Illustrating Arguments with Images Using Aspect-Aware Prompting

Maximilian Heinrich
Bauhaus-Universität Weimar
Weimar, Germany
maximilian.heinrich@uni-weimar.de

Sharat Anand
Bauhaus-Universität Weimar
Weimar, Germany
sharat.anand@uni-weimar.de

Johannes Kiesel
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany
johannes.kiesel@gesis.org

Benno Stein
Bauhaus-Universität Weimar
Weimar, Germany
benno.stein@uni-weimar.de

Abstract

Images can powerfully strengthen arguments, conveying ideas more immediately and compellingly than text alone. With the rise of text-to-image models, a broad audience can now generate custom visuals to illustrate their arguments. Yet a fundamental mismatch undermines this potential: these models are trained on concrete scene descriptions, while arguments operate at the level of general, abstract principles. Naively prompting such a model with an argumentative text therefore rarely produces images that genuinely illustrate the argument. To address this challenge, we propose an aspect-aware image generation approach. Given an argument, our method first identifies the key aspects that an illustrative image should convey, then constructs a detailed scene description grounded in both the argument and those aspects, and finally generates an image using that scene description as the prompt. A human-assessment evaluation demonstrates that this approach yields images that illustrate arguments significantly better than those produced by naive prompting.

1 Introduction

Images are ubiquitous in contemporary life. From social media and news feeds to advertising billboards and political campaigns, images have become a prominent mode of communication and argumentation (Groarke, 2024). Images can help to understand and retain complex information (Clark and Paivio, 1991). Moreover, images can evoke strong emotional responses—both positive and negative—and suggest or motivate particular actions (Freedberg, 1989; Dunaway, 2018). In advertising, political campaigns, and public discourse, images are frequently used to support judgments, justify positions, and motivate action. In this sense, they serve as visual arguments that can be as persuasive as their verbal counterparts (Groarke, 1996). However, until recently, most people could not illustrate their arguments with specially tailored images due to the high costs of generating images.[1]

Despite recent advances in text-to-image generation, the generation of images that effectively convey arguments remains a challenge. Whereas arguments are typically expressed at an abstract and conceptual level, text-to-image models are trained on concrete descriptions of scenes involving identifiable objects, actions, and settings. As a result of this mismatch, images generated with text-to-image models using an argument as a literal prompt are incomplete, inconsistent, or unrelated to the intended claim. The columns "Argument as literal prompt" in Table 1 (on the right-hand side) exemplify this problem: key aspects of the arguments are missing in the respective images.

To address this challenge, we propose aspect-aware prompt generation for generating images that illustrate arguments. The proposed approach comprises three steps from an argument to the image: (1) identify the argument's key aspects—visually interpretable words and phrases that capture the key ideas of an argument; (2) generate a descriptive scene-level prompt from the argument and the identified aspects; and (3) generate the image using the generated prompt. Column "Aspect-aware

[1] Internet memes are somewhat of an exception, as they – to put it simply – allow a generic visual pattern to be adapted to a specific argument. But memes cannot cover all arguments.

Proceedings of the 13th Workshop on Argument Mining and Reasoning, pages 72–85

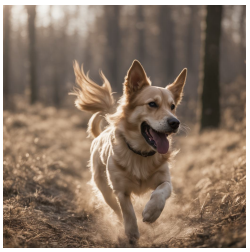
Argument	Aspect-aware prompt	Argument as literal prompt	
Sugar makes food more enjoyable.			
Dogs require regular outdoor activity which promotes a healthier lifestyle for their owners.			
Risk of crime is higher in public transport.			
	Stable Diffusion 1.0 XL	Stable Diffusion 1.0 XL	Stable Diffusion 3.5 Medium

Table 1: Turning an argument into an image: Using the literal argument (first column) as an image prompt results in images that do not reflect the essence of the argument (third and fourth columns), while the second column shows the generated images for which the prompt was created using our aspect-aware approach to prompt generation. Table 2 (third column) shows the prompts that were generated. The image generation models are Stable Diffusion 1.0 XL (second and third columns) and Stable Diffusion 3.5 Medium (fourth column).

prompt" in Table 1 exemplifies the images created for the different arguments mentioned above, showing a clear improvement over the images generated with naive prompting.[2]

This paper is structured as follows: Section 3 discusses the task and challenges of image generation to illustrate arguments in detail. Section 4 presents a new dataset of arguments and associated images for evaluating image generation algorithms. Section 5 introduces seven automatic quality indicators capturing different properties of prompts and generated images—such as imageability—and uses them alongside human assessment to compare images generated from literal prompts against those generated with aspect-aware prompts. The section further evaluates whether the automatically identified aspects align with those identified by humans. Section 6 then discusses the results of the experiments. Our experiments show that prompts derived from argument aspects produce images that convey the argument more effectively than prompts generated directly from the literal argument text.

[2]Code and data available at: https://github.com/webis-de/ArgMining-26

2 Related Work

The persuasive and cognitive impact of visuals has been documented across numerous domains. A historical overview of the influence of images on human emotions and societal practices is provided by Freedberg (1989). For comprehensive surveys of text-to-image generation methods, we refer the reader to Alhabeeb and Al-Shargabi (2024); Bie et al. (2025). The capability of generative models to respond to and interpret different prompts has been systematically investigated by Liu and Chilton (2022), who analyze how prompt design influences the quality and characteristics of generated images.

