Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 8.2 (2026) 1–35
HTTPS://DOI.ORG/10.5117/CCR2026.2.3.KIM

Visual framing at scale: A theory-driven computational framework for analyzing protest imagery with generative AI

Sang Jung Kim
School of Journalism and Mass Communication, University of Iowa, USA.

Lei Chen
School of Journalism and Mass Communication, University of Iowa, USA.
School of Communication, Xiamen University, Malaysia.

Abstract
This study presents a theory-driven, three-stage computational framework for analyzing visual framing in protest imagery. Focusing on the Black Lives Matter movement, we examine how visual elements contribute to two well-established frames in protest media coverage: the protest paradigm and solidarity framing. Leveraging GPT-4o and OpenCV, our framework extracts denotative and semiotic features—such as police presence, contestation, solidarity actions, and color contrast—and links these features to higher-order frame classifications using interpretable logistic regression models. The framework includes: (1) feature definition and validation through generative AI and a feature extraction tool, supported by human coders; (2) model training; and (3) predictive application to unseen images. Results show strong alignment between human and machine annotations, as well as high predictive accuracy in identifying the protest paradigm or solidarity frame in BLM images. We also introduce an intra-prompt stability score for the generative AI model to help mitigate hallucination and enhance the reliability of its outputs. This study offers a scalable, replicable, and interpretable approach to visual framing analysis, bridging communication theory with advanced computational tools in the study of visual political communication.

Keywords: Generative AI, visual framing, protest imagery, protest paradigm, solidarity framing, computer vision

Introduction

Visual imagery plays a central role in shaping how the public, policymakers, and the press perceive protests and protesters (Corrigall-Brown & Wilkes, 2012). In media coverage of protests such as Black Lives Matter (BLM),

images can frame participants as either disruptive threats or empowered political actors—amplifying or undermining the legitimacy of their claims. Communication scholars have long theorized these journalistic practices through concepts like the protest paradigm, which emphasizes conflict and deviance and tends to marginalize protesters (McLeod & Detenber, 1999), and solidarity framing, which highlights empathy, legitimacy, and unity among demonstrators (Masullo et al., 2024).

While theoretical advances in visual framing are well established, much of the field still relies on manual content analysis—a method that is labor-intensive, limited in scale, and prone to subjectivity, particularly given the emotional impact of imagery on coders who classify these images (Messaris & Abraham, 2001). Recent breakthroughs in computer vision and generative AI offer opportunities to scale visual analysis, yet existing computational approaches often lack theoretical grounding or transparency in how visual meaning is constructed. This study addresses that gap by introducing a three-stage computational framework for theory-driven visual framing analysis.

Our approach combines generative AI, computer vision-based feature extraction, and interpretable statistical modeling. One of the generative AI models used in this study is GPT-4o, a large vision-language model (LVLM) that processes both visual and textual inputs to extract contextually meaningful features from images. Leveraging both the large vision-language model GPT-4o and the feature extraction tool OpenCV, we identify key denotative and semiotic features theorized to shape visual frames—such as police presence, visible contestation, signs of protester solidarity, and color contrast—and link them to higher-order frame classifications: the protest paradigm and solidarity framing. The framework includes: (1) denotative feature definition and validation using LVLMs and human coders, along with semiotic feature extraction using a computer vision tool; (2) model training via logistic regression; and (3) predictive application to unseen imagery. Applied to BLM protest images, our method demonstrates how validated denotative and semiotic features classified by the generative AI model and feature extraction tool can reliably predict frame categories.

The goal of this study is to advance a scalable, interpretable, and theory-rich approach to visual analysis that contributes to the growing intersection of communication theory and AI-assisted computational social science research. It illustrates how generative AI can be responsibly integrated into multi-stage workflows for computational content analysis, advancing both methodological rigor and theoretical insight in the study of visual political

communication.

Auditing Bias in Protest Coverage Through the Lens of Generative AI

The visual framing of real-world events in journalism often carries entrenched biases. As a key component of news reporting, photojournalism does not simply document reality—it constructs narratives through selective framing, composition, and editorial choices. As Sontag (2010) and Debord et al. (2014) have critically observed, photojournalistic images are not without artifice. Traditional photojournalism distorts the audience's perception of events through selective editing, framing, and manipulation.

While scholarship on biases in visual framing has often focused on the misrepresentation or underrepresentation of particular social identities, such as gender, race, and migrants (Abraham & Appiah, 2006; Ferrucci et al., 2013; Gibbons, 2022; Lester & Ross, 2011; Lutz & Collins, 1993), a growing body of literature examines biases in social protest coverage and their effects on public opinion (Corrigall-Brown & Wilkes, 2012; Geise et al., 2023; L. Lu et al., 2025). L. Lu et al. (2025) found that, compared to visual coverage of the Black Lives Matter protests that emphasized conflict, audiences who were shown visual coverage emphasizing solidarity had more positive perceptions of the protesters and heightened engagement intentions. Geise et al. (2023) also found that while conflict between protesters and authorities increased negative emotions in audiences, protest coverage that excluded interactions between protesters and police while emphasizing protester solidarity elicited positive emotions in audiences. These studies reveal two key insights: biases are exhibited in the visual framing of protest coverage, and public opinion can be swayed by these framing techniques.

While these visual narratives—both framing and editorial choices—shape public perception, they also serve as training data for models that generate AI-driven imagery and videos. Recent advances in AI-powered news production, such as automated image selection or AI-generated illustrations, indicate that generative models are increasingly embedded in journalistic workflows (Mahadevan, 2024; Matich et al., 2025). If these models are trained on photojournalistic datasets exhibiting entrenched biases, they risk perpetuating and even amplifying existing framing tendencies in protest coverage. Holding photojournalism accountable and critically examining its biases is a necessary first step in addressing the ethical challenges posed by AI-generated imagery. By first revealing how traditional news photography constructs and reinforces particular narratives, we can better assess whether

generative AI perpetuates or diverges from these established visual biases.

Although visual framing operates on multiple levels—from overt editorial choices to subtle compositional cues (Rodriguez & Dimitrova, 2011)—humans may struggle to distinguish these layers without an emotional response guiding their interpretation. Because images are often processed through an affective lens before they are critically evaluated, the realism heuristic makes it difficult for audiences to recognize the constructed nature of visual framing (Olofsson et al., 2008; Sundar et al., 2021). As a result, subtle compositional cues—such as the placement of police officers in dominant positions or the use of specific color schemes to evoke emotion—can shape interpretation without viewers actively noticing these elements. This perceptual limitation underscores the need for a systematic approach to dissecting framing in photojournalism.

To address this challenge, this study employs a three-stage computational framework that integrates a generative multimodal AI model, a computer vision-based feature extraction tool, and logistic regression to classify protest imagery into protest paradigm and solidarity frames. By analyzing protest visuals at scale, the framework enables the systematic detection of framing biases—such as the overrepresentation of law enforcement, the emphasis on violent interactions over peaceful protest, or portrayals of activists as disruptors rather than political agents. It further provides a fine-grained and reliable classification of subtle visual cues discussed earlier—for example, color tone—that may influence how audiences perceive protest but are often too ambient or nuanced to be examined reliably by human coders, thereby more accurately revealing how they operate within photojournalistic framing.

