Assessing the Persuasive Effect of AI-Generated Image Support of Arguments

Mackwyn Quadras[1] **Manfred Stede**[1] **Henning Wachsmuth**[2]
[1]USF Cognitive Science, University of Potsdam, Germany
[2]L3S Research Center & Institute of Artificial Intelligence, Leibniz University Hannover, Germany

mackwyn.quadras@uni-potsdam.de, stede@uni-potsdam.de, h.wachsmuth@ai.uni-hannover.de

Abstract

Argumentation is, at its core, an inherently verbal activity. Yet, other modalities may support arguments, one of which are images. In the argument mining community, this combination has not received much attention yet. While a few previous works studied whether images can make argumentative texts more effective in persuading people, the images that were considered matched the texts loosely only, or they were heavily text-based themselves. In this paper, we take the step to study to what extent the persuasive effect of textual arguments can be supported by images specifically created for this purpose. For a consistent experiment design, we combine NLP with image generation to synthesize both arguments and images with generative AI, for five controversial topics and for two rhetorical strategies. In two consecutive user studies, we first determine the best-matching image for each argument and then compare the perceived effect of bare textual arguments to those that are supported by an image. Our results suggest that the images may increase the persuasive effect of argumentative texts, but with variance across topics.

Keywords: argument, persuasion, multimodality, images

1. Introduction

According to argumentation theory, the act of producing an argument is a verbal, social, and rational activity (van Eemeren and Grootendorst, 2004). Undoubtedly, the use of language (the verbal aspect) is at the core of human argumentation, and similarly, logical reasoning (the rational aspect) plays an important role. Yet, in everyday life, we can also find additional media taking part in argument production. In some social contexts, *images* in particular can contribute to argumentative effects — often not primarily for rational reasons, but rather for the rhetorical side of argumentation, that is, to achieve persuasion. Persuasion is related to, but not the same, as argumentation: It refers to intentional attempts to influence someone's attitudes or behaviors, and can occur through reasoned discourse or through emotional appeals (O'Keefe, 2002).

Consider the two arguments on school uniforms in Figure 1. Both texts make reasonable points toward the stance they support. However, the images placed next to the arguments draw attention. The research question that we will address in this paper is to what extent images can add to the persuasive effect of textual arguments, either for supporting the reasoning or for creating emotional appeal.

While most NLP work on persuasive argument has focused on text (Lukin et al., 2017; El Baff et al., 2020; Alshomary et al., 2022), a few recent studies explored the role of multimodal elements, such as images, in supporting the persuasiveness of arguments. Kiesel et al. (2021) presented first approaches to the retrieval of images that generally fit to an argument's topic and stance. Liu et al.

(a) Pro school uniforms

Promotes School Unity and Pride – Wearing a uniform fosters a sense of belonging, making students feel part of a team. It strengthens school identity and encourages discipline, preparing students for professional life.

(b) Con school uniforms

Freedom of Expression – Every child is unique, and forcing them into the same uniform suppresses their individuality. Letting students wear their own clothes allows them to express their personality and boosts their confidence.

Figure 1: Two pathos-oriented arguments (one pro, one con) on whether to have uniforms at school, along with images to support them. Both the arguments and the images have been created with generative AI for our study.

(2022) created a dataset with persuasiveness annotations of image-supported argumentative tweets that was later used in a shared task (Liu et al., 2023). However, the latter images are largely text-based, doubling what the textual argument conveys rather than supporting the text with visual impressions.

In this paper, we study to what extent the per-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8085–8095

suasive effect of arguments can be increased by images specifically created to support a given topic and stance. To obtain images that match the topic and stance of an argument well, we exploited the capabilities of generative AI, namely by prompting Dall-E (Ramesh et al., 2021). For a consistent experiment design, we also generated the textual arguments synthetically using GPT-4 (OpenAI, 2023).[1] Specifically, we created six pro and six con arguments for each of five topics. Following related studies (Wachsmuth et al., 2018), half of these arguments are meant to appeal to reason (logos in Aristotle's sense), the other to emotion (pathos). For image support, we generated three candidates each for the pro side and the con side of a topic. Conceptually, the two generation steps could be pipelined to obtain a fully automatic approach.

For the purpose of this paper, the generated arguments and images defined the basis for two consecutive user studies. First, we asked 26 participants to decide for all candidate images which of them best supports a logical perspective on the respective topic and stance, and which supports an emotional perspective. Pairing the best images with two of the three arguments per topic, stance, and rhetorical strategy (logos, pathos) we then carried out the second study to assess the persuasive effect of (a) the argument alone versus (b) the argument combined with the image.

Our results suggest that, overall, AI-generated images can increase the perceived persuasive effect of arguments. Matching expectation, we observe that the effect is bigger for pathos-oriented arguments than for logos-oriented ones. However, it differs across topics, which may possibly be due to varying difficulty in visualizing them in convincing ways. Generally, our findings provide a ground for studying how to effectively make use of images within multimodal argumentation technologies.