Beyond standard generation tasks, image generation has also been employed as an intermediate reasoning step for language models solving complex inferential problems (Chern et al., 2025). Related

work further investigates to what extent models can identify visual elements that are relevant to an argument. For instance, Chung et al. (2024) introduce a benchmark for visual argument understanding that reveals selective vision—identifying which visual elements are relevant to an argument—as a key challenge for current models.

At the intersection of argumentation and visualization, one line of research focuses on detecting persuasive techniques in multimodal content such as memes (Dimitrov et al., 2021). Another thread concerns the visual representation of argumentative structures: Khartabil et al. (2021) design visualization techniques for exploring individual argument structures, while Kiesel (2025) develops visual analytics approaches for argumentation at the corpus level. A related direction addresses the generation of complex statistical infographics from unstructured text, requiring models to plan chart types, layouts, and visual structure from textual content alone (Ghosh et al., 2025).

The most direct precedent for our work is the Touché Shared Task on Argument Image Retrieval and Generation (Kiesel et al., 2025, 2024; Bondarenko et al., 2023, 2022), which asks participants to retrieve or generate images that effectively convey a given argument. Existing approaches focus almost exclusively on retrieval from a predefined image pool, with generative models playing only an auxiliary role — for instance, to support ranking via comparison with generated images (Moebius et al., 2023) or to produce synthetic training data for retrieval models (Janusko et al., 2024). We participated in the most recent edition of this shared task with the aspect-aware generation approach presented in this paper (Kiesel et al., 2025; Anand and Heinrich, 2025). However, our shared-task contribution is limited to a system description and does not include a systematic evaluation of the approach — a gap that the present paper addresses.

3 From Arguments to Image Prompts

This section discusses the conceptual foundations of generating images for arguments, first examining the relationship between images and arguments, and then contrasting abstract argumentative statements with concrete scene descriptions.

Images and Arguments It is a fundamental characteristic of images that they are ambiguous yet rich in information (Kjeldsen, 2015). Consider a person standing on a ladder. An image of this situation conveys far more details than a verbal description, such as the person's appearance, the surrounding environment, or the color and material of the ladder. While visual richness provides extensive information, it also introduces ambiguity: it may be unclear whether the person is climbing up or down the ladder, leaving the situation open to interpretation. Another challenge concerns the interpretation of symbolic elements in images. Understanding such symbols often depends on cultural and contextual knowledge that is not explicitly represented in the image itself. Because images are rich in detail yet ambiguous, their meaning is often guided by accompanying textual elements that help anchor the viewer's interpretation and direct attention toward the intended message (Kjeldsen, 2015). The ambiguity of images raises the question of whether visuals can function as arguments at all. Critics argue that identifying argument structures in visuals—such as premise-conclusion relations—depends on background knowledge and interpretive processes that are fundamentally verbal in nature (Johnson, 2003). Others argue that similar ambiguity is also present in textual arguments (Groarke, 1996), and therefore visual arguments can also exist. Still others contend that visual representations can function as evidence supporting a claim (Dove, 2012).

Moreover, visualizing an argument does not necessarily mean producing a single image. Arguments may naturally evoke a sequence of mental images — for example, the claim *"Jack missed the shot on goal because he was blinded by sunlight"* is most readily understood as a dynamic situation unfolding across multiple moments. While Champagne and Pietarinen (2020) suggest tackling this with a sequence of visual steps that together convey premises and conclusions, in this work we focus on the single-image approach.

Abstract Statements versus Scene Descriptions Modern text-to-image models are trained on large datasets of image-caption pairs. Captions describe the visible content of an image, i.e., concrete scenes. Consequently, the models primarily learn a mapping between scene descriptions and visual representations (Radford et al., 2021; Segalis et al., 2023). A scene depicts a specific situation, assembled from an actor, an action, specific objects, and visible emotions. An argument, on the other hand, is not a scene description but often a rule, a proverb, or an abstract statement; it is therefore generic (timeless), does not describe a specific situ-

Argument	Identified aspects	Aspect-aware prompt
Sugar makes food more enjoyable.	sugar, enjoyment, eating	A young child is happily eating a bowl of ice cream, savoring every delicious bite. The child's eyes light up with joy as they taste the sweet, creamy goodness of the ice cream, enjoying the sweet, sugary taste that brings them pleasure and happiness.
Dogs require regular outdoor activity which promotes a healthier lifestyle for their owners.	outdoor activity, owner, healthy lifestyle	A playful dog and its owner are enjoying a brisk walk through a park on a sunny day, both of them smiling and full of energy. The dog is wearing a colorful bandana and the owner has a water bottle in their hand. The park is filled with people enjoying the beautiful weather, and there are benches and picnic tables throughout the park for people to relax.
Risk of crime is higher in public transport.	crime, public transport, risk	A crime scene where a violent crime has taken place in a public transport, with police officers investigating the area and taking statements from witnesses.

