

# Text-Guided Image Clustering

Andreas Stephan<sup>1,2,5</sup>, Lukas Miklautz<sup>1,2</sup>, Kevin Sidak<sup>1,2</sup>, Jan Philip Wahle<sup>4</sup>,  
Bela Gipp<sup>4</sup>, Claudia Plant<sup>1</sup> and Benjamin Roth<sup>1,3</sup>

<sup>1</sup> Faculty of Computer Science, University of Vienna, Austria

<sup>2</sup> UniVie Doctoral School Computer Science, University of Vienna, Austria

<sup>3</sup> Faculty of Philological and Cultural Studies, University of Vienna, Austria

<sup>4</sup> Georg-August-Universität Göttingen, Germany

<sup>5</sup>andreas.stephan@univie.ac.at

## Abstract

Image clustering divides a collection of images into meaningful groups, typically interpreted post-hoc via human-given annotations. Those are usually in the form of text, begging the question of using text as an abstraction for image clustering. Current image clustering methods, however, neglect the use of generated textual descriptions. We, therefore, propose *Text-Guided Image Clustering*, i.e., generating text using image captioning and visual question-answering (VQA) models and subsequently clustering the generated text. Further, we introduce a novel approach to inject task- or domain knowledge for clustering by prompting VQA models. Across eight diverse image clustering datasets, our results show that the obtained text representations often outperform image features. Additionally, we propose a counting-based cluster explainability method. Our evaluations show that the derived keyword-based explanations describe clusters better than the respective cluster accuracy suggests. Overall, this research challenges traditional approaches and paves the way for a paradigm shift in image clustering, using generated text<sup>1</sup>.

## 1 Introduction

Psychologists, neuroscientists, and linguists have long studied the dependence of vision and language in humans (Pinker and Bloom, 1990; Nowak et al., 2002; Corballis, 2017). Although the relationship between these modalities is not fully understood, there is a consistent finding: the brain generates a condensed representation to transmit visual information between brain regions (Cavanagh, 2021). A widely discussed type of representation is often referred to as “visual language” or “language of thought” (Fodor, 1975; Jackendoff et al., 1996). Studies based on these concepts suggest that language can be a crucial driver of visual understanding. For example, children remember conjunctions

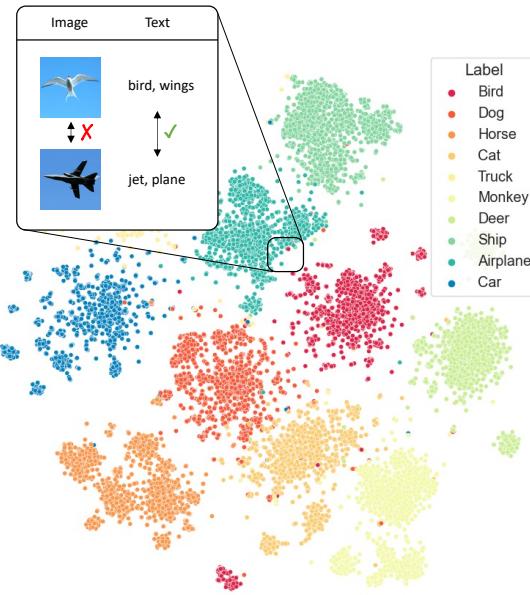


Figure 1: A t-SNE visualization of the BLIP-2 image embeddings for the STL10 dataset. While the images are highly similar (blue background), text such as bird and jet clearly distinguishes objects (and clusters).

of visual features better when accompanied by a textual description (Dessalegn and Landau, 2013), e.g., “the yellow is left of the black”. Given this relationship between visual perception and language comprehension, the question arises whether an abstract textual representation benefits image clustering.

With the significant growth of visual content created online, image clustering has become essential in, e.g., retrieval systems, image segmentation, or medical applications (Mittal et al., 2021; Pandey and Khanna, 2016; Kart et al., 2021). Language offers dense, human-interpretable information, providing multiple benefits when clustering (Figure 1). Emerging multi-modal foundation models and large language models (LLMs), e.g., Blip2 (Li et al., 2023) or GPT-3 (Davidson et al., 2018), allow to derive a “visual language” from images.

<sup>1</sup>Github repo: [https://github.com/AndSt/text\\_guided\\_cl](https://github.com/AndSt/text_guided_cl)

In this paper, we propose *text-guided image clustering*, i.e., deriving a textual representation from images to perform clustering purely based on their text representation. In Figure 2, we outline three approaches to text-guided image clustering. These approaches are structured by the degree of external knowledge introduced into the clustering process.

First, *caption-guided clustering* uses image captioning models to generate brief descriptions of the image content, requiring no external knowledge. In order to inspect the qualities of image and text representations, we compare vision encoder embeddings with TF-IDF (Sparck Jones, 1972) and SentenceBERT (SBERT, Reimers and Gurevych, 2019) representations of the generated text. Our experiments show that on a broad set of eight image clustering datasets, text representations on average outperform the image representations of three state-of-the-art (SOTA) models. Second, *keyword-guided clustering* injects knowledge about the clustering task by prompting visual question-answering (VQA) models to generate keywords, using the assumption that only a few keywords of interest are necessary to describe each image sufficiently. Interestingly, we observe an average performance increase of 5% for TF-IDF-based clusterings. Third, *prompt-guided clustering* introduces domain knowledge in the form of tailored prompts for VQA models. Quantitatively, we observe another performance increase and qualitatively show that clusters related to the question are formed better. Further, we propose to use the generated text for a straightforward counting-based cluster explainability method, generating a keyword-based description for each cluster.

Our contributions can be summarized as follows:

- We propose text-guided image clustering, a novel paradigm leveraging generated text for image clustering.
- We introduce a new way of image clustering by injecting task- and domain knowledge via prompting visual question-answering models.
- We show in our experiments that text-guided image clustering is competitive and often outperforms clustering solely based on images on several datasets.
- We propose a counting-based method to generate a description for each cluster, often exhibiting stronger interpretability than the cluster accuracy suggests.

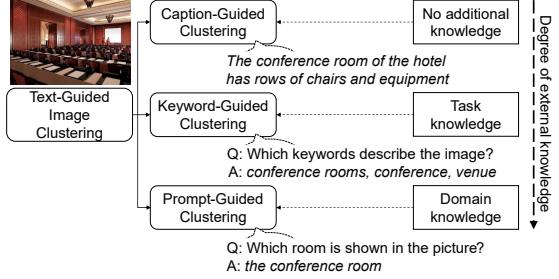


Figure 2: Taxonomy of the text generation processes, structured by the degree of external knowledge. Text is generated BLIP-2 (Li et al., 2023).

## 2 Related Work

We approach image clustering in a novel way by generating more abstract text descriptions using image-to-text models. Therefore, we discuss how our approach relates to earlier work in image clustering (Section 2.1), text clustering (Section 2.2) and give an overview of the enabling technology of image-to-text models in Section 2.3.

### 2.1 Image Clustering

Clustering is the task of grouping similar objects together while keeping dissimilar ones apart. A key problem for unsupervised clustering of images is finding a good similarity measure. Deep learning-based clustering methods approach this problem by learning a representation that maps semantically similar images closer together (Xie et al., 2016; Yang et al., 2017; Niu et al., 2020; Caron et al., 2018; Zhou et al., 2022b). A downside of unsupervised methods is that relying only on image information can suffer from the *blue sky problem* (Häusser et al., 2018). For example, in Figure 1, the blue background pixels make up most of the images. Our approach circumvents this downside by generating a concise textual description of an image. Multi-view clustering methods like (Jin et al., 2015; Chaudhary et al., 2019; Yang et al., 2021; Xu et al., 2022) combine heterogeneous views of data instances into a single clustering. In contrast to our work, they assume the availability of all modalities.

