Visual Framing for News Stance Detection via Image Generation

Dahyun Lee
Soongsil University
hyundai@soongsil.ac.kr

Jiyoung Han
KAIST
jiyoung.han@kaist.ac.kr

Kunwoo Park
Soongsil University
kunwoo.park@ssu.ac.kr

Abstract

Article-level news stance detection aims to identify the perspective of news articles toward social issues. Despite advances in stance detection and its importance for trustworthy media environments, news articles pose distinct challenges because their stances are often implicit, subtly conveyed through journalistic framing, and embedded in long, structurally complex texts. To address these challenges, we introduce VFSTANCE, which leverages visual framing to make implicit stance cues more explicit via image generation. In evaluation experiments, we demonstrate the effectiveness of VFSTANCE over existing methods and the contribution of visual framing to its performance. Finally, a controlled user study ($N = 200$) in a snippet-based news consumption setting further demonstrates that VFSTANCE can make stance signals visually salient and highlights its potential use beyond automated stance detection.

1 Introduction

News articles can present different perspectives on the same issue through selection and emphasis of particular interpretive frames (Gentzkow and Shapiro, 2010). Automatically identifying these perspectives is important for analyzing media bias (Hamborg et al., 2019), understanding public opinion (Chong and Druckman, 2007), and supporting informed news consumption (Park et al., 2009). Stance detection, the task of identifying an author's expressed attitude toward a specific target from text, has been extensively studied in social media domains (Mohammad et al., 2016). More recently, the task has been extended to news articles, where the goal is to classify whether a full article's position toward a target issue is supportive, neutral, or oppositional (Mascarell et al., 2021).

Despite advances in NLP and large language models (LLMs), article-level news stance detection remains challenging for two key reasons. First, professional journalistic norms generally favor detached and fact-oriented reporting (Kovach and Rosenstiel, 2021); thus, a news article's stance toward a target is often implicit and subtly expressed. Second, news articles are long and structurally complex. Stance detection methods developed primarily for short texts therefore do not readily transfer to article-level news stance detection. Even human readers may find it difficult to infer stance from subtle framing cues distributed across lengthy, rhetorically elaborate news articles.



Figure 1: Key idea behind VFSTANCE: making implicit stance cues in a lengthy, rhetorically complex news article explicit through image generation grounded in visual framing.

These challenges motivate the use of visual information, which can convey attitudinal cues more immediately and intuitively than text (Mehrabian, 1981; Paivio, 1986). Visual representations may therefore provide a complementary means of making implicit and dispersed stance cues in news articles more accessible. Publisher-provided news images (i.e., the original images published alongside the articles), however, are not always available and, even when present, may not closely align with an article's stance, as press photographs are typically produced under documentary norms that may constrain overt evaluative signaling (Schwartz, 1992). This motivates our first research question: *Can image generation technology enable more effective news stance detection?*

Recent advances in text-to-image (T2I) generation and its growing adoption across diverse domains suggest a new possibility: generated images may serve as intermediate visual representations

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 57393–57418
October 24-29, 2026 ©2026 Association for Computational Linguistics

that make stance cues already present in news text more explicit to downstream models (Zhang et al., 2025b). Generating such representations, however, remains challenging because the cues most relevant to article-level news stance detection are themselves subtle, implicit, and distributed throughout the text. Naïvely generating an image from a news article may therefore fail to capture the framing choices most indicative of its stance. This leads to our second research question: *How can we generate images that make stance cues more explicit?*

To address these questions, we draw on framing theory from communication and media studies, particularly research on visual framing in news (Messaris and Abraham, 2001; Rodriguez and Dimitrova, 2011). We hypothesize that image generation grounded in visual framing can transform implicit stance cues in a news article into visually salient stance signals, thereby enabling more effective stance detection. VFSTANCE (**V**isual **F**raming for **Stance** Detection) is a multi-stage, modular framework designed to test this hypothesis (Figure 1). As illustrated in Figure 2, VFSTANCE first uses an LLM to derive visual framing specifications from the article, which are then used to prompt a T2I model to generate an image that makes the corresponding stance cues more explicit. Finally, an instruction-following large vision-language model (LVLM) predicts the article-level stance jointly from the article and the generated image.

We evaluate VFSTANCE on two article-level stance detection datasets in Korean and German and compare its performance with existing methods. The results show that VFSTANCE outperforms existing approaches across both datasets. Ablation experiments further demonstrate the contributions of visual framing and image generation to its performance. Finally, we conduct a controlled user study ($N = 200$) in a snippet-based news consumption setting to examine whether the generated images also help human readers identify article stance. Participants exposed to images generated by VFSTANCE identify article stance more accurately than those exposed to alternative image conditions. Taken together, these findings suggest that visual framing and image generation can make otherwise implicit stance cues more readily discernible to both computational models and human readers.

The contributions of this study are summarized as follows:

- We propose VFSTANCE, a multi-stage framework grounded in visual framing that transforms implicit stance cues in news articles into more explicit visual representations for article-level stance detection.
- We evaluate VFSTANCE on two article-level stance detection datasets and demonstrate its effectiveness across languages, while ablation experiments establish the contributions of both visual framing and image generation.
- We conduct a controlled user study ($N = 200$) in a snippet-based news consumption setting, showing that images generated by VFSTANCE help readers identify article stance more accurately and suggesting potential applications beyond automated stance detection.

2 Related Work

Article-Level News Stance Detection Much of the stance detection literature has examined short-form social media text, particularly tweets. Recent approaches have included fine-tuning pre-trained language models (Liang et al., 2022), leveraging LLM reasoning (Gatto et al., 2023; Zhang et al., 2024) and employing in-context learning (Zhu et al., 2023; Cruickshank and Ng, 2026). Multimodal stance detection methods that jointly utilize text and images have also emerged, including target-aware multimodal prompt tuning (Liang et al., 2024), cross-modal fusion (Weinzierl and Harabagiu, 2023), multimodal alignment (Zhang et al., 2025a), and image generation (Zhang et al., 2025b). Recent studies have further explored zero-shot detection with LVLMs (Weinzierl and Harabagiu, 2024; AlShenaifi and Alangari, 2026).

News stance detection remains relatively underexplored. Existing work has largely centered on headline- or sentence-level stance in contexts such as fake news detection or rumor verification (Ferreira and Vlachos, 2016; Pomerleau and Rao, 2017; Hanselowski et al., 2018; Conforti et al., 2020), while studies of article-level stance toward social issues remain scarce (Mascarell et al., 2021; Lee et al., 2025). Building on recent advances in multimodal stance detection (Liang et al., 2024; Zhang et al., 2025b), we address this gap by generating visual representations that make implicit stance signals in news articles more explicit and leveraging them for article-level stance detection.

Visual Framing Analysis Framing involves selecting and emphasizing certain aspects of real-

ity to promote particular interpretations (Entman, 1993), and visual framing concerns how such interpretive emphasis is conveyed through visual elements (Messaris and Abraham, 2001). Empirical studies in communication have operationalized frames through manual content analysis using predefined categories (Coleman, 2010). In the computational domain, framing analysis has emerged as an active research area, as documented by recent surveys (Ali and Hassan, 2022; Otmakhova et al., 2024; Vallejo et al., 2024). Earlier computational approaches primarily analyzed the textual modality using annotated corpora and supervised methods. Widely used resources include the Media Frames Corpus with fifteen generic policy frames (Card et al., 2015), the Gun Violence Frame Corpus (Liu et al., 2019a), and the multilingual SemEval-2023 benchmark covering nine languages (Piskorski et al., 2023). More recent research has explored multimodal approaches, such as fusion-based approaches using paired headlines and lead images (Tourni et al., 2021). Studies have also leveraged LLMs and LVLMs for frame analysis of news text and imagery (Arora et al., 2025; Lu et al., 2026; El Damanhoury et al., 2026). Our study extends this line of research by using visual framing to guide image generation for article-level stance detection.

3 Problem and Dataset

3.1 Target Problem

For a news article A that covers an issue T, the goal of article-level stance detection is to determine the stance of A toward T as *supportive*, *neutral*, or *oppositional* using a detection model $M(\cdot)$. We investigate whether specifying visual framing elements and generating images based on these specifications can lead to more effective stance detection.

3.2 Datasets

We use two article-level news stance detection datasets: one accompanied by news images and one without images. The first is K-News-Stance-MM, a novel multimodal extension of the existing article-level stance detection dataset (Lee et al., 2025), comprising 1,816 Korean news articles with accompanying news images. As the first dataset to provide article-level stance labels together with news images, it serves as the primary testbed for this study. The second is CheeSE (Mascarell et al., 2021), which contains 1,762 German news articles. We use this dataset to examine the applicability of our method for predicting article-level stance in a text-only setting. Dataset sizes and label distributions are provided in Table A1, and detailed dataset information is provided in Appendix B.

For broader accessibility and supplementary analysis, we also provide LLM-translated multilingual extensions of K-News-Stance-MM in English, Chinese, Indonesian, and Arabic. The target languages were selected to span different resource levels (Joshi et al., 2020).

4 Proposed Method

We introduce VFSTANCE (**V**isual **F**raming for **Stance** Detection), a multi-stage, modular framework for article-level stance detection. The central idea is to transform implicit textual stance cues into more explicit framing signals that can be rendered as visually salient elements by a T2I model, thereby enabling an instruction-following LVLM to detect article-level stance through multimodal reasoning.

News articles often express their stance toward a target issue implicitly, consistent with professional journalistic norms favoring detached and fact-oriented reporting (Kovach and Rosenstiel, 2021). Rather than expressing stance through explicit evaluative claims, articles may convey it through framing choices, such as selective emphasis on particular actors, causes, consequences, or interpretations of an issue (Entman, 1993). These cues are often subtle and distributed across long documents, making them difficult to detect from text alone.

