# 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 text. Previous studies have primarily focused on the text, which may not fully capture the implicit stance conveyed by the author. To address this limitation, we propose a novel approach that transforms original texts into *artificially generated images* and uses the visual representation to enhance stance detection. Our approach first employs a text-to-image model to generate candidate images for each text. These images are carefully crafted to adhere to three specific criteria: textual relevance, target consistency, and stance consistency. Next, we introduce a comprehensive evaluation framework to select the optimal image for each text from its generated candidates. Subsequently, we introduce a multimodal stance detection model that leverages both the original textual content and the generated image to identify the author's stance. Experiments demonstrate the effectiveness of our approach and highlight the importance of artificially generated images for stance detection.

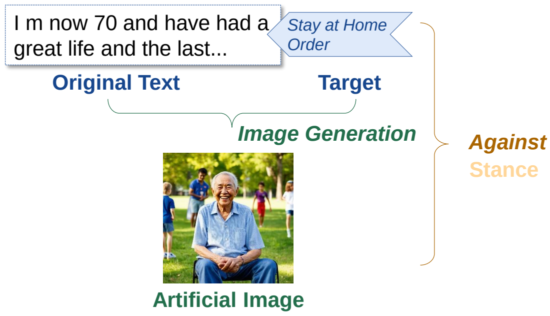


Figure 1: An example of artificial image generation for stance detection.

## 1 Introduction

Stance detection is a pivotal task in natural language processing, aiming to identify authors' attitudes from text. The automatic and accurate categorization of stances in complex linguistic contexts remains a significant research challenge.

Deep learning has substantially advanced stance detection. 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).

Although these deep learning models have shown strong capabilities in stance detection, they still mainly rely on text alone, 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 and language (Baltrušaitis et al., 2018). As a result, methods that incorporate multimodal information are able to capture complex patterns in the data more comprehensively. Weinzierl and Harabagiu (2023) proposes a synthetic data generation method based on text-image relation inference. Liang et al. (2024) proposes an effective targeted multimodal prompt tuning framework.

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 elements, 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.

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

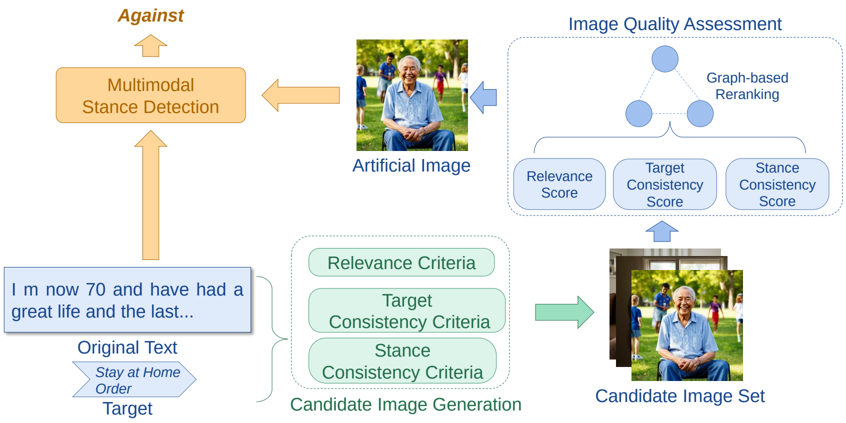


Figure 2: Overview of proposed model.

image format compared to the text alone. Nevertheless, creating such an image poses a challenge. It is imperative that the image encapsulates both the target and stance information from the original text, while simultaneously being easily comprehensible.

In our study, we initially employ a text-to-image model to generate candidate images for each text. The generation process adheres to three specific criteria: relevance, target consistency, and stance consistency. 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 images, 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. Experimental results demonstrate that our proposed approach significantly enhances performance.