An important problem in clustering is explainability (Fraiman et al., 2011; Moshkovitz et al., 2020), aiming to describe the content of the individual clusters. In general, there are clustering algorithms designed such that the resulting clustering is explainable (Dao et al., 2018), or post-processing methods that explain a given clustering. Existing methods use interpretable features such as

semantic tags (Sambaturu et al., 2020; Davidson et al., 2018), especially when textual explainability is considered. For instance, Zhang and Davidson (2021) uses integer linear programming to assign tags to clusters. Contrary to our approach, these methods assume given textual tags.

## 2.2 Text Clustering

Typically, the text is transformed into a vector representation, and then a clustering algorithm, e.g., K-Means, is applied. Early text representation approaches use counting-based representations such as Bag-of-Words (BoW) or TF-IDF (Sparck Jones, 1972; Zhang et al., 2011). The field moved away from frequency-based approaches as they neglect word order and cannot represent contextualized information, e.g., computer ‘mouse’ vs. the animal ‘mouse’ (Peters et al., 2018). In recent years, the focus in Natural Language Processing (NLP) shifted towards contextualized neural network-based vector encodings, dominated by transformer-based methods (Vaswani et al., 2017). The first breakthrough in transformer-based sentence representations was Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), a siamese network architecture fine-tuning BERT (Devlin et al., 2019) on supervised datasets, e.g. NLI. Following SBERT, text representation techniques are mostly trained using contrastive learning where the choice of positive and negative pairs is unsupervised, e.g., SimCSE (Gao et al., 2021), or weakly-supervised, e.g., E5 (Wang et al., 2022b).

## 2.3 Image-To-Text Models

Image captioning provides textual descriptions for given images. NIC (Vinyals et al., 2015) introduces the now common use of an image encoder and a language decoder. Subsequent models (Radford et al., 2021; Yuan et al., 2021) additionally allow multi-modal inputs, integrating both image and textual information to improve captioning and support tasks like Visual Question Answering (VQA) (Antol et al., 2015). Wang et al. (2022a) use only one image encoder and one text decoder, and perform image /video captioning and VQA in one simplified architecture. Flamingo (Alayrac et al., 2022) allows interleaving images and text by introducing Perceiver Resamplers on top of pre-trained image and language models. BLIP-2 (Li et al., 2023) is a state-of-the-art model that takes fixed pre-trained language and image models and only fine-tunes a so-called Query-Transformer, which only uses

a few trainable parameters. This is useful in our experiments because the underlying models are not trained on multimodal data, ensuring a fair comparison of the respective representations.

## 3 Methodology

We describe the formal setup, the experimental setup, and the chosen datasets.

### 3.1 Problem Definition

Let  $\mathbf{X} = \mathbf{x}_1, \dots, \mathbf{x}_n \subset \mathcal{X}$  denote the set of images in our dataset. The goal of image clustering is to obtain a clustering  $h : \mathcal{X} \rightarrow \mathcal{Y}$  that assigns images to their respective clusters. We propose to employ image-to-text models which typically consist of an image encoder  $f : \mathcal{X} \rightarrow \mathcal{Z}$ , embedding images into a latent space  $\mathcal{Z} \subset \mathbb{R}^d$ , and a text decoder, i.e. a LLM,  $g : \mathcal{Z} \rightarrow \mathcal{T}$ , where  $\mathcal{T}$  is some text space. The text is subsequently embedded  $t : \mathcal{T} \rightarrow \mathcal{V} \subset \mathbb{R}^l$  and clustered, e.g., with K-Means.

### 3.2 Experimental Setup

The central goal of this paper is to compare representations based on images and generated text for the task of image clustering. The following describes the choices and evaluation criteria common to all experiments.

**Clustering.** To shed light on the question of whether text is a (more) suitable representation for image clustering, we compare the performance of a clustering on the image space  $\mathbf{Z} = f(\mathbf{X})$  and of a clustering on the vectorization of the generated text  $\mathbf{T} = t(g(\mathbf{Z}))$ . Following the deep clustering (Xie et al., 2016; Yang et al., 2017) and self-supervised learning (Zhou et al., 2022a) literature, we use K-Means to evaluate the suitability of the respective image and text embeddings for clustering. We run K-Means 50 times in all experiments and report the mean outcome to get robust results. Whenever we need a single run, e.g., for qualitative analysis, the run with the lowest K-Means loss, also called inertia, is used.

**Vectorization.** In order to employ clustering algorithms, images and texts need to be represented as vectors. For image vectorization, we use the latent space of an image encoder. We experiment with multiple models introduced in Section 4.1. For text vectorization, one frequency-based and one neural algorithm are considered. TF-IDF (Sparck Jones, 1972) is a standard counting-based representation. Using the scikit-learn (Pedregosa et al., 2011) im-

plementation, English stop-words are removed, and a maximum vocabulary of 2000 words is set. No additional preprocessing is performed. Since nowadays transformer-based text representations are the standard, we experiment with SBERT<sup>2</sup> (Reimers and Gurevych, 2019) as it was the first BERT-based sentence representation. Note that larger, newer, and better transformer-based models are available. We deliberately choose a widely used, competitive, small model as this strengthens our claim that clusterings based on generated text often outperform clusterings based on image representations.

**Metrics.** To measure clustering performance, the Normalized Mutual Information (NMI) (Vinh et al., 2010) and the Cluster Accuracy (Acc) (Yang et al., 2010) are computed. Both metrics take values between 0 and 1, where higher numbers indicate a better match with the ground truth labels. For the sake of readability, we multiply them by 100.

### 3.3 Datasets

We consider a diverse collection of datasets, separated into three groups according to various challenges. Partially, there is an overlap between the properties of the datasets. Nevertheless, our selection of datasets is motivated by this grouping. Note that this is a more diverse set of datasets as typically used (Cai et al., 2022; Qian, 2023). An overview of the dataset statistics and samples of each dataset are depicted in Tables 6 and 7 in the Appendix, respectively.

**Standard Datasets.** We utilize three widely-used image clustering benchmarking datasets: STL10 (Coates et al., 2011), Cifar10 (Krizhevsky and Hinton, 2009) and ImageNet10 (Deng et al., 2009).

**Background Datasets.** To assess the robustness of our proposed method against background noise, we include Sports10 (Trivedi et al., 2021) and iNaturalist2021 (Grant Van Horn, 2021), two datasets containing high-resolution images of sports scenes in video games and natural environments.

**Human Interpretable Datasets.** Three datasets focusing on human concepts rather than individual objects are included. LSUN (Yu et al., 2015), showing, e.g., a living room or a kitchen, Human Activity Recognition (HAR) (Nagadia, 2022), containing scenes such as running and Facial Expression Recognition (FER2013) (Barsoum et al., 2016), e.g., surprise, are considered.

<sup>2</sup><https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

## 4 Text-Guided Image Clustering

We explore the potential of generated text for image clustering. First, we use standard image captioning and observe that the text representation outperforms the image representation of several models. Second, we guide the text generation using VQA models to generate keywords, which we call *keyword-guided clustering*, and introduce *prompt-guided clustering*, where we use domain-specific prompts to elicit relevant properties. Third, we use the generated text for cluster explainability, obtaining keyword-based descriptions for each cluster.

### 4.1 Caption-Guided Image Clustering

Modern foundation models provide the possibility to work with multiple modalities. In particular, image captioning models describe images with text. Thus, as a first experiment, we investigate how well text clustering on captioned text works in comparison to image clustering, and establish a consistent experimental setup.