Incorporating publisher images is a promising direction for improving stance detection performance (Appendix D.1). However, relying on conventional news images presents two limitations. First, not all news articles include an image. Second, even when images are available, they may provide limited stance-relevant information because publisher news images often serve documentary and contextual functions rather than explicitly foregrounding an article's interpretive stance (Reuters, 2008; Associated Press, 2024). Image generation using a T2I model offers a possible solution, but naïvely prompting the model with a full article or its summary may fail to capture stance-relevant framing cues effectively. The model must first identify which subtle and dispersed framing cues are most indicative of the article's stance before rendering them visually (see Table A10 for examples of

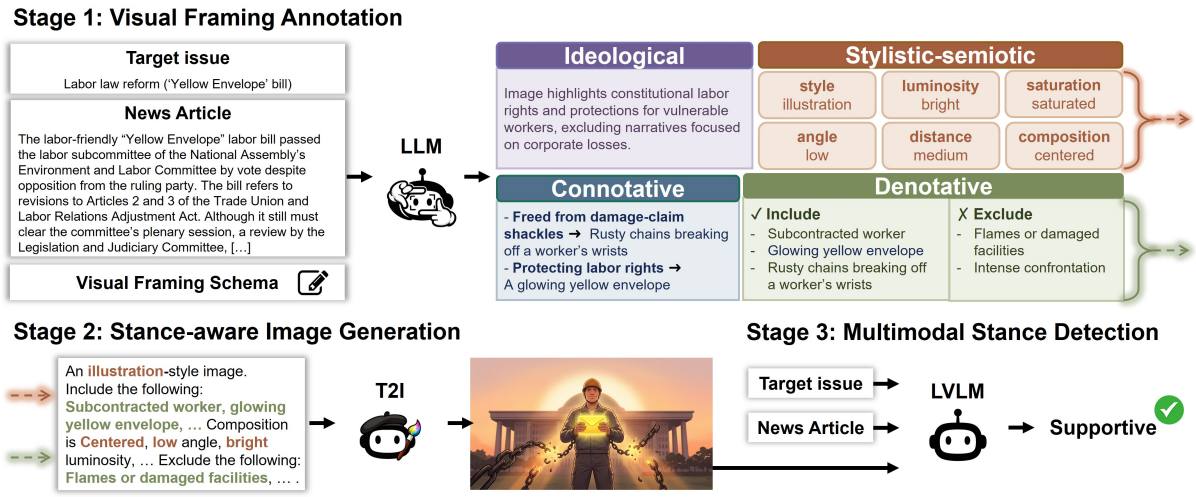


Figure 2: Overview of VFSTANCE: a multi-stage, modular framework grounded in visual framing.

naïve generation using a straightforward prompt).

To address these challenges, VFSTANCE draws on visual framing (Rodriguez and Dimitrova, 2011) to make subtle stance cues more explicit for article-level stance detection. As illustrated in Figure 2, VFSTANCE is a multi-stage, modular framework built on an LLM, a T2I model, and an LVLM. The framework first uses an LLM to derive a structured visual framing specification (**Stage 1**). A T2I model then generates a news image based on this specification (**Stage 2**), making the underlying stance cues more visually evident. Finally, an LVLM uses the article and the generated image to predict article-level stance (**Stage 3**). Stage 2 can also be skipped, in which case the visual framing specification is provided directly to Stage 3 in textual form; we refer to this variant as VFSTANCE (TEXT).

Stage 1: Visual Framing Annotation We first employ an LLM to specify visual framing features. The specification captures both the image content—that is, which actors, objects, or scenes should be included or excluded—and visual presentation, including style, composition, angle, distance, saturation, and luminosity. To construct this specification, we adapt Rodriguez and Dimitrova's (2011) model of visual framing as an annotation schema for image generation. Because the original model does not define variables specifically for image generation, we use prior visual framing literature (Kress and van Leeuwen, 1996; Hall, 1966; Barthes, 1977; Messaris and Abraham, 2001) to define annotatable features suitable for text-to-image generation. The theoretical grounding and operationalization of these features are described in Appendix C.2.

The resulting schema consists of four levels, as illustrated in Stage 1 of Figure 2. At the *ideological* level, the LLM identifies whose perspective or interests the image should serve. At the *connotative* level, it specifies the interpretive associations that the image should evoke beyond its literal depiction. For instance, when an article frames inter-ministerial cooperation as effective coordination, the image may include symbolic elements such as interlocking gears to convey integration and collective action. At the *stylistic-semiotic* level, the LLM assigns values to six features: style, composition, angle, distance, saturation, and luminosity. Together, these features shape the image's visual presentation and tone. At the *denotative* level, the model determines which subjects, objects, or scenes should be included or excluded.

Given a news article A and its target issue T as input, an LLM annotates the ten features across the four levels and returns the visual framing specification in JSON format. Stage 1 derives this specification solely from these inputs, without predicting the article's stance or accessing the gold label. Figure A7 shows the full prompt, and Figure A8 provides an example output.

Stage 2: Stance-Aware Image Generation Given the visual framing specification from Stage 1, we generate a news image using a T2I model. Among the ten features across four levels, we use eight features from the stylistic-semiotic and denotative levels. These levels provide explicit, visually renderable specifications, whereas the ideological and connotative levels capture more abstract interpretive information. Image I is generated using the

Category	Method	ACC	mF1	$F1_{Supportive}$	$F1_{Neutral}$	$F1_{Oppositional}$
VFSTANCE	Gemini-3-flash	**0.746 ± 0.002**	**0.747 ± 0.002**	**0.78 ± 0.003**	**0.649 ± 0.002**	**0.813 ± 0.003**
	Claude-4.6-sonnet	0.694 ± 0.002	0.696 ± 0.002	0.732 ± 0.002	0.556 ± 0.002	0.8 ± 0.001
	GPT-5.4-mini	0.671 ± 0.003	0.675 ± 0.003	0.69 ± 0.004	0.562 ± 0.003	0.774 ± 0.004
Multimodal	Gemini-3-flash	**0.719 ± 0.001**	**0.726 ± 0.001**	**0.73 ± 0.002**	**0.659 ± 0.001**	**0.788 ± 0.001**
	Claude-4.6-sonnet	0.669 ± 0.002	0.674 ± 0.002	0.691 ± 0.005	0.555 ± 0.002	0.778 ± 0.002
	GPT-5.4-mini	0.658 ± 0.002	0.664 ± 0.002	0.652 ± 0.001	0.569 ± 0.004	0.769 ± 0.002
	RoBERTa+ViT	0.619 ± 0.033	0.62 ± 0.036	0.633 ± 0.022	0.597 ± 0.015	0.63 ± 0.075
	CLIP	0.364 ± 0.021	0.345 ± 0.023	0.25 ± 0.071	0.364 ± 0.039	0.421 ± 0.032
	TMPT	0.347 ± 0.019	0.335 ± 0.013	0.272 ± 0.054	0.366 ± 0.045	0.368 ± 0.047
	T-MAD	0.332 ± 0.017	0.306 ± 0.023	0.306 ± 0.033	0.403 ± 0.044	0.208 ± 0.08
Textual	Gemini-3-flash	**0.712 ± 0.002**	**0.719 ± 0.002**	**0.711 ± 0.004**	**0.663 ± 0.002**	**0.781 ± 0.004**
	GPT-5.4-mini	0.653 ± 0.003	0.66 ± 0.003	0.653 ± 0.005	0.563 ± 0.004	0.765 ± 0.005
	Claude-4.6-sonnet	0.636 ± 0.002	0.629 ± 0.002	0.659 ± 0.002	0.451 ± 0.003	0.776 ± 0.003
	PT-HCL	0.621 ± 0.007	0.621 ± 0.005	0.638 ± 0.02	0.604 ± 0.021	0.62 ± 0.015
	LKI-BART	0.618 ± 0.021	0.614 ± 0.027	0.596 ± 0.059	0.578 ± 0.065	0.669 ± 0.025
	RoBERTa	0.604 ± 0.038	0.602 ± 0.04	0.61 ± 0.063	0.604 ± 0.024	0.594 ± 0.038
	CoT Embeddings	0.583 ± 0.07	0.569 ± 0.088	0.608 ± 0.069	0.504 ± 0.196	0.593 ± 0.132
Visual	Gemini-3-flash	**0.46 ± 0.004**	**0.456 ± 0.005**	0.418 ± 0.006	**0.48 ± 0.005**	**0.471 ± 0.005**
	Claude-4.6-sonnet	0.435 ± 0.002	0.433 ± 0.002	**0.425 ± 0.004**	0.453 ± 0.003	0.42 ± 0.004
	GPT-5.4-mini	0.38 ± 0.002	0.35 ± 0.002	0.346 ± 0.004	0.457 ± 0.002	0.246 ± 0.003
	ResNet	0.345 ± 0.007	0.32 ± 0.018	0.218 ± 0.064	0.316 ± 0.042	0.427 ± 0.018
	SwinT	0.337 ± 0.018	0.31 ± 0.026	0.206 ± 0.05	0.39 ± 0.03	0.334 ± 0.104
	ViT	0.317 ± 0.014	0.305 ± 0.018	0.259 ± 0.07	0.363 ± 0.037	0.292 ± 0.025

Table 1: Article-level stance detection performance on the test split of K-News-Stance-MM, measured by accuracy (ACC) and macro F1 (mF1). Categories and models are sorted by accuracy, with the best-performing model in each category highlighted in bold. Models in the VFSTANCE category differ only in the LVLM used as the Stage 3 detector. Methods in the multimodal and visual categories use publisher-provided news images.

template-based prompt shown in Figure A5.

Stage 3: Multimodal Stance Detection The final step is to predict the stance of article A toward a target issue T. The detector model M receives both the article text and the stance-aware image I generated in Stage 2. We use an LVLM as M, with the prompt shown in Figure A4. VFSTANCE aims to improve stance detection by combining the article text, which contains implicit stance cues, with a generated image designed to externalize stance-relevant framing signals. For VFSTANCE (TEXT), the visual framing specification from Stage 1 is provided to the detector instead of I.

Model Configurations We use Gemini-3-flash as the LLM in Stage 1 and Gemini-3.1-flash-image (also known as Nano Banana 2) as the T2I model in Stage 2. The latter was selected for its multilingual support and image generation quality. For Stage 3, we employ Gemini-3-flash as the LVLM because it achieved the best stance detection performance among the three proprietary models evaluated in this study—Gemini-3-flash, GPT-5.4-mini, and Claude-4.6-sonnet. A comparison with open-model alternatives is provided in Appendix D.6.

5 Evaluation Results

We present evaluation results on article-level stance detection, measured by accuracy (ACC) and macro F1 (mF1). We report the average performance over five runs, along with standard errors. The Mann–Whitney U test was used to assess the statistical significance of the differences. Detailed experimental settings are provided in Appendix A.

Comparison with Existing Methods Table 1 presents the article-level stance detection results on the K-News-Stance-MM test split, comparing VFSTANCE with the baseline methods. We evaluated eleven fine-tuned baselines grouped into textual, visual, and multimodal methods, all of which were proposed in previous studies. These models were fine-tuned on samples from the training split. Further details are provided in Appendix C.

Textual baselines include **RoBERTa** (Liu et al., 2019b), a fine-tuned classifier based on a masked language model; **CoT Embeddings** (Gatto et al., 2023), which trains RoBERTa on chain-of-thought

reasoning traces generated by an LLM; **LKI-BART** (Zhang et al., 2024), which injects text–target relational knowledge extracted by an LLM into a BART model; and **PT-HCL** (Liang et al., 2022), which separates target-invariant and target-specific features via contrastive learning.

Visual baselines include **ResNet** (He et al., 2016), a convolutional neural network; **ViT** (Dosovitskiy et al., 2021), a vision transformer; and **SwinT** (Liu et al., 2021), a hierarchical vision transformer with shifted window self-attention.