Table 2: From left to right: The original argument, the identified aspects, and the aspect-aware prompt describing a scene. Aspect identification and prompt generation are performed using LLaMA-3.2-3B-Instruct and Mistral-7B-v0.1, respectively. Using the aspect-aware prompts, the images in Table 1 (second column) were generated.

ation, and lacks identifiable actors.

This difference has direct implications for the concept of "visualizability", which can be defined as follows:

> Visualizability is the degree to which a linguistic description can be directly translated into a concrete visual scene. It depends on referentiality, concreteness, event structure, actor-object relationships, and spatial information.

In our approach we increase the visualizability of the original argument by reducing its ambiguity and subjectivity. For this purpose we identify a small number of aspects (4–8) for a given argument and, based on the two most salient aspects, formulate a scene description as a prompt that conveys the essence of the argument. We refer to our approach as "aspect-aware image generation". Table 2 gives examples for arguments, the identified aspects, and the generated prompts with which the images in Table 1 were generated.

The fact that more concrete descriptions lead to clearer and more coherent visual results is also confirmed by Liu and Chilton (2022).

4 Dataset Construction

To evaluate argumentative image generation, we construct a dataset of generated images for argumentative claims. The claims are sourced from an existing dataset by Heinrich et al. (2025). Each claim is a standalone assertion without supporting premises (we therefore use claim and argument interchangeably throughout), making it well suited for representation in a single image. The claims are argumentative in that they function as enthymemes: arguments whose supporting premises are implicit and reconstructable by the audience from common knowledge (Walton, 2008).

For instance, the claim *"Sugar makes food more enjoyable"* rests on unstated premises such as *sugar makes food sweeter* and *sweeter food is more enjoyable*. This enthymematic structure is characteristic of nearly all claims in our dataset, which range from causal assertions (*"Automation increases work efficiency"*) to evaluative judgments (*"Street art beautifies urban areas"*). It is also what makes image generation challenging: rather than depicting a literal scene description, the generated image must visually convey the implicit reasoning behind the claim—precisely the gap that our aspect-aware approach addresses.

We generate images for each claim using three approaches: Stable Diffusion XL Base 1.0 (Rombach et al., 2022), Stable Diffusion 3.5 Medium (Esser et al., 2024), both of which take the claim directly as input, and our aspect-aware generation approach, which proceeds in three stages. First, five candidate aspects are automatically identified from each argument using LLaMA 3.2 (3B-Instruct) (Grattafiori et al., 2024). Next, a human annotator selects the three aspects that best correspond to the claim. These selected aspects are then passed to Mistral (7B-v0.1) (Jiang et al., 2023) to generate a detailed image prompt, which Stable Diffusion XL Base 1.0 subsequently uses to produce the final image. The choice of models for each stage was guided by preliminary experiments during development. The prompts used for identification and generation are provided in Appendix A. After generation, 127 claims covering 27 topics remain in our dataset.

	Human assessment	
Interval	Aspects	Images
[0.0; 0.5)	429 (11%)	68 (4%)
[0.5; 1.0)	314 (8%)	141 (7%)
[1.0; 1.5)	631 (17%)	582 (31%)
[1.5; 2.0]	2,436 (64%)	1,114 (58%)
Σ	**3,810 (100%)**	**1,905 (100%)**

Table 3: Distribution of the human assessment scores across 1,905 images (127 claims × 3 systems × 5 images). Each image was rated by two annotators on two aspects using a three-point scale (0 = not depicted, 1 = partial, 2 = clear). The two ratings per aspect and the four ratings per image were averaged and assigned to the corresponding interval (first column).

For each argument–generation-approach pair, five images are generated, yielding 1,905 images in total[3].

For each argument, we define two ground-truth aspects that should be identifiable in a generated image to convey the underlying argument. Each image is then checked for the presence of both aspects: every aspect is independently annotated by two annotators using a three-point scale (0–2), where 0 denotes no coverage, 1 denotes partial coverage, and 2 denotes full coverage, resulting in 7,620 annotation points. The final aspect score for an image is computed as the mean of the two annotators' scores, and the final image score is obtained by averaging across the two aspects. Dataset statistics are summarized in Table 3.

Inter-annotator agreement, measured via Cohen's κ at the aspect level, yields $\kappa = 0.34$ — a moderate score that reflects the task's inherent subjectivity. Judging how strongly an image conveys a given argumentative aspect requires interpreting the visual scene in light of the broader claim, rather than applying purely perceptual criteria. Representative annotation examples illustrating this challenge are provided in Appendix B.

5 Experimental Analysis

We conduct two experiments. First, we evaluate our aspect-aware image generation approach against baselines that use the raw argument as a prompt, measuring image quality through human assessment and automated quality indicators. Second, we assess whether the aspects required for aspect-aware generation can be automatically extracted from a given argument.

5.1 Aspect-Aware Image Generation

We evaluate our aspect-aware image generation method against the two Stable Diffusion baselines on our dataset. Representative outputs from all three approaches are shown in Table 1. For each approach, quality indicators are computed per generated image, averaged across the five images generated per argument, and then aggregated across all arguments to obtain the overall score. Results are reported in Table 4. Beyond human assessments, we evaluate image quality using the indicators described below. Since prompt quality directly influences the generated image, two indicators — concreteness and imageability — target the prompt itself; the remaining indicators operate on the generated image. Several of the latter require textual image descriptions, for which we sample five descriptions per image using LLaVA (7B) (Liu et al., 2023) (temperature 0.3) to account for variability in model outputs.