**Setup.** The commonality between current image captioning models is that they consist of an image encoder and a generative LLM to generate text conditioned on the latent image space. As described in Section 3.2 we assess the quality of image and generated text by comparing the clustering performance of the vision encoder embeddings with TF-IDF and SBERT representations using K-Means. We benchmark three SOTA image-to-text models, namely a community-trained version of Flamingo<sup>3</sup> (Alayrac et al., 2022), GIT<sup>4</sup> (Wang et al., 2022a), and BLIP-2<sup>5</sup> (Li et al., 2023), all available within the Huggingface Transformers library (Wolf et al., 2020). Note that we abstain from including dedicated clustering methods (Cai et al., 2022; Qian, 2023; Gao et al., 2021) because they are based on a much weaker image encoder, thus achieving much lower performance. Furthermore, it is not straightforward to train transformer-based image models using clustering objectives. We probabilistically sample a maximum of 80 tokens, without any additional parameters. Only for Flamingo, we set top-K to 8, following the original repository. Experiments were performed on a single A100 40GB and took about 40h hours.

<sup>3</sup><https://huggingface.co/dhansmair/flamingo-mini>

<sup>4</sup><https://huggingface.co/microsoft/git-large>

<sup>5</sup><https://huggingface.co/Salesforce/blip2-flan-t5-xl>

Model	Representation	Standard				Background				Human				Avg					
		STL10		Cifar10		ImageNet10		Sports10		iNaturalist2021		FER2013		LSUN					
		Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Avg			
Flamingo	Image	95.0	<b>95.13</b>	84.0	84.19	<b>99.38</b>	<b>98.85</b>	<b>75.87</b>	<b>81.61</b>	40.8	58.09	<b>36.79</b>	<b>17.33</b>	60.67	60.98	50.07	43.67	67.82	<b>67.48</b>
	TF-IDF	82.22	77.0	81.85	76.23	94.32	89.57	54.16	49.86	34.27	43.63	25.77	2.91	<b>70.58</b>	64.04	40.92	35.52	60.51	54.85
	SBERT	<b>97.74</b>	94.68	<b>93.64</b>	<b>86.15</b>	98.36	96.05	60.32	55.89	<b>44.93</b>	<b>58.99</b>	29.79	9.77	68.96	<b>68.41</b>	<b>51.37</b>	<b>46.84</b>	<b>68.14</b>	64.6
GIT	Image	51.15	63.62	66.37	64.87	95.41	<b>93.78</b>	71.17	75.69	42.47	53.0	24.1	<b>2.15</b>	52.06	51.78	38.81	33.18	55.19	54.76
	TF-IDF	79.92	74.71	74.0	66.73	82.69	76.78	<b>87.42</b>	84.6	36.12	42.84	25.24	1.66	65.34	57.68	42.87	36.05	61.7	55.13
	SBERT	<b>96.58</b>	<b>93.34</b>	<b>86.79</b>	<b>76.97</b>	<b>96.37</b>	92.72	85.73	<b>88.14</b>	<b>46.04</b>	<b>58.78</b>	<b>26.61</b>	1.95	<b>69.82</b>	<b>61.95</b>	<b>48.11</b>	<b>42.66</b>	<b>69.51</b>	<b>64.56</b>
BLIP-2 (*)	Image	<b>99.65</b>	<b>99.16</b>	<b>98.69</b>	<b>97.59</b>	<b>99.8</b>	<b>99.35</b>	91.31	93.22	44.97	<b>62.7</b>	35.97	<b>21.2</b>	62.07	64.47	<b>52.65</b>	<b>47.06</b>	73.14	73.09
	TF-IDF	83.3	79.35	89.0	84.75	93.54	88.81	<b>99.38</b>	<b>98.65</b>	34.17	39.07	31.86	6.89	76.69	71.05	50.51	46.09	69.81	64.33
	SBERT	98.03	96.27	97.31	94.07	98.22	96.63	99.07	98.47	<b>47.43</b>	61.63	<b>38.21</b>	20.53	<b>81.11</b>	<b>74.37</b>	50.85	46.68	<b>76.28</b>	<b>73.58</b>

Table 1: Comparison of Clustering Accuracy and NMI of image space and generated captions, using TF-IDF and SBERT representations, of multiple Image-to-Text models. For each combination of dataset and metric, underlined numbers represent the best overall performance, and bold numbers the best performance per model. (\*) Note that BLIP-2 is pre-trained on ImageNet21K (Deng et al., 2009), which STL10 and ImageNet10 are subsets of.

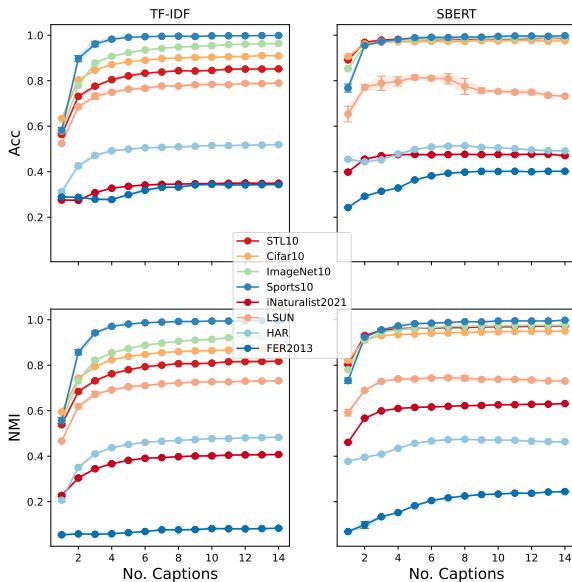


Figure 3: Effect of the number of captions sampled per image for BLIP-2. The number of captions is depicted on the X-axis, mean and standard deviation of clustering performance are on the Y-axis.

We start by studying the effect of the number of captions generated per image. For each amount of captions, we sample 6 versions and report the mean and standard error in Figure 3.

**Results.** We observe that, for TF-IDF, with a growing number of captions, the performance increases monotonically, whereas SBERT saturates for many datasets. Being counting-based, we think that the reason is that TF-IDF is better at reducing the effect of outlier captions, i.e. single bad captions. For all following experiments, we choose to sample 6 text generations as a trade-off between sampling efficiency and clustering performance.

The full image captioning results are shown in Table 1. The average scores (Avg) show that SBERT outperforms the other two representations

across all model types on almost all datasets, while the TF-IDF representation performs worst. Note that we abstain from sophisticated preprocessing such as lemmatization or stemming, which is common for frequency-based representations such as TF-IDF, to keep the setup simple and depend on text information as purely as possible. This might (to a certain degree) explain the worse performance.

Further, we observe that BLIP-2 is the best-performing model. It performs especially well on the standard datasets, which we think is due to the fact that it was pre-trained on ImageNet21k in a self-supervised fashion.

In summary, the results show that text representations, obtained only based on (latent) image representations, provide competitive clustering performance, often outperforming the corresponding image representation.

## 4.2 Knowledge Injection

Now we investigate the potential of guiding the text generation so that it is specifically suited for clustering. By using modern VQA models, it is possible to elicit dedicated information from images. In the following, we introduce two ways to make use of VQA models.

**Keyword-Guided Clustering.** Given that it is common to (verbally) describe clusters using keywords, we hypothesize that it is beneficial to prompt the model to generate keywords. The reasons are: 1) keywords provide useful inputs for simpler, traditional count-based representations such as TF-IDF, 2) keywords are useful for count-based analysis methods, such as the proposed cluster explainability algorithm in section 4.3, and 3) ground truth cluster labels (as given by classification datasets used in the clustering literature) are typically described using only a few keywords.