Multimodal baselines include **RoBERTa+ViT**, which concatenates textual and visual [CLS] representations; **CLIP** (Radford et al., 2021), which concatenates text and image embeddings from pre-trained CLIP encoders; **TMPT** (Liang et al., 2024), a target-aware multimodal prompt-tuning method; and **T-MAD** (Zhang et al., 2025a), a target-driven multimodal alignment method. Additionally, we evaluated the three proprietary LVLMs used as Stage 3 backbones in the proposed method using a straightforward prompting strategy.

Three key observations emerge from Table 1. First, the three LVLMs exhibited strong performance across the three baseline categories, outperforming all fine-tuned methods (p<0.01). Among them, Gemini-3-flash consistently achieved the best performance across categories (p<0.01), supporting its use as the backbone for VFSTANCE. Second, VFSTANCE with Gemini-3-flash as the backbone achieved the best overall performance, with an accuracy of 0.746 and a macro F1 score of 0.747. The proposed method outperformed all baselines (p<0.01), including both LVLM-based and fine-tuned methods. This result provides empirical support for the effectiveness of the proposed framework for article-level stance detection. Third, in terms of class-wise prediction performance, VFSTANCE with Gemini-3-flash achieved higher F1 scores for the supportive and oppositional labels than the multimodal baseline, increasing $F1_{\text{Supportive}}$ from 0.73 to 0.78 (p<0.01) and $F1_{\text{Oppositional}}$ from 0.788 to 0.813 (p<0.01). By contrast, $F1_{\text{Neutral}}$ slightly decreased from 0.659 to 0.649 (p<0.01). These findings suggest that the proposed method is particularly effective at making directional stance cues more explicit.

Ablation: Visual Framing in Image Generation To investigate the contribution of visual framing to image generation, we compared VFSTANCE with variants that use alternative image generation methods in Stage 2; the results are presented in Table 2. *Meta-Prompting* instructs an LLM to generate a prompt for a T2I model by specifying the task goal without using a visual framing schema. *Direct T2I* passes the news text directly to the T2I model to generate stance-aware images for prediction. *EAIG4SD* is an image generation framework proposed for stance detection on tweets (Zhang et al., 2025b) and, to our knowledge, is the only prior method that uses image generation for stance detection. All three methods achieved lower accuracy and macro F1 scores than VFSTANCE, with gaps of at least 0.025 in accuracy and 0.02 in macro F1 (p<0.01). These results provide empirical support for the role of visual framing to image generation and, consequently, to the performance gains achieved by VFSTANCE.

Method	ACC	mF1
VFSTANCE	**0.746 ± 0.002**	**0.747 ± 0.002**
Direct T2I	0.721 ± 0.001	0.723 ± 0.001
EAIG4SD	0.72 ± 0.002	0.727 ± 0.002
Meta-Prompting	0.712 ± 0.004	0.709 ± 0.004

Table 2: Ablation results on the impact of visual framing in image generation on stance detection performance.

Method	ACC	mF1
VFSTANCE	**0.746 ± 0.002**	**0.747 ± 0.002**
VFSTANCE (TEXT)	0.73 ± 0.003	0.725 ± 0.003
1-Step Prompting	0.688 ± 0.007	0.677 ± 0.008

Table 3: Ablation results on the impact of image generation on stance detection performance.

Ablation: Use of Image Generation Given the effectiveness of visual framing in image generation identified above, we next investigated whether image generation itself is necessary. We compare VFSTANCE with two alternative methods that do not use image generation but remain grounded in visual framing; their performance is reported in Table 3. VFSTANCE (TEXT) uses the visual framing specification from Stage 1 as input context for the LVLM in Stage 3 while skipping image generation in Stage 2. *1-Step Prompting* instructs an LVLM, Gemini-3-flash, to perform visual framing annotation and stance prediction in a single step.

The results show that VFSTANCE achieves statistically significant improvements over both alternatives without image generation (p<0.01), indicating that image generation provides an addi-

D	S	C	I	ACC	mF1
✓	✓			**0.746 ± 0.002**	**0.747 ± 0.002**
✓				0.732 ± 0.002	0.736 ± 0.002
✓	✓	✓		0.725 ± 0.002	0.728 ± 0.002
✓	✓	✓	✓	0.722 ± 0.003	0.723 ± 0.003

Table 4: Stance detection performance according to the visual framing levels adopted in Stage 2 (D: denotative, S: stylistic-semiotic, C: connotative, and I: ideological).

D	S	C	I	ACC	mF1
✓	✓	✓	✓	**0.746 ± 0.002**	**0.747 ± 0.002**
✓	✓			0.739 ± 0.003	0.737 ± 0.003

Table 5: Stance detection performance according to the visual framing levels used in Stage 1 (D: denotative, S: stylistic-semiotic, C: connotative, and I: ideological).

Method	ACC	mF1
VFSTANCE		
Gemini-3-flash	**0.618 ± 0.002**	**0.62 ± 0.002**
Claude-4.6-sonnet	0.6 ± 0.003	0.595 ± 0.003
GPT-5.4-mini	0.598 ± 0.003	0.59 ± 0.003
Textual		
Gemini-3-flash	0.605 ± 0.001	0.58 ± 0.001
GPT-5.4-mini	0.584 ± 0.003	0.578 ± 0.003
Claude-4.6-sonnet	0.577 ± 0.003	0.566 ± 0.003
RoBERTa	0.526 ± 0.014	0.442 ± 0.034
PT-HCL	0.517 ± 0.021	0.39 ± 0.048
LKI-BART	0.466 ± 0.01	0.34 ± 0.011
CoT Embeddings	0.424 ± 0.01	0.198 ± 0.003

Table 6: Stance detection performance on CheeSE, a German dataset without original news images.

tional performance gain beyond the contribution of visual framing identified in Table 2. Another notable finding is the strong performance of VFSTANCE (TEXT), which outperforms all baseline methods in Table 1, as well as *1-Step Prompting*. These results provide empirical evidence that visual framing contributes substantially even when represented only as a textual specification rather than rendered as an image. Considering the cost–accuracy trade-off between the two variants (Table A4), VFSTANCE and VFSTANCE (TEXT) may serve different practical needs: users may prefer VFSTANCE (TEXT) when computational cost is the primary concern and thus omit image generation, whereas VFSTANCE is preferable when maximizing detection accuracy is the priority.

Ablation: Visual Framing Level Selection We present ablation studies to support our choice of the denotative and stylistic-semiotic levels in Stage 2, selected from the four levels produced by the LLM-based annotation in Stage 1. In each comparison, only the features from the selected levels are used in the image-generation prompt for Stage 2, while all four levels are annotated identically in Stage 1. Table 4 presents the level-wise ablation results, showing that the two levels adopted in VFSTANCE are critical for stance detection performance, whereas the connotative and ideological levels are ineffective and even reduce accuracy and macro F1 when included. This finding suggests that the abstract nature of the two excluded levels makes them difficult to render visually in a way that makes stance signals more explicit, consistent with prior findings on the difficulty of generating abstract concepts (Liao et al., 2024).

Given the effectiveness of these two levels, we further examined whether annotating the visual framing specification across all four levels in Stage 1 is necessary. Specifically, we measured the performance of VFSTANCE when only the stylistic-semiotic and denotative levels were included in the Stage 1 annotation schema. Table 5 shows that additionally annotating the connotative and ideological features improves the effectiveness of the resulting annotations for the eight denotative and stylistic-semiotic features, yielding a 0.01 increase in macro F1. According to the hierarchical structure of Rodriguez and Dimitrova (2011)'s model, the connotative and ideological levels provide interpretive context for lower visual elements, which may partially explain this performance gain.

Effectiveness Across Languages To assess whether VFSTANCE is effective across datasets and languages, we evaluated it on CheeSE (Mascarell et al., 2021), a German article-level news stance detection dataset. Since CheeSE does not provide original news images, VFSTANCE was compared against seven textual baseline methods: three LVLM-based and four fine-tuned methods. As shown in Table 6, VFSTANCE achieved higher accuracy and macro F1 scores across multiple LVLM backbones. Gemini-3-flash again performed best, achieving an accuracy of 0.618 and a macro F1 score of 0.62, outperforming all baseline methods by a substantial margin ($p<0.01$). These results,

together with the Korean-language findings above, suggest that VFSTANCE can be effective across languages. To further support this finding, we provide supplementary results on translated versions of K-News-Stance-MM in four languages in Appendix D.4.

6 User Study

We conduct a controlled user study to examine whether images generated by VFSTANCE help news readers identify article stance. Specifically, we investigate whether participants can accurately discern article stance in a snippet-based news consumption setting where only limited textual information is available, reflecting common patterns of news consumption in online information environments and social feeds (Gabielkov et al., 2016). We recruited 200 native Korean speakers through PMI Research & Consulting (PMI)[1], with the sample balanced by gender and age.

For the user study, we selected nine articles covering three issues, with one supportive, one neutral, and one oppositional article per issue. These articles were published after June 2024, outside the period covered by K-News-Stance-MM. Each experimental snippet consisted of the article headline, the first two to three sentences of its lead paragraph, and a visual treatment determined by the experimental condition. For each article, we created four presentation versions with identical text: (1) **Text-only**, with no accompanying image; (2) **Original**, with the publisher's original image; (3) **Naïve**, with an image generated using a straightforward prompt without a visual framing specification; and (4) **Proposed**, with an image generated by VFSTANCE.

Each participant viewed all nine articles in randomized order, with each article randomly assigned to one of the four presentation conditions. This design yielded approximately 50 observations per condition for each article and approximately 450 observations per condition overall. After viewing each snippet, participants classified the article's stance toward the target issue as supportive, neutral, or oppositional through a web survey interface, a screenshot of which is provided in Figure A10b.

Figure 3 reports stance identification accuracy across conditions. VFSTANCE achieved the highest accuracy of 0.378, exceeding the text-only, original, and naïve conditions by 0.071, 0.098, and 0.076, respectively. A mixed-effects logistic regression, with presentation condition as a fixed effect and random intercepts for participants and articles, further showed that all three comparison conditions had significantly lower odds of correct stance identification than VFSTANCE: text-only ($\mathrm{OR} = 0.708$, $p = 0.001$), original ($\mathrm{OR} = 0.620$, $p < 0.0001$), and naïve ($\mathrm{OR} = 0.697$, $p = 0.0006$).

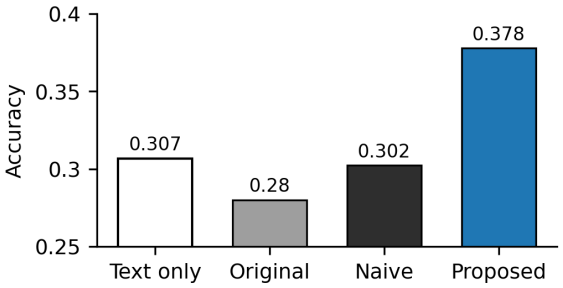


Figure 3: Stance identification accuracy across presentation conditions in a controlled user study under snippet-based news consumption.

