A Replicable Visual LLM Workflow for Emotion Coding in Political Events: Evidence from the 2019 Hong Kong Protests

Chi Shun (Gary) Fong
PhD, Department of Political Science, Pennsylvania State University
King-Wa, Fu
Professor, School of Future Media, The University of Hong Kong

Abstract

Existing research highlights the role of visual elements in triggering emotions such as anger and solidarity, thereby helping to overcome the high costs of repression in autocracies. However, large-scale studies on visually triggered protest emotions have remained scarce due to methodological challenges in coding these relatively subjective attributes, especially after the rise of social media, which has scaled the problem. The paper proposes using open-weight multimodal Large Language Models (LLMs) with visual analysis capabilities to code protest-related emotions in images. The results show that this LLM approach can achieve near-human-level accuracy in coding social media messages with images from the 2019 Hong Kong protests and can be applied to an alternative dataset of Twitter images related to the Black Lives Matter movement, retrieving most positive cases. Using mid-tier commercial-grade hardware and cloud computing services, these models can typically process thousands of images per day on a single machine. This paper explores the potential of using LLMs to annotate the semantic meanings of visual elements, which could be applied to the study of contentious politics and beyond, such as the mobilization effects of images, the emotional strategies of protest leaders, framing and misinformation strategies, and themes in online propaganda.

Keywords: social media, protest emotion, visual analysis, large language model

Introduction

Images play a vital role in political communication. They shape people's political attitudes and guide their behavior by providing information shortcuts for evaluating complex issues (Popkin, 1994) and by eliciting affective responses (Casas & Williams, 2019) that operate beyond narrowly rational calculation (Arceneaux, 2012). Images are particularly important to the study of activism, as visuals are essential to the expression of dissent (Mattoni & Teune, 2014) and help trigger emotions present at every stage of activism (Jasper, 1998). Yet, studies of the visual aspects of activism have remained scarce in sociology and political science, despite the centrality of social movements and contention in these disciplines (Mattoni & Teune, 2014). Most of the existing studies examine how images are utilized for facilitating social movements of various forms and causes, ranging from peaceful protest (Hoffmann & Neumayer, 2025; Vance & Potash, 2022), labor movement (Morrison & Isaac, 2012), environmental activism (Molder et al., 2022), uprising against authoritarianism (Gerbaudo, 2015; Lim, 2013), feminist and gender movement (Palczewski, 2005; Ramsey, 2000; Safaian & Teune, 2022) and far-right movement (Caiani, 2024), and representing these movements in the mainstream media (Arpan et al., 2006; Brown & Harlow, 2019; Corrigall-Brown & Wilkes, 2012; Cowart et al., 2016; Wetzel, 2012) by manually and selectively reading a small number of samples. While these earlier studies provide valuable insights into the potential effects of images on activism, they rarely systematically investigate how images are perceived by citizens *en masse* and how they causally impact their patterns of political participation in observational settings.

The rise of social media partially addresses this issue by providing a systematic collection of visual data with engagement for quantitative analysis, but it also introduces measurement challenges. On the one hand, social media allows users to post multimodal content beyond text, including audio, images, and videos, and provides objective metrics such as likes, shares, reactions, and other engagement indicators to study how visual content is perceived by other users, often mediated by social media algorithms (Neumayer & Rossi, 2018). On the other hand, the abundance of online data requires more efficient and accurate methods for studying image properties. While advances in computational methods like deep learning have enabled thousands of images to be processed within days, the process still entails significant technical barriers and computational costs. Further, these methods focus on image classification, object detection, and

face and person analysis (Joo & Steinert-Threlkeld, 2022), leaving the study of more subjective properties, like emotions (Casas & Williams, 2019; Hoffmann & Neumayer, 2025; Jenzen et al., 2020), which are the focus of this paper, largely reliant on qualitative reading of small, selected samples, as in the pre-internet and social media era.

We propose that recent advances in large language models (LLMs) with visual analysis capabilities may offer a more efficient and practical approach to image analysis, even for the relatively subjective task of emotion detection in protest settings. We demonstrate this using three open-weight models—LLaVA (Liu et al., 2023), Gemma3 (Kamath et al., 2025), and Qwen3-VL (Bai et al., 2025)—to analyze protest images related to the 2019 Hong Kong protests. These models represent mainstream medium-sized LLMs that can typically be supported by mid-level commercial-grade hardware or cloud computing services at an acceptable cost, with different advantages given their model architecture.[1] By revising the prompt iteratively based on the model outputs on the results, certainty of the estimation, and reasoning behind the label, the results show that these models can achieve near-human levels accuracy in identifying two major protest emotions, anger and solidarity, particularly the ability to correctly retrieve true cases, across 684 top-viewed human-labeled samples. This approach is validated internally by examining the explanations provided by LLMs and externally by studying another set of human-labeled protest images from the Black Lives Matter movement (Casas & Williams, 2019).

This paper advances the study of the visual dimension of activism by offering a hands-on, efficient, and replicable workflow for using LLMs to code protest images available on social media with near human-level accuracy, with the expectation of future improvement from newer models. By applying state-of-the-art medium-sized open-weight LLMs on mainstream cloud computing services, a single graphics processing unit (GPU), the main hardware that supports LLMs' computation, could typically code at least a few thousand images within a day, at an hourly cost of around 0.13 to 0.45 USD. Researchers could adopt this approach with minimal guidance if they possess basic programming literacy and access to online cloud computing services, without the need to engage in a complicated fine-tuning process. Furthermore, using these models locally instead of calling commercial LLM services via Application Programming

[1] See Appendix D for the comparison of model performance

Interfaces (APIs) makes results highly replicable (Barrie et al., 2024) and reduces the risk of privacy breaches (Balluff et al., 2026). Besides methodological advancement, this paper opens the potential to study substantive questions in contentious politics and beyond that require quantification of semantic meanings behind visual elements, such as the relationship between emotions and protest mobilization, the effectiveness of visual framing on misinformation, and the themes and strategies of propaganda and their persuasiveness.

Study of visually triggered emotions in contention

Emotions are broadly defined as socially recognizable patterns of felt response and disposition that transcend individual experiences under a specific societal context (Shah, 2024). Emotions are integral to all social actions. Emotions lead individuals to depart from their normal behavior and remain true to their most deeply held values and attitudes by paying less attention to available information and placing greater reliance on heuristics (Marcus, 2000). As Marcus (2000) noted, Lasswell (2009 [1948]) long ago held that politics is the expression of personal emotions. As such, the study of emotions has been applied across virtually all subfields of political science involving different actors, ranging from issues like voting behavior, partisan polarization, nationalism, and policy process (Marcus, 2000; Shah, 2024; Webster & Albertson, 2022)

Protest participation is no exception (Goodwin et al., 2000; Jasper, 1998; Van Troost et al., 2013). While overlooked by social movement theories that depict protesters as either irrational mobs (Le Bon, 2009) or rationally motivated actors (McCarthy & Zald, 1977), emotions are present at every stage of protest. For instance, prior to a protest, organizers often evoke anger among the public and lead enraged citizens to act against the actor they believe is responsible for their grievance (van Stekelenburg & Klandermans, 2013). After the outbreak of protest, group-based emotions like solidarity and enthusiasm help forge bonds among participants from different backgrounds and sustain their participation (Van Troost et al., 2013). More recent studies in social and political psychology challenge the assumption underlying the distinction between emotion and rationality (Aminzade et al., 2001; Pearlman, 2013; L. E. Young, 2019), demonstrating that emotions alter the utility citizens derive from protest

participation by changing their assessment of values, such as security and certainty, perception of reality, risk acceptance, etc.

In this regard, “emotional work” (Hochschild, 1979) – the cultivation of protest-enabling emotions (e.g., anger, solidarity, joy, etc.) and the suppression of protest-disabling emotions (e.g., fear, sadness, etc.) – is a key to organizing successful protests (Gamson, 1990; Jasper, 1998; Pearlman, 2013). Compared with text, visual elements can be perceived universally without prior knowledge (Joo & Steinert-Threlkeld, 2022). Further, visuals can be processed by and are processed more quickly by the brain (Popkin, 1994) and, more importantly, evoke stronger emotional responses than text alone (Casas & Williams, 2019). A large cross-national study of protest-related social media messages demonstrates that images and videos attract more audience engagement than their textual counterparts (Lu & Peng, 2024). In fact, visual elements are widely adopted in protest contexts, either through the explicit efforts of protest leaders or the co-production by online users in the social media era (Neumayer & Rossi, 2018).

While protest may involve more than 20 subtypes of emotions (Jasper, 1998), anger and solidarity are two of the most commonly discussed protest-enabling emotions (Goodwin et al., 2001; Pearlman, 2013)

Anger originates from grievances aroused by perceived injustice – “one is not receiving what one deserves” (Jasper, 1998; van Stekelenburg & Klandermans, 2013), which often involves a responsible out-group, notably institutional actors, as the perpetrator and a quest for the restoration of justice (Lambert et al., 2019). Following these studies, we defined anger in this paper as "reactive moral outrage against injustice or perceived violations of rights, typically targeting authorities to frame a clear 'common enemy."[2] It urges citizens to act out of moral outrage despite the risk of repression by making them more risk-averse and optimistic (Pearlman, 2013). Protest organizers often strategically evoke anger through highly visualized symbolic representations of unjustified violence (Olesen, 2015). During the Arab Spring, two well-recognized injustice symbols were Mohamed Bouazizi, a Tunisian street vendor with a college degree who set himself on fire after being repeatedly harassed by the police, and Khaled Said, an

[2] This definition is also employed in the prompt to instruct LLMs, see Appendix A.

Egyptian businessman who was beaten to death by the police after he exposed corruption of the police. Activists circulated their stories and images widely on Arab social media and employed their images as symbols of police brutality and more systemic injustice (Kharroub & Bas, 2016; Lim, 2013; Olesen, 2013). Similarly, in the U.S., activists and social media users spread social media memes and street art targeting police brutality, curating them as symbols of injustice in events like the UC Davis pepper spray incident (Bayerl & Stoynov, 2016) and the Black Lives Matter movement (Casas & Williams, 2019; Vance & Potash, 2022). Across the globe, political parties (Caiani, 2024; Hoffmann & Neumayer, 2025) and civil society organizations (Morrison & Isaac, 2012) frequently employ images to amplify and channel citizens' grievances towards particular targets.

Solidarity, on the other hand, is a group-based emotion that bridges protest participants from different social groups, claims, and preferred tactics (Summers-Effler, 2007) and even turns bystanders into sympathizers (L. Lu et al., 2025). It mobilizes citizens not only through direct persuasion but also by raising the expectation of mass participation (Chwe, 2013; Edmond, 2013). Often, solidarity is nurtured by creating a common identity that allows participants to look past their differences (Neufeld et al., 2019). Based on these attributes, we define solidarity as "positive collective emotion from shared identity, mutual aid, and hopeful commitment that sustains engagement across diverse protest factions."[3] On social media, protesters from the Egyptian revolution, the Spanish indignados, and Occupy Wall Street adopted inclusive, viral, and symbolic protest images as their profile pictures to showcase solidarity with one another, giving rise to a new form of online collective identification (Gerbaudo, 2015). Solidarity is especially important for marginalized groups to consolidate their in-group identification and attract alliances. Gay rights activists in Germany in the 1970s designed posters that encouraged gay men to display their homosexual orientation in public as a symbol of solidarity (Safaian & Teune, 2022). US Suffragists in the 1910s, on the other hand, used cartoons to cultivate solidarity between men and women, both of whom contributed to the war's victory (Ramsey, 2000). Similarly, their counterparts threatened that men would become feminized by women's suffrage in postcards (Palczewski, 2005). Solidarity is also intertwined with and mutually reinforcing of another protest emotion, enthusiasm, which emphasizes pride in the collective

[3] This definition is also employed in the prompt to instruct LLMs, see Appendix A.

identity and hopeful envisioning of the movement (Gerbaudo, 2016; Jasper, 1998). For instance, a study examining Greta Thunberg's Instagram reveals that nearly 80% of her posts invoked a hope narrative, an important component of enthusiasm (Molder et al., 2022).

Methodological challenges in detecting protest emotions in visual elements

In contentious politics and social movement studies, research on emotion and visual elements remains scant, despite substantive growth in the past two decades. One simple reason is that the role of emotions in protest was not recognized in earlier social movement theories. While pioneers like Le Bon (2009) framed protest as irrational, impulsive, and driven by a "herd mentality", dominant social movement theories developed from the 1970s, like resource-mobilization (McCarthy & Zald, 1977) and political process (McAdam, 1999), portray protesters as rational players and highlight the strategic deployment of resources. This paradigm posits that emotionality is incompatible with rationality, leaving little space for discussion of the role of emotions in social movements (Goodwin et al., 2000).