The Protest Paradigm vs. Solidarity Framing in Media Coverage

One of the biases commonly found in journalistic coverage of protests is the protest paradigm. The protest paradigm refers to the hegemonic journalistic practices of mainstream media outlets in covering protests and social movements that often reinforce the status quo (Gitlin, 2005; Herman & Chomsky, 2010). This paradigm includes several framing devices used by journalists, such as voiding the deeper causes of protest, downplaying the scale of protest, emphasizing the violence of protests, highlighting clashes between police and protesters, stressing the strangeness of protest, and over-relying on officials as information sources—thereby delegitimizing protesters by focusing on their actions, tactics, appearance, behaviors, and

legitimizing authorities (K. Kim & Shahin, 2020; McLeod & Detenber, 1999). Early studies further identified several sub-frames—such as the riot frame (emphasizing violence, chaos, and lawlessness), the confrontation frame (highlighting clashes between protesters and authorities), and the spectacle frame (focusing on the drama or sensational aspects of protest)—to distinguish them from other ways of reporting protest, such as the debate frame, which highlights competing viewpoints and arguments (Harlow & Brown, 2023). Empirical research has also shown that these sub-frames appear more frequently than the debate frame in mainstream media coverage (Harlow et al., 2020). Although most studies examining the protest paradigm have focused on analyzing textual frames in news articles covering protests, research on visual framing within the protest paradigm has also found a tendency to delegitimize protesters in demonstrations while upholding authorities' viewpoints (Corrigall-Brown & Wilkes, 2012; Geise et al., 2023; M. Kim & Bas, 2023; L. Lu et al., 2025; Neumayer & Rossi, 2018).

Yet images, as indexical signs (Ball, 2017), do not function in the same way as words. For instance, unlike verbal reporting, static news images lack a propositional syntax that can explicitly convey causality or generalization (M. Kim & Bas, 2023). This means that certain text-based strategies of delegitimizing or legitimizing protest, such as providing or withholding background causes of protest or relying heavily on official sources, cannot be directly applied to visuals. Instead, in visual practice the protest paradigm is conveyed primarily through the selection of denotative and stylistic/semiotic features—for example, by highlighting scenes of confrontation between protesters and police, or by depicting the presence of weapons and other visual cues that readily link protest to violence—thereby delegitimizing protest (Corrigall-Brown & Wilkes, 2012; Neumayer & Rossi, 2018).

Recent studies have found that these negative portrayals of protests have diminished (Harlow & Brown, 2023; Lee et al., 2013), reflecting increased self-awareness among journalists and the growing maturity of social media platforms, which have enabled more positive portrayals of protesters (Literat et al., 2023). One alternative framing of media coverage that researchers have suggested to portray protesters in a positive light is "solidarity framing," which highlights unity and peaceful protest scenes among protesters (L. Lu et al., 2025). In contrast to the protest frame, this counter-frame typically follows the WUNC model (Geise et al., 2023), emphasizing protest worthiness and unity, the presence of substantial numbers, and demonstrable commitment. Although the indexical nature of photojournalism limits the direct use of certain text-based framing strategies in counter-framing, it

also makes images especially effective in conveying solidarity cues—such as protesters marching with linked arms, carrying joint banners, participating in peaceful gatherings, and other visuals that communicate collective strength and mutual support.

Although scholars have proposed solidarity framing as an alternative to the protest paradigm, and empirical studies demonstrate its positive influence on public perceptions of protest (Geise et al., 2023; L. Lu et al., 2025; Masullo et al., 2024), there remains a gap in the literature regarding systematic analyses of the existence and prevalence of solidarity framing in news coverage compared to the protest paradigm. As digital media platforms have enabled the significant production and circulation of news coverage (Chadwick, 2017), it is becoming increasingly difficult for human researchers to examine visual frames in protest coverage alone. This study proposes a computational method for investigating the prevalence of visual frames reflecting the protest paradigm versus solidarity framing, integrating various computer vision techniques, including multimodal models and feature extraction tools. We introduce a three-stage computational framework to systematically analyze protest images at scale, contributing to advancing computational communication research on visual framing.

A Computational Driven Approach to Analyzing Imagery

Limitations of traditional computer vision approaches

Recent advances in computer vision have enabled researchers to analyze and classify visual content in increasingly sophisticated ways, offering new tools for studying images and videos at scale (Peng et al., 2024). Yet despite these technological developments, automated approaches still fall short of capturing the depth, subtlety, and interpretive nuance that characterize traditional visual framing analyses conducted by human coders. Drawing on foundational work in visual framing research, Rodriguez and Dimitrova (2011) outline four interrelated levels of visual framing: the denotative level, which refers to the (1) literal, observable content of an image; (2) the stylistic/semiotic level, encompassing formal visual elements such as lighting, color, and composition; (3) the connotative level, which addresses the symbolic or emotional meanings derived from denotative and stylistic cues; (4) and the ideological level, which reflects broader cultural, political, or normative values embedded within visual representations. These levels are not distinct but layered—lower-level features support the construction of

higher-level meanings.

Most current computer vision approaches only partially align with this layered structure. They typically focus on (1) detecting denotative features (e.g. Liu et al., 2018), (2) identifying stylistic and semiotic cues (e.g. Sharma & Peng, 2024), and (3) inferring connotative or ideological meanings from visual data (e.g. Y. Lu & Peng, 2024). However, these models often either classify only low-level features or attempt to make high-level inferences without linking them to lower-level elements, thereby skipping the intermediate steps involved in meaning construction.

First, *denotative approaches* focus on detecting the literal content of an image—answering the question, "Who or what is being depicted here?" (Rodriguez & Dimitrova, 2011, p.53)—such as identifying objects, scenes, or people using image classification and object detection models. Classical supervised and deep learning models—such as DeepFace[1] or YOLO[2] —detect objects in visual media. These systems support the identification of denotative elements (e.g., police officers, protest signs, flags), contributing to a surface-level understanding of visual frames. However, such models typically require manual annotation or classification to connect individual objects to broader frames. For example, developing a supervised machine learning classifier to identify a particular object requires 1,000 to 2,000 annotated images (Krizhevsky et al., 2017). These significant labor demands limit the accessibility of such analyses for researchers without sufficient human or computational resources. Generative AI models can ease these challenges by enabling denotative classification with fewer labeled examples, leveraging pre-trained architectures and large-scale datasets to identify objects more efficiently. Researchers can prompt such models directly to detect objects, reducing the burden of extensive manual annotation. Still, beyond resource constraints, denotative approaches often lack contextual sensitivity—e.g., identifying a police officer does not convey whether their presence is framed as protective or threatening—limiting interpretive depth. Still, beyond resource constraints, denotative approaches often lack contextual sensitivity—e.g., identifying a police officer does not convey whether their presence is framed as protective or threatening, which limits interpretive depth.

Second, *stylistic or semiotic approaches* focus on visual modalities such as depth and color contrast (Bell, 2012). Feature extraction tools or aesthetic

[1]DeepFace is a deep learning based facial recognition system developed by Meta researchers that conducts face verification tasks (Taigman et al., 2014).

[2]YOLO is a state-of-the-art object detection model that applies deep learning to identify multiple objects within images in real time.

analysis methods are commonly used here. For instance, Chen et al. (2022) used OpenCV to compare the roles of color and brightness in conspiracy and debunking YouTube videos related to COVID-19 and found significant differences in the color and brightness of their thumbnails. In contrast, generative AI models struggle with such cues because they are typically trained on datasets labeled for denotative (i.e., object/scene) tasks rather than symbolic or stylistic annotation (Dahou et al., 2025; J. Lu et al., 2023). While these approaches capture aesthetic trends, they neither classify denotative elements nor reliably distinguish between stylistic choices made for artistic purposes and those carrying symbolic or ideological weight, limiting their interpretive value for visual framing research. Indeed, the distinction between artistic and symbolic uses of color is often blurred in practice, as aesthetic treatments (e.g., saturation, hue, or contrast) may simultaneously carry symbolic functions. This overlap underscores the difficulty of fully disentangling the two in applied analysis.

Third, *connotative or ideological approaches* aim to associate visual symbols with social meanings and infer higher-order symbolic or emotional interpretations that emerge from the interaction of denotative and stylistic/semiotic cues (Lule, 2004). Some models employ deep learning or multimodal architectures (e.g., combining image and text inputs) to predict latent themes, emotional tone, or ideological framing (e.g., Yadav & Vishwakarma, 2023).

More recently, generative AI models—such as large language models or large multimodal models—have been explored as tools for directly classifying text or images into broader connotative or ideological frames, although most applications have focused on text classification (Pastorino et al., 2024). Generative AI models often leap from low-level features to high-level interpretations without explicitly modeling the layered meaning-making process central to visual framing theory.