Despite modest overall accuracy, these findings indicate that VFSTANCE-generated images facilitate stance identification under abbreviated news exposure. Their advantage over text-only, publisher-provided and naïvely generated images further highlights their potential utility beyond automated stance detection.

7 Conclusion

This study applies framing theory (Entman, 1993) and its visual extension (Rodriguez and Dimitrova, 2011) to article-level news stance detection. Despite advances in NLP and LLMs, the task remains challenging because stance cues in long, structurally complex news articles are often implicit and dispersed throughout the text. VFSTANCE is a multi-stage, modular stance detection framework in which an LLM produces a visual framing specification, a T2I model generates a stance-aware image, and an LVLM predicts article-level stance using the article and the generated image.

Evaluation results demonstrate that VFSTANCE outperforms existing stance detection methods and that both visual framing and image generation contribute to its performance. A controlled user study in a snippet-based news consumption setting further shows that images generated by VFSTANCE improve stance identification, shedding light on potential applications beyond automated stance detection. More broadly, the findings suggest that visual framing can function as an intermediate representa-

[1] https://pmirnc.com/

tional layer, transforming dispersed textual stance cues into signals that are more readily accessible to both computational models and human readers.

Taken together, these findings point to the potential of generative AI grounded in visual framing to make media perspectives more transparent, thereby supporting the identification of news bias and contributing to more pluralistic media environments. Future work could extend this approach to other areas of AI and NLP that involve implicit framing or evaluative signals, such as argument mining, bias analysis, and model bias auditing. The data, code, and prompts are available at https://github.com/ssu-humane/VFStance.

Limitations

Computational Costs VFSTANCE employs three models across its corresponding stages. Encouragingly, as discussed in Section 5, VFSTANCE (TEXT) offers a computationally efficient alternative by omitting image generation while still outperforming all baseline methods. This further demonstrates the effectiveness of the visual framing schema used in the proposed method. Considering the cost–accuracy trade-off (Appendix D.2), VFSTANCE (TEXT) could be used when computational cost is the primary concern, whereas VFSTANCE remains the strongest choice when maximizing detection accuracy is the priority.

Multilingual Evaluation The primary testbed, K-News-Stance-MM, is a Korean corpus, which limits the scope of multilingual evaluation in this study. This choice was necessary because K-News-Stance-MM is the first and only dataset to provide publisher images for article-level stance detection, which are required to compare VFSTANCE with visual and multimodal baselines. To provide additional evidence of multilingual effectiveness, we conducted experiments on CheeSE, a German text-only dataset, and on LLM-translated versions of K-News-Stance-MM in four languages (Appendix D.4). Future studies could construct article-level stance detection datasets with accompanying news images by adapting the guidelines provided by Lee et al. (2025).

Model Selection Proprietary models were selected as primary backbones for VFSTANCE because of their stronger language understanding, reasoning, and generation capabilities. We provide supplementary results for open-weight models in Appendix D.6, where they achieved lower performance. Because the framework is modular and model-agnostic, future work could evaluate a broader range of models and incorporate additional training to further improve detection accuracy.

Ethical Considerations

This study was approved by the Institutional Review Board at Soongsil University (SSU-202604-HR-804-1).

Copyright and Privacy Issues K-News-Stance-MM extends an existing dataset (Lee et al., 2025) built from news articles distributed through the Naver News platform. To respect the intellectual property rights of the original news publishers, K-News-Stance-MM is released under gated access with a custom Data Use Agreement: prospective users must submit an access request describing their research purpose and agree to terms that restrict use to non-commercial academic research and prohibit redistribution. The news data include the names of public figures, but private individuals are anonymized when present; thus, the dataset does not contain any personally identifiable information about private individuals, as manually verified for all samples in K-News-Stance-MM. CheeSE (Mascarell et al., 2021) is a publicly released benchmark and is used under its original release terms.

User Study Participants We recruited 200 Korean native speakers residing in South Korea through a survey panel provider. The task took approximately 10 minutes, with compensation of about USD 3.6, which exceeded the Korean statutory minimum hourly wage. All participants provided informed consent and could withdraw at any time. The study collected no sensitive personal information as defined by the Korean Personal Information Protection Act. Two potential risks were disclosed in advance: minimal-risk discomfort from politically contested content and the indirect inference of attitudes from stance judgments. Responses contained only randomly assigned IDs; therefore, individual records could not be selectively deleted after submission, a limitation that was also disclosed before consent.

Risks Associated with Image Generation Images generated by VFSTANCE could be repurposed for reader-facing applications, as demonstrated in

the user study in Section 6. However, such use requires careful consideration because T2I models may reproduce or amplify social biases. Before its use, generated images should be clearly labeled as synthetic and reviewed to ensure compliance with applicable defamation, personality-rights, and synthetic-media regulations.

AI Assistant Use We used AI-assisted language editing tools, primarily ChatGPT, exclusively for grammar checking and improving readability.

Acknowledgements

This research was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP), funded by the Korea government (MSIT) (IITP-2026-RS-2022-00156360, IITP-2026-RS-2024-00430997, and IITP-2026-RS-2020-II201602), and by the National Research Foundation of Korea (NRF), funded by the Korea government (MSIT) (RS-2023-00252535). KP and JH are the corresponding authors.

References

Mohammad Ali and Naeemul Hassan. 2022. A survey of computational framing analysis approaches. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9335–9348, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nouf AlShenaifi and Nourah Alangari. 2026. Beyond text: Multimodal stance detection in arabic tweets. *Machine Learning with Applications*, 23:100823.

Anthropic. 2026. Claude sonnet 4.6 system card. https://www-cdn.anthropic.com/78073f739564e986ff3e28522761a7a0b4484f84.pdf. Published: February 17, 2026. Accessed: 2026-05-24.

Arnav Arora, Srishti Yadav, Maria Antoniak, Serge Belongie, and Isabelle Augenstein. 2025. Multi-modal framing analysis of news. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31531–31553, Suzhou, China. Association for Computational Linguistics.

Associated Press. 2024. Ap news values and principles. https://www.ap.org/wp-content/uploads/2024/02/ap-news-values-and-principles-1.pdf. Accessed: 2026-05-18.

Roland Barthes. 1977. Rhetoric of the image. In Stephen Heath, editor, *Image, Music, Text*, pages 32–51. Hill and Wang, New York.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, Beijing, China. Association for Computational Linguistics.

Dennis Chong and James N Druckman. 2007. Framing theory. *Annu. Rev. Polit. Sci.*, 10(1):103–126.

Renita Coleman. 2010. Framing the pictures in our heads: Exploring the framing and agenda-setting effects of visual images. In *Doing news framing analysis*, pages 249–278. Routledge.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. STANDER: An expert-annotated dataset for news stance detection and evidence retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4086–4101, Online. Association for Computational Linguistics.

Iain Cruickshank and Lynnette Ng. 2026. Prompting and fine-tuning open source large language models for stance classification. *ACM Transactions on Intelligent Systems and Technology*, 17:1–25.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Kareem El Damanhoury, Carol Winkler, Ayse D. Lokmanoglu, and Keyu Alexander Chen Glanz. 2026. Visual framing in the AI era: Lessons from manual approaches for computational methods. *Computational Communication Research*, 8(1):1–41.

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

William Ferreira and Andreas Vlachos. 2016. Emergent: a novel data-set for stance classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1163–1168.

Maksym Gabielkov, Arthi Ramachandran, Augustin Chaintreau, and Arnaud Legout. 2016. Social clicks: What and who gets read on twitter? In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Science*, SIGMETRICS '16, page 179–192, New York, NY, USA. Association for Computing Machinery.

Joseph Gatto, Omar Sharif, and Sarah Preum. 2023. Chain-of-thought embeddings for stance detection on social media. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4154–4161, Singapore. Association for Computational Linguistics.

Matthew Gentzkow and Jesse M. Shapiro. 2010. What drives media slant? evidence from u.s. daily newspapers. *Econometrica*, 78(1):35–71.

Google. 2026a. Gemini 3.1 flash image preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-flash-image. Accessed: 2026-05-25.

Google. 2026b. Gemini 3.1 Pro Preview. https://ai.google.dev/gemini-api/docs/models/gemini-3.1-pro-preview. Accessed: 2026-05-26.

Google DeepMind. 2025. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Published: December 2025. Accessed: 2026-05-24.

Edward T. Hall. 1966. *The Hidden Dimension*. Doubleday, Garden City, NY.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles : an interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Bill Kovach and Tom Rosenstiel. 2021. *The elements of journalism, revised and updated 4th edition: What newspeople should know and the public should expect*. Crown.

Gunther Kress and Theo van Leeuwen. 1996. *Reading Images: The Grammar of Visual Design*. Routledge, London and New York.

Dahyun Lee, Jonghyeon Choi, Jiyoung Han, and Kunwoo Park. 2025. Journalism-guided agentic in-context learning for news stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15393–15416, Suzhou, China. Association for Computational Linguistics.

Bin Liang, Zixiao Chen, Lin Gui, Yulan He, Min Yang, and Ruifeng Xu. 2022. Zero-shot stance detection via contrastive learning. In *Proceedings of the ACM Web Conference 2022*, pages 2738–2747. Association for Computing Machinery.

Bin Liang, Ang Li, Jingqian Zhao, Lin Gui, Min Yang, Yue Yu, Kam-Fai Wong, and Ruifeng Xu. 2024. Multi-modal stance detection: New datasets and model. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12373–12387, Bangkok, Thailand. Association for Computational Linguistics.

Jiayi Liao, Xu Chen, Qiang Fu, Lun Du, Xiangnan He, Xiang Wang, Shi Han, and Dongmei Zhang. 2024. Text-to-image generation for abstract concepts. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019a. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019b. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Linqi Lu, Zihan Wan, Hyerin Kwon, Sang Jung Kim, Jiwon Kang, Laila Abbas, Jiawei Liu, and Douglas Mcleod. 2026. Evaluating large vision-language models for visual framing analysis in news imagery: A theory-driven benchmark.

Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. 2021. Stance detection in German news articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic. Association for Computational Linguistics.

Albert Mehrabian. 1981. *Silent Messages: Implicit Communication of Emotions and Attitudes*, 2nd edition. Wadsworth, Belmont, CA.

Paul Messaris and Linus Abraham. 2001. The role of images in framing news stories. In Stephen D. Reese, Oscar H. Gandy, and August E. Grant, editors, *Framing Public Life: Perspectives on Media and Our Understanding of the Social World*, pages 215–226. Lawrence Erlbaum Associates, Mahwah, NJ.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California. Association for Computational Linguistics.

OpenAI. 2026. Gpt-5.4 mini. `https://developers.openai.com/api/docs/models/gpt-5.4-mini`. OpenAI API model documentation. Accessed: 2026-05-24.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

Allan Paivio. 1986. *Mental Representations: A Dual Coding Approach*. Oxford University Press, New York.

Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. 2009. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 443–452.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Ji Yoon Han, Jangwon Park, Chisung Song, Junseong Kim, Youngsook Song, Taehwan Oh, and 1 others. 2021. Klue: Korean language understanding evaluation.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multilingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2343–2361, Toronto, Canada. Association for Computational Linguistics.

Dean Pomerleau and Delip Rao. 2017. Fake news challenge stage 1 (FNC-I): Stance detection. `http://www.fakenewschallenge.org`.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Reuters. 2008. *Reuters Handbook of Journalism*. Thomson Reuters.

Lulu Rodriguez and Daniela V Dimitrova. 2011. The levels of visual framing. *Journal of visual literacy*, 30(1):48–65.

Dona Schwartz. 1992. To tell the truth : codes of objectivity in photojournalism. *Communicatio*, 13:95–109.

Isidora Tourni, Lei Guo, Taufiq Husada Daryanto, Fabian Zhafransyah, Edward Edberg Halim, Mona Jalal, Boqi Chen, Sha Lai, Hengchang Hu, Margrit Betke, Prakash Ishwar, and Derry Tanti Wijaya. 2021. Detecting frames in news headlines and lead images in U.S. gun violence coverage. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4037–4050, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Gisela Vallejo, Timothy Baldwin, and Lea Frermann. 2024. Connecting the dots in news analysis: Bridging the cross-disciplinary disparities in media bias and framing. In *Proceedings of the Sixth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS 2024)*, pages 16–31, Mexico City, Mexico. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2023. Identification of multimodal stance towards frames of communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12597–12609, Singapore. Association for Computational Linguistics.

Maxwell Weinzierl and Sanda Harabagiu. 2024. Tree-of-counterfactual prompting for zero-shot stance detection. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–880, Bangkok, Thailand. Association for Computational Linguistics.

Zhao Zhang, Yiming Li, Jin Zhang, and Hui Xu. 2024. LLM-driven knowledge injection advances zero-shot and cross-target stance detection. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 371–378, Mexico City, Mexico. Association for Computational Linguistics.

ZhaoDan Zhang, Jin Zhang, Xueqi Cheng, and Hui Xu. 2025a. T-MAD: Target-driven multimodal alignment for stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 580–595, Suzhou, China. Association for Computational Linguistics.

Zhengkang Zhang, Zhongqing Wang, and Guodong Zhou. 2025b. Exploring artificial image generation for stance detection. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19846–19861, Suzhou, China. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can ChatGPT reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Experimental Setups

This section provides the experimental details of our study. Each result is reported as the average over five runs with the standard error: trainable models were run using random seeds 42–46, while API-based models were evaluated five times under the same inference configuration because not all backbones support user-specified seeds. Experiments were conducted on three NVIDIA RTX A6000 GPUs (48GB each) with 128GB RAM, using Python 3.10, PyTorch 2.4.1, Transformers 4.57.1, and CUDA 12.1.

We accessed GPT-5.4-mini (OpenAI, 2026), Claude-4.6-sonnet (Anthropic, 2026), and Gemini-3-flash (Google DeepMind, 2025) via API, with the thinking level set to low, a maximum of 16,000 output tokens, and a temperature of 1 for all models, since Claude-4.6-sonnet fixes the temperature to 1 when reasoning is enabled. Images were generated with Gemini-3.1-flash-image (Google, 2026a) under the default configuration (1K resolution), and automatic image quality assessment used Gemini-3.1-Pro (Google, 2026b) with default settings.

For textual baselines, we used KLUE-RoBERTa-large (Park et al., 2021) for K-News-Stance-MM and XLM-RoBERTa-large for CheeSE as the backbones for RoBERTa, CoT Embeddings, and PT-HCL. For LKI-BART, we used KoBART-base-v2 and BART-qg-German for K-News-Stance-MM and CheeSE, respectively. We set the learning rate to 3×10^{-5}, used a batch size of 32 for CoT Embeddings and 16 for the other baselines, froze the bottom seven layers, and used GPT-4o-mini as the LLM for CoT Embeddings and LKI-BART.

For visual baselines, we used ResNet-50, ViT-B/16, and SwinV2-Base with a batch size of 32, a linear scheduler with a warmup ratio of 0.1, and early stopping with a patience of 3. The learning rate was set to 1×10^{-4} for ResNet-50 and 5×10^{-5} for ViT-B/16 and SwinV2-Base, using the default image preprocessing provided with each checkpoint. For multimodal baselines, we used KLUE-RoBERTa-large and ViT-B/16 as the text and visual encoders, respectively, for RoBERTa+ViT, TMPT, and T-MAD, while a Korean CLIP variant was used for the CLIP baseline. We used a learning rate of 2×10^{-5} and a batch size of 16 for these baselines.

For EAIG4SD (Zhang et al., 2025b), we followed the original setup (Stable Diffusion 3 Medium, 28 denoising steps, and a guidance scale of 7), replacing unsupported components with Korean-capable alternatives: Qwen2.5-VL-7B-Instruct for the intermediate stance and sentiment prediction used in prompt construction (LoRA fine-tuned with rank 16, alpha 32, learning rate 1×10^{-4}, and 3 epochs) and a Korean CLIP variant for image–text similarity, with images selected via PageRank (damping factor 0.85) over the similarity graph.

For the open-weight experiments (Appendix D.6), InternVL3-14B-Instruct and Gemma3-12B-Instruct were evaluated zero-shot with the same prompts used for the proprietary models, using a temperature of 1 with a maximum of 1,024 output tokens for annotation and 16 for stance prediction. Stable Diffusion 3.5 Large generated one image per article with 28 denoising steps and a guidance scale of 4.5.

All fine-tuned models were trained with the AdamW optimizer and a weight decay of 0.01. Unless otherwise specified, hyperparameters were selected using validation subsets held out from the training split, and other method-specific settings followed the original studies.

The model IDs and parameter sizes used in the experiments are provided below.

- GPT-5.4-mini: `gpt-5.4-mini-2026-03-17` (Parameter size: unknown)
- Claude-4.6-sonnet: `claude-sonnet-4-6` (Parameter size: unknown)
- Gemini-3-flash: `gemini-3-flash-preview` (Parameter size: unknown)
- Gemini-3.1-flash-image: `gemini-3.1-flash-image-preview` (Parameter size: unknown)
- GPT-4o-mini: `gpt-4o-mini` (Parameter size: unknown)
- Qwen2.5-VL-7B-Instruct: https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct (Parameter size: 7B)
- InternVL3-14B-Instruct: https://huggingface.co/OpenGVLab/InternVL3-14B-Instruct (Parameter size: 15B)

- Gemma3-12B-Instruct: https://huggingface.co/google/gemma-3-12b-it (Parameter size: 12B)
- KLUE-RoBERTa-large: https://huggingface.co/klue/roberta-large (Parameter size: 337M)
- XLM-RoBERTa-large: https://huggingface.co/FacebookAI/xlm-roberta-large (Parameter size: 561M)
- BART-qg-German: https://huggingface.co/su157/bart-qg-german (Parameter size: 139M)
- KoBART-base-v2: https://huggingface.co/gogamza/kobart-base-v2 (Parameter size: 124M)
- ResNet-50: https://huggingface.co/microsoft/resnet-50 (Parameter size: 26M)
- ViT-base-patch16-224: https://huggingface.co/google/vit-base-patch16-224 (Parameter size: 86M)
- SwinV2-Base: https://huggingface.co/microsoft/swinv2-base-patch4-window12-192-22k (Parameter size: 88M)
- KoCLIP: https://huggingface.co/Bingsu/clip-vit-large-patch14-ko (Parameter size: 428M)
- Stable-diffusion-3-medium: https://huggingface.co/stabilityai/stable-diffusion-3-medium-diffusers (Parameter size: 2B)
- Stable-diffusion-3.5-large: https://huggingface.co/stabilityai/stable-diffusion-3.5-large (Parameter size: 8B)
- Gemini-3.1-pro: `gemini-3.1-pro-preview` (Parameter size: unknown)

Split	Total	Supportive	Neutral	Oppositional
Train	909	279	310	320
Test	907	289	305	313

(a) K-News-Stance-MM

Split	Total	In Favor	Discussing	Against
Train	1000	399	439	162
Test	762	303	335	124

(b) CheeSE

Table A1: Dataset sizes and label distributions across the data splits.

B Dataset Details

Table A1 presents the dataset sizes and label distributions across the data splits. Below we provide detailed information about each dataset.

K-News-Stance-MM is a multimodal extension of the existing article-level stance detection dataset (Lee et al., 2025). The original dataset consists of 2,000 news articles in Korean, with 999 articles for training and 1,001 for testing. Each article's stance toward a social issue (e.g., "The passage of the Yellow Envelope Act in a National Assembly standing committee") is labeled as *supportive*, *neutral*, or *oppositional*. We extended the dataset by crawling news images from the original webpages of each article, selecting the lead image (i.e., the first image displayed) when multiple were available. We successfully collected images for 1,816 articles (90.8%), comprising 909 training samples and 907 testing samples. Following the original dataset, we preserved the issue-level train/test split to prevent issue-level leakage across the splits. To our knowledge, this is the first dataset to provide article-level stance labels together with news images, and it serves as the primary testbed for this study. An example instance is provided in Table A9 with an English translation. For its broader accessibility, we provide translations into four widely spoken languages: English, Chinese, Indonesian, and Arabic through the GitHub repository. Gemini-3-flash was used for translation, while the stance labels and data splits were kept unchanged. A preliminary comparison of VFSTANCE with a text-only baseline is provided in Appendix D.4.

CheeSE is a news stance detection dataset consisting of German news articles (Mascarell et al., 2021). The original dataset contains 3,693 news articles with article-level stance annotations with

respect to debate questions (e.g., "Are abortions morally acceptable?"). We excluded 503 *unklar* (unclear) samples and 1,428 *Kein Bezug* (unrelated) samples, resulting in 1,762 samples. We split the resulting dataset into 800/200/762 training, validation, and test samples, respectively, while preserving label distributions across the splits. We use this text-only dataset to evaluate the proposed method in a setting without original news images.

C Method Details

This section provides details on the baseline and proposed methods. Experimental settings for all methods are summarized in Appendix A.

C.1 Baseline Methods

Textual Baselines For RoBERTa, the input sequence is constructed by concatenating the target issue, headline, and article text as `[CLS] issue [SEP] headline [SEP] article [SEP]`, and the `[CLS]` representation is used for stance classification. CoT Embeddings (Gatto et al., 2023) augments this input with an LLM-generated rationale explaining the article's stance toward the target issue. LKI-BART (Zhang et al., 2024) first generates stance-relevant background knowledge from the target issue and article and incorporates it into its generation-based stance prediction framework. PT-HCL (Liang et al., 2022) uses the same textual input format and jointly optimizes stance classification with hierarchical contrastive objectives. The zero-shot LVLM baseline in this category uses the prompt shown in Figure A2.

Visual Baselines ResNet, ViT, and SwinT predict stance using only the original publisher image associated with each article. The zero-shot LVLM baseline uses the prompt shown in Figure A3.

