Exploring Artificial Image Generation for Stance Detection

Anonymous ACL submission

Abstract

Stance detection is a task aimed at identifying and analyzing the author's stance from textual data. Previous studies primarily rely on analyzing the text itself, which may not fully capture the implicit stance conveyed by the author. To address this limitation, we propose a novel approach that involves transforming the original text into an *artificially generated image* and using this visual representation to aid in stance detection. Our approach begins by employing a large vision-language model to generate potential images for a given text. These images are carefully crafted to adhere to three specific criteria: relevance to the text, consistency with the target of the stance, and consistency with the stance itself. Next, we introduce a comprehensive evaluation framework to select the optimal image from the set of generated candidates. Once the optimal image has been selected, we introduce a multimodal stance detection model that leverages both the original textual content and the generated image to identify the author's stance. The experimental results demonstrate the effectiveness of our proposed approach, and also indicates the importance of artificially generated images for stance detection.

1 Introduction

Stance detection is a pivotal task in the field of natural language processing, aimed at identifying and analyzing the author's opinion or stance from textual data. Within this complex linguistic environment, the challenge of automatically and accurately categorizing these stances has emerged as a significant research question.

With the advancements in deep learning technologies, stance detection methods have made significant progress. These models learn rich linguistic representations through pre-training on large text datasets (Stodden et al., 2023; Arakelyan et al., 2023; Saha et al., 2024). Besides single-task stance detection models, multi-task learning

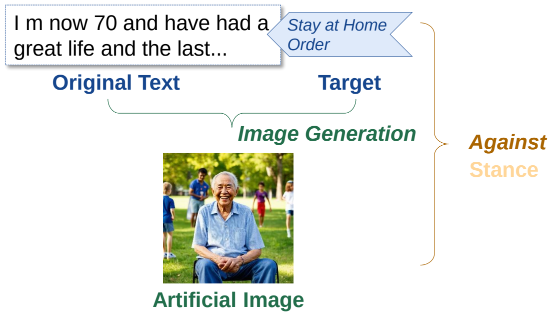


Figure 1: An example of artificial image generation for stance detection.

frameworks have also been proposed to enhance model performance (Chai et al., 2022). More recently, Gatto et al. (2023) introduced chain-of-thought embedding, which embeds reasoning text into stance detection process, enhancing the model's ability to identify implicit stances.

Although these deep learning models have shown strong capabilities in stance detection, they still mainly rely on single text modality, which may not fully capture the implicit stance. When humans understand and interpret the world, they often rely on the integrated information of multiple sensory inputs, such as vision, hearing, language, etc. As a result, methods that incorporate multimodal information are able to capture complex patterns in the data more comprehensively (Baltrušaitis et al., 2018; Hu et al., 2022; Liang et al., 2024).

However, these multimodal stance detection methods depend on user-generated images, which may not consistently capture the crucial expressions of stance or the intended target. This issue arises when the images include irrelevant components, like celebrities or objects that are not directly pertinent to the topic at hand. Furthermore, not all posts include user-generated images, which limits the applicability of these multimodal approaches in certain cases.

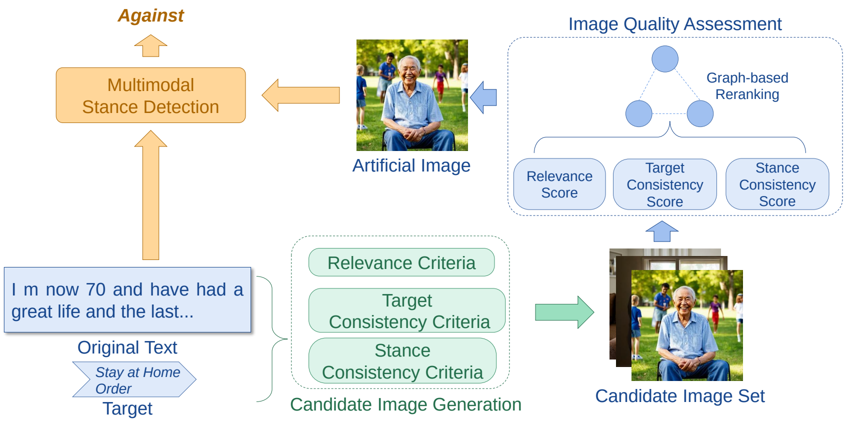


Figure 2: Overview of proposed model.

To address the above limitations, we propose transforming the original text into an *artificially generated image* and utilizing this visual representation for stance detection. As shown in Figure 1, a skillfully designed artificially generated image effectively communicates the same stance as the original text. As a result, comprehending the stance becomes significantly more straightforward when presented in an artificially generated image format compared to the text alone. Nevertheless, creating such an image poses a challenge. It is imperative that the image encapsulates both the stance and target information from the original text, while simultaneously being easily comprehensible.

