# Evaluating Large Vision-Language Models for Visual Framing Analysis in News Imagery: A Theory-Driven Benchmark


Linqi Lu
University of North Dakota
linqi.lu@und.edu

Zihan Wan
Carnegie Mellon University
zwan2@alumni.cmu.edu

Hyerin Kwon
University of Wisconsin-Madison
hkwon53@wisc.edu

Sang Jung Kim
University of Iowa
sangjung-kim@uiowa.edu

Jiwon Kang
University of Wisconsin-Madison
jkang76@wisc.edu

Laila Abbas
University of Wisconsin-Madison
labbas@wisc.edu

Jiawei Liu
University of Florida
jiaweiliu@ufl.edu

Douglas M. McLeod
University of Wisconsin-Madison
dmmcleod@wisc.edu



## Abstract

*This study evaluates and compares the effectiveness of multiple large vision-language models (LVLMs) for automated visual framing analysis in the context of news imagery about social movements. Specifically, we evaluate LVLMs (Gemma3-27B, GPT-4.1, InternVL3-14B, InternVL3-38B, and Qwen2.5-VL-72B) against human-annotated ground truth data, using both baseline prompts and a range of Chain-of-Thought (CoT) prompting strategies with increasing complexity (i.e., from simple to detailed to expert). Model performance is assessed across visual framing categories: conflict, peace, and solidarity, using standard evaluation metrics including F1-score and Cohen's kappa. Our findings show that (1) CoT prompting improves model alignment with human annotations across most framing categories, especially for complex social cues like solidarity; (2) expert-level CoT prompts show the highest agreement with human coders; and (3) model performance varies by the specific model in focus, with InternVL3-38B consistently outperforming others. This study provides a scalable and theory-driven framework for applying LVLMs to visual content analysis in social science research.*



**Keywords:** large vision-language models, prompt engineering, Chain-of-Thought, visual framing, news imagery


## 1. Introduction

Historically**,** message framing has been one of the most prolific areas of communication research (Chung et al., 2013). At the core of message framing is an approach to study the meaning embedded in communication messages. Framing research is based on four fundamental assumptions: (a) For any given subject matter, there are virtually infinite ways to construct messages; (b) In the process of creating a message, the author makes choices about the building blocks that are used to construct the message as well as about the ultimate form that the message takes; (c) These message construction decisions convey preferred meanings that shape a message receivers' understanding of the events and issues at hand; and (d) The impact of that message is not uniform, but is moderated by a variety of factors including the receiver's predispositions. These assumptions have guided framing and framing effects researchers from a variety of different social science disciplines, including communication, psychology, sociology, among many others (Entman, 1993).

While considerable attention has been paid to developing the theoretical underpinning and methodological approaches to textual message framing (McLeod et al., 2022), less attention has been devoted to assessing the meaning embedded within visual messages. In the spirit of the adage, "A picture is worth a thousand words," it is imperative to develop theoretical and methodological approaches to visual framing as the images (such as news photographs) that accompany textual messages can accentuate or alter the impact of the textual messages (such as news stories). Moreover, readers may attend to visual images without even reading the accompanying textual message, further underscoring the importance of visual frames. The visuals that accompany a news story may set the tone for how the message is perceived. They may also provide substantiating evidence to support assertions made in the text. For instance, a photograph showing a clash between protesters and police may substantiate a news story's assertion that a protest was violent. In essence, visuals can serve as "framing devices," helping audiences make sense of, interpret, and form attitudes toward complex events and issues, such as social movements (Geise & Baden, 2015).

Along this line, the fact that images circulate rapidly and proliferate on platforms, like news websites and social media, emphasizes the need to develop scalable visual framing analysis methods that contribute to advancing both theory and computational methodology in mass communication research. However, analyzing visual frames faces challenges. Compared to texts, images are more ambiguous and more dependent upon the specific context for interpretation. Textual content consists of words, sentences, paragraphs, with relatively stable meanings that can be more systematically coded. In contrast, images contain visual information at different levels, from objects, people, colors, to spatial relationships that may be difficult to classify if only focusing on a single level/dimension without considering others. Thus, interpreting visual frames involves, not only recognizing objects or scenes, but also social relationships, shared symbols, and embedded meanings. This complexity inevitably requires interpretive and analysis tools that extend beyond object detection or scene classification to infer symbolic meaning.

While natural language models (NLMs) have largely revolutionized the analysis of textual frames through in-context learning (i.e., learning based on examples) and fine-tuning (i.e., adjusting the prompts for improved accuracy), visual frame analysis lacks a clear path forward. Large vision-language models (LVLMs) represent a promising opportunity to advance visual frame analysis through their capacity to bridge image understanding with semantic inference. However, the application of LVLMs in social science research is still in its initial stage as integrating large-scale image datasets into communication analysis requires robust theoretical frameworks, scalable annotation methods, and reliable performance benchmarks.

To advance the methodological toolkit for visual framing analysis, this study proposes a systematic and comprehensive framework to evaluate the performance of LVLMs and prompt engineering strategies. Specifically, we assess how effectively these models can identify different visual frames and their underlying components, such as actors, actions, objects, and relational dynamics. Using a human-coded benchmark dataset as ground truth, we evaluate the degree of alignment between model-generated outputs and human interpretation. Our empirical findings provide insights on the extent to which current LVLMs can be applied to automated visual content analysis and illuminate best practices for integrating these tools into future research on visual framing in mass communication.

