Political Analysis (2026), **00**, 1–20
doi:10.1017/pan.2026.10062

PA

ARTICLE

An Image Is Worth *K* Topics: A Visual Structural Topic Model with Pretrained Image Embeddings

Matías Piqueras[1], Alexandra Segerberg[2], Matteo Magnani[1], Måns Magnusson[3] and Nataša Sladoje[1]

[1]Department of Information Technology, Uppsala University, Sweden; [2]Department of Government, Uppsala University, Sweden; [3]Department of Statistics, Uppsala University, Sweden

Corresponding author: Matías Piqueras; Email: matias.piqueras@it.uu.se

(Received 20 May 2025 UTC; revised 21 April 2026 UTC; accepted 24 April 2026 UTC)

Abstract

Political scientists are increasingly interested in analyzing visual content at scale. However, the existing computational toolbox is still in need of methods and models attuned to the specific challenges and goals of social and political inquiry. In this article, we introduce a visual structural topic model (vSTM) that combines continuous image embeddings with a structural topic model. This has important advantages compared to existing approaches. First, continuous embeddings allow the model to capture the semantic complexity of images, which is relevant to the study of politics. A consequence of using continuous embeddings is that we can directly use off-the-shelf pretrained deep learning models, offering state-of-the-art performance in image analysis tasks while saving time, cost and resources. Second, the structural topic model provides the ability to analyze how topics and covariates are related, while maintaining a nuanced representation of images as a mixture of multiple topics. In our experiments, we show that the vSTM is able to identify topics that are interpretable, coherent and relevant to the study of online political communication.

Keywords: visual topic modeling; political communication; computer vision; bayesian inference; unsupervised learning; computational social science

Edited by: Daniel J. Hopkins and Brandon M. Stewart

1. Introduction

The study of politics needs a robust capacity to analyze visual data at large scale. Visual content is potent in conveying political ideas and mobilizing support, and the deluge of images and videos in the contemporary media environment, if anything, amplifies its role. Since images may convey something different from text (Coleman 2010), it can be crucial to analyze visual data directly, instead of via its accompanying text. However, large-scale visual political analysis often has to skip the visual content to instead rely on the accompanying text (e.g., captions or transcripts). While the text-as-data literature has developed a range of specialized computational methods tailored to social scientific purposes, the visual equivalent is still underdeveloped (Bucy and Joo 2021; Magnani and Segerberg 2021; Peng, Lock, and Ali Salah 2024; Torres and Cantú 2022; Williams, Casas, and Wilkerson 2020).

Large-scale image analysis presents a distinctive methodological frontier in computational social science. Images in political analysis such as photos, memes and data visualizations present their own challenge as compared with words in text-as-data but also as compared with how visual features such as color and texture are approached in medical image analysis. Images in social contexts can entail complex, multilayered semantics that invoke a semantic and cultural gap between what is easily extracted and how the image is understood (van Noord 2022; Smeulders *et al.* 2000). Intriguingly, recent developments in computer vision offer potential to broach this frontier directly.

© The Author(s), 2026. Published by Cambridge University Press on behalf of The Society for Political Methodology
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/pan.2026.10062

This article proposes a *visual* structural topic model (vSTM) that is designed to deal with the semantic complexity of images in political analysis, while enabling mixed membership modeling and the capacity to capture the relationship between a visual topic and other variables of interest. The ability to automatically categorize unlabeled image data is central to a number of analytical scenarios, several of which pivot on the relationship between visual content and other covariates. Examples can be found in studies analyzing how various actors use visual strategies to shape public agendas and engage audiences with their cause (e.g., Farkas and Bene 2021; Jing and Tian 2021; Kim and Bas 2023; Laaksonen *et al.* 2022; Mooseder *et al.* 2023; Peng 2021; Qian *et al.* 2024; Torres 2024), or research that analyzes key drivers in processes of discursive politicization or protest mobilization (e.g., Knüpfer, Schwemmer, and Heft 2023; Lu and Peng 2024; Oak and Nielsen 2024).

The vSTM makes three key contributions. First, it incorporates state-of-the-art image representation techniques that significantly advance the ability to capture semantic complexity in image data for the purpose of visual topic modeling. Second, it extends the statistical underpinnings of traditional text topic models to accept continuous embeddings to leverage the signature strength of mixed membership models for the analysis of image data. Third, we provide an easy-to-use software package designed for scalability and flexibility.

2. Visual topics and image representation

We conceive of a *visual topic* as a meaningful collation of *elements* recurring in images. Elements are features derived from the arrangement of pixels. They may range from low-level features such as color to high-level entities such as "person," "protester," or "Greta Thunberg." Technically, a visual topic is a distribution over these elements, analogous to how textual topics are distributions over words in text topic models. However, the elements constituting images are not necessarily discrete (as words in text are), and individual elements do not necessarily carry a self-contained meaning (as words do). Moreover, in text, units are organized in a natural hierarchy of meaning (e.g., letters form words and words form sentences), such that the underpinning of a topic is intuitive and direct—a set of words that co-occur with high probability. By contrast, digital images store information as pixels, which lack explicit semantic structure. Defining a visual topic in terms of pixels would be similar to defining a text topic in terms of letters.

This means that different types of visual topics are possible: the (combination of) elements that we use to represent an image shape the meaning we can extract and what the resulting topics emphasize. Visual content can involve one or several semantic layers. Broadly, these include a graphical layer of low-level features extracted from pixels, such as color, brightness and shapes, which can project emotion and meaning (Kress and Van Leeuwen 2020), and a literal layer that brings into play the denotative representational meaning of high-level features, the objects depicted in the image (Barthes 1977). Yet another layer of connotative meaning emerges in elements beyond the depiction, in the cultural, ideological, and emotional associations that are invoked for specific audiences (Barthes 1977).

Visual content relevant for political analysis will typically assume a complex interaction of several layers, even if a single layer may constitute the key analytical focus in particular studies. For example, some may focus on low-level features, such as color, brightness, and graphics and their role in content and its effects, such as how they frame content (e.g., Chen *et al.* 2022; Curini, Moffitt, and Zulianello 2024; Qian *et al.* 2024; Torres 2024), engage audiences (e.g., Lu and Pan 2022; Qian *et al.* 2024) or cue credibility and trust (e.g., Billard and Moran 2023; Peng, Lu, and Shen 2023; Von Sikorski 2022). In other cases, focus centers on the persuasive use of high-level features, such as symbols, faces and casual settings (e.g., Farkas and Bene 2021; León, Negredo, and Erviti 2022; Peng 2021; Segerberg and Magnani 2025). Pursuing simple semantics with focus on one layer can be parsimonious and powerful, but in other cases, it is important to capture a more complex interaction of layers. This can be the case when trying to capture content themes that depend on, but are not reducible to, specific visual features. Examples of this could be a breast cancer topic with images of women and pink ribbons symbolizing cancer awareness and support, or a visual topic of climate science criticism consisting of conceptually

https://doi.org/10.1017/pan.2026.10062

related, yet visually distinct content, such as data visualizations, historical news clippings and weather photographs.

The distinctive character of visual semantics also means that computer vision can offer a variety of approaches to represent images as numerical vectors (*image embeddings*) to retain meaning for use in statistical analysis. The choice of embedding is a key operationalization, as it defines what it means for images to be similar and likely to share a topic. Depending on the embedding, topics can differently emphasize the various layers of meaning and the complexity with which they interact.

2.1. Three paradigms of image representation

Computer vision approaches to image representation have evolved profoundly over recent decades. For our purposes, three paradigm shifts are relevant. Each offers conceptual and performance advances that support distinct paths to modeling visual topics. Continuous embeddings are important in these developments: the more we move toward complex semantic layers, the more we need to draw on continuous embeddings.

2.1.1. Feature engineering

Early approaches focused on extracting low-level features by designing algorithms to detect specific patterns in images, like textures and edges (Dalal and Triggs 2005; Ojala, Pietikäinen, and Harwood 1996). A key example is the bag-of-visual-words (BoVW) model, which extracts distinctive local patches, clusters them into a visual vocabulary and represents each image by the frequency of these visual "words" (Csurka *et al.* 2004; Sivic and Zisserman 2003). Using low-level descriptors like SIFT (Lowe 2004), BoVW provides a discrete representation designed to identify textural elements and simple shapes. This is useful for literally themed topics based on distinctive graphical details, such as the presence of barbed wire fences in media coverage of migrants (see Torres 2024). However, because representations produced by BoVW ignore spatial relationships and higher-level semantic concepts, images that share local elements but are fundamentally different (e.g., a protest, parade or "migrant caravan") risk being associated with the same topic.

2.1.2. Supervised feature learning

The shift to deep learning marked a significant leap for computational image analysis. Architectures like VGG and ResNet learn hierarchical features in a supervised manner, optimized to classify images into a predefined set of human-annotated labels (e.g., 1,000 classes in ImageNet; Krizhevsky, Sutskever, and Hinton 2012). Similarly, large datasets (e.g., COCO; Lin *et al.* 2014) have been gathered to train deep object detection models, to identify if and where the object categories in the training data are present in an input image.