In summary, the main contributions of our work are: (1) We propose a novel approach using artificially generated images to eliminate the dependency on user-generated image in conventional multimodal methods; (2) We develop a systematic image generation workflow guided by three criteria to create more informative image; (3) We design comprehensive evaluation metrics to systematically evaluate and select the optimal image.

## 2 Related Works

With the rapid advancement of deep learning, stance detection has achieved significant progress. 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.

Beyond textual analysis, multimodal approaches have emerged. Weinzierl and Harabagiu (2023) synthesizes multimodal examples from existing posts to reveal prototypical text–image stance relations. 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) introduces 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. Weinzierl and Harabagiu (2024) used counterfactual prompting for zero-shot multimodal reasoning.

Our proposed approach differs from previous studies that rely on user-generated images. In-

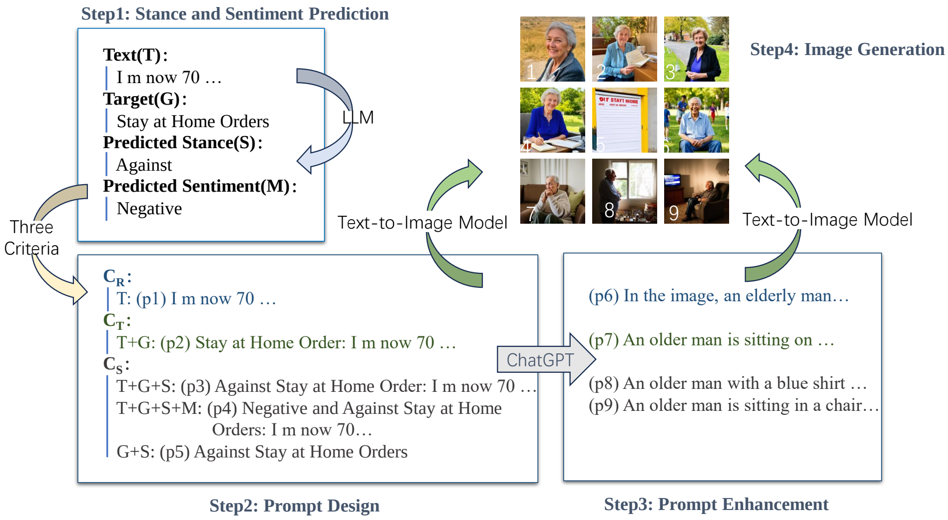


Figure 3: Dynamic Multi-output Generation Algorithm. $p$ represents image prompt. More details can be found in Appendix A.

stead, 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(VLM) 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 approach for stance detection, designed to generate an *artificial image* from the original text. Our approach consists of three key stages. First, we generate a diverse set of candidate images by applying criteria that ensure consistency with the stance and target in the original text. Second, we introduce a comprehensive evaluation framework with a graph re-ranking algorithm to assess the quality and accuracy of the candidate images, and select the optimal artificial image. Finally, we utilize a multimodal stance detection model to detect the stance through the original text and the selected image.

### 3.1 Candidate Image Generation

We first present the method for generating candidate images from the original text. To ensure semantic consistency between the generated candidate images and the original text, we formulate the following three criteria:

- **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. This aligns with the notion of attitude consistency proposed by Weinzierl and Harabagiu (2022).

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.

The first step of the algorithm is to predict the stance and sentiment of the text using a finetuned LLM. It is important to note that we maintain strict data separation. The stance labels used for image generation come from the predictions of the LLM (which is finetuned only on the training/validation set), and we never use the true labels from the test set.

The second step is to enrich the content of the image from multiple dimensions by integrating the text, target, and the predicted stance and sentiment, based on the three criteria. For example, according to the $C_T$ criterion, we concatenate the Target and Text to form the image prompt. The detailed prompts are shown in Table 7.

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. Therefore, the third step is to enhance the image prompts obtained in the second step by ChatGPT, in order to generate clearer and more comprehensible image prompts, thereby enhancing the understandability of the image content.