In the following, we first discuss related work (Section 2). Section 3 then describes how we produced both the argumentative texts and the images. In Sections 4 and 5, we present design and results of the two studies on image selection and persuasiveness rating. Section 6 concludes the paper.

2. Related Work

As discussed in the introduction, the vast majority of theoretical research on argumentation has centered on rational and critical thought (Toulmin, 1958; van Eemeren and Grootendorst, 2004; Walton et al., 2008). Accordingly, argument mining has largely built on these ideas; see the surveys by Stede et al. (2026) and Lawrence and Reed (2019). Still, increasing the readers' or listeners' *engagement* and *emotional appeal* can play a crucial role in persuasive arguing, as has long been known in rhetoric research (Aristotle, ca. 350 B.C.E./ translated 2007), and also found its way into some NLP research on argumentation.

In early work, Anand et al. (2011) annotate persuasive acts in blog posts, focusing on three mental states in persuasion: belief revision, attitude change, and compliance gaining. They propose 16 rhetorical strategies, such as appeals to emotion, credibility, and logic, utilizing methods like analogies, examples, and rhetorical questions. Duthie et al. (2016) extract positive and negative ethotic statements from UK parliamentary debates, emphasizing ethos over the logos-focused approach common in argument mining. Their work provides tools to analyze credibility in communication, revealing political dynamics such as cliques and coalitions. It highlights the role of interpersonal dynamics in structured debates, bridging sentiment analysis and argument mining. Unlike these works, we assess persuasive effects rather than identifying their basis in natural language.

El Baff et al. (2018) introduce a novel model for evaluating argument quality in news editorials, focusing on how arguments challenge or empower readers with opposing views. By analyzing 1,000 editorials from the New York Times, they reveal that high-quality argumentation can bridge ideological divides. Their work emphasizes the need to examine how different persuasive elements influence readers' agreement with arguments. In later work, El Baff et al. (2020) analyze to what extent linguistc style affects the persuasive effect of the editorials, finding that this differs across readers' ideologies. The results are in line with earlier work by Lukin et al. (2017) who reveal that different types of arguments (emotional vs. rational) affect readers differently depending on their personality.

Wachsmuth et al. (2018) propose a model for synthesizing arguments that integrate rhetorical strategies based on Aristotle's concepts of logos, ethos, and pathos. The model involves selecting content, arranging argument structure, and phrasing style to maximize persuasive effectiveness. By analyzing argumentative units and their roles, the study provides a framework for generating texts that align with specific rhetorical strategies. A prototypical realization was done by El Baff et al. (2019).

Wang et al. (2019) emphasize the importance of personalized persuasive dialogue systems for social good by tailoring strategies to individuals' psychological and demographic profiles. Analyzing 1,017 dialogues, the authors identify ten persuasion strategies and demonstrate how personalized approaches enhance persuasive communication.

[1] Despite limited reproducibility, we opted for closed-source models to maximize quality. All arguments and images can be found at `https://github.com/discourse-lab/ImageArguments-LREC26`

Their results highlight the need to understand the characteristics of the audience to maximize the persuasive effect of the argument. Similarly, Alshomary et al. (2022) show that morally-framed arguments are more effective, if they are aligned with the audience's beliefs. Findings like these hint at the idea that *images* can play an important role in argumentation, too. If carefully selected, they may align with the emotional or logical components of an argument to enhance its persuasive power, in particular when they resonate with the audience's values. In the argument mining community, multimodality has recently gained more attention, for instance, with the release of platforms for dealing with multimodal (written, auditory) information (Mancini et al., 2024). As for images, the VisArgs benchmark (Chung et al., 2024) explores the connection between image content and reasoning; the role of image/text combination for persuasion, however, is still underexplored.

Kiesel et al. (2021) propose to retrieve "argumentative images" using stance-aware query expansion. By incorporating stance-indicating terms, their approach effectively retrieves images that align with pro and con positions on various topics. Their research offers a useful framework for selecting images that enhance argument persuasiveness, demonstrating the role of targeted visual elements in persuasive communication. Liu et al. (2022) present *ImageArg*, a multimodal dataset of tweets annotated for image persuasiveness. By benchmarking image persuasiveness tasks and analyzing how visual elements strengthen arguments, the study provides valuable insights into the role of images in argumentative effectiveness. The few other works in this context also look at the how to ensure that an image actually matches the stance of an argument (Carnot et al., 2023).

While our work shares the broader goal of exploring the persuasive effect of images, it differs in scope and focus. Unlike ImageArg, which involves images containing textual arguments or slogans, our study specifically investigates the persuasive power of standalone images, devoid of any embedded text or overt contextual information. This approach enables us to isolate the visual impact of imagery on persuasion. By building on the previous studies, our research seeks to understand not only whether images can enhance persuasion but also what type of connection exists between text and image when they work together in an argument.