In our study, we initially employ a large vision-language model to generate potential images for a given original text. The generation process adheres to three specific criteria: relevance, target consistency, and stance consistency with the original text. By following these criteria, the generated candidate images are designed to contain the same stance as the original text while also being comprehensive and easy to understand. To select the optimal artificially generated image from the candidate set, we introduce a comprehensive evaluation framework. This framework incorporates multiple scoring metrics and a sophisticated graph-based method. The scoring metrics assess the quality and accuracy of the candidate images, while the graph-based method considers their semantic relationships. Through this meticulous evaluation process, we identify and choose the most optimal image that best retain the core meaning of the original text.

Furthermore, we introduce a multimodal stance detection model that leverages both the original textual content and the accompanying generated image to identify stance. The experiment results indicate that our proposed approach significantly enhances performance. This results also underscores the advantageous role played by the generated image in improving stance detection accuracy.

2 Related Works

With the rapid advancement of deep learning technologies, stance detection has witnessed substantial improvements. These sophisticated models acquire extensive linguistic knowledge by pre-training on vast text datasets (Stodden et al., 2023; Arakelyan et al., 2023; Saha et al., 2024).

Recent research has focused on addressing these limitations. For instance, Stodden et al. (2023) employed a masked language model to predict the likelihood of conjunctions within the text by temporarily removing (masking) them and then estimating their probabilities. Furthermore, Saha et al. (2024) combined stance detection with explanation generation by constructing argumentation dependency trees.

In addition to textual analysis, multimodal stance detection has also gained momentum. Liang et al. (2024) creates five new multimodal stance detection datasets of different domains based on Twitter, in which each example consists of a text and an image. Unlike datasets that focus on individual text-image pairs, Niu et al. (2024) introduced a new multimodal multi-turn conversational stance detection dataset that captures the natural multi-party conversational context occurring on social media. In terms of methodology, Liang et al. (2024) designed specific prompts tailored to the target and input them into pretrained language and vision models. This allowed them to extract features that contain target-specific stance information, further enhancing the performance of stance detection.

Our proposed method differs from previous studies that rely on user-generated images. Instead, we utilize a text-to-image model to generate artificially generated images related to the text. These generated images, along with the original text, are then fed into a large vision-language model for stance detection. By doing so, we aim to fully exploit the complementary information provided by both text and artificially generated images, thereby improving the accuracy and robustness of stance detection.

3 Proposed Model

As shown in Figure 2, we present a novel method for stance detection, designed to generate an *artificial image* from the original text. Our approach is structured into three key stages: In the initial stage, we generate a diverse array of candidate images. We achieve this by employing a range of criteria, ensuring that the resulting images maintaining consistency with the stance and target expressed in the original text. We then introduce a comprehensive evaluation framework with a graph re-ranking algorithm to assess the quality and accuracy of the candidate images, and select the most optimal artificial image that best retains the core meaning of the original text. In the final stage, we propose multimodal stance detection model to detect the stance through the original text and generated image.

3.1 Candidate Image Generation

We commence by delineating the methodology for generating a candidate image set from original text for stance detection. Given an original text, we employ a robust vision-language model to produce potential images that conform to three explicit criteria. These criteria are enumerated as follows:

- **Relevance (C_R):** The generated candidate images must be pertinent to the content of the original text.
- **Target Consistency (C_T):** The generated candidate images should incorporate the identical target information as that found in the original text.
- **Stance Consistency (C_S):** The stance conveyed by the generated candidate images should align with those expressed in the original text.

By strictly adhering to these criteria, the resultant candidate images are ensured to encapsulate the same stance information as the original text, while also being comprehensive and easily comprehensible.

We further design a dynamic multi-output algorithm aimed at generating a set of candidate images that adhere to the above criteria. The central tenet of this algorithm is the dynamic generation of images at various stages throughout the process. This methodology enables the candidate images to diverge and encapsulate distinct aspects of the criteria. By employing this approach, our algorithm ensures that each candidate image is tailored to meet specific criteria, thereby fostering a diverse and comprehensive set of images. The algorithm is shown as Algorithm 1.

The algorithm enriches the content of images from multiple dimensions by integrating text, target, and predicted stance. However, due to the lack of necessary contextual information and background knowledge in tweets, it is challenging for text-to-image models to accurately grasp the topics discussed in the tweets. To address this issue, we employ ChatGPT to generate clearer and more comprehensible image descriptions, thereby enhancing the understandability of the image content. Specifically, we first generate the stance and sentiment of the tweet using a finetuned LLM. Then, based on the criteria of C_R, C_T, and C_S, we design a prompt into which we inject the original text along with target, stance, and sentiment information. We further input this into ChatGPT to obtain suitable image descriptions. We can design different prompts based on different criteria to obtain images that contain different information. The implement detail of candidate images generation can be found in Appendix A.