Finally, we input the image prompts obtained from both the second and third steps into the text-to-image model to generate the corresponding candidate images.

The implement detail of candidate images generation can be found in Appendix A.

### 3.2 Image Quality Assessment

To select the optimal candidate image, we propose three evaluation metrics guided by the criteria in section3.1.

**Relevance Score $S_R$:** Initially, leveraging cross-modal alignment prients from large-scale pretrained models, we employ CLIP (Radford et al., 2021) Score as the Relevance Score to quantify text-image correspondence. We encode both the candidate image and the text into vector representations by 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.

In the first step, the candidate image, the original text and the target of the text are input into the VLM 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 VLM to generate a caption for the image. The reason for regenerating the caption of the image instead of using the image prompt generated in Section 3.1 is that text-to-image models may not fully reflect all the details in the image prompt onto the generated image. This caption, the original text and the target of the 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 the sum of $R_{T_I}$ and $R_{T_C}$. Each "yes" response is scored 5 points, and each "no" is scored 0 points. The formula for $S_T$ is:

$$S_T = R_{T_I} + R_{T_C} \tag{2}$$

**Stance Consistency Score $S_S$:** Furthermore, we also employ the VLM to evaluate the stance consistency of the candidate images. Initially, both the candidate image and the original text are input into the model to determine if the image accurately reflects the stance. The VLM then generates 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 th sum of $R_{S_I}$ and $R_{S_C}$. Each "yes" response is scored 5 points and and each "no" is scored 0 points. The formula for $S_S$ is as follows:

$$S_S = R_{S_I} + R_{S_C} \tag{3}$$

### 3.3 Graph-based Image 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 1.

In this method, each node represents a candidate image. Its value is derived from the average of three point-wise metrics. The weight of the edge between two nodes corresponds to the semantic similarity between the two candidate images. The semantic similarity between image pairs is quantified through CLIP-based metric computations. To identify the most optimal image, we em-

**Algorithm 1** 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.

Initialize $PR(v)$ as the average of $S_R(v)$, $S_T(v)$, and $S_S(v)$, $\forall v \in V$

**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** $v^*$

ploy 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 image re-ranking can be found in Appendix B.

### 3.4 Multimodal Stance Detection

After obtaining the appropriate image, we utilize 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 = Given the **Tweet** and the **Image**, what is the stance towards **Target**.

We then employ a LLM 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 representation $y = y_1, \ldots, y_{|y|}$, where $y$ denotes the tokenized representation of the target text sequence (stance labels such as "favor"). At the i-th step of generation, the decoder predicts the $i$-th token $y_i$ , 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 function 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 $\theta$ is the model parameters, and $(X, S)$ is a input-output pair in training set $\tau$, 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.

## 4 Experiments

### 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. The Semeval-16 dataset contains five targets: Atheism, Climate Change is a Real Concern, Feminist Movement, Hillary Clinton, and Legalization of Abortion. The Covid-19 dataset contains four targets:

| Modality | Model | Semeval-16 | | | Covid-19 | | |
|---|---|---|---|---|---|---|---|
| | | $F_{favor}$ | $F_{against}$ | $F_{avg}$ | $F_{favor}$ | $F_{against}$ | $F_{avg}$ |
| Textual | 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 | 0.585 |
| | 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 |
| | TR-Tweet-COT | 0.701 | 0.787 | 0.744 | 0.844 | 0.804 | 0.824 |
| Multimodal | GFMAP | 0.655 | 0.763 | 0.709 | 0.735 | 0.685 | 0.710 |
| | TMPT | 0.689 | 0.781 | 0.735 | 0.782 | 0.774 | 0.778 |
| | Ours | **0.804** | **0.833** | **0.818** | **0.882** | **0.847** | **0.865** |

Table 1: Comparison with different baselines.

| Dataset | Split | Total | Favor | Against | None |
|---|---|---|---|---|---|
| Semeval16 | train | 2,620 | 678 | 1,254 | 688 |
| | dev | 294 | 75 | 141 | 78 |
| | test | 1,249 | 304 | 715 | 230 |
| Covid19 | train | 4,532 | 1,464 | 1,442 | 1,627 |
| | dev | 800 | 263 | 243 | 294 |
| | test | 800 | 263 | 243 | 294 |

Table 2: Statistics of dataset.