(a) Logos pro argument

Resource Efficiency – Proper waste segregation ensures that recyclable materials, such as paper, plastics, and metals, are processed correctly, reducing the demand for raw materials. This conserves natural resources and decreases energy consumption in manufacturing.

(b) Logos con argument

Limited Recycling Feasibility – Not all materials can be effectively recycled due to mixed compositions, contamination, or lack of processing facilities. In some cases, segregating waste does not lead to meaningful environmental or economic benefits.

(c) Pathos pro argument

Protecting Our Children's Future – Imagine your children growing up in a world filled with overflowing landfills and polluted oceans. By simply segregating waste, we ensure a cleaner, safer environment for future generations to thrive.

(d) Pathos con argument

Frustration with Inefficiency – Many people carefully separate their waste only to see it all mixed together at collection sites. If the system is flawed, why should individuals bear the emotional guilt of improper disposal?

Figure 2: Two pro and two con example arguments (one logos-oriented and one pathos-oriented each), generated by ChatGPT for the controversial topic "Waste segregation is necessary."

3. Data

In this section, we describe the creation process of the AI-generated data that defines the basis of the user studies carried out in this work: textual arguments and candidate images to support them.

3.1. AI-Generated Arguments

While a sizable variety of argument corpora have been introduced to NLP research, the arguments they supply are often time-specific (Al Khatib et al., 2016), genre-specific (Ng et al., 2020), platform-specific (Skitalinskaya et al., 2021), or created by humans with a specific background (Toledo et al., 2019). To refrain from such specificities and dependencies, but also for a consistent experiment design, we decided to not only generate the images, but also the arguments they are meant to support, by (separate) AI models. To imitate a real-word setting, we started from five different controversial issues that we borrowed from the 'argumentative microtext' corpus of Peldszus and Stede (2016):

- Promote *waste* segregation as necessary
- Allow supermarkets to operate on *weekends*
- Increase *fines* for dog owners who don't clean up
- Restrict *meat* consumption for better health
- Have school *uniforms* for discipline and unity

As we expect that the effect of image support may differ depending on an argument's style, we distinguish two general rhetorical strategies for the arguments, in line with Wachsmuth et al. (2018): the appeal to reason, called *logos* by Aristotle (ca. 350 B.C.E./ translated 2007), and the appeal to emotion, *pathos*. For both, we intended to cover both stances (pro, con) on each controversial issue. Further, we aimed to produce three arguments per issue/strategy/stance combination, so that we capture some diversity in viewpoints. With these goals in mind, we prompted ChatGPT for each topic, as exemplified here for waste segregation:

- **Logos prompt:** "I need Logos-oriented arguments for some particular topics. Logos can be defined as - Argue based on logical reasoning, which means to make rational and logical conclusions towards the intended stance of the given topic. Generate three arguments in favor and three arguments against waste segregation."
- **Pathos prompt:** "I need Pathos-oriented arguments for some particular topics. Pathos can be defined as - Argue based on emotional reasoning, which means to appeal to the emotions of the reader regarding the topic within your argument. Generate three arguments in favor and three arguments against waste segregation."

In total, we hence created 60 arguments: three pro and three con logos-oriented and pathos-oriented arguments each for each of the five topics. Figure 2 above shows one example argument on waste segregation for each pair of stance and rhetorical strategy. As we explain in Section 4, we later reduced the sets from size three to size two for the final study. The full set of arguments generated for the *waste segregation* topic are shown in the appendix (and all other materials are in the anonymous repository mentioned in footnote 1).

3.2. AI-Generated Images

For image generation, we relied on Dall-E (Ramesh et al., 2021), due to its ability to create diverse, high-quality visuals based on text prompts. To achieve some variety in image style, we used three distinct prompt styles for producing images on a topic:

- *Argument-based prompt.* We provide the complete argument text to the image generator. To avoid "AI circularity", at this point we did not use one of the AI-generated arguments, but one from the 'microtext' corpus mentioned above (Peldszus and Stede, 2016).
- *Generally-descriptive prompt.* We merely provide a general description of the pro viewpoints and con viewpoints.
- *Emotionally-charged prompt.* Same as general-descriptive but with an additional instruction to add emotional appeal.

Each prompt was adapted for the pro and con stances, which resulted in six images per topic. This multi-style prompting approach allows us to explore in the subsequent second study how different visual styles may resonate with rhetorical strategies, while maintaining consistency across topics and stances. To illustrate, these are the prompts for the topic of waste segregation:

Figure 3: The three pro and three con images generated by Dall-E for the controversial topic "Waste segregation is necessary."

- **Argument-based prompt:** "Yes, it's annoying and cumbersome to separate your rubbish properly all the time. Three different bin bags stink away in the kitchen and have to be sorted into different wheelie bins. But still Germany produces way too much rubbish and too many resources are lost when what actually should be separated and recycled is burnt. We should take the chance and become pioneers in waste separation!
Give me an image in support of this argument and another one opposing it."
- **Generally-descriptive prompt:** "Now generate an image showing waste segregation as a step toward sustainability, and another image showing the challenges and inefficiencies of waste segregation."
- **Emotionally-charged prompt:** "Now generate an image showing waste segregation as a step toward sustainability, and another image showing the challenges and inefficiencies of waste segregation. Both with more emotional appeal."