Concreteness Concreteness reflects the degree to which a word refers to a perceptible, physical entity (e.g., *dog* scores high, *justice* scores low), with ratings from 1 (very abstract) to 5 (very concrete). We measure prompt concreteness using the psycholinguistic lexicon of Brysbaert et al. (2014). Since concreteness is defined at the word level, extending it to multi-word prompts is non-trivial (Wu and Smith, 2023). Rather than averaging over all words, we instead aggregate over the five highest-scoring tokens per prompt, focusing the measure on the most content-bearing terms. This is particularly important for aspect-aware prompts, which tend to be longer and thus contain more low-scoring function words (e.g., prepositions) that would otherwise dilute the score.

Imageability Imageability measures how readily a word evokes a mental image (Coltheart, 1981; Wilson, 1988), operationalized by looking up words in a dedicated psycholinguistic lexicon.[4] Although closely related to concreteness, the two constructs can diverge: *infinity*, for instance, may readily evoke a mental image (e.g., of endless space) yet receives a low concreteness score, as it lacks a tangible physical referent. Ratings range from

[3]The images and data were generated during our participation in, and organization of, the Touché shared task (Kiesel et al., 2025; Anand and Heinrich, 2025).

[4]Dataset available at https://huggingface.co/datasets/StephanAkkerman/MRC-psycholinguistic-database

System	Human assessment	Prompt quality		Image quality (SBERT)			Image quality (CLIP)	
		Concreteness	Imageability	Interpret-ability	Prompt fidelity	Aspect coverage	Prompt fidelity	Aspect coverage
Ours	0.800 ±0.15	0.924 ±0.04	0.796 ±0.04	0.924 ±0.01	0.774 ±0.05	0.670 ±0.04	0.658 ±0.01	0.619 ±0.01
SD 1.0	0.668 ±0.20	0.594 ±0.12	0.558 ±0.08	0.922 ±0.01	0.668 ±0.05	0.675 ±0.04	0.644 ±0.01	0.620 ±0.01
SD 3.5	0.671 ±0.22	0.594 ±0.12	0.558 ±0.08	0.918 ±0.01	0.676 ±0.05	0.679 ±0.03	0.645 ±0.01	0.618 ±0.01

Table 4: Comparison of our image generation approach using aspect-aware prompts (row "Ours") with Stable Diffusion 1.0 XL and 3.5 Medium using literal argument prompts (rows "SD 1.0" and "SD 3.5"). The second column (Human assessment) shows the achieved ground-truth scores (cf. the overall distribution in Table 3, third column). Columns 3–9 show algorithmically measured quality indicators for both the prompts and the images. For better readability, all scores have been scaled to a range between 0 and 1. The computation of the quality indicators is explained in the text.

1 (low imageability) to 7 (high imageability). As with concreteness, this measure is defined at the word level; to obtain a single imageability score for a prompt, we aggregate over the five highest-scoring tokens.

Interpretability Interpretability measures how consistently an image is described, estimated via a caption-consistency score inspired by Wu and Smith (2023). We compute pairwise cosine similarity over the five generated image descriptions using Sentence-BERT (SBERT, all-MiniLM-L6-v2) (Reimers and Gurevych, 2019), then average the resulting scores. Values range from -1 to 1, where higher scores indicate that independently generated descriptions converge on similar content, suggesting the depicted image is clear and unambiguous.

SBERT Prompt Fidelity Prompt fidelity measures how well the generated image reflects its prompt. Using SBERT, we compute the cosine similarity between the prompt and each of the five image descriptions and average the resulting scores. Values range from -1 to 1, with higher scores indicating closer alignment.

SBERT Aspect Coverage To evaluate whether the image descriptions capture the underlying ground-truth aspects, we compute the cosine similarity between each aspect and every sentence in the description using SBERT, retaining the maximum similarity across sentences. This ensures that an aspect is credited if any single sentence reflects it, rather than diluting the score by averaging over all words. The per-aspect scores are then averaged across both ground-truth aspects and all five image descriptions to yield a single score per image, ranging from -1 to 1.

CLIP Prompt Fidelity and Aspect Coverage Analogous to their SBERT counterparts, these measures use CLIP (ViT-B/32) embeddings (Radford et al., 2021) and cosine similarity, but operate directly on images rather than on textual descriptions, eliminating the need for an intermediate captioning step. Prompt fidelity is computed as the cosine similarity between the generated image embedding and the prompt embedding, following the image–text alignment paradigm of Hessel et al. (2021). Since CLIP truncates text inputs exceeding 77 tokens, longer aspect-aware prompts are partially cut off, which may reduce the measured prompt fidelity for these prompts. Aspect coverage is computed analogously, as the average cosine similarity between the image embedding and the embeddings of each ground-truth aspect.