Face Masks, Fauci, School Closures, and Stay at Home Orders.

We finetune InternLM-XComposer2-VL (Dong et al., 2024) using the Low-Rank Adaptation technique as our base model. The learning rate in the main experiment is set to 1e-4. We select 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. **GFMAP** (Soltani and Romberg, 2023), which extracts text and image features through pre-trained models and then inputs these features into a classification model. **TMPT** (Liang et al., 2024), which learns multimodal stance features from text and visual modalities using target information.

As shown in Table 1, these text-based stance detection models exhibited commendable performance in stance detection tasks. Notably, the InternLM-TextOnly model stands out due to its exceptional language understanding and representation capabilities. Furthermore, our approach achieves consistent and stable performance improvements over baseline models within the multimodal fusion framework($p < 0.05$), demonstrating the effectiveness of our approach. These results underscore the importance of integrating

[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.

generated images into stance detection, highlighting the added value of multimodal information in enhancing the accuracy of stance detection tasks. In the Semeval-16 dataset, the inclusion of images led to an increase of 35, 27, and 29 correctly predicted instances for the Favor, Against, and None categories, respectively, accounting for 11.5%, 3.8 and 12.6% of their respective totals. For the Covid-19 dataset, the corresponding improvements were 19 17 and 9 additional correct predictions, representing 7.2%, 7.0%, and 3.1% relative improvements for each category. This demonstrates the complementary role of generated images in enhancing stance detection across diverse label distributions.

### 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

| 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.

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 artifi-

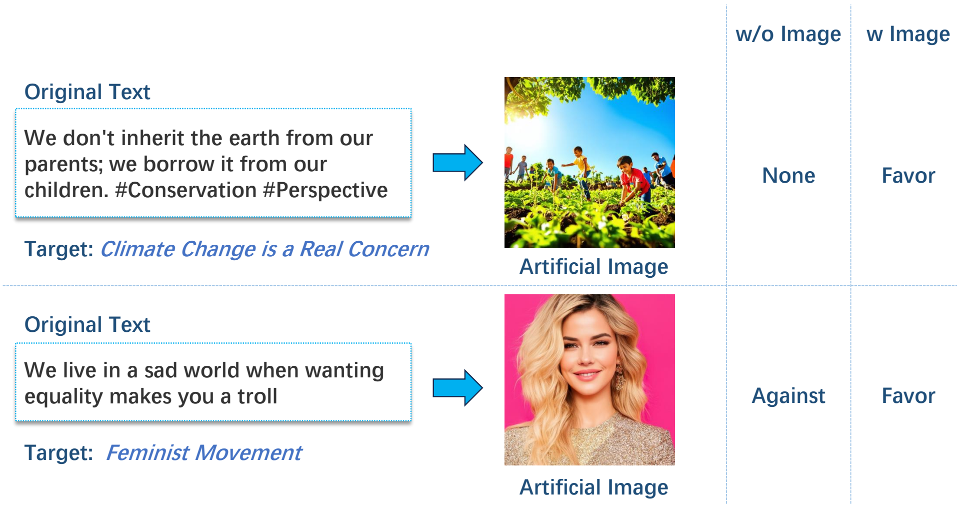


Figure 4: Examples of case study.

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

Table 6: Comparison with original images.

cially generated images outperform the basic textual models across all the multimodal models. This not only highlights the effectiveness of the 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 4.

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 by transforming the abstract stance in the text into concrete visual elements, 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 the generated image to enhance stance detection. Our approach begins by employing a text-to-image model to generate candidate images for a given text. 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 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 candidate images using a large VLM 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.