Figure 3 shows all six candidate images that Dall-E generated for the waste segregation topic.

		Logos-oriented						Pathos-oriented					
		Image A		Image B		Image C		Image A		Image B		Image C	
Topic	Stance	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio	Count	Ratio
Waste	Pro	**22**	4.400	7	1.000	1	0.055	5	0.227	7	1.000	**18**	18.000
	Con	10	0.909	**20**	4.000	0	0.000	11	1.100	5	0.250	**14**	N/A
Weekends	Pro	**17**	8.500	11	2.75	2	0.083	2	0.117	4	0.363	**24**	12.00
	Con	12	1.200	16	**1.454**	2	0.222	10	0.833	11	0.687	9	**4.500**
Fines	Pro	13	**0.866**	12	N/A	5	0.333	15	1.153	0	0.000	15	**3.000**
	Con	11	2.750	**14**	1.750	5	0.277	4	0.363	8	0.571	**18**	3.600
Meat	Pro	6	1.200	18	**1.384**	6	0.500	5	0.833	13	0.722	12	**2.000**
	Con	10	0.588	**14**	7.000	6	0.545	**17**	1.700	2	0.142	11	1.833
Uniforms	Pro	**17**	3.400	12	1.333	1	0.062	5	0.294	9	0.750	**16**	16.000
	Con	8	2.666	**17**	1.545	5	0.312	3	0.375	11	0.647	**16**	3.200

Table 1: Pre-study results: Participant vote distribution (counts and ratios) of the three image options for each topic-stance pair, separately for the logos- and pathos-oriented perspective. 'N/A' refers to "not applicable" because of division by zero. Per distribution, the highest value – either the decisive count or the decisive ratio (see Section 4.2) – is in bold.

4. Pre-Study: Aligning Images with Rhetorical Strategies

As a preparation for our main experiment on the persuasive effect of supporting arguments with images (Section 5), we did not simply trust that Dall-E produces all the varying images successfully. Instead, we ran an experiment with human raters for selecting from the generated image candidates those that are a good fit to the topic, the intended stance, and the rhetorical strategy, respectively.

4.1. Experimental Setup

To avoid overlap between the participants in the two experiments, here we worked with 30 master's and PhD students with a computer science background and/or a linguistics background, but not necessarily particular expertise in argumentation. We conducted the selection study with an online questionnaire using MS Office Forms.

For each topic, we showed the participants the three generated images with the respective stance and gave them the following instructions:

- **Logos-oriented:** "Please select the image that best aligns with a logical or rational perspective on the topic."
- **Pathos-oriented:** "Please choose the image that most effectively evokes an emotional response relevant to the topic."

In this way, we collected votes for each individual image, regarding how well it combines with a topic and stance and intended rhetorical effect.

4.2. Results

Table 1 shows the distribution of participants' votes over the image options of each topic-stance pair, broadly distinguished by rhetorical strategy (logos on the left, pathos on the right). Votes across the three image candidates add up to 30. To illustrate, for the selection on *Waste/Pro/Logos*, a clear majority of 22 participants voted for Image A, which we thus selected, as indicated by the boldface marking.

A complication for the evaluation arises from the fact that we offered participants the same images for both the logos and the pathos strategy – we deliberately did not pre-select images for strategy. Accordingly, it could happen that the same image receives the most votes both for pathos and for logos. Example: For the topic *Weekends*, image B receives the most votes both for *Contra/Logos* (16) and for *Contra/Pathos* (11). In order to not use the same image in both strategies, in such cases we select the image based not on the count but on the ratios of logos/pathos votes and of pathos/logos votes instead. In our example, the highest logos/pathos ratio for *Weekends/Contra* is 16/11=1.45 (image B), and the highest pathos/logos ratio is 9/2=4.5 (image C).

4.3. Final filtering stage

Following the image selection, we conducted an internal review to further refine the quality and coherence of our argument-image pairs. Recall that our participants matched images only to the general topics; not to the actual argument texts (which would have required many more participants). As Figures 2 and 3 demonstrate, images and texts can address the topic in quite different ways, by emphasizing different aspects, so that combinations do not always align to the same degree.

As described in the previous section, we had initially prepared three arguments for each topic, stance, and rhetorical strategy, resulting in 12 arguments per topic. To ensure a good match between argument texts and their corresponding images, we reviewed the data and rated each argument's alignment with its assigned image. Specifically, each author of this paper rated every text/image combination on a three-level scale: *A (strong match)*, *B (moderate match)*, and *C (mismatch)*.