However, much of the inattention was also methodological. Measuring affective states accurately and reliably presents a core methodological challenge not only in the study of protest but also in social science in general, as researchers must balance operational scale against psychological depth. To measure individual emotional state at scale, they often employ surveys with structured self-reported instruments like multi-item Likert batteries, providing high precision and construct validity on respondents' conscious affects (Marcus et al., 2017; Watson et al., 1988). However, in protest contexts, it is challenging for researchers to follow the protests using survey methods – with questions ranging from human resources, the spontaneity of protest, identification of protestors, and establishing representativeness (Walgrave et al., 2016; Yuen et al., 2022). Building a representative sample after the protest is also difficult, not to mention concerns about reliability when protesters need to recall their memories (Bradburn et al., 1987) and are susceptible to preference falsification (Kuran, 1991). Occasionally, emotions could be extracted from a comprehensive set of text, such as party manifestos (Crabtree et al., 2020; Jung, 2020; Kosmidis et al., 2019; Widmann, 2021), political speeches (Cochrane et al., 2022; Kosmidis et al., 2019; Osnabrügge et al., 2021), and newspaper articles (L. Young & Soroka, 2012) , notably using dictionary-based lexicons or machine learning algorithms (Grimmer &

Stewart, 2013). For the study of protest emotion, this is less common due to the lack of systematic collection of text that reflects the emotional stage of protesters, though some studies have been successful in achieving this by analyzing discussions on online forums and social media (Shahin & Ng, 2022; Zhu et al., 2022). However, studying text alone ignores the many visual elements used in emotional mobilization. Even worse, relying solely on text may miss how cues operate implicitly to trigger psychological responses. A related famous case is the campaign material used by the Republicans in the 1988 U.S. presidential election that features Willie Horton, an imprisoned NBA player, to signal the stereotyped fear of Black Americans as criminals. Mendelberg's (1997) experimental study on the campaign material found that footage about Horton's case would activate underlying racial prejudice among voters but not the salience of the crime issue, and such effects would disappear when a more explicit cue was presented (Valentino et al., 2002). Such a conclusion would not be reached if these studies solely analyze the text in the campaign material, as both racial and issue cues are presented simultaneously.

For the study of visually triggered protest emotions, researchers manually read protest materials and media representations of protests to infer the intended emotional effects on protesters or the public (Mattoni & Teune, 2014). This approach relies on subjective interpretations and creates observational data points that cannot reliably establish causal connections between emotional and protest-related outcomes, such as success and failure. Further, although manual reading of visual materials provides high-quality, thick descriptions of images, interpretations are also prone to selection bias originating from sampling error.

Table 1. Comparison of Image Analysis Methods

Method	Qualitative Reading	Neural Network Based Models
Speed	Slow	Could be fast, dependent on computational power
Cost	Generally Low, dependent on sample size	High set-up cost, dependent on availability of cloud computing services
Technicality	Relatively simple	High
Sample Size	Relatively small	Large sample, contingent upon computational resource

Quality	High but prone to selection and personal biases.	Dependent on the properties to be coded, and quality of human coders if required
Research Focus	Framing in selected media outlets and protest material	Objective features of images, summary statistics, and audiences' receptivity

The digital era has made it easier to study visually triggered emotions. Social media platforms are increasingly visual, allowing users to share multimodal content such as images and videos. At the same time, researchers have unprecedented access to these data through platform APIs, web scraping, and commercial data providers, enabling large-scale analyses of visual political communication. As a result, much of the recent research on visually triggered protest emotions relies on social media data rather than traditional print sources (Caiani, 2024; Casas & Williams, 2019; Gerbaudo, 2015; Hoffmann & Neumayer, 2025; Kharroub & Bas, 2016; Lim, 2013; Molder et al., 2022; Olesen, 2013). More importantly, social media provides behavioral measures of audience engagement, such as likes, shares, comments, and reposts, which serve as observable proxies for responses to visual content (Neumayer & Rossi, 2018). For example, Casas and Williams (2019) analyzed Twitter activity surrounding the Black Lives Matter movement and found that tweets containing images—particularly those evoking enthusiasm and fear—were more likely to be retweeted, providing observational evidence that emotionally evocative visuals are associated with greater online political participation.

However, the abundance of available data also creates new challenges. Manually coding hundreds of thousands of images is labor-intensive and often requires large teams of coders, including crowd-sourced workers from platforms such as Amazon Mechanical Turk (MTurk). Without adequate training and quality control, this approach can reduce both the reliability and validity of the resulting data (Lind et al., 2017), particularly if bot or fraudulent accounts are recruited (Wang et al., 2024). For example, Casas and Williams (2019) trained undergraduate research assistants to code only the 1,000 most-retweeted images, while the remaining images were labeled by 1,259 crowdsourced workers. Although a diverse pool of coders may improve the validity of image interpretations, emotional responses to images are inherently subjective, making consistent labeling difficult. Even among the trained research assistants, the intraclass

correlation coefficients for emotion coding ranged from 0.19 to 0.54, indicating only low to moderate agreement (Shrout & Fleiss, 1979).

Thus, social scientists increasingly adopt advanced computational methods—notably neural network-based models—to analyze large-scale image datasets. For instance, convolutional neural networks (CNNs), the primary models used in computer vision, recognize local spatial patterns by first converting image pixels into matrices of color intensity. The network then scans these matrices by sliding a filter across the image to extract low-level features, ultimately passing these outputs through successive computational layers to assemble smaller features into complex, large-scale objects (Torres & Cantú, 2022). These models can perform a few common computer vision tasks, including image classification, object detection, and face and person analysis (Joo & Steinert-Threlkeld, 2022). A handful of recent studies apply these advanced models to study protest-related images, such as detecting protest events (Scholz et al., 2025; Sobolev et al., 2020; Steinert et al., 2022; Won et al., 2017; Zhang & Pan, 2019), identifying protest themes (Kim & Bas, 2023; Y. Lu & Peng, 2024; Zhang & Peng, 2024), measuring violence (Rossi et al., 2023; Steinert et al., 2022; Won et al., 2017), and analyzing facial expressions (Prasad & Steinert-Threlkeld, 2025).

Fig. 1 Examples of images analysis in selected studies

a) Fig. 5 from Zhang & Pan (2019) – using image and text from social media messages for protest detection

b) Fig.1 from Steinert-Threlkeld et. al (2022) – rating on presence of protest (A), state violence (B) and protester violence (C) based on image features

These studies demonstrate that researchers can process large numbers of images in a reasonable period of time using commercial image analysis services such as Google Vision, Microsoft Face, and Amazon Rekognition, or pre-trained open-source models such as VGGNet (Simonyan & Zisserman, 2015) and ResNet (He et al., 2015). Trained on millions of annotated images, these models can identify thousands of attributes, including persons, objects, scenes, and faces, enabling researchers to generate summary statistics and observe patterns in objective features of visual elements (Peng, Lock, & Ali Salah, 2024). Even when researchers need to analyze images for more specific purposes, such as assessing the level of violence or detecting protest events, they may annotate a small sample of images and use it to update the parameters of existing models, a technique which is called fine-tuning. For instance, Scholz et. al. (2025) employ commercial-grade computers to fine-tune existing models for the detection of protest events from around 140,000 protest-related images. The fine-tuning process took 15 to 30 hours for each model, while the actual labeling process took an hour. Similarly, Zhang & Peng (2024) built their own high-performance computer to categorize 300,000 images using different approaches, with model fine-tuning times ranging from 5 minutes to 24 hours.

These studies face several challenges, however. First, the attributes they label remain objective, relying on features extracted from images. Yet, they can hardly be used to evaluate the semantic meaning of images, the abstract ideological concepts, and the subtleties entailed in a specific social-historical context that are more directly linked to theoretical concepts (Peng, Lock, & Ali Salah, 2024). For instance, a computer vision model can easily detect an everyday object like a yellow umbrella or a three-finger salute from The Hunger Games, but it could not capture how these non-violent elements carry dense symbolic meaning that signals resistance and solidarity in protest movements across Asia (Lertchoosakul, 2021; Veg, 2016). One notable exception is Prasad & Steinert-Threlkeld (2025), who classify facial expressions in protest images as proxies for protest emotions to examine how emotions influence mobilization. However, even this deeper attempt to extract affective states from human faces still falls short of measuring genuine protest emotions. A facial expression, such as a grin, could mean sarcasm or laughing out of anger toward the state, solidarity, or joy toward fellow protesters, depending on context or the actors involved. For instance, a laughing police officer when teargassing protesters, or a politician smiling when responding to protest demands, triggers anger in protesters out of injustice rather than joy.

Fig.2 Example of images the emotion triggered goes against facial expression in 2019 Hong Kong protests

a) Hong Kong Police smiled when teargassing protesters

b) Carrie Lam, the Chief Executive (Mayor) of Hong Kong in 2019, responded to journalist during press conference

Second, these approaches focus on images' facial features without referencing the specific context from which these images are drawn. While neural network models can identify visual features, this does not necessarily mean that contextual information can be input into the analysis process. Zhang & Pan (2019), as an exception, use both the message text and the corresponding image to detect protest but treat text and visual analysis as two distinct processes. In our case of detecting protest emotion, analyzing the message text is particularly useful, as automated text analysis is a well-established method of sentiment analysis, especially machine learning methods such as neural network models (Rudkowsky et al., 2018; van Atteveldt et al., 2021). This drawback is more theoretical than methodological. When users scroll past a protest image, carefully crafted textual context anchors, redirects, and/or amplifies their emotional response in ways that cannot be captured solely with the image's pixels. This interplay between text and image is even more critical for analyzing protest, where the text is not an accessory but an integral part of the image itself.

Finally, while online tutorials for applying advanced neural networks are now widely available, some technical barriers remain. First, researchers must possess a solid background in neural network architecture and machine learning algorithms, including selecting appropriate models, ensuring hardware compatibility with software platforms, and fine-tuning and training models for specific tasks. Second, the preprocessing and inference pipelines remain technical and tedious: converting raw social media graphics or posts to standardized image tensors, handling visual noise, and interpreting what the model's layers extract are still "black box" processes. Third, the economic start-up costs pose a steep barrier to entry; locally deploying or fine-tuning these image-processing neural networks requires high-end, dedicated GPU infrastructure that easily exceeds a $3000 USD investment[4], effectively gatekeeping these advanced visual methodologies for resource-constrained researchers, especially those from the developing world.

[4] Taking the set-up of Zhang and Peng (2024) as an example, they used two mid-tier GPU of Nvidia RTX 2080 with a Intel Core i9-9900K CPU. These hardware cost around 2,300 USD in 2022, and similar grade hardware would still cost around 2,000 USD in 2026. Other essential hardware including Motherboard, System Ram, Storage and Power Supply that are compatible with the CPU and GPU would cost at least 1,000 USD

To address these issues, we propose employing open-weight LLMs with visual capabilities to analyze protest images. This approach addresses the key challenges in visual analysis of protest social media. First, LLMs reason through their rich abstraction capabilities for understanding the semantic meanings behind images, especially when the models store large amounts of information relevant to specific cases, like the history of Black civil rights mobilization when studying images related to the Black Lives Matter movement, or the complex history of relations between Hong Kong and Mainland China for analyzing images related to the 2019 Hong Kong protests. These qualities suggest that LLMs could measure more subjective attributes of images, such as emotion, based on information and reasoning stored beyond the pixels (Peng, Lock, & Ali Salah, 2024). Second, some LLMs can integrate information for different modalities, such as text, visual, and audio, in their analysis. Text related to images, such as image captions or social media messages, provides useful context for understanding the semantic meaning behind them, especially for abstract images, thereby improving model performance (Zhang & Leung, 2026).

Finally, the application of LLMs is more hands-on when employing a commercial cloud computing service. The researcher needs only to write proper instructions, called a prompt, to guide the model through the training process, and revise it iteratively by evaluating the results. The steps can be completed online by loading the template code and using a commercial cloud computing service at an affordable price. For example, users could subscribe to Google Colab Pro Plus at a monthly rate of 50 USD. As illustrated in this paper, processing an image would take around 20 to 60 seconds on the platform, and a one-month subscription to the service could purchase enough computational power for analyzing roughly 6,000 – 45,000 images

Application: Analyzing images from the 2019 Hong Kong protests

Overview

This paper illustrates how to apply open-weight LLMs to analyze images from the 2019 Hong Kong protests to extract protest emotions. we use this case because activists employ both online and offline visual elements for mobilization (Fung, 2020). The messages were originally posted on Telegram - a social media cum instant messaging platform widely adopted by protesters in Hong Kong and in many other authoritarian regimes for its high security and privacy settings

(Herasimenka, 2022; Urman et al., 2021; Wijermars & Lokot, 2022). A core feature of Telegram is the channel, which allows administrators to spread messages to subscribers to facilitate tactical coordination. In early 2020, the author and other researchers at the University of Hong Kong collected around 2 million messages from nearly 1,000 protest-related Telegram channels (Teo & Fu, 2021). For this paper, we focus on a sample from a channel called @hkstandstrong_promo, which posted around 35,000 messages with images during the protests and was the core of the network of channels – it has been mentioned by around 600 channels and once attracted at least 160,000 subscribers. Similar to Casas and Williams (2019), we work on the most influential messages - 684 out of around 12,000 (~5%) manually coded images with view counts above 30,000. For the emotions, we focus on anger and solidarity, the two well-established protest emotions, as well as Joy. Joy is rarely discussed in the existing literature (Halperin, 2025; Ransan-Cooper et al., 2018), but has the potential to mobilize citizens because, like anger and solidarity, it makes protesters feel relieved, optimistic, and pleased about the movement's progress (Pearlman, 2013). We chose these posts because of their higher interrater reliability, so the evaluation of the LLMs' performance would not be biased by the poor quality or high subjectivity of human coders[5].

