Saarland University

Department of Language Science & Technology

Seminar: **Argumentation Mining**

Evaluating VLMs on Multimodal Aristotelian Persuasion Tasks

Author:
Khondoker Ittehadul Islam
Matriculation: 7085810

Supervisors:
Dr. Olga Petukhova

May 31, 2026

Abstract

Vision Language Models (VLMs) have demonstrated exceptional performance across various tasks. However, they have not yet been thoroughly evaluated on more complex tasks. The Persuasion Model, conceived by Aristotle, resembles a triangle shape, which highlights its inherent challenges related to personal biases. To assess the progress of VLMs on these complex tasks, we use the `ImageArg` datasets, focusing on the Logos, Ethos, and Pathos detection tasks. Our findings indicate that models from the Qwen family achieve improved F1 scores, with Qwen3 performing exceptionally well on the Logos and Pathos tasks, while Qwen2 exhibits competitive performance on the more complex Ethos detection task. We release the code to foster research in this direction [1].

[1] `https://github.com/KhondokerIslam/ArgMin.git`

Contents

1 Introduction

1.1 Persuasion Mode

Persuasion Mode is the study of persuasive strategies categorized as Ethos, Logos, and Pathos, which are typically employed in persuasive tasks. First developed by Aristotle, Ethos refers to credibility, Logos represents reasoning, and Pathos pertains to emotion. As illustrated in Figure 1, these three modes overlap within a triangle, indicating that a persuasive text can incorporate one, two, or all three of them. For example, the statement, *"Thousands of homeless children will sleep in the cold tonight unless we act now,"* has a strong emotional appeal that may compel a specific group to take immediate action. Conversely, another group might feel sympathy but prefer to suggest long-term policy solutions. This demonstrates that different persuasion modes resonate with people's various logical, emotional, and ethical reasoning.

Figure 1. The persuasion triangle, illustrating how logos, pathos, and ethos overlap and jointly contribute to persuasive message.

1.2 Computational Persuasion Mode Mining

In the field of computational mining, the goal is to automatically identify persuasion structures and relationships. Argumentation mining and fallacy detection both require an understanding of persuasion modes. Research by Higgins and Walker (2012); Carlile et al. (2018) examined Aristotelian persuasion strategies (i.e., Ethos, Logos, and Pathos), focusing primarily on reports and student essays in textual formats. However, they overlooked the potential of utilizing other modalities, such as images, which could enhance persuasiveness (Liu et al., 2022). Although the persuasive mode has been studied in the audio-text modality within the context of fallacy detection shared tasks (Mancini et al., 2025), this approach yielded poorer results compared to the purely textual modality. The authors highlighted several challenges, including noise that impeded the effective use of audio for this task. Consequently, we evaluate the computational paradigm of persuasion

modes using the ImageArg dataset (Liu et al., 2022), which incorporates images as an additional modality alongside text.

1.3 Vision Language Models

Vision-language models initially followed a similar approach to that of language models by using token masking during training (Bordes et al., 2024). The key difference was the incorporation of image processing through encoders, which extract features from images, while decoders primarily handle textual inputs. An important development in this area was the introduction of CLIP (Radford et al., 2021), which proposed that both image and text modalities could share the same representational space, allowing them to enhance each other. Subsequent research aligned continuous image data with discrete textual data and introduced a new training strategy called Supervised Fine-tuning (SFT), where models are fine-tuned based on user instructions. As the field has evolved, there has been a concerted effort to optimize the loss function to effectively balance the relevant and irrelevant information during training, ultimately resulting in better generative models. These advancements have led to improved shape bias and alignment with human judgment (Jaini et al., 2024). Consequently, VLMs have been extensively explored in simple visual question-answering tasks (Antol et al., 2015). However, there is limited literature on persuasion from a multi-modal perspective. A study by (Zhou et al., 2025) integrated pragma-dialectical steps with paralinguistic cues from audio elements into the in-context learning of the Qwen-3 reasoning model (Yang et al., 2025). Nevertheless, they identified significant limitations in the models' ability to effectively process long prompts, which could affect their performance. To address this issue, we will use the coding manual provided by the ImageArg authors as our prompt. More details can be found in Section 3.1.