## 2. Research Background

### 2.1. Visual framing analysis of social movements

In the context of social movements, media framing plays a critical role as it has been demonstrated to influence people's perceptions about the legitimacy of the movements (Boyle & McLeod, 2018; McLeod & Hertog, 1992), attitudes toward the underlying political and social issues, and audience engagement (Casas & Williams, 2019; Lu & Peng, 2024). Past literature showed that news articles tend to support of the status quo, delegitimizing protests and marginalizing protesters (Boyle & McLeod, 2018; McLeod & Hertog, 1992): news coverage of social protests were more likely to feature conflicts between protesters and police, and as a result, the emphasis on conflicts might trigger negative perceptions about the protesters and disapproval of their issue positions among the news audience.

What is often overlooked in the analysis of news coverage is that social movements may also strengthen solidarity within social groups through shared identity and goals (Coser, 1956; Sangiovanni & Viehoff, 2023). Moreover, tensions between groups may also reinforce solidarity within each group (Coser, 1956). Also, while conflicts are often highlighted, most of the social movements are peaceful (Mansoor, 2020). Consequently, highlighting solidarity and peace in social movements may better advance the causes advocated by the movements. Taken together, despite that past literature is highly conflict focused, visual framing of social movements from news media may either depict: (a) conflicts between protesters and police, (b) solidarity among protesters, (c) solidarity among police, or (d) highlighting peace (Lu et al., 2025). Analytical strategies for large-scale pattern detection are needed for investigating the news images of social movements (Joo & Steinert-Threlkeld, 2022; Neumayer & Rossi, 2018).

Compared to text, visuals of social movements might be more likely to be recalled due to their capacity to trigger affective responses and encode vivid mental representations, which enhances the sense of proximity to events/issues (Fahmy & Johnson, 2007). Scholars have more recently dedicated their attention to the analysis of visual elements (Rodriguez & Dimitrova, 2011) due to the growing dominance of audiovisual centered communication in the digital era. Geise (2017) defines visual framing as "the process of selecting some aspects of a perceived reality, highlighting them above others by means of visual communication … so that certain attributions, interpretations, or evaluations of the issue or item described are visually promoted" (p. 1). Scholars regard visual framing as an ongoing process, which includes the production and selection of visuals,

visual design and news values (e.g., Kress & Van Leeuwen, 2020), the presentation of news images (e.g., Fahmy, 2010; Grabe & Bucy, 2009), and audience reception: how viewers receive, interpret, and are impacted by the visuals (e.g., Iyer et al., 2014).

Advancements in analytical strategies can facilitate comprehensive visual framing analysis, enabling scalable discovery of visual patterns across large datasets (Joo & Steinert-Threlkeld, 2022) while preserving contextual nuance and symbolic depth. To systematically examine visuals, Rodriguez and Dimitrova's (2011) four-tiered visual framing model (denotative, semiotic, connotative, and ideological) provides a foundational framework that allows for a layered approach to the decoding of visual meaning across multiple levels. The *denotative* level focuses on identifying the basic representational content depicted, including the setting, objects as well as the actors; the *semiotic* level explores how an image is composed, including the camera angles as well as the actors' body posture, facial expressions (Forgas & East, 2008); the *connotative* level reveals the metaphorical message of a visual and the overarching meaning it reflects; the *ideological* level reflects the communicator's beliefs and motives behind the visual, including their sociopolitical and religious worldviews (Feng, 2013). Through capturing denotative and semiotic elements, the analysis can then be used to infer the connotative and ideological levels of visual framing. This model has been widely used to analyze visual content in diverse contexts, including social movements (Fahmy, 2010).

## 2.2. Computer vision techniques for image analysis

Computer vision (CV) techniques provide the foundational tools necessary for analyzing visual content. Image preprocessing techniques such as scanning, sampling, and quantization prepare raw image data for feasible analysis by standardizing pixel structures and reducing noise (Sharma et al., 2010). Feature extraction methods like edge detection, texture analysis, color histogram, are then used to identify distinctive patterns, shapes, and stylistic features that can signal visual salience (Lowe, 2004; Dalal & Triggs, 2005). Object detection and segmentation algorithms, including modern deep learning architectures enable the precise localization and categorization of key entities within images (Redmon et al., 2016; Ren et al., 2015).

Building on these foundations, deep learning architectures, particularly convolutional neural networks (CNNs), have become the backbone of semantic image understanding, enabling more robust and scalable object and scene classification (Joo & Steinert-Threlkeld, 2018). One widely used application of CNNs is facial expression and emotion recognition. For example, Joo et al. (2019) developed a multi-task CNN model to automatically detect facial displays of anger, threat, and happiness, as well as other visual cues like defiance and affiliative gestures in presidential debates.

## 2.3. Computational visual framing analysis with vision language models

Large vision-language models (LVLMs) have opened new avenues for scholars to explore the symbolic and ideological dimensions of images through prompting: querying visual content using natural language (Nayak et al., 2024). This is possible because LVLMs are pre-trained on billions of image-text pairs, which enables them to associate visual elements (e.g., raised fists in social movement images), not only with their literal form, but also with contextual meanings like solidarity, as these often co-occur in captions or surrounding text (Zhang et al., 2024). Many modern LVLMs integrate transformer-based visual encoders (Radford et al., 2021) with large language models to enable multimodal reasoning (Zhou et al., 2024).

Although recent advancements in various LVLMs enable scholars to examine higher-order visual frames (such as connotative and ideological frames) through prompting, it is crucial to compare different LVLMs to determine whether these models can accomplish the visual understanding tasks that human researchers perform. This is especially important because different LVLMs have different underlying mechanisms for processing image-text pairs; while CLIP uses a dual-encoder architecture trained with contrastive learning to align image and text embeddings in a shared space, GPT-4.1 integrates visual inputs directly into a unified transformer-based language model, allowing for more contextualized and generative reasoning (Achiam et al., 2023; Radford et al., 2021).