3.2 Image Quality Assessment

After generating a set of candidate images, we must ascertain which one is the most optimal. Based on the specific criteria delineated in the preceding subsection, we propose a suite of scoring

Algorithm 1 Dynamic multi-output algorithm for candidate image generation

Input: Original text T, Target G
Output: Set of candidate images S
Predict stance S and sentiment M from T by a finetuned LLM
Define combinations from three criteria:
    C_R: Input T to generate candidate images related to the original text content
    C_T: Input T and M to generate candidate images containing the target information
    C_S: Input T, G, S or T, G, S, M to generate candidate images conveying the stance information
For each combination C:
    Input C to a text-to-image model to generate images S_1
    Input C to ChatGPT to generate corresponding image descriptions
    Input the generated image descriptions to the text-to-image model to generate images S_2
    $S \leftarrow S + S_1 + S_2$
Return: Candidate images S

metrics to evaluate these candidate images and ultimately select the most suitable one. These scoring metrics are designed to consider multiple factors, including clarity, conciseness, accuracy, and fidelity to the original meaning.

Relevance Score S_R: Initially, we leverage a large vision-language models to gauge the relevance between the candidate image and the original text. In particular, we employ CLIP (Radford et al., 2021), a multimodal model that learns the association between images and text through contrastive learning. We encode both the candidate image and the text into vector representations using CLIP, and then calculate the cosine similarity between these two vectors,

$$S_R = \frac{\mathbf{I} \cdot \mathbf{T}}{\|\mathbf{I}\|_2 \|\mathbf{T}\|_2} \tag{1}$$

where $\mathbf{I}$ and $\mathbf{T}$ represent the vector representations of the image and text, respectively.

Target Consistency Score S_T: is a metric designed to evaluate how well a candidate image aligns with the target information described in the original text. This score is determined through a two-step process involving a vision-language model.

In the first step, the candidate image and the original text are input into the vision-language model to determine if the image accurately reflects the target information in the text. This results in a binary response, R_{T_I}, where "yes" indicates consistency and "no" indicates inconsistency.

In the second step, the candidate image is again input into the vision-language model to generate a caption for the image. This caption and the original text are then input into ChatGPT, which assesses whether the caption is relevant to the target information in the text. This also results in a binary response, R_{T_C}, with "yes" indicating relevance and "no" indicating irrelevance.

The final Target Consistency Score S_T is calculated as a weighted sum of R_{T_I} and R_{T_C}, with the weight determined by a trainable parameter α. The formula for S_T is:

$$S_T = \alpha R_{T_I} + (1 - \alpha) R_{T_C} \tag{2}$$

Stance Consistency Score S_S: Furthermore, we employ a vision-language model to evaluate the stance consistency of the candidate images. Initially, both the candidate image and the original text are input into the vision-language model to determine if the image accurately reflects the stance. The vision-language model then provides a binary response, denoted as R_{S_I} (yes/no), which serves as an indicator of the candidate image's appropriateness in terms of its stance alignment with the original text.

Subsequently, we adopt a similar approach to that used in calculating S_T to obtain captions for the image. These captions, along with the original text, are then input into ChatGPT to assess whether they reflect the stance expressed in the original text. This process yields another binary response, denoted as R_{S_C} (yes/no).

The final Stance Consistency Score S_S is computed as a weighted sum of R_{S_I} and R_{S_C}, with the weight determined by a trainable parameter α. The formula for S_S is as follows:

$$S_S = \beta R_{S_I} + (1 - \beta) R_{S_C} \tag{3}$$

Algorithm 2 Graph-based candidate images re-ranking

Input: Graph $G = (V, E)$, damping factor $\mu = 0.85$

Output: The image node $v^* \in V$ with the highest value.

$PR(v) \leftarrow \frac{S_R(v)+S_T(v)+S_S(v)}{3}, \forall v \in V$

repeat

$PR(v) \leftarrow (1-\mu) + \mu \sum_{u \in In(v)} \frac{PR(u)}{out(u)}$ $\forall$v $\in V$, where $In(v)$ is the set of incoming nodes and $out(u)$ is the number of outgoing edges from node u

until

$\max_{v \in V} |PR_{new}(v) - PR_{old}(v)| < 10^{-6}$

$v^* \leftarrow \arg\max_{v \in V} PR(v)$

return v^*

3.3 Graph-based Images Re-ranking

After evaluating the candidate images using the above three types of point-wise metrics, we propose a graph-based approach (Page et al., 1999) to collectively re-rank all the candidates and choose the most optimal artificial image . The algorithm is shown as Algorithm 2.

In this method, each node in the graph represents a candidate image, and its value is determined by the sum of the three point-wise metrics for that image. The weight of the edge between two nodes corresponds to the semantic similarity between the two candidate images. To identify the most optimal image, we employ a graph-based algorithm that involves a random walk across the entire graph. The goal of this random walk is to locate the node with the highest score, which is then selected as the most optimal artificial image. This approach takes into account not only the individual scores of the images but also their semantic relationships with other images in the set, providing a more comprehensive evaluation of the candidates. The implement detail of image quality assessment and graph-based images re-ranking can be found in Appendix B.