		Sports10		iNaturalist2021		FER2013		LSUN		HAR		Avg	
		Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI	Acc	NMI
Image	ViT	91.31	93.22	44.97	<b>62.70</b>	35.97	21.2	62.07	64.47	52.65	47.06	57.39	57.73
Caption-Guided	TF-IDF	<b>99.38</b>	<b>98.65</b>	34.17	39.07	31.86	6.89	<b>76.69</b>	<b>71.05</b>	50.51	46.09	58.52	52.35
	SBERT	99.07	<u>98.47</u>	47.43	61.63	38.21	20.53	<b>81.11</b>	<b>74.37</b>	50.85	46.68	63.33	60.34
Keyword-Guided	TF-IDF	<u>99.08</u>	97.82	42.13	48.25	<b>47.05</b>	27.34	76.2	69.28	51.35	45.47	63.16	57.63
	SBERT	96.89	96.87	<u>48.44</u>	59.48	46.44	29.96	70.63	70.82	<u>55.66</u>	<u>50.07</u>	<u>63.61</u>	<u>61.44</u>
Prompt-Guided	TF-IDF	84.83	94.46	38.01	47.61	<b>46.86</b>	<u>34.25</u>	66.4	59.92	52.74	47.96	57.77	56.84
	SBERT	98.70	98.12	<b>48.57</b>	<u>62.23</u>	45.60	<b>36.04</b>	71.59	63.54	<b>60.93</b>	<b>52.94</b>	<b>65.08</b>	<b>62.57</b>

Table 2: Comparison of clustering performance of the BLIP-2 image encoder features, and examined types of generated text. For prompt-guided clustering, the clusterings belonging to the prompt with the lowest K-Means are evaluated. For each dataset and metric combination, the best performance is bold, and the second-best performance is underlined.

**Prompt-Guided Clustering.** In real-world scenarios, often, some domain knowledge about the given data is available. The ability of VQA models to retrieve dedicated information from images opens up the possibility of using domain knowledge in the natural form of text. An example is to ask "Which activity is performed in the picture?". Note, crucially, that this is not possible using standard image clustering models.

**Setup.** Due to resource constraints, we only use the best-performing (cf. Table 1) image-to-text model, BLIP-2, for the subsequent experiments. Based on the results depicted in Figure 3, we sample  $k = 6$  texts for each image.

For keyword-guided clustering, we use the question "Which keywords describe the image?". To perform prompt-guided clustering, we create four questions for each of the datasets. The questions were created by naively transforming the name of the dataset into a question, e.g. for human action recognition "Which activity is performed?" is asked. Note, that no additional prompt engineering efforts were made, as we are not aware of a more principled way to design such prompts. Find all questions in Table 8 in Appendix B.

BLIP-2 solves the “standard” datasets with almost 100% and they exhibit only a collection of objects, making it difficult to pose interesting questions other than ‘What objects are described?’ . Thus, they are excluded in the following experiments. It is well known that current LLMs possibly generate very different texts, even though the prompt has the same meaning (Elazar et al., 2021). Therefore, in Table 2 we use an unsupervised heuristic to decide which prompt works best by taking the prompt belonging to the clustering with the lowest K-Means loss.

Modality / Question	SBERT	
	Acc	NMI
Image	52.65	47.06
Which keywords describe the image?	55.66	50.07
What type of motion is depicted in the picture?	49.20	42.54
Which activity is shown in the picture?	56.03	49.69
Which action is shown in the picture?	58.68	52.86
What is the person doing in the picture?	<b>60.93</b>	<b>52.94</b>

Table 3: A case study for prompt-guided image clustering on Human Action Recognition, using the SBERT representation. Find the full table in Appendix B.

**Results.** In Table 2 we observe that the average performance (Avg) for caption-guided image clustering and SBERT-based keyword-guided clustering is similar. Using keywords, TF-IDF improves on average by 5% for both cluster accuracy and NMI, closing the gap to SBERT. This result is in line with our hypothesis that keywords are a useful representation for image clustering.

As a case study, Table 3 holds the results for the HAR dataset. We observe a notable variance in the performance of multiple prompts. This is a common phenomenon for prompting-based methods (Zhao et al., 2021). Using the K-Means loss as a proxy for selecting the best prompt leads to the best average performance in Table 2.

Interestingly, the confusion matrices in Figure 4 show different assignment patterns depending on the question posed to the VQA model. For instance, when posing the question ‘What room is shown in the picture?’, all room clusters are formed well, but the others, e.g. bridge or tower, are worse. We argue that this variation is not an issue but a feature of prompt-guided image clustering, e.g., during exploratory data analysis, where one might want to investigate different aspects of a dataset.

In summary, we demonstrate that it is possible

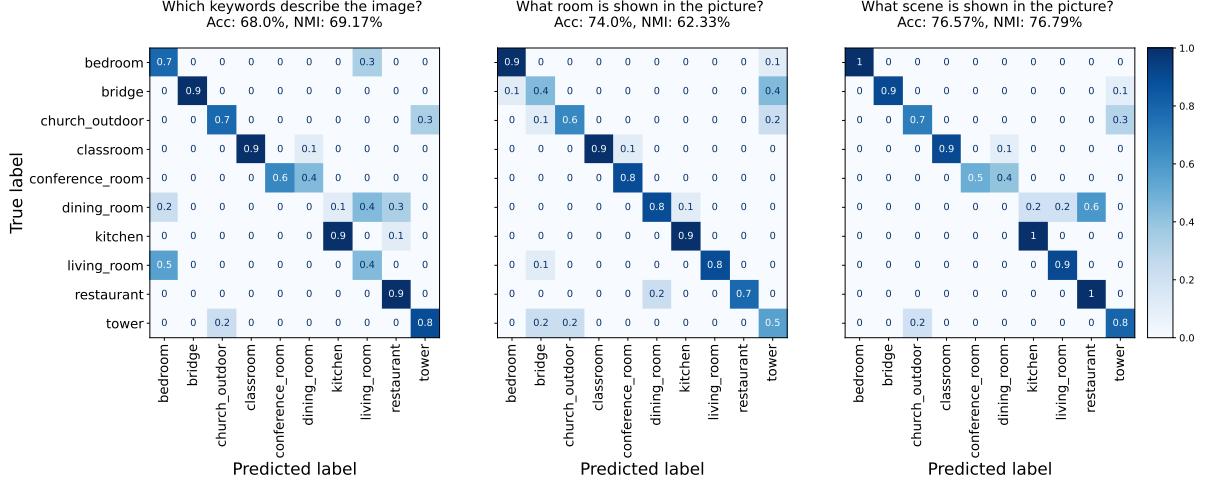


Figure 4: Confusion matrices based on three different clustering results from text generated with three different VQA prompts. While a similar cluster accuracy is achieved, we observe that the clustering relates to the prompt. In the middle all room clusters are clustered well, on the right side the clustering is not able to distinguish well between dining room, kitchen and restaurant (see corresponding dining room row), but leads to better overall accuracy.

to improve clustering performance by injecting domain knowledge in the form of text and that the clustering changes according to the posed questions. Further examples of the impact of different prompts on the embedded space and clustering are shown in t-SNE embeddings in Figures 6 and 7 in the Appendix.

### 4.3 Cluster Explainability

So far, we use the generated text solely to form clusters. But given the (built-in) interpretability of text, a natural extension is to use text as an explanation of the formed clusters. Explainability for image clustering is an important issue, as it provides insights into how the clustering algorithm groups the images, helping users understand the underlying patterns and relationships. The availability of textual descriptions for each cluster sample allows us to extrapolate to textual descriptions of each cluster as a whole. Note that this is not possible using models considering only images.

We hypothesize that a concise way to describe a cluster is to use a small set of keywords. This is based on the fact that the considered datasets use keyword-based labels. Thus, we introduce the following algorithm to obtain keywords for each cluster from the generated text.

**Explainability Algorithm.** For each predicted cluster, the keywords are sorted by their number of occurrences in the generated texts. The algorithm returns the most frequent keywords per cluster. If a keyword occurs in multiple cluster descriptions,

it is not considered, and the next most occurring is chosen. We take the two most occurring keywords based on an initial screening of the LSUN dataset. Find the Pseudocode in the Appendix C.