After assigning these ratings individually, we excluded the argument that aligned to the weakest degree with the image for each topic-stance-strategy triple. In particular, among the 60 arguments, 8 received three A's, 5 two A's, 4 one A and two B's, and 14 further at least two B's. This was sufficient to select two argument-image pairs for 14 of the 20 topic-stance-strategy combinations. We discussed the remaining pairs jointly to reach a consensus. In one single case (*weekend*, *con*, *logos*), we had to keep one pair where all three had chosen C.

This step reduced the dataset to two arguments per topic/stance/rhetoric combination and hence to 40 arguments in total. It helped to ensure that the final set of arguments for the main experiment is indeed well-suited for evaluating the effect of images on persuasive communication.

5. Main Study: Assessing the Persuasive Effect of Image Support

We now turn to the main research question, that is, whether and how much the persuasive effect of a textual argument can be influenced by a suitable image. We conducted a second user study, this time in a crowdsourcing setting using the *Prolific* platform. with Google Forms as backend. Participants were recruited according to these criteria:

- **Country of residence:** United Kingdom, United States, Ireland
- **First and primary language:** English
- **Highest education level:** Graduate degree (MA, MSc, MPhil, other), Doctorate degree (PhD, other)
- **Approval rate:** At least 95%

Participants were paid on the basis of an hourly rate of GBP 9.00, with completion time for the experiment estimated at 10 minutes. Before agreeing to the experiment, participants were informed that payment was conditional on a short comprehension check question as the last item in the questionnaire (in order to reduce the probability for inattentive answering). The question asked for a short description of the perceived difference between 'reasonable' and 'convincing' arguments. In the end we had to reject only one out of 80 participants, on the grounds of very quick response time and an insufficient answer to the comprehension question.

5.1. Experimental Setup

In the experiments, participants rate either *text-only* arguments, or combinations of *text and image*. In the following, we use the term *item* to refer to both, and call them *T-items* and *TI-items* respectively.

Recall that our arguments pertain to five different topics (*waste* segregation, supermarkets operating on *weekends*, *fines* for dog owners, *meat* consumption, and school *uniforms*). For each topic, we built eight argument-image pairs: two *pro* and two *con* logos-oriented arguments; and likewise 2+2 pathos-oriented arguments. This results in 8 TI-items per topic, and hence in 40 TI-items in total. Adding the corresponding versions without image, i.e., the T-items, the complete set consist of 80 items. We aimed at collecting five responses per item.

We decided that each participant should rank one item per topic, and hence five items in total. Further, they should see either only T-items or only TI-items during the survey, so no one rated both T-items and TI-items. The rationale for this decision was to avoid bias arising from judgments made earlier on a different type of item. We created 16 Google forms, where each contained one item for every topic. In total, 80 people participated in the study, with five participants each being assigned the same of the 16 forms. Combinations of topic, stance and rhetoric were balanced and randomized across surveys to ensure variety and avoid repetition.

For rating each individual item, we focus on high-level criteria, as the limited scale of the study may make fine-grained criteria too much affected by small phrasing deviations. In particular, participants were given three questions that we derived from notions of rhetorical and dialectical quality in the argument quality taxonomy of Wachsmuth et al. (2017). They had to be answered by choosing from a 5-point Likert scale:

Q1. Effective to oneself: To what extent is the argument [along with the accompanying image] persuasive to you on the given topic?

1. Not at all persuasive
2. Rather not persuasive
3. Partly persuasive, partly not
4. Rather persuasive
5. Highly persuasive

Q2. Effective to an undecided person: Imagine someone who is undecided about being in favor or against the given topic. To what extent would you think the argument [along with the

Question	Image	Results by Topic					Results by Strategy		Overall
		Waste	Weekends	Fines	Meat	Uniform	Logos	Pathos	
Q_1 (effective to oneself)	Without	3.95	3.47	3.86	3.79	**3.70**	3.96	3.54	3.75
	With	**4.05**	**3.76**	**4.21**	**3.90**	3.69	**4.05**	**3.80**	**3.92**
Q_2 (effective to others)	Without	3.47	3.33	3.72	**3.74**	**3.70**	3.79	3.39	3.59
	With	**3.79**	**3.60**	**3.86**	3.57	3.62	**3.83**	**3.55**	**3.69**
Q_3 (reasonableness)	Without	**4.23**	3.51	4.07	**4.14**	**3.93**	**4.20**	3.75	3.98
	With	4.21	**3.83**	**4.17**	3.81	3.86	4.13	**3.82**	3.98

Table 2: Main study results: Mean ratings (scale 1–5, higher is better) of the arguments across the five *topics*, the two rhetorical *strategies*, and *overall* for the three questions on the persuasive effect and reasonableness, Q_1–Q_3. Bold face indicates whether *without* and *with* image support has the higher rating in each case.

accompanying image] is likely to sway the person towards its position?

1. Not at all likely to sway
2. Rather not likely to sway
3. I really cannot judge
4. Rather likely to sway
5. Highly likely to sway

Q3. Reasonable: Irrespective of your personal or others' views, to what extent would you think the argument [along with the accompanying image] is reasonable for the given topic?

