature:0.25, natural:0.18, landscape:0.17, human:0.17, snow:0.16, beauty:0.16, solitude:0.13, resilience:0.13, isolation:0.13, sky:0.13
31	environmental:0.19, climate:0.16, human:0.15, change:0.15, natural:0.13, nature:0.12, message:0.11, action:0.1, scene:0.1, global:0.09

Table 1: Cluster descriptions.

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31
0	5																															5
1		4																														
2			5																													
3				5																												
4					5																											
5						4																										
6							5																					2				
7								5																								
8		1							5																							
9										5																						
10											5																					
11												2																				
12													5	1														1				
13												1		4		1									1							
14						1									5																	
15																4																
16												2					5															
17																		5														
18																			5													
19																				1												
20																					5				1							
21																						5										
22																							3									
23																				1				3					1	1		
24																				3					3							
25																										5						
26																											5					
27																							2					1				
28																													4			
29																								2						4		
30																												1			5	
31																																

Table 2: Confusion matrix for interpretability test. Rows: human assignment. Columns: algorithmic assignment.



(a) Connotative clustering quality

(b) Denotative clustering quality

Figure 6: Kmeans Clustering Quality Score for connotative (Figure 6a) and denotative (Figure 6b) clustering for various models, embedding strategies, and dimensionality reduction methods. We visualize violin plots to report the results for different dimensionality reduction parameters and minimum cluster size.

CL	Pr	Rec	CL	Pr	Rec
0	0.50	1.00	16	0.71	1.00
1	1.00	0.80	17	1.00	1.00
2	1.00	1.00	18	1.00	1.00
3	1.00	1.00	19	1.00	0.20
4	1.00	1.00	20	0.83	1.00
5	1.00	0.80	21	1.00	1.00
6	0.71	1.00	22	1.00	0.60
7	1.00	1.00	23	0.50	0.60
8	0.83	1.00	24	0.50	0.60
9	1.00	1.00	25	1.00	1.00
10	1.00	1.00	26	1.00	1.00
11	1.00	0.40	27	0.33	0.20
12	0.71	1.00	28	1.00	0.80
13	0.57	0.80	29	0.67	0.80
14	0.83	1.00	30	0.83	1.00
15	1.00	0.80	31	0.00	0.00

Table 3: Interpretability by cluster (precision and recall)