Multimodal Baselines For RoBERTa+ViT, RoBERTa encodes the target issue, headline, and article, ViT encodes the original publisher image, and the resulting `[CLS]` representations are concatenated for stance prediction. For CLIP, text and image embeddings from the pretrained encoders are concatenated and passed through a classification layer. TMPT (Liang et al., 2024) augments both inputs with target-aware prompts before multimodal fusion, and T-MAD (Zhang et al., 2025a) uses the separately encoded target representation to derive target-aligned multimodal features. The zero-shot LVLM baseline uses the prompt shown in Figure A4, which is identical to the Stage 3 prompt of VFSTANCE except that the original news image is provided.

Level	Feature	Type
Ideological	-	Open-ended
Connotative	-	Open-ended
Stylistic-Semiotic	Style	Categorical (photo, illustration)
	Composition	Categorical (centered, split, asymmetric, crowded)
	Angle	Categorical (low, eye-level, high)
	Distance	Categorical (close-up, medium, long)
	Saturation	Categorical (saturated, neutral, desaturated)
	Luminosity	Categorical (bright, neutral, dark)
Denotative	Inclusion	Open-ended
	Exclusion	Open-ended

Table A2: Visual framing schema used in Stage 1.

EAIG4SD EAIG4SD (Zhang et al., 2025b) generates multiple stance-aware candidate images from prompts constructed using the article content, target issue, predicted stance, and sentiment, and selects the final image via PageRank using text–image similarity, target consistency, and stance consistency signals. Since we use EAIG4SD only as an image-generation baseline, we apply its image-generation and selection procedures and feed the selected image to the LVLM used in Stage 3 of VFSTANCE, with the same prompt (Figure A4), for final stance prediction.

C.2 VFSTANCE

Visual Framing Features We describe the literature grounding each feature in our schema (Table A2). The four-level structure of the schema follows Rodriguez and Dimitrova's (2011) model of visual framing, in which lower levels capture concrete visual elements and higher levels capture their interpretive meanings. At the denotative level, the *inclusion* and *exclusion* features operationalize the framing functions of selection, emphasis, and omission (Rodriguez and Dimitrova, 2011; Messaris and Abraham, 2001; Entman, 1993): they specify which actors, objects, and scenes are foregrounded in or omitted from the image. At the stylistic-semiotic level, the *composition* and *angle* features derive from the grammar of visual design (Kress and van Leeuwen, 1996), particularly the spatial organization of visual elements and the interpersonal meanings of viewing position; the *distance* feature additionally reflects proxemic and social-distance cues (Hall, 1966; Kress and van Leeuwen, 1996). The *saturation* and *luminosity* features correspond to color intensity and brightness as cues of visual modality and expressive tone (Kress and van Leeuwen, 1996). The *style*

feature is a generation-oriented operationalization of representational modality (Messaris and Abraham, 2001; Kress and van Leeuwen, 1996). The connotative level captures symbolic or associative meaning beyond literal depiction (Rodriguez and Dimitrova, 2011; Barthes, 1977), and the ideological level concerns the perspectives and interests served by the image (Rodriguez and Dimitrova, 2011); both are annotated in Stage 1 but not rendered in Stage 2, providing the interpretive context for the lower-level features.

Prompts and Examples We describe the prompts used in each stage of VFSTANCE using a running example. This manuscript presents English translations of the prompts, while the original Korean prompts and examples are made publicly available in our GitHub repository.

In Stage 1, the LLM receives a news article as input together with the prompt shown in Figure A7, and produces the visual framing specification shown in Figure A9 (original Korean output in Figure A8). In Stage 2, the eight features from the stylistic-semiotic and denotative levels of this specification are inserted into a template-based prompt: each of the six stylistic-semiotic features—style, composition, angle, distance, saturation, and luminosity—fills its corresponding slot, and the `include_subjects` and `exclude_subjects` lists from the denotative level are inserted into the subject slots, yielding the T2I prompt shown in Figure A5. This prompt is then used to generate one image per article (Figure A6), with the API call retried if necessary until a valid image is returned. In Stage 3, the prompt in Figure A4 provides the target issue, news headline, article, and generated image to the LVLM, which returns a single-word stance label.

D Supplementary Results

D.1 Performance with Publisher Images

We conducted a supplementary analysis of the effects of publisher-provided news images and examined whether their inclusion improved overall detection performance. On the test split of K-News-Stance-MM, we compared the zero-shot performance of the three proprietary LVLMs used in the main experiments under three input settings: (1) image only, (2) text only, and (3) text and image. These modality-controlled experiments allow us to assess the contribution of publisher images to stance detection performance.

Image	Text	ACC	mF1
Gemini-3-flash			
✓		0.458	0.453
	✓	0.713	0.72
✓	✓	**0.718**	**0.725**
Claude-4.6-sonnet			
✓		0.437	0.434
	✓	0.637	0.63
✓	✓	**0.668**	**0.675**
GPT-5.4-mini			
✓		0.383	0.354
	✓	0.649	0.656
✓	✓	**0.656**	**0.662**

Table A3: Performance with publisher-provided images.

Method	ACC	API cost
VFSTANCE	0.746 ± 0.002	\$0.0378
VFSTANCE (TEXT)	0.73 ± 0.003	\$0.00392
1-Step Prompting	0.688 ± 0.007	\$0.00169

Table A4: Accuracy–cost trade-offs between VFSTANCE and its efficient alternatives grounded in visual framing. API costs are averaged per sample.

Table A3 presents the results. When only images were used, all models performed substantially worse than in the text-only setting. GPT-5.4-mini showed the lowest image-only performance, with an accuracy of 0.383 and a macro F1 of 0.354. This finding highlights the primary role of news text in article-level stance prediction. In contrast, all models achieved their best performance when both text and image modalities were incorporated. While the improvement was most pronounced for Claude-4.6-sonnet, only modest gains were observed for Gemini-3-flash and GPT-5.4-mini. These findings suggest that publisher images can provide useful complementary signals for stance detection, while also highlighting their limited added value, potentially because stance-relevant cues in conventional news images are often implicit.

D.2 Accuracy–Cost Trade-offs

We examined the trade-off between detection performance and API cost by comparing VFSTANCE with more efficient alternatives that remain grounded in visual framing. Table A4 presents the accuracy and API cost of the three methods, clearly demonstrating the resulting trade-offs.

Configuration	ACC	mF1
VFSTANCE	**0.746 ± 0.002**	**0.747 ± 0.002**
w/o compositional	0.738 ± 0.004	0.74 ± 0.004
w/o exclusion	0.736 ± 0.002	0.738 ± 0.003
w/o style	0.733 ± 0.002	0.736 ± 0.002
w/o tone	0.731 ± 0.002	0.733 ± 0.002

Table A5: Performance differences after ablating visual framing features in Stage 2.

Method	ACC	mF1
English		
VFSTANCE	**0.738 ± 0.002**	**0.74 ± 0.002**
Text	0.697 ± 0.003	0.704 ± 0.003
Chinese		
VFSTANCE	**0.735 ± 0.001**	**0.736 ± 0.001**
Text	0.7 ± 0.002	0.707 ± 0.002
Indonesian		
VFSTANCE	**0.725 ± 0.003**	**0.728 ± 0.003**
Text	0.68 ± 0.004	0.688 ± 0.004
Arabic		
VFSTANCE	**0.722 ± 0.003**	**0.724 ± 0.003**
Text	0.673 ± 0.001	0.68 ± 0.001

Table A6: Article-level stance detection performance on translated versions of K-News-Stance-MM.

D.3 Visual Framing Feature Selection

We conducted a feature-level ablation study to support the selection of eight features in Stage 2 for VFSTANCE. For the features at the stylistic-semiotic level, we grouped composition, angle, and distance into *compositional* features, and saturation and luminosity into *tone* features, based on their conceptual relatedness (Kress and van Leeuwen, 1996). At the denotative level, we ablated only the exclusion feature because the inclusion feature specifies the content to be rendered. Table A5 reports the results, where "w/o X" indicates VFSTANCE without X in Stage 2. For macro F1, removing the tone features yielded the largest drop of 0.014, followed by the style feature with a drop of 0.011. Despite differences in the number and granularity of the feature groups, the results suggest that tone and style are among the most effective cues for conveying stance, likely because T2I models can render them straightforwardly and LVLMs can readily perceive them.



(a) Misleading case



(b) Beneficial case

Figure A1: Representative cases where images generated by VFSTANCE changed stance predictions.

D.4 Multilingual Evaluation

We conduct a preliminary experiment to evaluate the effectiveness of VFSTANCE on multilingual versions of K-News-Stance-MM in English, Chinese, Indonesian, and Arabic. As shown in Table A6, VFSTANCE achieved higher accuracy and macro F1 scores than the text-only setting across all four languages, suggesting that the proposed method is effective across languages. Performance was highest in English and Chinese, followed by Indonesian and Arabic; however, the original Korean version achieved the best overall performance, with an accuracy of 0.746 and a macro F1 score of 0.747. The LLM-translated versions may lose some stance cues from the original Korean text because of their implicit nature. Future studies could further evaluate the multilingual robustness of VFSTANCE by constructing new resources that capture language- and culture-specific expressions of stance.

D.5 Error Case Analysis

We manually analyzed cases where images generated by VFSTANCE changed the stance prediction, with representative examples shown in Figure A1. First, among 55 cases where the textual LVLM baseline was correct but VFSTANCE was wrong, 51 had gold-neutral labels, and most errors shifted toward supportive rather than oppositional. In the example in Figure A1a, the article neutrally reports on reintroducing conscripted police, but the generated image depicts officers patrolling a bright

Stage 1	Stage 2	Stage 3	ACC	mF1
Gemini-3-flash	Gemini-3.1-flash-image	Gemini-3-flash	**0.746 ± 0.002**	**0.747 ± 0.002**
Gemini-3-flash	Gemini-3.1-flash-image	InternVL3-14B-Instruct	0.619 ± 0.001	0.622 ± 0.001
Gemini-3-flash	Stable Diffusion 3.5 Large	Gemini-3-flash	0.708 ± 0.004	0.712 ± 0.005
InternVL3-14B-Instruct	Gemini-3.1-flash-image	Gemini-3-flash	0.703 ± 0.002	0.706 ± 0.002
InternVL3-14B-Instruct	Stable Diffusion 3.5 Large	InternVL3-14B-Instruct	0.616 ± 0.002	0.619 ± 0.002

Table A7: Stage-level model substitution results with open-weight models on K-News-Stance-MM. The first row shows the proprietary-model configuration used in the main experiments. Each subsequent row replaces the model at a single stage with an open-weight model, while the final row uses open-weight models across all stages.

street, and its positive tone shifted the prediction from neutral to supportive.