Scholz *et al.* (2024) suggest that pretrained object detectors can be used to produce discrete image representations whose elements are the detected objects, to then train an image classifier. Used for topic modeling, this representation would discern visual topics based on the objects the model was trained to detect, which is both a strength—since the elements are easily interpretable—and a constraint, since it is limited to these categories. Training data may lack key categories, or they may be too coarse (e.g., reducing "politicians shaking hands" to "people"). Additionally, this representation overlooks spatial arrangement and composition (e.g., shot angles that convey power dynamics) or other graphical properties that may carry salience in political analysis (e.g., color grading that shifts emotional resonance).

Alternatively, continuous embeddings can be extracted from a network's intermediate layer. With models pretrained on large datasets, these embeddings can outperform earlier representations used in social science (Zhang and Peng 2022). Unlike the former, continuous embeddings capture not only the denotative objects but also their hierarchical and relational properties. This means that topics could be defined in terms of, and differently emphasize, graphical elements (e.g., color) or more semantically

https://doi.org/10.1017/pan.2026.10062

meaningful categories (e.g., people), depending on which layer is used (Yosinski *et al.* 2014; Zeiler and Fergus 2014). However, their elements (the values in the numerical vector constituting the embedding) are inherently bespoke as they are encouraged to capture variability in the categories of the training data, thereby often emphasizing denotative semantics (Arminio *et al.* 2026).

2.1.3. Self-supervised and weakly supervised feature learning

A third paradigm shifts toward task-agnostic representations learned through self-supervised or weakly supervised methods on much larger datasets, eliminating the need for expensive human annotations. These off-the-shelf continuous embeddings significantly advance the capacity to capture semantically complex visual topics, including ones that emphasize the connotative layer. Still, each approach encodes distinct assumptions.

One set of models, often based on the Vision Transformer architecture (ViT; Dosovitskiy *et al.* 2021), excel at complex topics with graphical and literal emphasis. Self-supervised models like DINO learn exclusively from images by producing similar embeddings for different augmented views (e.g., crops and color shifts) of the same image (Caron *et al.* 2021). This process leads to high-quality semantic segmentation and a remarkable pixel-level understanding of how objects are arranged and interact in a scene. For instance, in the image of politicians shaking hands, DINO embeddings may capture rich details beyond the objects themselves, including the background (e.g., flag and podium), composition (e.g., close-up and wide shot), interaction, attire, posture, facial expression and other graphical details (e.g., color grading).

A second approach uses text as a supervisory signal. Models like CLIP (Radford *et al.* 2021) are trained on a massive number of image–text pairs to learn a shared embedding space, where corresponding images and text are close together. This allows CLIP to learn comprehensive visual representations that move beyond typical object labels. Consequently, CLIP can in principle capture semantically complex visual topics based on conceptual and abstract themes that integrally involve the image. It can associate objects and symbols with larger entities, such as geographic location (e.g., the Eiffel Tower with Paris), and detect projected emotion, human traits, abstract concepts (e.g., "LGBTQ+") and visual styles like anime (Goh *et al.* 2021).

A final, emerging, paradigm involves generative models. Large language models (LLMs) increasingly incorporate a translation layer connecting image encoders (like CLIP or DINO) to a language model, allowing diverse inputs to be processed through a common textual interface (e.g., Li *et al.* 2023; Liu *et al.* 2023). Instead of a fixed representation, LLMs enable researchers to generate highly problem-specific and tailored textual descriptions. For example, the model can be prompted to focus on visual framing or emotional tone, or contextualize the image using accompanying text (e.g., a social media post). This approach can be used to cluster images based on their connotative meaning by first embedding the descriptions using a sentence transformer (Arminio *et al.* 2026). While offering unparalleled flexibility, a drawback is the potential semantic gap, as the textual description may omit image-native visual details.

2.2. Previous approaches to visual topic modeling

Pioneering work to tackle political image data with topic models has largely focused on the first two paradigms, remaining in the domain of graphical and literal visual topics. These approaches fall into two tracks.

2.2.1. Discrete feature approaches

This track uses a logic familiar from and compatible with traditional topic models, which operate on a document-term matrix. Torres (2024) skillfully exploits the discrete representation of images in BoVW to combine it with a structural topic model. Kim and Bas (2023) use a commercial object detection model to extract and count objects as input into a topic model, which could similarly be done using open-source models as suggested by Scholz *et al.* (2024) for supervised classification. The advantage

https://doi.org/10.1017/pan.2026.10062

here is smooth integration into existing topic model toolkits. The drawback is the difficulty in rendering semantically complex topics or being tied to pre-existing categories.

2.2.2. Single-membership clustering

The other track circumvents the topic modeling framework by using pretrained image embeddings for single-membership clustering (Joshi and Buntain 2024; Mooseder *et al.* 2023; Muise *et al.* 2022; Peng 2021). This track has been shown to produce more visually coherent clusters when compared to BoVW on image datasets based on political memes, protest, and climate change (Zhang and Peng 2022). Nevertheless, the approach has two limitations. One is that in the case of supervised pretrained embeddings, the clusters still emphasize graphical and literal similarity over connotative or conceptual similarity. For example, a cluster might effectively group "image types" (e.g., infographics), but risks overlooking distinct, ideologically relevant, meanings within those types (Joshi and Buntain 2024). As this can in principle be addressed using other types of embeddings, a more fundamental limitation is the assignment of images to a single cluster. Semantically rich visual content can refer to several themes, and single-membership may obscure analytical insights, for example, when graphical format topics (e.g., "memes" or "posters") dominate over literal content topics (e.g., "people"), even if the latter is present in the former.

2.3. Toward a vSTM with continuous embeddings

We contend that it is desirable for visual topic modeling in the context of political analysis to draw on approaches that can offer a spectrum of possible emphases (i.e., can capture graphical and literal, but also connotative similarity) and have the capacity to handle complex semantics in which multiple semantic layers of visual meaning come into play—that is, approaches based on continuous embeddings. We also hold that it is important to retain the flexibility of mixed membership modeling in semantically complex contexts.

There are two potential obstacles to moving in this direction. The first centers on interpretability. A critique of continuous embeddings is their opacity when used as model inputs compared to more straightforward discrete features (Scholz *et al.* 2024; Torres 2024). However, the balance is reversed when interpreting output-side topics. In the case of semantically complex images, a representation that is too simple risks producing topics that are less coherent and meaningful, and consequently difficult to interpret. We acknowledge that embracing continuous embeddings shifts the onus of interpretability to robust post hoc interpretation. As reported below, the vSTM therefore provides a number of outputs that can be used in this manner. We also provide evidence for the claim that the vSTM with continuous embeddings provides comparatively more interpretable and meaningful topics.

The other potential obstacle is that incorporating continuous embeddings requires an alternative formulation of the statistical underpinnings of the topic model. In the next section, we therefore provide such a formulation, extending a mixed-membership model that enables us to think of image embeddings as a weighted sum of topics and topic proportions. Our contribution is two fold. First, we introduce a hierarchical prior for topic proportions, linking covariates to the prevalence of topics through a generalized linear model, analogous to how STM extends LDA and CTM (Blei and Lafferty 2006; Blei, Ng, and Jordan 2003). Second, we focus on scalability, implementing and demonstrating how the model can be estimated using stochastic variational inference (Hoffman *et al.* 2013; Kucukelbir *et al.* 2017). This makes it feasible to explore several topic models in large image datasets, while also facilitating extensions and modifications of the model.

3. The visual STM

The vSTM is similar in spirit to the structural topic model for text in that the prevalence of topics depends on covariates through a generalized linear model (Roberts *et al.* 2014). To accommodate

https://doi.org/10.1017/pan.2026.10062

continuous embeddings, we draw on a set of related methods for modeling data as convex combinations of real-valued mixture components (Cutler and Breiman 1994; Giesen *et al.* 2023; Heller, Williamson, and Ghahramani 2008). We refer to the components as *topics* and their mixture weights as *topic proportions*. The vSTM differs from word-embedding topic models (e.g., Das, Zaheer, and Dyer 2015; Dieng, Ruiz, and Blei 2020) as it operates on a single embedding per observation, and unlike multimodal topic models (Bianchi, Terragni, and Hovy 2021; Zosa and Pivovarova 2022), our model does not require training neural networks or paired text–image data.

3.1. Statistical model

Formally, our data consist of N image embeddings $\boldsymbol{z}_i \in \mathbb{R}^D$ and image-level covariates $\boldsymbol{x}_i \in \mathbb{R}^P$, assumed to influence the prevalence of visual topics. We model each image embedding $\boldsymbol{z}_i$ using a normal distribution

$$\boldsymbol{z}_i \sim \text{normal}\left(\sum_{k=1}^{K} \theta_{i,k}\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_z\right), \tag{1}$$

where K is the number of topics, $\boldsymbol{\theta}_i$ is a vector of *topic proportions*, such that $\theta_{i,k} \geq 0$ and $\sum_k \theta_{i,k} = 1$, and $\boldsymbol{\beta} = [\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K]^T$ is a matrix of *topics*, where (Figure 1) $\boldsymbol{\beta}_k \in \mathbb{R}^D$. The image-specific topic proportions thus define how much each topic contributes to $\boldsymbol{z}_i$. A specific topic $\boldsymbol{\beta}_k$ is a point in the embedding space. If $\boldsymbol{z}_i$ is close to $\boldsymbol{\beta}_k$, then the corresponding topic proportion $\theta_{i,k}$ will also tend to be "large." $\boldsymbol{\Sigma}_z$ is the covariance matrix.