1. Not at all reasonable
2. Rather not reasonable
3. Partly reasonable, partly not
4. Rather reasonable
5. Highly reasonable

The phrase in brackets appeared only in TI-items. The rationale for posing the three different questions was to assess whether the responses differed when trying to eliminate personal "taste". For example, when a participant finds, an argument and/or image unattractive but is aware this is probably just their own subjective judgment, they could provide different ratings for the three questions.

Overall, with this setup, we aimed to test whether the presence of a carefully selected and rhetorically-aligned image could enhance the persuasive effect of a short argument. By working with five different topics, we mitigate the effect of potential biases that participants might have on specific topics/issues.

5.2. Results

Table 2 reports the mean scores separately for Q_1–Q_3. In each case, we show the results per topic, per rhetorical strategy, and overall—once without and once with image support. This way, we not only see overall trends, but also more detailed effects, such as whether logos arguments with images were rated more effective than those without, or how specific topics are affected by visual support.

Overall Across the three questions, arguments presented with images generally received higher or equal mean scores compared to those without images, suggesting a modest but consistent positive effect of image support on the impact of arguments. The difference is biggest for Q1 (effective to oneself), where arguments with images scored 3.92, compared to 3.75 without. While it is a bit smaller for Q2 (effective to others), with 3.69 vs. 3.59, for Q3 (reasonable), both conditions received the same mean score of 3.98. This might suggest that image support really has a rhetorical rather than a dialectical function.

Per Topic Image-supported arguments scored higher in most cases, but with variations. In case of *weekends* and *fines*, they consistently outperformed those without, across all three questions. In some cases, the difference was large, as for *fines* on Q1 (4.21 vs. 3.86). For *waste* and *meat*, partly the images scored better, partly the text-only arguments. Only for school *uniforms*, text-only arguments scored marginally higher across all questions (e.g., 3.70 vs. 3.62 in case of Q2).

Per Strategy First of all, when comparing the two strategies, we observe that logos-oriented arguments consistently scored higher than pathos-oriented ones, both with and without images. In case of Q1, for example, with image we see a value of 4.05 for logos compared to 3.80 for pathos, and without image 3.96 compared to 3.54.

However, the effect of adding an image was different for the two rhetorical strategies: Pathos arguments benefited from images across all questions, with higher scores in the image condition compared to text-only. For logos arguments, the impact was mixed: images led to slightly higher scores in Q1 and Q2, indicating improved persuasiveness. However, in Q3, the version with an image scored lower (4.13) than the version without (4.20), suggesting that the image may have slightly weakened the perceived reasonableness of the argument.

We calculated statistical significance of our comparisons using an independent t-test. For most differences, we could not prove significance, likely attributed to the limited scale of our study. An exception is the strategy dimension (logos vs. pathos), where the differences between the settings with and without image are all significant at $p < .05$.

5.3. Discussion

The results suggest that, while the inclusion of images overall tends to improve the perceived impact of arguments, the gain is not dramatic and varies across topics and rhetorical strategies. Pathos-oriented arguments benefitted more consistently from visual support, showing gains across all three evaluated questions. In contrast, logos-oriented arguments, although rated higher overall, showed a drop in perceived reasonableness (Q3) when accompanied by an image, suggesting that logical arguments may suffer from visuals that do not align with their tone or content.

A topic-level analysis further highlights this sensitivity. For 'school *uniforms*', participants rated the text-only versions higher than image-supported ones throughout the three questions, while 'supermarkets operating on *weekends*' and '*fines* for dog owners' show the opposite situation. We hypothesize that these differences may in part be due the particular style of the school uniform images (see Figure 1), which can be perceived as more "exaggerated" than the images we used for other topics. We leave the assessment of such effects (general appeal of an image, irrespective of its use context), as well as a more detailed analysis of argument-image alignment, to future work. In any case, the exceptions from the general trend indicate that images can at times detract from an argument's strength if they add noise or distract from the core message of the argument.

6. Conclusion

To the best of our knowledge, we conducted the first systematic study of pairing AI-generated text and images, and of assessing the ensuing persuasive effects. Using five different topics and two rhetorical strategies (logos, pathos), we find that visual support, particularly when emotionally aligned, can modestly improve the perceived impact. However, this effect is neither universal nor uniformly positive; it is contingent on both the rhetorical strategy and the semantic fit between image and text.

Pathos-oriented arguments benefitted most consistently from the addition of images, while logos-style arguments appeared more vulnerable to disruption when paired with ill-fitting visuals. In particular, participants rated such arguments as less reasonable when the image failed to reinforce the logical tone, as observed in neutral-view evaluations (Q3).

For two of the five topics, image inclusion occasionally led to lower ratings in reasonableness or others' perspective, showing that mismatched or distracting visuals can undercut an argument's effectiveness. We conclude that images must not only be relevant to the topic but also align well with the rhetorical and semantic content of the argument. We intend to study this relationship in more depth in our future work.