Pre-trained LVLMs can also differ significantly based on their data sources, regional development contexts, and design goals. For instance, Qwen2.5-VL-72B is primarily trained on data curated in Chinese contexts, potentially reflecting the cultural norms specific to its region of origin (Bai et al., 2025). Similarly, Google's Gemma3-27B, though multilingual and multimodal, reflects training priorities aligned with Western-centric datasets. These underlying differences can lead to cultural biases in how models interpret symbolic imagery (Ananthram et al., 2025). For instance, while one model might interpret a raised fist as a sign of solidarity, another might associate it primarily with aggression, depending on its training corpus. As such, comparing how different LVLMs interpret the same visual input is essential for understanding their

symbolic framing capacities and potential sociopolitical blind spots.

Another important feature of LVLMs that warrants careful examination is their tendency to adapt visual classification outputs in response to changes in researchers' prompting strategies. While existing research has explored how various prompting techniques such as zero-shot prompting (i.e., providing instructions without examples) and chain-of-thought prompting (i.e., guiding the model through logical reasoning steps) enhance the performance of large language models (Kojima et al., 2023), there remains a notable lack of studies examining how these strategies influence LVLM outputs. Thus, it is essential to identify which prompting strategies yield the most reliable and interpretable results from LVLMs.

Finally, manual coding serves as a validation in the process of automated visual analysis (Araujo et al., 2020). Peng and Lu (2023) emphasize the importance of incorporating human validation into automated visual analysis, especially when readily accessible computer vision tools are used for their convenience, sometimes at the expense of accuracy. For instance, detecting emotions in facial expressions still requires human validation. Building on this, the current study uses LVLMs to identify visual frames of social movements and compares these classifications to human-coded annotations to assess the extent to which model-generated frame classifications align with those coded by human annotators.

Taken together, this paper aims to (a) assess the capabilities of various LVLMs for scalable visual framing research, (b) compare prompting approaches to identify the most effective techniques for maximizing model performance, and (c) evaluate LVLMs to recommend the most suitable model for interpreting connotative and ideological frames in social movement imagery. We propose the following research questions.

> **RQ1**: How can state-of-the-art LVLMs be used as effective tools for visual framing research?
> **RQ2**: How do specific prompt designs (e.g., baseline Non-CoT, simple CoT, detailed CoT, expert CoT) work in analyzing visual frames?
> **RQ3**: How closely do model-detected frames align with ground truth from manual coding, and if the alignment differs between frame types?

# 3. Methodology

## 3.1. Data collection & preprocessing

This study collected a corpus of news articles and associated images related to four large-scale social movements in recent years: the Black Lives Matter (BLM) movement, far-right mobilizations, anti-war movements, and the Antifa (anti-fascist) movement. Articles were retrieved via the LexisNexis News API, which provides access to digital and print news content across major U.S. news organizations. Thirteen news outlets were selected to reflect a range of political orientations, including *Daily Kos, MSNBC, Rolling Stone, CNN, The New York Times, Politico, The Hill, CBS News, ABC News, The Wall Street Journal, Fox News, The Washington Times,* and *The Daily Caller*. These outlets span the ideological spectrum from progressive to conservative, ensuring variation in both visual and textual framing approaches. The LexisNexis API was used to retrieve articles containing keywords related to social movements published between 01/01/2018 and 03/01/2025 with filters applied to remove duplicate entries, inaccessible articles, and irrelevant articles ($N = 48{,}135$ after filtering). Following the article collection, GNews API and Scrapingdog API were utilized for image retrieval. This scraping process yielded a corpus of 8,979 high-resolution images, each linked to the corresponding article metadata (e.g., source, headline, publication date, and news content).

## 3.2. Human annotation

To establish ground truth benchmarks for model validation, a two-stage human annotation procedure was implemented to evaluate (a) news relevance and (b) visual content for framing analysis. Two human-annotated datasets were subsequently used for model comparison: the first for assessing headline-image relevance against the LLaMA-3-8B model, and the second for evaluating visual framing predictions from a set of LVLMs: Gemma3-27B, GPT-4.1, InternVL3-14B, InternVL3-38B, and Qwen2.5-VL-72B.

In the first stage, three trained annotators independently coded a random subsample of 200 images to assess whether the associated headline and article content were relevant to social movements. Relevance was coded dichotomously as "1" (relevant) or "0" (irrelevant), yielding high inter-coder agreement (Krippendorff's $\alpha = .91$).

In the second stage, the annotators conducted a detailed visual content analysis on a separate set of 200 randomly selected social movement images. This round focused on identifying four key semantic framing categories: conflict, peace, protester solidarity, and police solidarity, using a structured codebook informed by past literature (Boyle & McLeod, 2018; McLeod & Hertog, 1992; Lu et al., 2025). Human coding of the four frames served as the gold standard evaluating the performance of LVLMs, achieving a mean Krippendorff's alpha of .86.

To ensure clarity and replicability, annotators received extensive training using a shared codebook that included operational definitions, decision rules, and multiple annotated examples for each frame category. Pilot rounds were conducted prior to formal coding to calibrate interpretations, and disagreements were resolved through discussion until consensus was reached.

We further identified 79 specific visual elements that mapped onto the four key frames according to a theoretical framework encompassing actors, actions, objects, environment, and relationships. In addition to the presence or absence of these elements, we also coded the degree of conflict and solidarity in the visuals to capture variation in intensity. This framework builds on Rodriguez and Dimitrova's (2011) four-tiered model of visual framing extended through the social semiotic framework (Jewitt & Oyama, 2004; Kress & van Leeuwen, 2020) and computational scene understanding models (Krishna et al., 2017). The comprehensive codebook, annotated examples, and full prompt designs are available via Appendix (https://osf.io/vnuep/?view_only=191ee836cf974b5ebaf96c11c261d61e).