Topic proportions are sampled from a logistic-normal distribution, whose mean depends on the covariates $\boldsymbol{x}_i \in \mathbb{R}^P$ and a set of topic prevalence coefficients $\boldsymbol{\Gamma} \in \mathbb{R}^{P\times(K-1)}$

$$\begin{aligned} \boldsymbol{\eta}_i &\sim \text{normal}(\boldsymbol{x}_i\boldsymbol{\Gamma}, \boldsymbol{\Sigma}_\theta) \\ \boldsymbol{\theta}_i &= f(\boldsymbol{\eta}_i). \end{aligned} \tag{2}$$

This reflects the idea that images with similar covariates are also more likely to use the same topics. To ensure statistical identifiability, $\boldsymbol{\eta}_i$ is sampled as a $(K-1)$-dimensional Gaussian and mapped to the K-dimensional simplex via a transformation function f, rather than using an overparameterized K-dimensional softmax. This projection resolves the translation invariance of the standard softmax by anchoring the latent space. Our software defaults to the isometric log-ratio (ILR) transformation for this mapping, although other options like appending a fixed reference value (e.g., zero) to $\boldsymbol{\eta}_i$ achieve the same identifiability. Unlike the Dirichlet distribution, this construction allows correlations between topics to be modeled explicitly (Aitchison and Shen 1980; Blei and Lafferty 2006). We additionally implement a sparsemax transformation (Martins and Astudillo 2016) as an alternative to the ILR link, which can generate sparser topic proportions.

We also place priors on $\boldsymbol{\beta}_k, \boldsymbol{\Sigma}_z, \boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Gamma}$. We use student-t distributions for

$$\gamma_{p,k} \sim \text{student-t}(\nu_\gamma, 0, \sigma_\gamma)$$

and $\beta_{k,d} \sim \text{student-t}(\nu_\beta, 0, \sigma_\beta)$. Similar to Gaussian mixture models, multiple parameterizations are possible for the likelihood covariance $\boldsymbol{\Sigma}_z$, in this case, reflecting different assumptions about the error

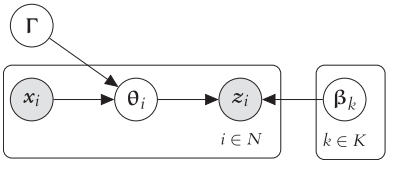


Figure 1. Plate diagram of the generative model of vSTM. $\boldsymbol{\beta}_k$ is a topic, $\boldsymbol{\theta}_i$ is the vector of topic proportions for image i, $\boldsymbol{\Gamma}$ is the matrix of topic prevalence coefficients, $\boldsymbol{z}_i$ is the image embedding and $\boldsymbol{x}_i$ are image-level covariates.

https://doi.org/10.1017/pan.2026.10062

structure of the embeddings (see also Gruhl and Erosheva 2014; Heller, Williamson, and Ghahramani 2008). A simple covariance (e.g., $\boldsymbol{\Sigma}_z = \sigma_z \boldsymbol{I}$) with fixed σ_z can work well in practice. Crucially, the choice must be considered in light of the specific geometry of the embeddings. In our analysis in Section 5.1, we found L2-normalizing the embeddings prior to fitting the model, and calibrating the variance to the hypersphere by scaling the standard deviation by $1/\sqrt{D}$ (since the expected magnitude of any single dimension on a unit hypersphere decreases as dimensionality D increases) to be robust and work well across different embeddings.

Finally, we decompose the topic proportion covariance $\boldsymbol{\Sigma}_\theta$ into a correlation matrix $\boldsymbol{\Omega}_\theta$ and standard deviations $\boldsymbol{\sigma}_\theta$, such that $\boldsymbol{\Sigma}_\theta = \text{diag}(\boldsymbol{\sigma}_\theta)\boldsymbol{\Omega}_\theta \text{diag}(\boldsymbol{\sigma}_\theta)$. We use the Lewandowski–Kurowicka–Joe (LKJ) prior on $\boldsymbol{\Omega}_\theta \sim \text{LKJ}(\eta_\theta)$, where $\eta_\theta > 0$ controls the degree of correlation between the visual topics. A value of $\eta_\theta = 1$ corresponds to a uniform distribution over correlation matrices, while values greater than one concentrate more mass around the diagonal (Lewandowski, Kurowicka, and Joe 2009). More specific details on the priors and hyperparameters used in Sections 4 and 5 are deferred to the Supplementary Material.

3.2. Inference

We implement model inference to be flexible and scalable. Specifically, vSTM utilizes the automatic differentiation and parallelization capabilities of the Python library JAX (Bradbury *et al.* 2018). This enables the user of our software to defer heavy computations to a GPU (if available), and to leverage existing implementations of state-of-the-art inference algorithms implemented in libraries like BlackJAX and NumPyro (Cabezas *et al.* 2024; Phan, Pradhan, and Jankowiak 2019). We make this easy by defining the model's log joint density $p(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_\theta, \boldsymbol{\sigma}_\theta, \boldsymbol{\sigma}_z | \boldsymbol{z}_i, \boldsymbol{x}_i)$ and gradient $\nabla \log p(\boldsymbol{\theta}_i, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \boldsymbol{\Omega}_\theta, \boldsymbol{\sigma}_\theta, \boldsymbol{\sigma}_z | \boldsymbol{z}_i, \boldsymbol{x}_i)$, conditioned on the data, in terms of parameters in an unconstrained space. Necessary reparameterizations and Jacobian adjustments (due to a change of variable) for constrained parameters (e.g., topic proportions $\boldsymbol{\theta}_i$) are handled internally. This is similar to the approach in probabilistic programming languages like Stan (Stan Development Team 2024) and is what enables compatibility with a wide range of general-purpose optimization and sampling methods that are designed for unconstrained spaces.

In this article, we use mean-field variational inference (MF-VI) throughout. We describe the details of our implementation and how we make it scalable to large data in Section D of the Supplementary Material. Since VI is known to underestimate the uncertainty of the posterior distribution, but can provide good point estimates of the posterior mean (Wang and Blei 2019), our focus is on the latter. These quantities are often of primary interest, so we argue that VI offers a practical trade-off that is often reasonable in topic modeling contexts where the researcher is likely to explore many models. Furthermore, recent work has shown that making accurate statistical and causal inferences using latent quantities inferred from unstructured data, requires additional considerations at a more fundamental level than the choice of a particular inference method (Egami *et al.* 2022, 2023).

3.3. Covariates and topic prevalence

A key output of the model is the relation between the covariates and topic prevalence. Since the image-level covariates $\boldsymbol{x}_i$ map to topic proportions $\boldsymbol{\theta}_i$ nonlinearly, we characterize the relationship more interpretably via the marginal distribution of topic proportions

$$\mathbb{E}[\boldsymbol{\theta}_i | \boldsymbol{x}_i] = \mathbb{E}[f(\boldsymbol{\eta}_i)], \tag{3}$$

where $\boldsymbol{\eta}_i \sim \text{normal}(\boldsymbol{x}_i \widehat{\boldsymbol{\Gamma}}, \widehat{\boldsymbol{\Sigma}}_\theta)$, f is the function mapping $\boldsymbol{\eta}_i$ to the simplex and $\widehat{\boldsymbol{\Gamma}}$ and $\widehat{\boldsymbol{\Sigma}}_\theta$ are point estimates of the prevalence coefficients and covariance matrix of topic proportions.

In practice, the researcher can use Equation (3) to predict the value of topic proportions at some specific covariate values of interest, use predictions at different values to make comparisons or compute an average prediction over a range of values. We show a concrete example in Section 4.

https://doi.org/10.1017/pan.2026.10062

3.4. Interpreting the topic-element distribution

The vSTM can in principle be combined with many types of image representations, some of which are more easily interpretable, while others contain elements that do not individually carry a clear meaning. As discussed in Section 2.1, less interpretable representations can be desirable for capturing complex semantics. In both cases, we need a procedure to summarize the topic distribution $\boldsymbol{\beta}_k$.

When the input is already based on interpretable elements, the interpretation is straightforward, as we can simply look at the elements that are the most characteristic for the images associated with each topic.

When image embeddings are difficult to interpret directly, it is still possible to obtain interpretable summaries of the topics post hoc using an auxiliary representation of the images. Let $\tilde{\boldsymbol{z}}_i$ be an auxiliary representation of $\boldsymbol{z}_i$, with interpretable elements. The probability of any particular element appearing in topic k is then

$$\tilde{\beta}_{kd} = \frac{\sum_{i=1}^{N} \hat{\theta}_{ik} \tilde{z}_{id}}{\sum_{d=1}^{D} \sum_{i=1}^{N} \hat{\theta}_{ik} \tilde{z}_{id}}, \tag{4}$$

where $\hat{\boldsymbol{\theta}}_i$ are the topic proportions obtained from the vSTM. As $\tilde{\boldsymbol{\beta}}_k$ is now a distribution over interpretable elements, it can be analyzed directly or be used for calculating other quantities, such as frequent and exclusive (FREX) keywords (Bischof and Airoldi 2012).