**Setup.** We provide a quantitative analysis of the generated descriptions by applying two metrics. First, we introduce the subset exact match (SEM) metric, for which we lowercase each string and check whether the ground truth cluster name appears in the predicted keywords. No further standardization, such as stemming or lemmatization, is performed. Second, SBERT embeddings are used to check the similarity between cluster names and keywords obtained by the explainability algorithm. According to our initial investigation, we use a cosine similarity of 0.4 as the threshold to indicate a match between ground truth and explanation. For each dataset, we provide the cluster accuracy and the explainability performance given the ground truth (*Truth*) clustering and the predicted (*Pred*) clustering, corresponding to the cluster accuracy. Out of the 50 conducted K-Means runs, we use the clustering with the lowest K-Means loss for the analysis.

**Results.** Table 5 depicts the quantitative evaluation of our algorithm. We observe that the SBERT metric is always equal to or higher than the SEM metric, which makes sense as SEM is a rather strict metric, not understanding synonyms or syntactical changes, e.g., "TableTennis" vs. "table tennis". Interestingly, in most cases, the SBERT metric is higher than the clustering accuracy. Table 4 shows

Ground Truth	Explanation	SEM	SBERT Sim.
Sports10			
AmericanFootball	football, nfl	0	1
Basketball	basketball, basketball game	1	1
BikeRacing	motorcycle, rider	0	1
CarRacing	car, speed	0	0
Fighting	fight, boxing	0	1
Hockey	hockey, hockey game	1	1
Soccer	soccer, soccer game	1	1
TableTennis	ping pong, table tennis	0	0
Tennis	tennis, tennis game	1	1
Volleyball	volleyball, beach	1	1
LSUN			
bedroom	bedroom, bed	1	1
bridge	bridge, river	1	1
church_outdoor	church, cathedral	0	1
classroom	classroom, teacher	1	1
conference_room	meeting, conference	0	1
dining_room	dining room, dining table	1	1
kitchen	kitchen, wood	1	1
living_room	living room, living	1	1
restaurant	restaurant, bar	1	1
tower	tower, city	1	1

Table 4: Examples of generated explanations for Sports10 and LSUN. If a value in the SEM or SBERT Sim. column is 1, it means that the metric says ground truth and explanation match.

an example of generated descriptions and metrics. We observe that both metrics cannot understand that “TableTennis” and “ping pong, table tennis” have the same meaning, but still, all cluster descriptions of Sports10 are correct. For iNaturalist2021 and FER2013, we observe that the generated text is often of bad quality, resulting in low-quality descriptions. We conclude that the generated descriptions provide a good overview of the content of the generated clusters and in most cases, describe the dataset better than clustering accuracy suggests.

## 5 Broader Impact

We believe there is a lot of unused potential for text as an abstraction in image clustering.

**Text as a proxy for “meaningful” clustering.** Clustering research aims to find meaningful clusters. In general, it is an open question to define what meaningful exactly stands for, some researchers even call it an ill-posed problem. We argue that text is a good proxy to express meaningfulness as it is based on the natural human form of communication. This is a novel viewpoint on the task of image clustering aligning with research methodologies in the clustering community, where clustering methods are commonly benchmarked with datasets that have human-annotated textual labels as ground truth. Our research contributes to the discussion about meaningful clustering by showing

	Cluster Acc		SEM		SBERT Sim.	
	TF-IDF	SBERT	Truth	Pred	Truth	Pred
STL10	87	98	100	<b>100</b>	100	<b>100</b>
ImageNet10	94	99	30	30	100	<b>100</b>
CIFAR10	91	97	90	90	100	<b>100</b>
Sports10	99	98	50	50	80	80
iNaturalist2021	40	48	0	0	91	45
LSUN	75	68	70	<b>80</b>	100	<b>100</b>
HAR	51	56	20	13	87	<b>87</b>
FER2013	46	46	12	12	38	25

Table 5: Evaluation of our explainability method. In “Truth”, the explainability method is applied to the ground truth clustering whereas in “Pred” it is applied to the clustering of the given clustering accuracy. Numbers are boldened if the explainability score of a found clustering (“Pred” columns) outperforms clustering accuracies.

that generated text improves the interpretability of the detected clusters.

**Knowledge Injection.** Furthermore, it can be highly subjective what determines a meaningful clustering. For a given dataset, different people are interested in different types of information. For example, in real-world scenarios, an expert might have several questions about a dataset based on their domain knowledge. We show that these questions can be used to guide the clustering process by prompting VQA models. Given the current speed of research, we believe that the increasing ability to use more detailed prompts will drastically improve our knowledge injection method. This, in turn, will open up new research avenues for injecting knowledge into the clustering process.

## 6 Conclusion

In this work, we introduce *Text-Guided Image Clustering*, using image-captioning and VQA models to automatically generate text, and subsequently cluster only the generated text. After applying multiple captioning models on eight diverse datasets, our experiments show that representations of generated text outperform image representations on many datasets. Further, we use text to include task- and domain knowledge by prompting VQA models, resulting in additional improvements in clustering performance. We find that it is possible to shape the clustering favorably according to the information given by a specific prompt. Additionally, we use the generated text to obtain a keyword-based description for each cluster and show their usefulness quantitatively and qualitatively.

While it is difficult to identify background noise or

irrelevant features in the pixel space, text is discrete and interpretable. We show that text-guided image clustering often outperforms clustering purely on image information, and provides interpretability. Therefore, our research provides insights into the role of text in determining meaningful clusterings.

## 7 Limitations

While our proposed approach shows promising results, several limitations apply.

Text-guided image clustering is dependent on the quality and effectiveness of the generated text. In cases where the generated text is incomplete, misleading, or fails to capture the essential features of the images, the clustering algorithm may struggle to accurately group similar images. Current image-to-text models are mostly trained on data obtained from the internet. For example, because of licensing and other restrictions, many domain-specific images are not represented appropriately in the training data, resulting in poor text generation abilities for those domains. Nevertheless, our experiments are performed on a wide variety of datasets, more diverse than in common image clustering research, proving the general applicability of the method.

While we show that our approach is effective for image clustering, we do not include results for other visual modalities, such as video or 3D point clouds. We show that it is worthwhile to investigate the possibility of clustering images using generated text and generating textual cluster explanations. The rapid advancement of machine learning models will also enable the same approach for other modalities.

The approach of prompt-guided image clustering is based on the assumption that domain knowledge is readily accessible, allowing the generation of specific questions to guide VQA models. While we show that leveraging domain knowledge can prove advantageous, clustering methods are frequently employed for exploratory data analysis. Introducing domain knowledge may limit the discovery of novel insights or alternative interpretations due to biased prompts.

## Acknowledgements

We thank the anonymous reviewers for their constructive feedback. This research has been funded by the Vienna Science and Technology Fund (WWTF)[10.47379/VRG19008]

“Knowledge-infused Deep Learning for Natural Language Processing”. We thank the European High Performance Computing initiative for providing the computational resources that enabled this work. EHPC-DEV-2022D10-051, EHPC-DEV-2023D11-017.

## References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. [Flamingo: a visual language model for few-shot learning](#). In *Advances in Neural Information Processing Systems*.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Emad Barsoum, Cha Zhang, Cristian Canton Ferrer, and Zhengyou Zhang. 2016. Training deep networks for facial expression recognition with crowd-sourced label distribution. In *ACM International Conference on Multimodal Interaction (ICMI)*.
- Jinyu Cai, Jicong Fan, Wenzhong Guo, Shiping Wang, Yunhe Zhang, and Zhao Zhang. 2022. [Efficient deep embedded subspace clustering](#). In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21–30.
- Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. 2018. Deep clustering for unsupervised learning of visual features. In *ECCV(14)*, volume 11218 of *Lecture Notes in Computer Science*, pages 139–156. Springer.
- Patrick Cavanagh. 2021. [The language of vision. Perception](#), 50(3):195–215.
- Chandramani Chaudhary, Poonam Goyal, Siddhant Tuli, Shuchita Banthia, Navneet Goyal, and Yiping Phoebe Chen. 2019. A novel multimodal clustering framework for images with diverse associated text. *Multimedia Tools and Applications*, 78:17623–17652.
- Adam Coates, Andrew Ng, and Honglak Lee. 2011. [An analysis of single-layer networks in unsupervised feature learning](#). In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 215–223, Fort Lauderdale, FL, USA. PMLR.