5.2 Identification of Argument Aspects

We evaluate whether LLMs can automatically recover the ground-truth aspects for a given argumentative claim. To this end, each model is prompted to generate five candidate aspects per claim, and we run each model five times per claim to account for output variability. We experiment with LLaMA 3.2 (3B-Instruct) and GPT-5-Nano (2025-04-07) (Singh et al., 2025). LLaMA is run at temperature 0.3; GPT-5-Nano at its default of 1.0, as the parameter is not user-adjustable. Since a model may return semantically redundant aspects within a single output, we deduplicate the five generated aspects before matching. We report recall over ground-truth aspects — the fraction of the two

Model	Aspect selection	τ=.40	τ=.50	τ=.60	τ=.70
LLaMA 3.2	All	0.909	0.852	0.728	0.595
	Top-2	0.797	0.715	0.607	0.490
GPT-5-Nano	All	0.976	0.964	0.881	0.757
	Top-2	0.859	0.821	0.743	0.617

Table 5: Recall over ground-truth aspects under Hungarian 1-to-1 assignment with SBERT cosine similarity $\geq \tau$, where recall is defined as the fraction of the two ground-truth aspects matched by at least one generated aspect. Results are shown for LLaMA 3.2 (3B-Instruct) and GPT-5-Nano across two aspect selection strategies: "All" considers all five generated aspects, "Top-2" retains only the two most similar to the claim by SBERT cosine similarity. Averaged over five runs.

ground-truth aspects matched by at least one generated aspect — using Hungarian assignment (Kuhn, 1955) to enforce strict one-to-one correspondence between generated and ground-truth aspects. A match is established when SBERT cosine similarity exceeds a threshold. We compare two aspect selection strategies. The first applies Hungarian matching directly against all deduplicated aspects, discarding unmatched ones. The second first reduces the candidate set to the two aspects most similar to the claim by SBERT cosine similarity before matching, constituting a fully automatic setting that requires no human input. Recall across varying thresholds is reported in Table 5.

6 Results and Discussion

Human assessments in Table 4 confirm that the aspect-aware generation approach consistently outperforms both baselines, with Figure 1 providing a holistic view across all quality dimensions. The comparatively high standard deviation of the human scores underscores the subjectivity inherent in evaluating argumentative images, consistent with the moderate inter-annotator agreement reported in Section 4—even though predefined aspects serve as an explicit reference frame for the annotators. Concreteness and imageability scores are notably higher for the aspect-aware approach, suggesting that its prompts more effectively evoke concrete, visually grounded scenes. Regarding interpretability, differences between systems are small, as all three approaches generally produce coherent and interpretable images. This aligns with annotators' qualitative observations: while baseline images are typically understandable, they tend to highlight only a single argumentative aspect while neglecting others, reducing their overall relevance to the underlying claim.

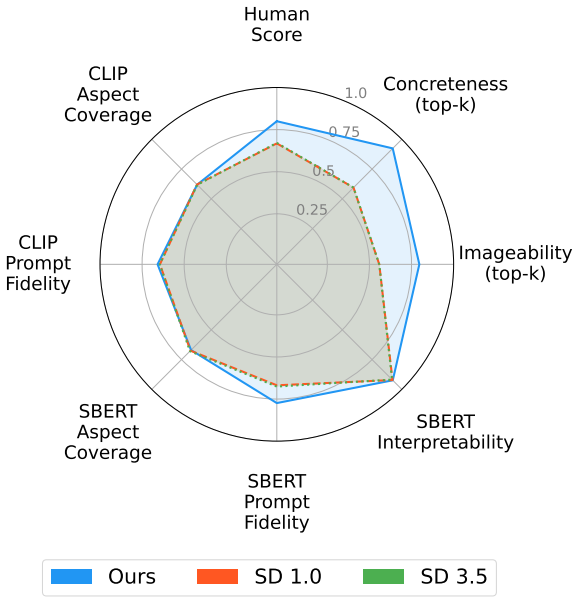


Figure 1: Radar chart comparing aspect-aware generation ("Ours") against Stable Diffusion XL 1.0 ("SD 1.0") and Stable Diffusion 3.5 Medium ("SD 3.5") across eight quality indicators. All scores have been scaled to a range between 0 and 1.

A notable difference is observed for prompt fidelity under both SBERT and CLIP: aspect-aware prompts align considerably more closely with the generated descriptions than raw argumentative claims, confirming that they are more descriptive and yield images that better reflect the intended content. Additionally, CLIP similarity scores tend to cluster within a more compressed range than their SBERT counterparts, a characteristic attributable to the anisotropic nature of CLIP representations (Tyshchuk et al., 2023).

The baseline systems score marginally higher on aspect coverage under SBERT, and under CLIP only Stable Diffusion XL 1.0 achieves a marginally higher score. However, higher coverage does not indicate superior image quality. The reason lies in how the two approaches depict aspects. The baseline uses the claim verbatim as a prompt and tends to depict one concrete aspect literally while neglecting more abstract ones. Consider the claim *"Sugar makes food more enjoyable"* with ground-truth aspects *sugar* and *enjoyable food*: the baseline generates images of candy and sweets whose captions contain words like *candy, sugar, sweet, treats, lollipops* that overlap directly with the *sugar* aspect, yielding high coverage for that aspect, but

largely ignore *enjoyable food*. Since the final coverage score averages over both aspects, the high score on the literally depicted aspect can still outweigh the low score on the neglected one. The aspect-aware approach instead constructs a scene description that integrates both aspects indirectly — for instance, a child happily eating ice cream conveys both *sugar* and *enjoyable food*. The resulting captions (*child, ice cream, bowl, joy, dessert*) are distributionally further from the original aspect terms, so the individual coverage scores tend to be lower, and their average falls below that of the baseline despite arguably conveying the argument more effectively. This reveals a limitation of the coverage metric: it rewards literal depiction of aspect terms over holistic scene descriptions that integrate multiple aspects into a coherent visual narrative.