We view this approach as complementary to the necessary step of inspecting representative images, regardless of the embedding type. Sections 4 and 5 exemplify the trickier case when embeddings are less interpretable. In both cases, we use CLIP to produce the input embeddings ($\boldsymbol{z}_i$) and an LLM to generate text descriptions, which are encoded into an image-term matrix ($\tilde{\boldsymbol{z}}_i$). We note that this approach, in principle, can be used for any type of (input or auxiliary) interpretable elements, for example, object detection labels (Scholz *et al.* 2024) or visual words (Torres 2024).

4. Applying the model: Visual topics and their interpretation

This section uses an example from online climate communication to showcase the steps involved in applying the vSTM. We provide the code to fit the model efficiently on large-scale data as an easy-to-use software package, which also includes tools to aid interpretation, as shown in this section.

4.1. Data and setup

We focus on Twitter/X communication around United Nations Climate Change Conferences (COP) 21–27. The data consist of 154,177 unique images posted with the #COP2X hashtag by the most active users in the hashtag stream (N = 1,707), defined by the number of posts. We use the COP year, stakeholder type, and stance as our image-level covariates, assumed to influence the prevalence of visual topics. Each account was manually coded according to stakeholder type based on profile description, following a coding scheme developed from Mooseder *et al.* (2023). We also manually coded the account stance toward progressive climate action (for/against/unclear), based on profile description and recent (multimodal) post content.

4.1.1. Step 1: Choosing the embedding model

A first step in applying the vSTM is choosing which embeddings to use. Many factors influence how the *meaning* of images is encoded.

Table 1 gives an overview of example embeddings, following our discussion in Section 2. If research goals require connotative and complex topics, we recommend a "CLIP-style" embedding. These embeddings are easy and fast to generate and can represent semantic layers that extend beyond literal depiction. For more tailored representations, using an LLM as in Arminio *et al.* (2026) is also a good option. Pretrained embeddings from models trained under a supervised objective (e.g., VGG16 on ImageNet)

https://doi.org/10.1017/pan.2026.10062

Table 1. Image embedding options.

Paradigm	Example model	Embedding semantics
Supervised	Pretrained VGG16	Bespoke[1]
Self-supervised	DINO	Graphical/Literal
Weakly supervised	CLIP	Graphical/Literal/Connotative
Generative AI	Gemma 3	Tailored[2]

Note: The table shows example models that can be used to extract embeddings. 1. By *bespoke*, we mean that the semantics are tied to categories in pretraining data (e.g., ImageNet). 2. *Tailored* refers to the possibility to steer the focus of the embedding using a prompt.

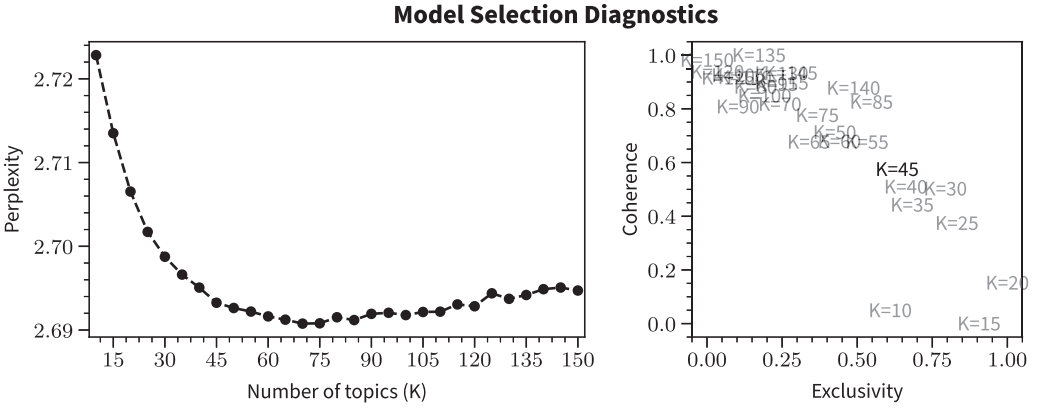


Figure 2. Diagnostics for model selection. The left panel shows perplexity, averaged over five held-out sets. The right panel shows coherence and exclusivity for each model, averaged over all topics.

should be considered primarily if the training data and learning task closely resemble the analysis data. If the goals of the research call for graphically and/or literally defined topics, self-supervised models (e.g., DINO) are likely to offer good off-the-shelf performance. We provide initial evidence of these claims in Section B of the Supplementary Material, by evaluating the coherence of topics resulting from different embeddings, in terms of denotation (what is depicted) and connotation (the implied meaning), under different embeddings.

In this section, we use CLIP's image encoder as an example to show how the model is run step by step.[1]

4.1.2. Step 2: Selecting the number of topics

The next step is to select the number of topics. To this end, two sets of diagnostics can provide practical guidance in narrowing down a range or selection of models to consider for further inspection.

The first concerns the predictive performance of the model. In this regard, we use five-fold cross-validation to compute *perplexity*, which is the exponential of the negative log likelihood on held-out data, normalized by the number of held-out observations and embedding dimensions. Lower perplexity indicates that the model, relatively speaking, is better at predicting unseen data. Since topic proportions are local to each image, we fit new ones for the held-out set, while keeping global parameters fixed.

In addition, *coherence*, and *exclusivity* scores (details in Section A.3 of the Supplementary Material) can serve as a surrogate for semantic validity—how internally coherent and yet distinctive the topics are (Quinn *et al.* 2010). Figure 2 shows the diagnostics for several models, varying the number of topics

[1] Additional details on the training and specific choice of hyperparameters are provided in Section A.3 of the Supplementary Material.

https://doi.org/10.1017/pan.2026.10062

between 10 and 150. The left panel shows perplexity, averaged over five held-out sets. It initially decreases sharply, before leveling off around $K = 50$, and later increases again. The right panel reports coherence and exclusivity averaged over all topics. Low exclusivity suggests that many topics are similar to each other, while low coherence indicates that the images within a topic are more heterogeneous.

In this example, we select $K = 45$ as a model that offers a balance between coherence, exclusivity and perplexity. We emphasize that this decision is ultimately made in a specific research context, where priorities can differ. The figure should in this sense not be read prescriptively, but rather as a tool to navigate how the model behaves as the resolution (broad vs. fine-grained topics) increases or decreases.

4.1.3. Step 3: Validation of topic coherence

As with any unsupervised method, manual validation is crucial (Grimmer, Roberts, and Stewart 2021; Quinn *et al.* 2010). Step 2 provides a range of K's where the model, relatively speaking, performs well with respect to a mathematical operationalization of what "well" means (e.g., topics are well separated in Euclidean space). Although useful given the computational and time constraints faced in practice, a good score does not necessarily imply that topics are meaningful to humans. Thus, to test more systematically if topics are coherent, we adapt the word-intrusion task to images (Chang *et al.* 2009; Ying, Montgomery, and Stewart 2022). We tasked four researchers to perform 100 evaluations, where each task consists of identifying the "intruder" in a set of four images, where three are associated with the same topic, and the fourth ("intruder") is from a different topic. The results show that evaluators are able to identify the intruder around 65% of the time, even as the same-topic images are drawn randomly from the pool of images where the topic is the *largest* (i.e., $k = \arg\max_k \theta_{ik}$), rather than from the top of the distribution. This indicates that evaluators intuitively perceived topics as representing the same theme or concept. Further details regarding the procedure and additional considerations on the task design are provided in Section A.4 of the Supplementary Material.

4.1.4. Step 4: Interpreting and labeling topics

Step 4 is to interpret and label topics. We maintain that it is important to include an image-native approach to interpreting visual topics. In contrast to reading representative documents, inspecting a large number of images is fast and intuitive. In order to label topics, we recommend visualizing images that are representative for each topic (see also Torres 2024; Zhang and Peng 2022). One can consider either the top N images with the highest topic proportion θ_{ik} for each k, or a random sample of images with high proportions. The former provides a sense of what prototypical images look like; the latter gives a better sense of the variety of images included, and if the meaning is still coherent in less pronounced cases.

As a complement, the approach in Section 3.4 can be used to generate an interpretable representation of topics post hoc, using both images and text. Here, we use the Gemma 3 model (Gemma Team *et al.* 2025) to generate a description for each image, using a simple open-ended prompt.[2] A part-of-speech tagger is then used to extract relevant word classes (nouns, verbs, and adjectives) from the descriptions, which are transformed into an image-term matrix $\tilde{\boldsymbol{Z}} \in \mathbb{N}^{N \times V}$, where V is the number of unique terms. We then compute the topic-term distribution $\tilde{\boldsymbol{\beta}}_k$ using Equation (4) and use this to derive FREX scores (Bischof and Airoldi 2012).

We assign topic labels by inspecting 50 top images and a random sample of 200 images for each topic. We also look at the top 25 FREX nouns, verbs and adjectives for each topic. Figure 3 shows an example of what this looks like in practice, for a selection of topics. In this case, we find that topics are generally coherent but some are more difficult to label than others, mainly due to being similar to other topics. We include an overview of all topics and broader macro-topic categories in Table 1 in the Supplementary Material.

[2] "Describe the image in a coherent paragraph of three to four sentences."

https://doi.org/10.1017/pan.2026.10062



Figure 3. Top six images from a selection of topics. The top FREX nouns, verbs, and adjectives of a topic are shown above the images and are identified automatically using a part-of-speech tagger on the LLM image descriptions.