- Michael C Corballis. 2017. Language evolution: a changing perspective. *Trends in cognitive sciences*, 21(4):229–236.
- Thi-Bich-Hanh Dao, Chia-Tung Kuo, S. S. Ravi, Christel Vrain, and Ian Davidson. 2018. Descriptive clustering: Ilp and cp formulations with applications. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1263–1269. International Joint Conferences on Artificial Intelligence Organization.
- Ian Davidson, Antoine Gourru, and S Ravi. 2018. The cluster description problem - complexity results, formulations and approximations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.
- J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.
- Banchiamlack Dessalegn and Barbara Landau. 2013. Interaction between language and vision: It’s momentary, abstract, and it develops. *Cognition*, 127(3):331–344.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhi Isha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Jerry A Fodor. 1975. *The language of thought*, volume 5. Harvard university press.
- Ricardo Fraiman, Badih Ghattas, and Marcela Svarc. 2011. Interpretable clustering using unsupervised binary trees. *Advances in Data Analysis and Classification*, 7:125–145.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- macaodha Grant Van Horn. 2021. inat challenge 2021 - fgvc8.
- Philip Häusser, Johannes Plapp, Vladimir Golkov, Elie Aljalbout, and Daniel Cremers. 2018. Associative deep clustering: Training a classification network with no labels. In *GCPR*, volume 11269 of *Lecture Notes in Computer Science*, pages 18–32. Springer.
- Ray Jackendoff, Paul Bloom, Mary A Peterson, Lynn Nadel, and Merrill F Garrett. 1996. Language and space. chapter “The Architecture of the Linguistic-Spatial Interface”, pages 1–30.
- Cheng Jin, Wenhui Mao, Ruiqi Zhang, Yuejie Zhang, and Xiangyang Xue. 2015. Cross-modal image clustering via canonical correlation analysis. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 151–159. AAAI Press.
- Turkay Kart, Wenjia Bai, Ben Glocker, and Daniel Rueckert. 2021. Deepmcat: Large-scale deep clustering for medical image categorization. In *Deep Generative Models, and Data Augmentation, Labelling, and Imperfections*, pages 259–267, Cham. Springer International Publishing.
- Alex Krizhevsky and Geoffrey Hinton. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.
- Himanshu Mittal, Avinash Pandey, Mukesh Saraswat, Sumit Kumar, Raju Pal, and Garv Modwel. 2021. A comprehensive survey of image segmentation: clustering methods, performance parameters, and benchmark datasets. *Multimedia Tools and Applications*, 81.
- Michal Moshkovitz, Sanjoy Dasgupta, Cyrus Rashtchian, and Nave Frost. 2020. Explainable k-means and k-medians clustering. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7055–7065. PMLR.
- Meet Nagadia. 2022. Human action recognition (har) dataset.
- Chuang Niu, Jun Zhang, Ge Wang, and Jimin Liang. 2020. Gatcluster: Self-supervised gaussian-attention network for image clustering. In *ECCV (25)*, volume 12370 of *Lecture Notes in Computer Science*, pages 735–751. Springer.
- Martin A Nowak, Natalia L Komarova, and Partha Niyogi. 2002. Computational and evolutionary aspects of language. *Nature*, 417(6889):611–617.
- Shreelekha Pandey and Pritee Khanna. 2016. Content-based image retrieval embedded with agglomerative clustering built on information loss. *Computers & Electrical Engineering*, 54:506–521.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos,

- D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.
- Steven Pinker and Paul Bloom. 1990. Natural language and natural selection. *Behavioral and Brain Sciences*, 13(4):707–727.
- Qi Qian. 2023. Stable cluster discrimination for deep clustering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16645–16654.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.
- Prathyush Sambaturu, Aparna Gupta, Ian Davidson, S. S. Ravi, Anil Vullikanti, and Andrew Warren. 2020. Efficient algorithms for generating provably near-optimal cluster descriptors for explainability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1636–1643.
- Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.
- Chintan Trivedi, Antonios Liapis, and Georgios N Yannakakis. 2021. Contrastive learning of generalized game representations. In *2021 IEEE Conference on Games (CoG)*. IEEE.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Nguyen Xuan Vinh, Julien Epps, and James Bailey. 2010. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(95):2837–2854.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition*.
- Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022a. GIT: A generative image-to-text transformer for vision and language. *Transactions on Machine Learning Research*.
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Dixin Jiang, Rangan Majumder, and Furu Wei. 2022b. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.
- Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierrette Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- Junyuan Xie, Ross B. Girshick, and Ali Farhadi. 2016. Unsupervised deep embedding for clustering analysis. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 478–487. JMLR.org.
- Jie Xu, Huayi Tang, Yazhou Ren, Liang Peng, Xiaofeng Zhu, and Lifang He. 2022. Multi-level feature learning for contrastive multi-view clustering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16030–16039.
- Bo Yang, Xiao Fu, Nicholas D. Sidiropoulos, and Mingyi Hong. 2017. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3861–3870. PMLR.
- Sean T Yang, Kuan-Hao Huang, and Bill Howe. 2021. Jecl: Joint embedding and cluster learning for image-text pairs. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 8344–8351. IEEE.
- Yi Yang, Dong Xu, Feiping Nie, Shuicheng Yan, and Yueteng Zhuang. 2010. Image clustering using local discriminant models and global integration. *IEEE Trans. Image Process.*, 19(10):2761–2773.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel C. F. Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luwei Zhou, and Pengchuan Zhang. 2021. Florence: A new foundation model for computer vision. *ArXiv*, abs/2111.11432.

Hongjing Zhang and Ian Davidson. 2021. Deep descriptive clustering. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 3342–3348. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Wen Zhang, Taketoshi Yoshida, and Xijin Tang. 2011. A comparative study of tf\* idf, lsi and multi-words for text classification. *Expert Systems with Applications*, 38(3):2758–2765.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Ci-hang Xie, Alan L. Yuille, and Tao Kong. 2022a. Image BERT pre-training with online tokenizer. In *ICLR*. OpenReview.net.

Sheng Zhou, Hongjia Xu, Zhuonan Zheng, Jiawei Chen, Zhao Li, Jiajun Bu, Jia Wu, Xin Wang, Wenwu Zhu, and Martin Ester. 2022b. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions. *ArXiv preprint*, abs/2206.07579.

## A Dataset Description

Here, we provide some additional information about the datasets. An overview of the datasets is given in Table 6, including name, number of classes, number of images, and size, given in pixels. You can find examples of images of each dataset in Table 7.

In the following, there is a small description of the datasets, including the class labels, provided in their original form which we also use in the evaluation of our explainability algorithm.

**STL10 (Coates et al., 2011).** This traditional dataset consists of 10 classes, namely “deer, horse,

bird, cat, ship, airplane, car, truck, monkey, dog”. We use the full dataset, i.e. train and test split. Note, that it is inspired by Cifar10 and attempts to be more complicated because it contains fewer images.

**Cifar10 (Krizhevsky and Hinton, 2009).** The dataset is comprised of 10 similar object classes: “deer, horse, bird, automobile, airplane, cat, ship, truck, dog, frog”. Again, we use the full dataset.

**ImageNet10.** Imagenet-10 is a subset of the larger ImageNet dataset, containing 10 classes. Given the hierarchical nature of of ImageNet, each class is described by multiple keywords: ‘trailer truck, tractor trailer, trucking rig, rig, articulated lorry, semi’, ‘snow leopard, ounce, *Panthera uncia*’, ‘airliner’, ‘Maltese dog, Maltese terrier, Maltese’, ‘sports car, sport car’, ‘orange’, ‘soccer ball’, ‘airship, dirigible’, ‘container ship, containership, container vessel’, ‘king penguin, *Aptenodytes patagonica*’

**Sports10 (Trivedi et al., 2021).** The Sports-10 dataset provides labeled images from 175 video games across 10 sports genres. The labels are “Car-Racing, Tennis, AmericanFootball, BikeRacing, TableTennis, Fighting, Basketball, Hockey, Soccer, Volleyball”.