Practical Guide

Step 1: Prompt Writing

The first step is to write a prompt that instructs LLMs to perform image analysis, specifically to provide binary labels indicating whether each of the three emotions is present (1) or absent (0). We also asked the model to describe the image, provide a short explanation for each coding decision, and rate its certainty in each decision on a scale of 0 to 10. Following some best practices (Giray, 2023), we structured the prompt into three parts: 1) Task; 2) Context; and 3) Output, which was retained despite having rounds of revisions. First, we specify the task as emotion coding of images from the 2019 Hong Kong protests and provide the exact step-by-step instructions. Second, we provide additional context for the coding by listing general rules,

[5] The human coders are trained undergraduate students, and each image was coded by the agreement of two coders. If disagreement occurs, a third coder is assigned to make the final judgment. The three emotions, anger, solidarity, and joy achieved Fleiss' kappa values of 0.77, 0.84, and 0.62, respectively.

definitions, and cues for specific emotions, and examples. Finally, we detailed the model's output schema in JSON format to eliminate output errors.

Step 2: Setting up the LLMs environment

The second step involved setting up the environment to run the local LLMs. Two hardware environments were evaluated: a local high-performance computer and a Google Colab virtual machine with a Pro+ subscription. Rather than relying on API calls to cloud-based LLMs like Gemini and Claude, which pose data privacy risks and suffer from a reliability crisis due to constant updates (Balluff et al., 2026; Barrie et al., 2024), the LLMs were deployed locally. We used Ollama, an open-source framework that enables efficient execution of local models (Marcondes et al., 2025). It supports an extensive library of open-weight LLMs, allowing users to interact with the models locally in interactive notebook environments such as Jupyter and Google Colab. Three medium-sized LLMs with visual analysis capabilities were selected to strike an optimal balance between computational efficiency and task performance (Huang & Wang, 2025), including LLaVA:34B (Liu et al., 2023), Gemma3:27B (Kamath et al., 2025), and Qwen3-VL:32B (Bai et al., 2025).[6]

Step 3: Evaluation and Optimization

The next step is to run the models to analyze the image based on the prompt sequentially and evaluate the results accordingly. Just like the evaluation of standard machine learning tasks, we evaluate the performance of the models on three metrics: 1) Precision, 2) Recall, and 3) F1 score[7]. Precision measures how accurate a model is when making positive predictions, recall measures the model's ability to identify all positive cases in the dataset, and F1 is a balanced metric that combines the two. Then we extract results that do not match the manual labels and read the qualitative description of the image and explanation for each emotion code. In particular, we adopt uncertainty sampling in active learning by focusing on cases the model is least certain about, i.e., those with an uncertainty score close to 0.5 (Rouzegar & Makrehchi,

[6] In the LLM setting, the adjacent "B" indicates the billions of parameters, or variables, that a model uses to process information and make predictions. Larger models generally perform better but require high-tier hardware and consume much more energy. See Appendix D for the comparison of these models

[7] Precision = True Positives/ (True Positives + False Positives), Recall = True Positives/(True Positives + False Negatives), and F1 = 2 x (Precision * Recall)/(Precision + Recall)

2024). In a traditional supervised machine learning task, these most uncertain cases are labeled by human coders to retrain the model. For LLM inference, we instead updated the prompt by adding new rules and cues for each emotion, while keeping the update as general as possible. This practice helps avoid overfitting, a phenomenon in classical machine learning tasks where a model learns too many nuances rather than general patterns in the current data, leading it to perform poorly on new, unseen data (Giray, 2023).

Table 2: vLLMs Optimization Strategy by parts

Category	**Parameter / Variable**	**Impact on Runtime**	**Optimization Advice & Guidelines**
Model	Parameter Count	**Massive.** Scaling parameters linearly increases compute requirements in Phase 1 and memory bandwidth load in Phase 2.	Use the smallest model that passes the quality threshold. Medium-sized models spanning 20B to 40B provide a good balance on performance and efficiency (Huang & Wang, 2025), especially quantized versions that compute with lower numerical precision (Das et al., 2026).
Hardware	GPU Architecture & VRAM Bandwidth	**Massive.** Phase 1 speed scales with GPU's FLOPS and Phase 2 speed scales directly with VRAM Bandwidth	Set up the system with at least 24 Gigabyte of VRAM that could complete load the model with enough space for the data process, such as a physical Nvidia 3090, or Nvidia L4 or Nvidia A100 on Google Colab.
Prompt	Image Resolution	**Moderate:** Higher resolution images are split into larger number of visual patches for input in Phase 1	Scale down images to the minimum resolution needed, a 768*768 would be an optimal choice – taking up 200 to 800 tokens
	Prompt Formatting	**Moderate:** Longer prompts increase Phase 1 time and lead the model to iterate longer in Phase 2	Enforce direct output instructions and keep the prompt as neat as possible (normally 1 word takes up 1.5 token)
Hyperparameters	Context Window Size	**High:** Pre-allocates VRAM for the cache in the computation. Larger	Set it to comfortably fits your input and output size. Avoid

		size means heavier memory load in Phase 2	defaulting to maximum window limits
	Max Output Limit	**High:** Determines the maximum number of iterations in Phase 2	Set it as tightly as possible around your expected response length to prevent infinite loop
	Temperature	**Moderate:** Determines the amount randomness of the output in Phase 2	Set it low (0 – 0.2) to reduce the possible variation in the output to reduce the latency and increase the replicability

Beyond performance, researchers must evaluate computational efficiency when applying LLMs. LLMs' runtime is divided into two distinct computational phases, each governed by different hardware bottlenecks and token parameters—the fundamental units of data processing in LLMs. The first phase processes the input instructions and vision data and is limited by the GPU's floating-point operations per second (FLOPS) and the total length of the input tokens, which, in multimodal contexts, includes textual prompt tokens as well as visual tokens that scale with image resolution and granularity of the LLMs' visual encoders. The second phase generates the output by processing each output sequentially and is bound by the bandwidth of the GPU's memory. Other than this hardware constraint, the efficiency of this second phase is constrained by two factors: the maximum context window, which dictates cache size in GPU VRAM and increases per-token memory retrieval overhead, and the total output token size, which determines the number of iterations required to complete generation. Researchers should use a prompt with optimal size and enough clarity and adjust the maximum context window and total output token size accordingly. Table 3 summarizes how different settings affect LLM runtime.

Results

Performance

Table 3 evaluates the performance of three LLMs after three rounds of active learning across two conditions: plain image (evaluating the visual stimulus in isolation) and with message (evaluating the visual stimulus alongside contextual text or message). Performance is measured using Precision, Recall, and F1-score, with baseline ground-truth prevalence rates (T) provided for each emotion.

Table 3: Performance of the LLMs coding by emotions

Plain image	**LLaVA:34B**			**Gemma 3:27B**			**Qwen3-VL:32B**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Anger (T = 31%)	0.45	0.39	0.42	0.38	0.76	0.51	0.61	0.81	0.69
Solidarity (T = 65%)	0.71	0.58	0.64	0.74	0.80	0.77	0.83	0.80	0.81
Joy (T=14%)	0.35	0.18	0.24	0.46	0.38	0.41	0.38	0.56	0.45
With message	**LLaVA:34B**			**Gemma 3:27B**			**Qwen3-VL:32B**		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Anger (T = 31%)	0.54	0.50	0.52	0.43	0.84	0.57	0.62	0.79	0.70
Solidarity (T = 65%)	0.70	0.62	0.66	0.70	0.87	0.77	0.80	0.79	0.79
Joy (T=14%)	0.47	0.26	0.33	0.55	0.40	0.46	0.40	0.60	0.48

Overall, model performance varies significantly across emotions, contextual conditions, and underlying architectures, shedding light on the challenges and potential of using LLMs for automated analysis of the semantic meaning of visual elements.

First, providing textual context improves model performance, notably on both LLaVa:34B and Gemma3:27B for labeling Anger and Joy. This improvement was not as pronounced on Qwen3-VL:32B and for the labeling of Solidarity, possibly because there is less room for improvement. Second, the results demonstrate a pronounced performance hierarchy tied to both statistical prevalence and affective complexity. Solidarity—the most prevalent emotion in the benchmark—consistently yields the highest predictive accuracy across all models, peaking at an F1-score of 0.81 with Qwen3-VL:32B. In contrast, Joy (T = 14%) proves substantially harder to detect, topping out at an F1-score of just 0.48. While class imbalance partially accounts for this disparity, there could be a deeper socio-linguistic factor. High-arousal

social emotions like Anger and Enthusiasm often manifest through standardized visual gestures and explicit linguistic markers. Conversely, nuanced positive emotions like Joy are heavily context-dependent and culturally mediated, making them far more evasive for LLMs to reliably identify without rich contextual knowledge. The relatively poorer performance could also be attributed to the class imbalance of Joy, which will be further discussed in the validation section below.

Finally, performance gaps exist between models with different architectures[8]. Qwen3-VL:32B emerges as the most robust model, maintaining a strong balance between Precision and Recall for both anger and solidarity, with an F1 score > 0.70, a widely accepted threshold of a good classifier for a balanced dataset[9]. This is because this model is heavily trained on joint-modality analysis, a more fine-grained visual encoder, and more sophisticated reasoning capabilities (Bai et al., 2025). The drawback, as shown in Table 6, is that the model needs a longer runtime to capture more visual details and perform more detailed reasoning. And given the moderate imbalance of the emotion of Anger (T = 31%), having a lower precision and thus an F1 of around 0.50 to 0.60, as in the case of Gemma3:27B, is also considered acceptable and *en par* with recent studies using various machine learning methods to detect emotions in text (Demszky et al., 2020; Koufakou et al., 2024; Mohammad et al., 2018). Based on the results, the performance of LLaVA, which is better adapted to clear photos with obvious subjects, is deemed unacceptable. As emotion is a subjective attribute with ambiguity in human annotation, LLMs' performance is partially bounded by the performance of human coders (Xu et al., 2026). Considering the inter-coder reliability for the three emotions ranged from 0.62 to 0.82, the performance of Qwen3-VL:32B, and to a lesser extent Gemma3:27B, could arguably be regarded as close to human performance when textual data is provided, as in the human coding process.

Reliability and Replicability

[8] Refer to Appendix D for the comparison of different models

[9] In machine learning and statistics, a balanced dataset means that the classes or outcomes you are trying to predict are represented in roughly equal proportions.

Beyond predictive performance of individual models, we also explored the agreement between models as a proxy of reliability of results - in other words, are models making similar predictions? As reiterated in the literature, replicability of LLMs matters. So, we compared whether results from local HPC and cloud settings are similar.

Table 4: Inter-model and Intra-model agreement by emotions

Inter-model (Cohen's Kappa)	**Anger**	**Solidarity**	**Joy**
plain image	0.13	0.16	0.15
with message	0.22	0.18	0.25
Intra-model (Percentage Agreement)	**Anger**	**Solidarity**	**Joy**
LLaVA:34B: plain image	73%	69%	90%
LLaVA:34B: with message	71%	68%	89%
Gemma3:27B: plain image	64%	76%	85%
Gemma3:27B: with message	68%	84%	88%
Qwen3-VL:32B: plain image	83%	83%	87%
Qwen3-VL:32B: with message	84%	80%	86%

As shown in Table 4, inter-model agreement, measured via Cohen's Kappa, remains low across all emotions, ranging from 0.13 to 0.16 for plain images and slightly improving to 0.22 to 0.25 when incorporating contextual text messages. This divergence primarily stems from architectural disparities among model families. In Appendix C, we show that Gemma4:31B (Team et al., 2026), which used similar training logic and visual encoding to Qwen3-VL:32B, achieved moderate agreement with Qwen3-VL:32 B's results. In terms of intra-model consistency across execution environments, we focused on the percentage of perfect agreement to measure exact deterministic replicability. All models exhibited higher reproducibility, especially Qwen3-VL:32B (>80% across all emotions). This is especially impressive when the local HPC and Google Colab use different hardware and software configurations, despite using the same context window size and output token count.

Monetary and Time Cost

Table 5 benchmarks the inference runtime per image across three vision-language models under two distinct hardware setups: a local high-performance computing (HPC) environment equipped with an Nvidia RTX PRO 6000 Blackwell GPU (96GB GDDR7), and Google Colab cloud environments (using an Nvidia A100(40GB) for Qwen3-VL:32B, and an Nvidia L4(24GB) for LLaVA:34B and Gemma 3:27B).

Table 5: Runtimes in second per image by different models and settings

	LLaVA:34B	**Gemma 3:27B**	**Qwen3-VL:32B[10]**
HPC: plain image	16s	11s	37s
HPC: with message	16s	11s	42s
Google Colab: plain image	28s	22s	61s
Google Colab: with message	29s	22s	73s

Across both settings, Gemma3:27B was the fastest model, followed by LLaVa:34B, and then the best-performing Qwen3-VL:32B. Considering the performance of the Gemma model in Table 3, it balances performance and efficiency well. Appending textual message prompts created negligible overhead for Gemma 3:27B and LLaVA:34B but added noticeable additional processing time to Qwen3-VL:32 B, reflecting its deeper cross-modal attention mechanisms and reasoning ability. Across all models, the local HPC workstation consistently executed tasks almost 2 times faster than the cloud-hosted Google Colab, given the stronger computational power of the GPU in the HPC. Yet, one should note that GPUs have been severely marked up in recent years given the rising popularity of AI. Hardware capable of running these models can easily exceed 3,000 USD, not to mention that a high-end GPU like the Nvidia RTX PRO 6000 Blackwell in the HPC now costs over 10,000 USD. If the task could be optimized with a budgeted GPU that provides satisfactory performance, using a cloud computing service like Google Colab would be more cost-effective. For instance, a single Nvidia L4 GPU costs only around 0.14 USD per hour and could run nearly 4,000 images per day.