4.1.5. Step 5: Analyzing covariates and topic prevalence

While steps 1–4 are generally applicable, the use of further parameters for analysis depends on the research question. This said, a key output of the model is $\boldsymbol{\Gamma}$, which characterizes how *image-level covariates* influence the *prevalence* of topics.

Figure 4 exemplifies this in the context of a question that concerns the online discourse of actors for and against progressive climate action. It shows how, in our case, topics vary by actor type, stance, and COP event for a selection of topics, computed using the marginal distribution of topic

https://doi.org/10.1017/pan.2026.10062

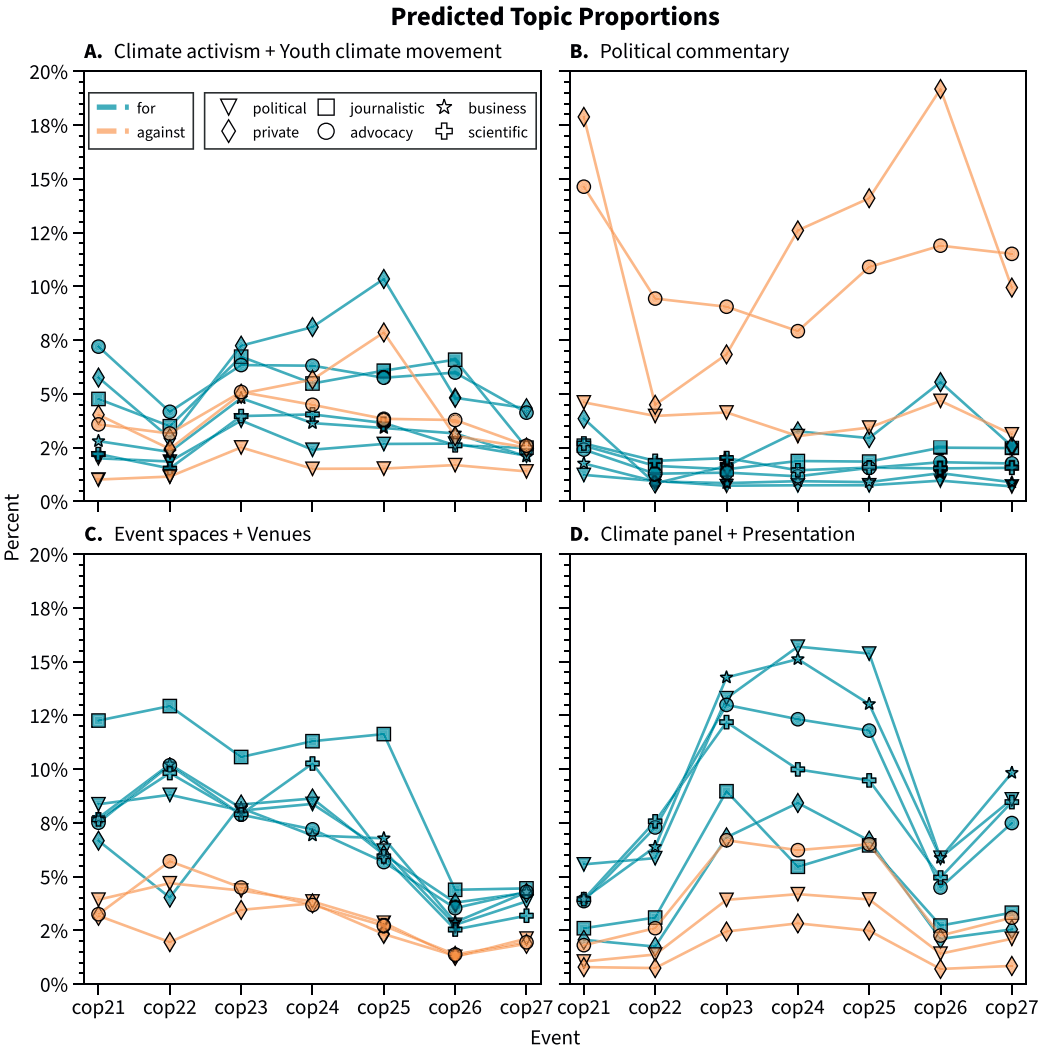


Figure 4. Predicted topic proportions by actor type and stance, across each COP for a selection of topics. Topics with a "+" are merged into one.

proportions (Equation 3). Top-producing private individual accounts are likely to post images of "climate activism," but diverge in their rendition: "For" accounts post undramatic images of turnout in street demonstrations; "against" accounts post images that follow a delegitimizing "protest paradigm" frame (Chan and Lee 1984) by highlighting police confrontations, disruption (e.g., blocking traffic) or spectacle (e.g., activists dressed in costumes). As a group, top-producing politicians, business actors, and scientists post a different set of topics, such as witnessing conference scenes (e.g., "climate panels"). Journalistic actors document summit scenes, political elites, and protests. Their visual coverage of summit scenes (e.g., "event spaces") decreases, correlating with the pandemic and increasing constraints on media access (Hayes and O'Neill 2025).

Note that our focus here is to show the model in application, and that this figure should be interpreted with caution. Recall that the data consist of unique images posted by highly productive accounts, not the most popular content. There were only a handful of accounts with a stance "against" climate action in this data.

https://doi.org/10.1017/pan.2026.10062

5. Comparison to existing approaches

We now proceed to compare vSTM to other approaches to visual structural topic modeling both quantitatively and qualitatively. Currently, the BoVW representation model combined with STM is the most established approach and constitutes the state of the art in a social science context (Torres 2024). In the quantitative comparison, we also compare vSTM to STM applied to object detection labels and LLM-generated image descriptions. While not previously evaluated for topic modeling, these discrete visual elements have been proposed as useful representations for classification and clustering tasks (Arminio *et al.* 2026; Scholz *et al.* 2024).

5.1. Semantic validity of topics

We compare the semantic validity of topics produced by vSTM to discrete approaches with STM. We use a dataset of almost 12,000 images related to climate communication from Zhang and Peng (2022), for which Arminio *et al.* (2026) produced expert annotations for 500 pairs of images that were annotated for their denotative and connotative similarity on a three-point scale (0 = unrelated, 1 = somewhat related and 2 = closely related). To quantify semantic validity, we use the connotative annotations to adopt a measure for clustering quality from Grimmer and King (2011). We assign each image to its dominant topic and calculate a quality score Q as the average expert-rated similarity for same-topic pairs, minus the average expert-rated similarity for different-topic pairs. The maximum possible score is 2, the minimum is −2, and a score of 0 suggests that the model performs no better than random assignment.

We fit several configurations of vSTM and STM to the same set of images. For robustness, we test vSTM with two embeddings: CLIP and GPT4o-generated image descriptions, embedded using a sentence transformer.[3] For STM, we use BoVW, the same GPT4o descriptions, but represented in a document-feature matrix (DFM), and object detection labels produced by the YOLOE model (Wang *et al.* 2025). For BoVW, we must set the number of visual words that form the vocabulary. We test several sizes (500, 1,000, 2,500, 5,000, 7,500, and 10,000) and report here the one producing the best results, which is 2,500 visual words. For the object-based STM, we use YOLOE as it has a comparably large vocabulary (1,000+ objects). We represent the images by the number of times an object appears in the image, akin to a DFM. A threshold on the confidence of the predictions must be set to determine which of the detected objects should be kept. This is a trade-off between accuracy and coverage. We use 0.1, as we find that a lower threshold generally leads to better results on the evaluation, even as the labels are noisier. A higher threshold quickly leads to many images without a single label, which would force us to remove them from the analysis.[4]

We vary the number of topics between 10 and 100 to avoid the risk of bias in favor of any particular approach. For each configuration, we fit the model 10 times. We use short-run LDA to initialize STM (Roberts, Stewart, and Tingley 2019) and draw the initial parameters from a uniform distribution for vSTM.

Figure 5 reports the average quality score Q with a 95% confidence interval. We find that more complex representations (CLIP+vSTM, GPT4o+vSTM, and GPT4o+STM) significantly outperform BoVW+STM and YOLOE+STM across all K. The results indicate that vSTM performs marginally better than STM with GPT4o representations for most K. This aligns with our expectation that converting a rich textual description into a simpler DFM leads to loss of information, which is avoided by vSTM's use of transformer embeddings. While on average scoring lower than the GPT4o configurations, CLIP+vSTM performs remarkably well compared to BoVW+STM and YOLOE+STM. This is important. CLIP embeddings are available off the shelf, are generally much faster to obtain and do not require designing a prompt. Moreover, the GPT4o prompt was in this case explicitly tuned to

[3] These descriptions (and embeddings) are taken directly from Arminio *et al.* (2026).
[4] Additional details are provided in Section C of the Supplementary Material.