Table 5 shows that the two ground-truth aspects of each claim are consistently recovered among the five LLM-generated aspects. This holds even under Top-2 selection, where retaining only the two aspects most similar to the claim still yields high recall. Recall increases as the similarity threshold τ becomes more permissive, and across all conditions, GPT-5-Nano outperforms LLaMA 3.2. Notably, even at the strictest threshold ($\tau = .70$), GPT-5-Nano retains strong recall scores of 0.757 (All) and 0.617 (Top-2), while LLaMA 3.2 degrades more sharply—suggesting that GPT-5-Nano produces aspects that align more closely with the ground truth.

7 Conclusion

Generating images for arguments poses a fundamental challenge: arguments are typically abstract and rarely describe concrete scenes, while images are inherently ambiguous and open to interpretation. We address this through aspect-aware image generation, which decomposes the process into three explicit steps: aspect identification, prompt construction, and image generation. This grounds abstract argumentative content in a concrete visual scene, improving visualizability. Experiments on a dataset with human relevance assessments show that aspect-aware image generation consistently outperforms baselines that use raw argument text as a prompt. We further demonstrate that aspect-aware prompts are substantially more concrete and imageable than raw argument text, and that LLM-generated aspects can reliably recover the ground-truth aspects defined in the dataset. In principle, an argument could be illustrated without explicitly identifying aspects — for instance, by directly prompting a language model to generate a scene description from the argument. However, the explicit identification of aspects serves two important functions beyond prompt construction: it provides a controllable interface for steering the generation process toward specific argumentative dimensions, and it establishes interpretable criteria against which the resulting images can be evaluated. Future work investigating alternative decomposition strategies or additional aspect types may yield further insights into how argumentative content can be most effectively translated into visual form, while improved evaluation methods could better capture whether generated images convey argumentative intent holistically rather than through literal depiction of individual aspects.

Beyond improving the illustration of arguments, the explicit modeling of argumentative aspects opens further possibilities. By selecting aspects that emphasize specific persuasive goals, images can be tailored to the characteristics and preferences of a target audience. Another promising direction is a more thorough investigation of automated image evaluation. Building on prior work on argument quality (Wachsmuth et al., 2017), future approaches could assess whether generated images effectively convey dimensions such as credibility or emotional appeal, enabling more systematic analysis of how well images support argumentative goals and guiding the optimization of generation approaches.

A further research direction concerns the illustration of more complex arguments — ones that consist of premises and conclusions rather than a single claim, and that therefore require multiple coordinated images. As discussed in Section 3, Champagne and Pietarinen (2020) suggest that understanding certain arguments requires a sequence of visual steps that together convey premises and conclusions. Generating such image sequences introduces the challenge of maintaining semantic and stylistic consistency across images. Techniques for personalization and concept preservation (Ruiz et al., 2023; Gal et al., 2023) may help ensure cross-image coherence, enabling richer forms of visual argumentation. In such settings, the generation process could be decomposed into a sequence of subgoals, with intermediate visual outputs combined post hoc to construct a coherent argumentative narrative (Chern et al., 2025).

8 Limitations and Ethical Considerations

The subjective nature of images makes standardized evaluation inherently difficult, and our annotation does not assess dimensions such as emotional impact, perceived realism, or persuasive power — all of which are highly relevant to images used in argumentative contexts. Furthermore, the evaluation considers aspects in isolation rather than examining whether multiple aspects are coherently integrated within a single image. Additionally, while our pipeline uses specific models for each stage based on preliminary experiments, a systematic evaluation of how different language and image generation models affect each stage of the pipeline — from aspect identification to prompt construction to image generation — remains an open question.

Regarding image generation, potential biases in generated images must also be acknowledged: generative models are trained on large-scale datasets that reflect cultural and societal influences, and are therefore far from neutral (Morales et al., 2025). More broadly, images function as powerful communicative tools: they are often perceived as direct reflections of reality, lending them an appearance of inherent truth and credibility (Grancea, 2017). Synthetic or manipulated images can therefore destabilize established notions of visual evidence and trust (Momeni, 2025), particularly when deployed to frame events or reinforce specific interpretations of complex issues. As generative models make the production of persuasive visual content increasingly accessible, the generation and use of argument-supporting images demand careful and responsible application.

Acknowledgments

This work was supported by the German Federal Ministry of Research, Technology and Space (BMFTR) through the project "DIALOKIA: Überprüfung von LLM-generierter Argumentation mittels dialektischem Sprachmodell" (01IS24084A-B).