For instance, when generative AI models such as GPT-4o are prompted to label ideological or connotative frames directly from images, there is little transparency about which denotative or semiotic features underlie their predictions. GPT-4o may, for example, label an image as reflecting a protest paradigm based solely on the denotative presence of police officers. In doing so, the model infers an ideological frame directly from surface cues while bypassing intermediate layers of meaning—such as whether the police appear in scenes of confrontation or in moments of coordination and empathy. Although researchers can ask the model to provide justifications, it is unclear whether these reflect genuine reasoning or post hoc

rationalizations, raising the risk of hallucination (Ji et al., 2023). Few-shot prompting (i.e., providing the model with a handful of labeled examples to guide its inferences) and lowering the temperature (i.e., a parameter adjustment that reduces randomness in model outputs) may encourage the model to link higher-level frames to lower-level cues more consistently. However, unlike traditional statistical models, the internal weighting of features in GPT-4o remains opaque and less generalizable across contexts. This limitation reflects broader critiques of large language models: while they produce fluent outputs, they often lack real understanding or systematic reasoning (Bender et al., 2021; Marcus & Davis, 2020). As a result, these models generate predictions that are difficult to interpret or justify, raising concerns about transparency, replicability, and theoretical alignment with established communication frameworks.

In some cases, unsupervised machine learning techniques—such as clustering, topic modeling, or dimensionality reduction—are used to group images based on visual or stylistic similarity. Additionally, transfer learning from multimodal models (e.g., CLIP [3]) has further enabled systems to infer latent themes from images without requiring labeled data, offering a more scalable alternative to manual annotation (Radford et al., 2021). These emergent groupings provide a data-driven, inductive path toward identifying connotative themes, with the potential to surface patterns that may not be apparent through deductive coding. However, the interpretability of such inferences remains a subject of debate, and the reliance on human interpretation in post hoc analysis introduces subjectivity. Moreover, these methods may struggle to account for context, cultural nuance, or ideological complexity—factors that are central to framing theory but difficult to capture through unsupervised clustering alone. Table 1 provides an overview of key computer vision approaches used in visual framing analysis.

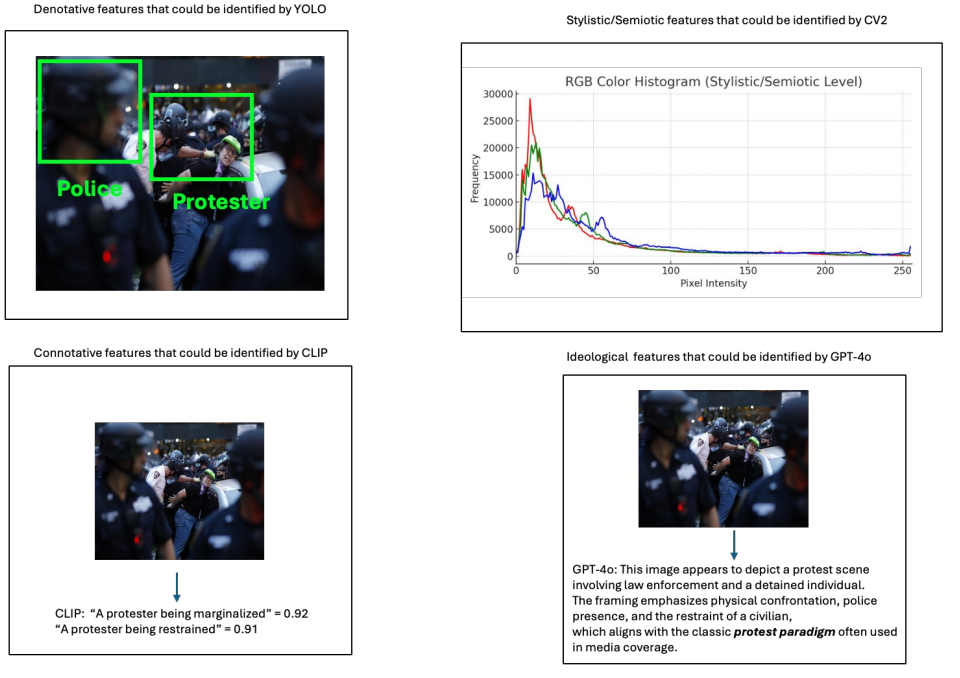
Figure 1 illustrates how each level of visual framing can be identified using YOLO (denotative), OpenCV (semiotic), CLIP (connotative), and GPT-4o (ideological) as representative tools.

Because existing approaches often require substantial manual annotation or rely on researcher-driven interpretation to connect visual elements to frames, there is a growing need for innovative methods that bridge low-level visual features (i.e., denotative and stylistic/semiotic) and higher-order frames (i.e., connotative and ideological) with greater efficiency and conceptual clarity. To address this gap, we propose a three-stage framework.

[3] CLIP (Contrastive Language–Image Pre-training) is a multimodal AI that associates representations of images with text by training on large-scale image–caption pairs.

Levels of visual framing	Key questions	Computer vision approaches	Example CV tools	Strengths	Limitations	Examples of framing elements
Denotative	Who or what is being depicted?	Supervised deep learning models	YOLO, DeepFace, ResNet	High accuracy in identifying people or objects	Requires substantive manual labeling, lacks contextual sensitivity	Protest signs, police officers
Stylistic/Semiotic	How is it visually presented?	Feature extraction, aesthetic analysis	OpenCV	Captures visual modality (contrast, brightness)	Limited interpretive insight	Color, texture, and contrast-based feature analysis
Connotative	What symbolic or emotional meaning is conveyed in images?	Supervised deep learning models, multimodal models	CLIP, fine-tuned CNN models	Can infer themes from latent features	Potential interpretive jumps	Victimization of protesters; marginalization of immigrants
Ideological	What broader values or ideologies are implied?	Latent frame detection, generative multimodal AI, unsupervised clustering	ResNet-based unsupervised ML	Identifies patterns across large datasets, scalable	Often hard to infer the relationships between denotative and stylistic/semiotic elements	Protest paradigm; solidarity framing

Table 1: Overview of computer vision approaches to visual framing analysis



Note. In the upper-left panel, YOLO ("You Only Look Once") is used as a state-of-the-art object detection model to identify and localize police and protesters within the image by drawing bounding boxes around the regions. The upper-right panel presents an RGB color histogram, which shows the distribution of pixel intensities across red, green, and blue channels; this is used at the stylistic/semiotic level to capture color contrast and tonal features. The lower-left panel demonstrates how CLIP (Contrastive Language–Image Pre-training), a vision–language model trained to align images and text, produces similarity scores between the detected visual regions and textual frame labels (e.g., "protester," "police").

Figure 1: Mapping denotative, semiotic, connotative, and ideological frames using computer vision tools.

In the first stage, we detect denotative and semiotic elements grounded in visual framing theory. In the second stage, these features are used to train a statistical model that explains how lower-level visual cues contribute to higher-order framing. In the third stage, the trained model is applied to new images to infer connotative and ideological meanings—such as the protest paradigm or solidarity framing. By aligning with the layered nature of visual framing, this framework more accurately reflects how meaning is constructed in human visual interpretation and enables theory-driven computational classification of visual frames.

A three-stage computational framework for theory-driven visual framing research

A major shift in computer vision approaches to identifying denotative elements has emerged with the introduction of generative multimodal AI models—such as large vision-language models (LVLMs)—which enable researchers to move beyond descriptive, inductive methods. These models process both image and text data, facilitating a new class of deductive, theory-driven computational analyses. Specifically, LVLMs allow researchers to detect targeted denotative elements—such as the presence of police or protest signs—through prompting, eliminating the need for extensive human labeling and enabling more scalable and theoretically grounded visual coding.