3.4 Multimodal Stance Detection

After obtaining the appropriate image, we propose a multimodal stance detection model to integrate the original text and the generated image jointly. Specifically, we design prompt P based on the content of the text and the generated image:

P = <ImageHere> Given the **Tweet** and the **Target** picture, what is the stance towards Target.

Here, <ImageHere> represents the placeholder for the image.

We then employ a Large Language Model as a text encoder to obtain textual hidden representations H_t.

$$H_t = Encode(P) \tag{4}$$

We utilize the Vision Transformer (ViT) to obtain visual hidden representations H_v from the image I.

$$H_v = ViT(I) \tag{5}$$

We concatenate them to form our multimodal hidden representations H.

$$H = H_v \oplus H_t \tag{6}$$

Given the fused hidden representations $H = h_1, \ldots, h_{|h|}$ as input, the model outputs the linearized representation $y = y_1, \ldots, y_{|y|}$. At the i-th step of generation, the decoder predicts the i-th token y_i in linearized form, and the decoder state h_i^d is as follows,

$$y_i, h_i^d = Decoder\left(\left[h_1^d, \ldots, h_{i-1}^d\right], y_{i-1}\right) \tag{7}$$

The overall conditional probability $p(y \mid x)$ is computed by multiplying the probabilities of each $p(y_i \mid y_{<i}, x)$:

$$p(y \mid x) = \prod_{i=1}^{|y|} p(y_i \mid y_{<i}, x) \tag{8}$$

3.5 Objective Function

The objective functions is to maximize the output stance S probability given the text X. Therefore, we optimize the negative log-likelihood loss function:

$$\mathcal{L} = -\frac{1}{|\tau|} \sum_{(X,S) \in \tau} \log p(S \mid X; \theta) \tag{9}$$

where θ is the model parameters, and (X, S) is a input-output pair in training set τ, then

$$\log p(T \mid X; \theta) = \sum_{i=1}^{m} \log p(t_i \mid t_1, t_2, \ldots, t_{i-1}, X; \theta) \tag{10}$$

where $p(t_i \mid t_1, t_2, \ldots, t_{i-1}, X; \theta)$ is calculated by the decoder.

Model	Semeval-16			Covid-19		
	F_{favor}	$F_{against}$	F_{avg}	F_{favor}	$F_{against}$	F_{avg}
BERT	0.640	0.757	0.698	0.729	0.676	0.703
RoBERTa	0.651	0.773	0.712	0.768	0.762	0.765
Flan-T5	0.666	0.768	0.717	0.793	0.744	0.769
LlaMA3	0.796	0.808	0.802	0.846	0.850	0.848
GPT-4o-mini	0.717	0.715	0.716	0.544	0.626	0.858
InternLM-TextOnly	0.772	0.811	0.792	0.865	0.821	0.843
MTIN	-	-	0.703	-	-	0.679
Stanceformer	0.653	0.776	0.715	0.779	0.769	0.774
PSDCOT	0.672	0.775	0.723	0.786	0.776	0.781
TR-Tweet-COT	0.701	0.787	0.744	0.844	0.804	0.824
Ours	**0.804**	**0.833**	**0.818**	**0.882**	**0.847**	**0.865**

Table 1: Comparison with different baselines.

	Semeval-16	**Covid-19**
Train	2,520	4,532
Dev	294	800
Test	1,249	800

Table 2: Statistics of dataset.

4 Experiments

In this section, we will introduce the dataset, experimental setup, evaluation metrics, and baselines. We also report the main findings in this section and conduct additional experiments to demonstrate the effectiveness of the method.

4.1 Data and Setting

We conducted experiments on two stance detection datasets: Semeval-16 (Mohammad et al., 2016) and Covid-19 (Glandt et al., 2021). Both datasets categorize stances into three classes: favor, against, and none. Since the original Semeval-16 dataset does not have a validation set, we follow the setup of Barbieri et al. (2020) to partition the validation set. The statistical information of the datasets is shown in Table 2.

We finetune InternLM-XComposer2-VL (Dong et al., 2024) using the LoRA technique as our base model. This model is also used for the calculation of S_T and S_S. The learning rate in the main experiment is set to 1e-4. We selected Stable-Diffusion-3 (Esser et al., 2024) as the text-to-image model. Our experiments are carried out with one NVIDIA GeForce RTX 4090 GPU.

Following Mohammad et al. (2016), we record F_{avg}, where F_{avg} is the macro average of the F1 scores for favor and against. We report results averaged over three runs.