Validation

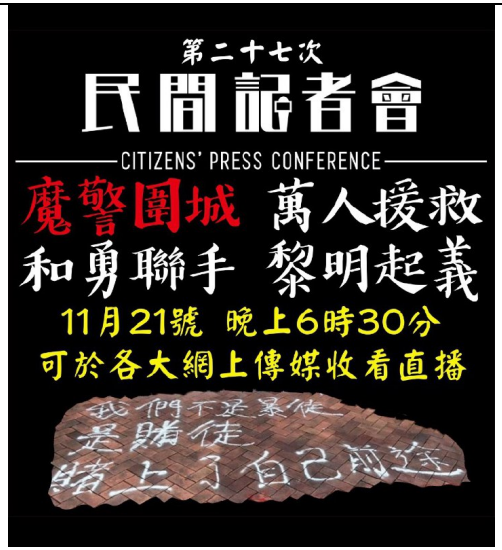
[10] For Qwen3-vl, we use longer context windows of 16,384 and 8,192 for the output token size. For the two models, we use a shorter context window of 4,192 and output token size of 1,024

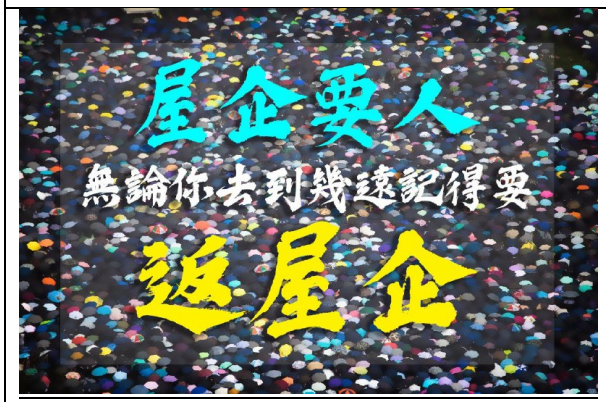
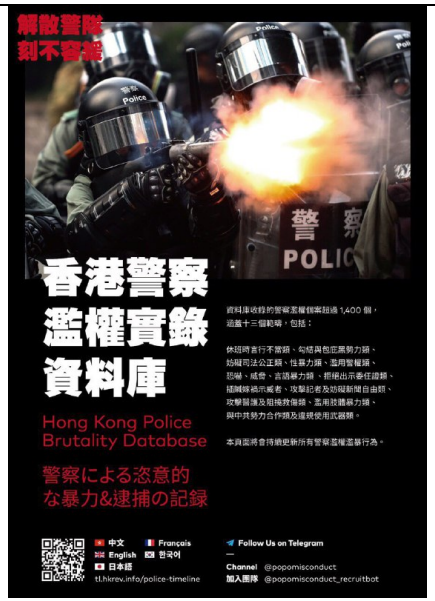
As stated in the literature (Grimmer & Stewart, 2013), machine learning results must be validated repeatedly. The following session discusses face validity, convergent and divergent validity, construct validity, and external validity, and ecological validity.

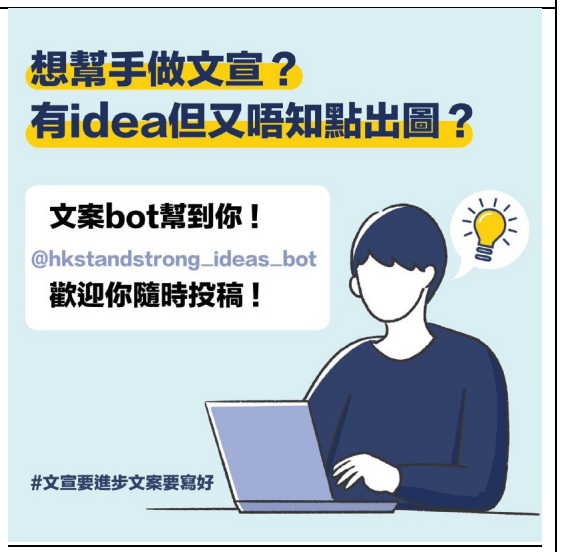
Face Validity

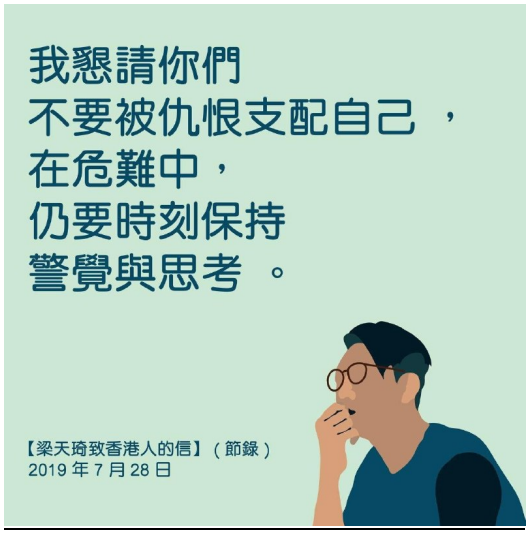
One way to evaluate the content validity of LLM output is to examine the results and assess whether they make sense (Birkenmaier et al., 2024). Thus, we randomly selected cases from the top-performing Qwen3-VL:32B model with text-message context for true-positive, true-negative, false-positive, and false-negative predictions. For the positive cases, they illustrate the model's ability to accurately describe the image, including text that may not be English, and support their decision with proper reasoning. In this regard, LLMs offer a unique advantage over average human coders by enabling applications beyond English-speaking countries. Regarding the negative cases, Qwen3-vl:32B maintains strong reasoning capabilities. Most mispredictions stem from either discrepancies in subjective judgment or subtle contextual knowledge that generally exceeds the scope of a typical large language model (LLM). False positives are primarily due to disagreements; while clues indicating anger and solidarity are present in the images, human coders may judge these cues as insufficiently explicit. Conversely, false negatives are often linked to more definitive details: in the anger image, the eye symbolizes a protester who lost sight due to police gunfire; in the solidarity image, the man is Edward Leung, a notable Hong Kong independence advocate respected by younger activists and jailed before the 2019 protests; and the upper section of the joy image attempts to connect a public indecency incident involving a mainland dancer soliciting tips from elderly citizens in parks with local protests advocating against mainland influence.

Fig 3. True positive and true negative prediction – Qwen3-VL:32B

a) True Positive		
Anger	Solidarity	Joy
		
Description: Black poster with white, red, and yellow text announcing a press conference, featuring protest slogans and a handwritten note at the bottom.	Description: A large crowd of protesters at night with many holding lights and American flags, overlaid with Chinese text announcing the US Senate passing the Hong Kong Human Rights and Democracy Act.	Description: Blue poster with six yellow icons and text listing practical items for protesters, emphasizing preparation and the five demands, with a P.S. encouraging sharing via Airdrop.
Reasoning: The image uses the term 'demon police' (魔警) to directly criticize the police, and the context of the event (in response to attacks and blockades by the regime) expresses moral outrage against authorities.	Reasoning: The image depicts a large, unified crowd and highlights international support (US Senate passing the bill), which are key cues for solidarity (shared identity, collective strength, and international backing).	Reasoning: The playful P.S. ("Go crazy and Airdrop...") and lighthearted phrasing (e.g., "Save every dollar, get every dollar") and fun, cartoonish design (with the 'six treasures' as a game-like list) indicate joy and playfulness.

b) True Negative		
Anger	Solidarity	Joy
		 香港人日程表APP，上線啦~~~ 追蹤最新香港民主自由運動，點擊下載支持前線，目前僅提供Andriod版本。 http://app.hkrevolution.club/HKcalander.apk
Description: Colorful confetti background with three lines of text in blue, white, and yellow: '屋企要人', '無論你去到幾遠記得要', '返屋企' – Go back home no matters how far you are away	Description: A poster showing a police officer firing a weapon, with text in multiple languages about a 'Hong Kong Police Brutality Database' and details of alleged police misconduct.	Description: A mobile app interface for a protest calendar showing event listings and a December calendar, promoting the app for tracking pro-democracy protests.
Reasoning: The image does not contain direct criticism of authorities, documentation of misconduct, or framing of systemic injustice; it is a unifying call without anger.	Reasoning: The image focuses on documenting police brutality and does not explicitly show or reference collective protest unity, mutual aid, or shared identity among protesters. The database is a tool for accountability but lacks direct expressions of solidarity.	Reasoning: The image lacks playful, satirical, or humorous elements; the design is serious and informational, with no evidence of joy.

c) False Positive		
Anger	Solidarity	Joy
		
Description: A large crowd of people in black, many with faces covered, with overlaid text in Chinese and English calling for a protest on October 1st with full gear and unity.	Description: Purple background with two lines of black Chinese text: top line bold and larger, bottom line smaller. Text is in traditional Chinese font.	Description: Light blue background with cartoon of person using laptop. Text in Chinese promotes a bot for submitting propaganda ideas. Lightbulb icon present.
Reasoning: The text explicitly references 'fearless of suppression, arrest, terror attack', which is a direct reaction to the authorities' actions (suppression and arrests) and expresses moral outrage against the government and police.	Reasoning: The image is a guide for the 'struggle' (a collective term for the movement) and the post text includes #手足 (comrades), indicating shared identity and mutual aid within the protest community.	Reasoning: The image uses a cartoon style, a lightbulb icon (a playful symbol for an idea), and a lighthearted, conversational tone to encourage participation, which serves as a form of comic relief and morale-boosting.

d) False Negative		
Anger	Solidarity	Joy
		
Description: A hand with an eye in the palm, text 'What did it cost?', and below in red: 'Hong Kong International Airport 8.12'.	Description: Light green background with dark blue Chinese text and an illustration of a man with glasses in a thoughtful pose on the right.	Description: A protest poster for two 9.21 events: 'Recover Tuen Mun' and 'Yuen Long Sit-in', with a bottom photo of a crowd and the slogan 'Liberate Hong Kong, Revolution of Our Times, Five Demands, Not One Less'.
Reasoning: The image does not contain direct criticism of authorities, documentation of misconduct, or framing of systemic injustice; it is a reflective question without angry cues.	Reasoning: The text is a cautionary message to avoid hatred and maintain alertness, but it lacks explicit calls for mutual aid, shared identity, or hopeful commitment as defined for solidarity.	Reasoning: The image lacks any playful, satirical, or lighthearted elements. The top image has some smiling but not in a protest context, and the bottom image is serious. No memes or humor are present.

Convergent and Divergent Validity

Next, to evaluate whether the LLMs reliably capture distinct affective constructs, we analyze the divergence and convergence of constructs in the model's generated qualitative explanations. By computing cosine similarity between text explanations for different emotions, this tests the conceptual boundaries of the model's classifications, ensuring that its multimodal reasoning maintains clear semantic separation between distinct emotions. we again focused on the best-performing Qwen3-vl:32B model.

Table 6: Cosine similarity of explanation across emotions (Qwen3-VL32B with message)

Emotion Pair Comparison	Mean Cosine Similarity	High Overlap Cases (>0.85)	High Overlap Share (%)
Solidarity vs. Anger	0.5066	116 / 684	16.96%
Solidarity vs. Joy	0.4885	116 / 684	16.96%
Anger vs. Joy	0.5000	115 / 684	16.81%

The results confirm that the LLM maintains a healthy conceptual separation, yielding an average explanation cosine similarity of about 0.50 across all emotion pairings. A similarity score of 0.50 suggests that the model captures the shared domain context of the 2019 Hong Kong protests while maintaining distinct semantic trajectories for each emotion. Furthermore, high-overlap cases were observed in approximately 17% of the dataset. This accurately reflects the data's complexity, as visual elements in protests could allow multiple emotional framing strategies to coexist, such as simultaneous moral outrage against authorities and communal solidarity.

To evaluate the construct validity of the LLMs' output, we apply BERTopic to the model's generated image descriptions grouped by each emotion subset. The rationale is that if the model is truly detecting distinct emotions, the underlying visual motifs extracted should reflect theoretically distinct cues for each emotion. BERTopic accomplishes this by first converting text descriptions into vector embeddings and then clustering semantically similar descriptions. Table 7 again shows the results from Qwen3-VL:32B with message. Table 7 provides robust empirical

support for construct validity, demonstrating that the visual pipeline successfully differentiates the thematic content unique to each emotion. Some discovered topics, however, consist of generic background descriptions rather than emotion-specific cues and do not necessarily match the emotions.

Table 7: BERTopic (Qwen3-VL:32B with message)

Emotion Construct	Topic ID	Count	Discovered visual motifs (Keywords)
Solidarity (N = 442)	0	35	Protest demands (demands_protest_airport_protest poster)
	1	35	Rights framing (human rights_rights_poster_human)
	2	27	Uncertain (chinese_black_white_line)
	3	24	Schedules and locations of protests (events_schedule_2019_locations)
	4	19	Uncertain (cartoon_hats_protester_details)
Anger (N = 272)	0	82	Police and protester confrontations (police_protesters_officer_showing)
	1	18	Protective gear & protest slogan (mask_gas mask_gas_liberate)
	2	17	Uncertain (chinese_line_person_black)
	3	15	Contextual Markers (kong_hong kong_hong_2019)
	4	13	Criticism (poster_protest poster_protest_criticizing)
Joy (N = 140)	0	37	Visual Elements (chinese_police_red_white)
	1	20	General Event Context (2019_protest_hong kong_kong)
	2	15	Cartoons and gear (cartoon_protester_hard hat_hard)
	3	12	Movement claims (demands_protest_poster_protest poster)
	4	12	Assembly site (poster_edinburgh_place_edinburgh place)

Construct validity

We demonstrate the construct validity of my workflow and that it performs better than similar measurement methods by comparing LLM outputs with those of a facial emotion extraction model and a text-based BERT model.