Another general takeaway from the study is that AI image generation already has the potential to be used, with relatively straightforward prompting strategies, in tasks where persuasive effects are at stake. Ultimately, this work contributes to the growing understanding of visual-verbal interaction in persuasive discourse, highlighting both the potential and the limitations of visual augmentation in argument generation.

One of the potentially interesting steps for future work is to assess specifically the effects of AI-generated images, by comparing them to "normal" images playing the same role alongside the texts. A hypothesis here may be that when audiences have become used to seeing lots of AI-generated images – and to recognizing them as such – the AI-generated images are less prone to be convincing than the "natural" images can be. We leave this to future work.

7. Limitations

While our study is an important first step toward in-depth analysis of persuasive effects of text/image analysis, it has some limitations. Image selection was done by rating candidate images for their alignment with topic, stance and rhetorical strategy, but not for their suitability to the specific argumentative text; this aspect was handled in the filtering stage by the authors of the study. In follow-up experiments, it can be done by recruited participants (which, however, requires finding a design that does not pre-empt the "main" experiment).

The AI-generated argumentative texts have been used in the experiment "as is"; the authors of the paper judged them as well-suited. This pre-check, targeting coherence, naturalness, and argumentative adequacy, could also be done in a separate study with participants, to reduce the influence of our subjective judgement.

8. Ethical Considerations

As is well-known, AI-generated text and images can in principle contain material that audience members may consider as offensive or inappropriate. We screened all our material for such aspects and did not find any. Also, the participants of our two experiments did not report any concerns.

Our pre-study was conducted with members of our research labs who volunteered to participate in the experiment. For the second experiment, we aimed to provide a fair compensation to the crowdworkers (see the description in Sct. 5). The hourly rate we used to calculate payment is well above the UK minimum wage; Prolific states it is the "average" payment made in such studies. Duration of the experiment (the second factor for payment calculation) was estimated on the basis of our previous experiences with several similar text understanding experiments.

Most importantly: Pairing automatic text and image generation with the notion of persuasion is a sensitive matter *Disinformation* is, by definition, being produced in order to influence audiences, which also includes the goal of persuasion. We are well-aware that there is a "dual use" issue for the type of text/image production that we are studying. The aim of our research is to better understand the cognitive mechanisms of "convincingness", to advance knowledge on how human readers react to differently-framed information. Being aware of these mechanisms is an important precondition for recognizing and trying to combat disinformation. We strongly denounce the use of technologies similar to those that we study for the production of harmful content of any kind.

9. Acknowledgments

We thank the 30 master's and PhD students from University of Potsdam and Leibniz University Hannover for their contribution to the pre-study. Furthermore, we thank the anonymous reviewers for their helpful feedback. This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under project number TRR 318/3 2026 – 438445824.

10. Bibliographical References

Khalid Al Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443. The COLING 2016 Organizing Committee.

Milad Alshomary, Roxanne El Baff, Timon Gurcke, and Henning Wachsmuth. 2022. The moral debater: A study on the computational generation of morally framed arguments. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8782–8797, Dublin, Ireland. Association for Computational Linguistics.

Pranav Anand, Joseph King, Jordan Boyd-Graber, Earl Wagner, Craig Martell, Doug Oard, and Philip Resnik. 2011. Believe me: we can do this! annotating persuasive acts in blog text. In *Proceedings of the 10th AAAI Conference on Computational Models of Natural Argument*, AAAIWS'11-10, page 11?15. AAAI Press.

Aristotle. ca. 350 B.C.E./ translated 2007. *On Rhetoric: A Theory of Civic Discourse*. Oxford University Press, Oxford, UK. Translated by George A. Kennedy.

Miriam Louise Carnot, Lorenz Heinemann, Jan Braker, Tobias Schreieder, Johannes Kiesel, Maik Fröbe, Martin Potthast, and Benno Stein. 2023. On stance detection in image retrieval for argumentation. In *46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023)*, pages 2562–2571. ACM.

Jiwan Chung, Sungjae Lee, Minseo Kim, Seungju Han, Ashkan Yousefpour, Jack Hessel, and Youngjae Yu. 2024. Selective vision is the challenge for visual reasoning: A benchmark for visual argument understanding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2423–2451, Miami, Florida, USA. Association for Computational Linguistics.

Rory Duthie, Katarzyna Budynska, and Chris Reed. 2016. Mining ethos in political debate. In *Proceedings of the Sixth International Conference on Computational Models of Argument*, pages 299–310.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, Manfred Stede, and Benno Stein. 2019. Computational argumentation synthesis as a language modeling task. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 54–64, Tokyo, Japan. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464. Association for Computational Linguistics.

Roxanne El Baff, Henning Wachsmuth, Khalid Al Khatib, and Benno Stein. 2020. Analyzing the persuasive effect of style in news editorial argumentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3154–3160, Online. Association for Computational Linguistics.