4.2 Main Results

We initially benchmarked our method against several established baselines, including **BERT** (Devlin et al., 2019), **RoBERTa** (Loureiro et al., 2022), and **T5** (Chung et al., 2022), which have demonstrated strong performance across various NLP tasks. Subsequently, we evaluated our approach against Large Language Models (LLMs) such as **LlaMA3** (Dubey et al., 2024), **InternLM-TextOnly**, and **GPT-4o-mini**[1], which are known for their extensive capabilities and scalability. Finally, we contrasted our method with state-of-the-art stance detection techniques, specifically **MTIN** (Chai et al., 2022), which incorporates a multi-task interaction module to capture word-level interactions between tasks, **TR-Tweet-COT** (Gatto et al., 2023), which integrates Chain-of-Thought (COT) reasoning into a RoBERTa-based stance detection framework by introducing COT embeddings, **Stanceformer** (Garg and Caragea, 2024), which introduces a target-awareness matrix into the transformer architecture to enhance attention to targets, and **PSDCOT** (Ding et al., 2024), which acquires knowledge through the chain-of-thought method and integrates it using a multi-prompt learning network.

As shown in Table 1, these text-based stance

[1] https://openai.com/index/hello-gpt-4o/

Model	Semeval-16	Covid-19
Text Only	0.792	0.843
Text		
+ Image(Oringial)	0.800	0.852
+ Image(C_R)	0.802	0.853
+ Image(C_T)	0.805	0.855
+ Image(C_S)	0.811	0.855
Ours	**0.818**	**0.865**

Table 3: The influence of different candidate image generation methods. "Original" represents images generated solely from the original text.

Model	Semeval-16	Covid-19
Ours	**0.818**	**0.865**
-S_R	0.811	0.859
-S_T	0.808	0.851
-S_S	0.811	0.856
-GraphRanking	0.807	0.860

Table 4: The contribution of image quality assessment.

Model	Semeval-16	Covid-19
Text Only	0.792	0.843
Ours	0.818	0.865
Bert	0.698	0.703
Bert+ResNet	0.707	0.715
T5	0.717	0.769
T5+ViT	0.726	0.783

Table 5: Influence of different multimodal models.

detection models exhibited commendable performance in stance detection tasks. Notably, the InternLM-TextOnly model stand out due to its exceptional language understanding and representation capabilities. Furthermore, our proposed model significantly outperform all the baselines ($p < 0.05$), demonstrating the effectiveness of our approach. These results underscore the importance of integrating generated images into stance detection, highlighting the added value of multimodal information in enhancing the accuracy of stance detection tasks.

4.3 Impact of Candidate Image Generation Methods

We subsequently investigated the impact of various candidate image generation methods, as discussed in Section 3.1.

As shown in Table 3, the incorporation of artificially generated image, whether derived directly from the original text or processed through diverse image generation methods, markedly enhances the model's performance. This underscores the crucial role of generated image in boosting model accuracy. Notably, when stance consistency C_S is considered, the model outperforms all other criteria. Furthermore, our proposed model, which integrates all types of criteria, achieves the optimal performance. This suggests that effectively combining multiple candidate image generation methods is essential for obtaining the best results.

4.4 Impact of Image Assessment Strategies

We subsequently carried out a series of ablation experiments to delve deeper into the importance of image quality assessment strategies, where S_R, S_T, S_S have been discussed in Section 3.2, and *GraphRanking* has been discussed in Section 3.3.

As shown in Table 4, the results highlight the beneficial impact of these assessment strategies and the graph re-ranking algorithm in boosting the model's overall performance. If any one of these components is removed, the performance of the model decreases compared to the complete version.

4.5 Results of Different Multimodal Models

We conducted further experimental research to investigate the efficacy of artificially generated images with different multimodal models. The experiment are categorized into three distinct groups. In the first group, all models are based on the InternLM-XComposer2-VL architecture. The primary difference among them is the type of input data utilized: one subset of models receives only text data, while the other subset receives both text and image data (Ours). For the second group, the text-based model employs Bert, while the multimodal model combined Bert with ResNet (He et al., 2016). In this configuration, Bert is responsible for extracting text features, and ResNet was used to extract image features. In the third group, the text-based model utilizes T5, and the multimodal model is a combination of T5 and ViT (Dosovitskiy, 2020).

The experimental results, presented in Table 5, demonstrate that the models incorporating artificially generated images outperform the basic textual models across all the multimodal models. This not only highlights the effectiveness of the

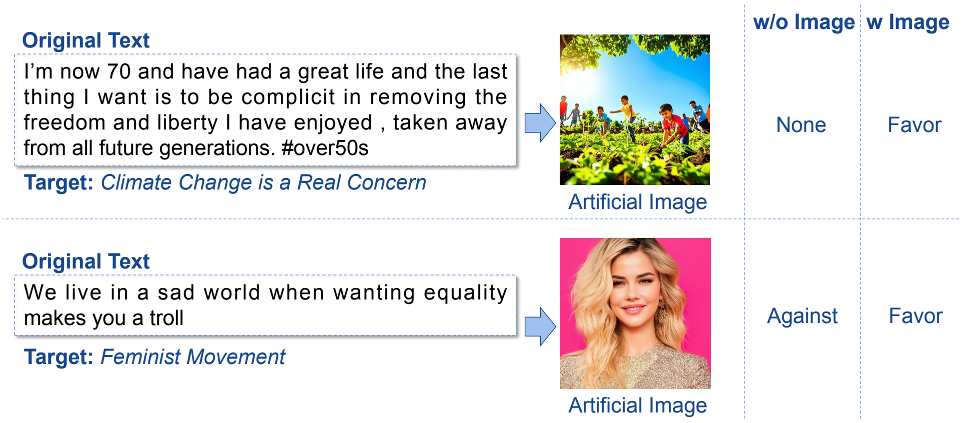


Figure 3: Examples of case study.