## 3.3. LVLMs annotation

To evaluate the performance of different LVLMs in classifying the four framing categories, the standardized multimodal prompt first requires LVLMs to classify the unique visual elements, mostly at the denotative and semiotic levels, to infer connotative and ideological levels of visual framing. Then, the prompt provides definitions of four semantic framing categories (conflict, peace, protester solidarity, and police solidarity) and asks how, based on the LVLM's classification of the 79 visual elements and the provided definitions, the model would determine the appropriate frame(s). Specifically, the prompt instructs each model to analyze an image and return a JSON-formatted output indicating whether each framing category is present or not (true/false), along with the supporting visual elements. For instance, the prompt asks whether the image depicts "conflict" based on indicators such as aggressive gestures, riot gear, or confrontations between protesters and police.

This schema not only aligns with principles from multimodal discourse and social semiotics where images are read as texts with layered meaning systems but also echoes object-relation models in computer vision, particularly scene graph approaches that represent images as structured triplets of subjects, predicates, and objects (Krishna et al., 2017). By bridging humanistic and computational traditions, the framework enables both qualitative human annotation and automated large-scale analysis of visual frames, offering a theoretically grounded and operationally robust tool for multimodal framing research.

Table 1. Conceptual Visual Analysis Framework

| Visual Category | Social Semiotic Function | Example |
|---|---|---|
| Actor | Representational participants (e.g., roles or identities) | Protester, police officer, bystander, etc |
| Action | Process types (e.g., material, verbal, mental) | Marching, shouting, kneeling, etc |
| Object | Circumstantial elements; symbolic cues | Signs, shields, flags, weapons, etc |
| Environment | Locational and compositional meaning | Indoor, outdoor, etc |
| Relationship | Interactive meaning (e.g., gaze, proximity, power dynamics) | Protester-police standoff, group cohesion, etc |

## 3.4. Benchmark experiment

To evaluate the performance of leading LVLMs on visual framing detection tasks, we conducted a benchmark experiment using their outputs generated from a standardized multimodal prompt. The selected models for comparison were Gemma3-27B, GPT-4.1, InternVL3-14B, InternVL3-38B, and Qwen2.5-VL-72B. Model selection was guided by practical accessibility, implementation feasibility, and performance rankings on OpenCompass benchmarks (OpenCompass, 2024). InternVL3-38B, for instance, ranked second among open-source models and third overall in combined rankings during the study period. These models represent a diverse cross-section of contemporary LVLMs varying in architecture, parameter size, and visual reasoning capabilities.

To quantify model performance, we compared each model's classifications of the four semantic framing categories to a gold-standard dataset of images manually annotated by expert human coders trained in visual communication research. Evaluation was performed using four widely adopted classification metrics: precision, recall, F1-score (Zhang et al., 2019), and Cohen's kappa (κ; Ananda et al., 2021; Cohen, 1960). InternVL3-38B achieved the highest performance on the conflict frame (F1=.92, κ=.87), followed by GPT-4.1 (F1=.88, κ=.81). For the peace frame, InternVL3-38B also led with strong performance (F1=.84, κ=.70), with Qwen2.5-VL-72B (F1=.83, κ=.67) and GPT-4.1 (F1=.79, κ=.64) performing competitively. For the police solidarity frame, Qwen2.5-VL-72B (F1= .88, κ=.84) performed the best, while protester solidarity framing was most accurately identified by InternVL3-38B (F1=.87, κ=.78). Overall, performance varied across models and frames, but

InternVL3-38B consistently demonstrated strong performance across multiple framing types (see Appendix).

### 3.5. Chain-of-Thought prompting for visual framing

This study employed a structured prompt engineering approach to enhance the performance of LVLMs in detecting visual frames. After establishing a baseline using the standardized multimodal prompt for classification (see section 3.3), we explored whether prompting models to reason step-by-step, known as Chain-of-Thought (CoT) prompting, would improve their capacity to interpret social movement images across four framing categories.

Building on prior work demonstrating the utility of CoT in natural language tasks (Wei et al., 2022; Wang et al., 2022), we developed three progressively elaborated CoT prompts. Each version was designed to support model reasoning through multiple interpretive stages based on framing theory and visual semiotics. The aim was to reduce ambiguity, encourage contextual reasoning, and yield more reliable predictions aligned with human-coded ground truth.

In simple CoT condition, models were prompted to make basic observations, highlight salient features, and draw a conclusion. The detailed version extended this with guided prompts about emotional indicators, symbolic elements, and spatial relationships, components theorized to underlie moderate to strong framing effects (Kress & van Leeuwen, 2020). The expert-level CoT prompt mimicked a professional analyst's workflow by asking the model to assess spatial hierarchies, power dynamics, communicative symbolism, and group cohesion before assigning confidence ratings to its final output.

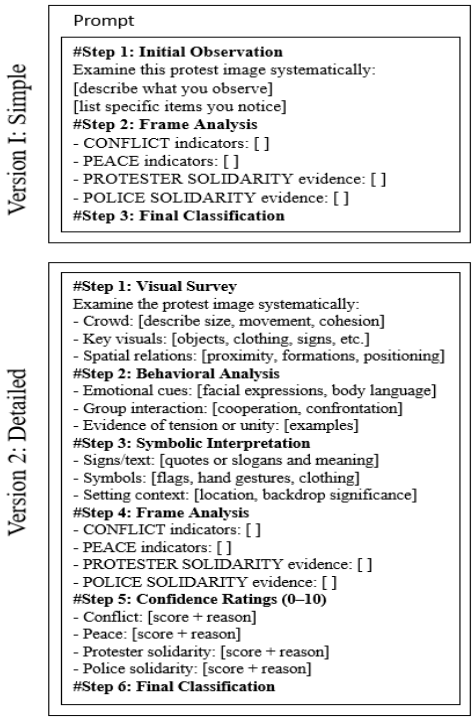


**Figure 1. Example of CoT prompting (Simple & Detailed)**

Each CoT variant was delivered as a single API request, embedding background knowledge, stepwise instructions, and structured output expectations. Below is an excerpt of the expert-level prompt used in this study:

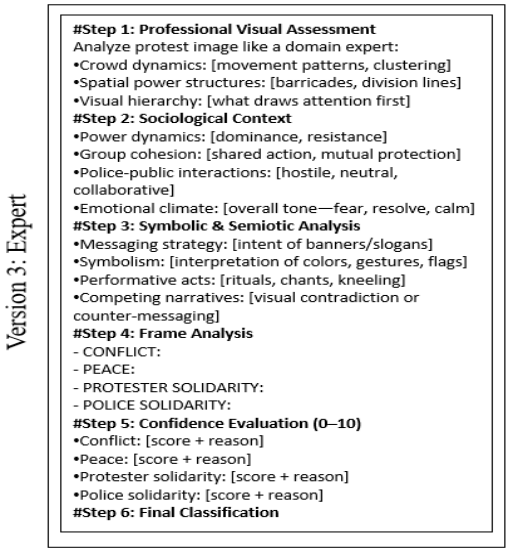


**Figure 2. Example of CoT prompting (Expert)**

All open-source models were deployed on a dedicated server equipped with two NVIDIA RTX 6000 Ada Generation GPUs (48 GB each), supporting efficient batched inference and multimodal input

processing at scale. Bootstrap resampling (a robust method widely endorsed for uncertainty estimation with limited sample sizes) was employed to ensure statistical rigor (Austin & Tu, 2004; Carpenter & Bithell, 2000). Predictions from three independent runs for each model-prompt combination were consolidated using majority voting, yielding a single representative prediction set per condition (Kuncheva, 2004). This aggregation approach accounts for run-to-run variability and has demonstrated improved estimator stability in binary classification tasks (Zhou et al., 2002). Bootstrap resampling with 2,000 iterations was subsequently applied to these consolidated prediction sets to derive empirical confidence intervals for F1-score.

Results demonstrated that CoT prompting improved model performance compared to the baseline condition. For the conflict frame, nearly all models benefited from CoT prompting. InternVL3-38B demonstrated the highest performance overall with F1 of 0.93 [0.89, 0.97]. Qwen2.5-VL-72B and GPT 4.1 also achieved consistent high F1 of 0.89 [0.83, 0.94] across CoT variants. The peace frame was the most stable across conditions, with relatively less sensitivity to prompt variation. Qwen2.5-VL-72B achieved the highest F1=0.85 [0.79, 0.90] with CoT prompting. For the protester solidarity frame, all models experienced modest gains. GPT-4.1 showed the most substantial improvement, increasing from a baseline F1=0.63 [0.53, 0.72] to 0.91 [0.86, 0.95] under the simple CoT prompt, with performance stabilizing around 0.88 in other CoT prompting conditions. InternVL3-14B with the highest F1=0.89 [0.83, 0.93] under Expert CoT prompting showed improvement, while InternVL3-38B found no performance increase. Police solidarity framing exhibited the most pronounced performance improvements under CoT prompting. InternVL3-14B advanced from baseline F1=0.48 [0.33, 0.63] to F1=0.92 [0.86, 0.97] under expert prompting, while InternVL3-38B improved from F1=0.55 [0.39, 0.68] to F1=0.93 [0.86, 0.97] under detailed CoT prompting. GPT-4.1 showed substantial enhancement from F1= 0.28 [0.12, 0.44] at baseline to F1=0.65 [0.52, 0.78] in the simple CoT condition. Qwen2.5-VL-72B maintained high performance across all conditions (from F1=0.83 [0.74, 0.91] to F1=0.90 [0.82, 0.96]).

Overall, InternVL3-38B was selected for subsequent analysis (see Section 3.7) due to its consistently high performance across all framing types, with confidence intervals indicating reliable measurement precision across experimental conditions.