Therefore, the first stage of our three-stage computational framework involves identifying denotative and stylistic/semiotic elements based on the literature review, developing a coding book to guide both human coders and prompts for generative multimodal AI models, and applying these models to a subset of images to calculate human–machine intercoder reliability. To support this process, a consensus on the "ground truth" should be established among human coders, and the stability of AI-generated coding outputs should be tested across multiple iterations of the generative multimodal models.

For instance, if we decide to code "the presence of police" as a denotative feature of the protest paradigm, there should be: (1) agreement among human coders on which images contain police presence, (2) consistency across different prompting outputs from the generative AI model in detecting police presence, and (3) high reliability between human and AI outputs for this feature—before applying the AI model to code this element across the full dataset. This ensures the reliability and stability of the model's classification of denotative features. For stylistic/semiotic features, we encourage

researchers to use feature extraction tools (e.g., OpenCV) over generative AI models to determine color contrast and saturation, as feature extraction tools provide precise, objective pixel-level measurements that are consistent and not affected by interpretive variation.

While the identification and validation of denotative and stylistic/semiotic features are essential for grounding visual analysis in theory, this alone does not explain how these elements combine to shape broader narrative constructions. Although visual framing theory suggests that such features contribute to higher-order frames—such as the protest paradigm or solidarity framing—determining the relative influence of each element is challenging for human coders. This is due in part to the indexical nature of images, which evoke meaning intuitively and holistically in human perception (Barthes & Heath, 2006), making it difficult to analytically isolate and evaluate the contribution of individual components.

To address this limitation, the second stage of our framework employs a logistic regression model to estimate how combinations of denotative and semiotic elements predict the presence of higher-order frames. We construct a training dataset in which these lower-level features, extracted through generative AI and feature extraction tools, are paired with frame labels (e.g., protest paradigm, solidarity framing) assigned by human coders. This statistical approach enables us to quantify the relationship between visual components and the frames they construct, offering a transparent, interpretable, and replicable methodology for visual framing research.

The third stage involves using the trained model to analyze a new set of images. By applying the relationships learned in the previous step—such as how the presence of police or certain color tones relate to specific frames—we can see how well the model can predict the type of framing in images it has not been trained before. This step helps us test whether the combination of features identified by generative AI and tools like OpenCV can be used to automatically classify images into higher-level frames like the protest paradigm or solidarity framing. In doing so, this stage demonstrates the potential for fully automated analysis of visual framing—from detecting low-level visual elements to interpreting broader ideological meanings. Table 2 summarizes the key tasks, tools, and levels of human involvement across each stage of the proposed three-stage computational framework for visual framing analysis.

In sum, our three-stage computational framework responds to a lingering gap in visual framing research: the need for a scalable yet theory-driven method that connects low-level visual features with higher-order frame in-

Stage	Key tasks	Tools	Human involvement
1. Identify denotative and semiotic elements	• Develop coding book based on the literature review • Identify denotative and stylistic/semiotic elements • Apply generative AI models • Extract visual features	• Generative multimodal AI models • Feature extraction tools	• Define ground truth • Test intercoder reliability between human and the AI models
2. Estimate relationship between lower-level and higher-level frames	• Pair images with higher frame labels from human coders (e.g., protest paradigm, solidarity framing) • Train a regression model to link visual features to frames	• Logistic regression model	• Train and evaluate model using human-labeled data
3. Predict frames in a new set of images	• Apply trained model to new, unlabeled images • Predict higher-order frames from images	• Apply trained logistic regression model to the test dataset	• Minimal (model application only)

Table 2: A three-stage computational framework for visual framing analysis.

terpretation. Existing approaches often treat denotative, semiotic, and ideological elements in isolation or rely on human intuition without systematic modeling. By combining generative AI models, feature extraction tools, and interpretable statistical techniques, our framework enables a layered analysis of visual meaning that mirrors how frames are constructed in human perception. This approach not only improves consistency and scalability but also advances the theoretical integration of computational methods in framing research. The following section outlines how the proposed framework is applied to classify protest paradigm and solidarity framing in the context of Black Lives Matter protest imagery.

Current Study

Building on the three-stage computational framework outlined above, the current study applies this approach to the analysis of protest imagery related to the Black Lives Matter movement. While prior work has examined protest framing through either manual coding or limited object detection tools, few studies have integrated theory-driven visual features, generative AI models, and statistical modeling in a unified, scalable framework. This study aims to fill that gap by demonstrating how denotative and stylistic/semiotic elements—extracted through multimodal generative AI models and feature extraction tools—can be used to predict higher-order frames such as the protest paradigm and solidarity framing.

We focus on two datasets comprising images from U.S. news sources that covered Black Lives Matter protests. The first dataset, drawn from the GDELT Visual Global Knowledge Graph ($n = 245$), is used to define and validate visual features, develop a coding book, test reliability between human coders and generative AI outputs, and train a logistic regression model to estimate the relationship between visual features and framing categories. The second dataset, compiled from Google Images ($n = 1{,}536$), allows us to apply the trained model to unseen images and evaluate its predictive performance in classifying protest images along theoretically meaningful dimensions. Through this implementation, the study not only tests the effectiveness of a novel computational pipeline but also demonstrates how a theory-driven, partially automated approach can improve the consistency, scalability, and interpretability of visual framing analysis in protest research.

Method and Results

Data sources

GDELT dataset

We extracted 245 images from the Global Database of Events, Language, and Tone (GDELT), a large-scale open-source platform continuously monitoring global news content across multiple formats. We conducted a targeted search using the combination of keywords "Black Lives Matter" and "protest," restricted the image tag to "protest," and limited the source domains to sixteen major U.S. news outlets classified along an ideological spectrum (Faris et al., 2020; Wang et al., 2024)(see Appendix A of supplementary materials for specifics). The search covered the period from January 1, 2017, to August 1, 2024, yielding 459 news articles. After excluding 61 articles whose images were unavailable at the time of access and 153 images unrelated to BLM protests, we retained 245 images.

Google Image Search Dataset

To assess the effectiveness and scalability of the approach, we collected a large sample of 1,536 images from Google Image Search using Serp API, again focusing on the keyword "Black Lives Matter" across sixteen major U.S. news outlets classified along an ideological spectrum (see Appendix B of supplementary materials for specifics). This dataset was used for out-of-sample prediction and model evaluation.

Stage 1. Feature extraction and coding validation

To begin, we developed a theory-informed coding book grounded in visual framing and protest imagery literature (Rodriguez & Dimitrova, 2011), identifying key denotative (i.e., presence of police, protester presence, contestation between police and protesters, existence of weapon, existence of solidarity actions, existence of masked officials) and semiotic (i.e., color contrast, saturation) elements. We used GPT-4o, a large vision-language model (LVLM), to extract denotative features via structured prompts. For stylistic and semiotic features, we employed OpenCV, a Python-based image processing library, to objectively extract color information at the pixel level and compute measures of color saturation and contrast. These values were calculated using standardized image processing techniques to ensure consistency across the dataset.

To validate the reliability of denotative feature coding, we randomly selected 40 images from the GDELT dataset. Two trained human coders independently annotated these images, and disagreements were resolved through discussion to establish a consensus "ground truth" (see Appendix C of supplementary materials for details). To assess the intra-prompt stability of GPT-4o's performance (Barrie et al., 2024), we ran the model on the same 40 images across 20 iterations. We then compared GPT-4o's most frequently generated outputs, determined through majority voting, to the human-coded ground truth to evaluate its classification accuracy and consistency (see Appendix D of supplementary materials for details). Table 3 illustrates the denotative and semiotic elements used to classify protest imagery.

Cohen's Kappa was used to assess agreement across coding items between GPT-4o's outputs and the human-coded ground truth. The results indicated strong consistency, with Kappa values ranging from 0.75 to 1.00, reflecting substantial to almost perfect agreement. Table 4 presents the intercoder reliability scores (Cohen's Kappa) between the ground truth established by human coders and the most representative outputs from the GPT-4o model. After establishing reliability across the denotative coding categories, the GPT-4o model was applied to code the denotative elements of all 245 images [4].