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.

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.

Mohammad Soltani and Julia Romberg. 2023. A general framework for multimodal argument persuasiveness classification of tweets. In *Proceedings of the 10th Workshop on Argument Mining*, pages 148–156, Singapore. Association for Computational Linguistics.

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.

Maxwell Weinzierl and Sanda Harabagiu. 2022. Identifying the adoption or rejection of misinformation targeting covid-19 vaccines in twitter discourse. In *Proceedings of the ACM Web Conference 2022*, WWW '22, page 3196–3205, New York, NY, USA. Association for Computing Machinery.

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.

## 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

Step1: We first finetune InternLM-XComposer2-VL on the training and validation sets with a learning rate of 1e-4 to predict the stance and sentiment of each text.

Step2: We design different image prompts based on three distinct criteria. For the $C_R$ criterion, we use the original text as image prompt1. For the $C_T$ criterion, we concatenate the Target and Text to form image prompt2. For the $C_R$ criterion, we concatenate the Stance, Target, and Text to form image prompt3, the Sentiment, Stance, Target, and Text to form image prompt4, and the Stance and Target to form image prompt5. As shown in Table 7, these correspond to prompts 1 to 5.

Step3: To make the image prompts more comprehensible for the text-to-image model, we enhance the image prompts obtained in the second step using ChatGPT. Since the results of enhancing prompt5 with ChatGPT are almost identical, we do not enhance prompt5 using ChatGPT. The model we used for ChatGPT is GPT-4o-mini, with parameters set as $temperature = 0$, $top_p = 1.0$, and $top_k = 50$. The prompts input to ChatGPT are shown in each row of Table 8, and the results are shown as prompts 6 to 9 in Table 7.

Step4: We select Stable-Diffusion-3 as our text-to-image model, with parameters set as num_inference_steps=28 and guidance_scale=7. We input the image prompts obtained from both the second and third steps into Stable-Diffusion-3. The generated images are shown in Table 7.

## B Implement Details of Image Quality Assessment

After obtaining the candidate images, we evaluate the quality of the generated images using a multi-modal assessment framework. The evaluation system is based on $C_R$, $C_T$, and $C_S$. We InternLM-XComposer2-VL as the VLM. The ChatGPT used in $S_T$ and $S_S$ is GPT-4o-mini, with parameters set as $temperature = 0$, $top_p = 1.0$, and $top_k = 50$.

### 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 VLM is as follows:

> 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, the target of the text and the image into the VLM to ask whether the image is related to the target:

> Given the **text**, **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:

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

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:

> Given the **text**, **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'.

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 Image 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 1, we obtain the optimal image.

## C Error Case Analysis

We conduct an error case analysis and find that in the Semeval-16 dataset, the incorporation of images results in 3, 32, and 19 misclassified instances for the "Favor," "Against," and "None" categories, accounting for 1%, 4.5%, and 8.3% of their respective total instances. In the Covid-19 dataset, the incorporation of images results in 7, 7, and 18 prediction errors for the "Favor," "Against," and "None" categories, corresponding to 2.7%, 2.9%,

| | 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 Ṡizz-a-fast phone orders - Lost freedom |

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

| | 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 Sherman 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 Sizz-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 image re-ranking.

Figure 5: One error case.

and 6.1% of their respective totals. This indicates that the "None" label is more negatively impacted by visual information, likely because images inherently convey a certain stance. As shown in Figure 5, the original text is "Seriously considering writing an article on the ban of school skirts at a secondary school, any thoughts would be appreciated", the target is "Feminist Movement" and the stance is "None". The generated image depicts a girl wearing a neat dress with a bright smile, conveying positive emotions that bias the model toward the "Favor" label. Additionally, we observed that images containing multiple human figures often exhibit distorted anatomies, which may further interfere with stance detection.