https://doi.org/10.1017/pan.2026.10062

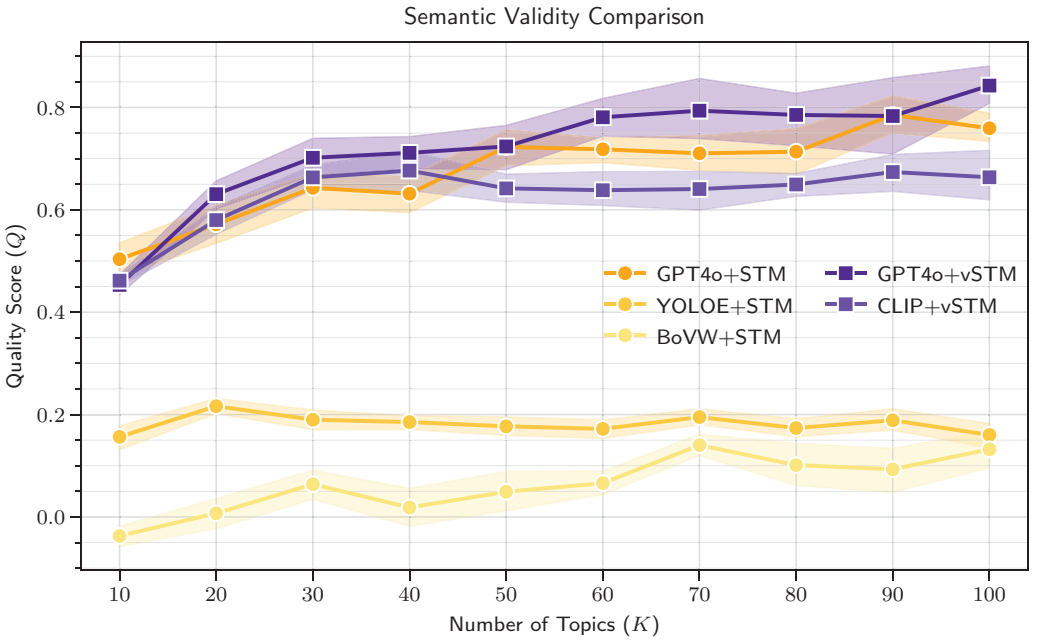


Figure 5. Quality score for different discrete and continuous configurations using STM and vSTM, respectively.

summarize the connotative meaning of the images, whereas CLIP's embeddings are general-purpose. Our main takeaway is that modern, complex image representations are crucial for good performance in visual topic modeling. The continuous vSTM framework is uniquely positioned to fully leverage the strength of these representations.

5.2. Qualitative inspection of vSTM and bag of visual words + STM topics

We test vSTM's capacity to generate topics that are coherent, interpretable, and accurate in capturing concepts of substantive interest in comparison to BoVW + STM. Therefore, we replicate Torres's (2024) analysis of the extent to which media ideology affects coverage of a "caravan" of migrants attempting to enter the United States on foot. An STM is fitted to 688 images represented using a BoVW and identifies a topic of interest: the extent to which dense crowds are present in an image. The "Dense crowd" topic is assumed to be a frame that casts the size of the migrant group as a threat.

We applied vSTM using the same number of topics (15) as in the original study and identified those portraying large crowds and migrants. We again use CLIP to generate the image embeddings and Gemma 3 to produce image descriptions and generate topic keywords in the same way as in Section 4. Figures 6 and 7, respectively, show the three vSTM topics that portray migrants and the "Dense crowd" topic from Torres (2024). As the method for generating keywords is agnostic to how the topics were generated, we use the most representative images and top FREX nouns, verbs, and adjectives to label the topics in both cases.

The results show that the vSTM migrant topics are comparatively more coherent with respect to the semantic meaning, both in terms of how similar representative images are, and the interpretability of keywords. The topic "Migrant caravan (distant view)" consistently shows images of the caravan: people walking in a dense crowd, often taken from above or from a distance. The keywords contain words, such as *caravan, dense*, and *migrant*. Meanwhile, the "Dense crowd" topic discovered by the BoVW+STM approach is more ambiguous. Among the top images, we find pictures of protests, Trump rallies and soldiers. While the first two include dense crowds, they are arguably less relevant for the particular frame under study.

https://doi.org/10.1017/pan.2026.10062



Figure 6. Top 18 most representative images for three migrant topics. The top FREX keywords are computed using the method in Section 3.4.

The upshot is that while the "Dense crowd" topic may better capture how *much* of a picture is filled with people, it is less likely to adequately capture what is *actually* portrayed. This suggests that relying on low-level visual features can produce topics that are well defined according to specific graphical or literal criteria but fail to capture meaning in a broader sense. In contrast, by relying on a more powerful

https://doi.org/10.1017/pan.2026.10062



Figure 7. The top 18 most representative images for the "Dense crowd" topic from Torres (2024). The top FREX keywords are computed using the method in Section 3.4.

representation (here CLIP), vSTM can produce semantically valid topics that are easier to interpret. The other topics in Figure 6 are similarly meaningful yet distinct in portraying migrant groups. For example, the topic we label "Children and circumstances" emphasizes difficult circumstances and vulnerable individuals, especially mothers with children, rather than the size of the caravan.

6. Concluding remarks

Topic models have become a standard tool for large-scale content analysis of text in the social sciences. As visual data become more central in political life, there is a growing need for similar methods that can systematically categorize images. To meaningfully help address questions on the basis of large-scale visual digital communication and trace data, such methods need to encompass a spectrum of visual semantic layers and have the capacity to capture the complex interaction between them.

We propose the vSTM, a visual topic model that adapts traditional probabilistic topic models to the image domain by leveraging pretrained image embeddings. The article shows empirically that vSTM produces coherent topics, outperforming traditional approaches based on low-level visual features. It also demonstrates how topics can be interpreted and labeled efficiently using a combination of visualizations and topical keywords, despite the embeddings themselves not being directly interpretable. Ultimately, by jointly estimating latent visual themes and their relationship with document-level covariates, the vSTM provides a novel way to test and advance theories of visual political communication at scale. Specifically, the model's ability to capture multilayered semantics, from graphical components and literal depictions to more affective and abstract themes, allows researchers to systematically analyze how, when and with what consequences political actors of various kinds use and respond to nontextual content and cues for persuasion and mobilization. Consequently, the vSTM enables a deeper theoretical integration of visuality into models of agenda setting, public opinion and political behavior.

This work suggests several directions for future research. First, there is a need to develop and evaluate robust validation procedures that are tailored to visual topic models and can account for the multilayered semantics of images. Relatedly, while we provide conceptual and practical guidance on the choice of pretrained embeddings, future work should further examine their implications for measurement and potential biases. Additionally, a current limitation when using variational inference to estimate the model is reliably quantifying the uncertainty of the covariate-topic relationships, which remains an important point for future work. Finally, several extensions of the model are possible, including jointly modeling multiple modalities, such as text and images, to include semi-supervised components

https://doi.org/10.1017/pan.2026.10062

guiding the topic discovery toward key concepts of interest, or altering the likelihood to explicitly model directional data, as embeddings often are.

Acknowledgments. We thank Michelle Torres, Victoria Yantseva, Joakim Lindblad, Sophie Mainz, Emelie Karlsson, Cristóbal Moya, Xin Shen, Diletta Goglia, Inga Wohlert, Davide Vega, Simon Herrmann, and Väinö Yrjänäinen for feedback and comments. We also appreciated comments from participants of the 2024 workshop "How Image-As-Data Approaches Can Help Analysing Protest and Its Organisation: Methods and Applications" at Freie Universität Berlin. We thank members of the Center for Social Media and Politics (CSMaP), New York University. The feedback from Benjamin Guinaudeau, Jason Greenfield, Melina Munch, Chris Schwarz, Amanda Drucker, and Kylan Rutherford greatly helped improve the final draft of the article.

Funding statement. This research has been funded by the Swedish Research Council (Grant No. 2021-02769) and by the Swedish Research Council for Health, Working Life and Welfare (FORTE) (Grant No. 2021-01646) under the CHANSE ERA-NET Co-fund Programme, which has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No. 101004509. Experiments were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through Grant Agreement No. 2022-06725. M.M. was funded by the Swedish Research Council (Grant No. 2022-03381). M.P. acknowledges support from The Karl Staaff Foundation.

Data availability statement. Replication material for this article is available at the Harvard Dataverse at https://doi.org/10.7910/DVN/WATYCB (Piqueras *et al.* 2026).

Competing interests. The authors declare no competing interests.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/pan.2026.10062.

References

Arminio, L., M. Magnani, M. Piqueras, L. Rossi, and A. Segerberg. 2026. "Leveraging VLLMs for Visual Clustering: Image-to-Text Mapping Shows Increased Semantic Capabilities and Interpretability." *Social Science Computer Review* 44 (3): 572-91. https://doi.org/10.1177/08944393251376703

Aitchison, J., and S. M. Shen. 1980. "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika* 67 (2): 261–272. https://doi.org/10.1093/biomet/67.2.261

Barthes, R. 1977. *Image, Music, Text: Essays*, 13. [Dr.] Edn. London: Fontana.

Bianchi, F., S. Terragni, and D. Hovy. 2021. "Pre-Training Is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence." In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, edited by C. Zong, F. Xia, W. Li, and R. Navigli, 759–766. Online: Association for Computational Linguistics.

Billard, T. J., and R. E. Moran. 2023. "Designing Trust: Design Style, Political Ideology, and Trust in Fake News Websites." *Digital Journalism* 11 (3): 519–546. https://doi.org/10.1080/21670811.2022.2087098

Bischof, J., and E. M. Airoldi. 2012. "Summarizing Topical Content with Word Frequency and Exclusivity." In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, edited by J. Langford and J. Pineau, 201–208. Omniprint.