First, we follow Prasad & Steinert-Threlkeld (2025), who study the relationship between online protest emotion and mobilization, to extract emotional facial expressions from the data using the Python package Py-feat (Jin Hyun Cheong et al., 2024) – an open-source package that integrates different machine learning classifiers and neural network models that could be applied to static images and videos. The Py-feat package allows extraction of a range of emotions and related facial attributes, including anger and happiness, which is closely related to Joy. Table 8 below shows the precision, recall, and F-1 score for applying the model to the 684 coded images from the 2019 Hong Kong protests, using different emotion score cut-off thresholds ranging from 1 to 0. The results show that the model could predict anger with moderate precision but not Joy, and that recall for both emotions was extremely low. While the extremely low level of recall is potentially caused by a lack of real human faces in the images included in the samples, the low to moderate level of precision, especially for the more subjective emotion of Joy, suggests the deep semantic meaning behind protest images could not be simply extracted through identification of facial expressions.

Table 8: Performance of emotional facial extraction model (Py-feat)

Threshold = 0.5	Precision	Recall	F1
Anger (T= 31%)	0.53	0.05	0.08
Joy (T= 14%)	0.16	0.05	0.08
Threshold = 0.3	Precision	Recall	F1
Anger (T = 31%)	0.56	0.06	0.12
Joy (T= 14%)	0.14	0.08	0.11

To provide another benchmark for LLMs' performance, we also ran a BERT model (Devlin et al., 2019), specifically the mDeBERTa-v3-base, which supports multilingual analysis, to extract emotions from the text messages in the images. Because images without embedded text were excluded, the effective sample size was reduced to 484 observations. we used an 80/20

train-test split across 5 cross-validation rounds, each with a different random seed. Across the 5 runs, the fine-tuned text-only model excelled at detecting dominant and explicit categories—achieving an average F1-score of 0.83 for Solidarity, outperforming even the best-performing multimodal model, Qwen3-VL with text message (F1 = 0.79). This superior performance likely stems from the fact that solidarity-oriented images frequently contain explicit calls to action and administrative details regarding protest logistics, which are easily captured via text alone. For Anger, mDeBERTa-v3 achieved a solid F1-score of 0.56, though it still fell short of Qwen3-VL with text message (F1 = 0.70). However, the text-only model struggled significantly with Joy (F1 = 0.15, Recall = 0.09), in contrast to Qwen3-VL with text message (F1 = 0.48, recall = 0.6). All in all, LLMs maintained far higher recall and more balanced performance across all three emotion labels. This indicates that zero-shot LLMs possess superior contextual understanding when interpreting subjective visual attributes, particularly given that roughly 30% of the original visual corpus contained no text at all.

Table 9: Performance of the BERT model on text messages

80% as training data (N= 97)	Precision	Recall	F1
Anger (T ~29%)	0.74	0.46	0.56
Solidarity (T ~56%)	0.78	0.88	0.83
Joy (T ~16%)	0.56	0.09	0.15

External Validity

For external validation in an alternative context, we apply the above workflow to the Casas & Williams (2019) dataset using the best-performing model, Qwen3-VL:32B. We worked on the top-retweeted 866 images, which were coded by trained undergraduate coders, in addition to MTurk workers who were employed to code. This ensures higher coding quality and, thus, reduces bias in estimating model performance. We focus on the more prevalent and theoretically discussed protest-enabling emotions of anger, sadness, and solidarity-related enthusiasm.[11] As the original dataset is coded on a scale of 0 to 10, we binarized the data into 0 and 1 with a

[11] We also combined anger and disgust in the data as they are highly theoretically related.

passing threshold of 2, given that the emotional scores were extremely right-skewed and zero-inflated with a mean and median of less than 2. Table 10 presents the results of the model after three rounds of active learning.

Table 10: Performance of the LLM coding by emotions – BLM

Plain image – Qwen3-VL:32B	**Precision**	**Recall**	**F1**	**Runtime per image**
Anger (T = 43%)	0.59	0.71	0.65	
Sadness (T = 40%)	0.56	0.43	0.49	34s
Enthusiasm (T = 41%)	0.49	0.67	0.56	
With message– Qwen3-VL:32B	**Precision**	**Recall**	**F1**	**Runtime per image**
Anger (T = 43%)	0.61	0.71	0.66	
Sadness (T = 40%)	0.56	0.42	0.48	52s
Enthusiasm (T = 41%)	0.51	0.66	0.57	

As shown in Table 10, the model performance is slightly weaker than the application on the 2019 Hong Kong protests. The classification of Anger and Enthusiasm is marginally acceptable with an F1 of 0.66 and 0.57 when text messages are included in the data. Substantively, high-arousal mobilizing emotions like Anger and Enthusiasm continue to yield stronger overall performance than Sadness, confirming that distinct visual cues remain easier for LLMs to identify than lower-intensity, subtle emotional signals.

The lower accuracy, however, does not necessarily indicate the inherent incompetence of LLMs in alternative contexts. Instead, the results stem from a combination of platform dynamics, data curation strategies, and human annotation biases. First, the initial Hong Kong protest dataset was drawn from a highly curated, pro-protest Telegram channel specifically designed for political mobilization and "emotional work". These images were explicitly constructed with high signal, emotionally charged iconography, making emotion detection straightforward. By contrast, this application relied on individual Twitter posts from Casas & William (2019) gathered via broad keyword queries with noisy visual signals. This difference in data curation also resulted in a severe zero-inflation of emotions, as they are less prevalent on Twitter than in the specially tasked Telegram channel. From a statistical standpoint, zero-inflation

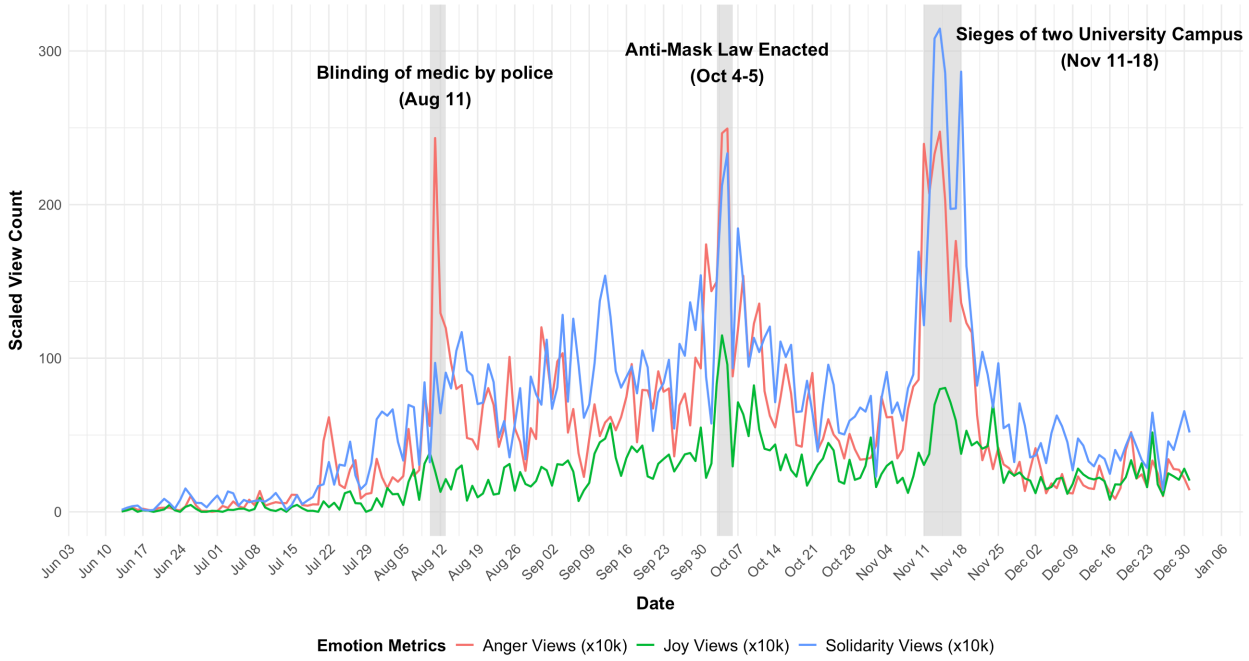
disproportionately penalizes precision by increasing the likelihood of false positives, a component in the denominator of the decision formula. This vulnerability is further compounded by human annotation errors. When relying on crowdsourced workers (e.g., MTurk) for subjective visual tasks, inter-coder reliability tends to be lower. As demonstrated in Monte Carlo simulations by Song et al. (2020), lower inter-coder reliability is more likely to lead to a high false-negative rate. It makes sense as low-effort or uncertain coders rarely assign positive and negative labels with equal probability - systematically defaulting to zero is a safer, harder-to-detect option. Finally, emotional coding is inherently subjective and conditioned by coders' perception of the issue. For a more complex issue with diverse online discussion, like BLM, stronger disagreement over emotional interpretation is possible.

However, there are still ways to improve the model's performance even when the data is noisy, coders' agreement is low, and coding quality is concerning, though the space is limited. First, researchers can leverage the model's self-reported certainty scores to filter out ambiguous predictions and discard low-confidence inferences where visual or textual noise is high to assign a default score, or to continue the active learning process. Second, incorporating a few well-chosen "few-shot" examples (illustrating hard positives and confusing negatives) directly into the prompt text can firmly anchor the model's classification boundaries without requiring any complex architectural fine-tuning. Finally, implementing a lightweight consensus or majority-voting approach by running inference across two models can help smooth out stochastic errors and mitigate the impact of unreliable human baseline annotations.

Ecological Validity

While we have provided evidence of the models' validity across different dimensions, evaluating ecological validity remains necessary. Specifically, does the automated visual emotional coding accurately track the real-world trajectory of the 2019 Hong Kong protests?

Fig. 4 Emotional Trends Over Time with Key Events in 2019 Hong Kong Protests



To examine the temporal pattern of visual emotions, we applied the Qwen-VL model to all messages with images posted on the Hong Kong Stand Strong Promote Telegram channel from early June to late December 2019 and calculated the daily aggregated view counts for each emotion. Because the dataset contains over 30,000 images, we treat the message as the unit of analysis and apply the model exclusively to the first image per message. This design assumes that online users' attention focuses primarily on the first image, which also captures the dominant emotional frame. Furthermore, we filter out messages forwarded from external users or channels, as their view metrics reflect reach specifically within this channel rather than across the entire online network, making them systematically lower than those of messages natively posted by this channel.[12] The final dataset contains only 14,354 messages with images. The final analytical sample consists of 14,354 original image-containing messages. Figure 1 displays the resulting temporal patterns of visual emotions alongside major protest milestones.

As shown in Figure 4, the visual emotional dynamics generated by Qwen-VL closely align with critical offline turning points, providing strong ecological validation. First, anger and

[12] Telegram's architecture allows view counts of messages originated from a particular user, group, and channel to be aggregated when displayed in their source. The forwarded messages have an average view count of 407, while the original messages have an average view count of 15,002.

solidarity show high contemporaneity with similar magnitude, while Joy shows a much weaker trend, confirming the dominant role of anger and solidarity in protest emotions. Second, the spikes in emotions correspond to three major milestones in the movement: Anger spiked on Aug 11, matching the blinding of a female first-aider by police projectiles; the enactment of the anti-mask law in early October by the government triggered serious clashes among citizens and mobilization; and the same applies to the sieges of two university campuses by the Hong Kong police in mid-November. Finally, it shows the surge of emotions at the very beginning of the movement and their gradual decay after November, as citizens become increasingly fatigued after a prolonged period of mobilization, especially amid an outbreak of a then-unknown plague of COVID-19. Overall, the tight alignment between LLMs-detected emotional trends and protest patterns confirms that the model's classifications reliably mirror the real-world affective trajectory of the protests.

Discussion

This paper illustrates how state-of-the-art open-weight LLMs with visual capabilities could meet the growing demand for multimodal data analysis, driven by the rise of social media, in both accuracy and efficiency. By offering a way to quantify emotions in protests from systematically collected social media data, this study opens a new line of research on the causal relationship between emotions and dissent, a relationship long suggested in the literature but rarely examined beyond experimental settings (Geise et al., 2021; L. E. Young, 2019), and on related questions, such as how protest organizers manage protest emotions alongside state repression. By demonstrating how to annotate semantic meaning in visual elements with LLMs, this paper also unleashes the potential to study beyond contentious politics at scales enabled by visual data. For instance, Peng et al. (2023) suggest going beyond objective features when studying visual misinformation by examining subjective attributes that LLMs can assess, such as image framing and credibility, and the reasoning behind them. Similarly, LLMs could be applied to the study of propaganda, as in Lu et al. (2026), to examine the themes of propaganda videos on Chinese social media beyond small human-annotated samples. This is especially important when popular social media platforms, such as TikTok and Instagram, prioritize visual content over pure text, and memes, short-form video frames, and infographics have become the primary vehicles for political communication(Peng, Lock, & Salah, 2024).