**Table 2. Bootstrap F1 Scores with 95% CIs for News Framing Classification by Model and Prompt**

| | | Non-CoT | CoT | | |
|---|---|---|---|---|---|
| **Framing** | **Model** | **Baseline** | **Simple** | **Detailed** | **Expert** |
| Conflict | Gemma3-27B | 0.79 [0.71, 0.86] | 0.85 [0.78, 0.91] | 0.84 [0.78, 0.90] | 0.82 [0.75, 0.88] |
| | GPT-4.1 | 0.89 [0.83, 0.94] | 0.86 [0.80, 0.92] | 0.88 [0.82, 0.94] | 0.89 [0.83, 0.94] |
| | InternVL3-14B | 0.66 [0.55, 0.76] | 0.82 [0.74, 0.89] | 0.83 [0.75, 0.89] | 0.81 [0.73, 0.88] |
| | **InternVL3-38B** | **0.91 [0.85, 0.95]** | **0.92 [0.87, 0.96]** | **0.93 [0.88, 0.97]** | **0.93 [0.89, 0.97]** |
| | Qwen2.5-VL-72B | 0.83 [0.75, 0.90] | 0.88 [0.81, 0.93] | 0.89 [0.83, 0.94] | 0.89 [0.83, 0.94] |
| Peace | Gemma3-27B | 0.71 [0.62, 0.78] | 0.81 [0.74, 0.86] | 0.80 [0.74, 0.86] | 0.84 [0.77, 0.89] |
| | GPT-4.1 | 0.75 [0.68, 0.82] | 0.83 [0.76, 0.88] | 0.80 [0.73, 0.86] | 0.81 [0.74, 0.86] |
| | InternVL3-14B | 0.78 [0.70, 0.84] | 0.81 [0.74, 0.87] | 0.81 [0.74, 0.87] | 0.80 [0.73, 0.86] |
| | **InternVL3-38B** | **0.83 [0.77, 0.88]** | 0.82 [0.76, 0.88] | 0.80 [0.73, 0.86] | 0.82 [0.76, 0.88] |
| | **Qwen2.5-VL-72B** | 0.82 [0.76, 0.88] | **0.85 [0.79, 0.90]** | **0.84 [0.78, 0.89]** | **0.85 [0.79, 0.90]** |
| Protester solidarity | Gemma3-27B | 0.80 [0.73, 0.86] | 0.80 [0.73, 0.86] | 0.79 [0.72, 0.85] | 0.82 [0.76, 0.88] |
| | **GPT-4.1** | 0.63 [0.53, 0.72] | **0.91 [0.86, 0.95]** | **0.88 [0.83, 0.93]** | 0.88 [0.83, 0.93] |
| | **InternVL3-14B** | 0.84 [0.77, 0.90] | 0.89 [0.83, 0.94] | 0.87 [0.81, 0.92] | **0.89 [0.83, 0.93]** |
| | **InternVL3-38B** | **0.87 [0.82, 0.92]** | 0.84 [0.78, 0.89] | 0.84 [0.78, 0.90] | 0.84 [0.78, 0.90] |
| | Qwen2.5-VL-72B | 0.83 [0.78, 0.89] | 0.86 [0.80, 0.91] | 0.86 [0.80, 0.91] | 0.83 [0.77, 0.88] |
| Police solidarity | Gemma3-27B | 0.80 [0.71, 0.87] | 0.82 [0.74, 0.89] | 0.80 [0.72, 0.88] | 0.79 [0.70, 0.87] |
| | GPT-4.1 | 0.28 [0.12, 0.44] | 0.65 [0.52, 0.78] | 0.47 [0.31, 0.62] | 0.64 [0.49, 0.76] |
| | **InternVL3-14B** | 0.48 [0.33, 0.63] | 0.86 [0.77, 0.93] | 0.85 [0.76, 0.92] | **0.92 [0.86, 0.97]** |
| | **InternVL3-38B** | 0.55 [0.39, 0.68] | **0.88 [0.80, 0.95]** | **0.93 [0.86, 0.97]** | 0.89 [0.82, 0.95] |
| | **Qwen2.5-VL-72B** | **0.90 [0.82, 0.96]** | 0.83 [0.74, 0.91] | 0.90 [0.82, 0.96] | 0.90 [0.83, 0.96] |

## 3.6. Reliability validation

To assess the robustness of model outputs beyond standard classification metrics, we evaluated LVLMs' reliability using Krippendorff's alpha (α) (Lee et al., 2024). For each of the five models (Gemma3-27B, GPT-4.1, InternVL3-14B, InternVL3-38B, Qwen2.5-VL-72B) and each of the four prompt conditions (Baseline, Simple, Detailed, Expert), we conducted three independent inference runs using identical inputs but separate randomized seeds. All five LVLMs across prompts demonstrated high reliability, with median $\alpha > .80$ (Landis & Koch, 1977). All α values were above .60, the minimum acceptable reliability threshold, underscoring the stability of frame predictions across prompt designs (see Figure 3).

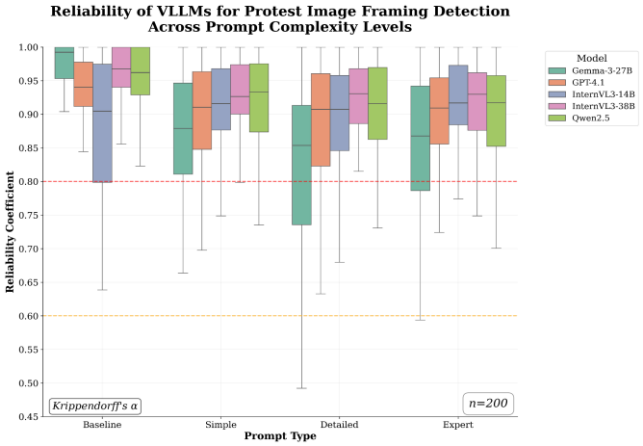


**Figure 3. Reliability performance across multiple tests**

## 3.7. Visual indicators of framing

We analyzed 8,979 social movement images in terms of their capacity to predict four types of framing (i.e., conflict, peace, protester solidarity, and police solidarity) using 77 binary predictors regarding actors, objects, actions, environments, and relationships. Both

logistic regression (LR) and random forest (RF) classifiers were trained with stratified bootstrap resampling (N = 1,000) (see Appendix).

Conflict framing was most reliably predicted, with both LR and RF converging on emotion intensity and violence related indicators. LR showed that tense protester emotions (OR = 39.61 [9.77, 136.98]), visible smoke or fire (OR = 37.39 [9.56, 135.35]), and visible property damage (OR = 36.95 [6.02, 147.25]) were the strongest predictors. RF similarly ranked tense protester emotions (Imp = 0.13 [0.09, 0.16]), tense police emotions (Imp = 0.09 [0.07, 0.13]), and visible property damage (Imp = 0.07 [0.05, 0.09]) as among the most important predictors. By comparison, peace framing was captured through indicators of calmness and memorialization. LR revealed high odds ratios for peaceful gatherings (OR = 375.59 [254.29, 560.10]), followed by calm protester emotions (OR = 17.86 [13.56, 24.01]) and memorial elements (OR = 13.36 [5.70, 29.01]). RF results echoed the pattern, with peaceful gatherings (Imp = 0.27 [0.22, 0.32]) and calm emotions (Imp = 0.21 [0.17, 0.25]) emerging as dominant predictors.