Stage 2. Frame labeling and model construction

Next, two human coders classified each of the 245 images according to whether they reflected the protest paradigm (0/1) or solidarity framing (0/1). To avoid conflating denotative or stylistic cues with these higher-order connotative and ideological frames, we adopted a holistic coding approach that followed the conceptual definitions of each frame in the literature.

For the protest paradigm, images were coded if they emphasized confrontation (e.g., violence or disruption), framed protests as criminal or deviant, relied heavily on official sources over protesters, or delegitimized protesters by portraying them as ignorant, strange, or ineffective. By contrast, solidarity framing was coded when images centered affected voices, challenged established power structures, fostered unity and shared goals,

[4] Although GPT achieved high intercoder reliability in our application to protest imagery, this result should not be interpreted as a guarantee of reliability across different datasets or framing contexts. Because denotative and semiotic cues can be interpreted differently across domains, intercoder reliability must be re-established each time the framework is applied to new areas of research. Our findings therefore demonstrate feasibility in this domain, while highlighting the need for replication to assess generalizability.

Category	Definition	Coding Scheme	Computational Tools	Intercoder Reliability within GPT-4o
Police presence	code the police existence in the image.	• 0 = no existence • 1 = existence	GPT-4o	Krippendorff's Alpha: 0.90
Protester presence	code the protester existence in the image.	• 0 = no existence • 1 = existence	GPT-4o	Krippendorff's Alpha: 0.84
Contestation between police and protesters	code when there is a conflict between the protester and the police.	• 0 = no existence • 1 = existence	GPT-4o	Krippendorff's Alpha: 0.85
Existence of weapon	code the existence of the weapon in the image.	• 0 = no existence • 1 = police holds the weapon • 2 = protester holds the weapon	GPT-4o	Krippendorff's Alpha: 0.78
Existence of solidarity actions	code when there is a solidarity action.	• 0 = no existence • 1 = existence	GPT-4o	Krippendorff's Alpha: 0.77
Existence of masked officials	code when there is a masked official.	• 0 = no existence • 1 = existence	GPT-4o	Krippendorff's Alpha: 0.77
Color contrast	calculate the contrast of the image.	numeric	OpenCV	N/A
Saturation	calculate the saturation of the image.	numeric	OpenCV	N/A

Table 3: Denotative and stylistic/semiotic elements for classifying protest imagery into protest paradigm and solidarity framing

Category	Intercoder Reliability between Human Coders and GPT-4o
Police presence	Cohen's Kappa: 1
Protester presence	Cohen's Kappa: 1
Contestation between police and protesters	Cohen's Kappa: 0.75
Existence of weapon	Cohen's Kappa: 0.88
Existence of solidarity actions	Cohen's Kappa: 0.88
Existence of masked officials	Cohen's Kappa: 0.83

Table 4: Intercoder reliability between human coders and GPT-4o across denotative elements.

highlighted mutual support among participants, or addressed systemic injustices beyond individual events.

We also created a third category, "other", to classify images that did not clearly align with either the protest paradigm or solidarity framing. For example, one image only showed the back of a single protester's head and body, while another depicted a blurred crowd of protesters. In both cases, the images provided insufficient information to determine whether the protest was delegitimized (e.g., through a focus on conflict) or legitimized (e.g., through collective solidarity). The three coding categories were mutually exclusive. When an image contained potentially contradictory elements, we prioritized the most salient ones (foregrounded or dominating the visual focus) to determine the overall frame classification. During coding, images were separated from their source information to avoid potential coder bias linked to specific media outlets. In addition, both coders' racial and national identities were situated outside the dynamics at the center of the BLM protests. This positional distance helped reduce the likelihood that U.S.-specific identity heuristics shaped coding decisions, allowing judgments to remain anchored in the codebook's frame definitions and frame-relevant visual cues.

Initial coding agreement between the two coders was 95.5% for the protest paradigm (Cohen's Kappa = 0.87) and 91.8% for solidarity framing (Cohen's Kappa = 0.81). Disagreements between two human coders were subsequently resolved through discussion to reach consensus and establish the ground truth. These frame labels were then used as dependent variables in logistic regression models, with the denotative and stylistic/semiotic visual features—coded by the GPT-4o model and OpenCV—serving as predictors.

The models were designed to quantify the contribution of specific denotative and semiotic elements to the likelihood of an image being framed according to the protest paradigm or solidarity framing.

Logistic regression explaining the protest paradigm.

To assess how specific visual elements contribute to the presence of the protest paradigm, we used denotative and semiotic features extracted from 245 GDELT images as independent variables, and the existence of the protest paradigm in the image (0/1) was a binary dependent variable for a logistic regression. This approach allows for estimating the relative contribution of each feature—such as the presence of police, protest signs, contestation between the police and protester(s), dominant color, and saturation—to the likelihood that an image reflects the protest paradigm.

We found that the presence of police and the presence of masked officials were highly correlated in the logistic regression model predicting the protest paradigm, with variance inflation factor (VIF) scores exceeding 7. To address multicollinearity, we excluded the presence of masked officials from the models. Table 5 presents the logistic regression results explaining variance in the protest paradigm frame. Specifically, the presence of police ($B = 3.45$, $SE = 0.60$, $p < .001$) and the contested action between the police and the protester(s) ($B = 3.74$, $SE = 0.81$, $p < .001$) emerged as significant predictors of the protest paradigm. Specifically, images featuring police were over 30 times more likely to be framed within the protest paradigm compared to those without, and contestation between the police and protester(s) was associated with more than a 43-fold increase in the likelihood of such framing. The logistic regression model demonstrated excellent explanatory power, with a McFadden's R^2 of 0.55, indicating that the model accounts for a substantial portion of the variance in visual framing classifications.

For instance, as illustrated in Figure 2—a real protest image annotated with key visual elements—the likelihood that the image is classified within the protest paradigm increases when the GPT-4o model detects police presence and identifies contestation between the police and protester(s).

Logistic regression explaining the solidarity framing.

A logistic regression model was also employed to predict the presence of the solidarity framing (coded as 0/1). We also found that the presence of police and the presence of masked officials were highly correlated in the logistic regression model predicting the solidarity framing, with variance

Predictor	B(estimate)	SE	z	OR(expB)	p-value
Intercept*	-5.69	2.38	-2.39	0.01	0.02
Police presence***	3.45	0.60	5.79	31.41	$<$ .001
Protester presence	-1.24	1.04	-1.19	0.29	0.24
Contestation between police and protesters***	3.74	0.81	4.63	43.32	$<$ .001
Existence of police with a weapon (0 = no existence)	0.02	1.15	0.02	1.02	0.99
Existence of a protester with a weapon (0 = no existence)	18.35	1150	0.02	94M	0.99
Existence of solidarity action	-0.48	0.76	-0.64	0.62	0.53
Color contrast	0.05	0.03	1.92	1.05	0.05
Color saturation	0.01	0.01	1.11	1.01	0.27
McFadden's R^2 = 0.55					

Table 5: Multivariate logistic regression results explaining the presence of the protest paradigm frame.

Figure 2: Example of protest paradigm framing with detected visual features.

inflation factor (VIF) scores exceeding 7. To address multicollinearity, we excluded the presence of masked officials from the models. Table 6 presents the logistic regression results explaining the variance in the solidarity framing. Specifically, the presence of police (B = -3.13, SE = 0.55, $p < .001$), the contested action between the police and the protester(s) (B = -2.16, SE = 0.69, $p < .001$), the existence of solidarity action (B = 1.91, SE = 0.68, p = 0.005) emerged as significant predictors of the solidarity framing. Images featuring police decreased the odds of solidarity framing by 96%, and the contestation also significantly reduced the odds by 88%. Images depicting solidarity actions were nearly seven times more likely to be framed within the solidarity frame than those without such actions. The logistic regression model demonstrated excellent explanatory power, with a McFadden's R^2 of 0.46, suggesting that the model captures a substantial portion of the variance in framing classifications.