Of course, this paper also has a few limitations. First, the current workflow relies heavily on iterative prompt engineering to optimize inference. Although the results stabilize after 3 rounds of active learning, they may still be sensitive to subtle changes in the input process, such as instruction phrasing, temperature, and context length, which may hinder true out-of-domain generalizability. Researchers with stronger training in machine learning should explore Low-Rank Adaptation (LoRA): fine-tuning LLMs with small, high-quality human-annotated samples can permanently encode domain-specific affective cues directly into weight adapters without compounding runtime costs with a long list of instructions. Second, while active learning via uncertainty sampling refines prompt rules for a specific movement, tailoring rules around edge cases risks overfitting prompt contexts to localized political symbols. Researchers should keep that in mind when revising the prompt. Third, the current workflow focuses on single-model performance and is prone to bias. In the future, researchers could consider formalizing multi-model ensembles to build a more representative decision boundary, improving the reliability and precision of the measurement.

Despite LLMs' ability to analyze visual elements efficiently and with fair accuracy, researchers must still consider several factors when deploying LLMs locally, especially the choice between setting up a higher- performance computer (HPC) and using a cloud computing service.

First, consider the time and monetary costs. Cloud computing services offer a more flexible solution – you can access them as long as you have an internet connection, adjust your usage based on your computational needs, and do not need to maintain HPC day-to-day. By contrast, setting up an HPC requires a certain level of knowledge of hardware and system operation, and a fixed cost that easily exceeds 3,000 dollars. With 3,000 dollars, you can get 30,000 compute units on Google Colab, equivalent to around 20,000 hours of runtime on an Nvidia L4 GPU, or 5,500 hours of runtime on an Nvidia A100 (40GB) GPU. These amounts of runtime should be enough to analyze at least 600,000 to 240,000 images using medium-sized LLMs, assuming each image takes around 30 seconds to process. Moreover, you need not worry

about updating the hardware to fit the needs of newer models, as commercial cloud computing services will update their hardware from time to time.

Another concern is replicability. Although LLMs have shown low output variance when running locally with dedicated settings, such as temperature, they remain sensitive to both hardware and software changes. Local HPC provides greater stability in both hardware and software, whereas researchers have no control over the latter when using cloud computing services. In this regard, LLM annotation is like human coding in that exact replication is impossible, but variation should and can be controlled to a certain extent, encouraging researchers to run the data multiple times.

Finally, on the privacy and ethics front—particularly when researching contentious political topics like protests, collective action, and civil disobedience—data sensitivity strongly favors local deployment. Running the model on local HPC, of course, provides stronger data privacy than using a commercial cloud computing service when the image contains identifiable information, such as faces, geolocation data, and metadata, even though the data in this study is publicly available. Nonetheless, the security issue is multilayered; for instance, data from an HPC could be hacked, and LLMs may have backdoors that send data elsewhere unless the device can be kept completely offline, which may complicate daily maintenance. Researchers must assess the risk of a privacy breach and the associated risk when applying LLMs.

Reference

Aminzade, R. R., Goldstone, J. A., McAdam, D., Perry, E. J., Sewell, W. H., Tarrow, S., & Tilley, C. (2001). *Silence and Voice in the Study of Contentious Politics* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511815331

Arceneaux, K. (2012). Cognitive Biases and the Strength of Political Arguments. *American Journal of Political Science*, *56*(2), 271–285. https://doi.org/10.1111/j.1540-5907.2011.00573.x

Arpan, L. M., Baker, K., Lee, Y., Jung, T., Lorusso, L., & Smith, J. (2006). News Coverage of Social Protests and the Effects of Photographs and Prior Attitudes. *Mass Communication and Society*, *9*(1), 1–20. https://doi.org/10.1207/s15327825mcs0901_1

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., … Zhu, K. (2025). *Qwen3-VL Technical Report* (arXiv:2511.21631). arXiv. https://doi.org/10.48550/arXiv.2511.21631

Balluff, P., Ho, J. C., Gruber, J. B., Palicki, S., Palmer, A., Rossi, L., Shklovski, I., & Chan, C. (2026). Newer, Larger, Better? A Critique of the Unreflective LLM Adoption in Communication Research. *Political Communication*, *43*(3), 572–581. https://doi.org/10.1080/10584609.2026.2618486

Barrie, C., Palmer, A., & Spirling, A. (2024). *Replication for Language Models*. https://arthurspirling.org/documents/BarriePalmerSpirling_TrustMeBro.pdf

Bayerl, P. S., & Stoynov, L. (2016). Revenge by photoshop: Memefying police acts in the public dialogue about injustice. *New Media & Society*, *18*(6), 1006–1026. https://doi.org/10.1177/1461444814554747

Birkenmaier, L., Lechner, C. M., & Wagner, C. (2024). The Search for Solid Ground in Text as Data: A Systematic Review of Validation Practices and Practical Recommendations for Validation. *Communication Methods and Measures*, *18*(3), 249–277. https://doi.org/10.1080/19312458.2023.2285765

Bradburn, N. M., Rips, L. J., & Shevell, S. K. (1987). Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science*, *236*(4798), 157–161. https://doi.org/10.1126/science.3563494

Brown, D. K., & Harlow, S. (2019). Protests, Media Coverage, and a Hierarchy of Social Struggle. *The International Journal of Press/Politics*, *24*(4), 508–530. https://doi.org/10.1177/1940161219853517

Caiani, M. (2024). Visual Analysis and the Contentious Politics of the Radical Right. *Sociology Compass*, *18*(9), e13267. https://doi.org/10.1111/soc4.13267

Casas, A., & Williams, N. W. (2019). Images that Matter: Online Protests and the Mobilizing Role of Pictures. *Political Research Quarterly*, *72*(2), 360–375. https://doi.org/10.1177/1065912918786805

Chwe, M. S.-Y. (2013). *Rational ritual: Culture, coordination, and common knowledge* (Reissue with a new afterword by the author). Princeton University Press.

Cochrane, C., Rheault, L., Godbout, J.-F., Whyte, T., Wong, M. W.-C., & Borwein, S. (2022). The Automatic Analysis of Emotion in Political Speech Based on Transcripts. *Political Communication*, *39*(1), 98–121. https://doi.org/10.1080/10584609.2021.1952497

Corrigall-Brown, C., & Wilkes, R. (2012). Picturing Protest: The Visual Framing of Collective Action by First Nations in Canada. *American Behavioral Scientist*, *56*(2), 223–243. https://doi.org/10.1177/0002764211419357

Cowart, H. S., Saunders, L. M., & Blackstone, G. E. (2016). Picture a Protest: Analyzing Media Images Tweeted From Ferguson. *Social Media + Society*, *2*(4), 2056305116674029. https://doi.org/10.1177/2056305116674029

Crabtree, C., Golder, M., Gschwend, T., & Indriđason, I. H. (2020). It Is Not Only What You Say, It Is Also How You Say It: The Strategic Use of Campaign Sentiment. *The Journal of Politics*, *82*(3), 1044–1060. https://doi.org/10.1086/707613

Das, G., La, V., Lau, E., Shrivastava, A., & Gwilliam, M. (2026). *Towards Understanding Best Practices for Quantization of Vision-Language Models* (arXiv:2601.15287). arXiv. https://doi.org/10.48550/arXiv.2601.15287

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G., & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.372

Edmond, C. (2013). Information Manipulation, Coordination, and Regime Change*. *The Review of Economic Studies*, *80*(4), 1422–1458. https://doi.org/10.1093/restud/rdt020

Fung, W. L. C. (2020). *Canvas of dissent: A study of visuals and its significance in group formations and communications during the 2019 Hong Kong Anti-ELAB movement* [Master Thesis, Lund University]. https://lup.lub.lu.se/student-papers/search/publication/9009383

Gamson, W. A. (1990). *The strategy of social protest* (2. ed). Wadsworth Publ. Co.

Geise, S., Panke, D., & Heck, A. (2021). Still Images—Moving People? How Media Images of Protest Issues and Movements Influence Participatory Intentions. *The International Journal of Press/Politics*, *26*(1), 92–118. https://doi.org/10.1177/1940161220968534

Gerbaudo, P. (2015). Protest avatars as memetic signifiers: Political profile pictures and the construction of collective identity on social media in the 2011 protest wave. *Information, Communication & Society*, *18*(8), 916–929. https://doi.org/10.1080/1369118X.2015.1043316

Gerbaudo, P. (2016). Constructing Public Space| Rousing the Facebook Crowd: Digital Enthusiasm and Emotional Contagion in the 2011 Protests in Egypt and Spain. *International Journal of Communication*, *10*(0), Article 0.

Giray, L. (2023). Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of Biomedical Engineering*, *51*(12), 2629–2633. https://doi.org/10.1007/s10439-023-03272-4

Goodwin, J., Jasper, J. M., & Polletta, F. (2001). *Passionate Politics: Emotions and Social Movements*. University of Chicago Press.

Goodwin, J., Jasper, J., & Polletta, F. (2000). *The Return of The Repressed: The Fall and Rise of Emotions in Social Movement Theory*. https://doi.org/10.17813/maiq.5.1.74u39102m107g748

Grimmer, J., & Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Halperin, L. (2025). Combining emotions: Hope, anger, joy, and love in Israeli peace movements. *Social Movement Studies*, *24*(3), 379–394. https://doi.org/10.1080/14742837.2022.2155628

He, K., Zhang, X., Ren, S., & Sun, J. (2015). *Deep Residual Learning for Image Recognition* (arXiv:1512.03385). arXiv. https://doi.org/10.48550/arXiv.1512.03385

Herasimenka, A. (2022). Movement Leadership and Messaging Platforms in Preemptive Repressive Settings: Telegram and the Navalny Movement in Russia. *Social Media + Society*, *8*(3), 20563051221123038. https://doi.org/10.1177/20563051221123038

Hoffmann, M., & Neumayer, C. (2025). Images of protest in movement parties' social media communication. *New Media & Society*, *27*(8), 4929–4952. https://doi.org/10.1177/14614448241243352

Huang, D., & Wang, Z. (2025). LLMs at the Edge: Performance and Efficiency Evaluation with Ollama on Diverse Hardware. *2025 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN64981.2025.11228317

Jasper, J. M. (1998). The Emotions of Protest: Affective and Reactive Emotions In and Around Social Movements. *Sociological Forum*, *13*(3), 397–424. https://doi.org/10.1023/A:1022175308081

Jenzen, O., Erhart, I., Eslen-Ziya, H., Korkut, U., & McGarry, A. (2020). The symbol of social media in contemporary protest: Twitter and the Gezi Park movement. *Convergence*, 1354856520933747. https://doi.org/10.1177/1354856520933747

Joo, J., & Steinert-Threlkeld, Z. (2022). Image as Data: Automated Content Analysis for Visual Presentations of Political Actors and Events. *Computational Communication Research*, *4*(1), Article 1.

Jung, J.-H. (2020). The Mobilizing Effect of Parties' Moral Rhetoric. *American Journal of Political Science*, *64*(2), 341–355. https://doi.org/10.1111/ajps.12476

Kamath, A., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., Rouillard, L., Mesnard, T., Cideron, G., Grill, J., Ramos, S., Yvinec, E., Casbon, M., Pot, E., Penchev, I., Liu, G., Visin, F., … Hussenot, L. (2025, March 25). *Gemma 3 Technical Report*. arXiv.Org. https://arxiv.org/abs/2503.19786v1

Kharroub, T., & Bas, O. (2016). Social media and protests: An examination of Twitter images of the 2011 Egyptian revolution. *New Media & Society*, *18*(9), 1973–1992. https://doi.org/10.1177/1461444815571914

Kim, M., & Bas, O. (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), 20563051231195582. https://doi.org/10.1177/20563051231195582

Kosmidis, S., Hobolt, S. B., Molloy, E., & Whitefield, S. (2019). Party Competition and Emotive Rhetoric. *Comparative Political Studies*, *52*(6), 811–837. https://doi.org/10.1177/0010414018797942

Koufakou, A., Nieves, E., & Peller, J. (2024). Towards a new Benchmark for Emotion Detection in NLP: A Unifying Framework of Recent Corpora. In D. Hupkes, V. Dankers, K. Batsuren, A. Kazemnejad, C. Christodoulopoulos, M. Giulianelli, & R. Cotterell (Eds.), *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP* (pp. 196–206). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.genbench-1.13

Kuran, T. (1991). Now Out of Never: The Element of Surprise in the East European Revolution of 1989. *World Politics*, *44*(1), 7–48. https://doi.org/10.2307/2010422

Lambert, A. J., Eadeh, F. R., & Hanson, E. J. (2019). Anger and its consequences for judgment and behavior: Recent developments in social and political psychology. In *Advances in*