Model	ImageArg
Text	0.852
+Original Image	0.864
+Generated Image	0.870

Table 6: Comparison with original images.

artificially generated images but also proves their universality, as they can be successfully utilized in various multimodal models.

4.6 Comparison with Original Images

We further validate the efficacy of the artificially generated images against original images using the ImageArg dataset (Liu et al., 2022), a multimodal stance detection dataset that categorizes stances into two classes: Support and Oppose.

As illustrated in Table 6, the results indicate that the model's performance was enhanced with the incorporation of images. Notably, the use of artificially generated images outperformed the use of original images. This outcome suggests that the artificially generated images are able to more comprehensively represent or complement the textual content, thereby conveying the user's stance more accurately.

4.7 Case Study

To gain a more intuitive understanding of the benefits of generated artificial images in stance detection tasks, we conduct a case study in Figure 3.

In the first example, the inclusion of the generated artificial image visually strengthen the link between the original text and the target, making the stance more apparent and thus leading to a correct classification as *favor* when the image was included. In the second example, the words "sad" and "troll" in the original text might suggest a stance of *against*; however, an image depicting a confident woman provides additional context, resulting in a correct prediction. From the provided examples, it is evident that generated artificial images can resolve ambiguities and enrich the information available, thereby enhancing the model's classification effectiveness.

5 Conclusion

In this study, we propose a novel approach that involves transforming the original text into an artificially generated image and using this visual representation to aid in stance detection. Our approach begins by employing a large vision-language model to generate potential images for a given text. Next, we introduce a comprehensive evaluation framework to select the optimal artificial image from the set of generated candidates. Once the optimal artificial image has been selected, we introduce a multimodal stance detection model that leverages both the original textual content and the accompanying generated image to identify the author's stance. The experimental results demonstrate the effectiveness of our proposed approach, and also indicates the importance of generated images for stance detection.

Limitations

The proposed study, which involves transforming original text into artificially generated images to aid in stance detection, represents a novel and innovative approach. However, the limitation of this study is the computational resources required to generate and evaluate the artificial images. The process of generating potential images using a large vision-language model and then selecting the optimal image through a comprehensive evaluation framework can be computationally intensive.

References

Erik Arakelyan, Arnav Arora, and Isabelle Augenstein. 2023. Topic-guided sampling for data-efficient multi-domain stance detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13448–13464, Toronto, Canada. Association for Computational Linguistics.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Heyan Chai, Siyu Tang, Jinhao Cui, Ye Ding, Binxing Fang, and Qing Liao. 2022. Improving multi-task stance detection with multi-task interaction network. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2990–3000, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv e-prints*, pages arXiv–2210.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daijun Ding, Xianghua Fu, Xiaojiang Peng, Xiaomao Fan, Hu Huang, and Bowen Zhang. 2024. Leveraging chain-of-thought to enhance stance detection with prompt-tuning. *Mathematics*, 12(4).

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, et al. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*.

Krishna Garg and Cornelia Caragea. 2024. Stanceformer: Target-aware transformer for stance detection. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4969–4984, Miami, Florida, USA. Association for Computational Linguistics.

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.

Kyle Glandt, Sarthak Khanal, Yingjie Li, Doina Caragea, and Cornelia Caragea. 2021. Stance detection in COVID-19 tweets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1596–1611, Online. 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.

Mengyang Hu, Pengyuan Liu, Weikang Wang, Hu Zhang, and Chengxiao Lin. 2022. Msdd: A multimodal language dateset for stance detection. In *Workshop on Chinese Lexical Semantics*, pages 112–124. Springer.

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 and virtual meeting. Association for Computational Linguistics.

Zhexiong Liu, Meiqi Guo, Yue Dai, and Diane Litman. 2022. ImageArg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the*

9th Workshop on Argument Mining*, pages 1–18, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

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.

Fuqiang Niu, Zebang Cheng, Xianghua Fu, Xiaojiang Peng, Genan Dai, Yin Chen, Hu Huang, and Bowen Zhang. 2024. Multimodal multi-turn conversation stance detection: A challenge dataset and effective model. *arXiv preprint arXiv:2409.00597*.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rudra Ranajee Saha, Laks V. S. Lakshmanan, and Raymond T. Ng. 2024. Stance detection with explanations. *Computational Linguistics*, 50(1):193–235.

Regina Stodden, Laura Kallmeyer, Lea Kawaletz, and Heidrun Dorgeloh. 2023. Using masked language model probabilities of connectives for stance detection in English discourse. In *Proceedings of the 10th Workshop on Argument Mining*, pages 11–18, Singapore. Association for Computational Linguistics.