For instance, as illustrated in Figure 3—a real protest image annotated with key visual elements—the likelihood that the image is classified within the solidarity frame increases when the GPT-4o model detects solidarity actions by protesters and does not detect police presence or contestation between police and protesters.

Stage 3. Out-of-sample prediction and evaluation

To evaluate these logistic regression models' predictive capacity, we applied these models to the second dataset of 1,527 images collected from Google Images. The same GPT-4o and OpenCV pipelines were used to extract features from this dataset. The trained logistic regression models were then

Predictor	B(estimate)	SE	z	OR(expB)	p-value
Intercept	0.41	1.81	0.23	1.50	0.82
Police presence***	-3.13	0.55	-5.74	0.04	$<$ .001
Protester presence	1.82	1.06	1.71	6.14	0.09
Contestation between police and protesters**	-2.16	0.69	-3.13	0.12	0.002
Existence of police with a weapon (0 = no existence)	1.24	1.15	1.07	3.44	0.29
Existence of a protester with a weapon (0 = no existence)	-16.15	1239.74	-0.01	0.00	0.99
Existence of solidarity action**	1.91	0.68	2.83	6.78	0.005
Color contrast	-0.01	0.02	-0.68	0.99	0.50
Color saturation	-0.01	0.01	-1.55	0.99	0.12
McFadden's R^2 = 0.46					

Table 6: Multivariate logistic regression results explaining the presence of the solidarity frame.

Figure 3: Example of protest paradigm framing with detected visual features.

used to generate predicted frame labels for each image. We evaluated the model's performance by comparing its predictions to manually coded labels for a validation subset of 200 images, consisting of 100 images predicted as a protest paradigm and 100 predicted as a solidarity frame. This allowed us to assess classification accuracy and generalizability across both categories. This final stage demonstrates the feasibility of a fully or semi-automated pipeline for classifying visual frames based on theoretically grounded visual cues. Performance was evaluated using standard metrics, including precision, recall, F1 score, and overall accuracy for each frame classification.

For protest paradigm classification, the model achieved a precision of 0.88, recall of 0.98, F1 score of 0.93, and accuracy of 0.93. For solidarity frame classification, precision was 0.86, recall 0.98, F1 score 0.91, and accuracy 0.92. These results indicate strong performance for both classifications, with particularly high recall and F1 scores suggesting the model is effective at correctly identifying images within each frame. The high accuracy (above 90%) further supports the model's reliability in distinguishing between protest paradigm and solidarity frame imagery. Figure 3 presents representative images classified under the protest paradigm and solidarity frame, based on the regression models. These examples visually illustrate how the model operationalizes frame detection using theoretically grounded cues such as police presence, contestation, and solidarity actions. Together, these findings demonstrate the feasibility of a semi-automated pipeline for visual frame classification, with robust predictive performance and interpretable decision criteria grounded in framing theory. Figure 4 illustrates how logistic regression models classify example images into the protest paradigm or solidarity framing.

Figure 4: Visual examples of predicted protest paradigm and solidarity framing by logistic regression models.

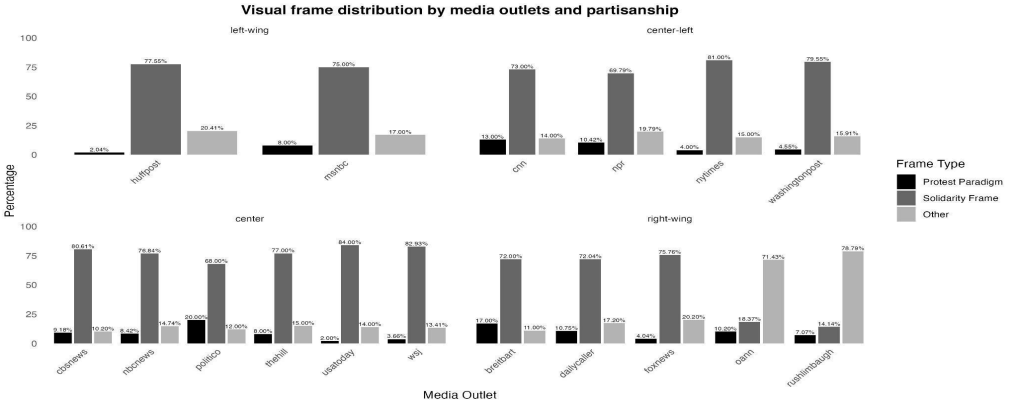
Beyond model performance, this semi-automated classification approach —combining a multimodal generative AI model with computer vision-based feature extraction—also enables scholars to investigate broader research questions, such as whether the partisanship of news outlets is associated with differences in the prevalence of protest paradigm versus solidarity frame usage. Figure 5 illustrates that solidarity framing appears more frequently than the protest paradigm across different news outlets in Google Images search results—regardless of partisanship—when reporting on Black Lives Matter. This pattern reflects a broader shift in visual protest coverage, as recent scholarship has noted: a movement away from protest paradigm framing and toward solidarity framing, in which protesters are portrayed as more empowered and legitimate actors (Literat et al., 2023).

A Pearson's chi-square test of independence indicated that the distribution of frame types (protest paradigm vs. solidarity framing) differed significantly by partisanship, $\chi^2(3, N = 1{,}173) = 16.62$, $p < .001$. Standardized residuals revealed that right-wing outlets relied on protest paradigm frames more than expected, whereas left-wing outlets emphasized solidarity frames more than expected. Center and center-left outlets exhibited distributions that were closer to the overall average.

Discussion

This study presents a novel methodological framework that leverages generative AI and computer vision-based feature extraction to conduct theory-driven visual framing analysis with high interpretability and empirical rigor. Moving beyond one-shot applications of generative AI, our approach integrates it into a multi-stage pipeline that combines feature extraction, human validation, and logistic regression. Building on a three-stage computational framework—feature definition and validation, model training, and predictive application—this study demonstrates how generative AI can be effectively embedded within a scalable and replicable approach to empirical visual framing research.

When applied to protest imagery related to the Black Lives Matter movement, this framework reveals that the presence of police and contested action between police and protester(s) are key denotative elements contributing to the construction of the protest paradigm in visual framing. Interestingly, these same elements also influence solidarity framing, as their absence increases the likelihood of an image being classified as solidarity-framed, particularly when accompanied by visible acts of protester unity. The logistic regression models also demonstrate strong predictive power



Note. Although the Rush Limbaugh Show was primarily a radio program, its companion website provided article-style content with photographs, which served as the basis for visual framing coverage.

Figure 5: Visual frame distribution by media outlets and associated partisanship.

when applied to a new set of images, showing promise for the scalability and replicability of computer-assisted visual coding. Theoretically, this study demonstrates how the four levels of visual framing proposed by Rodriguez and Dimitrova (2011), along with their interdependent relationships, can be explicated through computational methods. Despite extensive research on visual framing, few studies have bridged this literature with computer vision. This disconnection has limited the development of computational approaches grounded in theory. This study opens up an exciting new avenue for revisiting and explicating established communication concepts through advanced computational methods, particularly by leveraging generative AI tools to bridge theoretical frameworks with scalable, data-driven analysis.

In addition to introducing a new visual analytic approach, our study provides fresh insights into how contemporary media frame protest events. Notably, across all three partisan categories—left, center, and right—the share of solidarity frames consistently exceeded that of protest frames, although right-wing outlets relied more heavily on protest-paradigm frames than did left-wing outlets.

These findings resonate with recent arguments about the protest paradigm (e.g., Hayes & O'Neill, 2021), suggesting that the long-standing dominance of protest framing may be increasingly challenged in the social media era. Several factors may account for this shift. The legitimacy of anti-racist movements has gained broad social and moral recognition in Western societies; at the same time, the circulation of these movements through digital activism—such as hashtag campaigns—has enabled them to mobilize large-scale participation and solidarity in short periods, thereby reinforcing their perceived legitimacy and visibility (Duvall & Heckemeyer, 2018). As a result, even conservative media outlets, when confronted with such highly visible movements, often refrain from overtly relying on the traditional protest paradigm to undermine their legitimacy. Future research should also consider the temporal dynamics of these frames, examining whether solidarity and protest framing vary over the course of protest cycles—for instance, between moments of peak mobilization and subsequent decline—in order to capture how framing evolves in relation to political and media contexts.