Experimental Social Psychology (Vol. 59, pp. 103–173). Elsevier. https://doi.org/10.1016/bs.aesp.2018.12.001
Lasswell, H. D. (2009). *Power and Personality*. Transaction Publishers. https://books.google.com/books?id=1y69mxixRYQC
Le Bon, G. (2009). *Psychology of crowds*. Sparkling Books.
Lertchoosakul, K. (2021). The white ribbon movement: High school students in the 2020 Thai youth protests. *Critical Asian Studies*, *53*(2), 206–218. https://doi.org/10.1080/14672715.2021.1883452
Lim, M. (2013). Framing Bouazizi: 'White lies', hybrid network, and collective/connective action in the 2010–11 Tunisian uprising. *Journalism*, *14*(7), 921–941. https://doi.org/10.1177/1464884913478359
Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content Analysis by the Crowd: Assessing the Usability of Crowdsourcing for Coding Latent Constructs. *Communication Methods and Measures*, *11*(3), 191–209. https://doi.org/10.1080/19312458.2017.1317338
Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). *Visual Instruction Tuning* (arXiv:2304.08485). arXiv. https://doi.org/10.48550/arXiv.2304.08485
Lu, L., Tao, R., Kwon, H., Kang, J., Zhou, Y., Xin, H., Duncan, J., & McLeod, D. (2025). Visual Constructs of Conflict and Solidarity: The Role of Visual Framing on Public Perceptions and Engagement Intentions with Social Protests. *Visual Communication Quarterly*, *32*(1), 17–32. https://doi.org/10.1080/15551393.2025.2452959
Lu, Y. (2026). Performative Propaganda Engagement: How Celebrity Fans Engage with State Propaganda on Weibo. *Political Communication*, *43*(1), 99–127. https://doi.org/10.1080/10584609.2025.2584999
Lu, Y., & Peng, Y. (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
Marcondes, F. S., Gala, A., Magalhães, R., Perez de Britto, F., Durães, D., & Novais, P. (2025). Using Ollama. In F. S. Marcondes, A. Gala, R. Magalhães, F. Perez de Britto, D. Durães, & P. Novais (Eds.), *Natural Language Analytics with Generative Large-Language Models: A Practical Approach with Ollama and Open-Source LLMs* (pp. 23–35). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-76631-2_3
Marcus, G. E. (2000). Emotions in Politics. *Annual Review of Political Science*, *3*(Volume 3, 2000), 221–250. https://doi.org/10.1146/annurev.polisci.3.1.221
Marcus, G. E., Neuman, W. R., & MacKuen, M. B. (2017). Measuring Emotional Response: Comparing Alternative Approaches to Measurement. *Political Science Research and Methods*, *5*(4), 733–754. https://doi.org/10.1017/psrm.2015.65
Mattoni, A., & Teune, S. (2014). Visions of Protest. A Media-Historic Perspective on Images in Social Movements. *Sociology Compass*, *8*(6), 876–887. https://doi.org/10.1111/soc4.12173
McAdam, D. (1999). *Political process and the development of Black insurgency, 1930-1970* (2nd edition). University of Chicago Press.
McCarthy, J. D., & Zald, M. N. (1977). Resource Mobilization and Social Movements: A Partial Theory. *American Journal of Sociology*, *82*(6), 1212–1241. https://doi.org/10.1086/226464
Mohammad, S., Bravo-Marquez, F., Salameh, M., & Kiritchenko, S. (2018). SemEval-2018 Task 1: Affect in Tweets. In M. Apidianaki, S. M. Mohammad, J. May, E. Shutova, S.

Bethard, & M. Carpuat (Eds.), *Proceedings of the 12th International Workshop on Semantic Evaluation* (pp. 1–17). Association for Computational Linguistics. https://doi.org/10.18653/v1/S18-1001

Molder, A. L., Lakind, A., Clemmons, Z. E., & Chen, K. (2022). Framing the Global Youth Climate Movement: A Qualitative Content Analysis of Greta Thunberg's Moral, Hopeful, and Motivational Framing on Instagram. *The International Journal of Press/Politics*, *27*(3), 668–695. https://doi.org/10.1177/19401612211055691

Morrison, D. R., & Isaac, L. W. (2012). Insurgent Images: Genre Selection and Visual Frame Amplification in IWW Cartoon Art. *Social Movement Studies*, *11*(1), 61–78. https://doi.org/10.1080/14742837.2012.640530

Neufeld, K. H. S., Starzyk, K. B., & Gaucher, D. (2019). Political Solidarity: A Theory and a Measure. *Journal of Social and Political Psychology*, *7*(2), 726–765. https://doi.org/10.5964/jspp.v7i2.1058

Neumayer, C., & Rossi, L. (2018). Images of protest in social media: Struggle over visibility and visual narratives. *New Media & Society*, *20*(11), 4293–4310. https://doi.org/10.1177/1461444818770602

Olesen, T. (2013). "We are all Khaled Said": Visual Injustice Symbols in the Egyptian Revolution, 2010–2011. In N. Doerr, A. Mattoni, & S. Teune (Eds.), *Advances in the Visual Analysis of Social Movements* (Vol. 35, pp. 3–25). Emerald Group Publishing Limited. https://doi.org/10.1108/S0163-786X(2013)0000035005

Olesen, T. (2015). *Global Injustice Symbols and Social Movements*. Palgrave Macmillan US. https://doi.org/10.1057/9781137481177

Osnabrügge, M., Hobolt, S. B., & Rodon, T. (2021). Playing to the Gallery: Emotive Rhetoric in Parliaments. *American Political Science Review*, *115*(3), 885–899. https://doi.org/10.1017/S0003055421000356

Palczewski, C. H. (2005). The Male Madonna and the Feminine Uncle Sam: Visual Argument, Icons, and Ideographs in 1909 Anti-Woman Suffrage Postcards. *Quarterly Journal of Speech*, *91*(4), 365–394. https://doi.org/10.1080/00335630500488325

Pearlman, W. (2013). Emotions and the Microfoundations of the Arab Uprisings. *Perspectives on Politics*, *11*(2), 387–409. https://doi.org/10.1017/S1537592713001072

Peng, Y., Lock, I., & Ali Salah, A. (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., Lock, I., & Salah, A. A. (2024). Automated Visual Analysis for the Study of Social Media Effects: Opportunities, Approaches, and Challenges. *Communication Methods and Measures*. (world). https://www.tandfonline.com/doi/abs/10.1080/19312458.2023.2277956

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

Popkin, S. L. (1994). *The reasoning voter: Communication and persuasion in presidential campaigns* (2nd ed). University of Chicago Press.

Prasad, I. S., & Steinert-Threlkeld, Z. C. (2025). Taken at face value: Emotion expression and protest dynamics. *Research & Politics*, *12*(3), 20531680251366095. https://doi.org/10.1177/20531680251366095

Ramsey, E. M. (2000). Inventing citizens during World War I: Suffrage cartoons in the woman citizen. *Western Journal of Communication*, *64*(2), 113–147. https://doi.org/10.1080/10570310009374668

Ransan-Cooper, H., A. Ercan, S., & Duus, S. (2018). When anger meets joy: How emotions mobilise and sustain the anti-coal seam gas movement in regional Australia. *Social Movement Studies*, *17*(6), 635–657. https://doi.org/10.1080/14742837.2018.1515624

Rossi, L., Neumayer, C., Henrichsen, J., & Beck, L. K. (2023). Measuring Violence: A Computational Analysis of Violence and Propagation of Image Tweets From Political Protest. *Social Science Computer Review*, *41*(3), 905–925. https://doi.org/10.1177/08944393211055429

Rouzegar, H., & Makrehchi, M. (2024). Enhancing Text Classification through LLM-Driven Active Learning and Human Annotation. In S. Henning & M. Stede (Eds.), *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)* (pp. 98–111). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.law-1.10

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than Bags of Words: Sentiment Analysis with Word Embeddings. *Communication Methods and Measures*, *12*(2–3), 140–157. https://doi.org/10.1080/19312458.2018.1455817

Safaian, D., & Teune, S. (2022). Visual framing in the German movements for gay liberation and against nuclear energy. *Visual Studies*, *37*(4), 272–283. https://doi.org/10.1080/1472586X.2021.1940265

Scholz, S., Weidmann, N. B., Steinert-Threlkeld, Z. C., Keremoğlu, E., & Goldlücke, B. (2025). Improving Computer Vision Interpretability: Transparent Two-Level Classification for Complex Scenes. *Political Analysis*, *33*(2), 107–121. https://doi.org/10.1017/pan.2024.18

Shah, T. M. (2024). Emotions in Politics: A Review of Contemporary Perspectives and Trends. *International Political Science Abstracts*, *74*(1), 1–14. https://doi.org/10.1177/00208345241232769

Shahin, S., & Ng, Y. M. M. (2022). Connective action or collective inertia? Emotion, cognition, and the limits of digitally networked resistance. *Social Movement Studies*, *21*(4), 530–548. https://doi.org/10.1080/14742837.2021.1928485

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Simonyan, K., & Zisserman, A. (2015). *Very Deep Convolutional Networks for Large-Scale Image Recognition* (arXiv:1409.1556). arXiv. https://doi.org/10.48550/arXiv.1409.1556

Sobolev, A., Chen, M. K., Joo, J., & Steinert-Threlkeld, Z. C. (2020). News and Geolocated Social Media Accurately Measure Protest Size Variation. *American Political Science Review*, *114*(4), 1343–1351. https://doi.org/10.1017/S0003055420000295

Song, H., Tolochko, P., Eberl, J.-M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S., & Boomgaarden, H. G. (2020). In Validations We Trust? The Impact of Imperfect Human Annotations as a Gold Standard on the Quality of Validation of Automated Content Analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Steinert, -Threlkeld Zachary C., Chan, A. M., & Joo, J. (2022). How State and Protester Violence Affect Protest Dynamics. *The Journal of Politics*, *84*(2), 798–813. https://doi.org/10.1086/715600

Summers-Effler, E. (2007). The emotional significance of solidarity for social movement communities: Sustaining Catholic worker community and service. In *Emotions and social movements* (pp. 135–149). Routledge. https://books.google.com/books?hl=zh-TW&lr=&id=LTOefxu2lUkC&oi=fnd&pg=PA135&dq=The+emotional+significance+of+solidarity+for+social+movement+communities&ots=CxGh0m6abi&sig=zitXjgqzh5pohoxC3Mx3mNZJwIA

Team, G., Abd, S. E., Aggarwal, V., Algayres, R., Andreev, A., Bachem, O., Ballantyne, I., Brick, C., Cărbune, V., Casbon, M., Chaturvedi, M., Cotruta, V., Coucke, A., Culliton, P., Dadashi, R., Dixon, L., Elhawaty, M., Evci, U., Farabet, C., … Zou, C. (2026). *Gemma 4 Technical Report* (arXiv:2607.02770). arXiv. https://doi.org/10.48550/arXiv.2607.02770

Teo, E., & Fu, K. (2021). A novel systematic approach of constructing protests repertoires from social media: Comparing the roles of organizational and non-organizational actors in social movement. *Journal of Computational Social Science*, *4*(2), 787–812. https://doi.org/10.1007/s42001-021-00101-3

Torres, M., & Cantú, F. (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

Urman, A., Ho, J. C., & Katz, S. (2021). Analyzing protest mobilization on Telegram: The case of 2019 Anti-Extradition Bill movement in Hong Kong. *PLOS ONE*, *16*(10), e0256675. https://doi.org/10.1371/journal.pone.0256675

Valentino, N. A., Hutchings, V. L., & White, I. K. (2002). Cues that Matter: How Political Ads Prime Racial Attitudes During Campaigns. *American Political Science Review*, *96*(1), 75–90. https://doi.org/10.1017/S0003055402004240

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, *15*(2), 121–140. https://doi.org/10.1080/19312458.2020.1869198

van Stekelenburg, J., & Klandermans, B. (2013). The social psychology of protest. *Current Sociology*, *61*(5–6), 886–905. https://doi.org/10.1177/0011392113479314

Van Troost, D., Van Stekelenburg, J., & Klandermans, B. (2013). Emotions of Protest. In N. Demertzis (Ed.), *Emotions in Politics* (pp. 186–203). Palgrave Macmillan UK. https://doi.org/10.1057/9781137025661_10

Vance, L. D., & Potash, J. S. (2022). Black Lives Matter protest art: Uncovering explicit and implicit emotions through thematic analysis. *Peace and Conflict: Journal of Peace Psychology*, *28*(1), 121–129. https://doi.org/10.1037/pac0000584

Veg, S. (2016). Creating a Textual Public Space: Slogans and Texts from Hong Kong's Umbrella Movement. *Journal of Asian Studies*, *75*(3), 673–702. https://doi.org/10.1017/S0021911816000565

Walgrave, S., Wouters, R., & Ketelaars, P. (2016). Response Problems in the Protest Survey Design: Evidence from Fifty-One Protest Events in Seven Countries*. *Mobilization: An International Quarterly*, *21*(1), 83–104. https://doi.org/10.17813/1086/671X-21-1-83

Wang, S., Jui, I. J., & Thorpe, J. (2024). Is Crowdsourcing a Puppet Show? Detecting a New Type of Fraud in Online Platforms. *Proceedings of the New Security Paradigms Workshop*, 84–95. https://doi.org/10.1145/3703465.3703472

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Webster, S. W., & Albertson, B. (2022). Emotion and Politics: Noncognitive Psychological Biases in Public Opinion. *Annual Review of Political Science*, *25*(Volume 25, 2022), 401–418. https://doi.org/10.1146/annurev-polisci-051120-105353

Wetzel, C. (2012). Envisioning Land Seizure: Diachronic Representations of the Occupation of Alcatraz Island. *American Behavioral Scientist*, *56*(2), 151–171. https://doi.org/10.1177/0002764211419354

Widmann, T. (2021). How Emotional Are Populists Really? Factors Explaining Emotional Appeals in the Communication of Political Parties. *Political Psychology*, *42*(1), 163–181. https://doi.org/10.1111/pops.12693