Second, among 57 cases where VFSTANCE was correct but VFSTANCE (TEXT) was wrong, 48 had gold-neutral labels, with errors again skewed toward supportive. Specification terms such as *bright* or *dark* may act as lexical shortcuts when provided directly as text: in Figure A1b, the specification containing *bright* led VFSTANCE (TEXT) to a supportive misclassification, whereas rendering the same specification as a photographic image yielded a correct prediction. This suggests that perceptual cues are less susceptible to such lexical shortcuts than their textual counterparts.

D.6 Applicability to Open-Weight Models

We conducted two supplementary experiments on K-News-Stance-MM to examine whether VFSTANCE remains effective when implemented with open-weight models. Since VFSTANCE is a modular framework, each stage can be instantiated with alternative models beyond the proprietary ones used in the main experiments. We used two open-weight LVLMs, InternVL3-14B-Instruct and Gemma3-12B-Instruct, and an open-weight T2I model, Stable Diffusion 3.5 Large. Detailed experimental settings are provided in Appendix A.

Open-Weight Backbone Comparison To verify that the performance gains of VFSTANCE are not limited to the tested proprietary backbones, we compared VFSTANCE with the corresponding textual, visual, and multimodal baselines, using each open-weight LVLM as the Stage 3 detector. Table A8 shows that VFSTANCE outperformed all corresponding baselines under both backbones, indicating that its effectiveness extends beyond proprietary models. However, the absolute performance remained below the proprietary configuration, which may be partly attributable to the smaller sizes of the open-weight models available under our computational constraints.

Method	ACC	mF1
VFSTANCE		
Gemini-3-flash	**0.746 ± 0.002**	**0.747 ± 0.002**
InternVL3-14B-Instruct	0.619 ± 0.001	0.622 ± 0.001
Gemma3-12B-Instruct	0.622 ± 0.001	0.604 ± 0.001
Multimodal		
InternVL3-14B-Instruct	0.605 ± 0.001	0.608 ± 0.001
Gemma3-12B-Instruct	0.594 ± 0.001	0.586 ± 0.001
Textual		
InternVL3-14B-Instruct	0.592 ± 0.001	0.588 ± 0.001
Gemma3-12B-Instruct	0.586 ± 0.002	0.565 ± 0.002
Visual		
Gemma3-12B-Instruct	0.355 ± 0.001	0.298 ± 0.001
InternVL3-14B-Instruct	0.354 ± 0.002	0.282 ± 0.002

Table A8: Stance detection performance on K-News-Stance-MM with open-weight LVLMs for Stage 3, compared against the corresponding baselines under each backbone. The first row shows the proprietary-model configuration used in the main experiments.

Stage-Level Ablation To identify which stage is most sensitive to model substitution, we replaced the model at each stage of VFSTANCE with its open-weight counterpart, as listed in Table A7. Substitution reduced performance at all three stages, with the largest drop observed when the Stage 3 LVLM was replaced (0.127 in accuracy and 0.125 in macro F1). The fully open-weight variant performed comparably to the Stage 3-only substitution, suggesting that Stage 3 is the primary bottleneck: effective stance detection depends particularly on the LVLM's language understanding, visual perception, and multimodal reasoning.

D.7 Generated Image Examples

Table A10 compares three image sources—original publisher photographs, naïvely generated images (*Direct T2I* in Table 2), and VFSTANCE-generated

images—for three Korean articles with different stances toward the same target issue.

E User Study Details

This section provides additional details on the user study described in Section 6. Participants were balanced by gender and age: 100 female and 100 male, with 20 of each gender in each of five age groups (20–29, 30–39, 40–49, 50–59, and 60+). English translations of the nine articles are shown in Table A11, the condition images are compared in Table A12, and the instructions and survey interface are shown in Figures A10a and A10b. The original Korean articles are available in our GitHub repository.

Prompt – Textual Stance Detection

Based on the given article text, analyze the article text's stance toward the target issue.

The response should be in the form of a single word: `'supportive'`, `'neutral'`, or `'oppositional'`.

Target Issue: *{issue}*
News headline: *{headline}*
News article: *{article}*

Figure A2: Textual prompt used for LVLM-based stance detection. Blue italic text highlights the input.

Prompt – Visual Stance Detection

Based on the given image accompanying the article, infer the article text's stance toward the target issue.

The response should be in the form of a single word: `'supportive'`, `'neutral'`, or `'oppositional'`.

Target Issue: *{issue}*

Figure A3: Visual prompt used for LVLM-based stance detection. Blue italic text highlights the input.

Prompt – Multimodal Stance Detection (Stage 3)

Based on the given article text and an image accompanying this article, analyze the article text's stance toward the target issue.

The response should be in the form of a single word: `'supportive'`, `'neutral'`, or `'oppositional'`.

Target Issue: *{issue}*
News headline: *{headline}*
News article: *{article}*

Figure A4: Multimodal prompt used for LVLM-based stance detection. Blue italic text highlights the input.

Prompt – Stance-aware Image Generation (Stage 2)

An *illustration*-style image.
Include the following: *silhouettes of empty provincial cities, a huge Seoul-centered black hole, a ballot box with election-campaign slogans, a chaotically mixed map of Gyeonggi administrative districts, place names Gimpo, Guri, Hanam, Goyang, Bucheon, and Gwangmyeong*.
Composition is *asymmetric*, *high* angle and *long* shot.
Saturation is *desaturated*, luminosity is *dark*.
Exclude the following: *Rep. Cho Kyung-tae, Lee Jae-myung, brightly smiling citizens, a glamorous Seoul nightscape*.
Generate only one image at a time.

Figure A5: The English-translated text-to-image prompt used for the T2I model in Stage 2, shown with an illustrative input. Blue italic text highlights placeholders filled using the LLM-based annotations from Stage 1; the remaining text constitutes the fixed template.



Figure A6: Image generated by the T2I model in Stage 2 from the prompt in Figure A5.

Target Issue	Democratic Party-Led Revision of the Grain Management Act Put Directly to Plenary Session for Vote
Headline	Lee's 'Bill No. 1' Grain Act likely to become Yoon's 'Veto No. 1'
Article	The revision of the Grain Management Act, which mandates the government to purchase excess rice production, passed the plenary session of the National Assembly on the 23rd, led by the Democratic Party of Korea and the Justice Party. President Yoon Suk-yeol is reportedly reviewing the option of exercising his veto power over the revision. As a result, the Grain Management Act, which Democratic Party leader Lee Jae-myung had designated as his 'Bill No. 1,' is increasingly likely to become the subject of President Yoon's first-ever veto. If President Yoon does exercise his veto, it would be the first such action in approximately seven years, since former President Park Geun-hye vetoed the National Assembly Act revision centered on 'standing hearings' in May 2016. The revision was approved at the plenary session with 169 votes in favor, 90 against, and 7 abstentions out of 266 members present. The core provision requires the government to purchase the entire excess production when rice output exceeds demand by 3–5% or when rice prices fall by 5–8% compared to the previous year. The Democratic Party pushed the Grain Management Act forward, emphasizing rice price stabilization, farmer protection, and food sovereignty. However, the government and the People Power Party opposed it on grounds of rice oversupply, fiscal burden, and weakened agricultural competitiveness, but were unable to overcome their numerical disadvantage. The government expressed deep regret over the passage of the revision. Minister of Agriculture, Food and Rural Affairs Chung Hwang-keun stated at an emergency briefing at the Government Complex Seoul that "the side effects are all too obvious" and declared the revision unacceptable. With the opposition's forced passage of the revision through the National Assembly, concerns about the bill's side effects are growing. Following the passage, the government is expected to purchase an additional average of approximately 200,000 tons of rice per year, with the estimated cost reaching around 1 trillion won. As demand for rice declines due to changing dietary habits, critics argue that the mandatorily purchased rice will pile up in government warehouses and ultimately be sold at a loss for purposes such as producing makgeolli. There are also significant concerns that the expansion of mandatory government purchases could lead to increased rice production and a chronic decline in rice prices. President Yoon is reportedly leaning toward exercising his veto over the revision. A senior presidential office official stated, "Once the revision is sent to the government next week, there will be a period of public consultation and deliberation."
Image	https://imgnews.pstatic.net/image/005/2023/03/24/2023032321490368287_1679575743_0924293423_20230324041205219.jpg?type=w860
Stance	Oppositional

Table A9: English-translated example article from K-News-Stance-MM. Publisher-provided original images are omitted to respect copyright; their URLs are provided instead.

Issue: Court Recognizes Same-Sex Partners as Legal Dependents for Health Insurance
이슈: 법원 동성 부부 배우자도 건강보험 피부양자 인정

Stance	Original	Naïve	VFSTANCE
Headline: Court Recognizes Same-Sex Partner's Health Insurance... A Step Forward for Minority Rights **제목:** 법원, 동성반려자 건보 인정... 소수자 인권 진일보			
Sup.	[See: https://imgnews.pstatic.net/image/469/2023/02/22/0000724756_001_20230222061125823.jpg?type=w860]		
Headline: First Recognition of Same-Sex Partner as Health Insurance Dependent... Court Rules "No Discrimination Based on Sexual Orientation" **제목:** 동성커플 건보 피부양자 첫 인정... 법원 "성적지향 이유로 차별 안돼"			
Neu.	[See: https://imgnews.pstatic.net/image/020/2023/02/22/0003481184_001_20230222031302451.jpg?type=w860]		
Headline: Ruling Recognizing Same-Sex Couple's Insurance Eligibility — Supreme Court Should Rectify **제목:** '동성커플 건보 자격' 인정 판결, 대법원이 바로잡으라			
Opp.	[See: https://imgnews.pstatic.net/image/005/2023/02/22/2023022118390482487_1676972344_0924288416_20230222040306049.jpg?type=w860]		

Table A10: Examples from three image sources: publisher-provided original images, images generated with a straightforward prompt, and VFSTANCE-generated images for three articles with different stances on the same issue. Original images are omitted to respect copyright; their URLs are provided instead.