However, despite the strengths of this framework, potential concerns about bias and hallucination in generative AI outputs must be considered. To assess and mitigate bias, we evaluated human–machine reliability using Cohen's Kappa, which showed substantial agreement between human-coded and AI-generated denotative elements. While this suggests that the model's outputs align well with human judgment, it is essential to acknowledge that

both human and machine annotations may carry underlying biases that cannot be fully eliminated. To mitigate hallucination, a known limitation of large generative models, we implemented a stability measure that involved running 20 iterations of the same prompt, calculating Krippendorff's alpha to assess intra-prompt consistency, and selecting the most frequently occurring elements as the model's final output via majority voting. This approach substantially reduces the influence of outlier responses and increases the reliability of GenAI-assisted annotation, offering a methodologically grounded way to counteract hallucination in computational content analysis.

While this study demonstrates the potential of generative AI to advance theory-driven visual analysis, our study also has limitations. Our analysis focuses exclusively on two specific framing categories—protest paradigm and solidarity framing—within the context of Black Lives Matter protests in the United States. This narrow scope limits the generalizability of our findings to other protest movements, cultural contexts, or visual framing dimensions. In addition, because protest imagery is often politically charged and emotionally resonant, automating its interpretation raises ethical considerations related to bias, cultural representation, and power. Although our use of human–machine reliability measures and stability protocols mitigates some of these concerns, context-dependent assumptions shape both AI models and human coders.

Relatedly, because our coding scheme is designed to capture the core denotative and stylistic/semiotic features of protest and solidarity frames, it necessarily leaves out certain elements that may also shape perception. These include contextual cues external to the image (e.g., headlines or accompanying narratives), which audiences often interpret alongside visuals (Brantner et al., 2011). Within the visual domain, we also recognize additional dimensions that can influence public perceptions of protest imagery.

For example, stylistic or semiotic features such as shot scale and vantage point (long/wide vs. close-up; eye-level vs. high/low angles) or facial expressions (e.g., surprise, happiness, anger, sadness, fear, disgust) may further shape interpretations in ways that either marginalize protesters (as in the protest paradigm) or invite empathy and solidarity (as in solidarity framing). From a social semiotic perspective (Feng & O'Halloran, 2012), such features play a meaningful role in how images construct meaning. A distant, high-angle shot of a large crowd, for instance, can evoke detachment and accentuate disorder, reinforcing protest-paradigm narratives. By contrast, a close-up at eye level that emphasizes linked arms or empathetic expressions can humanize protesters, foreground solidarity cues, and foster more

supportive interpretations (Hayes & O'Neill, 2021).

While these additional denotative and semiotic elements open promising avenues for future research, our analysis centers on features most firmly established in the literature on visual framing in protest imagery, as our aim is to advance a theory-driven computational framework for analyzing protest visuals in scale. Our coding scheme therefore concentrates on features more clearly established in prior research (e.g., emphasis on conflict), ensuring analytical consistency and comparability. Future research should extend beyond these established categories to examine how additional visual dimensions interact with core features in shaping perceptions of protest within the protest paradigm and solidarity framing.

Additionally, our use of logistic regression—though valuable for its interpretability—models only linear relationships between visual features and higher-order frames. While this enables a clear assessment of how denotative and semiotic elements influence framing classifications, more sophisticated modeling approaches (e.g., ensemble methods or neural networks) could improve predictive performance and capture complex, nonlinear dynamics. Despite these limitations, the study establishes a strong foundation for extending theory-informed, AI-assisted visual analysis into new domains and methodological frontiers.

Building on this foundation, future work can extend our framework to a wider range of communication phenomena—including political advertising, health campaigns, and user-generated content—to test the flexibility of generative AI in different domains. Because our model was trained on a specific domain (protest imagery) within a particular context (Black Lives Matter), its predictive performance may not generalize uniformly across other datasets or contexts. However, the methodological pipeline (i.e., the three-stage framework) proposed in this paper is adaptable and can be applied to other domains with appropriate contextual grounding.

Applying our framework in such cases would require scholars to identify domain-specific visual elements and map them onto the four interrelated levels we outlined: denotative, stylistic/semiotic, connotative, and ideological, as well as to understand how denotative and stylistic/semiotic elements comprise connotative and ideological frames through logistic regression models. For example, in studying protests or political advertisements that emphasize patriotism, national flags could serve as denotative cues, close-up shots of those flags as stylistic/semiotic features, unity or sacrifice as connotative meanings, and appeals to nationalism as ideological commitments. This demonstrates how applying the framework in a contextually

grounded way can improve classification accuracy while preserving its theoretical structure. Similarly, extending the framework to social media would require accounting for the distinctive visual repertoires of lay photographers, including selfies, memes, and amateur documentation.

The model's predictive accuracy can also be used to assess the need for incorporating additional visual framing elements. If predictions of denotative and semiotic features do not adequately support the inference of higher-level connotative and ideological frames, this gap signals the need to expand the feature space. Such gaps may be reflected in low power within logistic regression models or reduced predictive accuracy when tested on larger datasets. In these cases, incorporating stylistic or compositional cues—such as camera angle, crowd density, or facial expressions—can help strengthen the link between surface-level features and deeper framing processes.

Beyond advancing automated visual content analysis, our approach also contributes to the study of framing effects. Because the pipeline produces transparent, theory-anchored classifications of visual frames, it can be used to generate experimental stimuli and survey materials that systematically vary in different levels of visual features. This opens opportunities for researchers to test how audiences respond to denotative, semiotic, and connotative cues—for instance, whether close-up images of protesters elicit more empathy than distant crowd shots, or whether police presence consistently triggers threat perceptions. In this way, the framework supports cumulative research that links content patterns to their downstream effects on public opinion.

Methodologically, the integration of generative AI with transparent modeling techniques like logistic regression offers a path forward for scholars seeking to combine computational scale with theoretical depth. Further exploration of multi-label classification, classification using multimodal information, or fine-tuning domain-specific vision-language models could enhance both the precision and nuance of automated visual analysis. As the field of communication increasingly engages with AI tools, our study contributes to an emerging research agenda that seeks not only to adopt new technologies but to embed them within theory-rich, ethically aware frameworks for empirical inquiry.

Acknowledgement

This work was supported by an Arts and Humanities Initiative (AHI) Standard Grant from the University of Iowa Office of the Vice President for Research. This manuscript was developed through the University of Iowa's

Center for Social Science Innovation's Write on Target program, and we especially thank Ethan Rogers for his suggestions and contributions to improving the manuscript.

Supplementary materials for this work may be found at:
https://osf.io/8ymuw/overview?view_only=1559d2e39010470b96712478e94cdad6

References

Abraham, L., & Appiah, O. (2006). Framing news stories: The role of visual imagery in priming racial stereotypes. *Howard Journal of Communications*, *17*(3), 183–203. https://doi.org/10.1080/10646170600829584

Ball, C. (2017). Realisms and indexicalities of photographic propositions. *Signs and Society*, *5*, S154–S177. https://doi.org/10.1086/690032

Barrie, C., Palaiologou, E., & Törnberg, P. (2024). Prompt stability scoring for text annotation with large language models [Version Number: 2]. https://doi.org/10.48550/ARXIV.2407.02039

Barthes, R., & Heath, S. (2006). *Image - music - text: Essays* (1. paperback ed., 28.[print.]). Hill; Wang.

Bell, P. (2012). Content analysis of visual images. In J. Hughes (Ed.), *Sage visual methods* (pp. 31–57). SAGE Publications.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

Chadwick, A. (2017). *The hybrid media system: Politics and power* (2nd edition). Oxford university press.