References

Sarah K. Alhabeeb and Amal A. Al-Shargabi. 2024. Text-to-image synthesis with generative models: Methods, datasets, performance metrics, challenges, and future direction. *IEEE Access*, 12:24412–24427.

Sharat Anand and Maximilian Heinrich. 2025. Hanuman at Touché: Image Generation with Argument-Aspect Fusion. In *Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025*, volume 4038 of *CEUR Workshop Proceedings*, pages 4571–4579. CEUR-WS.org.

Fengxiang Bie, Yibo Yang, Zhongzhu Zhou, Adam Ghanem, Minjia Zhang, Zhewei Yao, Xiaoxia Wu, Connor Holmes, Pareesa Ameneh Golnari, David A. Clifton, Yuxiong He, Dacheng Tao, and Shuaiwen Leon Song. 2025. Renaissance: A survey into AI text-to-image generation in the era of large model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 47(3):2212–2231.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Ferdinand Schlatt, Valentin Barrière, Brian Ravenet, Léo Hemamou, Simon Luck, Jan Heinrich Reimer, Benno Stein, Martin Potthast, and Matthias Hagen. 2023. Overview of Touché 2023: Argument and Causal Retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 14th International Conference of the CLEF Association, CLEF 2023, Thessaloniki, Greece, September 18-21, 2023, Proceedings*, volume 14163 of *Lecture Notes in Computer Science*, pages 507–530. Springer.

Alexander Bondarenko, Maik Fröbe, Johannes Kiesel, Shahbaz Syed, Timon Gurcke, Meriem Beloucif, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2022. Overview of touché 2022: Argument retrieval. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5-8, 2022, Proceedings*, volume 13390 of *Lecture Notes in Computer Science*, pages 311–336. Springer.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, 46(3):904–911.

Marc Champagne and Ahti-Veikko Pietarinen. 2020. Why images cannot be arguments, but moving ones might. *Argumentation*, 34(2):207–236.

Ethan Chern, Zhulin Hu, Steffi Chern, Siqi Kou, Jiadi Su, Yan Ma, Zhijie Deng, and Pengfei Liu. 2025. Thinking with generated images. arXiv 2505.22525.

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, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 2423–2451. Association for Computational Linguistics.

James M. Clark and Allan Paivio. 1991. Dual coding theory and education. *Educational Psychology Review*, 3(3):149–210.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 33A(4):497–505.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Semeval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of the 15th International Workshop on Semantic Evaluation, SemEval@ACL/IJCNLP 2021, Virtual Event / Bangkok, Thailand, August 5-6, 2021*, pages 70–98. Association for Computational Linguistics.

Ian J. Dove. 2012. On Images as Evidence and Arguments. In Frans H. van Eemeren and Bart Garssen, editors, *Topical Themes in Argumentation Theory: Twenty Exploratory Studies*, Argumentation Library, pages 223–238. Springer Netherlands, Dordrecht.

Finis Dunaway. 2018. Images, Emotions, Politics. *Modern American History*, 1(3):369–376.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 12606–12633. PMLR / OpenReview.net.

David Freedberg. 1989. *The Power of Images: Studies in the History and Theory of Response*. University of Chicago Press, Chicago.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. 2023. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Akash Ghosh, Aparna Garimella, Pritika Ramu, Sambaran Bandyopadhyay, and Sriparna Saha. 2025. Infogen: Generating complex statistical infographics from documents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 20552–20570. Association for Computational Linguistics.

Ioana Grancea. 2017. Types of Visual Arguments. *Argumentum. Journal of the Seminar of Discursive Logic, Argumentation Theory and Rhetoric*, 15(2):16–34.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Leo Groarke. 1996. Logic, art and argument. *Informal Logic*, 18:105.

Leo Groarke. 2024. Informal Logic. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2024 edition. Metaphysics Research Lab, Stanford University.

Maximilian Heinrich, Johannes Kiesel, Moritz Wolter, Martin Potthast, and Benno Stein. 2025. Touché25: Image Retrieval and Generation for Arguments.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 7514–7528. Association for Computational Linguistics.

Tamás Janusko, Aaron Kämpf, Denis Keiling, Jessica Knick, David Schäfer, and Maik Thiele. 2024. HTW-DIL at Touché: Multimodal Dense Information Retrieval for Arguments. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024*, volume 3740 of *CEUR Workshop Proceedings*, pages 3401–3406. CEUR-WS.org.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Ralph H. Johnson. 2003. Why "visual arguments" aren't arguments. In *OSSA Conference Archive*.

Dana Khartabil, Christopher Collins, S. Wells, Benjamin Bach, and Jessie Kennedy. 2021. Design and evaluation of visualization techniques to facilitate argument exploration. *Comput. Graph. Forum*, 40(6):447–465.

Dora Kiesel. 2025. *Effective Visual Analytics for Exploring Argumentation and Deliberation in Text Corpora*. Ph.D. thesis, Bauhaus-Universität Weimar, Weimar.

Johannes Kiesel, Çagri Çöltekin, Marcel Gohsen, Sebastian Heineking, Maximilian Heinrich, Maik Fröbe, Tim Hagen, Mohammad Aliannejadi, Sharat Anand, Tomaz Erjavec, Matthias Hagen, Matyás Kopp, Nikola Ljubesic, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevicius, Harrisen Scells, Moritz Wolter, Ines Zelch, and 2 others. 2025. Overview of Touché 2025: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain,*

September 9-12, 2025, Proceedings, volume 16089 of *Lecture Notes in Computer Science*, pages 486–508. Springer.