Blei, D., and J. Lafferty. 2006. "Correlated Topic Models." In *Advances in Neural Information Processing Systems*, Vol. 18, edited by Y. Weiss, B. Schölkopf, and J. Platt, 147–154. Cambridge, MA: MIT Press.

Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3: 993–1022.

Bradbury, J., et al. 2018. "JAX: Composable Transformations of Python+NumPy Programs. V. 0.3.13." http://github.com/jax-ml/jax

Bucy, E. P., and J. Joo. 2021. "Editors' Introduction: Visual Politics, Grand Collaborative Programs, and the Opportunity to Think Big." *The International Journal of Press/Politics* 26 (1): 5–21. https://doi.org/10.1177/1940161220970361

Cabezas, A., et al. 2024. "BlackJAX: Composable Bayesian Inference in JAX." Preprint, arXiv:2402.10797.

Caron, M., et al. 2021. "Emerging Properties in Self-Supervised Vision Transformers." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 9650–9660. New York: IEEE. https://doi.org/10.1109/ICCV48922.2021.00951

Chan, J. M., and C.-C. Lee. 1984. "The Journalistic Paradigm on Civil Protests: A Case Study of Hong Kong." In *The News Media in National and International Conflict*, edited by A. Arno and W. Dissanayake, 183–202. Boulder, CO: Westview Press.

Chang, J., S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." In *Advances in Neural Information Processing Systems*, Vol. 22, edited by Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, 288–296. Red Hook, NY: Curran Associates, Inc. https://doi.org/10.5555/2984093.2984126

Chen, K., S. J. Kim, Q. Gao, and S. Raschka. 2022. "Visual Framing of Science Conspiracy Videos: Integrating Machine Learning with Communication Theories to Study the Use of Color and Brightness." *Computational Communication Research* 4 (1): 98-134.

https://doi.org/10.1017/pan.2026.10062

Coleman, R. 2010. "Framing the Pictures in our Heads: Exploring the Framing and Agenda-Setting Effects of Visual Images." In *Doing News Framing Analysis*, edited by P. D'Angelo and J. Kuypers, 233-261. New York: Routledge.

Csurka, G., C. Dance, L. Fan, J. Willamowski, and C. Bray. 2004. "Visual Categorization with Bags of Keypoints." In Proceedings of the Workshop on Statistical Learning in Computer Vision (ECCV 2004), 1–22. Berlin: Springer.

Curini, L., B. Moffitt, and M. Zulianello. 2024. "The Colors of the Populist Radical Right: The Strategic Use of Hue and Saturation in Party Logos." *The International Journal of Press/Politics*. https://doi.org/10.1177/19401612241229216

Cutler, A., and L. Breiman. 1994. "Archetypal Analysis." *Technometrics: A Journal of Statistics for the Physical, Chemical, and Engineering Sciences* 36 (4): 338–347.

Dalal, N., and B. Triggs. 2005. "Histograms of Oriented Gradients for Human Detection." In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, Vol. 1, 886–893. New York: IEEE. https://doi.org/10.1109/CVPR.2005.177

Das, R., M. Zaheer, and C. Dyer. 2015. "Gaussian LDA for Topic Models with Word Embeddings." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, edited by C. Zong, and M. Strube, 795–804. Beijing: Association for Computational Linguistics. https://doi.org/10.3115/v1/P15-1077

Dieng, A. B., F. J. R. Ruiz, and D. M. Blei. 2020. "Topic Modeling in Embedding Spaces." *Transactions of the Association for Computational Linguistics* 8: 439–453.

Dosovitskiy, A., et al. 2021. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale." In *Proceedings of the 9th International Conference on Learning Representations (ICLR 2021)*. OpenReview. https://openreview.net/forum?id=YicbFdNTTy.

Egami, N., C. J. Fong, J. Grimmer, M. E. Roberts, and B. M. Stewart. 2022. "How to Make Causal Inferences Using Texts." *Science Advances* 8 (42): eabg2652. https://doi.org/10.1126/sciadv.abg2652

Egami, N., M. Hinck, B. Stewart, and H. Wei. 2023. "Using Imperfect Surrogates for Downstream Inference: Design-Based Supervised Learning for Social Science Applications of Large Language Models." In *Advances in Neural Information Processing Systems*, Vol. 36, edited by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, 68589–68601. Red Hook, NY: Curran Associates, Inc.

Farkas, X., and M. Bene. 2021. "Images, Politicians, and Social Media: Patterns and Effects of Politicians' Image-Based Political Communication Strategies on Social Media." *The International Journal of Press/Politics* 26 (1): 119–142. https://doi.org/10.1177/1940161220959553

Gemma Team, et al. 2025. "Gemma 3 Technical Report." Preprint. https://doi.org/0.48550/arXiv.2503.19786.

Giesen, J., P. Kahlmeyer, S. Laue, M. Mitterreiter, F. Nussbaum, and C. Staudt. 2023. "Mixed Membership Gaussians." *Journal of Multivariate Analysis* 195: 105141. https://doi.org/10.1016/j.jmva.2022.105141

Goh, G., et al. 2021. "Multimodal Neurons in Artificial Neural Networks." *Distill* 6 (3): e30. https://doi.org/10.23915/distill.00030

Grimmer, J., and G. King. 2011. "General Purpose Computer-Assisted Clustering and Conceptualization." *Proceedings of the National Academy of Sciences* 108 (7): 2643–2650. https://doi.org/10.1073/pnas.1018067108

Grimmer, J., M. E. Roberts, and B. M. Stewart. 2021. "Machine Learning for Social Science: An Agnostic Approach." *Annual Review of Political Science* 24 (1): 395–419. https://doi.org/10.1146/annurev-polisci-053119-015921

Gruhl, J., and E. A. Erosheva. 2014. "A Tale of Two (Types of) Memberships: Comparing Mixed and Partial Membership with a Continuous Data Example." In Handbook of Mixed Membership Models and Their Applications, edited by E. M. Airoldi, D. M. Blei, E. A. Erosheva, and S. E. Fienberg, 15–38. Boca Raton, FL: Chapman & Hall/CRC.

Hayes, S., and S. O'Neill. 2025. "Visual Politics, Protest, and Power: Who Shaped the Climate Visual Discourse at COP26?" *Journalism Studies* 26 (4): 441–463. https://doi.org/10.1080/1461670X.2024.2436912

Heller, K. A., S. Williamson, and Z. Ghahramani. 2008. "Statistical Models for Partial Membership." In *Proceedings of the 25th International Conference on Machine Learning ICML '08*, 392–399. New York, NY: Association for Computing Machinery. https://doi.org/10.1145/1390156.1390206

Hoffman, M. D., D. M. Blei, C. Wang, and J. Paisley. 2013. "Stochastic Variational Inference." *Journal of Machine Learning Research* 14 (40): 1303–1347.

Jing, L., and Y. Tian. 2021. "Self-Supervised Visual Feature Learning with Deep Neural Networks: A Survey." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43 (11): 4037–4058. https://doi.org/10.1109/TPAMI.2020.2992393

Joshi, A., and C. Buntain. 2024. "Examining Similar and Ideologically Correlated Imagery in Online Political Communication." In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 18, edited by Y.-R. Lin, Y. Mejova, and M. Cha, 774–786.

Kim, M., and O. Bas. 2023. "Seeing the Black Lives Matter Movement through Computer Vision? An Automated Visual Analysis of News Media Images on Facebook." *Social Media + Society* 9 (3). https://doi.org/10.1177/20563051231195582

Knüpfer, C. B., C. Schwemmer, and A. Heft. 2023. "Politicization and Right-Wing Normalization on YouTube: A Topic-Based Analysis of the Alternative Influence Network." *International Journal of Communication* 17 (2023): 23-23.

Kress, G., and T. Van Leeuwen. 2021. *Reading Images: The Grammar of Visual Design*. 3rd ed. London: Routledge.

https://doi.org/10.1017/pan.2026.10062

Krizhevsky, A., I. Sutskever, and G. E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." In *Advances in Neural Information Processing Systems*, In Advances in Neural Information Processing Systems, Vol. 25, edited by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, 1097–1105. Red Hook, NY: Curran Associates, Inc.

Kucukelbir, A., D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. 2017. "Automatic Differentiation Variational Inference." *Journal of Machine Learning Research* 18 (14): 1–45.

Laaksonen, S.-M., J. Hokka, M. Nelimarkka, and K. Nikunen. 2022. "Affective Visual Rhetoric and Discursive Practices of the Far-Right across Social Media." In *The Far-Right Discourse of Multiculturalism in Intergroup Interactions: A Critical Discursive Perspective*, edited by K. Pettersson and E. Nortio, 189–216. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-89066-7_8

León, B., S. Negredo, and M. C. Erviti. 2022. "Social Engagement with Climate Change: Principles for Effective Visual Representation on Social Media." *Climate Policy* 22 (8): 976–992. https://doi.org/10.1080/14693062.2022.2077292

Lewandowski, D., D. Kurowicka, and H. Joe. 2009. "Generating Random Correlation Matrices Based on Vines and Extended Onion Method." *Journal of Multivariate Analysis* 100, no.9: 1989–2001. https://doi.org/10.1016/j.jmva.2009.04.008

Li, J., D. Li, S. Savarese, and S. Hoi. 2023. "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models." In *Proceedings of the 40th International Conference on Machine Learning*, edited by A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, 19730–19742. Cambridge, MA: PMLR.