**Inaturalist2021 (Grant Van Horn, 2021).** The full dataset contains images of 10,000 species separated into 10 classes, which are “Animalia, Arachnids, Amphibians, Birds, Insects, Ray-finned Fishes, Plants, Mollusks, Reptiles, Fungi, Mammals”. We experiment with the validation set.

Dataset Group	Name	No. of classes	No. of Images	Size (pixels)
Standard	STL10	10	13000	96x96
	ImageNet10	10	13000	500x364
	CIFAR10	10	60000	32x32
Background	Sports10	10	3000	1280x720
	iNaturalist 2021	11	100000	284x222
Human	LSUN	10	3000	341x256
	Human Action Recognition	15	18000	240x160
	FER2013	8	35488	48x48

Table 6: Overview over some basic dataset statistics.

**LSUN (Yu et al., 2015).** The Large-Scale Scene Understanding (LSUN) dataset offers labeled images depicting scenes from the following categories: “conference\_room, dining\_room, bedroom, church\_outdoor, bridge, tower, restaurant, living\_room, classroom, kitchen”. We experiment with the test set.

**HAR (Nagadia, 2022).** contains images of human activities. They are “running, sleeping, listening\_to\_music, texting, drinking, clapping, fighting, eating, sitting, using\_laptop, cycling, calling, laughing, hugging, dancing”.

**FER2013 (Barsoum et al., 2016).** The Facial Expression Recognition 2013 dataset consists of labeled grayscale images depicting human facial expressions, which are “surprise, anger, contempt, happiness, fear, disgust, sadness, neutral”.

## B Knowledge Injection

In section 4.2 we introduce prompt-guided clustering. For each dataset, multiple prompts are tested. They are generated by adapting the dataset name and transforming them into a question. Table 8 encompasses all prompts used in our experimental setup, accompanied by the corresponding evaluation performance metrics, namely Cluster Accuracy (Acc) and Normalized Mutual Information (NMI) for the image encoder representation, and the TF-IDF and SBERT representations. The used model is BLIP-2. Further, we provide a visual inspection of the same numbers in Figure 5.

In order to get a better understanding of the comparison of embedding structure, and how generated text relates to that, we provide two examples. In Figure 6 there is an example of the LSUN dataset and in Figure 7 there is a corresponding example of the Sports10 dataset.

## C Explainability

In this section, we provide pseudo-code for the algorithm in section 4.3. As described previously, it counts the number of keyword occurrences per cluster. Afterwards, it takes the top two exclusive keywords.

---

### Algorithm 1 Explainability

---

```

Require:
1:  $X = \{X_1, X_2, \dots, X_m\}$  : be the set of keyword lists for each sample,
2:  $Y = \{Y_1, Y_2, \dots, Y_m\}$  : be the set of (predicted) cluster labels for each
sample,
3:  $n$  : Number of output keywords per cluster.
Ensure: List
4: procedure SIMPLEXAI( $X, Y$ )
5:    $A, O \leftarrow [], []$  ▷ Active keywords, and others
6:   for  $i$  in unique( $Y$ ) do
7:      $K \leftarrow$  count-ordered list of keywords cluster  $i$ 
8:      $A[i] \leftarrow K[0 : n]$ 
9:      $O[i] \leftarrow K[n :]$ 
10:    end for
11:    while  $\bigcap_i A[i] \neq \emptyset$  do ▷ Remove duplicates
12:       $D \leftarrow \bigcap_i A[i]$ 
13:       $A[i] \leftarrow A[i] \setminus D$ 
14:       $A[i] \leftarrow A[i] \cup O[0 : |D|]$ 
15:       $O[i] \leftarrow O[2|D| :]$ 
16:    end while
17:    return  $A$ 
18: end procedure

```

---

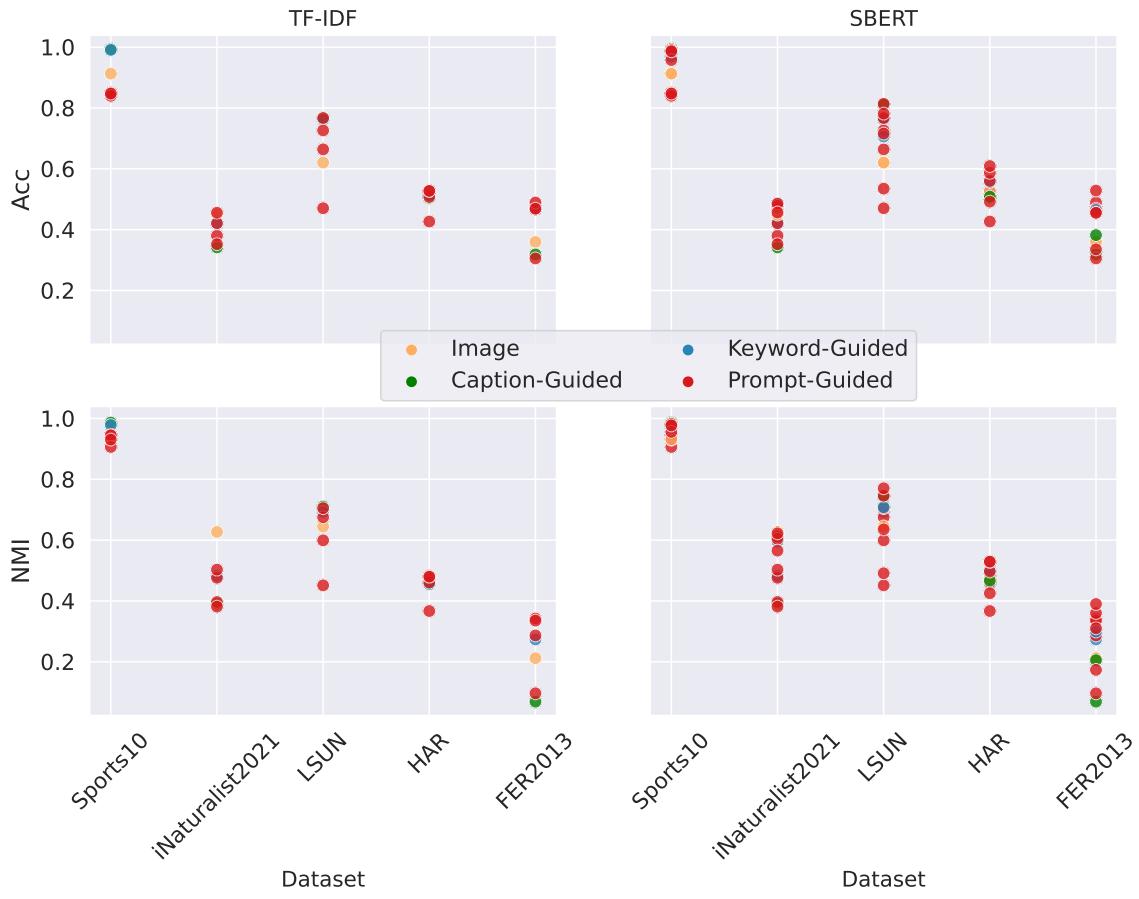


Figure 5: Comparison of all used strategies. Find the questions for prompt-guided clustering in Table 8.