A Implement Detail of Candidate Image Generation

We demonstrate the implement detail of candidate image generation process with the following data:

> **Text:** I'm now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s
> **Target:** Stay at Home Orders

We first fine-tuned InternLM-XComposer2-VL to predict the stance and sentiment of each text. Then, based on the criteria C_R, C_T, and C_S, as well as the original text and the predicted stance and sentiment, we prepare five types of prompts, namely $prompt_1$-$prompt_5$, as shown in Table 7

In addition, we expanded these prompts using ChatGPT to obtain more detailed prompts namely $prompt_6$-$prompt_9$, as shown in Table 7. The expansion process is shown in Table 8. Subsequently, we input these prompts into Stable-Diffusion-3 and generate nine images.

B Implement Details of Image Quality Assessment

After obtaining the candidate images, we evaluate the quality of the generated images using a multimodal assessment framework. The evaluation system is based on C_R, C_T, and C_S. The VLM mentioned below is InternLM-XComposer2-VL.

Relevance Score S_R

We first extract the features of the original text and the generated images using the CLIP model, then compute their cosine similarity to obtain the score S_R.

Target Consistency Score S_T

Next, we generate a caption for each image using the VLM. The prompt input to the VLMis as follows:

> <ImageHere> Please provide a caption for the image that includes details about the scene, people, actions, expressions, and background. If there is any text in the image, please incorporate that into the caption as well.

The generated caption is as shown in the Table 9. Then, based on the target consistency criteria, we input the original text and the image into the VLM to ask whether the image is related to the target:

> <ImageHere> Given the tweet text, target target and the image, whether the image is related to the target, reply with 'yes' or 'no'.

We will receive a 'yes' or 'no' response R_{T_I}. Subsequently, we input the generated caption into ChatGPT to ask whether the caption is related to the target:

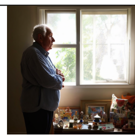
	Criteria	Prompt	Generated Image
$prompt_1$	C_R	I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s	
$prompt_2$	C_T	Stay at Home Orders: I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s	
$prompt_3$	C_S	Against Stay at Home Orders: I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s	
$prompt_4$	C_S	Negative and Against Stay at Home Orders: I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s	
$prompt_5$	C_S	Against Stay at Home Orders	
$prompt_6$	C_R	In the image, an elderly man is sitting on a chair in a park, smiling and looking directly at the camera. He is wearing a blue shirt and jeans. In the background, there are children playing, with one of them holding a frisbee. The scene appears to be a sunny day, as indicated by the bright lighting and shadows.	
$prompt_7$	C_T	An older man is sitting on a brown couch, wearing a gray sweater and green pants. He has his hands clasped together in front of him, looking deep in thought with a serious expression on his face. The room features white walls and a window with blinds, providing a neutral background for the scene.	
$prompt_8$	C_S	An older man with a blue shirt stands in front of a window, looking out at the view. He is standing next to a table filled with various items such as figurines and picture frames. The room has a warm atmosphere, and the sunlight streaming through the window illuminates the scene	
$prompt_9$	C_S	An older man is sitting in a chair, watching the news on a television that is placed on a wooden cabinet. The TV screen displays an advertisement for a phone order service with the text Šizz-a-fast phone orders - Lost freedom	

Table 7: The prompts and generated images in candidate image generation process.

prompt	output
Please expand the following Tweet into a detailed visual description. Describe the scene in terms of colors, objects, people, and other visual elements.Tweet: **I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s**	In the image, an elderly man is sitting on a chair in a park, smiling and looking directly at the camera. He is wearing a blue shirt and jeans. In the background, there are children playing, with one of them holding a frisbee. The scene appears to be a sunny day, as indicated by the bright lighting and shadows.
Please expand the following Tweet related to Target **Stay at Home Orders** into a detailed visual description. Describe the scene in terms of colors, objects, people, and other visual elements. The generated image description should maintain the same stance towards the Target as the original Tweet.Tweet: **I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s**	An older man is sitting on a brown couch, wearing a gray sweater and green pants. He has his hands clasped together in front of him, looking deep in thought with a serious expression on his face. The room features white walls and a window with blinds, providing a neutral background for the scene.
Please expand the following Tweet related to Target **Stay at Home Orders** into a detailed visual description. The Tweet's stance on target Stay at Home Orders is **Against**. Describe the scene in terms of colors, objects, people, and other visual elements. The generated image description should maintain the same stance towards the Target as the original Tweet.Tweet: **I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s**	An older man with a blue shirt stands in front of a window, looking out at the view. He is standing next to a table filled with various items such as figurines and picture frames. The room has a warm atmosphere, and the sunlight streaming through the window illuminates the scene
Please expand the following Tweet related to Target **Stay at Home Orders** into a detailed visual description. The Tweet's stance on target Stay at Home Orders is **Against**, and its sentiment is **negative**. Describe the scene in terms of colors, objects, people, and other visual elements. The generated image description should maintain the same stance towards the Target as the original Tweet.Tweet: **I m now 70 and have had a great life and the last thing I want is to be complicit in removing the freedom and liberty I have enjoyed , taken away from all future generations. #over50s**	An older man is sitting in a chair, watching the news on a television that is placed on a wooden cabinet. The TV screen displays an advertisement for a phone order service with the text Ësizz-a-fast phone orders - Lost freedom

Table 8: The process of prompt expanding for candidate image generation.