Johannes Kiesel, Nico Reichenbach, Benno Stein, and Martin Potthast. 2021. Image retrieval for arguments using stance-aware query expansion. In *Proceedings of the 8th Workshop on Argument Mining*, pages 36–45, Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Zhexiong Liu, Mohamed Elaraby, Yang Zhong, and Diane Litman. 2023. Overview of ImageArg-2023: The first shared task in multimodal argument mining. In *Proceedings of the 10th Workshop on Argument Mining*, pages 120–132, Singapore. 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.

Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 742–753, Valencia, Spain. Association for Computational Linguistics.

Eleonora Mancini, Federico Ruggeri, Stefano Colamonaco, Andrea Zecca, Samuele Marro, and Paolo Torroni. 2024. MAMKit: A comprehensive multimodal argument mining toolkit. In *Proceedings of the 11th Workshop on Argument Mining (ArgMining 2024)*, pages 69–82, Bangkok, Thailand. Association for Computational Linguistics.

Lily Ng, Anne Lauscher, Joel Tetreault, and Courtney Napoles. 2020. Creating a domain-diverse corpus for theory-based argument quality assessment. In *Proceedings of the 7th Workshop on Argument Mining*, pages 117–126, Online. Association for Computational Linguistics.

Daniel J. O'Keefe. 2002. *Persuasion: Theory and research*, 2nd edition. Sage Publications.

OpenAI. 2023. GPT-4 technical report.

Andreas Peldszus and Manfred Stede. 2016. An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon 2015 / Vol. 2*, pages 801–816. College Publications, London.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation.

Gabriella Skitalinskaya, Jonas Klaff, and Henning Wachsmuth. 2021. Learning from revisions: Quality assessment of claims in argumentation at scale. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1718–1729, Online. Association for Computational Linguistics.

Manfred Stede, Jodi Schneider, and Henning Wachsmuth. 2026. *Argument Mining (Second Edition)*. Synthesis Lectures on Human Language Technologies. Springer, Cham.

Assaf Toledo, Shai Gretz, Edo Cohen-Karlik, Roni Friedman, Elad Venezian, Dan Lahav, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Automatic argument quality assessment - New datasets and methods. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5625–5635. Association for Computational Linguistics.

Stephen Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge.

Frans H. van Eemeren and Rob Grootendorst. 2004. *A Systematic Theory of Argumentation: The Pragma-Dialectical Approach*. Cambridge University Press, Cambridge, UK.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187. Association for Computational Linguistics.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation synthesis following rhetorical strategies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3753–3765. Association for Computational Linguistics.

D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation schemes*. Cambridge University Press, Cambridge.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

A. Appendix A: Generated Texts for Waste Segregation

Logos-Oriented Arguments in Favor

- Resource Efficiency – Proper waste segregation ensures that recyclable materials, such as paper, plastics, and metals, are processed correctly, reducing the demand for raw materials. This conserves natural resources and decreases energy consumption in manufacturing.

- Economic Benefits – Recycling and composting industries create jobs and generate revenue. Countries with effective waste management systems have lower disposal costs and higher material recovery rates, leading to economic growth.

- Reduced Environmental Impact – Segregated waste minimizes landfill overflow and prevents hazardous materials from contaminating soil and water. This leads to a healthier environment, reducing long-term cleanup and health costs.

Logos-Oriented Arguments against

- High Implementation Costs – Establishing an efficient waste segregation system requires investments in infrastructure, specialized collection vehicles, and public awareness campaigns, which can be financially burdensome, especially for developing regions.

- Compliance Issues – Many individuals and businesses do not consistently follow waste segregation guidelines, leading to contamination of recyclable materials. This reduces the efficiency of recycling programs and increases processing costs.

- Limited Recycling Feasibility – Not all materials can be effectively recycled due to mixed compositions, contamination, or lack of processing facilities. In some cases, segregating waste does not lead to meaningful environmental or economic benefits.

Pathos-Oriented Arguments in Favor

- Protecting Our Children's Future – Imagine your children growing up in a world filled with overflowing landfills and polluted oceans. By simply segregating waste, we ensure a cleaner, safer environment for future generations to thrive.

- Respect for Our Communities – No one wants to walk through streets littered with trash or breathe in the stench of rotting waste. Proper segregation keeps our neighborhoods clean and dignified, showing respect for those around us.

- A Small Act with a Big Impact – It takes just a few seconds to separate recyclables, yet this small effort can prevent tons of waste from destroying ecosystems. Knowing that you're making a difference can bring a sense of pride and responsibility.

Pathos-Oriented Arguments against

- Burden on Everyday People – Life is already stressful with work, family, and responsibilities. Adding the extra task of sorting waste feels like an unfair burden, especially when large corporations produce far more waste than individuals.

- Frustration with Inefficiency – Many people carefully separate their waste only to see it all mixed together at collection sites. If the system is flawed, why should individuals bear the emotional guilt of improper disposal?

- Feeling of Pointlessness – Despite individual efforts, global pollution and environmental destruction continue. It's disheartening to see that no matter how much waste is segregated, the larger problem remains unchanged.