Protester solidarity framing was characterized by gestures of unity and determination. LR showed strong predictive power from comforting or hugging (OR = 39.62 [18.35, 83.40]), organized crowds (OR = 15.45 [10.24, 23.53), and determined emotions (OR = 15.14 [10.24, 22.31]). RF results also highlighted organized crowds (Imp = 0.19 [0.14, 0.24]), peaceful gatherings (Imp = 0.17 [0.13, 0.21]), and determined emotions (Imp = 0.13 [0.09, 0.17]). Police solidarity was predicted by indicators of police presence and demeanor. LR identified calm police emotions (OR = 21.93 [8.38, 61.53]), riot gear (OR = 16.82 [5.82, 41.09]), and police presence (OR = 11.25 [4.18, 40.99]) as significant predictors. Similarly, RF emphasized police presence (Imp = 0.24 [0.18, 0.30]), determined emotions (Imp = 0.13 [0.09, 0.18]), and regular gear (Imp = 0.087 [0.06, 0.12]) as predictors.

## 4. Discussion and Conclusion

This study provides a systematic evaluation of state-of-the-art LVLMs for visual framing analysis in the context of social movement news imagery. In response to RQ1, our findings confirm that contemporary LVLMs, including Gemma3-27B, GPT-4.1, InternVL3-14B, InternVL3-38B, and Qwen2.5-VL-72B, exhibit substantial potential for automating large-scale content analysis, particularly when guided by theory-driven prompts. Across models, InternVL3-38B and Qwen2.5-VL-72B consistently achieved high baseline agreement with human coding, while GPT-4.1 benefited most dramatically from prompt enhancements, suggesting that model architecture interacts meaningfully with prompting strategies.

Addressing RQ2, we found that CoT prompting significantly improved model alignment with human annotations, especially for frames requiring interpretive nuance, such as solidarity. Expert-level CoT prompts, which guided models through multi-step reasoning, consistently outperformed baseline and simple prompts. These results underscore the value of prompt engineering for enhancing the interpretability and reliability of LVLMs, supporting prior work in language modeling and extending it into the visual domain.

To answer RQ3, we identified persistent challenges in frame detection accuracy. While conflict and peace frames showed relatively high alignment across models, solidarity frames were more difficult to detect, possibly due to their reliance on subtle cues and symbolic markers. Unlike conflict, which is marked by explicit cues of confrontation, solidarity is typically implicit, relying on subtle gestures of unity, shared symbols, or group alignments. For example, a raised fist may signify solidarity, but could also be read as aggression or celebration, depending on the context. These cues are often relational, polysemous, and culturally specific, making them harder for LVLMs to consistently identify than the more straightforward visual markers of conflict or peace. The largest differences in performance appeared in baseline conditions, emphasizing the need for structured guidance in complex visual interpretation tasks.

This study has limitations. First, model performance is highly dependent on structured prompts and may not generalize reliably to unguided inference settings. Second, while the coding framework improves replicability, it cannot fully capture connotative or metaphorical dimensions of visual meaning without further instructions.

In sum, this study offers critical insights into how LVLMs can be deployed in mass communication research. It contributes new evidence on model reliability, prompt efficacy, and the differences in model performance between visual frames, providing a foundation for future framing research. As visual content plays an increasingly important role in public discourse, tools like CoT-enhanced LVLMs can help scale visual framing analysis while preserving theoretical rigor and human interpretability.

## 5. References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). GPT-4 technical report. arXiv preprint arXiv:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Ananda, A., Ngan, K. H., Karabağ, C., Ter-Sarkisov, A., Alonso, E., & Reyes-Aldasoro, C. C. (2021). Classification and visualisation of normal and abnormal radiographs; a comparison between eleven convolutional neural network architectures. *Sensors, 21*(16), 5381. https://doi.org/10.3390/s21165381

Ananthram, A., Stengel-Eskin, E., Bansal, M., & McKeown, K. (2025). See it from my perspective: How language affects cultural bias in image understanding. arXiv preprint arXiv:2406.11665 https://doi.org/10.48550/arXiv.2406.11665

Araujo, T., Lock, I., & van de Velde, B. (2020). Automated visual content analysis (AVCA) in communication research: A protocol for large scale image classification with pre-trained computer vision models. *Communication Methods and Measures*, *14*(4), 239-265. https://doi.org/10.1080/19312458.2020.1810648

Austin, P. C., & Tu, J. V. (2004). Bootstrap Methods for Developing Predictive Models. *The American Statistician*, *58*(2), 131–137. http://www.jstor.org/stable/27643521

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., ... & Lin, J. (2025). Qwen2. 5-vl technical report. arXiv preprint arXiv:2502.13923.

Boyle, M. P., & McLeod, D. M. (2018). News framing and social protest: Toward a comprehensive model. In *Doing News Framing Analysis II* (pp. 295-319). Routledge.

Carpenter, J., & Bithell, J. (2000). Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Statistics in Medicine, 19*(9), 1141–1164.

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

Chung, C.J., Barnett, G. A., Kim, K., & Lackaff, D. (2013). An analysis of communication theory and discipline. *Scientometrics, 95*(3), 985-1002.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37-46.

Coser, L. (1956). *The functions of social conflict*. Free Press.

Dalal, N., & Triggs, B. (2005, June). Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)* (Vol. 1, pp. 886-893). IEEE. https://doi.org/10.1109/CVPR.2005.177

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51–58.

Fahmy, S. (2010). Contrasting visual frames of our times: A framing analysis of English-and Arabic-language press coverage of war and terrorism. *International Communication Gazette*, *72*(8), 695-717.