Wijermars, M., & Lokot, T. (2022). Is Telegram a "harbinger of freedom"? The performance, practices, and perception of platforms as political actors in authoritarian states. *Post-Soviet Affairs*, *38*(1–2), 125–145. https://doi.org/10.1080/1060586X.2022.2030645

Won, D., Steinert-Threlkeld, Z. C., & Joo, J. (2017). Protest Activity Detection and Perceived Violence Estimation from Social Media Images. *Proceedings of the 25th ACM International Conference on Multimedia, MM '17*, 786–794. https://doi.org/10.1145/3123266.3123282

Xu, Z., Khatri, V., Dai, Y., Liu, X., Li, S., Zhang, X., & Yu, R. (2026). *Enhancing LLM-Based Data Annotation with Error Decomposition* (arXiv:2601.11920). arXiv. https://doi.org/10.48550/arXiv.2601.11920

Young, L. E. (2019). The Psychology of State Repression: Fear and Dissent Decisions in Zimbabwe. *American Political Science Review*, *113*(1), 140–155. https://doi.org/10.1017/S000305541800076X

Young, L., & Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, *29*(2), 205–231. https://doi.org/10.1080/10584609.2012.671234

Yuen, S., Tang, G., Lee, F. L. F., & Cheng, E. W. (2022). Surveying Spontaneous Mass Protests: Mixed-mode Sampling and Field Methods. *Sociological Methodology*, *52*(1), 75–102. https://doi.org/10.1177/00811750211071130

Zhang, H., & Leung, R. (2026). Joint Text-and-Image Clustering for Social Science Research. *Sociological Methodology*, *56*(1), 1–30. https://doi.org/10.1177/00811750251382929

Zhang, H., & Pan, J. (2019). CASM: A Deep-Learning Approach for Identifying Collective Action Events with Text and Image Data from Social Media. *Sociological Methodology*, *49*(1), 1–57. https://doi.org/10.1177/0081175019860244

Zhang, H., & Peng, Y. (2024). Image Clustering: An Unsupervised Approach to Categorize Visual Data in Social Science Research. *Sociological Methods & Research*, *53*(3), 1534–1587. https://doi.org/10.1177/00491241221082603

Zhu, Y., Cheng, E. W., Shen, F., & Walker, R. M. (2022). An Eye for an Eye? An Integrated Model of Attitude Change Toward Protest Violence. *Political Communication*, *39*(4), 539–563. https://doi.org/10.1080/10584609.2022.2053915

Appendix A Prompt for coding data from 2019 Hong Kong Protests

prompt_hk_v4 = """
Objective: Analyze the provided image from the 2019 Hong Kong Anti-Extradition Bill Protests within its historical framework (mass protests demanding democratic reforms, police accountability, and protection of civil liberties). Evaluate visual evidence for Solidarity, Anger, and Joy from the pro-protest perspective. Output strictly in JSON.

1. STEP-BY-STEP WORKFLOW
1. TRANSCRIBE & TRANSLATE: Locate and transcribe all salient text/slogans. Translate Cantonese into English internally.
2. DESCRIBE: Provide a concise description of the visual layout and key subjects (max 30 words).
3. EVALUATE & CODE LAST: Write detailed qualitative explanations for each emotion FIRST (max 30 words). Then, assign the binary presence code ("0" or "1") and certainty score [0-10] based strictly on your analysis.

2. DEFINITIONS & CUES
DEFINITIONS
- Solidarity: Positive collective emotion from shared identity, mutual aid, and hopeful commitment that sustains engagement across diverse protest factions.
- Anger: Reactive moral outrage against injustice or perceived violations of rights, typically targeting authorities to frame a clear "common enemy."
- Joy: Delight, playfulness, or relief used to counter stress, boost morale, and sustain engagement through humor and lighthearted expression.

CUES
- Solidarity: Shared memories, core protest slogans (5 demands), calls for mutual support/safety, petitions, calls from unions and narratives highlighting collective strength and international backing.
- Anger: Direct criticism of government/police, documentation of misconduct (force, arrests, abuses), and framing of systemic institutional injustice. Other targets may include pro-government citizens and businessmen.
- Joy: Satire, memes, caricatures of authority, pop culture references, and cartoonish/playful visuals used for political expression or comic relief.

3. STRICT RULES OF ANALYSIS
1. Holistic Evaluation: Analyze the image as a complete unit (all visuals and text together) and focus on cues salient to a human observer.
2. Diagnostic Cues: Use the visual cues listed, but remain open to other evidence.
3. Word aesthetics: Account for artistic, handwritten, non-standard, stylized typography and varied orientations when doing the word collocation
4. Image Priority: If the text message (if provided) contradicts the image, rely on the image.
5. Salience & Primary Function: While protest materials essentially are about grievances, only code anger when the dominant tone is explicit hostility, violent confrontation, or aggressive outrage.

6. Interpretation Limits: Default to "0". Only mark an emotion as present ("1") if clear, explicit evidence exists. Do not make highly abstract or speculative interpretations.
7. Salience Prioritization: Prioritize visual cues that command attention (size, contrast, position, headlines).
8. Factual-Emotional Weight: Acknowledge that factual images (e.g., tallies, casualties, news clips) can carry emotion.
9. Accurate Subject Identification: Distinguish between police, protesters, and neutral citizens.
10. Intent and Complexity: Do not assume an image is restricted to only one dominant emotion.
11. Broad Movement Context: Evaluate images considering the history of the movement and Hong Kong.

4. EXPECTED CODING LOGIC EXAMPLES
- Anger: "1" if showing police firing tear gas at citizens; "0" if text only lists future rally schedules without triggering clues.
- Solidarity: "1" if showing large crowds and messages of resilience; "0" if focused purely on an isolated act of police brutality.
- Joy: "1" if utilizing a playful meme mocking authority; "0" if the tone is entirely serious or informational.

5. OUTPUT SCHEMA
Perform the workflow internally. Output ONLY a single, raw, valid JSON object matching this structure exactly (use "0" or "1" as your decision variables, and "0" to "10" for the certainty level):

```
{
  "Image_Description": "Text description string",
  "Explanation_Solidarity": "Short explanation string",
  "Solidarity": "0",
  "Certainty_Solidarity": "0",
  "Explanation_Anger": "Short explanation string",
  "Anger": "0",
  "Certainty_Anger": "0",
  "Explanation_Joy": "Short explanation string",
  "Joy": "0",
  "Certainty_Joy": "0"
}
```

IMPORTANT: Do not include markdown code blocks (like ```json) or conversational text.
"""

Appendix B: Prompt for coding data from Casas & William (2019)

prompt_blm_v4 = """
Objective: Analyze the provided social media image from the Black Lives Matter (BLM) and #ShutdownA14 protests (April 2015) within its historical framework (mass protests demanding police accountability, an end to police brutality, and racial justice). Evaluate visual evidence for Sadness, Anger, and Enthusiasm from the pro-protest perspective. Output strictly in JSON.

1. STEP-BY-STEP WORKFLOW
1. TRANSCRIBE & TRANSLATE: Locate and transcribe all salient text/slogans. Translate non-English text internally if applicable.
2. DESCRIBE: Provide a concise description of the visual layout and key subjects (max 30 words).
3. EVALUATE & CODE LAST: Write detailed qualitative explanations for each emotion FIRST (max 30 words). Then, assign the binary presence code ("0" or "1") and certainty score [0-10] based strictly on your analysis.

2. DEFINITIONS & CUES
DEFINITIONS
- Sadness: Heavy grief, sympathy, or mourning stemming from tragic loss of life, police violence, and structural factors like systemic violence and negligence toward affected families and communities.
- Anger: Reactive moral outrage against injustice or perceived violations of rights, typically targeting authorities to frame a clear "common enemy.
- Enthusiasm: A collective emotion stemming from shared identity, dynamic action, hopeful commitment, and active engagement among protesters, sustaining movement energy.

CUES
- Sadness: Active memorialization (vigils, candles, spontaneous shrines, flowers), headshots/portraits/obituaries of victims, public listings of casualty names, and authentic family mourning.
- Anger: Direct criticism of law enforcement/government, documentation of police violence/misconduct (weapons, arrests, physical force), hostile signs targeting authorities, and aggressive anti-brutality rhetoric.
- Enthusiasm: Protest flyers, calls to action, hopeful/empowering language, active marching, dynamic crowds, energetic banner-holding, and active postures signaling hopeful participation in social change.

3. STRICT RULES OF ANALYSIS
1. Holistic Evaluation: Analyze the image as a complete unit.
2. Diagnostic Cues: Use the visual cues listed, but remain open to other evidence.
3. Image Priority: If the text message (if provided) contradicts the image, rely on the image.
4. Ecological Grounding: Ground coding on the shared grievances of the pro-protest public.
5. Interpretation Limits: Avoid overly speculative or highly abstract interpretations.
6. Salience Prioritization: Prioritize visual cues that command attention (size, contrast, position, headlines).

7. Factual-Emotional Weight: Acknowledge that factual images (e.g., tallies, casualties, news clips) can carry emotion.
8. Accurate Subject Identification: Distinguish between police, protesters, and neutral citizens.
9. Intent and Complexity: Do not assume an image is restricted to only one dominant emotion.
10. Broad Movement Context: Evaluate images considering the history of the BLM movement and Black people.

4. EXPECTED CODING LOGIC EXAMPLES
- Sadness: "1" if displaying a candle-lit vigil or victim portrait; "0" if the tone is purely energetic, victorious, or informational.
- Anger: "1" if showing police aggressively deploying force or direct hostility toward officers; "0" if text only lists future march logistics or rally schedules.
- Enthusiasm: "1" if showing dynamic marching crowds or empowering calls to action; "0" if focused purely on a static memorial portrait without active engagement.

5. OUTPUT SCHEMA
Perform the workflow internally. Output ONLY a single, raw, valid JSON object matching this structure exactly (use "0" or "1" as your decision variables, and "0" to "10" for the certainty level):

```
{
  "Image_Description": "Text description string",
  "Explanation_Sadness": "Short explanation string",
  "Sadness": "0",
  "Certainty_Sadness": "0",
  "Explanation_Anger": "Short explanation string",
  "Anger": "0",
  "Certainty_Anger": "0",
  "Explanation_Enthusiasm": "Short explanation string",
  "Enthusiasm": "0",
  "Certainty_Enthusiasm": "0"
}
```

IMPORTANT: Do not include markdown code blocks (like ```json) or conversational text.
"""

Appendix C: Results from alternative model on the 2019 Hong Kong Protests Data

Table A1. Alternative Model – Results from Gemma4:31B models

		Runtime	
Plain image		20s	
With message		21s	
Performance: Plain image	**Anger**	**Solidarity**	**Joy**
Precision	0.58	0.77	0.46
Recall	0.89	0.93	0.70
F1-Score	0.70	0.84	0.56
Performance: With Message	**Anger**	**Solidarity**	**Joy**
Precision	0.59	0.76	0.49
Recall	0.87	0.95	0.76
F1-Score	0.70	0.85	0.59
Cohen's Kappa with Qwen3-VL:32B	**Anger**	**Solidarity**	**Joy**
Plain image	0.70	0.44	0.48
With Message	0.61	0.47	0.47

Appendix D: Comparison of LLMs

Table A2. Comparing features and performance of the three LLMs employed

Areas of focus	Gemma 3	Llava	Qwen3-vl
1. Vision encoding	**Big-picture summaries:** Shrinks every photo to a small, fixed summary size. Great at identifying general scenes and reading standard text across 140+ languages.	**Clear, front-and-center images:** Chops images into square tiles to keep details clear. Works well for clear photos with obvious subjects	**Micro-details & bad handwriting:** Excels at reading small, blurry, or tilted text on protest signs. Can pinpoint exact locations of objects in using bounding boxes.
2. How smart is its reasoning?	**Solid general logic:** Good at answering straightforward questions and giving standard descriptions but misses tiny visual clues hidden in crowded backgrounds.	**Basic step-by-step logic:** Relies on its text memory to make sense of what it sees. Good for simple questions but can hallucinate details in chaotic scenes.	**Detective-level visual logic:** Uses "chain-of-thought" reasoning to think step-by-step. It looks closely at symbols, signs, and spatial relationships before reaching a conclusion.
3. How fast and cheap is it to run?	**Balanced speed & high precision:** Because every image is compressed into a fixed, lightweight size, it processes massive piles of photos quickly using minimal computer memory.	**Slower & needs more computer power:** Tiling high-resolution photos into multiple pieces uses up a lot of memory fast, making large-scale image processing slower and pricier.	**Slowest:** Adjusts how hard it works depending on image complexity. Requires decent computer power by studying details of images with reasoning
4. New dimension: customization & fine-tuning	**Moderate effort:** Supported by modern ai tools, but its fixed image size limits how much you can teach it to spot tiny new visual details.	**Easiest to customize:** Good for finetuning by connecting to many other open-source tools scoring system or categories.	**Advanced setup required:** Capable of specialized tasks, but demand higher computational power
5. Main takeaway	**The high-speed worker:** pick this if you have **data at scale t**hat could be grossly classify or describe without fine-grained visual markers	**The customizer:** pick this if you want to finetune a model on your own custom rules using easy, beginner-friendly open-source tools.	**The deep investigator:** pick this if you need to read tiny sign text, count objects in a crowd, or analyze complex political symbols.