Johannes Kiesel, Çagri Çöltekin, Maximilian Heinrich, Maik Fröbe, Milad Alshomary, Bertrand De Longueville, Tomaz Erjavec, Nicolas Handke, Matyás Kopp, Nikola Ljubesic, Katja Meden, Nailia Mirzakhmedova, Vaidas Morkevicius, Theresa Reitis-Münstermann, Mario Scharfbillig, Nicolas Stefanovitch, Henning Wachsmuth, Martin Potthast, and Benno Stein. 2024. Overview of Touché 2024: Argumentation Systems. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II*, volume 14959 of *Lecture Notes in Computer Science*, pages 308–332. Springer.

Jens E. Kjeldsen. 2015. The Rhetoric of Thick Representation: How Pictures Render the Importance and Strength of an Argument Salient. *Argumentation*, 29(2):197–215.

H. W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Vivian Liu and Lydia B. Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *CHI '22: CHI Conference on Human Factors in Computing Systems, New Orleans, LA, USA, 29 April 2022 - 5 May 2022*, pages 384:1–384:23. ACM.

Max Moebius, Maximilian Enderling, and Sarah T. Bachinger. 2023. Jean-Luc Picard at Touché 2023: Comparing Image Generation, Stance Detection and Feature Matching for Image Retrieval for Arguments. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 3111–3118. CEUR-WS.org.

Mina Momeni. 2025. Artificial intelligence and political deepfakes: Shaping citizen perceptions through misinformation. *Journal of Creative Communications*, 20(1):41–56.

Sergio Morales, Robert Clarisó, and Jordi Cabot. 2025. Imagebite: A framework for evaluating representational harms in text-to-image models. In *4th IEEE/ACM International Conference on AI Engineering - Software Engineering for AI, CAIN 2025, Ottawa, ON, Canada, April 27-28, 2025*, pages 95–106. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. 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 2019, Hong Kong, China, November 3-7, 2019*, pages 3980–3990. Association for Computational Linguistics.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10674–10685. IEEE.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 22500–22510. IEEE.

Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. A picture is worth a thousand words: principled recaptioning improves image generation. arXiv 2310.16656.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 465 others. 2025. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Kirill Tyshchuk, Polina Karpikova, Andrew Spiridonov, Anastasiia Prutianova, Anton Razzhigaev, and Alexander Panchenko. 2023. On isotropy of multimodal embeddings. *Inf.*, 14(7):392.

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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 176–187. Association for Computational Linguistics.

Douglas Neil Walton. 2008. The three bases for the enthymeme: A dialogical theory. *Journal of Applied Logic*, 6(3):361–379.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior Research Methods, Instruments, & Computers*, 20(1):6–10.

Si Wu and David Smith. 2023. Composition and deformance: Measuring imageability with a text-to-image model. In *Proceedings of the 5th Workshop on Narrative Understanding*, pages 106–117, Toronto, Canada. Association for Computational Linguistics.

A Prompts

```
List exactly 5 unique aspects (as single words or short phrases) from the following text.
Format them as a numbered list:

“{text}”

Only output the aspect names. Do not include explanations or sentences.
```

Figure 2: Aspect Generation Prompt to identify aspects out of argumentative claim (used with LLaMA 3.2 3B-Instruct).

```
You are a creative assistant tasked with generating descriptive prompts for image generation.
Given an argument and its aspects, craft a detailed, vivid prompt that combines them into a
single, cohesive description suitable for generating an image. The prompt should be rich in
detail, incorporating the argument and all aspects naturally.

Argument: {argument}
Aspects: {aspects}
```

Figure 3: Image-Prompt Generation Prompt to combine Arguments and Aspects into a descriptive prompt (used with Mistral 7B-v0.1).

B Annotation Examples

Argument	Aspect-aware prompt	Argument as literal prompt	
Violent video games promote aggression.			
Aspect: gaming Aspect: aggression	(2,2) (2,2)	(0,2) (2,2)	(2,2) (2,2)
Cultural traditions often rely on the use of animals.			
Aspect: animal Aspect: tradition	(2,2) (2,2)	(2,2) (2,2)	(1,0) (2,2)
	Stable Diffusion 1.0 XL	Stable Diffusion 1.0 XL	Stable Diffusion 3.5 Medium

Table 6: Examples of image generation for argumentative prompts. The left column lists an argument and the aspects that should be visually depicted. Images are generated either using an aspect-aware prompt or by using the argument itself as the prompt. Below each image, tuples denote the ratings assigned by Annotator 1 and Annotator 2 for the corresponding aspect, on a three-point scale (0 = not depicted, 1 = partially depicted, 2 = clearly depicted). Using the argument as a literal prompt can lead to ambiguous interpretations. For example, the Stable Diffusion 1.0 image for *Violent video games promote aggression* received scores of (0,2) for the *gaming* aspect, since the scene can be interpreted as real-world violence rather than a video game context.