> Given the following caption, please determine whether it is related to the target target. Only respond with 'yes' or 'no'. caption: caption

Similarly, we will receive a 'yes' or 'no' response R_{T_C}. For the responses R_{T_I} and R_{T_C}, each 'yes' is scored 5 points, and each 'no' is scored 0 points. The sum of these two responses constitutes the score S_T.

Stance Consistency Score S_S

Next, based on the stance consistency criteria, we input the original text and the image into the VLM to ask whether the stance of the image aligns with the stance of the original text:

> <ImageHere> Given the tweet text, target target and the image, whether the image can reflect the stance of the tweet towards the target, reply with 'yes' or 'no'.

We will receive a 'yes' or 'no' response R_{S_I}. Then, we input the generated caption into ChatGPT to ask whether the stance of the caption aligns with the original text:

> Given the following caption and text, please determine whether the stance of the caption towards the target is consistent with the original text. Only respond with 'yes' or 'no'. caption: caption. original text: text

We will receive a 'yes' or 'no' response R_{S_C}. For the responses R_{S_I} and R_{S_C}, each 'yes' is scored 5 points, and each 'no' is scored 0 points. The sum of these two responses constitutes the score S_S.

Graph-based Images Re-ranking

Finally, the arithmetic mean of S_R, S_T, and S_S is the score of the image. Specific scores are shown in Table 10. We treat each image as a vertex, where the value of the vertex is the image's score, and the edges represent the similarity between images. After applying Algorithm 2, we obtain the optimal image.

	caption
P_1	The image captures an older woman with gray hair, wearing a blue jacket and earrings, standing in a field of tall grass. She is smiling at the camera, exuding a sense of joy and contentment. The background reveals a vast landscape of dry grassland, suggesting that she might be enjoying a peaceful walk or appreciating nature's beauty
P_2	An older woman is sitting at a wooden table, holding a piece of paper that reads S̈herman Carpet, Genuine Pile, 70 % Polyester, 30% Polyamide, 1.5mm Pile Height
P_3	The image features an older woman standing in a park-like setting. She is wearing a blue blouse and a black jacket, with her hands clasped in front of her. The woman appears to be smiling as she looks towards the camera. In the background, there are trees and a building visible.
P_4	An older woman, wearing a blue sweater and a necklace, is sitting at a table with a glass of beer in front of her. She is writing in a book with a yellow pen while smiling. The background features lush greenery and trees.
P_5	The image features a white garage door with red text that reads ÖIT STAYT HOMIE ORDER ORDIT MI - BREICMR 14. DE.SATAT.̈ The door is closed, and there are no people or actions taking place in the scene. The background consists of a yellow wall and some greenery on the left side.
P_6	In the image, an elderly man is sitting on a chair in a park, smiling and looking directly at the camera. He is wearing a blue shirt and jeans. In the background, there are children playing, with one of them holding a frisbee. The scene appears to be a sunny day, as indicated by the bright lighting and shadows.
P_7	An older man is sitting on a brown couch, wearing a gray sweater and green pants. He has his hands clasped together in front of him, looking deep in thought with a serious expression on his face. The room features white walls and a window with blinds, providing a neutral background for the scene.
P_8	An older man with a blue shirt stands in front of a window, looking out at the view. He is standing next to a table filled with various items such as figurines and picture frames. The room has a warm atmosphere, and the sunlight streaming through the window illuminates the scene.
P_9	An older man is sitting in a chair, watching the news on a television that is placed on a wooden cabinet. The TV screen displays an advertisement for a phone order service with the text S̈izz-a-fast phone orders - Lost freedom

Table 9: The generated caption for S_T and S_S.

	S_R	R_{T_I}	R_{T_C}	R_T	R_{S_I}	R_{S_I}	R_S	score
P_1	18.43	no	no	0	no	no	0	6.14
P_2	18.02	no	no	0	yes	no	5	7.67
P_3	17.24	no	no	0	no	no	0	5.75
P_4	17.29	no	no	0	no	no	0	5.76
P_5	14.75	no	yes	5	no	no	0	6.58
P_6	17.95	no	no	0	no	no	0	5.98
P_7	17.46	no	no	0	no	no	0	5.82
P_8	18.41	no	no	0	no	no	0	6.14
P_9	14.82	no	no	0	yes	no	5	6.61

Table 10: The Initial scores of candidate images in graph-based images re-ranking.