Prompt – Visual Framing Annotation (Stage 1)

<role>You are a specialized assistant for generating image prompts based on Visual Framing.</role>
<instructions>Design an image that visually communicates the stance a given news article takes toward a specified target issue. The stance will be one of the following: supportive, neutral, oppositional. To effectively convey the stance, annotate the Visual Framing spec according to the instructions below. The image as a whole should construct a single coherent interpretation that aligns with and effectively communicates the article's interpretation toward the issue.
Visual Framing Levels
1. Ideological level: Analyze the core axis of conflict surrounding the issue, then determine — according to the article's stance — whose perspective and voice the image will support, whose perspective and voice it will marginalize or silence, or whether the image will instead maintain distance from the conflict axis and present the issue from a neutral/balanced viewpoint. State this in one sentence.
(whose interests does the image serve? whose voices are silenced?)
2. Connotative level: Specify the cultural or symbolic associations the image should evoke, along with the concrete visual elements that will evoke them, using the format "A → B".
- A: the concept, value, emotion, or idea to be conveyed.
- B: a concrete, visually representable object that would evoke A within the socio-cultural context of the country in which the article was written.
3. Stylistic-Semiotic level: Determine the stylistic conventions and technical transformations involved in conveying the stance.
- style: [photo, illustration] signals factual versus interpretive presentation.
- composition: [centered, split, asymmetric, crowded] structures the spatial arrangement of visual elements.
- angle: [low, eye-level, high] conveys power relations through viewpoint.
- distance: [close-up, medium, long] controls social distance from the subject.
- saturation: [saturated, neutral, desaturated] sets image tone through color intensity.
- luminosity: [bright, neutral, dark] sets image tone.
4. Denotative level: At the surface level of the image, decide what should be intentionally included and what should be intentionally excluded. For real individuals or groups, include the exact names mentioned in the article.
- include_subjects: specific subjects intentionally included.
- exclude_subjects: specific subjects intentionally excluded. </instructions>
<constraints>
- All outputs must be written in Korean.
- Every B in the Connotative level must also appear in Denotative subjects.
- Keep the decision space of each level separate.
</constraints>
<output_format>Respond only with a JSON object. No preface or explanation.
{
"ideological": "",
"connotative": ["A → B", ...],
"stylistic_semiotic": {
"style": "photo | illustration", "composition": "centered | split | asymmetric | crowded", "angle": "low | eye-level | high", "distance": "close-up | medium | long", "saturation": "saturated | neutral | desaturated", "luminosity": "bright | neutral | dark"
},
"denotative": {"include_subjects": ["", ...], "exclude_subjects": ["", ...]}
} </output_format>
<context>
- Target Issue: *People Power Party Launches Special Committee for Gimpo-Seoul Integration*
- News Headline: *All of them to Seoul? "Five Gyeonggi cities besides Gimpo also considered for incorporation"*
- News Article (excerpt): *Cho Kyung-tae: "Guri, Hanam, Goyang, Bucheon, and Gwangmyeong under review . . . at least five, and hopefully more, should join." The People Power Party's "Mega Seoul" plan, including Gimpo's incorporation into Seoul, is reportedly also considering five other cities in Gyeonggi Province, raising concerns over confusion across the metropolitan area. The People Power Party officially declared that it would pursue Seoul incorporation not only for Gimpo but also for cities across the Seoul living sphere. Rep. Cho Kyung-tae, chair of the party's Special Committee for Improving Resident Convenience in the Metropolitan Area (tentative name), told Channel A on the 2nd, "Beyond Gimpo, I hope at least five places, including Guri, Hanam, Goyang, Bucheon, and Gwangmyeong, will join Seoul." . . .*
</context>

Figure A7: English-translated system and user prompts used in Stage 1 for visual framing annotation in VFSTANCE, shown with an illustrative input. Blue italic text highlights the input.

Stage 1 Output – Visual Framing Specifications (Korean)


```
{
  "ideological": "수도권 집중화와 선거용 선심성 정책이 초래할 국가적 불균형과 혼란을 우려하는
  관점을 옹호하며, 무분별한 서울 확장의 부정적 파급력을 부각한다.",
  "connotative": [
    "수도권 집중으로 인한 지역 소멸의 위기감 → 텅 빈 지방 도시들의 실루엣",
    "계획 없는 무분별한 서울 확장 시도 → 거대한 서울 중심의 블랙홀",
    "선거 승리만을 목적으로 한 급조된 정책 → 선거용 홍보 문구가 적힌 투표함",
    "경기도 주요 도시들의 행정적 혼란 → 어지럽게 뒤섞인 경기도 행정 구역 지도"
  ],
  "stylistic_semiotic": {
    "style": "illustration", "composition": "asymmetric", "angle": "high", "distance": "long",
    "saturation": "desaturated", "luminosity": "dark"
  },
  "denotative": {
    "include_subjects": [
      "텅 빈 지방 도시들의 실루엣", "거대한 서울 중심의 블랙홀", "선거용 홍보 문구가 적힌 투표함",
      "어지럽게 뒤섞인 경기도 행정 구역 지도", "김포, 구리, 하남, 고양, 부천, 광명 지명"
    ],
    "exclude_subjects": [
      "조경태 의원", "이재명 대표", "밝게 웃는 시민들", "화려한 서울의 야경"
    ]
  }
}
```


Figure A8: Original Korean-language output of Stage 1 visual framing annotation in VFSTANCE.

Stage 1 Output – Visual Framing Specifications (English)


```
{
  "ideological": "Advocates a perspective concerned about the national imbalance and confusion
  that may result from metropolitan concentration and election-oriented populist policies,
  while highlighting the negative ripple effects of indiscriminate Seoul expansion.",
  "connotative": [
    "A sense of crisis over regional extinction caused by metropolitan concentration →
    silhouettes of empty provincial cities",
    "An unplanned attempt at indiscriminate Seoul expansion → a huge Seoul-centered black hole",
    "A hastily assembled policy aimed only at electoral victory → a ballot box with
    election-campaign slogans",
    "Administrative confusion among major Gyeonggi cities → a chaotically mixed map of Gyeonggi
    administrative districts"
  ],
  "stylistic_semiotic": {
    "style": "illustration", "composition": "asymmetric", "angle": "high", "distance": "long",
    "saturation": "desaturated", "luminosity": "dark"
  },
  "denotative": {
    "include_subjects": [
      "silhouettes of empty provincial cities", "a huge Seoul-centered black hole", "a ballot
      box with election-campaign slogans", "a chaotically mixed map of Gyeonggi administrative
      districts", "place names Gimpo, Guri, Hanam, Goyang, Bucheon, and Gwangmyeong"
    ],
    "exclude_subjects": [
      "Rep. Cho Kyung-tae", "Lee Jae-myung", "brightly smiling citizens", "a glamorous Seoul
      nightscape"
    ]
  }
}
```


Figure A9: English-translated output of Stage 1 visual framing annotation in VFSTANCE.

ID	Stance	Headline & Lead
Issue 1. Military to Resume Full-Scale Propaganda Broadcasts Following Repeated Trash Balloon Incidents		
A1	Sup.	**JCS to "fully implement propaganda broadcasts to the North"… countering N. Korean trash balloons** In response to North Korea's trash balloon launches, the military authorities have played the card of fully implementing propaganda broadcasts to the North. A hardline tit-for-tat standoff between the two Koreas through psychological warfare appears to be deepening.
A2	Neu.	**N. Korea launches 9th round of trash balloons… Military counters with "full-scale propaganda broadcasts to the North"** After North Korea once again released trash balloons toward the South on the morning of the 21st, the military authorities responded with the "full-scale implementation of propaganda broadcasts to the North," and military tensions between the two Koreas are escalating.
A3	Opp.	**How will the North respond to the full expansion of propaganda broadcasts?… Border tensions mount** Using domestic civic groups' anti-North leaflet drops as a pretext, North Korea continues to launch trash balloons, and our military authorities have repeatedly countered with propaganda broadcasts to the North.
Issue 2. Impeachment Motion Against the BAI Chairman Faces Vote Tomorrow		
B1	Sup.	**BAI, unable to question Kim Keon-hee about "21Gram," protests: "We can't be expected to torture it out of them"** The Board of Audit and Inspection (BAI), after failing to identify who recommended "21Gram"—the contractor awarded the presidential residence renovation in Hannam-dong, Seoul—protested that "it isn't something we can uncover by torturing people."
B2	Neu.	**Impeachment motions against the BAI Chairman and prosecutors reported to the plenary session… Budget bill put on hold** Speaker Woo urged the ruling and opposition parties to reach a budget agreement by the 10th. The impeachment motions against the BAI Chairman and prosecutors will be voted on on the 4th, creating a vacuum in the chain of command. The year-end political situation is becoming increasingly unpredictable.
B3	Opp.	**BAI: "There's no Plan B at this stage… We trust the impeachment will be withdrawn"** As the impeachment motion against BAI Chairman Choe Jae-hae, filed by the Democratic Party of Korea, was reported to the National Assembly's plenary session on the 2nd, the Board of Audit and Inspection stated that it "is not considering a Plan B at this stage" in preparation for the aftermath of the impeachment.
Issue 3. President Yoon Nominates Vice Justice Minister Shim Woo-jung for Prosecutor General		
C1	Sup.	**Prosecutor General nominee Shim Woo-jung, a "planning specialist": "I will do my utmost to earn the public's trust"** On the 11th, President Yoon Suk-yeol nominated Vice Justice Minister Shim Woo-jung (53, Judicial Research and Training Institute Class 26) as the next Prosecutor General candidate.
C2	Neu.	**Prosecutor General nominee on the "Kim Keon-hee handbag" allegations: "Law and principle are what matter"** Prosecutor General nominee Shim Woo-jung took a principled stance on the "luxury handbag" allegations involving First Lady Kim Keon-hee, former head of Covana Contents, stating that "it is important to uphold the law and principle."
C3	Opp.	**Emphasis on "tighter control of the prosecution"… Yongsan's "safe choice"** On the 11th, President Yoon Suk-yeol nominated Vice Justice Minister Shim Woo-jung as the next head of the prosecution. Commentators view this as a "safe choice" by Yoon—who has been unable to shake off various legal risks—made with an eye to the relationship between the presidential office in Yongsan and the prosecution, and to the stability of the prosecutorial organization.

Table A11: English translation of articles used in the user study. Each cell shows the headline (bold) and lead of the corresponding article. The original Korean articles are available in our GitHub repository.

ID	Stance	Original	Naïve	Proposed
A1	Sup.	[See: https://imgnews.pstatic.net/image/022/2024/07/21/20240721505503_20240721142308421.jpg?type=w860]		
A2	Neu.	[See: https://imgnews.pstatic.net/image/009/2024/07/21/0005337843_001_20240721130306464.jpg?type=w860]		
A3	Opp.	[See: https://imgnews.pstatic.net/image/079/2024/07/21/0003918478_001_20240721161010430.jpg?type=w860]		
B1	Sup.	[See: https://imgnews.pstatic.net/image/028/2024/12/02/0002719024_001_20241203153828650.jpg?type=w860]		
B2	Neu.	[See: https://imgnews.pstatic.net/image/656/2024/12/02/0000113206_001_20241202185211320.jpg?type=w860]		
B3	Opp.	[See: https://imgnews.pstatic.net/image/079/2024/12/02/0003965096_001_20241202160219972.jpg?type=w860]		
C1	Sup.	[See: https://imgnews.pstatic.net/image/020/2024/08/11/0003581252_001_20240811204909147.jpg?type=w860]		
C2	Neu.	[See: https://imgnews.pstatic.net/image/002/2024/08/11/0002345370_001_20240811180508603.jpg?type=w860]		
C3	Opp.	[See: https://imgnews.pstatic.net/image/032/2024/08/11/0003314304_001_20250514095121338.jpg?type=w860]		

Table A12: Images for the nine articles used in the user study (Table A11). Original images are omitted to respect copyright; their URLs are provided instead.



(a) Instructions (b) User interface

Figure A10: Instructions and user interface used in the user study.