2 Literature Review: ImageArg

2.1 Corpus Creation

The authors of the `ImageArg` dataset, as detailed in Liu et al. (2022), used the Twitter API[2] to gather multi-modal data related to gun control by employing specific keywords associated with the topic[3]. They only retained tweets that received a confidence score greater than 0.9 from the ArgumentText Classify API[4] for the purpose of annotation. Next, they set up a pipeline for annotating these instances, which included tasks ranging from basic stance detection to detailed persuasion mode detection. A notable aspect of their approach was the introduction of a metric called the *Persuasive Score Improvement*. Annotators were required to first assign a score based solely on the text of the tweets and then provide a second score after reviewing the accompanying images. Only those scores that surpassed a predetermined threshold were considered valid for each detected persuasion mode. Ultimately, out of an initial total of 1,003 instances, only 259 were found suitable for the persuasion mode detection task. Table 1 presents the inter-annotator agreement (IAA) scores (Krippendorff, 2011) for the persuasion mode, which ranged from 50% to 58%, indicating a moderate level of agreement among annotators. This variability highlights the challenges involved in the task, which are often influenced by the personal biases of the annotators. To enhance the effectiveness of the annotation process, the authors provided comprehensive instructions to the annotators. Figure 2, 3, and 4 each showcase examples of the different persuasion modes.

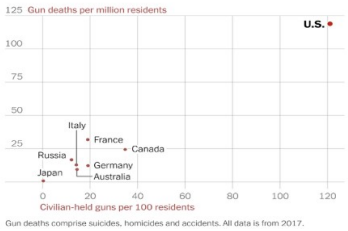
Task	Alpha	Count
Logos	55.3	259
Pathos	51.0	259
Ethos	57.8	259

Table 1. Alpha illustrating the inter-annotator agreement (IAA) score of `ImageArg` on the Persuasion Mode task.

[2] `https://developer.twitter.com/en/docs/twitter-api`

[3] The keywords were sourced from Guo et al. (2020).

[4] `https://api.argumentsearch.com`



Logos: Gun deaths and gun ownership by population - by country. Hmmm. Well, this doesn't take much effort to figure out why we've got such #GunViolence…

Figure 2. Example of logos in the `ImageArg` dataset. The image presents statistical evidence on gun ownership and gun deaths across countries, while the accompanying tweet text interprets the observed relationship to support an argument about gun violence.

Pathos: A personal narrative - Dr. Sonya Lewis "We must reject helplessness and complacency and we must allow ourselves to feel the raw, sick"

Figure 3. Example of ethos in the `ImageArg` dataset. The image depicts Abraham Lincoln in an American-themed outfit, with his hand covering his face in grief. The accompanying tweet conveys a personal narrative from an independent individual asking for strength.

Ethos: The US has 4.4 % of the world's population but 42% of gun violence. #guncontrol #gunviolence

Figure 4. Example of pathos in the `ImageArg` dataset. The image features a paper-cut headline from The New York Times alongside a paper-cut excerpt from the newspaper that provides facts supporting the headline. The accompanying tweets reiterate those facts.

2.2 Multi-modal Benchmarking

For benchmarking, Liu et al. (2022) utilized the image encoder ResNet50 and the text encoder BERT to fine-tune linear classifiers. They referred to ResNet50 as I-M (Image Modality; He et al. (2016)) and BERT as T-M (Textual Modality; Devlin et al. (2019)), with M-M representing multi-modality, where ResNet50 acts as the encoder and BERT as the decoder. They established a baseline with the random baseline, referred to as bm. The authors reported that bm achieved a higher F1 score in Logos and Pathos detection, while T-M performed competitively in Ethos detection. In terms of precision, T-M and I-M attained better scores in Pathos and Ethos detection, respectively. However, M-M showed poor performance across all tasks and evaluation metrics, suggesting that a more complex and better-trained M-M could lead to improved results. Subsequent works on this dataset focused exclusively on the stance detection task, raising the pressing question of:

(**Research Question**): *How are current Vision-Language Models (VLMs) are evolving to address this complex challenge.*

3 Experiment, Results & Discussion

3.1 Methodology

To answer this research question, we considered *Qwen2-VL-7B-Instruct* and *Qwen3-VL-8B-Instruct* of the same Qwen family. We set the models that achieved the best F1-score on respective persuasion modes on the ImageArg dataset as baseline, i.e., (bm for Logos and Pathos, and T-M for Ethos). We only evaluated on the test set on a zero-shot setting and adopted the same evaluation metric of this paper, i.e., Precision, Recall, and F1-score, and kept the same image size reported in the paper for fair evaluation. For prompt, we utilized the code manual and instructions used to build this dataset during annotation and set `generation_prompt` as `TRUE`. We left the remaining generation parameters of these models unchanged and ran on a single NVIDIA A100 Tensor Core GPU.