Chen, K., Kim, S. J., Gao, Q., & Raschka, S. (2022). Visual framing of science conspiracy videos: Integrating machine learning with communication theories to study the use of color and brightness. *Computational Communication Research*, *4*(1). https://doi.org/10.5117/CCR2022.1.003.CHEN

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

Dahou, Y., Huynh, N. D., Le-Khac, P. H., Para, W. R., Singh, A., & Narayan, S. (2025). Vision-language models can't see the obvious [Version Number: 1]. https://doi.org/10.48550/ARXIV.2507.04741

Debord, G., Knabb, K., & Debord, G. (2014). *The society of the spectacle*. Bureau of Public Secrets.

Duvall, S.-S., & Heckemeyer, N. (2018). #BlackLivesMatter: Black celebrity hashtag activism and the discursive formation of a social movement. *Celebrity Studies*, *9*(3), 391–408. https://doi.org/10.1080/19392397.2018.1440247

Faris, R., Clark, J., Etling, B., Kaiser, J., Roberts, H., Schmitt, C., Tilton, C., & Benkler, Y. (2020). Polarization and the pandemic: American political discourse, march – may 2020. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3721653

Feng, D., & O'Halloran, K. L. (2012). Representing emotive meaning in visual images: A social semiotic approach. *Journal of Pragmatics*, *44*(14), 2067–2084. https://doi.org/10.1016/j.pragma.2012.10.003

Ferrucci, P., Tandoc, E. C., Painter, C. E., & Leshner, G. (2013). A black and white game: Racial stereotypes in baseball. *Howard Journal of Communications*, *24*(3), 309–325. https://doi.org/10.1080/10646175.2013.805971

Geise, S., Heck, A., & Panke, D. (2023). "shiny happy people laughing": The protest paradigm, WUNC, and the visual framing of political activism. *Visual Communication Quarterly*, *30*(2), 90–105. https://doi.org/10.1080/15551393.2023.2196631

Gibbons, S. (2022). Gender on the agenda: Media framing of women and women of color in the 2020 u.s. presidential election. *Newspaper Research Journal*, *43*(1), 102–128. https://doi.org/10.1177/07395329221077253

Gitlin, T. (2005). *Whole world is watching: Mass media in the making and unmaking of the new left* (T. Gitlin, Ed.). University of California Press. https://doi.org/10.1525/9780520352438

Harlow, S., & Brown, D. K. (2023). A new protest paradigm: Toward a critical approach to protest news analyses. *The International Journal of Press/Politics*, *28*(2), 333–343. https://doi.org/10.1177/19401612231153377

Harlow, S., Brown, D. K., Salaverría, R., & García-Perdomo, V. (2020). Is the whole world watching? building a typology of protest coverage on social media from around the world. *Journalism Studies*, *21*(11), 1590–1608. https://doi.org/10.1080/1461670X.2020.1776144

Hayes, S., & O'Neill, S. (2021). The greta effect: Visualising climate protest in UK media and the getty images collections. *Global Environmental Change*, *71*, 102392. https://doi.org/10.1016/j.gloenvcha.2021.102392

Herman, E. S., & Chomsky, N. (2010). *Manufacturing consent: The political economy of the mass media* [OCLC: 1004572220]. Vintage Digital.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38. https://doi.org/10.1145/3571730

Kim, K., & Shahin, S. (2020). Ideological parallelism: Toward a transnational understanding of the protest paradigm. *Social Movement Studies*, *19*(4), 391–407. https://doi.org/10.1080/14742837.2019.1681956

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Lee, N.-J., Shah, D. V., & McLeod, J. M. (2013). Processes of Political Socialization: A Communication Mediation Approach to Youth Civic Engagement. *Communication Research*, *40*(5), 669–697. https://doi.org/10.1177/0093650212436712

Lester, P. M., & Ross, S. D. (Eds.). (2011). *Images that injure: Pictorial stereotypes in the media* (3rd ed). Praeger. https://doi.org/10.5040/9798400668470

Literat, I., Boxman-Shabtai, L., & Kligler-Vilenchik, N. (2023). Protesting the protest paradigm: TikTok as a space for media criticism. *The International Journal of Press/Politics*, *28*(2), 362–383. https://doi.org/10.1177/19401612221117481

Liu, C., Tao, Y., Liang, J., Li, K., & Chen, Y. (2018). Object detection based on YOLO network. *2018 IEEE 4th Information Technology and Mechatronics Engineering Conference (ITOEC)*, 799–803. https://doi.org/10.1109/ITOEC.2018.8740604

Lu, J., Rao, J., Chen, K., Guo, X., Zhang, Y., Sun, B., Yang, C., & Yang, J. (2023). Evaluation and enhancement of semantic grounding in large vision-language models [Version Number: 2]. https://doi.org/10.48550/ARXIV.2309.04041

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., & 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

Lule, J. (2004). War and its metaphors: News language and the prelude to war in iraq, 2003. *Journalism Studies*, *5*(2), 179–190. https://doi.org/10.1080/1461670042000211168

Lutz, C. A., & Collins, J. L. (1993). *Reading "national geographic"*. the University of Chicago press.

Mahadevan, A. (2024). Ai is already reshaping newsrooms, ap study finds. *Poynter*. https://www.poynter.org/ethics-trust/2024/artificial-intelligence-transforming-journalism/

Marcus, G., & Davis, E. (2020). *Gpt-3, bloviator: Openai's language generator has no idea what it's talking about* [Accessed: 2025-11-10]. https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion

Masullo, G. M., Brown, D. K., & Harlow, S. (2024). Shifting the protest paradigm? legitimizing and humanizing protest coverage lead to more positive attitudes toward protest, mixed results on news credibility. *Journalism*, *25*(6), 1230–1251. https://doi.org/10.1177/14648849231200135

Matich, P., Thomson, T. J., & Thomas, R. J. (2025). Old threats, new name? generative AI and visual journalism. *Journalism Practice*, *19*(10), 2402–2421. https://doi.org/10.1080/17512786.2025.2451677

McLeod, D. M., & Detenber, B. H. (1999). Framing effects of television news coverage of social protest. *Journal of Communication*, *49*(3), 3–23. https://doi.org/10.1111/j.1460-2466.1999.tb02802.x

Messaris, P., & Abraham, L. K. (2001). The role of images in framing news stories. https://api.semanticscholar.org/CorpusID:158869911

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

Olofsson, J. K., Nordin, S., Sequeira, H., & Polich, J. (2008). Affective picture processing: An integrative review of ERP findings. *Biological Psychology*, *77*(3), 247–265. https://doi.org/10.1016/j.biopsycho.2007.11.006

Pastorino, V., Sivakumar, J. A., & Moosavi, N. S. (2024). Decoding news narratives: A critical analysis of large language models in framing detection [Version Number: 3]. https://doi.org/10.48550/ARXIV.2402.11621

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

Rodriguez, L., & Dimitrova, D. V. (2011). The levels of visual framing. *Journal of Visual Literacy*, *30*(1), 48–65. https://doi.org/10.1080/23796529.2011.11674684

Sharma, M., & Peng, Y. (2024). How visual aesthetics and calorie density predict food image popularity on instagram: A computer vision analysis. *Health Communication*, *39*(3), 577–591. https://doi.org/10.1080/10410236.2023.2175635

Sontag, S. (2010). *On photography* (14. [Nachdr.]). Picador.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Taigman, Y., Yang, M., Ranzato, M., & Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Y., Kim, S. J., Shan, Y., Sun, Y., Jiang, X., Lee, H., Borah, P., Wagner, M., & Shah, D. (2024). Slant, extremity, and diversity: How the shape of news use explains electoral judgments and confidence. *Public Opinion Quarterly*, *88*, 708–734. https://doi.org/10.1093/poq/nfae031

Yadav, A., & Vishwakarma, D. K. (2023). A deep multi-level attentive network for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *19*(1), 1–19. https://doi.org/10.1145/3517139