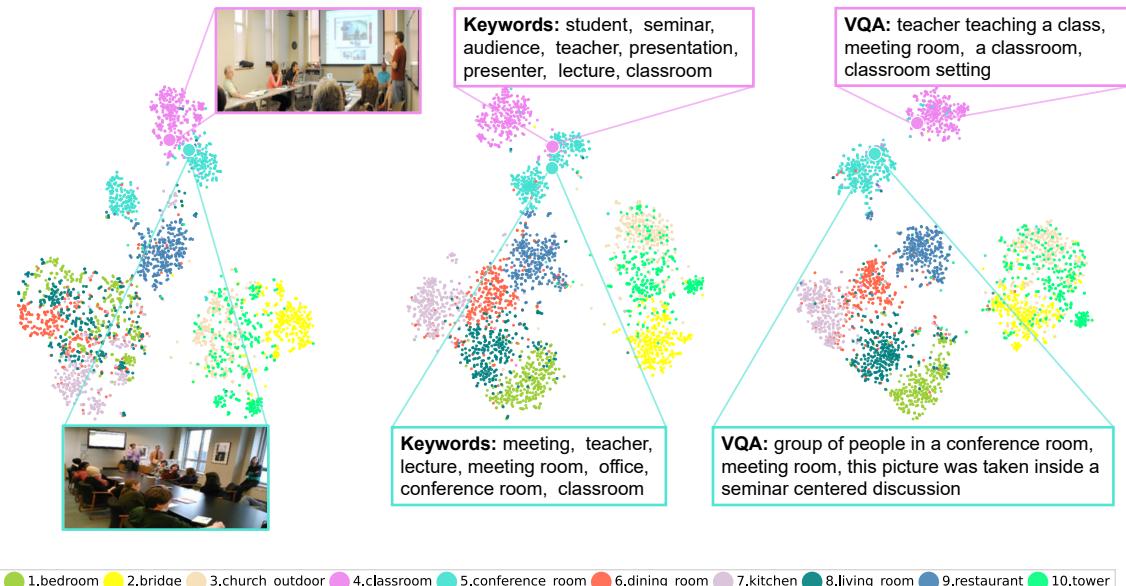


Figure 6: t-SNE embeddings of BLIP2 for the LSUN dataset. From left to right: Image embedding (Acc: 63.11), Keyword SBERT embedding (Acc: 71.12) and VQA SBERT embedding (Acc: 81.83 with prompt: "What environment is shown in the picture?"). The improvement in cluster accuracy corresponds to better separated clusters in the t-SNE embeddings.

Dataset	Image1	Label1	Image2	Label2
STL10		bird		car
CIFAR10		automobile		horse
ImageNet10		airship, dirigible		soccer ball
Sports10		CarRacing		BikeRacing
iNaturalist2021		Birds		Insects
LSUN		kitchen		bridge
Human Action Recognition		cycling		running
FER2013		anger		happiness

Table 7: Exemplatory images of the datasets. The images contain different properties, such as image quality or background noise. Also, the labels vary in their syntax and semantic meaning, e.g. objects vs. movements.

Dataset	Modality / Question	Image		TF-IDF		SBERT	
		Acc	NMI	Acc	NMI	Acc	NMI
Sports10	Image	91.31	93.22	99.38	98.65	99.07	98.47
	Caption			99.08	97.82	96.89	96.87
	Keyword			84.89	94.57	98.7	98.12
	Which sport is shown in the picture?			84.83	94.46	99.0	98.21
	What type of sport is shown in the picture?			84.0	90.64	95.77	95.58
	Which game is shown in the picture?			84.76	93.06	98.64	97.7
	Which sports contest is shown in the picture?						
iNaturalist2021	Image	44.97	62.7	34.17	39.07	47.43	61.63
	Caption			42.13	48.25	48.44	59.48
	Keyword			38.01	47.61	47.14	61.21
	What type of biological object is shown in the picture?			35.23	39.66	47.82	60.43
	What is the biological classification of the object in the picture?			42.1	50.3	48.57	62.23
	Which biological category is shown in the picture?			45.57	38.13	45.65	56.55
	Which species is shown in the picture?						
LSUN	Image	62.07	64.47	76.69	71.05	81.11	74.37
	Caption			76.2	69.28	70.63	70.82
	Keyword			47.04	45.12	53.49	49.11
	What location is shown in the picture?			72.63	67.52	81.37	74.6
	What kind of environment is shown in the picture?			66.4	59.92	71.59	63.54
	What room is shown in the picture?			76.71	70.5	78.15	77.05
	What scene is shown in the picture?						
HAR	Image	52.65	47.06	50.51	46.09	50.85	46.68
	Caption			51.35	45.47	55.66	50.07
	Keyword			42.68	36.69	49.2	42.54
	What type of motion is depicted in the picture?			50.77	46.04	56.03	49.69
	Which activity is shown in the picture?			52.75	48.13	58.68	52.86
	Which action is shown in the picture?			52.74	47.96	60.93	52.94
	What is the person doing in the picture?						
FER2013	Image	35.97	21.2	31.86	6.89	38.21	20.53
	Caption			47.05	27.34	46.44	29.96
	Keyword			30.53	9.64	33.53	17.34
	What type of countenance is shown in the picture?			46.86	34.25	45.6	36.04
	Which emotion is shown in the picture?			48.93	33.55	52.85	39.0
	Which facial expression is shown in the picture?			46.89	28.66	45.54	31.03
	Which mood is shown in the picture?						

Table 8: Full evaluation table for all prompts. All representations, image and text are based on the BLIP-2 model.

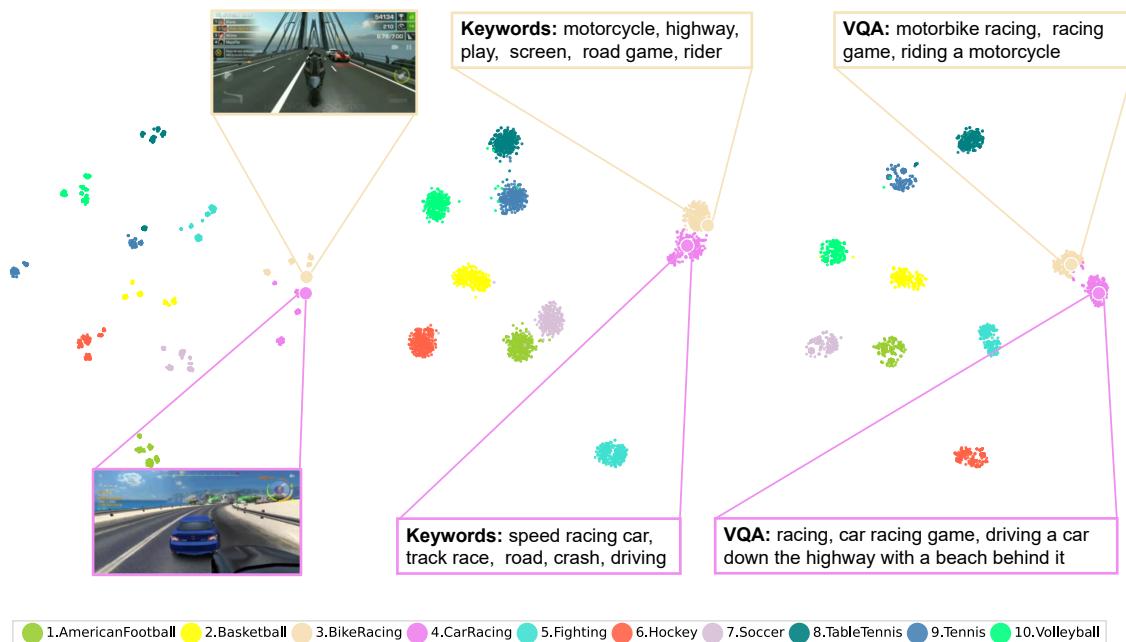


Figure 7: t-SNE embeddings of BLIP2 for the Sports10 dataset. From left to right: Image embedding (Acc: 91.31), Keyword SBERT embedding (Acc: 96.89) and VQA SBERT embedding (Acc: 99.00 with prompt: “What type of sport is shown in the picture?”). The improvement in cluster accuracy corresponds to better separated clusters in the t-SNE embeddings.