Fahmy, S., & Johnson, T. J. (2007). Show the truth and let the audience decide: A web-based survey showing support among viewers of Al-Jazeera for use of graphic imagery. *Journal of Broadcasting & Electronic Media*, *51*(2), 245-264. https://doi.org/10.1080/08838150701304688

Feng, D., & O'Halloran, K. L. (2013). The visual representation of metaphor: A social semiotic approach. *Review of Cognitive Linguistics, 11*(2), 320-335. https://doi.org/10.1075/rcl.11.2.07fen

Forgas, J. P., & East, R. (2008). On being happy and gullible: Mood effects on skepticism and the detection of deception. *Journal of Experimental Social Psychology*, *44*(5), 1362-1367. https://doi.org/10.1016/j.jesp.2008.04.010

Geise, S., & Baden, C. (2015). Putting the image back into the frame: Modeling the linkage between visual communication and frame-processing theory. *Communication Theory*, *25*(1), 46-69. https://doi.org/10.1111/comt.12048

Geise, S. (2017). Visual framing. In R. Holtz-Bacha, & C. Reinemann (Eds.), *Handbook of Political Communication* (pp. 1–14). De Gruyter.

Grabe, M. E., & Bucy, E. P. (2009). *Image bite politics: News and the visual framing of elections*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195372076.001.0001

Iyer, A., Webster, J., Hornsey, M. J., & Vanman, E. J. (2014). Understanding the power of the picture: The effect of image content on emotional and political responses to terrorism. *Journal of Applied Social Psychology*, *44*(7), 511–521.

Lee, N., Hong, J., & Thorne, J. (2024). Evaluating the Consistency of LLM Evaluators. arXiv preprint arXiv:2412.00543

Jewitt, C., & Oyama, R. (2004). Visual Meaning: A Social Semiotic Approach. In T. van Leeuwen & C. Jewitt (Eds.), *The Handbook of Visual Analysis* (pp. 134–156). Sage.

Joo, J., Bucy, E. P., & Seidel, C. (2019). Computational communication science| automated coding of televised leader displays: Detecting nonverbal political behavior with computer vision and deep learning. *International Journal of Communication*, *13*, 4044-4066.

Joo, J., & Steinert-Threlkeld, Z. C. (2018). Image as Data: Automated Visual Content Analysis for Political Science. arXiv preprint arXiv:1810.01544. https://doi.org/10.48550/arXiv.1810.01544

Joo, J., & Steinert-Threlkeld, Z. C. (2022). Image as data: Automated content analysis for visual presentations of political actors and events. *Computational Communication Research*, *4*(1).

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems, 35,* 22199-22213.

Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., ... & Fei-Fei, L. (2017). Visual

genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision, 123*(1), 32–73.

Kress, G., & Van Leeuwen, T. (2020). *Reading images: The grammar of visual design*. Routledge.

Kuncheva, L. I. (2004). Combining pattern classifiers: Methods and algorithms. John Wiley & Sons.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Lu, L., Tao, R., Kwon, H., Kang, J., Zhou, Y., Xin, H., ... & 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.

Mansoor (2020). 93% of Black Lives Matter protests have been peaceful, new report finds. *Time.* https://time.com/5886348/report-peaceful-protests/

McLeod, D. M., Choung, H., Su, M. H., Kim, S., Tao R., Liu, J., & Lee, B. (no). Navigating a diverse paradigm: A conceptual framework for experimental framing effects research. *Review of Communication Research, 10*, 1-58.

McLeod, D. M., & Hertog, J. K. (1992). The manufacture of public opinion by reporters: Informal cues for public perceptions of protest groups. *Discourse and Society, 3*(3), 259-275.

Nayak, S., Jain, K., Awal, R., Reddy, S., van Steenkiste, S., Hendricks, L. A., ... & Agrawal, A. (2024). *Benchmarking vision language models for cultural understanding.* arXiv preprint arXiv:2407.10920. https://doi.org/10.48550/arXiv.2407.10920

Neumayer, C., & Rossi, L. (2018). Images of protest in social media: Struggle over visibility and visual narratives. *New Media & Society*, *20*(11), 4293-4310.

OpenCompass. (2024). *Open VLM leaderboard*. Hugging Face Spaces. https://huggingface.co/spaces/opencompass/open_vlm_leaderboard

Peng, Y., & Lu, Y. (2023). Computational visual analysis in political communication. In *Research Handbook on Visual Politics* (pp. 42-54). Edward Elgar Publishing. https://doi.org/10.4337/9781800376939.00010

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). https://proceedings.mlr.press/v139/radford21a.html

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 779-788. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*, 91–99.

Rodriguez, L., & Dimitrova, D. V. (2011). The levels of visual framing. *Journal of visual literacy*, *30*(1), 48-65.

Sangiovanni, A., & Viehoff, J. (2023). *Solidarity in social and political philosophy.* Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/entries/solidarity/

Sharma, N., Aggarwal, L. M., & Kalra, P. K. (2010). Image preprocessing techniques for object detection in a video sequence: A review. *International Journal of Computer Applications, 3*(1), 28–34.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., & Zhou, D. (2022). Rationale-augmented ensembles in language models. arXiv preprint arXiv:2207.00747. https://doi.org/10.48550/arXiv.2207.00747

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824-24837.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2019). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM, 64*(3), 107–115.

Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *46*(8), 5625-5644.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. Artificial Intelligence, 137(1-2), 239-263.

Zhou, Q., Zhou, R., Hu, Z., Lu, P., Gao, S., & Zhang, Y. (2024). Image-of-thought prompting for visual reasoning refinement in multimodal large language models. arXiv preprint arXiv:2405.13872. https://doi.org/10.48550/arXiv.2405.13872