Lin, T.-Y., et al. 2014. "Microsoft Coco: Common Objects in Context." In *European Conference on Computer Vision*, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, 740–755. Cham: Springer.

Liu, H., C. Li, Q. Wu, and Y. J. Lee. 2023. "Visual Instruction Tuning." *Advances in Neural Information Processing Systems*, Vol. 36, edited by A. Oh, T. Neumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, 34892–34916. Red Hook, NY: Curran Associates, Inc.

Lowe, D. G. 2004. "Distinctive Image Features from Scale-Invariant Keypoints." *International Journal of Computer Vision* 60: 91–110.

Lu, Y., and J. Pan. 2022. "The Pervasive Presence of Chinese Government Content on Douyin Trending Videos." *Computational Communication Research* 4 (1): 68-97. https://doi.org/10.5117/CCR2022.2.002.LU

Lu, Y., and Y. Peng. 2024. "The Mobilizing Power of Visual Media across Stages of Social-Mediated Protests." *Political Communication* 41 (4): 531–558. https://doi.org/10.1080/10584609.2024.2317951

Magnani, M., and A. Segerberg. 2021. "On the Conditions for Integrating Deep Learning into the Study of Visual Politics." In *13th ACM Web Science Conference 2021*, 56–64. ACM. https://doi.org/10.1145/3447535.3462511

Martins, A., and R. Astudillo. 2016. "From Softmax to Sparsemax: A Sparse Model of Attention and Multi-Label Classification." In *Proceedings of the 33rd International Conference on Machine Learning, Vol. 48 of Proceedings of Machine Learning Research*, edited by M. F. Balcan and K. Q. Weinberger, 1614–1623. Cambridge, MA: PMLR.

Mooseder, A., C. Brantner, R. Zamith, and J. Pfeffer. 2023. "(Social) Media Logics and Visualizing Climate Change: 10 Years of #Climatechange Images on Twitter." *Social Media + Society* 9 (1). https://doi.org/10.1177/20563051231164310

Muise, D., L. Yingdan, J. Pan, and B. Reeves. 2022. "Selectively Localized: Temporal and Visual Structure of Smartphone Screen Activity across Media Environments." *Mobile Media & Communication* 10 (3): 487–509. https://doi.org/10.1177/20501579221080333

Oak, E. R., and R. A. Nielsen. 2024. How the Rhetoric of Women in the Alt-Right Broadens the Movements Appeal." https://www.mit.edu/~rnielsen/altright.pdf

Ojala, T., M. Pietikäinen, and D. Harwood. 1996. "A Comparative Study of Texture Measures with Classification Based on Featured Distributions." *Pattern Recognition* 29 (1): 51–59.

Peng, Y. 2021. "What Makes Politicians Instagram Posts Popular? Analyzing Social Media Strategies of Candidates and Office Holders with Computer Vision." *The International Journal of Press/Politics.* 26 (1): 143–166. https://doi.org/10.1177/1940161220964769

Peng, Y., I. Lock, and A. Ali Salah. 2024. "Automated Visual Analysis for the Study of Social Media Effects: Opportunities, Approaches, and Challenges." *Communication Methods and Measures* 18 (2): 163–185. https://doi.org/10.1080/19312458.2023.2277956

Peng, Y., Y. Lu, and C. Shen. 2023. "An Agenda for Studying Credibility Perceptions of Visual Misinformation." *Political Communication* 40 (2): 225–237. https://doi.org/10.1080/10584609.2023.2175398

Phan, D., N. Pradhan, and M. Jankowiak. 2019. "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro." Preprint, arXiv:1912.11554.

Piqueras, M., A. Segerberg, M. Magnani, M. Magnusson, and N. Sladoje. 2026. "Replication Data for: An Image is Worth K Topics: A Visual Structural Topic Model with Pretrained Image Embeddings." Version V1. https://doi.org/10.7910/DVN/WATYCB

Qian, S., L. Yingdan, Y. Peng, C. C. Shen, and X. Huacen. 2024. "Convergence or Divergence? A Cross-Platform Analysis of Climate Change Visual Content Categories, Features, and Social Media Engagement on Twitter and Instagram." *Public Relations Review* 50 (2): 102454.

https://doi.org/10.1017/pan.2026.10062

Quinn, K. M., B. L. Monroe, M. Colaresi, M. H. Crespin, and D. R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* 54 (1): 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Radford, A., et al. 2021. "Learning Transferable Visual Models from Natural Language Supervision." In *Proceedings of the 38th International Conference on Machine Learning, Vol. 139 of Proceedings of Machine Learning Research*, edited by M. Meila and T. ZhangIn , 8748–8763. Cambridge, MA: PMLR.

Roberts, M. E., B. M. Stewart, and D. Tingley. 2019. "Stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91: 1–40. https://doi.org/10.18637/jss.v091.i02

Roberts, M. E., et al. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* 58 (4): 1064–1082. https://doi.org/10.1111/ajps.12103

Scholz, S., N. B. Weidmann, Z. C. Steinert-Threlkeld, E. Keremoğlu, and B. Goldlücke. 2024. "Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes." *Political Analysis* 1–15. https://doi.org/10.1017/pan.2024.18

Segerberg, A., and M. Magnani. 2025. "Visual Digital Intermediaries and Global Climate Communication: Is Climate Change Still a Distant Problem on YouTube?" *PLoS One* 20 (4): 1–23. https://doi.org/10.1371/journal.pone.0318338

Sivic, and Zisserman. 2003. "Video Google: A Text Retrieval Approach to Object Matching in Videos." In *Proceedings Ninth IEEE International Conference on Computer Vision*, 1470–1477. New York: IEEE. https://doi.org/10.1109/ICCV.2003.1238663

Smeulders, A. W. M., M. Worring, S. Santini, A. Gupta, and R. Jain. 2000. "Content-Based Image Retrieval at the End of the Early Years." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (12): 1349–1380. https://doi.org/10.1109/34.895972

Stan Development Team. 2024. "*Stan Modeling Language Users Guide and Reference Manual*." Version 2.35.0. http://mc-stan.org/

Torres, M. 2024. "A Framework for the Unsupervised and Semi-Supervised Analysis of Visual Frames." *Political Analysis* 32 (2): 199–220. https://doi.org/10.1017/pan.2023.32

Torres, M., and F. Cantú. 2022. "Learning to See: Convolutional Neural Networks for the Analysis of Social Science Data." *Political Analysis* 30 (1): 113–131. https://doi.org/10.1017/pan.2021.9

van Noord, N. 2022. "A Survey of Computational Methods for Iconic Image Analysis." *Digital Scholarship in the Humanities* 37 (4): 1316–1338. https://doi.org/10.1093/llc/fqac003

Von Sikorski, C. 2022. "Visual Polarisation: Examining the Interplay of Visual Cues and Media Trust on the Evaluation of Political Candidates." *Journalism* 23 (9): 1900–1918. https://doi.org/10.1177/1464884920987680

Wang, A., L. Liu, H. Chen, Z. Lin, J. Han, and G. Ding. 2025. "YOLOE: Real-Time Seeing Anything." In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 24591–24602. New York: IEEE. https://doi.org/10.1109/ICCV51701.2025.02280

Wang, Y., and D. M. Blei. 2019. "Frequentist Consistency of Variational Bayes." *Journal of the American Statistical Association* 114 (527): 1147–1161. https://doi.org/10.1080/01621459.2018.1473776

Williams, N. W., A. Casas, and J. D. Wilkerson. 2020. *Images as Data for Social Science Research: An Introduction to Convolutional Neural Nets for Image Classification*, Elements in Quantitative and Computational Methods for the Social Sciences. https://doi.org/10.1017/9781108860741

Ying, L., J. M. Montgomery, and B. M. Stewart. 2022. "Topics, Concepts, and Measurement: A Crowdsourced Procedure for Validating Topics as Measures." *Political Analysis* 30 (4): 570–589. https://doi.org/10.1017/pan.2021.33

Yosinski, J., J. Clune, Y. Bengio, and H. Lipson. 2014. "How Transferable Are Features in Deep Neural Networks?" In *Advances in Neural Information Processing Systems*, Vol. 27, edited by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, 3320–3328. Red Hook, NY: Curran Associates, Inc.

Zeiler, M. D., and R. Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In Computer Vision – ECCV 2014, Vol. 8689 of Lecture Notes in Computer Science, edited by D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, 818–833. Cham: Springer.

Zhang, H., and Y. Peng. 2022. "Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research." *Sociological Methods & Research*. https://doi.org/10.1177/00491241221082603

Zosa, E., and L. Pivovarova. 2022. "Multilingual and Multimodal Topic Modelling with Pretrained Embeddings." In *Proceedings of the 29th International Conference on Computational Linguistics*, edited by N. Calzolari, et al., 4037–4048. Gyeongju: International Committee on Computational Linguistics.

Cite this article: Piqueras, M., Segerberg, A., Magnani, M., Magnusson, M. and Sladoje, N., (2026). An Image Is Worth *K* Topics: A Visual Structural Topic Model with Pretrained Image Embeddings. *Political Analysis*, 1–20. https://doi.org/10.1017/pan.2026.10062

https://doi.org/10.1017/pan.2026.10062