3.2 Results

Persu. Mode	**Model**	Prec ↑	Rec ↑	F1 ↑
Logos	BASE (*bm*)	0.405	**1.000**	0.575
	Qwen2	0.596	0.885	**0.709**
	Qwen3	**0.633**	0.813	**0.709**
Pathos	BASE (*bm*)	0.554	**1.000**	0.712
	Qwen2	**0.568**	0.969	**0.714**
	Qwen3	0.538	**1.000**	0.696
Ethos	T-M (*bm*)	0.168	0.817	0.272
	Qwen2	0.195	0.550	0.277
	Qwen3	**0.236**	**0.875**	**0.365**

Table 2. Performance of persuasion mode classification across Logos, Pathos, and Ethos using precision (Prec), recall (Rec), and F1-score metrics. The Qwen models achieved the strongest overall performance, attaining the highest F1-scores across persuasion modes.

Bold values indicate the best result for each metric within a persuasion mode.

Table 2 presents the results of this study. Overall, the Qwen models achieved the highest F1-scores across all persuasion modes. Specifically, the Qwen models demonstrated significantly improved precision in detecting Logos, indicating a better understanding of logical reasoning as it relates to interpreting statistical evidence in images. Additionally, both models showed competitive results in terms of precision and recall, with Qwen2 performing slightly better in comprehending persuasive texts and images. Interestingly,

Ethos is the only task where both models were very close to the baseline, highlighting the challenge of overcoming emotional bias, which also influenced the annotation process.

3.3 Discussion

To compare the two generations of the Qwen family, Qwen3 exhibited superior ethical and credibility comprehension compared to Qwen2. This suggests that Qwen3 has enhanced image processing capabilities and better utilization of numerical image data, while Qwen2 slightly excels in understanding hidden emotional appeals. Looking ahead, it would be interesting to investigate how much image modality influenced these performances, which closely aligns with this paper's other task, *Persuasive Score Improvement*.

4 Conclusion

In this research, we evaluate the effectiveness of Vision-Language Models (VLMs) in the complex task of detecting persuasion modes. Persuasion modes—Logos, Pathos, and Ethos—overlap within a persuasion triangle, complicating the influence of inherent personal biases. For our evaluation, we selected the `ImageArg` dataset, where the authors reported moderate annotator agreement due to the challenges associated with the task. To assess model performance, we focused on two generation models from the Qwen family to understand how these models are evolving to tackle this complex challenge. Overall, we found that Qwen models achieved the highest F1 scores when compared to benchmark models. Specifically, Qwen3 demonstrated superior performance in identifying the Logos and Pathos modes, while Qwen2 performed competitively in the more difficult Ethos detection task. Future research could explore the effectiveness of these VLMs in detecting more nuanced persuasion modes, such as *Ad Hominem* and *Appeal to Authority*, as well as their application to related tasks such as Fallacy Detection.

5 Limitations

5.1 Alternative VLM Family

Currently, numerous VLM model families are available to address this research question. However, we utilize Qwen because of its popularity and comprehensive open-source report. Since different VLMs are trained with distinct objectives, their performance may vary if another VLM were chosen.

5.2 Evaluation Setup

In our study, we evaluate only in a zero-shot setting. Since the community representing these language families does not release their training data, we assume that the training instances for these tasks were included in the models' training. However, this may not necessarily be the case. If it is not, an n-shot setup could be employed to compare the results against the same baseline, which was trained with significantly more samples.

References

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., and Parikh, D. (2015). Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Bordes, F., Pang, R. Y., Ajay, A., Li, A. C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., et al. (2024). An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*.

Carlile, W., Gurrapadi, N., Ke, Z., and Ng, V. (2018). Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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.

Guo, M., Hwa, R., Lin, Y.-R., and Chung, W.-T. (2020). Inflating topic relevance with ideology: A case study of political ideology bias in social topic detection models. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4873–4885.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Higgins, C. and Walker, R. (2012). Ethos, logos, pathos: Strategies of persuasion in social/environmental reports. In *Accounting forum*, volume 36, pages 194–208. Elsevier.

Jaini, P., Clark, K., and Geirhos, R. (2024). Intriguing properties of generative classifiers. In *International Conference on Learning Representations*, volume 2024, pages 13898–13923.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Liu, Z., Guo, M., Dai, Y., and Litman, D. (2022). Imagearg: A multi-modal tweet dataset for image persuasiveness mining. In *Proceedings of the 9th Workshop on Argument Mining*, pages 1–18.

Mancini, E., Ruggeri, F., Villata, S., and Torroni, P. (2025). Overview of mm-argfallacy2025 on multimodal argumentative fallacy detection and classification in political debates. In *Proceedings of the 12th Argument mining Workshop*, pages 358–368.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhou, H., Westerdijk, H., and Islam, K. I. (2025). Joint effects of argumentation theory, audio modality and data enrichment on llm-based fallacy classification. *arXiv preprint arXiv:2509.11127*.